SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 1–29

SEMIDISCRETE FINITE ELEMENT APPROXIMATIONS OF A LINEAR FLUID-STRUCTURE INTERACTION PROBLEM*

Q. DU†, M. D. GUNZBURGER‡, L. S. HOU§, AND J. LEE¶

Abstract. Semidiscrete finite element approximations of a linear fluid-structure interaction problem are studied. First, results concerning a divergence-free weak formulation of the interaction problem are reviewed. Next, semidiscrete finite element approximations are defined, and the existence of finite element solutions is proved with the help of an auxiliary, discretely divergence-free formulation. A discrete inf-sup condition is verified, and the existence of a finite element pressure is established. Strong a priori estimates for the finite element solutions are also derived. Then, by passing to the limit in the finite element approximations, the existence of a strong solution is demonstrated and semidiscrete error estimates are obtained.

Key words. fluid-structure interactions, finite element methods, error estimates

AMS subject classifications. 65M60, 76M10, 76D07, 73V05, 73C02

DOI. 10.1137/S0036142903408654

1. Introduction.

Fluid-structure interaction problems have been extensively studied in recent years both analytically and computationally. The book [28] and the special issue [30] give accounts of the state of the art from the engineering points of view. In addition, a short discussion of the literature can be found in [10]. The references in [10] include [4, 18, 29] for fluid-structure interactions involving elementary fluids, [2, 3, 32] for fluid-structure interactions involving inviscid fluids, and [6, 7, 8, 9, 11, 13, 14, 15, 16, 17, 20, 21, 22, 26, 27, 33, 34] for interactions between viscous, incompressible fluids and elastic solids.

In [10], we analyzed a model for the interactions between Stokesian fluids and linear elastic solids. This paper is devoted to the finite element analysis of that model. As in [10], we assume that the fluid and solid occupy two adjacent open Lipschitz domains, $\Omega_1 \subset \mathbb{R}^d$ and $\Omega_2 \subset \mathbb{R}^d$, respectively, where $d = 2$ or 3 is the space dimension. We denote by Ω the entire fluid-solid region under consideration; i.e., Ω is the interior of $\overline{\Omega}_1 \cup \overline{\Omega}_2$. Let $\Gamma_0 = \partial\Omega_1 \cap \partial\Omega_2$ denote the interface between the fluid and solid, and let $\Gamma_1 = \partial\Omega_1 \setminus \Gamma_0$ and $\Gamma_2 = \partial\Omega_2 \setminus \Gamma_0$ denote the parts of the fluid and solid boundaries, respectively, excluding the interface Γ_0. For obvious reasons we assume that $\text{meas}(\Gamma_1 \cup \Gamma_2) \neq 0$.

*Received by the editors May 31, 2002; accepted for publication (in revised form) March 7, 2003; published electronically January 6, 2004.

http://www.siam.org/journals/sinum/42-1/40865.html

†Department of Mathematics, Penn State University, State College, PA 16802 (qdu@math.psu.edu). The work of this author was supported in part by the National Science Foundation under grant DMS-0196522.

‡School of Computational Science and Information Technology, Florida State University, Tallahassee, FL 32306-4120 (gunzburg@csit.fsu.edu). The work of this author was supported in part by the National Science Foundation under grant DMS-9806358.

§Department of Mathematics, Iowa State University, Ames, IA 50011-2064 (hou@math.iastate.edu).

¶Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213 (jeehyunl@andrew.cmu.edu).

In the fluid region Ω_1, we apply the Stokes system

$$\begin{cases} \rho_1 \mathbf{v}_t + \nabla p - \mu_1 \nabla \cdot (\nabla \mathbf{v} + \nabla \mathbf{v}^T) = \rho_1 \mathbf{f}_1 & \text{in } \Omega_1, \\ \nabla \cdot \mathbf{v} = 0 & \text{in } \Omega_1, \\ \mathbf{v} = 0 & \text{on } \Gamma_1, \\ \mathbf{v}|_{t=0} = \mathbf{v}_0 & \text{in } \Omega_1, \end{cases} \tag{1.1}$$

where $\mathbf{v}$ denotes the fluid velocity, p the fluid pressure, $\mathbf{f}_1$ the given body force per unit mass, ρ_1 and μ_1 the constant fluid density and viscosity, and $\mathbf{v}_0$ the given initial velocity.

In the solid region, we apply the equations of linear elasticity

$$\begin{cases} \rho_2 \mathbf{u}_{tt} - \mu_2 \nabla \cdot (\nabla \mathbf{u} + \nabla \mathbf{u}^T) - \lambda_2 \nabla (\nabla \cdot \mathbf{u}) = \rho_2 \mathbf{f}_2 & \text{in } \Omega_2, \\ \mathbf{u} = 0 & \text{on } \Gamma_2, \\ \mathbf{u}|_{t=0} = \mathbf{u}_0 \quad \text{and} \quad \mathbf{u}_t|_{t=0} = \mathbf{u}_1 & \text{in } \Omega_2, \end{cases} \tag{1.2}$$

where $\mathbf{u}$ denotes the displacement of the solid, $\mathbf{f}_2$ the given loading force per unit mass, μ_2 and λ_2 the Lamé constants, ρ_2 the constant solid density, and $\mathbf{u}_0$ and $\mathbf{u}_1$ the given initial data.

Across the *fixed* interface Γ_0 between the fluid and solid, the velocity and stress vector are continuous. Thus, we have

$$\mathbf{u}_t = \mathbf{v} \qquad \text{on } \Gamma_0 \tag{1.3}$$

and

$$\mu_2 (\nabla \mathbf{u} + \nabla \mathbf{u}^T) \cdot \mathbf{n}_2 + \lambda_2 (\nabla \cdot \mathbf{u}) \mathbf{n}_2 = p \mathbf{n}_1 - \mu_1 (\nabla \mathbf{v} + \nabla \mathbf{v}^T) \cdot \mathbf{n}_1 \qquad \text{on } \Gamma_0, \tag{1.4}$$

where $\mathbf{n}_i$ is the outward-pointing unit normal vector along $\partial\Omega_i$, $i = 1, 2$.

The physical validity of the model (1.1)–(1.4) was explained in [10]. Previous work concerning this model include, as cited in [10], eigenmode analysis [34], homogenization [8], the one-dimensional case [11], and a numerical algorithm [13]. In [10], weak formulations for (1.1)–(1.4) were defined, and the existence of weak solutions was established. The proof for the existence result was based on Galerkin approximations using divergence-free basis functions, and the pressure term was absent in the Galerkin approximations.

The objective of this paper is to define semidiscrete finite element approximations, prove the convergence of finite element solutions, and derive error estimates for the finite element approximations. We point out that finite element basis functions in general are not divergence-free, and finite element formulations must be studied with the pressure term. The proof for the convergence of finite element solutions provides an alternative proof to that found in [10] for the existence of a weak solution; the results of this paper do not rely on those of [10] concerning the existence of a divergence-free weak solution. Moreover, the regularity and compatibility assumptions made on the data in this paper lead to a stronger solution. The details for the divergence-free Galerkin approximations of [10] and the discretely divergence-free finite element approximations are sufficiently different so that separate treatments are warranted.

A few technical aspects contained in this paper are particularly noteworthy: the finite element initial conditions are defined asymmetrically about the two subdomains Ω_1 and Ω_2; two inf-sup conditions are verified that facilitate the analysis of certain steady-state saddle point problems (these inf-sup conditions are also useful in dealing with approximations of mixed boundary value problems for the Stokes equations); and error estimates for a weighted L^2 projection onto discretely divergence-free spaces are derived.

The plan of the paper is as follows. In section 2, we recall relevant results of [10], in particular the weak formulations and the existence theorems. In section 3, we define semidiscrete finite element approximations and establish the existence of and a priori estimates for the finite element solutions. In section 4, we show the convergence of finite element solutions and derive error estimates.

2. Notations and results concerning divergence-free weak formulations. In this section we will recall the notation, weak formulations, and existence results of [10].

Throughout this paper, C denotes a positive constant, depending on the domains Ω, Ω_1, and Ω_2, whose meaning and value changes with context. $H^s(\mathcal{D})$, $s \in \mathbb{R}$, denotes the standard Sobolev space of order s with respect to the set $\mathcal{D}$ equipped with the standard norm $\|\cdot\|_{s,\mathcal{D}}$. Vector-valued Sobolev spaces are denoted by $\mathbf{H}^s(\mathcal{D})$, with norms still denoted by $\|\cdot\|_{s,\mathcal{D}}$. $H_0^1(\mathcal{D})$ denotes the space of functions belonging to $H^1(\mathcal{D})$ that vanish on the boundary $\partial\mathcal{D}$ of $\mathcal{D}$; $\mathbf{H}_0^1(\mathcal{D})$ denotes the vector-valued counterpart.

We will use the following L^2 inner product notations on scalar and vector-valued L^2 spaces:

$$[p,q]_{\mathcal{D}} = \int_{\mathcal{D}} pq\, d\mathcal{D} \quad \forall\, p,q \in L^2(\mathcal{D}), \qquad [\mathbf{u},\mathbf{v}]_{\mathcal{D}} = \int_{\mathcal{D}} \mathbf{u}\cdot\mathbf{v}\, d\mathcal{D} \quad \forall\, \mathbf{u},\mathbf{v} \in \mathbf{L}^2(\mathcal{D}),$$

where the spatial set $\mathcal{D}$ is Ω or Γ_0 or Ω_i, for $i = 1, 2$.

We introduce the function spaces

$$X_i = [\mathbf{H}_0^1(\Omega)]|_{\Omega_i} \quad \text{with the norm } \|\cdot\|_{X_i} = \|\cdot\|_{1,\Omega_i}, \quad i = 1,2,$$

and

$$\boldsymbol{\Psi} = \{\boldsymbol{\eta} \in \mathbf{H}_0^1(\Omega) : \operatorname{div}\boldsymbol{\eta} = 0 \text{ in } \Omega_1\} \quad \text{with the norm } \|\cdot\|_{1,\Omega}.$$

We define the weighted $\mathbf{L}^2(\Omega)$ inner product $[[\cdot,\cdot]]$ by

$$[[\boldsymbol{\xi},\boldsymbol{\eta}]] = [\rho_1\boldsymbol{\xi},\boldsymbol{\eta}]_{\Omega_1} + [\rho_2\boldsymbol{\xi},\boldsymbol{\eta}]_{\Omega_2} \qquad \forall\, \boldsymbol{\xi},\boldsymbol{\eta} \in \mathbf{L}^2(\Omega). \tag{2.1}$$

We denote by $\langle\langle\cdot,\cdot\rangle\rangle$ the duality pairing between $\boldsymbol{\Psi}^*$ and $\boldsymbol{\Psi}$ that is generated from the weighted $\mathbf{L}^2(\Omega)$ inner product $[[\cdot,\cdot]]$. The norm on the dual space $\boldsymbol{\Psi}^*$ is defined in the conventional manner:

$$\|\mathbf{g}\|_{\boldsymbol{\Psi}^*} = \sup_{\boldsymbol{\eta}\in\boldsymbol{\Psi},\, \|\boldsymbol{\eta}\|_{1,\Omega}\le 1} |\langle\langle \mathbf{g},\boldsymbol{\eta}\rangle\rangle| \qquad \forall\, \mathbf{g} \in \boldsymbol{\Psi}^*.$$

We define the bilinear forms

$$a_1[\mathbf{u},\mathbf{v}] = 2\int_{\Omega_1} \mu_1(\nabla\mathbf{u}+\nabla\mathbf{u}^T) : (\nabla\mathbf{v}+\nabla\mathbf{v}^T)\, d\Omega \qquad \forall\, \mathbf{u},\mathbf{v}\in X_1,$$

$$a_2[\mathbf{u},\mathbf{v}] = \int_{\Omega_2} \Big\{2\mu_2(\nabla\mathbf{u}+\nabla\mathbf{u}^T) : (\nabla\mathbf{v}+\nabla\mathbf{v}^T) + \lambda_2(\nabla\cdot\mathbf{u})(\nabla\cdot\mathbf{v})\Big\}\, d\Omega \ \ \forall \mathbf{u},\mathbf{v}\in X_2,$$

$$b[\mathbf{v},q] = -\int_{\Omega_1} q\nabla\cdot\mathbf{v}\, d\Omega \qquad \forall\, \mathbf{v}\in X_1,\ \forall\, q\in L^2(\Omega_1).$$

It can be verified with the help of Korn's inequalities [31, pp. 31, 120] that for $i = 1, 2$,

$$a_i[\boldsymbol{\eta},\boldsymbol{\eta}] \geq k_i\|\boldsymbol{\eta}\|^2_{1,\Omega_i} \quad \forall\, \boldsymbol{\eta}\in X_i \qquad \text{if meas } (\Gamma_i)\neq 0 \tag{2.2}$$

and

$$[\boldsymbol{\eta},\boldsymbol{\eta}]_{\Omega_i} + a_i[\boldsymbol{\eta},\boldsymbol{\eta}] \geq k_i\|\boldsymbol{\eta}\|^2_{1,\Omega_i} \quad \forall\, \boldsymbol{\eta}\in X_i \qquad \text{if meas } (\Gamma_i) = 0. \tag{2.3}$$

The bounded bilinear form $b[\cdot,\cdot]$ was shown in [10] to satisfy the inf-sup conditions

$$\inf_{q\in L^2(\Omega_1)} \sup_{\boldsymbol{\eta}\in\mathbf{H}^1_0(\Omega)} \frac{b[\boldsymbol{\eta},q]}{\|\boldsymbol{\eta}\|_{1,\Omega}\|q\|_{0,\Omega_1}} \geq k_b \tag{2.4}$$

and

$$\inf_{q\in L^2(\Omega_1)} \sup_{\mathbf{v}\in X_1} \frac{b[\mathbf{v},q]}{\|\mathbf{v}\|_{1,\Omega_1}\|q\|_{0,\Omega_1}} \geq k_b, \tag{2.5}$$

where $k_b > 0$ is a constant.

For functions that also depend on time, we introduce the space $L^2(0,T;X)$ that consists of L^2-integrable functions from $[0,T]$ into the space X and which is equipped with the norm

$$\left(\int_0^t \|f\|^2_X\, dt\right)^{1/2}.$$

Similarly, we introduce the space $C(0,T;X)$ that consists of continuous functions from $[0,T]$ into the space X and which is equipped with the norm

$$\sup_{t\in[0,T]} \|f\|_X.$$

The divergence-free weak formulation for (1.1)–(1.4) was defined in [10] as follows. Given

$$\begin{cases} \mathbf{f}_1\in C([0,T];\mathbf{L}^2(\Omega_1)), \quad \mathbf{f}_2\in C([0,T];\mathbf{L}^2(\Omega_2)), \quad \mathbf{u}_0\in X_2, \\ \mathbf{v}_0\in X_1, \quad \operatorname{div}\mathbf{v}_0 = 0 \text{ in } \Omega_1, \quad \mathbf{u}_1\in X_2, \quad \mathbf{v}_0|_{\Gamma_0} = \mathbf{u}_1|_{\Gamma_0}, \end{cases} \tag{2.6}$$

seek a pair $(\mathbf{v},\mathbf{u})$ such that

$$(\mathbf{v},\mathbf{u})\in L^2(0,T;X_1)\times L^2(0,T;X_2), \qquad \operatorname{div}\mathbf{v} = 0, \tag{2.7}$$

$$\begin{aligned} &\frac{d}{dt}\Big(\rho_1[\mathbf{v},\boldsymbol{\eta}]_{\Omega_1} + \rho_2[\partial_t\mathbf{u},\boldsymbol{\eta}]_{\Omega_2}\Big) + a_1[\mathbf{v},\boldsymbol{\eta}] + a_2[\mathbf{u},\boldsymbol{\eta}] \\ &\qquad = \rho_1[\mathbf{f}_1,\boldsymbol{\eta}]_{\Omega_1} + \rho_2[\mathbf{f}_2,\boldsymbol{\eta}]_{\Omega_2} \qquad \forall\, \boldsymbol{\eta}\in\boldsymbol{\Psi}, \end{aligned} \tag{2.8}$$

$$\mathbf{v}|_{t=0} = \mathbf{v}_0, \qquad \mathbf{u}|_{t=0} = \mathbf{u}_0, \qquad \mathbf{u}_t|_{t=0} = \mathbf{u}_1, \tag{2.9}$$

and

$$\int_0^t \mathbf{v}(s)|_{\Gamma_0}\, ds = \mathbf{u}(t)|_{\Gamma_0} - \mathbf{u}_0|_{\Gamma_0} \quad \text{a.e. } t. \tag{2.10}$$

The "natural" interface condition (1.4) is built into (2.8), and the "essential" interface condition (1.3) is enforced weakly in the sense of (2.10).

By defining

$$\boldsymbol{\xi} = \begin{cases} \mathbf{v} & \text{in } \Omega_1, \\ \mathbf{u}_t & \text{in } \Omega_2, \end{cases} \qquad \boldsymbol{\xi}_0 = \begin{cases} \mathbf{v}_0 & \text{in } \Omega_1, \\ \mathbf{u}_1 & \text{in } \Omega_2, \end{cases} \quad \text{and} \quad \mathbf{f} = \begin{cases} \mathbf{f}_1 & \text{in } \Omega_1, \\ \mathbf{f}_2 & \text{in } \Omega_2, \end{cases} \tag{2.11}$$

(2.7)–(2.10) was conveniently recast in [10] into the following equivalent, auxiliary divergence-free weak formulation: seek a $\boldsymbol{\xi}$ such that

$$\begin{aligned} &\boldsymbol{\xi} \in L^2(0,T;\mathbf{L}^2(\Omega)), \qquad \partial_t \boldsymbol{\xi} \in L^2(0,T;\boldsymbol{\Psi}^*), \\ &\boldsymbol{\xi}|_{\Omega_1} \in L^2(0,T;X_1), \qquad \operatorname{div} \boldsymbol{\xi}|_{\Omega_1} = 0, \qquad \textstyle\int_0^t \boldsymbol{\xi}(s)|_{\Omega_2}\, ds \in L^2(0,T;X_2), \end{aligned} \tag{2.12}$$

$$\langle\langle \boldsymbol{\xi}_t, \boldsymbol{\eta} \rangle\rangle + a_1[\boldsymbol{\xi}, \boldsymbol{\eta}] + a_2\left[\int_0^t \boldsymbol{\xi}(s)\, ds, \boldsymbol{\eta}\right] = [[\mathbf{f}, \boldsymbol{\eta}]] - a_2[\mathbf{u}_0, \boldsymbol{\eta}] \quad \forall\, \boldsymbol{\eta} \in \boldsymbol{\Psi}, \text{ a.e. } t, \tag{2.13}$$

$$\boldsymbol{\xi}(0) = \boldsymbol{\xi}_0, \tag{2.14}$$

and

$$\int_0^t (\boldsymbol{\xi}(s)|_{\Omega_1})\Big|_{\Gamma_0}\, ds = \int_0^t (\boldsymbol{\xi}(s)|_{\Omega_2})\Big|_{\Gamma_0}\, ds \qquad \text{a.e. } t. \tag{2.15}$$

The existence and uniqueness of a solution for the auxiliary problem (2.12)–(2.15) was proved in [10].

THEOREM 2.1. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2$, *and* $\mathbf{u}_0$ *satisfy* (2.6). *Then, there exists a unique solution* $\boldsymbol{\xi}$ *for* (2.12)–(2.15). *Moreover,* $\boldsymbol{\xi}$ *satisfies the estimates*

$$\begin{aligned} &\|\boldsymbol{\xi}(t)\|_{0,\Omega}^2 + \|\boldsymbol{\xi}\|_{L^2(0,T;\mathbf{H}^1(\Omega_1))}^2 + \left\|\int_0^t \boldsymbol{\xi}(s)\, ds\right\|_{\mathbf{H}^1(\Omega_2)}^2 \\ &\quad \le Ce^{CT}(\|\mathbf{f}\|_{L^2(0,T;\mathbf{L}^2(\Omega))}^2 + \|\mathbf{u}_0\|_{1,\Omega_2}^2 + \|\mathbf{v}_0\|_{1,\Omega_1}^2 + \|\mathbf{u}_1\|_{1,\Omega_2}^2) \quad \forall\, t \in [0,T] \end{aligned} \tag{2.16}$$

and

$$\begin{aligned} &\|\partial_t \boldsymbol{\xi}\|_{L^2(0,T;\boldsymbol{\Psi}^*)}^2 \\ &\quad \le Ce^{CT}(\|\mathbf{f}\|_{L^2(0,T;\mathbf{L}^2(\Omega))}^2 + \|\mathbf{u}_0\|_{1,\Omega_2}^2 + \|\mathbf{v}_0\|_{1,\Omega_1}^2 + \|\mathbf{u}_1\|_{1,\Omega_2}^2). \end{aligned} \tag{2.17}$$

Using relation (2.11) reversely, i.e., setting $\mathbf{v} = \boldsymbol{\xi}|_{\Omega_1}$ and $\mathbf{u} = \mathbf{u}_0 + \int_0^t \boldsymbol{\xi}(s)|_{\Omega_2}\, ds$, Theorem 2.1 immediately yields the following existence result for (2.7)–(2.10).

THEOREM 2.2. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6). *Then, there exists a unique solution* $(\mathbf{v}, \mathbf{u}) \in L^2(0,T;X_1) \times L^2(0,T;X_2)$ *for* (2.7)–(2.10), *where* (2.8) *holds in the sense of distributions on* $(0,T)$. *Moreover,*

(2.18)
$$\|\mathbf{v}(t)\|^2_{0,\Omega_1} + \|\mathbf{u}_t(t)\|^2_{0,\Omega_2} + \|\mathbf{v}\|^2_{L^2(0,T;\mathbf{H}^1(\Omega_1))} + \|\mathbf{u}(t)\|^2_{\mathbf{H}^1(\Omega_2)}$$
$$\leq Ce^{CT}(\|\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{1,\Omega_2} + \|\mathbf{v}_0\|^2_{0,\Omega_1} + \|\mathbf{u}_1\|^2_{0,\Omega_2}) \quad \forall\, t \in [0,T].$$

The existence of a stronger solution and an L^2-integrable pressure was also established in [10].

THEOREM 2.3. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6) *and*

$$\partial_t \mathbf{f}_i \in L^2(0,T;\mathbf{L}^2(\Omega_i)),\ i = 1,2, \quad \mathbf{v}_0 \in \mathbf{H}^2(\Omega_1), \quad \mathbf{u}_1 \in \mathbf{H}^1(\Omega_2), \quad \mathbf{u}_0 \in \mathbf{H}^2(\Omega_2).$$

Assume further that there exists a $p_0 \in H^1(\Omega_1)$ *such that*

$$(p_0\mathbf{n}_1 - \mu_1(\nabla \mathbf{v}_0 + \nabla \mathbf{v}_0^T)\cdot \mathbf{n}_1)|_{\Gamma_0} = (\mu_2(\nabla \mathbf{u}_0 + \nabla \mathbf{u}_0^T)\cdot \mathbf{n}_2 + (\lambda_2+\mu_2)(\operatorname{div} \mathbf{u}_0)\mathbf{n}_2)|_{\Gamma_0},$$

where $\mathbf{n}_i$ *denotes the outward-pointing normal along* $\partial\Omega_i$. *Then, the solution* $(\mathbf{v}, \mathbf{u})$ *to* (2.7)–(2.10) *satisfies*

$$\mathbf{v} \in L^\infty(0,T;\mathbf{L}^2(\Omega_1)) \cap L^2(0,T;X_1), \quad \mathbf{u} \in L^\infty(0,T;X_2),$$

$$\mathbf{v}_t \in L^\infty(0,T;\mathbf{L}^2(\Omega_1)) \cap L^2(0,T;X_1), \quad \mathbf{u}_t \in L^\infty(0,T;X_2), \quad \mathbf{u}_{tt} \in L^\infty(0,T;\mathbf{L}^2(\Omega_2)),$$

and

$$\|\partial_t \mathbf{v}(t)\|^2_{0,\Omega_1} + \|\partial_{tt}\mathbf{u}(t)\|^2_{0,\Omega_2} + \|\partial_t \mathbf{v}\|^2_{L^2(0,T;X_1)} + \|\partial_t \mathbf{u}(t)\|^2_{1,\Omega_2)}$$
$$\leq Ce^{CT}(\|\mathbf{f}\|^2_{H^1(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{2,\Omega_2} + \|\mathbf{v}_0\|^2_{2,\Omega_1} + \|p_0\|^2_{1,\Omega_1} + \|\mathbf{u}_1\|^2_{1,\Omega_2})\ \forall t \in [0,T].$$

Furthermore, there exists a unique $p \in L^2(0,T;L^2(\Omega_1))$ *such that*

(2.19)
$$\rho_1[\mathbf{v}_t, \boldsymbol{\eta}]_{\Omega_1} + b[\boldsymbol{\eta}, p] + a_1[\mathbf{v}, \boldsymbol{\eta}] + \rho_2[\mathbf{u}_{tt}, \boldsymbol{\eta}]_{\Omega_2} + a_2[\mathbf{u}, \boldsymbol{\eta}]$$
$$= \rho_1[\mathbf{f}_1, \boldsymbol{\eta}]_{\Omega_1} + \rho_2[\mathbf{f}_2, \boldsymbol{\eta}]_{\Omega_2} \qquad \forall\, \boldsymbol{\eta} \in \mathbf{H}^1_0(\Omega),\ a.e.\ t$$

and

$$\|p\|_{L^2(0,T;L^2(\Omega_1))} \leq Ce^{CT}(\|\mathbf{f}\|^2_{H^1(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{2,\Omega_2} + \|\mathbf{v}_0\|^2_{2,\Omega_1} + \|p_0\|^2_{1,\Omega_1} + \|\mathbf{u}_1\|^2_{1,\Omega_2}).$$

3. Semidiscrete finite element approximations. In this section we will define semidiscrete finite element approximations, prove the existence of finite element solutions on discretely divergence-free spaces and derive energy estimates, and establish the existence of a discrete pressure by verifying inf-sup conditions for finite element space pairs.

As alluded to previously, finite element solutions in general are not divergence-free, and finite element formulations should include the pressure term. Of course, the corresponding continuous weak formulation should also contain the pressure term. Such a weak formulation requires additional regularity on $\mathbf{v}_t$ and $\mathbf{u}_{tt}$. The continuous weak formulation we consider is as follows: given $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2$, and $\mathbf{u}_0$ satisfying (2.6), seek a triplet $(\mathbf{v}, p, \mathbf{u})$ such that

(3.1)
$$(\mathbf{v}, p, \mathbf{u}) \in L^2(0,T;X_1) \times L^2(0,T;L^2(\Omega_1)) \times L^2(0,T;X_2),$$

$$\mathbf{v}_t \in L^2(0,T;\mathbf{L}^2(\Omega_1)), \qquad \mathbf{u}_t \in L^2(0,T;X_2), \qquad \mathbf{u}_{tt} \in L^2(0,T;\mathbf{L}^2(\Omega_2)), \tag{3.2}$$

$$\begin{aligned} &\rho_1[\mathbf{v}_t,\boldsymbol{\eta}]_{\Omega_1} + b[\boldsymbol{\eta},p] + a_1[\mathbf{v},\boldsymbol{\eta}] + \rho_2[\mathbf{u}_{tt},\boldsymbol{\eta}]_{\Omega_2} + a_2[\mathbf{u},\boldsymbol{\eta}] \\ &\quad = \rho_1[\mathbf{f}_1,\boldsymbol{\eta}]_{\Omega_1} + \rho_2[\mathbf{f}_2,\boldsymbol{\eta}]_{\Omega_2} \qquad \forall\, \boldsymbol{\eta} \in \mathbf{H}_0^1(\Omega),\ \text{a.e. } t \in [0,T], \end{aligned} \tag{3.3}$$

$$b[\mathbf{v},q] = 0 \qquad \forall\, q \in L^2(\Omega_1),\ \text{a.e. } t \in [0,T], \tag{3.4}$$

$$\mathbf{v}|_{t=0} = \mathbf{v}_0, \qquad \mathbf{u}|_{t=0} = \mathbf{u}_0, \qquad \mathbf{u}_t|_{t=0} = \mathbf{u}_1, \tag{3.5}$$

$$\mathbf{v}|_{\Gamma_0} = \mathbf{u}_t|_{\Gamma_0} \quad \text{a.e. } t. \tag{3.6}$$

We will define finite element approximations to (3.3)–(3.6). By showing the convergence of finite element solutions, we establish the existence of a solution for (3.1)–(3.6). For reasons connected with the derivation of the regularity results (3.2), we will define finite element initial conditions in a nonstandard manner.

3.1. Finite element discretization. In what follows we assume that Ω_1 and Ω_2 are two-dimensional polygons or three-dimensional polyhedra. Let h denote a discretization parameter associated with the triangulation $\mathcal{T}^h(\Omega)$ of Ω. We assume that elements of $\mathcal{T}^h$ do not cross the interface Γ_0. We assume that the triangulation $\mathcal{T}^h$ consists of triangular elements in two dimensions or tetrahedral elements in three dimensions, though our results can be extended to other types of triangulations. Furthermore, we assume that there exists a triangulation $\mathcal{T}^{h_0}(\Omega)$ such that, for each $h < h_0$, $\mathcal{T}^h(\Omega)$ is a refinement of $\mathcal{T}^{h_0}(\Omega)$.

For each h, we choose $X^h \subset \mathbf{C}(\overline{\Omega}) \cap \mathbf{H}_0^1(\Omega)$ and $Q_1^h \subset L^2(\Omega_1)$ as finite element subspaces over the triangulation $\mathcal{T}^h(\Omega)$. We assume that X^h contains piecewise linear functions. We set

$$X_i^h = X^h|_{\Omega_i}, \quad i = 1,2,$$

and

$$\boldsymbol{\Psi}^h = \{\boldsymbol{\eta}_h \in X^h : b[\boldsymbol{\eta}_h, q_h] = 0\ \forall\, q_h \in Q_1^h\}.$$

We assume that the finite element spaces X_1^h, X_2^h, and Q_1^h satisfy the standard approximation properties [5]; i.e., there exist an integer $k > 0$ and constant $C > 0$ such that

$$\inf_{\mathbf{v}^h \in X_i^h} \|\mathbf{v} - \mathbf{v}^h\|_{0,\Omega_i} \le Ch^{r+1}\|\mathbf{v}\|_{r+1,\Omega_i} \qquad \forall\, \mathbf{v} \in \mathbf{H}^{r+1}(\Omega_i) \cap X_i,\ r \in [0,k], \tag{3.7}$$

$$\inf_{\mathbf{v}^h \in X_i^h} \|\mathbf{v} - \mathbf{v}^h\|_{1,\Omega_i} \le Ch^{r}\|\mathbf{v}\|_{r+1,\Omega_i} \qquad \forall\, \mathbf{v} \in \mathbf{H}^{r+1}(\Omega_i) \cap X_i,\ r \in [0,k], \tag{3.8}$$

and

$$\inf_{q^h \in Q_1^h} \|q - q^h\|_{0,\Omega_1} \le Ch^{r}\|p\|_{r,\Omega_1} \qquad \forall\, q \in H^r(\Omega_1),\ r \in [0,k]. \tag{3.9}$$

Also, X^h satisfies the approximation properties

$$\inf_{\boldsymbol{\eta}^h \in X^h} \|\boldsymbol{\eta} - \boldsymbol{\eta}^h\|_{0,\Omega} \le Ch^{r+1}\|\boldsymbol{\eta}\|_{r+1,\Omega} \qquad \forall\, \boldsymbol{\eta} \in \mathbf{H}^{r+1}(\Omega) \cap \mathbf{L}^2(\Omega),\ r \in [0,k], \tag{3.10}$$

and

$$\inf_{\boldsymbol{\eta}^h \in X^h} \|\boldsymbol{\eta} - \boldsymbol{\eta}^h\|_{1,\Omega_i} \le Ch^{r}\|\boldsymbol{\eta}\|_{r+1,\Omega} \qquad \forall\, \boldsymbol{\eta} \in \mathbf{H}^{r+1}(\Omega) \cap \mathbf{H}_0^1(\Omega),\ r \in [0,k]. \tag{3.11}$$

We assume that the finite element pair $\{\widetilde{X}_1^h, M^h\} \equiv \{X_1^h \cap \mathbf{H}_0^1(\Omega_1), Q_1^h \cap L_0^2(\Omega_1)\}$ satisfies the discrete inf-sup condition

$$\inf_{q^h \in M^h(\Omega_1)} \sup_{\mathbf{v}^h \in \widetilde{X}^h(\Omega_1)} \frac{b[\mathbf{v}^h, q^h]}{\|\mathbf{v}^h\|_{1,\Omega_1}\|q^h\|_{0,\Omega_1}} \ge C. \tag{3.12}$$

Choices of finite element spaces satisfying (3.12) are well known [19]. Note that functions in $\widetilde{X}_1^h$ vanish on Γ_0.

We also assume that triangulations are uniformly regular so that the following inverse inequalities hold:

$$\begin{aligned} &\|\mathbf{v}^h\|_{1,\Omega} \le Ch^{-1}\|\mathbf{v}^h\|_{0,\Omega} \quad \forall\, \mathbf{v}^h \in X^h; \\ &\|\mathbf{v}^h\|_{1,\Omega_i} \le Ch^{-1}\|\mathbf{v}^h\|_{0,\Omega_i} \quad \forall\, \mathbf{v}^h \in X_i^h,\ i = 1,2. \end{aligned} \tag{3.13}$$

Semidiscrete finite element approximations of the weak form (3.3)–(3.6) are defined as follows: seek $(\mathbf{v}^h, p^h, \mathbf{u}^h) \in C^1([0,T]; X_1^h) \times C([0,T]; Q_1^h) \times C^1([0,T]; X_2^h)$ such that

$$\begin{aligned} &\rho_1[\partial_t \mathbf{v}_h, \boldsymbol{\eta}_h]_{\Omega_1} + b[\boldsymbol{\eta}_h, p_h] + a_1[\mathbf{v}_h, \boldsymbol{\eta}_h] + \rho_2[\partial_{tt}\mathbf{u}_h, \boldsymbol{\eta}_h]_{\Omega_2} + a_2[\mathbf{u}_h, \boldsymbol{\eta}_h] \\ &\quad = \rho_1[\mathbf{f}_1, \boldsymbol{\eta}_h]_{\Omega_1} + \rho_2[\mathbf{f}_2, \boldsymbol{\eta}_h]_{\Omega_2} \qquad \forall\, \boldsymbol{\eta}_h \in X^h,\ \text{a.e. } t, \end{aligned} \tag{3.14}$$

$$b[\mathbf{v}_h, q_h] = 0 \qquad \forall\, q_h \in Q_1^h,\ \text{a.e. } t, \tag{3.15}$$

$$\mathbf{v}_h|_{\Gamma_0} = \partial_t \mathbf{u}_h|_{\Gamma_0} \qquad \text{a.e. } t \in [0,T], \tag{3.16}$$

$$\mathbf{v}_h|_{t=0} = \mathbf{v}_{0,h}, \qquad \mathbf{u}_h|_{t=0} = \mathbf{u}_{0,h}, \qquad \partial_t \mathbf{u}_h|_{t=0} = \mathbf{u}_{1,h}, \tag{3.17}$$

where $\mathbf{v}_{0,h} \in \boldsymbol{\Psi}^h|_{\Omega_1}$, $\mathbf{u}_{0,h} \in X_2^h$, and $\mathbf{u}_{1,h} \in X_2^h$ are finite element approximations of $\mathbf{v}_0$, $\mathbf{u}_0$, and $\mathbf{u}_1$, respectively. We assume that $(\mathbf{v}_{0,h}, \mathbf{u}_{1,h})$ satisfies

$$b[\mathbf{v}_{0,h}, q_h] = 0 \quad \forall\, q_h \in Q_1^h, \qquad \mathbf{v}_{0,h}|_{\Gamma_0} = \mathbf{u}_{1,h}|_{\Gamma_0} \tag{3.18}$$

and that $\mathbf{u}_{0,h}$ is defined by

$$a_2[\mathbf{u}_{0,h}, \mathbf{w}_h] = a_2[\mathbf{u}_0, \mathbf{w}_h] \qquad \forall\, \mathbf{w}_h \in X_2^h. \tag{3.19}$$

3.2. The existence of discretely divergence-free finite element solutions. The existence of finite element solutions $\{(\mathbf{v}^h, \mathbf{u}^h)\}$ can be established in a manner analogous to the analysis of the Galerkin approximations $\{(\mathbf{v}_m, \mathbf{u}_m)\}$ in [10]. However, it should be noted that finite element approximations are not special cases of the Galerkin approximations due to the fact that the basis functions used in the Galerkin approximations are divergence-free in Ω_1, whereas the finite element solutions are only discretely divergence-free in Ω_1 in the sense of (3.15), i.e., they belong to the space of discretely divergence-free functions $\boldsymbol{\Psi}^h$.

We first formulate auxiliary semidiscrete finite element approximations on the discretely divergence-free space $\boldsymbol{\Psi}^h$. Through the relation

$$\boldsymbol{\xi}_h = \begin{cases} \mathbf{v}_h & \text{in } \Omega_1, \\ \partial_t \mathbf{u}_h & \text{in } \Omega_2, \end{cases} \tag{3.20}$$

we see that (3.14)–(3.19) can be recast into the system

$$\begin{aligned} &\rho_1[\partial_t \boldsymbol{\xi}_h, \boldsymbol{\eta}_h]_{\Omega_1} + \rho_2[\partial_t \boldsymbol{\xi}_h, \boldsymbol{\eta}_h]_{\Omega_2} + a_1[\boldsymbol{\xi}_h, \boldsymbol{\eta}_h] + a_2\left[\int_0^t \boldsymbol{\xi}_h(s)\, ds, \boldsymbol{\eta}_h\right] \\ &\quad = \rho_1[\mathbf{f}_1, \boldsymbol{\eta}_h]_{\Omega_1} + \rho_2[\mathbf{f}_2, \boldsymbol{\eta}_h]_{\Omega_2} - a_2[\mathbf{u}_0, \boldsymbol{\eta}_h] \qquad \forall\, \boldsymbol{\eta}_h \in \boldsymbol{\Psi}^h,\ t \in [0, T] \end{aligned} \tag{3.21}$$

and

$$\boldsymbol{\xi}_h(0) = \boldsymbol{\xi}_{0,h} \equiv \begin{cases} \mathbf{v}_{0,h} & \text{in } \Omega_1, \\ \mathbf{u}_{1,h} & \text{in } \Omega_2. \end{cases} \tag{3.22}$$

Let $\{\boldsymbol{\psi}_j^h\}_{j=1}^{J_h}$ be a finite element basis for $\boldsymbol{\Psi}^h$. Assumption (3.18) implies that $\boldsymbol{\xi}_{0,h} \in \boldsymbol{\Psi}^h$, so that we can write

$$\boldsymbol{\xi}_{0,h} = \sum_{j=1}^{J_h} d_j \boldsymbol{\psi}_j^h.$$

The solution $\boldsymbol{\xi}^h \in C([0, T]; \boldsymbol{\Psi}^h)$ for (3.21)–(3.22) can be expressed in the form

$$\boldsymbol{\xi}_h = \sum_{j=1}^{J_h} g_j^h(t) \boldsymbol{\psi}_j^h(\mathbf{x}) \tag{3.23}$$

so that system (3.21)–(3.22) is equivalent to the following linear system of ordinary differential equations for $\{g_j^h\}_{j=1}^{J_h}$:

$$\begin{cases} \displaystyle\sum_{j=1}^{J_h} [[\boldsymbol{\psi}_j^h, \boldsymbol{\psi}_i^h]] \frac{d}{dt} g_j^h(t) + \sum_{j=1}^{J_h} a_1[\boldsymbol{\psi}_j^h, \boldsymbol{\psi}_i^h]\, g_j^h(t) + \sum_{j=1}^{J_h} a_2[\boldsymbol{\psi}_j^h, \boldsymbol{\psi}_i^h] \int_0^t g_j^h(s)\, ds \\ \qquad = [\rho_1 \mathbf{f}_1(t), \boldsymbol{\psi}_i^h]_{\Omega_1} + [\rho_2 \mathbf{f}_2(t), \boldsymbol{\psi}_i^h]_{\Omega_2} - a_2[\mathbf{u}_0, \boldsymbol{\psi}_i^h], \qquad i = 1, \ldots, J_h,\ t \in [0, T], \\ g_i^h(0) = d_i, \qquad i = 1, \ldots, J_h. \end{cases}$$

We have the following results concerning the existence of and a priori estimates for a finite element solution $\boldsymbol{\xi}_h$ of (3.21)–(3.22). The proof is the same as that in [10] for the Galerkin approximations and thus is omitted.

THEOREM 3.1. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6). *Then, there exists a unique function* $\boldsymbol{\xi}_h \in C^1([0,T];\boldsymbol{\Psi}^h)$ *which satisfies* (3.21)–(3.22) *and the estimate*

$$\begin{aligned}(3.24)\quad &\|\boldsymbol{\xi}_h(t)\|^2_{0,\Omega} + \|\boldsymbol{\xi}_h\|^2_{L^2(0,T;\mathbf{H}^1(\Omega_1))} + \left\|\int_0^t \boldsymbol{\xi}_h(s)\,ds\right\|^2_{\mathbf{H}^1(\Omega_2)} \\ &\le Ce^{CT}(\|\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{1,\Omega_2} + \|\mathbf{v}_{0,h}\|^2_{0,\Omega_1} + \|\mathbf{u}_{1,h}\|^2_{0,\Omega_2})\,\forall t \in [0,T].\end{aligned}$$

Setting $\mathbf{v}_h = \boldsymbol{\xi}_h|_{\Omega_1}$, $\mathbf{u}_h = \mathbf{u}_{0,h} + \int_0^t \boldsymbol{\xi}_h(s)|_{\Omega_2}\,ds$ and using (3.19), we immediately obtain the existence of a $(\mathbf{v}_h, \mathbf{u}_h)$ satisfying the discretely divergence-free version of (3.14)–(3.19), as follows.

THEOREM 3.2. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6). *Then, there exists a unique* $(\mathbf{v}_h, \mathbf{u}_h) \in C^1([0,T];\boldsymbol{\Psi}^h|_{\Omega_1}) \times C^2([0,T];X_2)$ *satisfying*

$$\begin{aligned}(3.25)\quad &\rho_1[\partial_t \mathbf{v}_h, \boldsymbol{\eta}_h]_{\Omega_1} + a_1[\mathbf{v}, \boldsymbol{\eta}_h] + \rho_2[\partial_{tt}\mathbf{u}_h, \boldsymbol{\eta}_h]_{\Omega_2} + a_2[\mathbf{u}_h, \boldsymbol{\eta}_h] \\ &\qquad = \rho_1[\mathbf{f}_1, \boldsymbol{\eta}_h]_{\Omega_1} + \rho_2[\mathbf{f}_2, \boldsymbol{\eta}_h]_{\Omega_2} \qquad \forall\, \boldsymbol{\eta}_h \in \boldsymbol{\Psi}^h,\ t \in [0,T]\end{aligned}$$

and (3.15)–(3.19). *Moreover, the following estimate holds:*

$$\begin{aligned}(3.26)\quad &\|\mathbf{v}_h(t)\|^2_{0,\Omega_1} + \|\partial_t \mathbf{u}_h(t)\|^2_{0,\Omega_2} + \|\mathbf{v}_h\|^2_{L^2(0,T;\mathbf{H}^1(\Omega_1))} + \|\mathbf{u}_h\|^2_{\mathbf{H}^1(\Omega_2)} \\ &\le Ce^{CT}(\|\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{1,\Omega_2} + \|\mathbf{v}_{0,h}\|^2_{0,\Omega_1} + \|\mathbf{u}_{1,h}\|^2_{0,\Omega_2})\ \forall t \in [0,T].\end{aligned}$$

3.3. The discrete inf-sup conditions and discrete pressure fields. We have proved the existence of a finite element solution in the discretely divergence-free formulation consisting of (3.25) and (3.15)–(3.19). We will show the existence of a discrete pressure p_h such that (3.14) holds. A crucial step towards this goal is the verification of discrete inf-sup conditions. The discrete inf-sup conditions will also play a role in deriving strong energy estimates in a subsequent section.

We rewrite (3.14) as

$$\begin{aligned}(3.27)\quad b[\boldsymbol{\eta}_h, p_h] &= -\rho_1[\partial_t \mathbf{v}_h, \boldsymbol{\eta}_h]_{\Omega_1} - a_1[\mathbf{v}, \boldsymbol{\eta}_h] - \rho_2[\partial_{tt}\mathbf{u}_h, \boldsymbol{\eta}_h]_{\Omega_2} - a_2[\mathbf{u}_h, \boldsymbol{\eta}_h] \\ &\quad + \rho_1[\mathbf{f}_1, \boldsymbol{\eta}_h]_{\Omega_1} + \rho_2[\mathbf{f}_2, \boldsymbol{\eta}_h]_{\Omega_2} \qquad \forall\, \boldsymbol{\eta}_h \in X^h,\ t \in [0,T].\end{aligned}$$

In terms of the auxiliary variable $\boldsymbol{\xi}_h$, (3.27) is equivalent to

$$\begin{aligned}(3.28)\quad b[\boldsymbol{\eta}_h, p_h] &= -[[\partial_t \boldsymbol{\xi}_h, \boldsymbol{\eta}_h]] - a_1[\mathbf{v}_h, \boldsymbol{\eta}_h] \\ &\quad - a_2[\mathbf{u}_h, \boldsymbol{\eta}_h] + [[\mathbf{f}, \boldsymbol{\eta}_h]] \qquad \forall\, \boldsymbol{\eta}_h \in X^h,\ \forall\, t \in [0,T].\end{aligned}$$

To show the existence of a $p_h \in C([0,T];Q_1^h)$ satisfying (3.27) or (3.28), we need to verify a discrete inf-sup condition for $b[\cdot,\cdot]$, which will be presented below; this will be the task of this subsection. To derive an estimate for p_h, we need an estimate for $\|\partial_t \boldsymbol{\xi}_h\|_{0,\Omega}$, or $\|\partial_t \mathbf{v}_h\|_{0,\Omega_1}$ and $\|\partial_{tt}\mathbf{u}_h\|_{0,\Omega_2}$; these will be derived in section 3.4.

The inf-sup condition we will verify is

$$(3.29)\qquad \inf_{q_h \in Q_1^h} \sup_{\boldsymbol{\eta}_h \in X^h} \frac{b[\boldsymbol{\eta}_h, q_h]}{\|\boldsymbol{\eta}_h\|_{1,\Omega}\|q_h\|_{0,\Omega_1}} \ge C.$$

This inf-sup condition was proved in [2] for a special choice of X_h and Q_1^h. We will establish (3.29) for the general case under assumption (3.12). To this end, we will

first need the following lemma, and we will need to prove the inf-sup condition

$$\inf_{q_h\in Q_1^h}\sup_{\mathbf{v}_h\in X_1^h}\frac{b[\mathbf{v}_h,q_h]}{\|\mathbf{v}_h\|_{1,\Omega_1}\|q_h\|_{0,\Omega_1}}\geq C. \tag{3.30}$$

LEMMA 3.3. *For each constant d, there exists a piecewise linear function $\mathbf{v}\in X_1^{h_0}$ such that*

$$\int_{\Gamma_0}\mathbf{v}\cdot\mathbf{n}\,d\Gamma=-d,\qquad b[\mathbf{v},d]=|d|^2,\qquad \textit{and}\qquad \|\mathbf{v}\|_{1,\Omega_1}\leq C|d|,$$

where $\mathbf{n}$ denotes the unit outward-pointing normal along $\partial\Omega_1$, and the constant C depends only on the coarse triangulation $\mathcal{T}^{h_0}(\Omega_1)$.

Proof. We give the complete proof for the two-dimensional case and discuss the ideas for the three-dimensional case in an ensuing remark.

We choose from $\mathcal{T}^{h_0}(\Omega)$ a layer of triangles $K\equiv\cup_{j=1}^{J_0}K_j\subset\overline{\Omega}_1$ adjacent to Γ_0, i.e., each K_j has either a side or a vertex on Γ_0. We denote the vertices on $\Gamma_0\cap\partial K$ by $A_j, j=0,1,\ldots,J_0$. We define the C^0, piecewise linear vector function $\mathbf{v}=(v_1,v_2)$ on K as follows:

$$\begin{cases}\mathbf{v}=\mathbf{0}\text{ at points }A_0\text{ and }A_{J_0},\\ \mathbf{v}=\mathbf{0}\text{ at all vertices of }K\text{ belonging to the interior of }\Omega_1,\\ \mathbf{v}\cdot\mathbf{n}_{j-1}=-\overline{d}\text{ and }\mathbf{v}\cdot\mathbf{n}_j=-\overline{d}\text{ at }A_j,\quad j=1,\ldots,J_0-1,\ \mathbf{n}_{j-1}\neq\mathbf{n}_j,\\ \mathbf{v}\cdot\mathbf{n}_{j-1}=-\overline{d}\text{ and }\mathbf{v}\cdot\boldsymbol{\tau}_j=0\text{ at }A_j,\quad j=1,\ldots,J_0-1,\ \mathbf{n}_{j-1}=\mathbf{n}_j,\end{cases}$$

where

$$\overline{d}=d\Bigg/\left(|\overline{A_0A_1}|/2+\sum_{j=2}^{J_0-1}|\overline{A_{j-1}A_j}|+|\overline{A_{J_0-1}A_{J_0}}|/2\right)$$

and $\mathbf{n}_j$ and $\boldsymbol{\tau}_j$ denote the unit, outward-pointing normal and unit tangent vectors, respectively, on $\partial\Omega_1\cap\overline{A_{j-1}A_j}$. Note that $\mathbf{n}_j$ and $\boldsymbol{\tau}_j$ are defined with respect to the *segment* $\overline{A_{j-1}A_j}$ so that they are well defined. Clearly, the values of $v_1(A_j)$ and $v_2(A_j)$ are proportional to $\overline{d}$. We can write

$$v_i(\mathbf{x})=\sum_{j=1}^{J_0-1}v_i(A_j)L_j^{h_0}(\mathbf{x}),\qquad i=1,2,$$

where for each j, $L_j^{h_0}(\mathbf{x})$ is the continuous piecewise linear basis function (the shape function) associated with the vertex A_j. Then,

$$\|v_i\|_{1,K}^2\leq C\sum_{j=1}^{J_0-1}|v_i(A_j)|^2\|L_j^{h_0}\|_{1,K}^2\leq C|\overline{d}|^2\sum_{j=1}^{J_0-1}\|L_j^{h_0}\|_{1,K}^2$$

so that

$$\|\mathbf{v}\|_{1,K}\leq C|\overline{d}|.$$

We extend $\mathbf{v}$ to Ω_1 by zero outside K and denote the extended function still by $\mathbf{v}$. Then we readily have $\mathbf{v} \in X_1^{h_0}$,

$$\|\mathbf{v}\|_{1,\Omega_1} = \|\mathbf{v}\|_{1,K} \le C|\overline{d}| \le C|d|,$$

and

$$\int_{\Gamma_0} \mathbf{v}\cdot\mathbf{n}\, d\Gamma = \sum_{j=1}^{J_0} \int_{\overline{A_{j-1}A_j}} \mathbf{v}\cdot\mathbf{n}\, d\Gamma$$
$$= -\overline{d}\left(|\overline{A_0A_1}|/2 + \sum_{j=2}^{J_0-1} |\overline{A_{j-1}A_j}| + |\overline{A_{J_0-1}A_{J_0}}|/2 \right) = -d.$$

Using Green's theorem and the last equality, we have

$$b[\mathbf{v}, d] = -d\int_{\Omega_1} \nabla\cdot\mathbf{v}\, d\Omega = -d\int_{\Gamma_0} \mathbf{v}\cdot\mathbf{n}\, d\Gamma = d^2. \qquad \square$$

Remark 1. In the three-dimensional case we merely need assume that $[\mathcal{T}^{h_0}(\Omega)]|_{\Gamma_0}$ contains a vertex P_0 shared by exactly three triangles. Indeed, in forming the coarse triangulation $\mathcal{T}^{h_0}(\Omega)$, we may simply choose a partition on a flat piece of Γ_0 to meet this requirement. Then, we define a $\mathbf{v}$ to satisfy $\mathbf{v}\cdot\mathbf{n} = \overline{d}$ and $\mathbf{v}\times\mathbf{n} = \mathbf{0}$ at P_0, and $\mathbf{v} = \mathbf{0}$ at all other vertices, where $\overline{d}$ is a suitable scaling of d.

Next we prove inf-sup condition (3.30) based on the inf-sup assumption (3.12) for the pair $\{\widetilde{X}_1^h, M_1^h\} \equiv \{X_1^h \cap \mathbf{H}_0^1(\Omega_1), Q_1^h \cap L_0^2(\Omega_1)\}$.

THEOREM 3.4. *The pair* $\{X_1^h, Q_1^h\}$ *satisfies inf-sup condition* (3.30).

Proof. Owing to [19, Remark 1.4, p. 118], the inf-sup condition (3.30) is equivalent to

$$\begin{aligned} &\forall\, q_h \in Q_1^h, \quad \text{there exists} \quad \mathbf{v}_h \in X_1^h \quad \text{such that} \\ &\quad b[\mathbf{v}_h, q_h] \ge C\|q_h\|_{0,\Omega_1}^2 \quad \text{and} \quad \|\mathbf{v}_h\|_{1,\Omega_1} \le C\|q_h\|_{0,\Omega_1}. \end{aligned} \tag{3.31}$$

Let $q_h \in Q_1^h$ be given. Set

$$\overline{q}_h = \frac{1}{|\Omega_1|}\int_{\Omega_1} q_h\, d\Omega \qquad \text{and} \qquad \widetilde{q}_h = q_h - \overline{q}_h.$$

Then $q_h = \widetilde{q}_h + \overline{q}_h$ in Ω_1 and $\|q_h\|_{0,\Omega_1}^2 = \|\widetilde{q}_h\|_{0,\Omega_1}^2 + \|\overline{q}_h\|_{0,\Omega_1}^2$. Obviously, $\widetilde{q}_h \in M_1^h \equiv Q_1^h \cap L_0^2(\Omega_1)$ so that, by inf-sup condition (3.12) for the pair $\{\widetilde{X}_1^h, M_1^h\}$, we may choose a $\widetilde{\mathbf{v}}_h \in \widetilde{X}_1^h$ such that

$$b[\widetilde{\mathbf{v}}_h, \widetilde{q}_h] = \|\widetilde{q}_h\|_{0,\Omega_1}^2 \qquad \text{and} \qquad \|\widetilde{\mathbf{v}}_h\|_{1,\Omega_1} \le C\|\widetilde{q}_h\|_{0,\Omega_1}.$$

By Lemma 3.3 with $d = \|\overline{q}_h\|_{0,\Omega_1}$, we may choose a $\overline{\mathbf{v}}_h \in X_1^h$ such that

$$b[\overline{\mathbf{v}}_h, \overline{q}_h] = \|\overline{q}_h\|_{0,\Omega_1}^2 \qquad \text{and} \qquad \|\overline{\mathbf{v}}_h\|_{1,\Omega_1} \le C\|\overline{q}_h\|_{0,\Omega_1}.$$

(We recall that we assumed that $\mathcal{T}^h(\Omega_1)$ is a refinement of a coarse triangulation $\mathcal{T}^{h_0}(\Omega_1)$ so that a piecewise linear function on $\mathcal{T}^{h_0}(\Omega_1)$ belongs to X_1^h.) Setting

$\mathbf{v}_h = \widetilde{\mathbf{v}}_h + \alpha\overline{\mathbf{v}}_h$ for some $\alpha > 0$ (to be determined), we have

$$\begin{aligned} b[\mathbf{v}_h, q_h] &= b[\widetilde{\mathbf{v}}_h, \widetilde{q}_h] + b[\widetilde{\mathbf{v}}_h, \overline{q}_h] + \alpha b[\overline{\mathbf{v}}_h, \widetilde{q}_h] + \alpha b[\overline{\mathbf{v}}_h, \overline{q}_h] \\ &\geq \|\widetilde{q}_h\|^2_{0,\Omega_1} + 0 - C\alpha\|\widetilde{q}_h\|_{0,\Omega_1}\|\overline{\mathbf{v}}_h\|_{1,\Omega_1} + \alpha\|\overline{q}_h\|^2_{0,\Omega_1} \\ &\geq \|\widetilde{q}_h\|^2_{0,\Omega_1} - C\alpha\|\widetilde{q}_h\|_{0,\Omega_1}\|\overline{q}_h\|_{0,\Omega_1} + \alpha\|\overline{q}_h\|^2_{0,\Omega_1} \\ &\geq \|\widetilde{q}_h\|^2_{0,\Omega_1} - [C\alpha\|\widetilde{q}_h\|^2_{0,\Omega_1} + \frac{\alpha}{2}\|\overline{q}_h\|^2_{0,\Omega_1}] + \alpha\|\overline{q}_h\|^2_{0,\Omega_1} \\ &= (1 - C\alpha)\|\widetilde{q}_h\|^2_{0,\Omega_1} + \frac{\alpha}{2}\|\overline{q}_h\|^2_{0,\Omega_1} \end{aligned}$$

so that by choosing a sufficiently small $\alpha > 0$ we obtain

$$b[\mathbf{v}_h, q_h] \geq \min\{1 - C\alpha, \alpha/2\}\left(\|\widetilde{q}_h\|^2_{0,\Omega_1} + \frac{1}{2}\|\overline{q}_h\|^2_{0,\Omega_1}\right) \geq C\|q_h\|^2_{0,\Omega_1}.$$

Also,

$$\|\mathbf{v}_h\|_{1,\Omega_1} \leq \|\widetilde{\mathbf{v}}_h\|_{1,\Omega_1} + \|\overline{\mathbf{v}}_h\|_{1,\Omega_1} \leq C\|\widetilde{q}_h\|_{0,\Omega_1} + C\|\overline{q}_h\|_{0,\Omega_1} \leq C\|q_h\|_{0,\Omega_1}.$$

Hence, we have proved (3.31) which is equivalent to (3.30). ∎

We now prove inf-sup condition (3.29) for $\{X^h, Q_1^h\}$.

THEOREM 3.5. *$\{X^h, Q_1^h\}$ satisfies the inf-sup condition* (3.29).

Proof. Let the discrete extension operator $E^h : X_1^h \to X^h$ be defined as follows: for any $\mathbf{v}_h \in X_1^h$, $(E^h\mathbf{v}_h)|_{\overline{\Omega}_1} = \mathbf{v}_h$ and $(E^h\mathbf{v}_h)|_{\Omega_2} \in X_2^h$ is the solution of

$$[\nabla(E^h\mathbf{v}_h), \nabla\mathbf{z}_h]_{\Omega_2} = 0 \quad \forall\, \mathbf{z}_h \in X_2^h \cap \mathbf{H}_0^1(\Omega_2), \quad (E^h\mathbf{v}_h)|_{\Gamma_2} = \mathbf{0}, \quad (E^h\mathbf{v}_h)|_{\Gamma_0} = \mathbf{v}_h|_{\Gamma_0}.$$

It is well known (see, e.g., [23] and [1]) that $\|E^h\mathbf{v}_h\|_{1,\Omega_2} \leq C\|\mathbf{v}_h\|_{1/2,\Gamma_0}$ so that

$$\begin{aligned} \|E^h\mathbf{v}_h\|_{1,\Omega} &\leq C\|(E^h\mathbf{v}_h)|_{\Omega_1}\|_{1,\Omega_1} + \|(E^h\mathbf{v}_h)|_{\Omega_2}\|_{1,\Omega_2} \\ &\leq C(\|\mathbf{v}_h\|_{1,\Omega_1} + \|\mathbf{v}_h\|_{1/2,\Gamma_0}) \leq C\|\mathbf{v}_h\|_{1,\Omega_1} \qquad \forall\, \mathbf{v}_h \in X_1^h. \end{aligned}$$

Then, for every $q_h \in Q_1^h$ we have

$$\begin{aligned} \sup_{\boldsymbol{\eta}_h \in X^h} \frac{b[\boldsymbol{\eta}_h, q_h]}{\|q_h\|_{0,\Omega_1}\,\|\boldsymbol{\eta}_h\|_{1,\Omega}} &\geq \sup_{\mathbf{v}_h \in X_1^h} \frac{b[E^h\mathbf{v}_h, q_h]}{\|q_h\|_{0,\Omega_1}\,\|E^h\mathbf{v}_h\|_{1,\Omega}} \\ &\geq C \sup_{\mathbf{v}_h \in X_1^h} \frac{b[E^h\mathbf{v}_h, q_h]}{\|q_h\|_{0,\Omega_1}\,\|\mathbf{v}_h\|_{1,\Omega_1}} = C \sup_{\mathbf{v}_h \in X_1^h} \frac{b[\mathbf{v}_h, q_h]}{\|q_h\|_{0,\Omega_1}\,\|\mathbf{v}_h\|_{1,\Omega_1}} \geq C, \end{aligned}$$

where the last step is valid because of (3.30). ∎

As a direct consequence of [19, Lemma 4.1, p. 58], Theorem 3.8, and the inf-sup condition (3.29), we obtain the following theorem concerning the existence of a discrete pressure. Note that an estimate for p_h will be established in section 3.4 only after we have derived strong energy estimates, particularly the estimate for $\|\partial_t\boldsymbol{\xi}_h\|_{L^2(0,T;\mathbf{L}^2(\Omega))}$.

THEOREM 3.6. *Assume that $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, and $\mathbf{u}_1$ satisfy* (2.6)*, and let $\boldsymbol{\xi}_h \in C^1([0,T]; \boldsymbol{\Psi}^h)$ be the solution of* (3.21)–(3.22)*. Let $(\mathbf{v}_h, \mathbf{u}_h) \in C^1([0,T]; X^h|_{\Omega_1}) \times C^1([0,T]; X_2^h)$ be the solution of* (3.25) *and* (3.15)–(3.19)*. Then there exists a unique $p_h \in C([0,T]; Q_1^h)$ satisfying* (3.28) *and* (3.15)*.*

Proof. The existence and uniqueness of a $p_h \in C([0,T]; Q_1^h)$ satisfying (3.28) follow directly from [19, Lemma 4.1, p. 58], Theorem 3.8, and the inf-sup condition (3.29). Since (3.28) is equivalent to (3.27), we also conclude that p_h satisfies (3.27) and is the unique such solution. ∎

3.4. Strong a priori energy estimates for the finite element solutions. In the finite element system (3.25) and (3.15)–(3.19) the discrete initial conditions are arbitrary approximations of the corresponding continuous initial data. We now make a particular choice of discrete initial data that will allow us to derive an estimate for $\|\partial_t \boldsymbol{\xi}_h\|_{0,\Omega}$ under additional assumptions on the data. Such an estimate can then be used to derive an estimate for $\|p_h\|_{L^2(0,T;L^2(\Omega_1))}$. (The existence of a discrete pressure p_h satisfying (3.14) was shown in section 3.3.) The estimates on p_h and $\partial_t \boldsymbol{\xi}_h$ will be needed in order to prove the convergence of finite element solutions, since finite element formulations involve the term $b[\boldsymbol{\eta}_h, p_h]$, which, in general, does not vanish for $\boldsymbol{\eta}_h \in X^h$.

We first study the approximation of the initial condition. We choose $(\mathbf{v}_{0,h}, \mathbf{u}_{1,h}) \in \boldsymbol{\Psi}^h$ and $p_{0,h} \in Q_1^h$ to be the solution of

$$\begin{aligned} &a_1[\mathbf{v}_{0,h}, \boldsymbol{\eta}_h] + [\mathbf{u}_{1,h}, \boldsymbol{\eta}_h]_{\Omega_2} + b[\boldsymbol{\eta}_h, p_{0,h}] \\ &\qquad = a_1[\mathbf{v}_0, \boldsymbol{\eta}_h] + [\mathbf{u}_1, \boldsymbol{\eta}_h]_{\Omega_2} + b[\boldsymbol{\eta}_h, p_0] \qquad \forall\, \boldsymbol{\eta}_h \in X^h, \end{aligned} \tag{3.32}$$

$$b[\mathbf{v}_{0,h}, q_h] = 0 \quad \forall\, q_h \in Q_1^h \qquad \text{and} \qquad \mathbf{v}_{0,h}|_{\Gamma_0} = \mathbf{u}_{1,h}|_{\Gamma_0}, \tag{3.33}$$

where p_0 is the initial pressure field associated with the initial velocity field $\mathbf{v}_0$.

LEMMA 3.7. *Assume that* $\mathbf{v}_0 \in X_1$, $p_0 \in L^2(\Omega_1)$, $\mathbf{u}_1 \in X_2$, *and* $\mathbf{v}_0|_{\Gamma_0} = \mathbf{u}_1|_{\Gamma_0}$. *Then there exists a unique triplet* $(\mathbf{v}_{0,h}, p_{0,h}, \mathbf{u}_{1,h}) \in X_1^h \times Q_1^h \times X_2^h$ *which satisfies* (3.32)–(3.33) *and*

$$\begin{aligned} &\|\mathbf{v}_{0,h} - \mathbf{v}_0\|_{1,\Omega_1} + \|\mathbf{u}_{1,h} - \mathbf{u}_1\|_{0,\Omega_2} + \|p_{0,h} - p_0\|_{0,\Omega_1} \\ &\quad \le C(\|\boldsymbol{\eta}_h - \mathbf{v}_0\|_{1,\Omega_1} + \|\boldsymbol{\eta}_h - \mathbf{u}_1\|_{0,\Omega_2} + \|q_h - p_0\|_{0,\Omega_1}) \qquad \forall\, (\boldsymbol{\eta}_h, q_h) \in X^h \times Q_1^h. \end{aligned} \tag{3.34}$$

If, in addition, $\mathbf{v}_0 \in \mathbf{H}^{r+1}(\Omega_1)$, $p_0 \in H^r(\Omega_1)$, *and* $\mathbf{u}_1 \in \mathbf{H}^{r+1}(\Omega_2)$ *for some* $r \in [0, k]$ (k *being the integer appearing in the approximation properties*), *then*

$$\begin{aligned} &\|\mathbf{v}_{0,h} - \mathbf{v}_0\|_{1,\Omega_1} + \|\mathbf{u}_{1,h} - \mathbf{u}_1\|_{0,\Omega_2} + \|p_{0,h} - p_0\|_{0,\Omega_1} \\ &\qquad \le C h^r (\|\mathbf{v}_0\|_{r+1,\Omega_1} + \|\mathbf{u}_1\|_{r+1,\Omega_2} + \|p_0\|_{r,\Omega_1}). \end{aligned} \tag{3.35}$$

Proof. We set $\widetilde{X} = \{\boldsymbol{\eta} \in \mathbf{L}^2(\Omega) : \boldsymbol{\eta}|_{\Omega_1} \in X_1, \operatorname{div} \boldsymbol{\eta}|_{\Omega_1} = 0\}$ and equip $\widetilde{X}$ with the inner product

$$[\boldsymbol{\xi}, \boldsymbol{\eta}]_{\widetilde{X}} = a_1[\boldsymbol{\xi}, \boldsymbol{\eta}] + [\boldsymbol{\xi}, \boldsymbol{\eta}]_{\Omega_2} \qquad \forall\, \boldsymbol{\xi}, \boldsymbol{\eta} \in \widetilde{X}.$$

It is easy to check that $\widetilde{X}$ is a Hilbert space. The continuous inf-sup condition (2.4) implies

$$\begin{aligned} \inf_{q \in L^2(\Omega_1)} \sup_{\boldsymbol{\eta} \in \widetilde{X}} \frac{b[\boldsymbol{\eta}, q]}{\|\boldsymbol{\eta}\|_{\widetilde{X}} \|q\|_{0,\Omega_1}} &\ge \inf_{q \in L^2(\Omega_1)} \sup_{\boldsymbol{\eta} \in \mathbf{H}_0^1(\Omega)} \frac{b[\boldsymbol{\eta}, q]}{\|\boldsymbol{\eta}\|_{\widetilde{X}} \|q\|_{0,\Omega_1}} \\ &\ge \inf_{q \in L^2(\Omega_1)} \sup_{\boldsymbol{\eta} \in \mathbf{H}_0^1(\Omega)} \frac{b[\boldsymbol{\eta}, q]}{\|\boldsymbol{\eta}\|_{1,\Omega} \|q\|_{0,\Omega_1}} \ge C. \end{aligned}$$

Thus, by [19, Theorem 1.1, p. 114], there exists a unique $(\widetilde{\boldsymbol{\xi}}_0, \widetilde{p}_0) \in \widetilde{X} \times L^2(\Omega_1)$ satisfying

$$[\widetilde{\boldsymbol{\xi}}_0, \boldsymbol{\eta}]_{\widetilde{X}} + b[\boldsymbol{\eta}, \widetilde{p}_0] = a_1[\mathbf{v}_0, \boldsymbol{\eta}] + [\mathbf{u}_1, \boldsymbol{\eta}]_{\Omega_2} + b[\boldsymbol{\eta}, p_0] \qquad \forall\, \boldsymbol{\eta} \in \widetilde{X}, \tag{3.36}$$

$$b[\widetilde{\boldsymbol{\xi}}_0, q] = 0 \qquad \forall\, q \in L^2(\Omega_1). \tag{3.37}$$

As $\boldsymbol{\xi}_0$ defined by (2.11) and p_0 constitute an obvious solution to (3.36)–(3.37), we have

$$\widetilde{\boldsymbol{\xi}}_0 = \boldsymbol{\xi}_0 = \left\{ \begin{array}{l} \boldsymbol{\xi}|_{\Omega_1} = \mathbf{v}_0, \\ \boldsymbol{\xi}|_{\Omega_2} = \mathbf{u}_1, \end{array} \right. \qquad \text{and} \qquad \widetilde{p}_0 = p_0. \tag{3.38}$$

Similarly, the discrete inf-sup condition (3.29) implies

$$\inf_{q_h \in Q_1^h} \sup_{\boldsymbol{\eta}_h \in X^h} \frac{b[\boldsymbol{\eta}_h, q_h]}{\|\boldsymbol{\eta}_h\|_{\widetilde{X}} \|q_h\|_{0,\Omega_1}} \geq \inf_{q_h \in Q_1^h} \sup_{\boldsymbol{\eta}_h \in X^h} \frac{b[\boldsymbol{\eta}_h, q_h]}{\|\boldsymbol{\eta}_h\|_{1,\Omega} \|q_h\|_{0,\Omega_1}} \geq C,$$

so that by [19, Theorem 1.1, p. 114] there exists a unique $(\boldsymbol{\xi}_{0,h}, p_{0,h}) \in X^h \times Q_1^h$ satisfying

$$[\boldsymbol{\xi}_{0,h}, \boldsymbol{\eta}_h]_{\widetilde{X}} + b[\boldsymbol{\eta}_h, p_{0,h}] = a_1[\mathbf{v}_0, \boldsymbol{\eta}_h] + [\mathbf{u}_1, \boldsymbol{\eta}_h]_{\Omega_2} + b[\boldsymbol{\eta}_h, p_0] \qquad \forall\, \boldsymbol{\eta}_h \in X^h, \tag{3.39}$$

$$b[\boldsymbol{\xi}_{0,h}, q_h] = 0 \qquad \forall\, q \in Q_1^h; \tag{3.40}$$

moreover, the following error estimate holds:

$$\begin{aligned} &\|\boldsymbol{\xi}_{0,h} - \boldsymbol{\xi}_0\|_{\widetilde{X}} + \|p_{0,h} - p_0\|_{0,\Omega_1} \\ &\qquad \leq C(\|\boldsymbol{\eta}_h - \boldsymbol{\xi}_0\|_{\widetilde{X}} + \|q_h - p_0\|_{0,\Omega_1}) \qquad \forall\, (\boldsymbol{\eta}_h, q_h) \in X^h \times Q_1^h. \end{aligned} \tag{3.41}$$

By setting

$$\mathbf{v}_{0,h} = \boldsymbol{\xi}_{0,h}|_{\Omega_1} \quad \text{and} \quad \mathbf{u}_{1,h} = \boldsymbol{\xi}_{0,h}|_{\Omega_2}, \tag{3.42}$$

we see that (3.41) is equivalent to (3.34) and that (3.32)–(3.33) are satisfied. The uniqueness of the solution $(\mathbf{v}_{0,h}, p_{0,h}, \mathbf{u}_{1,h})$ for (3.32)–(3.33) follows from the uniqueness of the solution $(\boldsymbol{\xi}_{0,h}, p_{0,h})$ for (3.39)–(3.40).

Next, assuming that $\mathbf{v}_0 \in \mathbf{H}^{r+1}(\Omega_1)$, $p_0 \in H^r(\Omega_1)$, and $\mathbf{u}_1 \in \mathbf{H}^{r+1}(\Omega_2)$ for some $r \in [1, k]$, we proceed to prove (3.35) by making a particular choice of $\boldsymbol{\eta}_h$ in (3.34). Let $(\overline{\mathbf{v}}_{0,h}, \overline{p}_{0,h}) \in X_1^h \times Q_1^h$ be the unique finite element solution of the following Stokes system on Ω_1:

$$a_1[\overline{\mathbf{v}}_{0,h}, \mathbf{z}_h] + b[\mathbf{z}_h, \overline{p}_{0,h}] = a_1[\mathbf{v}_0, \mathbf{z}_h] + b[\mathbf{z}_h, \overline{p_0}] \qquad \forall\, \mathbf{z}_h \in X_1^h \cap \mathbf{H}_0^1(\Omega_1),$$

$$b[\overline{\mathbf{v}}_{0,h}, q_h] = 0 \qquad \forall\, q_h \in Q_1^h \cap L_0^2(\Omega_1),$$

$$\overline{\mathbf{v}}_{0,h}|_{\Gamma_1} = \mathbf{0} \qquad \text{and} \qquad [\overline{\mathbf{v}}_{0,h}, \mathbf{s}_h]_{0,\Gamma_0} = [\mathbf{v}_0, \mathbf{s}_h]_{0,\Gamma_0} \quad \forall\, \mathbf{s}_h \in X_1^h|_{\Gamma_0},$$

where $\overline{p_0} = p_0 - (1/|\Omega_1|) \int_{\Omega_1} p_0 \, d\mathbf{x}$. Using the results of [23] concerning error estimates for the finite element approximations of the Stokes equations with inhomogeneous boundary conditions, we obtain

$$\begin{aligned} &\|\overline{\mathbf{v}}_{0,h} - \mathbf{v}_0\|_{1,\Omega_1} + \|\overline{p}_{0,h} - \overline{p_0}\|_{0,\Omega_1} \\ &\qquad \leq Ch^r(\|\mathbf{v}_0\|_{r+1,\Omega_1} + \|\overline{p_0}\|_{r,\Omega_1}) \leq Ch^r(\|\mathbf{v}_0\|_{r+1,\Omega_1} + \|p_0\|_{r,\Omega_1}). \end{aligned} \tag{3.43}$$

Analogously, let $\overline{\mathbf{u}}_{1,h} \in X_2^h$ be the unique finite element solution of the following elliptic system on Ω_2 with an inhomogeneous boundary condition:

$$
\begin{aligned}
&[\nabla\overline{\mathbf{u}}_{1,h}, \nabla\mathbf{w}_h]_{\Omega_2} = [\nabla\mathbf{u}_1, \nabla\mathbf{w}_h]_{\Omega_2} \qquad \forall\, \mathbf{w}_h \in X_2^h \cap \mathbf{H}_0^1(\Omega_2),\\
&\overline{\mathbf{u}}_{1,h}|_{\Gamma_2} = \mathbf{0} \qquad \text{and} \qquad [\overline{\mathbf{u}}_{1,h}, \mathbf{s}_h]_{0,\Gamma_0} = [\mathbf{u}_1, \mathbf{s}_h]_{0,\Gamma_0} \quad \forall\, \mathbf{s}_h \in X_2^h|_{\Gamma_0}.
\end{aligned} \tag{3.44}
$$

Then we have

$$
\|\overline{\mathbf{u}}_{1,h} - \mathbf{u}_1\|_{0,\Omega_2} \le \|\overline{\mathbf{u}}_{1,h} - \mathbf{u}_1\|_{1,\Omega_2} \le Ch^r \|\mathbf{u}_1\|_{r+1,\Omega_2}. \tag{3.45}
$$

The assumption $\mathbf{v}_0|_{\Gamma_0} = \mathbf{u}_1|_{\Gamma_0}$ implies $\overline{\mathbf{v}}_{0,h}|_{\Gamma_0} = \overline{\mathbf{u}}_{1,h}|_{\Gamma_0}$, so that the element $\overline{\boldsymbol{\eta}}_h$ defined by

$$
\overline{\boldsymbol{\eta}}_h|_{\Omega_1} = \begin{cases} \overline{\mathbf{v}}_{0,h} & \text{in } \Omega_1, \\ \overline{\mathbf{u}}_{1,h} & \text{in } \Omega_2 \end{cases}
$$

satisfies $\overline{\boldsymbol{\eta}}_h \in X^h$. By choosing $\boldsymbol{\eta}_h = \overline{\boldsymbol{\eta}}_h$ and $q_h = \overline{p}_{0,h} + (1/|\Omega_1|)\int_{\Omega_1} p_0\, d\mathbf{x}$ in (3.34) and using (3.43)–(3.45), we arrive at (3.35). ∎

We now derive a strong a priori energy estimate for the auxiliary finite element solution $\boldsymbol{\xi}_h$.

THEOREM 3.8. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6) *and*

$$
\partial_t \mathbf{f}_i \in L^2(0,T;\mathbf{L}^2(\Omega_i)),\ i = 1,2, \quad \mathbf{v}_0 \in \mathbf{H}^2(\Omega_1),\ \mathbf{u}_1 \in \mathbf{H}^2(\Omega_1),\ \mathbf{u}_0 \in \mathbf{H}^2(\Omega_2). \tag{3.46}
$$

Assume further that there exists a $p_0 \in H^1(\Omega_1)$ *such that*

$$
(p_0\mathbf{n}_1 - \mu_1 \nabla\mathbf{v}_0 \cdot \mathbf{n}_1)|_{\Gamma_0} = (\mu_2 \nabla\mathbf{u}_0 \cdot \mathbf{n}_2 + (\lambda_2 + \mu_2)(\operatorname{div}\mathbf{u}_0)\mathbf{n}_2)|_{\Gamma_0}, \tag{3.47}
$$

where $\mathbf{n}_i$ *denotes the outward-pointing normal along* $\partial\Omega_i$, $i = 1, 2$. *Then there exists a unique solution* $\boldsymbol{\xi}_h \in C^1([0,T];\boldsymbol{\Psi}^h)$ *for* (3.21)–(3.22) *with the initial condition* $\boldsymbol{\xi}_{0,h}$ *defined by*

$$
\boldsymbol{\xi}_{0,h}|_{\Omega_1} = \mathbf{v}_{0,h} \quad \textit{and} \quad \boldsymbol{\xi}_{0,h}|_{\Omega_2} = \mathbf{u}_{1,h}, \tag{3.48}
$$

where $\mathbf{v}_{0,h}$ *and* $\mathbf{u}_{1,h}$ *are determined by* (3.32)–(3.33). *Moreover,* $\boldsymbol{\xi}_h$ *satisfies the estimates*

$$
\begin{aligned}
&\|\boldsymbol{\xi}_h\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\boldsymbol{\xi}_h\|^2_{L^2(0,T;\mathbf{H}^1(\Omega_1))} + \left\| \int_0^t \boldsymbol{\xi}_h(s)\, ds \right\|^2_{L^\infty(0,T;\mathbf{H}^1(\Omega_2))} \\
&\quad \le Ce^{CT}(\|\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{1,\Omega_2} + \|\mathbf{v}_0\|^2_{1,\Omega_1} + \|\mathbf{u}_1\|^2_{0,\Omega_2} + \|p_0\|^2_{0,\Omega_1})
\end{aligned} \tag{3.49}
$$

and

$$
\begin{aligned}
&\|\partial_t\boldsymbol{\xi}_h\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))} + \|\partial_t\boldsymbol{\xi}_h\|^2_{L^2(0,T;\mathbf{H}^1(\Omega_1))} + \left\| \int_0^t \partial_t\boldsymbol{\xi}_h(s)\, ds \right\|^2_{L^\infty(0,T;\mathbf{H}^1(\Omega_2))} \\
&\quad \le Ce^{CT}(\|\mathbf{f}\|^2_{H^1(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{2,\Omega_2} + \|\mathbf{v}_0\|^2_{2,\Omega_1} + \|p_0\|^2_{1,\Omega_1} + \|\mathbf{u}_1\|^2_{2,\Omega_2}).
\end{aligned} \tag{3.50}
$$

Proof. By Theorem 3.1, there exists a unique solution $\boldsymbol{\xi}_h \in C^1([0,T];\boldsymbol{\Psi}^h)$ for (3.21)–(3.22) and (3.24). We note that, by virtue of Lemma 3.7, the initial condition $\boldsymbol{\xi}_{0,h} \in \boldsymbol{\Psi}^h$ satisfies the estimate

$$
\|\boldsymbol{\xi}_{0,h}\|_{1,\Omega_1} + \|\boldsymbol{\xi}_{0,h}\|_{0,\Omega_2} \le C(\|\mathbf{v}_0\|_{1,\Omega_1} + \|p_0\|_{0,\Omega_1} + \|\mathbf{u}_1\|_{0,\Omega_2}).
$$

Thus (3.49) follows from the last estimate and (3.24).

Defining $\boldsymbol{\zeta}_h = \partial_t \boldsymbol{\xi}_h$ and differentiating (3.21), we obtain that for each $t \in [0,T]$

$$\begin{aligned}(3.51)\qquad &\rho_1[\partial_t\boldsymbol{\zeta}_h,\boldsymbol{\eta}_h]_{\Omega_1} + \rho_2[\partial_t\boldsymbol{\zeta}_h,\boldsymbol{\eta}_h]_{\Omega_2} + a_1[\boldsymbol{\zeta}_h,\boldsymbol{\eta}_h] + a_2\left[\int_0^t \boldsymbol{\zeta}_h(s)\,ds,\boldsymbol{\eta}_h\right]\\ &\quad = \rho_1[\partial_t\mathbf{f}_1,\boldsymbol{\eta}_h]_{\Omega_1} + \rho_2[\partial_t\mathbf{f}_2,\boldsymbol{\eta}_h]_{\Omega_2} - a_2[\boldsymbol{\xi}_h(0),\boldsymbol{\eta}_h] \qquad \forall\,\boldsymbol{\eta}_h \in \boldsymbol{\Psi}^h.\end{aligned}$$

Setting $\boldsymbol{\eta}_h = \boldsymbol{\zeta}_h(t)$ in (3.51) and integrating in t, we obtain

$$\begin{aligned}&[[\boldsymbol{\zeta}_h(t),\boldsymbol{\zeta}_h(t)]] + \int_0^t a_1[\boldsymbol{\zeta}_h(s),\boldsymbol{\zeta}_h(s)]\,ds + a_2\left[\int_0^t \boldsymbol{\zeta}_h(s)\,ds,\int_0^t \boldsymbol{\zeta}_h(s)\,ds\right]\\ &\quad\le C(\|\boldsymbol{\zeta}_h(0)\|_{0,\Omega}^2 + \|\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))}) + a_2\left[\mathbf{u}_0,\int_0^t \boldsymbol{\zeta}_h(s)\,ds\right] + \int_0^t \|\boldsymbol{\zeta}_h(s)\|_{0,\Omega}^2\,ds\\ &\quad\le C(\|\boldsymbol{\zeta}_h(0)\|_{0,\Omega}^2 + \|\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{1,\Omega_2})\\ &\qquad + \frac{1}{2}a_2\left[\int_0^t \boldsymbol{\zeta}_h(s)\,ds,\textstyle\int_0^t \boldsymbol{\zeta}_h(s)\,ds\right] + \int_0^t \|\boldsymbol{\zeta}_h(s)\|_{0,\Omega}^2\,ds,\end{aligned}$$

so that

$$\begin{aligned}(3.52)\qquad &\|\boldsymbol{\zeta}_h(t)\|_{0,\Omega}^2 + \int_0^t a_1[\boldsymbol{\zeta}_h(t),\boldsymbol{\zeta}_h(t)]\,dt + a_2\left[\int_0^t \boldsymbol{\zeta}_h(s)\,ds,\int_0^t \boldsymbol{\zeta}_h(s)\,ds\right]\\ &\quad\le C(\|\boldsymbol{\zeta}_h(0)\|_{0,\Omega}^2 + \|\partial_t\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\boldsymbol{\xi}_h(0)\|^2_{1,\Omega_2}) + \int_0^t \|\boldsymbol{\zeta}_h(s)\|_{0,\Omega}^2\,ds.\end{aligned}$$

Dropping the second and third terms on the left-hand side of (3.52) and then applying the following version of Gronwall's inequality [12, p. 625],

$$(3.53)\qquad \text{if}\ \ r(t) \le C_1 + C_2\int_0^t r(s)\,ds,\ \ \text{then}\ \ r(t) \le C_1(1+C_2t)e^{C_2t},$$

we deduce

$$\|\boldsymbol{\zeta}_h(t)\|_{0,\Omega}^2 \le Ce^{CT}(\|\boldsymbol{\zeta}_h(0)\|_{0,\Omega}^2 + \|\boldsymbol{\xi}_h(0)\|^2_{1,\Omega_2} + \|\partial_t\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))}).$$

The last estimate and (3.52) yield

$$\begin{aligned}(3.54)\qquad &\|\boldsymbol{\zeta}_h(t)\|_{0,\Omega}^2 + \int_0^t a_1[\boldsymbol{\zeta}_h(t),\boldsymbol{\zeta}_h(t)]\,dt + a_2\left[\int_0^t \boldsymbol{\zeta}_h(s)\,ds,\int_0^t \boldsymbol{\zeta}_h(s)\,ds\right]\\ &\quad\le Ce^{CT}(\|\boldsymbol{\zeta}_h(0)\|_{0,\Omega}^2 + \|\partial_t\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\boldsymbol{\xi}_h(0)\|^2_{1,\Omega_2}).\end{aligned}$$

The term $\|\boldsymbol{\xi}_h(0)\|^2_{1,\Omega_2}$ on the right-hand side of (3.54) can be estimated with the help of inverse inequalities (3.13), (3.45), and (3.35) with $r = 1$:

$$\begin{aligned}\|\boldsymbol{\xi}_h(0)\|_{1,\Omega_2} &\le \|\boldsymbol{\xi}_{0,h} - \overline{\mathbf{u}}_{1,h}\|_{1,\Omega_2} + \|\overline{\mathbf{u}}_{1,h} - \mathbf{u}_1\|_{1,\Omega_2} + \|\mathbf{u}_1\|_{1,\Omega_2}\\ &\le \frac{C}{h}\|\boldsymbol{\xi}_{0,h} - \overline{\mathbf{u}}_{1,h}\|_{0,\Omega_2} + Ch\|\mathbf{u}_1\|_{2,\Omega_2} + \|\mathbf{u}_1\|_{1,\Omega_2}\\ &\le \frac{C}{h}\|\boldsymbol{\xi}_{0,h} - \mathbf{u}_1\|_{0,\Omega_2} + \frac{C}{h}\|\mathbf{u}_1 - \overline{\mathbf{u}}_{1,h}\|_{0,\Omega_2} + Ch\|\mathbf{u}_1\|_{2,\Omega_2} + \|\mathbf{u}_1\|_{1,\Omega_2}\\ &\le C(\|\mathbf{v}_0\|_{2,\Omega_1} + \|\mathbf{u}_1\|_{2,\Omega_2} + \|p_0\|_{0,\Omega_1}),\end{aligned}\qquad (3.55)$$

where $\overline{\mathbf{u}}_{1,h}$ is defined by (3.44). The term $\|\boldsymbol{\zeta}_h(0)\|_{0,\Omega}^2$ can be estimated as follows. Evaluating (3.21) at $t = 0$, then setting $\boldsymbol{\eta}_h = \partial_t \boldsymbol{\xi}_h(0)$ and using (3.52), we have

$$\begin{aligned}
&[[\partial_t \boldsymbol{\xi}_h(0), \partial_t \boldsymbol{\xi}_h(0)]] = [[\mathbf{f}(0), \partial_t \boldsymbol{\xi}_h(0)]] - a_2[\mathbf{u}_0, \partial_t \boldsymbol{\xi}_h(0)] - a_1[\boldsymbol{\xi}_h(0), \partial_t \boldsymbol{\xi}_h(0)] \\
&= [[\mathbf{f}(0), \partial_t \boldsymbol{\xi}_h(0)]] - a_2[\mathbf{u}_0, \partial_t \boldsymbol{\xi}_h(0)] - b[\partial_t \boldsymbol{\xi}_h(0), p_0] - a_1[\mathbf{v}_0, \partial_t \boldsymbol{\xi}_h(0)] \\
&\quad - [\mathbf{u}_1, \partial_t \boldsymbol{\xi}_h(0)]_{\Omega_2} + [\boldsymbol{\xi}_h(0), \partial_t \boldsymbol{\xi}_h(0)]_{\Omega_2} \\
&= [[\mathbf{f}(0), \partial_t \boldsymbol{\xi}_h(0)]] + [\Delta \mathbf{u}_0 + \nabla(\operatorname{div} \mathbf{u}_0), \partial_t \boldsymbol{\xi}_h(0)]_{\Omega_2} + [\Delta \mathbf{v}_0 - \nabla p_0, \partial_t \boldsymbol{\xi}_h(0)] \\
&\quad + \int_{\Gamma_0} (-\mu_2 \nabla \mathbf{u}_0 \cdot \mathbf{n}_2 - (\lambda_2 + \mu_2)(\operatorname{div} \mathbf{u}_0)\mathbf{n}_2 + p_0 \mathbf{n}_1 - \nabla \mathbf{v}_0 \cdot \mathbf{n}_1) \cdot \partial_t \boldsymbol{\xi}_h(0)\, d\Gamma \\
&\quad - [\mathbf{u}_1, \partial_t \boldsymbol{\xi}_h(0)]_{\Omega_2} + [\boldsymbol{\xi}_h(0), \partial_t \boldsymbol{\xi}_h(0)]_{\Omega_2}.
\end{aligned}$$

Applying assumption (3.47) and initial condition (3.32)–(3.33) to the last relation, we are led to

$$\begin{aligned}
[[\partial_t \boldsymbol{\xi}_h(0), \partial_t \boldsymbol{\xi}_h(0)]] &= [[\mathbf{f}(0), \partial_t \boldsymbol{\xi}_h(0)]] + [\Delta \mathbf{u}_0 + \nabla(\operatorname{div} \mathbf{u}_0), \partial_t \boldsymbol{\xi}_h(0)]_{\Omega_2} \\
&\quad + [\Delta \mathbf{v}_0 - \nabla p_0, \partial_t \boldsymbol{\xi}_h(0)] - [\mathbf{u}_1, \partial_t \boldsymbol{\xi}_h(0)]_{\Omega_2} + [\boldsymbol{\xi}_h(0), \partial_t \boldsymbol{\xi}_h(0)]_{\Omega_2} \\
&\le C(\|\mathbf{f}(0)\|_{\mathbf{L}^2(\Omega)}^2 + \|\mathbf{u}_0\|_{2,\Omega_2}^2 + \|\mathbf{v}_0\|_2^2 + \|\mathbf{u}_1\|_{2,\Omega_2}^2 + \|p_0\|_{1,\Omega_1}^2) \\
&\quad + C\|\boldsymbol{\xi}_{0,h}\|_{0,\Omega}^2 + \frac{1}{2}[[\partial_t \boldsymbol{\xi}_h(0), \partial_t \boldsymbol{\xi}_h(0)]],
\end{aligned}$$

so that, using (3.55), the last relation simplifies to

$$\|\partial_t \boldsymbol{\xi}_h(0)\|_{0,\Omega}^2 \le C(\|\mathbf{f}\|_{H^1(0,T;\mathbf{L}^2(\Omega))}^2 + \|\mathbf{u}_0\|_{2,\Omega_2}^2 + \|\mathbf{v}_0\|_{2,\Omega_1}^2 + \|p_0\|_{1,\Omega_1}^2 + \|\mathbf{u}_1\|_{2,\Omega_2}^2).$$

Combining (3.54), (3.55), and the last relation, we obtain (3.50). ∎

Remark 2. The particular choice of the initial condition (3.32)–(3.33) played a key role in the estimation of $[[\partial_t \boldsymbol{\xi}_h(0), \partial_t \boldsymbol{\xi}_h(0)]]$.

Using relation (3.20) in reverse, i.e., setting $\mathbf{u}_h = \mathbf{u}_{0,h} + \int_0^t \boldsymbol{\xi}_h(s)|_{\Omega_2}\, ds$ and $\mathbf{v}_h = \boldsymbol{\xi}_h|_{\Omega_1}$, we arrive at the following theorem.

THEOREM 3.9. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6) *and* (3.46). *Assume further that there exists a* $p_0 \in H^1(\Omega_1)$ *such that* (3.47) *holds. Then there exists a unique solution* $(\mathbf{v}_h, p_h, \mathbf{u}_h) \in C^1([0,T]; X_1^h) \times C([0,T]; Q_1^h) \times C^1([0,T]; X_2^h)$ *for* (3.14)–(3.19) *with the initial conditions* $(\mathbf{v}_{0,h}, \mathbf{u}_{1,h})$ *defined by* (3.32)–(3.33). *Moreover,* $(\mathbf{v}_h, p_h, \mathbf{u}_h)$ *satisfies the estimates*

$$\begin{aligned}
&\|\mathbf{v}_h\|_{L^\infty(0,T;\mathbf{L}^2(\Omega_1))}^2 + \|\partial_t \mathbf{u}_h\|_{L^\infty(0,T;\mathbf{L}^2(\Omega_2))}^2 \\
&\qquad + \|\mathbf{v}_h\|_{L^2(0,T;\mathbf{H}^1(\Omega_1))}^2 + \|\mathbf{u}_h\|_{L^\infty(0,T;\mathbf{H}^1(\Omega_2))}^2 \\
&\le Ce^{CT}(\|\mathbf{f}\|_{L^2(0,T;\mathbf{L}^2(\Omega))}^2 + \|\mathbf{u}_0\|_{1,\Omega_2}^2 + \|\mathbf{v}_0\|_{1,\Omega_1}^2 + \|p_0\|_{0,\Omega_1}^2 + \|\mathbf{u}_1\|_{0,\Omega_2}^2)
\end{aligned} \tag{3.56}$$

and

$$\begin{aligned}
&\|\partial_t \mathbf{v}_h\|_{L^\infty(0,T;L^2(\Omega_1))}^2 + \|\partial_{tt} \mathbf{u}_h\|_{L^\infty(0,T;L^2(\Omega_2))}^2 \\
&\qquad + \|\partial_t \mathbf{v}_h\|_{L^2(0,T;\mathbf{H}^1(\Omega_1))}^2 + \|\partial_t \mathbf{u}_h\|_{L^\infty(0,T;\mathbf{H}^1(\Omega_2))}^2 \\
&\le Ce^{CT}(\|\mathbf{f}\|_{H^1(0,T;\mathbf{L}^2(\Omega))}^2 + \|\mathbf{u}_0\|_{2,\Omega_2}^2 + \|\mathbf{v}_0\|_{2,\Omega_1}^2 + \|p_0\|_{1,\Omega_1}^2 + \|\mathbf{u}_1\|_{2,\Omega_2}^2).
\end{aligned} \tag{3.57}$$

Recall that Theorem 3.6 stated only the existence and uniqueness of a discrete pressure p_h satisfying (3.28), (3.27) and (3.14). By virtue of the strong energy estimates (3.57) and the discrete inf-sup conditions, we now can establish an estimate for p_h.

THEOREM 3.10. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6) *and* (3.46). *Assume further that there exists a* $p_0 \in H^1(\Omega_1)$ *such that* (3.47) *holds. Let* $(\mathbf{v}_h, p_h, \mathbf{u}_h) \in C^1([0,T]; X_1^h) \times C([0,T]; Q_1^h) \times C^1([0,T]; X_2^h)$ *be the solution for* (3.14)–(3.19) *with the initial conditions* $(\mathbf{v}_{0,h}, \mathbf{u}_{1,h})$ *defined by* (3.32)–(3.33). *Then* p_h *satisfies the estimate*

(3.58)
$$\|p_h\|^2_{L^2(0,T;L^2(\Omega_1))} \le Ce^{CT}\Big(\|\mathbf{f}\|^2_{H^1(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{2,\Omega_2} + \|\mathbf{v}_0\|^2_{2,\Omega_1} + \|p_0\|^2_{1,\Omega_1} + \|\mathbf{u}_1\|^2_{2,\Omega_2}\Big).$$

Proof. We observe that from (3.28) we have

$$\|p_h\|_{L^2(0,T;L^2(\Omega_1))} \le C\Big(\|\partial_t \boldsymbol{\xi}_h\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\boldsymbol{\xi}_h\|_{L^2(0,T;X_1)} + \|\textstyle\int_0^t \boldsymbol{\xi}_h(s)\,ds\|_{L^2(0,T;X_1)}\Big).$$

Thus, (3.58) follows from the last relation and energy estimate (3.50) for $\boldsymbol{\xi}_h$. ∎

Remark 3. Note that Theorems 3.8, 3.9, and 3.10 require the specification of an initial pressure p_0 and the initial interface stress condition (3.47). From a physical point of view, these requirements are entirely reasonable.

4. The convergence of finite element solutions and error estimates. Having proved the existence of finite element solutions $(\mathbf{v}_h, p_h, \mathbf{u}_h)$ for problem (3.14)–(3.19) and (3.32)–(3.33), we now prove the convergence of the finite element solutions and derive error estimates.

4.1. The convergence of finite element solutions. We first consider the convergence of the finite element approximations.

THEOREM 4.1. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6) *and* (3.46) *and that there exists a* $p_0 \in H^1(\Omega_1)$ *such that* (3.47) *holds. Let* $(\mathbf{v}_h, p_h, \mathbf{u}_h) \in C^1([0,T]; X_1^h) \times C([0,T]; Q_1^h) \times C^1([0,T]; X_2^h)$ *be the unique solution of* (3.14)–(3.19) *with the initial conditions* $(\mathbf{v}_{0,h}, \mathbf{u}_{1,h})$ *defined by* (3.32)–(3.33). *Assume further that the finite element meshes are nested, i.e., that the triangulation* $\mathcal{T}^{h_2}(\Omega)$ *is a refinement of the triangulation* $\mathcal{T}^{h_1}(\Omega)$ *whenever* $h_2 < h_1$. *Then, there exists a unique* $(\mathbf{v}, p, \mathbf{u})$ *such that*

$$\begin{cases} \mathbf{v} \in L^\infty(0,T;\mathbf{L}^2(\Omega_1)) \cap L^2(0,T;X_1), \\ \partial_t \mathbf{v} \in L^\infty(0,T;\mathbf{L}^2(\Omega_1)) \cap L^2(0,T;X_1), \quad p \in L^2(0,T;L^2(\Omega_1)), \\ \mathbf{u} \in L^\infty(0,T;X_2), \quad \partial_t \mathbf{u} \in L^\infty(0,T;X_2), \quad \partial_{tt}\mathbf{u} \in L^\infty(0,T;\mathbf{L}^2(\Omega_2)), \end{cases} \tag{4.1}$$

$$\mathbf{v}_h \rightharpoonup \mathbf{v} \quad \text{in } L^2(0,T;X_1), \qquad \mathbf{v}_h \overset{*}{\rightharpoonup} \mathbf{v} \quad \text{in } L^\infty(0,T;\mathbf{L}^2(\Omega_1)), \tag{4.2}$$

$$\partial_t \mathbf{v}_h \overset{*}{\rightharpoonup} \partial_t \mathbf{v} \quad \text{in } L^\infty(0,T;\mathbf{L}^2(\Omega_1)), \qquad \partial_t \mathbf{v}_h \rightharpoonup \partial_t \mathbf{v} \quad \text{in } L^2(0,T;X_1), \tag{4.3}$$

$$\mathbf{u}_h \overset{*}{\rightharpoonup} \mathbf{u} \quad \text{in } L^\infty(0,T;X_2), \tag{4.4}$$

$$\partial_t \mathbf{u}_h \overset{*}{\rightharpoonup} \partial_t \mathbf{u} \quad \text{in } L^\infty(0,T;\mathbf{L}^2(\Omega_1)), \qquad \partial_t \mathbf{u}_h \overset{*}{\rightharpoonup} \partial_t \mathbf{u} \quad \text{in } L^\infty(0,T;X_2), \tag{4.5}$$

$$\partial_{tt} \mathbf{u}_h \overset{*}{\rightharpoonup} \partial_{tt} \mathbf{u} \quad \text{in } L^\infty(0,T;\mathbf{L}^2(\Omega_2)), \tag{4.6}$$

and

$$p^h \rightharpoonup p \quad \text{weakly in } L^2(0,T;L^2(\Omega_1)). \tag{4.7}$$

Furthermore, $(\mathbf{v}, p, \mathbf{u})$ *satisfies* (3.3)–(3.6) *and the estimates*

$$\begin{aligned} &\|\mathbf{v}\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega_1))} + \|\partial_t \mathbf{u}\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega_2))} \\ &\qquad + \|\mathbf{v}\|^2_{L^2(0,T;\mathbf{H}^1(\Omega_1))} + \|\mathbf{u}\|^2_{L^\infty(0,T;\mathbf{H}^1(\Omega_2))} \\ &\quad \le Ce^{CT}(\|\mathbf{f}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{2,\Omega_2} + \|\mathbf{v}_0\|^2_{2,\Omega_1} + \|p_0\|^2_{1,\Omega_1} + \|\mathbf{u}_1\|^2_{2,\Omega_2}) \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} &\|\partial_t \mathbf{v}\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega_1))} + \|\partial_t \mathbf{v}\|^2_{L^2(0,T;\mathbf{H}^1(\Omega_1))} + \|p\|^2_{L^2(0,T;L^2(\Omega_1))} \\ &\qquad + \|\partial_{tt} \mathbf{u}(t)\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega_2))} + \|\partial_t \mathbf{u}\|^2_{L^\infty(0,T;\mathbf{H}^1(\Omega_2))} \\ &\quad \le Ce^{CT}(\|\mathbf{f}\|^2_{H^1(0,T;\mathbf{L}^2(\Omega))} + \|\mathbf{u}_0\|^2_{2,\Omega_2} + \|\mathbf{v}_0\|^2_{2,\Omega_1} + \|p_0\|^2_{1,\Omega_1} + \|\mathbf{u}_1\|^2_{2,\Omega_2}). \end{aligned} \tag{4.9}$$

Proof. We have that $\{(\mathbf{v}_h, p_h, \mathbf{u}_h)\}$ satisfies the estimates (3.56)–(3.57) and (3.58). Using these estimates, we may extract a subsequence $\{(\mathbf{v}_{h_n}, p_{h_n}, \mathbf{u}_{h_n})\}$ of $\{(\mathbf{v}_h, p_h, \mathbf{u}_h)\}$, with $\{h_n\}$ decreasing to 0 as $n \to \infty$, such that (4.2)–(4.7) hold for the subsequence $\{(\mathbf{v}_{h_n}, p_{h_n}, \mathbf{u}_{h_n})\}$ for a $(\mathbf{v}, p, \mathbf{u})$ satisfying (4.1).

Equation (3.17) holds for $h = h_n$, and thus, by passing to the limit as $n \to \infty$ in that equation, we obtain (3.6). Also, $\mathbf{u}(0) = \mathbf{u}_0$ trivially holds.

To prove that $(\mathbf{v}, p, \mathbf{u})$ satisfies (3.3) we begin from (3.14) with $h = h_n$. We arbitrarily fix an integer N and a function $\boldsymbol{\eta} \in C^1([0,T]; X^{h_N})$. For each $n > N$ we obtain from (3.14) and the nesting assumption on the triangulation family $\mathcal{T}^h(\Omega)$ that

$$\begin{aligned} &\int_0^T \Big(\rho_1[\partial_t \mathbf{v}_{h_n}, \boldsymbol{\eta}]_{\Omega_1} + a_1[\mathbf{v}_{h_n}, \boldsymbol{\eta}] + b[\boldsymbol{\eta}, p_{h_n}] + \rho_2[\partial_{tt} \mathbf{u}_{h_n}, \boldsymbol{\eta}]_{\Omega_2} + a_2[\mathbf{u}_{h_n}, \boldsymbol{\eta}]\Big)\, dt \\ &\qquad = \int_0^T \Big(\rho_1[\mathbf{f}_1, \boldsymbol{\eta}]_{\Omega_1} + \rho_2[\mathbf{f}_2, \boldsymbol{\eta}]_{\Omega_2}\Big)\, dt. \end{aligned} \tag{4.10}$$

Passing to the limit as $n \to \infty$, we find

$$\begin{aligned} &\int_0^T \Big(\rho_1[\partial_t \mathbf{v}, \boldsymbol{\eta}]_{\Omega_1} + a_1[\mathbf{v}, \boldsymbol{\eta}] + b[\boldsymbol{\eta}, p] + \rho_2[\partial_{tt} \mathbf{u}, \boldsymbol{\eta}]_{\Omega_2} + a_2[\mathbf{u}, \boldsymbol{\eta}]\Big)\, dt \\ &\qquad = \int_0^T \Big(\rho_1[\mathbf{f}_1, \boldsymbol{\eta}]_{\Omega_1} + \rho_2[\mathbf{f}_2, \boldsymbol{\eta}]_{\Omega_2}\Big)\, dt. \end{aligned} \tag{4.11}$$

Equality (4.11) then holds for all $\boldsymbol{\eta} \in L^2(0,T;\mathbf{H}^1_0(\Omega))$, as $\bigcup_{n=N}^{\infty} C([0,T];X^{h_n})$ is dense in $L^2(0,T;\mathbf{H}^1_0(\Omega))$ for the $L^2(0,T;\mathbf{H}^1_0(\Omega))$ norm. Hence,

$$\begin{aligned}\rho_1[\partial_t \mathbf{v},\boldsymbol{\eta}]_{\Omega_1} + a_1[\mathbf{v},\boldsymbol{\eta}] + b[\boldsymbol{\eta},p] + \rho_2[\partial_{tt}\mathbf{u},\boldsymbol{\eta}]_{\Omega_2} + a_2[\mathbf{u},\boldsymbol{\eta}]\\ = \rho_1[\mathbf{f}_1,\boldsymbol{\eta}]_{\Omega_1} + \rho_2[\mathbf{f}_2,\boldsymbol{\eta}]_{\Omega_2} \qquad \forall\, \boldsymbol{\eta} \in \mathbf{H}^1_0(\Omega),\ \text{a.e. } t,\end{aligned}$$

which is precisely (3.3).

From (3.15) we obtain

$$\int_0^T b[\mathbf{v}_{h_n}, q]\, ds = 0$$

for all $q \in L^2(0,T;Q_1^{h_N})$ and all $n \geq N$. Passing to the limit as $n \to \infty$ leads us to

$$\int_0^T b[\mathbf{v}, q]\, ds = 0 \tag{4.12}$$

for all $q \in L^2(0,T;Q_1^{h_N})$. Using the denseness (with respect to the $L^2(0,T;L^2(\Omega_1))$ norm) of $\bigcup_{n=N}^{\infty} L^2(0,T;Q_1^{h_n})$ in $L^2(0,T;L^2(\Omega_1))$, we see that (4.12) holds for all $q \in L^2(0,T;L^2(\Omega_1))$. In particular, this implies (3.4).

To verify the initial condition (3.5) we first note that the regularity results (4.1) imply that $\mathbf{v} \in C([0,T];\mathbf{L}^2(\Omega_1)) \cap C([0,T;X_1)$, $\mathbf{u} \in C([0,T];\mathbf{L}^2(\Omega_2)) \cap C([0,T;X_2)$, and $\partial_t \mathbf{u} \in C([0,T];\mathbf{L}^2(\Omega_2))$. For each $\boldsymbol{\eta} \in C^1([0,T];\mathbf{H}^1_0(\Omega))$ with $\boldsymbol{\eta}(T) = \mathbf{0}$ we obtain, from (4.11), by integration by parts that

$$\begin{aligned}\int_0^T \Big(-\rho_1[\mathbf{v},\partial_t\boldsymbol{\eta}]_{\Omega_1} - \rho_2[\partial_t\mathbf{u},\partial_t\boldsymbol{\eta}]_{\Omega_2} + a_1[\mathbf{v},\boldsymbol{\eta}] + b[\boldsymbol{\eta},\widehat{p}] + a_2[\mathbf{u},\boldsymbol{\eta}]\Big)\, dt\\ = \int_0^T [[\mathbf{f},\boldsymbol{\eta}]]\, dt + \rho_1[\mathbf{v}(0),\boldsymbol{\eta}(0)]_{\Omega_1} + \rho_2[\partial_t\mathbf{u}(0),\boldsymbol{\eta}(0)]_{\Omega_2}.\end{aligned} \tag{4.13}$$

On the other hand, from (4.10), we deduce that for all $\boldsymbol{\eta} \in C^1([0,T];X^{h_N})$ and all $n > N$,

$$\begin{aligned}\int_0^T \Big(-\rho_1[\mathbf{v}_{h_n},\partial_t\boldsymbol{\eta}]_{\Omega_1} - \rho_2[\partial_t\mathbf{u}_{h_n},\partial_t\boldsymbol{\eta}]_{\Omega_2}\\ + a_1[\mathbf{v}_{h_n},\boldsymbol{\eta}] + b[\boldsymbol{\eta},p_{h_n}] + a_2[\mathbf{u}_{h_n},\boldsymbol{\eta}]\Big)\, dt\\ = \int_0^T [[\mathbf{f},\boldsymbol{\eta}]]\, dt + \rho_1[\mathbf{v}_{h_n}(0),\boldsymbol{\eta}(0)]_{\Omega_1} + \rho_2[\partial_t\mathbf{u}_{h_n}(0),\boldsymbol{\eta}(0)]_{\Omega_2}.\end{aligned} \tag{4.14}$$

Holding N fixed and passing to the limit as $n \to \infty$ in (4.14) and utilizing (3.35), we arrive at

$$\begin{aligned}\int_0^T \Big(-\rho_1[\mathbf{v},\partial_t\boldsymbol{\eta}]_{\Omega_1} - \rho_2[\partial_t\mathbf{u},\partial_t\boldsymbol{\eta}]_{\Omega_2} + a_1[\mathbf{v},\boldsymbol{\eta}] + b[\boldsymbol{\eta},\widehat{p}] + a_2[\mathbf{u},\boldsymbol{\eta}]\Big)\, dt\\ = \int_0^T [[\mathbf{f},\boldsymbol{\eta}]]\, dt + \rho_1[\mathbf{v}_0,\boldsymbol{\eta}(0)]_{\Omega_1} + \rho_2[\mathbf{u}_1,\boldsymbol{\eta}(0)]_{\Omega_2}\end{aligned} \tag{4.15}$$

for all $\boldsymbol{\eta} \in C^1([0,T];X^{h_N})$. Comparing (4.13) and (4.15), we obtain

$$\rho_1[\mathbf{v}(0)-\mathbf{v}_0,\boldsymbol{\eta}(0)]_{\Omega_1} + \rho_2[\partial_t\mathbf{u}(0)-\mathbf{u}_1,\boldsymbol{\eta}(0)]_{\Omega_2} = 0 \tag{4.16}$$

for all $\boldsymbol{\eta}(0) \in X^{h_N}$. Since $\bigcup_{n=N}^{\infty} X^{h_n}$ is dense in $\mathbf{L}^2(\Omega)$ for the $\mathbf{L}^2(\Omega)$ norm, we derive

$$\mathbf{v}(0) = \mathbf{v}_0 \quad \text{in } \mathbf{L}^2(\Omega_1) \qquad \text{and} \qquad \partial_t \mathbf{u}(0) = \mathbf{u}_1 \quad \text{in } \mathbf{L}^2(\Omega_2).$$

To check $\mathbf{u}(0) = \mathbf{u}_0$ we first note that with regularity (4.1) we are justified to write

$$\mathbf{u} = \mathbf{u}(0) + \int_0^t \partial_t \mathbf{u}(s)\, ds. \tag{4.17}$$

From the compact embedding $H^1(0,T;B) \hookrightarrow\hookrightarrow L^2(0,T;B)$ for any Banach space B and the weak convergence (4.2)–(4.5) we deduce that for a further subsequence h_{n_j} we have

$$\partial_t \mathbf{u}_{h_{n_j}} \to \partial_t \mathbf{u} \quad \text{in } L^2(0,T;\mathbf{L}^2(\Omega_2)) \qquad \text{and} \qquad \mathbf{u}_{h_{n_j}} \to \mathbf{u} \quad \text{in } L^2(0,T;\mathbf{L}^2(\Omega_2)),$$

so that, passing to the limit in the relation

$$\mathbf{u}_{h_n} = \mathbf{u}_{0,h_n} + \int_0^t \partial_t \mathbf{u}_{h_n}(s)\, ds$$

and noting that $\|\mathbf{u}_{0,h} - \mathbf{u}_0\|_{0,\Omega_2} \to 0$ as $h \to 0$, we obtain

$$\mathbf{u} = \mathbf{u}_0 + \int_0^t \partial_t \mathbf{u}(s)\, ds. \tag{4.18}$$

A comparison of (4.17) and (4.18) yields $\mathbf{u}(0) = \mathbf{u}_0$.

Hence we have verified that $(\mathbf{v}, p, \mathbf{u})$ satisfies (3.1)–(3.6). Of course, $(\mathbf{v}, \mathbf{u})$ is also a solution for (2.7)–(2.10), so that, by Theorem 2.2, $(\mathbf{v}, \mathbf{u})$ is the unique solution of (2.7)–(2.10) and estimate (4.8) holds. Then, by Theorem 2.3, we obtain the uniqueness of p. Estimate (4.9) follows from (3.57) and (3.58).

Finally, it follows from the uniqueness of the limit $(\mathbf{v}, p, \mathbf{u})$ that the entire family of finite element solutions $(\mathbf{v}_h, p_h, \mathbf{u}_h)$ satisfies (4.2)–(4.7) as $h \to 0$. □

We also have the following strong convergence, the proof of which is contained in that of Theorem 4.1.

COROLLARY 4.2. *Assume that all hypotheses of Theorem* 4.1 *hold. Then*

$$\mathbf{v}_h \to \mathbf{v} \quad \textit{in } L^2(0,T;\mathbf{L}^2(\Omega_1)), \qquad \mathbf{u}_h \to \mathbf{u} \quad \textit{in } L^2(0,T;X_2)$$

and

$$\partial_t \mathbf{u}_h \to \partial_t \mathbf{u} \quad \textit{in } L^2(0,T;\mathbf{L}^2(\Omega_2)).$$

4.2. Error estimates for finite element approximations. We will estimate the error between the continuous solution defined by (3.3)–(3.6) and the finite element solution defined by (3.14)–(3.19) and (3.32)–(3.33). To this end we introduce the weighted $\mathbf{L}^2(\Omega)$ projection operator onto the discretely divergence-free space $\boldsymbol{\Psi}^h$. ($\boldsymbol{\Psi}^h$ is discretely divergence-free in Ω_1.)

The projection operator $\mathcal{P}^h : \mathbf{L}^2(\Omega) \to \boldsymbol{\Psi}^h$ with respect to the weighted $\mathbf{L}^2(\Omega)$ inner product is defined as follows: for every $\boldsymbol{\eta} \in \mathbf{L}^2(\Omega)$, $\mathcal{P}^h\boldsymbol{\eta} \in \boldsymbol{\Psi}^h$ is the solution of

$$[[\mathcal{P}^h\boldsymbol{\eta}, \mathbf{z}^h]] = [[\boldsymbol{\eta}, \mathbf{z}^h]] \qquad \forall\, \mathbf{z}^h \in \boldsymbol{\Psi}^h. \tag{4.19}$$

Note that the definition of $\mathbf{\Psi}^h$ implies

$$b[\mathcal{P}^h\boldsymbol{\eta}, q^h] = 0 \qquad \forall\, q^h \in Q_1^h. \tag{4.20}$$

We assume that the domains Ω_1 and Ω_2 satisfy the following regularity assumptions.

Hypothesis (H1). The problem

$$\begin{cases} (\bar{\mathbf{v}}, \bar{p}) \in \mathbf{H}_0^1(\Omega_1) \times L_0^2(\Omega_1), \\ [\nabla\bar{\mathbf{v}}, \nabla\mathbf{z}]_{\Omega_1} + b[\mathbf{z}, \bar{p}] = [\bar{\mathbf{f}}_1, \mathbf{z}]_{\Omega_1} \qquad \forall\, \mathbf{z} \in \mathbf{H}_0^1(\Omega_1), \\ b[\mathbf{v}, q] = 0 \qquad \forall\, q \in L^2(\Omega_1) \end{cases} \tag{4.21}$$

is $\mathbf{H}^{2-\epsilon_1}$ regular for an $\epsilon_1 \in (0,1)$; i.e., for every $\bar{\mathbf{f}}_1 \in \mathbf{L}^2(\Omega_1)$, the solution $(\bar{\mathbf{v}}, \bar{p})$ to problem (4.21) belongs to $\mathbf{H}^{2-\epsilon_1}(\Omega_1) \times H^{1-\epsilon_1}(\Omega_1)$, $-\bar{p}\mathbf{n}_1 + (\nabla\bar{\mathbf{v}} + \nabla\bar{\mathbf{v}}^T)\mathbf{n}_1 \in \mathbf{H}^{1/2-\epsilon_1}(\Gamma_0)$, and

$$\|\bar{\mathbf{v}}\|_{\mathbf{H}^{2-\epsilon_1}(\Omega_1)} + \|\bar{p}\|_{H^{1-\epsilon_1}(\Omega_1)}\| - \bar{p}\mathbf{n}_1 + (\nabla\bar{\mathbf{v}} + \nabla\bar{\mathbf{v}}^T)\mathbf{n}_1\|_{1/2-\epsilon_1,\Gamma_0} \le C\|\bar{\mathbf{f}}_1\|_{0,\Omega_1}.$$

Hypothesis (H2). The problem

$$\begin{cases} \bar{\mathbf{u}} \in \mathbf{H}_0^1(\Omega_1), \\ [\nabla\bar{\mathbf{u}}, \nabla\mathbf{w}]_{\Omega_1} = [\bar{\mathbf{f}}_2, \mathbf{w}]_{\Omega_1} \qquad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega_2) \end{cases} \tag{4.22}$$

is $\mathbf{H}^{2-\epsilon_2}$ regular for an $\epsilon_2 \in (0,1)$; i.e., for every $\bar{\mathbf{f}}_2 \in \mathbf{L}^2(\Omega_2)$, the solution $\bar{\mathbf{u}}$ to problem (4.22) belongs to $\mathbf{H}^{2-\epsilon_2}(\Omega_2)$, $\nabla\bar{\mathbf{u}} \cdot \mathbf{n}_2 \in \mathbf{H}^{1/2-\epsilon_2}(\Gamma_0)$, and

$$\|\bar{\mathbf{u}}\|_{\mathbf{H}^{2-\epsilon_2}(\Omega_2)}\|\nabla\bar{\mathbf{u}} \cdot \mathbf{n}_2\|_{1/2-\epsilon_2,\Gamma_0} \le C\|\bar{\mathbf{f}}_2\|_{0,\Omega_2}.$$

Remark 4. Hypotheses (H1)–(H2) are simply equivalent to angle conditions on Ω_1 and Ω_2 owing to the well-known regularity results on polygonal domains for boundary value problems (4.21) and (4.22); see [24] and [19]. In particular, if both Ω_1 and Ω_2 are convex (in which case Γ_0 is necessarily a straight line), then ϵ_1 and ϵ_2 can be chosen arbitrarily small.

Under Hypotheses (H1)–(H2), we may prove the following error estimates for the projection operator $\mathcal{P}^h$:

$$\begin{aligned} &\|\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}\|_{1,\Omega} \le Ch^{r-\epsilon}(\|\boldsymbol{\zeta}\|_{r+1,\Omega_1} + \|\boldsymbol{\zeta}\|_{r+1,\Omega_2}) \\ &\qquad \forall\, \boldsymbol{\zeta} \in \mathbf{\Psi} \quad \text{ with } \boldsymbol{\zeta}|_{\Omega_i} \in \mathbf{H}^{r+1}(\Omega_i), \quad i = 1, 2, r \in [0, k], \end{aligned} \tag{4.23}$$

and

$$\begin{aligned} &\|\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}\|_{0,\Omega} \le Ch^{r+1-\epsilon}(\|\boldsymbol{\zeta}\|_{r+1,\Omega_1} + \|\boldsymbol{\zeta}\|_{r+1,\Omega_2}) \\ &\qquad \forall\, \boldsymbol{\zeta} \in \mathbf{\Psi} \quad \text{ with } \boldsymbol{\zeta}|_{\Omega_i} \in \mathbf{H}^{r+1}(\Omega_i), \quad i = 1, 2, r \in [0, k]. \end{aligned} \tag{4.24}$$

The proof of (4.23)–(4.24) will be given in the appendix, Theorem A.3.

Now we prove the following error estimates for the semidiscrete finite element approximations of the fluid-solid interaction problem.

THEOREM 4.3. *Assume that* $\mathbf{f}_1, \mathbf{v}_0, \mathbf{f}_2, \mathbf{u}_0$, *and* $\mathbf{u}_1$ *satisfy* (2.6) *and* (3.46) *and that there exists a* $p_0 \in H^1(\Omega_1)$ *such that* (3.47) *holds. Assume also that* (H1)–(H2) *hold.*

Let $(\mathbf{v}, p, \mathbf{u})$ *be the solution of* (3.1)–(3.6)*, and* $(\mathbf{v}_h, p_h, \mathbf{u}_h)$ *be the solution of* (3.14)–(3.19) *and* (3.32)–(3.33)*. Assume that for some* $r \in [1, k]$*,* $\mathbf{v} \in L^2(0,T;\mathbf{H}^{r+1}(\Omega_1))$*,* $\partial_t \mathbf{v} \in L^2(0,T;\mathbf{H}^{r-1}(\Omega_1))$*,* $p \in L^2(0,T;H^r(\Omega_1))$*,* $\partial_t \mathbf{u} \in L^2(0,T;\mathbf{H}^{r+1}(\Omega_2))$*,* $\partial_{tt}\mathbf{u} \in L^2(0,T;\mathbf{H}^{r-1}(\Omega_2))$*,* $\mathbf{v}_0 \in \mathbf{H}^{r+1}(\Omega_1)$*,* $\mathbf{u}_1 \in \mathbf{H}^{r+1}(\Omega_2)$*,* $\mathbf{u}_0 \in \mathbf{H}^{r+1}(\Omega_2)$*, and* $p_0 \in H^r(\Omega_1)$*. Then,*

$$
\begin{aligned}
&\|\mathbf{v}(t) - \mathbf{v}_h(t)\|_{0,\Omega_1}^2 + \|\mathbf{v} - \mathbf{v}_h\|_{L^2(0,T;X_1)}^2 \\
&\quad + \|\partial_t \mathbf{u}(t) - \partial_t \mathbf{u}_h(t)\|_{0,\Omega_2}^2 + \|\mathbf{u}(t) - \mathbf{u}_h(t)\|_{1,\Omega_2}^2 \\
&\le Ce^{CT} h^{2r} \big(\|\mathbf{v}_0\|_{r+1,\Omega_1}^2 + \|\mathbf{u}_1\|_{r+1,\Omega_2}^2 + \|\mathbf{u}_0\|_{r+1,\Omega_2}^2 + \|p_0\|_{r,\Omega_1}^2 \\
&\quad + \|p\|_{L^2(0,T;H^r(\Omega_1))}^2 \big) + Ce^{CT} h^{2(r-\epsilon)} (\|\mathbf{v}\|_{L^2(0,T;\mathbf{H}^{r+1}(\Omega_1))}^2 \\
&\quad + \|\mathbf{u}_t\|_{L^2(0,T;\mathbf{H}^{r+1}(\Omega_2))}^2 + \|\partial_t \mathbf{v}\|_{L^2(0,T;\mathbf{H}^{r-1}(\Omega_1))}^2 + \|\partial_{tt}\mathbf{u}\|_{L^2(0,T;\mathbf{H}^{r-1}(\Omega_2))}^2)
\end{aligned} \tag{4.25}
$$

for all $t \in [0,T]$*.*

Proof. Let $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_h$ be defined by (2.11) and (3.20), respectively. We set $\widetilde{\mathbf{v}}_h(t) = [\mathcal{P}^h \boldsymbol{\xi}(t)]|_{\Omega_1}$ and $\widetilde{\mathbf{w}}_h(t) = [\mathcal{P}^h \boldsymbol{\xi}(t)]|_{\Omega_2}$.

By subtracting (3.14)–(3.15) from the corresponding equations of (3.3)–(3.4), we obtain the following "orthogonality conditions":

$$
\begin{aligned}
&\rho_1[\partial_t \mathbf{v} - \partial_t \mathbf{v}_h, \boldsymbol{\eta}_h]_{\Omega_1} + b[\boldsymbol{\eta}_h, p - p_h] + a_1[\mathbf{v} - \mathbf{v}_h, \boldsymbol{\eta}_h] \\
&\qquad + \rho_2[\mathbf{u}_{tt} - \partial_{tt}\mathbf{u}_h, \boldsymbol{\eta}_h]_{\Omega_2} + a_2[\mathbf{u} - \mathbf{u}_h, \boldsymbol{\eta}_h] = 0 \qquad \forall\, \boldsymbol{\eta}_h \in X^h, \text{ a.e. } t,
\end{aligned} \tag{4.26}
$$

$$
b[\mathbf{v} - \mathbf{v}_h, q_h] = 0 \qquad \forall\, q_h \in Q_1^h, \text{ a.e. } t. \tag{4.27}
$$

By adding/subtracting terms and using (4.26)–(4.27), we deduce that

$$
\begin{aligned}
&\rho_1[\partial_t \mathbf{v}_h - \partial_t \mathbf{v}_h, \mathbf{v} - \mathbf{v}_h]_{\Omega_1} + a_1[\mathbf{v} - \mathbf{v}_h, \mathbf{v} - \mathbf{v}_h] \\
&\qquad + \rho_2[\partial_{tt}\mathbf{u} - \partial_{tt}\mathbf{u}_h, \partial_t \mathbf{u} - \partial_t \mathbf{u}_h]_{\Omega_2} + a_2[\mathbf{u} - \mathbf{u}_h, \partial_t \mathbf{u} - \mathbf{u}_h] \\
&= \rho_1[\partial_t \mathbf{v} - \partial_t \mathbf{v}_h, \mathbf{v} - \widetilde{\mathbf{v}}_h]_{\Omega_1} + a_1[\mathbf{v} - \mathbf{v}_h, \mathbf{v} - \widetilde{\mathbf{v}}_h] \\
&\qquad + \rho_2[\partial_{tt}\mathbf{u} - \partial_{tt}\mathbf{u}_h, \partial_t \mathbf{u} - \widetilde{\mathbf{w}}_h]_{\Omega_2} + a_2[\mathbf{u} - \mathbf{u}_h, \partial_t \mathbf{u} - \widetilde{\mathbf{w}}_h] \\
&\qquad - b[\widetilde{\mathbf{v}}_h - \mathbf{v}_h, p - p_h] + \rho_1[\partial_t \mathbf{v} - \partial_t \mathbf{v}_h, \widetilde{\mathbf{v}}_h - \mathbf{v}_h]_{\Omega_1} \\
&\qquad + a_1[\mathbf{v} - \mathbf{v}_h, \widetilde{\mathbf{v}} - \mathbf{v}_h] + \rho_2[\partial_{tt}\mathbf{u} - \partial_{tt}\mathbf{u}_h, \widetilde{\mathbf{w}}_h - \partial_t \mathbf{u}_h]_{\Omega_2} \\
&\qquad + a_2[\mathbf{u} - \mathbf{u}_h, \widetilde{\mathbf{w}}_h - \partial_t \mathbf{u}_h] + b[\widetilde{\mathbf{v}}_h - \mathbf{v}_h, p - p_h] \\
&= \rho_1[\partial_t \mathbf{v} - \partial_t \mathbf{v}_h, \mathbf{v} - \widetilde{\mathbf{v}}_h]_{\Omega_1} + a_1[\mathbf{v} - \mathbf{v}_h, \mathbf{v} - \widetilde{\mathbf{v}}_h] \\
&\qquad + \rho_2[\partial_{tt}\mathbf{u} - \partial_{tt}\mathbf{u}_h, \partial_t \mathbf{u} - \widetilde{\mathbf{w}}_h]_{\Omega_2} + a_2[\mathbf{u} - \mathbf{u}_h, \partial_t \mathbf{u} - \widetilde{\mathbf{w}}_h] \\
&\qquad + b[\mathbf{v}_h - \widetilde{\mathbf{v}}_h, p - p_h].
\end{aligned} \tag{4.28}
$$

By the definition of $\widetilde{\mathbf{v}}_h$ and (4.20), we obtain

$$
b[\widetilde{\mathbf{v}}_h(t), p_h] = b[\mathcal{P}^h \boldsymbol{\xi}(t), p_h] = 0 = b[\mathcal{P}^h \boldsymbol{\xi}(t), q_h] = b[\widetilde{\mathbf{v}}_h(t), q_h] \qquad \forall\, q_h \in Q_1^h\,. \tag{4.29}
$$

Utilizing (3.15), we have

$$b[\mathbf{v}_h(t), p_h] = 0 = b[\mathbf{v}_h(t), q_h] \qquad \forall\, q_h \in Q_1^h. \tag{4.30}$$

Additionally,

$$\begin{aligned}
&\rho_1[\partial_t \mathbf{v} - \partial_t \mathbf{v}_h, \mathbf{v} - \widetilde{\mathbf{v}}_h]_{\Omega_1} + \rho_2[\partial_{tt}\mathbf{u} - \partial_{tt}\mathbf{u}_h, \partial_t \mathbf{u} - \widetilde{\mathbf{w}}]_{\Omega_2} \\
&\quad = [[\partial_t \boldsymbol{\xi}(t) - \partial_t \boldsymbol{\xi}_h(t), \boldsymbol{\xi}(t) - \mathcal{P}^h \boldsymbol{\xi}(t)]] = [[\partial_t \boldsymbol{\xi}(t), \boldsymbol{\xi}(t) - \mathcal{P}^h \boldsymbol{\xi}(t)]] \\
&\quad = [[\partial_t \boldsymbol{\xi}(t) - \partial_t \mathcal{P}^h \boldsymbol{\xi}(t), \boldsymbol{\xi}(t) - \mathcal{P}^h \boldsymbol{\xi}(t)]] \\
&\quad = \frac{1}{2}\frac{d}{dt}[[\boldsymbol{\xi}(t) - \mathcal{P}^h \boldsymbol{\xi}(t), \boldsymbol{\xi}(t) - \mathcal{P}^h \boldsymbol{\xi}(t)]] \\
&\quad = \frac{\rho_1}{2}\frac{d}{dt}\|\mathbf{v} - \widetilde{\mathbf{v}}_h\|^2_{0,\Omega_1} + \frac{\rho_2}{2}\frac{d}{dt}\|\partial_t \mathbf{u} - \widetilde{\mathbf{w}}_h\|^2_{0,\Omega_2}.
\end{aligned} \tag{4.31}$$

Combining (4.28)–(4.31), we deduce that for all $q_h \in L^2(0,T;Q_1^h)$

$$\begin{aligned}
&\frac{\rho_1}{2}\frac{d}{dt}\|\mathbf{v} - \mathbf{v}_h\|^2_{0,\Omega_1} + a_1[\mathbf{v} - \mathbf{v}_h, \mathbf{v} - \mathbf{v}_h] + \frac{\rho_2}{2}\frac{d}{dt}\|\partial_t \mathbf{u} - \partial_t \mathbf{u}_h\|^2_{0,\Omega_2} \\
&\qquad + \frac{1}{2}\frac{d}{dt}a_2[\mathbf{u} - \mathbf{u}_h, \mathbf{u} - \mathbf{u}_h] \\
&= \frac{\rho_1}{2}\frac{d}{dt}\|\mathbf{v} - \widetilde{\mathbf{v}}_h\|^2_{0,\Omega_1} + a_1[\mathbf{v} - \mathbf{v}_h, \mathbf{v} - \widetilde{\mathbf{v}}_h] + \frac{\rho_2}{2}\frac{d}{dt}\|\partial_t \mathbf{u} - \widetilde{\mathbf{w}}_h\|^2_{0,\Omega_2} \\
&\qquad + a_2[\mathbf{u} - \mathbf{u}_h, \partial_t \mathbf{u} - \widetilde{\mathbf{w}}^h] + b[\mathbf{v}_h - \widetilde{\mathbf{v}}_h, p - q_h] \\
&\le \frac{\rho_1}{2}\frac{d}{dt}\|\mathbf{v} - \widetilde{\mathbf{v}}_h\|^2_{0,\Omega_1} + \frac{k_1}{4}\|\mathbf{v}(t) - \mathbf{v}_h(t)\|^2_{1,\Omega_1} + C\|\mathbf{v}(t) - \widetilde{\mathbf{v}}_h(t)\|^2_{1,\Omega_1} \\
&\qquad + \frac{\rho_2}{2}\frac{d}{dt}\|\partial_t \mathbf{u} - \widetilde{\mathbf{w}}_h\|^2_{0,\Omega_2} + \|\mathbf{u}(t) - \mathbf{u}_h(t)\|^2_{0,\Omega_2} + C\|\partial_t \mathbf{u}(t) - \widetilde{\mathbf{w}}_h(t)\|^2_{1,\Omega_2} \\
&\qquad + C\|\mathbf{v}(t) - \widetilde{\mathbf{v}}_h(t)\|^2_{1,\Omega_2} + \frac{k_1}{4}\|\mathbf{v}(t) - \mathbf{v}_h(t)\|^2_{1,\Omega_2} + C\|p(t) - q_h\|^2_{0,\Omega_1}.
\end{aligned}$$

Applying (2.2)–(2.3) to the last relation and integrating in t, we obtain

$$\begin{aligned}
&\rho_1\|\mathbf{v}(t) - \mathbf{v}_h(t)\|^2_{0,\Omega_1} + k_1\|\mathbf{v} - \mathbf{v}_h\|^2_{L^2(0,T;\mathbf{H}^1(\Omega_1))} \\
&\qquad + \rho_2\|\partial_t \mathbf{u}(t) - \partial_t \mathbf{u}_h(t)\|^2_{0,\Omega_2} + \|\mathbf{u}(t) - \mathbf{u}_h(t)\|^2_{1,\Omega_2} \\
&\le C\Big(\|\mathbf{v}(0) - \mathbf{v}_{0,h}\|^2_{0,\Omega_1} + \|\partial_t \mathbf{u}(0) - \mathbf{u}_{1,h}\|^2_{0,\Omega_2} + \|\mathbf{u}_0 - \mathbf{u}_{0,h}\|^2_{1,\Omega_2} \\
&\qquad + \|\boldsymbol{\xi}_0 - \mathcal{P}^h \boldsymbol{\xi}_0\|^2_{0,\Omega} + \|\boldsymbol{\xi}(t_0) - \mathcal{P}^h \boldsymbol{\xi}(t_0)\|^2_{0,\Omega_1} + \|\boldsymbol{\xi} - \mathcal{P}^h \boldsymbol{\xi}\|^2_{L^2(0,T;\mathbf{H}^1(\Omega))} \\
&\qquad + \|p - q_h\|^2_{L^2(0,T;L^2(\Omega_1))}\Big) + \int_0^t \|\mathbf{u}(s) - \mathbf{u}_h(s)\|^2_{1,\Omega_2} ds
\end{aligned} \tag{4.32}$$

for all $q_h \in L^2(0,T;Q_1^h)$, where $t_0 \in [0,T]$ is such that

$$\|\boldsymbol{\xi}(t_0) - \mathcal{P}^h \boldsymbol{\xi}(t_0)\|^2_{0,\Omega} = \max_{t\in[0,T]} \|\boldsymbol{\xi}(t) - \mathcal{P}^h \boldsymbol{\xi}(t)\|^2_{0,\Omega}\,.$$

The error estimate (3.34) yields

$$\begin{aligned}\|\mathbf{v}_0-\mathbf{v}_{0,h}\|_{0,\Omega_1}^2+\|\mathbf{u}_1-\mathbf{u}_{1,h}\|_{0,\Omega_2}^2\\ =Ch^{2r}(\|\mathbf{v}_0\|_{r+1,\Omega_1}^2+\|\mathbf{u}_1\|_{r+1,\Omega_2}^2+\|p_0\|_{r,\Omega_1}^2).\end{aligned}\tag{4.33}$$

Equation (3.19) and the approximation properties imply

$$\|\mathbf{u}_0-\mathbf{u}_{0,h}\|_{1,\Omega_2}^2\le Ch^{2r}\|\mathbf{u}_0\|_{r+1,\Omega_1}^2.\tag{4.34}$$

Also, by virtue of (4.24), we have

$$\begin{aligned}\|\boldsymbol{\xi}(t_0)-\mathcal{P}^h\boldsymbol{\xi}(t_0)\|_{0,\Omega}^2&\le Ch^{2r-2\epsilon}\Big(\|\mathbf{v}(t_0)\|_{r,\Omega_1}^2+\|\partial_t\mathbf{u}(t_0)\|_{r,\Omega_2}^2\Big)\\ &\le Ch^{2r-2\epsilon}\Big(\|\mathbf{v}\|_{L^2(0,T;\mathbf{H}^{r+1}(\Omega_1))}^2+\|\partial_t\mathbf{v}\|_{L^2(0,T;\mathbf{H}^{r-1}(\Omega_1))}^2\\ &\qquad+\|\mathbf{u}\|_{L^2(0,T;\mathbf{H}^{r+1}(\Omega_2))}^2+\|\partial_t\mathbf{u}\|_{L^2(0,T;\mathbf{H}^{r-1}(\Omega_2))}^2\Big).\end{aligned}\tag{4.35}$$

Thus, utilizing (4.33)–(4.35), (4.23), and (3.9), we may simplify (4.32) to

$$\begin{aligned}&\rho_1\|\mathbf{v}(t)-\mathbf{v}_h(t)\|_{0,\Omega_1}^2+k_1\|\mathbf{v}-\mathbf{v}_h\|_{L^2(0,T;\mathbf{H}^1(\Omega_1))}^2\\ &\qquad+\rho_2\|\partial_t\mathbf{u}(t)-\partial_t\mathbf{u}_h(t)\|_{0,\Omega_2}^2+\|\mathbf{u}(t)-\mathbf{u}_h(t)\|_{1,\Omega_2}^2\\ &\le Ch^{2r}\Big(\|\mathbf{v}_0\|_{r+1,\Omega_1}^2+\|\mathbf{u}_1\|_{r+1,\Omega_2}^2+\|p_0\|_{r,\Omega_1}^2+\|\mathbf{u}_0\|_{r+1,\Omega_1}^2\\ &\qquad+\|p\|_{L^2(0,T;H^r(\Omega_1))}^2\Big)+Ch^{2r-2\epsilon}\Big(\|\mathbf{v}\|_{L^2(0,T;\mathbf{H}^{r+1}(\Omega_1))}^2\\ &\qquad+\|\partial_t\mathbf{v}\|_{L^2(0,T;\mathbf{H}^{r-1}(\Omega_1))}^2+\|\mathbf{u}\|_{L^2(0,T;\mathbf{H}^{r+1}(\Omega_2))}^2\\ &\qquad+\|\partial_t\mathbf{u}\|_{L^2(0,T;\mathbf{H}^{r-1}(\Omega_2))}^2\Big)+\int_0^t\|\mathbf{u}(s)-\mathbf{u}_h(s)\|_{1,\Omega_2}^2ds.\end{aligned}\tag{4.36}$$

By dropping the first three terms on the left-hand side of (4.36) and applying the Gronwall's inequality (3.53), we obtain

$$\begin{aligned}\|\mathbf{u}(t)-\mathbf{u}_h(t)\|_{1,\Omega_2}^2&\le Ce^{CT}h^{2r}\Big[\|\mathbf{v}_0\|_{r+1,\Omega_1}^2+\|\mathbf{u}_1\|_{r+1,\Omega_2}^2\\ &\qquad+\|p_0\|_{r,\Omega_1}^2+\|\mathbf{u}_0\|_{r+1,\Omega_1}^2+\|p\|_{L^2(0,T;H^r(\Omega_1))}^2\Big]\\ &\qquad+Ce^{CT}h^{2r-2\epsilon}\Big(\|\mathbf{v}\|_{L^2(0,T;\mathbf{H}^{r+1}(\Omega_1))}^2+\|\partial_t\mathbf{v}\|_{L^2(0,T;\mathbf{H}^{r-1}(\Omega_1))}^2\\ &\qquad+\|\mathbf{u}\|_{L^2(0,T;\mathbf{H}^{r+1}(\Omega_2))}^2+\|\partial_t\mathbf{u}\|_{L^2(0,T;\mathbf{H}^{r-1}(\Omega_2))}^2\Big).\end{aligned}\tag{4.37}$$

Hence, (4.25) follows from (4.36)–(4.37). ∎

Appendix. Error estimates for the weighted L^2 projection onto $\boldsymbol{\Psi}^h$. The objective of this subsection is to prove error estimates (4.23)–(4.24) for the weighted L^2 projection operator $\mathcal{P}^h$ defined by (4.19).

We introduce an operator $\mathcal{S}^h:\boldsymbol{\Psi}\to\boldsymbol{\Psi}^h$ as follows. For each $\boldsymbol{\zeta}\in\boldsymbol{\Psi}\subset\mathbf{H}_0^1(\Omega)$,

$$\mathcal{S}^h\boldsymbol{\zeta}=\begin{cases}\boldsymbol{\zeta}_{1,h} & \text{in } \Omega_1,\\ \boldsymbol{\zeta}_{2,h} & \text{in } \Omega_2,\end{cases}\tag{A.1}$$

where $\boldsymbol{\zeta}_{1,h} \in X_1^h$ together with some $\sigma_h \in Q_1^h$ is the finite element solution of

$$\begin{cases} a_1[\boldsymbol{\zeta}_{1,h}, \mathbf{z}_h] + b[\mathbf{z}_h, \overline{\sigma}_h] = [\boldsymbol{\zeta}, \mathbf{z}_h] \qquad \forall\, \mathbf{z}_h \in X_1^h \cap \mathbf{H}_0^1(\Omega_1), \\ b[\boldsymbol{\zeta}_{1,h}, q_h] = 0 \qquad \forall\, q_h \in Q_1^h \cap L_0^2(\Omega_1), \\ \boldsymbol{\zeta}_{1,h}|_{\Gamma_1} = \mathbf{0} \qquad \text{and} \qquad [\boldsymbol{\zeta}_{1,h}, \mathbf{s}_h]_{0,\Gamma_0} = [\boldsymbol{\zeta}, \mathbf{s}_h]_{0,\Gamma_0} \quad \forall\, \mathbf{s}_h \in X_1^h|_{\Gamma_0}, \end{cases}$$

and $\boldsymbol{\zeta}_{2,h} \in X_2^h$ is the finite element solution of

$$\begin{cases} [\nabla \boldsymbol{\zeta}_{2,h}, \nabla \mathbf{w}_h]_{\Omega_2} = [\nabla \boldsymbol{\zeta}, \nabla \mathbf{w}_h]_{\Omega_2} \qquad \forall\, \mathbf{w}_h \in X_2^h \cap \mathbf{H}_0^1(\Omega_2), \\ \boldsymbol{\zeta}_{2,h}|_{\Gamma_2} = \mathbf{0} \qquad \text{and} \qquad [\boldsymbol{\zeta}_{2,h}, \mathbf{s}_h]_{0,\Gamma_0} = [\boldsymbol{\zeta}, \mathbf{s}_h]_{0,\Gamma_0} \quad \forall\, \mathbf{s}_h \in X_2^h|_{\Gamma_0}. \end{cases}$$

Evidently, $\boldsymbol{\zeta}_{1,h}|_{\Gamma_0} = \boldsymbol{\zeta}_{2,h}|_{\Gamma_0}$, so that $\mathcal{S}^h\boldsymbol{\zeta}$ defined by (A.1) indeed satisfies $\mathcal{S}^h\boldsymbol{\zeta} \in \boldsymbol{\Psi}^h$.

Using the results of [23, 25] concerning error estimates for the finite element approximations of the Stokes equations (noting that $\operatorname{div} \boldsymbol{\zeta}|_{\Omega_1} = 0$) with inhomogeneous boundary conditions, we obtain

$$\|\boldsymbol{\zeta}_{1,h} - \boldsymbol{\zeta}\|_{1,\Omega_1} \le Ch^r \|\boldsymbol{\zeta}\|_{r+1,\Omega_1} \quad \text{if } \boldsymbol{\zeta}|_{\Omega_1} \in \mathbf{H}^{r+1}(\Omega_1). \tag{A.2}$$

Furthermore, under assumption (H1), we may adapt straightforwardly the proof in [23] for an Aubin–Nitsche-type result to obtain

$$\|\boldsymbol{\zeta}_{1,h} - \boldsymbol{\zeta}\|_{0,\Omega_1} \le Ch^{1-\epsilon_1} \|\boldsymbol{\zeta}_{1,h} - \boldsymbol{\zeta}\|_{1,\Omega_1}. \tag{A.3}$$

Likewise,

$$\|\boldsymbol{\zeta}_{2,h} - \boldsymbol{\zeta}\|_{1,\Omega_2} \le Ch^r \|\boldsymbol{\zeta}\|_{r+1,\Omega_2} \quad \text{if } \boldsymbol{\zeta}|_{\Omega_2} \in \mathbf{H}^{r+1}(\Omega_2), \tag{A.4}$$

and, under assumption (H2),

$$\|\boldsymbol{\zeta}_{2,h} - \boldsymbol{\zeta}\|_{0,\Omega_2} \le Ch^{1-\epsilon_2} \|\boldsymbol{\zeta}_{2,h} - \boldsymbol{\zeta}\|_{1,\Omega_2}. \tag{A.5}$$

To summarize, we have the following results.

PROPOSITION A.1. *If $\boldsymbol{\zeta} \in \boldsymbol{\Psi}$ and $\boldsymbol{\zeta}|_{\Omega_i} \in \mathbf{H}^{r+1}(\Omega_i)$ $(i = 1, 2)$ for some $r \in [0, k]$, then*

$$\|\mathcal{S}^h\boldsymbol{\zeta} - \boldsymbol{\zeta}\|_{1,\Omega} \le Ch^r (\|\boldsymbol{\zeta}\|_{r+1,\Omega_1} + \|\boldsymbol{\zeta}\|_{r+1,\Omega_2}). \tag{A.6}$$

If, in addition, assumptions (H1)–(H2) *hold, then*

$$\|\mathcal{S}^h\boldsymbol{\zeta} - \boldsymbol{\zeta}\|_{0,\Omega} \le Ch^{1-\epsilon} \|\mathcal{S}^h\boldsymbol{\zeta} - \boldsymbol{\zeta}\|_{1,\Omega}, \tag{A.7}$$

where $\epsilon = \max\{\epsilon_1, \epsilon_2\}$

The following proposition establishes relationships between approximation properties for the operator $\mathcal{P}^h$ and those for the operator $\mathcal{S}^h$.

PROPOSITION A.2. *Assume that* (H1)–(H2) *hold. Then,*

$$\|\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}\|_{1,\Omega} \le Ch^{-\epsilon} \|\boldsymbol{\zeta} - \mathcal{S}^h\boldsymbol{\zeta}\|_{1,\Omega} \qquad \forall\, \boldsymbol{\zeta} \in \boldsymbol{\Psi}. \tag{A.8}$$

Proof. Let $\boldsymbol{\zeta} \in \boldsymbol{\Psi}$ be given. The best approximation property of a projection operator implies that

$$\|\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}\|_{0,\Omega} \le \|\mathcal{S}^h\boldsymbol{\zeta} - \boldsymbol{\zeta}\|_{0,\Omega}. \tag{A.9}$$

Using the triangle inequality, the inverse inequality (3.13), and inequality (A.9), we deduce that

$$\begin{aligned}\|\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}\|_{1,\Omega} &\le \|\boldsymbol{\zeta} - \mathcal{S}^h\boldsymbol{\zeta}\|_{1,\Omega} + \|\mathcal{S}^h\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}\|_{1,\Omega} \\ &\le \|\boldsymbol{\zeta} - \mathcal{S}^h\boldsymbol{\zeta}\|_{1,\Omega} + \frac{C}{h}\|\mathcal{S}^h\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}\|_{0,\Omega} \\ &\le \|\boldsymbol{\zeta} - \mathcal{S}^h\boldsymbol{\zeta}\|_{1,\Omega} + \frac{C}{h}\|\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}\|_{0,\Omega} + \frac{C}{h}\|\mathcal{S}^h\boldsymbol{\zeta} - \boldsymbol{\zeta}\|_{0,\Omega} \\ &\le \|\boldsymbol{\zeta} - \mathcal{S}^h\boldsymbol{\zeta}\|_{1,\Omega} + \frac{C}{h}\|\mathcal{S}^h\boldsymbol{\zeta} - \boldsymbol{\zeta}\|_{0,\Omega}.\end{aligned}$$

Thus, (A.8) follows from the last inequality and (A.7). □

Finally, as obvious consequences of (A.8) and (A.6)–(A.7), we obtain the following error estimates for $\boldsymbol{\zeta} - \mathcal{P}^h\boldsymbol{\zeta}$:

THEOREM A.3. *Assume that* (H1)–(H2) *hold. Then the operator* $\mathcal{P}^h$ *satisfies the error estimates* (4.23) *and* (4.24).

REFERENCES

[1] M. BERGGREN, *Approximations of Very Weak Solutions to Boundary-Value Problems*, SIAM J. Numer. Anal., to appear.

[2] A. BERMÚDEZ, R. DURÁN, AND R. RODRÍGUEZ, *Finite element analysis of compressible and incompressible fluid-solid systems*, Math. Comp., 67 (1998), pp. 111–136.

[3] F. BLOM, *A monolithical fluid-structure interaction algorithm applied to the piston problem*, Comput. Methods Appl. Mech. Engrg., 167 (1998), pp. 369–391.

[4] J. BOUJOT, *Mathematical formulation of fluid-structure interaction problems*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 239–260.

[5] P. CIARLET, *The Finite Element Method For Elliptic Problems*, North–Holland, Amsterdam, 1978.

[6] C. CONCA AND M. DURÁN, *A numerical study of a spectral problem in solid-fluid type structures* Numer. Methods Partial Differential Equations, 11 (1995), pp. 423–444.

[7] C. CONCA, J. MARTÍN, AND J. TUCSNAK, *Motion of a rigid body in a viscous fluid*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 473–478.

[8] S. DASSER, *A penalization method for the homogenization of a mixed fluid-structure problem*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 759–764.

[9] B. DESJARDINS AND M. J. ESTEBAN, *On weak solutions for fluid-rigid structure interaction: Compressible and incompressible models*, Comm. Partial Differential Equations, 25 (2000), pp. 1399–1413.

[10] Q. DU, M. GUNZBURGER, L. HOU, AND J. LEE, *Analysis of a Linear Fluid-Structure Interaction Problem*, Disc. Cont. Dyn. Syst., 9 (2003), pp. 633–650.

[11] D. ERRATE, M. J. ESTEBAN, AND Y. MADAY, *Couplage fluid-structure, Un modéle simplifié en dimension* 1, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 275–281.

[12] L. C. EVANS, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[13] C. FARHAT, M. LESOINNE, AND P. LETALLEC, *Load and motion transfer algorithms for fluid/structure interaction problems with non-matching discrete interfaces: Momentum and energy conservation, optimal discretization and application to aeroelasticity*, Comput. Methods Appl. Mech. Engrg., 157 (1998), pp. 95–114.

[14] F. FLORI AND P. ORENGA, *On a fluid-structure interaction problem*, in Trends in Applications of Mathematics to Mechanics (Nice, 1998), Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 293–305.

[15] F. FLORI AND P. ORENGA, *Analysis of a nonlinear fluid-structure interaction problem in velocity-displacement formulation*, Nonlinear Anal., 35 (1999), pp. 561–587.

[16] F. FLORI AND P. ORENGA, *On a nonlinear fluid-structure interaction problem defined on a domain depending on time*, Nonlinear Anal., 38 (1999), pp. 549–569.

[17] F. FLORI AND P. ORENGA, *Fluid-structure interaction: Analysis of a* 3*-D compressible model*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 753–777.

[18] L. GASTALDI, *Mixed finite element methods in fluid structure system*, Numer. Math., 74 (1996), pp. 153–176.

[19] V. Girault and P. Raviart, *Finite Element Methods For Navier–Stokes Equations*, Springer, Berlin, 1986.
[20] C. Grandmont, *Existence and uniqueness for a two-dimensional steady-state fluid-structure interaction problem*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 651–656.
[21] C. Grandmont and Y. Maday, *Existence for an unsteady fluid-structure interaction problem*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 609–636.
[22] C. Grandmont and Y. Maday, *Existence of solutions for a two-dimensional unsteady fluid-structure interaction problem*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 525–530.
[23] M. D. Gunzburger and L. S. Hou, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424.
[24] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[25] L. Hou, *Error estimates for semidiscrete finite element approximations of the Stokes equations under minimal regularity assumptions*, J. Sci. Comput., 16 (2001), pp. 287–317.
[26] G. Hsiao, R. Kleinman and G. Roach, *Weak solutions of fluid-solid interaction problems*, Math. Nachr., 218 (2000), pp.139–163.
[27] P. LeTallec and S. Mani, *Numerical analysis of a linearized fluid-structure interaction problem*, Numer. Math., 87 (2000), pp. 317-354.
[28] H. Morand and R. Ohayon, *Fluid Structure Interaction: Applied Numerical Methods*, Wiley, New York, 1995.
[29] S. Micu and E. Zuazua, *Asymptotics for the spectrum of a fluid/structure hybrid system arising in the control of noise*, SIAM J. Math. Anal., 29 (1998), pp. 967–1001.
[30] R. Ohayon and C. Felippa, eds., *Special Issue on Fluid-Structure Interactions*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2977–3292.
[31] P.D. Panagiotopoulos, *Inequality Problems in Mechanics and Applications*, Birkhäuser Boston, Cambridge, MA, 1985.
[32] R. Rodríguez and J. Solomin, *The order of convergence of eigenfrequencies in finite element approximations of fluid-structure interaction problems*, Math. Comp., 65 (1996), pp. 1463–1475.
[33] M. Rumpf, *On equilibria in the interaction of fluids and elastic solids. Theory of the Navier-Stokes equations*, Ser. Adv. Math. Appl. Sci., 47 (1998), pp. 136–158.
[34] R. Schulkes, *Interactions of an elastic solid with a viscous fluid: Eigenmode analysis*, J. Comput. Phys., 100 (1992), pp. 270–283.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 30–48

ON THE NORMS OF INVERSES OF PSEUDOSPECTRAL DIFFERENTIATION MATRICES*

DAVID M. SLOAN†

Abstract. In this paper we give integral expressions for the elements of the inverses of second-order pseudospectral differentiation matrices. Simple upper bounds are given for the maximum norms of these inverse matrices when Chebyshev collocation points are used. Comment is made on the failure to obtain upper bounds that are uniform in the number of collocation points when the points are evenly spaced. We also give integral expressions for inverses of first-order Chebyshev pseudospectral differentiation matrices.

Key words. pseudospectral, differentiation matrices, Chebyshev

AMS subject classifications. 65D25, 65F35, 65N35

DOI. 10.1137/S0036142902414542

1. Introduction. Spectral collocation provides powerful methods—known as pseudospectral methods—for computing approximate solutions of ordinary and partial differential equations. Over the last two decades, pseudospectral methods have emerged as viable alternatives in many situations to finite difference and finite element methods. In the pseudospectral approach, the unknown solution of the differential equation is approximated by a global interpolant, such as an algebraic or trigonometric polynomial of high degree. The derivatives appearing in the differential equation are approximated by exact differentiation of the interpolant, and the unknown coefficients in the interpolant are then obtained by setting the residual to zero at an appropriate number of collocation points in the domain of the problem. Since interpolation and differentiation are linear operations, the process of obtaining approximations to the values of the derivative of a function at the set of collocation points can be expressed as a matrix-vector multiplication: the matrices involved are called *pseudospectral differentiation matrices*. A Chebyshev pseudospectral differentiation matrix is one for which the domain in the direction of differentiation is the finite interval $[-1, 1]$, the interpolant is an algebraic polynomial, and the collocation points are chosen to be the extrema or zeros of a Chebyshev polynomial.

The aim of this paper is to give integral expressions for the elements of the inverses of second-order pseudospectral differentiation matrices, and to show that these inverse matrices have simple upper bounds in the maximum norm if the collocation points are Chebyshev extrema or zeros. Integral expressions are also given for the elements of the inverses of first-order pseudospectral differentiation matrices. Comments are made on the failure to obtain upper bounds for the inverse matrices when the collocation points are evenly spaced and the interpolant is an algebraic polynomial.

The main advantage of pseudospectral methods for approximate solution of differential equations is their high accuracy for problems whose solutions are sufficiently smooth. Such methods converge exponentially fast [3, 11, 14], whereas finite differences and finite elements have algebraic rates of convergence. In practical computations this means that high accuracy can be achieved with relatively coarse

*Received by the editors September 16, 2002; accepted for publication (in revised form) May 21, 2003; published electronically January 6, 2004.
http://www.siam.org/journals/sinum/42-1/41454.html
†Department of Mathematics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland (d.sloan@strath.ac.uk).

discretizations. Disadvantages include the occurrence of full rather than sparse differentiation matrices, and less flexibility when dealing with irregular domains. For a full treatment of numerical solution of differential equations by pseudospectral methods, the reader is referred to [2, 3, 7, 8, 9].

The computation of pseudospectral differentiation matrices for derivatives of arbitrary order has been considered by Huang and Sloan [12] and by Welfert [16]. A MATLAB package by Weideman and Reddy [15] may be used to generate pseudospectral differentiation matrices and to solve differential equations by the pseudospectral method. FORTRAN packages for pseudospectral computations have been produced by Funaro [10] and by Don and Solomonoff [6].

The outline of this paper is as follows. Section 2 presents integral expressions for the elements of second-order pseudospectral differentiation matrices for algebraic polynomial interpolants and arbitrary distributions of collocation points. Upper bounds on the inverse matrices are given in the maximum norm when the collocation points are extrema or zeros of Chebyshev polynomials. It is also shown that the elements become unbounded as the number of collocation points tends to infinity when the points are evenly spaced. Section 3 presents integral expressions for the elements of first-order pseudospectral differentiation matrices, and conclusions and comments are given in section 4. The analysis leading to the upper bound on the inverses of second-order matrices depends on a negativity property of the elements, and this property is examined in the appendix.

2. Second-order differentiation matrices.

2.1. Background. We consider pseudospectral differentiation of nonperiodic functions on the domain $[-1, 1]$. Let $\{x_j\}_{j=0}^N$ be a set of $N+1$ distinct collocation points, or nodes, satisfying

$$-1 = x_0 < x_1 < \cdots < x_{N-1} < x_N = 1 \tag{2.1}$$

but otherwise arbitrary. If u is a real function defined on this set of nodes, the approximating polynomial interpolant $p_N \in P_N$ is defined by

$$p_N(x) = \sum_{j=0}^{N} u(x_j)\, L_j^{(N)}(x), \tag{2.2}$$

where P_N is the set of real polynomials of degree at most N and $L_j^{(N)}(x)$ is the Lagrange interpolation polynomial

$$L_j^{(N)}(x) = \prod_{\substack{i=0 \\ i\neq j}}^{N} \left(\frac{x - x_i}{x_j - x_i} \right). \tag{2.3}$$

$L_j^{(N)}$ is determined by the conditions

$$L_j^{(N)}(x_i) = \begin{cases} 0 & \text{if} \quad j \neq i, \\ 1 & \text{if} \quad j = i, \end{cases} \quad \text{for } 0 \leq i, j \leq N. \tag{2.4}$$

For $0 \leq i \leq N$, the second-order derivative of u at $x = x_i$ is approximated by $p_N''(x_i)$. Now set

$$\mathbf{u} := [u(x_0), u(x_1), \ldots, u(x_N)]^T \quad \text{and} \quad \mathbf{w} := [p_N''(x_0), p_N''(x_1), \ldots, p_N''(x_N)]^T,$$

and it follows from (2.2) that

$$\mathbf{w} = D_N^{(2)}\,\mathbf{u}, \tag{2.5}$$

where $D_N^{(2)} \in \mathbb{R}^{(N+1)\times(N+1)}$, with the element of row index i and column index j given by

$$D_N^{(2)}(i,j) = \left(\frac{d^2\,L_j^{(N)}}{dx^2}\right)(x_i) \quad \text{for } 0 \le i,j \le N. \tag{2.6}$$

The vector $\mathbf{w}$ is the discrete second derivative vector, and $D_N^{(2)}$ is the second-order pseudospectral differentiation matrix based on the nodes (2.1).

It is readily shown using (2.2) and (2.3) that $D_N^{(2)}\,\mathbf{v}_0 = D_N^{(2)}\,\mathbf{v}_1 = \mathbf{0}$, where $\mathbf{v}_0$ and $\mathbf{v}_1$ are vectors in $\mathbb{R}^{(N+1)}$ defined by $\mathbf{v}_0 = [1,1,\ldots,1]^T$ and $\mathbf{v}_1 = [x_0,x_1,\ldots,x_N]^T$, and it follows that $D_N^{(2)}$ is singular. To obtain the nonsingular differentiation matrix that arises in pseudospectral solution of differential equations, consider the simple one-dimensional Poisson equation

$$\frac{d^2u}{dx^2} = f(x), \quad -1 < x < 1, \quad u(\pm 1) = 0,$$

where $f \in C[-1,1]$. If u_j denotes the pseudospectral approximation to u at $x = x_j$ $(j = 0,1,\ldots,N)$, then $u_0 := 0$, $u_N := 0$, and the approximation to u'' at $x = x_i$ is given by (2.2) as $p_N''(x_i)$, where $p_N(x) = \sum_{j=1}^{N-1} u_j\,L_j^{(N)}(x)$. The unknown coefficients $\{u_j\}_{j=1}^{N-1}$ are obtained from the conditions

$$p_N''(x_i) = f(x_i), \quad 1 \le i \le N-1.$$

This system may be written as

$$\tilde{D}_N^{(2)}\,\mathbf{v} = \mathbf{f}, \tag{2.7}$$

where $\mathbf{f} = [f(x_1), f(x_2), \ldots, f(x_{N-1})]^T$, $\mathbf{v} = [u_1, u_2, \ldots, u_{N-1}]^T$, and $\tilde{D}_N^{(2)} \in \mathbb{R}^{(N-1)\times(N-1)}$ is obtained by deleting the first and last rows and columns from $D_N^{(2)}$. Henceforth we shall refer to $\tilde{D}_N^{(2)}$ as the second-order pseudospectral differentiation matrix based on the nodes (2.1). The matrix $\tilde{D}_N^{(2)}$ is nonsingular: the objective now is to give expressions for the elements of its inverse and to obtain an upper bound for the inverse when the nodes (2.1) are suitably chosen.

2.2. Formation of $(\tilde{D}_N^{(2)})^{-1}$. Let $\psi_j^{(N-2)} \in P_{N-2}$ be the Lagrange polynomial satisfying

$$\psi_j^{(N-2)}(x_i) = \begin{cases} 0 & \text{if} \quad j \ne i, \\ 1 & \text{if} \quad j = i, \end{cases} \quad \text{for} \;\; 1 \le i,\, j \le N-1. \tag{2.8}$$

Now define $\phi_j^{(N)} \in P_N$ by

$$\left.\begin{aligned} &\frac{d^2\phi_j^{(N)}}{dx^2}(x) = \psi_j^{(N-2)}(x), \quad j = 1,2,\ldots,N-1, \\ &\text{subject to} \;\; \phi_j^{(N)}(\pm 1) = 0. \end{aligned}\right\} \tag{2.9}$$

It is clear from the construction of $D_N^{(2)}$ that if $\phi_j \in P_N$ and $\boldsymbol{\phi}_j = [\phi_j(x_0), \phi_j(x_1), \ldots, \phi_j(x_N)]^T$, then $D_N^{(2)}\,\boldsymbol{\phi}_j = \boldsymbol{\psi}_j$, where $\boldsymbol{\psi}_j = [\phi_j''(x_0), \phi_j''(x_1), \ldots, \phi_j''(x_N)]^T$. In other words, pseudospectral differentiation of order 2 is exact for functions in P_N. (This is true for fixed N and differentiation of any order.)

Equation (2.5), with $\mathbf{w} := \boldsymbol{\psi}_j^{(N-2)} = [\psi_j^{(N-2)}(x_0),\ \psi_j^{(N-2)}(x_1), \ldots, \psi_j^{(N-2)}(x_N)]^T$ and $\mathbf{u} := \boldsymbol{\phi}_j^{(N)} = [\phi_j^{(N)}(x_0),\ \phi_j^{(N)}(x_1), \ldots, \phi_j^{(N)}(x_N)]^T$ becomes the identity

$$D_N^{(2)}\,\boldsymbol{\phi}_j^{(N)} = \boldsymbol{\psi}_j^{(N-2)}.$$

If $\Phi = [\boldsymbol{\phi}_1^{(N)}, \boldsymbol{\phi}_2^{(N)}, \ldots, \boldsymbol{\phi}_{N-1}^{(N)}]$ and $\Psi = [\boldsymbol{\psi}_1^{(N-2)}, \boldsymbol{\psi}_2^{(N-2)}, \ldots, \boldsymbol{\psi}_{N-1}^{(N-2)}]$, we may utilize conditions (2.8) in

$$D_N^{(2)}\,\Phi = \Psi$$

to obtain

$$D_N^{(2)}\,\Phi = \begin{bmatrix} \psi_1^{(N-2)}(x_0) & \psi_2^{(N-2)}(x_0) & \cdots & \psi_{N-1}^{(N-2)}(x_0) \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \psi_1^{(N-2)}(x_N) & \psi_2^{(N-2)}(x_N) & \cdots & \psi_{N-1}^{(N-2)}(x_N) \end{bmatrix}.$$

If we identify elements from rows 2 to N in the above equation, we obtain

$$\tilde{D}_N^{(2)}(\tilde{D}_N^{(2)})^{-1} = I,$$

where

$$(\tilde{D}_N^{(2)})^{-1}(i,j) = \phi_j^{(N)}(x_i), \quad 1 \le i,\, j \le N-1. \tag{2.10}$$

The elements of the inverse of the matrix $\tilde{D}_N^{(2)}$ are thus given by solutions of the differential equations (2.9).

The solution of (2.9) may be written as

$$\phi_j^{(N)}(x) = \int_{-1}^{1} G(z;x)\psi_j^{(N-2)}(z)dz, \quad 1 \le j \le N-1, \tag{2.11}$$

where $G(z;x)$ is the Green's function defined by

$$G(z;x) = \begin{cases} \dfrac{(z+1)(x-1)}{2}, & -1 \le z \le x, \\ \dfrac{(z-1)(x+1)}{2}, & x < z \le 1. \end{cases} \tag{2.12}$$

Note that $-\frac{1}{2} \le G(z;x) < 0$ for $z, x \in (-1,1)$.

An equivalent representation is

$$\phi_j^{(N)}(x) = \int_{-1}^{x} (x-z)\psi_j^{(N-2)}(z)dz + \frac{(x+1)}{2}\int_{-1}^{1} (z-1)\psi_j^{(N-2)}(z)dz. \tag{2.13}$$

The elements $(\tilde{D}_N^{(2)})^{-1}(i,j),\ 1 \le i,j \le N-1$, are now obtained by evaluating $\phi_j^{(N)}(x)$ at $x = x_i$.

2.3. Symmetry properties of $(\tilde{D}_N^{(2)})^{-1}$. Henceforth we shall assume that the nodes (2.1) are symmetrically located around $x = 0$. Thus $x_j = -x_{N-j}$ for $1 \le j \le N/2$, and there is a node at $x = 0$ if and only if N is even. With a nodal distribution of this type it is readily seen that $L_j^{(N)}(x) = L_{N-j}^{(N)}(-x)$. Hence

$$\begin{aligned}\tilde{D}_N^{(2)}(i,j) &= \frac{d^2 L_j^{(N)}}{dx^2}(x_i) = \frac{d^2 L_{N-j}^{(N)}}{dx^2}(-x_i) = \frac{d^2 L_{N-j}^{(N)}}{dx^2}(x_{N-i}) \\ &= \tilde{D}_N^{(2)}(N-i, N-j),\end{aligned} \tag{2.14}$$

which shows that $\tilde{D}_N^{(2)}$ is a *centrosymmetric matrix* [1]. The centrosymmetric property may be written as $\tilde{D}_N^{(2)} = P\tilde{D}_N^{(2)}P$, with

$$P = P^T = P^{-1} = \begin{bmatrix} & & & & 1 \\ & & & \cdot & \\ & & \cdot & & \\ & \cdot & & & \\ 1 & & & & \end{bmatrix}.$$

Clearly

$$(\tilde{D}_N^{(2)})^{-1} = P^{-1}(\tilde{D}_N^{(2)})^{-1}P^{-1} = P(\tilde{D}_N^{(2)})^{-1}P,$$

and it follows that $(\tilde{D}_N^{(2)})^{-1}$ is also centrosymmetric.

2.4. Upper bounds on $(\tilde{D}_N^{(2)})^{-1}$. We now consider three specific distributions of the collocation points (2.1):

$$\begin{aligned}&\text{(a)}\quad \textit{Chebyshev extrema} \quad : \quad x_j := t_j = -\cos\left(\frac{\pi j}{N}\right), \quad j = 0, 1, \ldots, N; \\ &\text{(b)}\quad \textit{Chebyshev zeros} \quad : \quad x_0 := s_0 = -1,\ x_N := s_N = +1, \quad \text{and} \\ &\qquad\qquad\qquad\qquad\quad x_j := s_j = -\cos\left(\frac{\pi(2j-1)}{2(N-1)}\right), \quad j = 1, 2, \ldots, N-1; \\ &\text{(c)}\quad \textit{Evenly spaced} \quad : \quad x_j := \rho_j = -1 + 2 \times \frac{j}{N}, \quad j = 0, 1, \ldots, N.\end{aligned} \tag{2.15}$$

The points $\{t_j\}_{j=0}^N$ are the extrema of the Chebyshev polynomial T_N in $[-1, 1]$, and $\{s_j\}_{j=1}^{N-1}$ are the zeros of T_{N-1} (see [13]). The distributions (a) and (b) are popular choices for pseudospectral solution of differential equations [2, 3, 7, 8].

In the case of nodes based on Chebyshev extrema, it is readily shown that (see [4, 13])

$$\psi_j^{(N-2)}(x) = \frac{T_N'(x)}{(x - t_j)\, T_N''(t_j)}, \quad j = 1, 2, \ldots, N-1,$$

and if this is substituted into (2.11), we obtain

$$\phi_j^{(N)}(x) = \frac{1}{T_N''(t_j)} \int_{-1}^{1} \frac{G(z;x) T_N'(z)}{(z - t_j)}\, dz, \quad j = 1, 2, \ldots, N-1. \tag{2.16}$$

With nodes based on Chebyshev zeros, we find

$$\psi_j^{(N-2)}(x) = \frac{T_{N-1}(x)}{(x - s_j) T_{N-1}'(s_j)}, \quad j = 1, 2, \ldots, N-1,$$

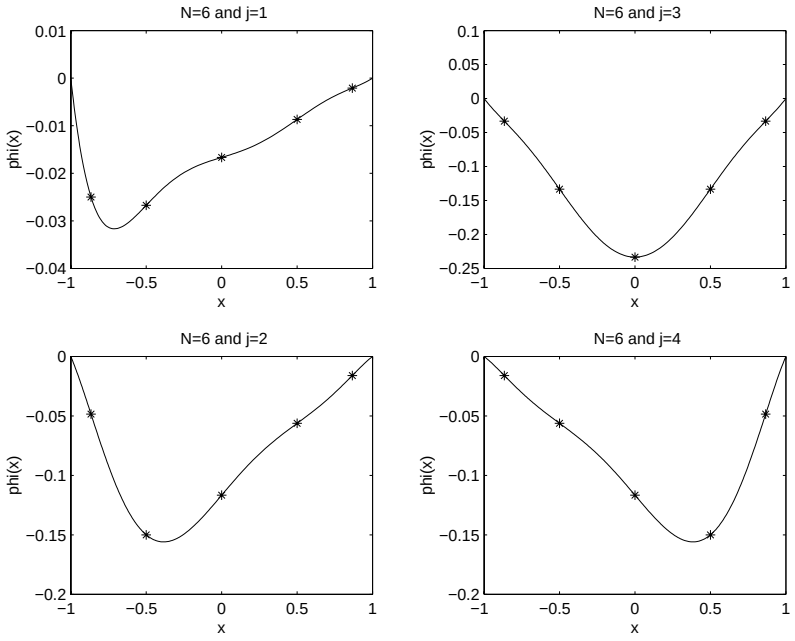


FIG. 1. $\phi_j^{(N)}(x)$ *corresponding to Chebyshev extrema nodes.* $N = 6$ *throughout, and* $j = 1, 3, 2,$ *and* 4 *as indicated above each panel.*

which yields

$$\phi_j^{(N)}(x) = \frac{1}{T'_{N-1}(s_j)} \int_{-1}^{1} \frac{G(z;x)T_{N-1}(z)}{(z - s_j)}\, dz, \quad j = 1, 2, \ldots, N-1. \tag{2.17}$$

The functions $\phi_j^{(N)}(x)$ were determined for a range of values of N and j, using the three nodal distributions given in (2.15). Evaluations are readily effected using the representation (2.13), with the MATLAB routines ode45 and quadl employed to compute the first and second integrals, respectively, in (2.13). Numerical experiments suggest that for both Chebyshev distributions of nodes, $\phi_j^{(N)}(x) < 0$ on $(-1, 1)$ for any positive integer N and for all $j = 1, 2, \ldots, N-1$. For evenly spaced nodes, this negativity result does not hold. To illustrate this observation, Figure 1 shows $\phi_j^{(N)}(x)$ corresponding to the Chebyshev extrema nodes (2.15(a)), with $N = 6$ and $j = 1, 2, 3,$ and 4. The profiles corresponding to the Chebyshev zero nodes have similar qualitative features. The centrosymmetry property of $(\tilde{D}_N^{(2)})^{-1}$ implies that $\phi_j^{(N)}(x) = \phi_{N-j}^{(N)}(-x)$ for $1 \le j \le N-1$, and this symmetry is seen in the plots corresponding to $N = 6$, $j = 2$ and $N = 6$, $j = 4$. In Figure 2, $\phi_j^{(N)}(x)$ is shown for the same values of N and j on evenly spaced nodes. Here again, the property $\phi_j^{(N)}(x) = \phi_{N-j}^{(N)}(-x)$ is illustrated by the plots $N = 6$, $j = 2$ and $N = 6$, $j = 4$. In the case of evenly spaced nodes, $\phi_j^{(N)}(x)$ can take positive values; in particular, elements of the inverse differentiation matrix are given by the values of $\phi_j^{(N)}(x)$ at the asterisks, and some of these values are positive. Thus, the computational results show that $(\tilde{D}_N^{(2)})^{-1}$ is a negative matrix for Chebyshev nodes, but not necessarily for evenly spaced nodes.

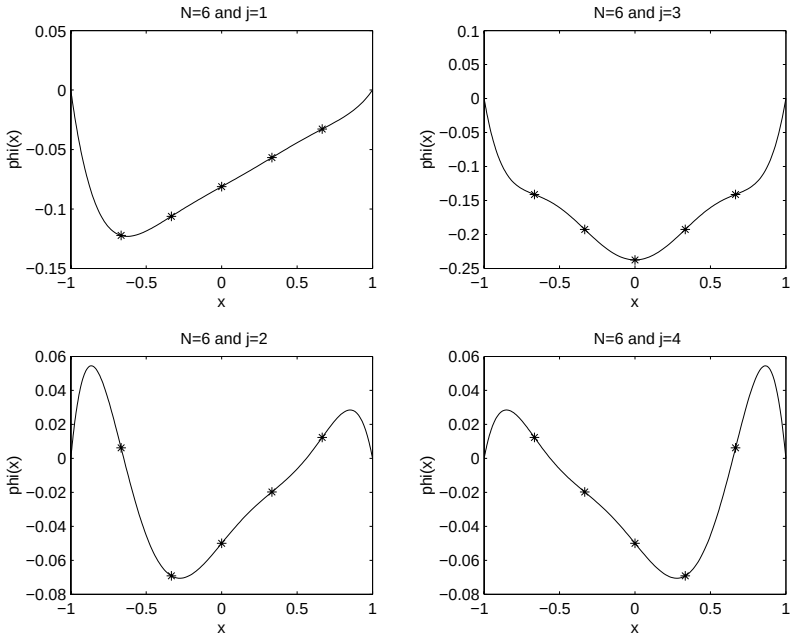


FIG. 2. $\phi_j^{(N)}(x)$ *corresponding to evenly spaced nodes.* $N = 6$ *throughout, and* $j = 1, 3, 2$ *and* 4 *as indicated above each panel.*

TABLE 1
$(\tilde{D}_6^{(2)})^{-1}$ *for Chebyshev extrema nodes.*

$$-10^{-1} \times \begin{bmatrix} 0.250 & 0.485 & 0.333 & 0.161 & 0.021 \\ 0.267 & 1.500 & 1.333 & 0.562 & 0.087 \\ 0.167 & 1.167 & 2.333 & 1.167 & 0.167 \\ 0.087 & 0.562 & 1.333 & 1.500 & 0.267 \\ 0.021 & 0.161 & 0.333 & 0.485 & 0.250 \end{bmatrix}$$

TABLE 2
$(\tilde{D}_6^{(2)})^{-1}$ *for evenly spaced nodes.*

$$-10^{-1} \times \begin{bmatrix} 1.223 & -0.062 & 1.412 & -0.123 & 0.328 \\ 1.062 & 0.691 & 1.926 & 0.198 & 0.568 \\ 0.812 & 0.500 & 2.375 & 0.500 & 0.812 \\ 0.568 & 0.198 & 1.926 & 0.691 & 1.062 \\ 0.328 & -0.123 & 1.412 & -0.062 & 1.223 \end{bmatrix}$$

The inverse matrices $(\tilde{D}_N^{(2)})^{-1}$ corresponding to $N = 6$ are given in Tables 1 and 2 for Chebyshev extrema nodes and evenly spaced nodes, respectively. The matrices displayed in Tables 1 and 2 were obtained by inverting the appropriate pseudospectral differentiation matrices computed using the MATLAB codes of Weideman and Reddy [15].

Note that $(\tilde{D}_6^{(2)})^{-1}$ is negative in Table 1, but there are positive elements in Table 2. The matrices are centrosymmetric for both sets of nodes.

To obtain upper bounds for $(\tilde{D}_N^{(2)})^{-1}$ we first consider the row sums of the inverse matrix. The sum of the elements in row i $(i = 1, 2, \ldots, N-1)$ of $(\tilde{D}_N^{(2)})^{-1}$, for any distribution of nodes (2.1), is given by

$$S_i^{(N)} = \sum_{j=1}^{N-1} \phi_j^{(N)}(x_i). \tag{2.18}$$

Now define $\phi^{(N)}(x) = \sum_{j=1}^{N-1} \phi_j^{(N)}(x)$ and $\psi^{(N-2)}(x) = \sum_{j=1}^{N-1} \psi_j^{(N-2)}(x)$. From (2.9) we see that $\phi^{(N)}$ is defined by

$$\left.\begin{array}{l} \dfrac{d^2\phi^{(N)}}{dx^2}(x) = \psi^{(N-2)}(x), \\ \text{with} \quad \phi^{(N)}(\pm 1) = 0. \end{array}\right\} \tag{2.19}$$

Clearly, $\psi^{(N-2)} \in P_{N-2}$ and $\psi^{(N-2)}(x_i) = 1$ for $i = 1, 2, \ldots, N-1$. The unique polynomial in P_{N-2} satisfying these conditions is $\psi^{(N-2)}(x) \equiv 1$, and it follows from (2.19) that

$$\phi^{(N)}(x) = \frac{1}{2}(x^2 - 1). \tag{2.20}$$

It follows from (2.18) and (2.20) that the row sum is given by

$$S_i^{(N)} = \frac{1}{2}(x_i^2 - 1) \quad \text{for} \quad i = 1, 2, \ldots, N-1. \tag{2.21}$$

To obtain upper bounds for the maximum norms of $(\tilde{D}_N^{(2)})^{-1}$ when the nodes are Chebyshev extrema or Chebyshev zeros, we note that $(\tilde{D}_N^{(2)})^{-1}(i,j) < 0$ for $1 \le i, j \le N-1$. This negativity property, which is established in the appendix, is equivalent to the conditions $\phi_j^{(N)}(t_i) < 0$ and $\phi_j^{(N)}(s_i) < 0$ for $1 \le i, j \le N-1$. The numerical experiments described above and the analysis in the appendix show that $\phi_j^{(N)}(x) < 0$ for $1 \le j \le N-1$ and $x \in (-1, 1)$, with each of the Chebyshev nodal distributions. Under this negativity assumption, it follows from (2.21) that

$$\sum_{j=1}^{N-1} \left| (\tilde{D}_N^{(2)})^{-1}(i,j) \right| = \frac{1}{2}(1 - x_i^2), \quad 1 \le i \le N-1.$$

The maximum value of this quantity occurs at $i = N/2$ if N is even and at $i = (N \pm 1)/2$ if N is odd. The key results are contained in the following theorem.

THEOREM 2.1. *For either of the two Chebyshev nodal distributions in* (2.15),

$$\|(\tilde{D}_N^{(2)})^{-1}\|_\infty = \frac{1}{2} \quad \textit{if} \quad N \quad \textit{is even.} \tag{2.22}$$

If N is odd, then

$$\|(\tilde{D}_N^{(2)})^{-1}\|_\infty = \frac{1}{2}(1 - t_{(N-1)/2}^2) = \frac{1}{2}\cos^2\left(\frac{\pi}{2N}\right) \tag{2.23}$$

for the Chebyshev extrema distribution, and

$$\|(\tilde{D}_N^{(2)})^{-1}\|_\infty = \frac{1}{2}(1 - s_{(N-1)/2}^2) = \frac{1}{2}\cos^2\left(\frac{\pi}{2(N-1)}\right) \tag{2.24}$$

TABLE 3
$\|(\tilde{D}_N^{(2)})^{-1}\|_\infty$ *for evenly spaced nodes.*

N	$\|(\tilde{D}_N^{(2)})^{-1}\|_\infty$
4	0.5000
5	0.4800
6	0.5000
7	0.4898
8	0.5926
⋮	⋮
15	14.2727
16	28.1331
⋮	⋮
32	3.5×10^5

for the Chebyshev zeros distribution. For distributions (2.15(a)) *or* (2.15(b)),

$$\|(\tilde{D}_N^{(2)})^{-1}\|_\infty \le \frac{1}{2}, \quad \textit{with equality if } N \textit{ is even.} \tag{2.25}$$

The values of the norms given in (2.22)–(2.24) are readily verified numerically.

If the nodes are evenly spaced, the inverse matrix does not satisfy the negativity property (unless $N \le 5$), and the bound given in (2.25) does not apply if $N \ge 8$. Table 3 gives $\|(\tilde{D}_N^{(2)})^{-1}\|_\infty$ for evenly spaced nodes at several values of N.

The results show that the norm increases rapidly with N. One expects that nodal distributions with clustering at the ends [8] will be a prerequisite for the existence of bounds on the inverse differentiation matrix. Numerical results show, for example, that if the collocation points $\{s_j\}_{j=1}^{N-1}$ in (2.15) denote the zeros of the Legendre polynomial of degree $N-1$, then $(\tilde{D}_N^{(2)})^{-1})$ is negative. Furthermore, $\|(\tilde{D}_N^{(2)})^{-1}\|_\infty$ takes values $\frac{1}{2}$ or $\frac{1}{2}(1 - s^2_{(N-1)/2})$ for N even or odd.

3. First-order differentiation matrices.

3.1. Background. Here we follow the method adopted in the preceding section to give a brief presentation of expressions for the elements of the inverses of first-order pseudospectral differentiation matrices. The element of row index i and column index j of the first-order pseudospectral differentiation matrix $D_N^{(1)} \in \mathbb{R}^{(N+1)\times(N+1)}$ associated with nodes (2.1) is given by

$$D_N^{(1)}(i,j) = \left(\frac{dL_j^{(N)}}{dx}\right)(x_i) \quad \text{for} \quad 0 \le i,j \le N.$$

In the notation of subsection 2.1, $D_N^{(1)}\mathbf{v}_0 = \mathbf{0}$, and $D_N^{(1)}$ is therefore singular. To obtain a nonsingular differentiation matrix that arises in pseudospectral solution of differential equations, we consider the simple differential problem

$$\frac{du}{dx} = f(x), \quad -1 < x \le 1, \quad u(-1) = 0,$$

where $f \in C[-1,1]$. The pseudospectral solution is obtained from

$$\tilde{D}_N^{(1)}\,\mathbf{v} = \mathbf{f}, \tag{3.1}$$

where $\mathbf{f} = [f(x_1), f(x_2), \ldots, f(x_N)]^T$, $\mathbf{v} = [u_1, u_2, \ldots, u_N]^T$, and $\tilde{D}_N^{(1)} \in \mathbb{R}^{N\times N}$ is obtained by deleting the first row and column from $D_N^{(1)}$. (If the boundary condition of the first-order differential problem is imposed at $x = +1$, then the last row and column of $D_N^{(1)}$ are deleted.)

3.2. Formation of $(\tilde{\boldsymbol{D}}_{\boldsymbol{N}}^{(\mathbf{1})})^{-\mathbf{1}}$. Let $\psi_j^{(N-1)} \in P_{N-1}$ be the Lagrange polynomial satisfying

$$\psi_j^{(N-1)}(x_i) = \begin{cases} 0 & \text{if} \quad j \neq i, \\ 1 & \text{if} \quad j = i, \end{cases} \quad \text{for} \;\; 1 \leq i, j \leq N. \tag{3.2}$$

Now define $\phi_j^{(N)} \in P_N$ by

$$\left.\begin{aligned} &\frac{d\phi_j^{(N)}}{dx}(x) = \psi_j^{(N-1)}(x), \quad j = 1, 2, \ldots, N, \\ &\text{subject to} \;\; \phi_j^{(N)}(-1) = 0. \end{aligned}\right\} \tag{3.3}$$

The notation here is inconsistent with that adopted in section 2; however, the aim in this choice of notation is to simplify the presentation.

Following the steps taken in subsection 2.1, we find that

$$(\tilde{D}_N^{(1)})^{-1}(i, j) = \phi_j^{(N)}(x_i), \quad 1 \leq i, j \leq N. \tag{3.4}$$

The matrix elements are given by the solution of (3.3), which is

$$\phi_j^{(N)}(x) = \int_{-1}^{x} \psi_j^{(N-1)}(z)dz, \quad 1 \leq j \leq N. \tag{3.5}$$

In the case of the Chebyshev extrema nodes (see [13]),

$$\psi_j^{(N-1)}(x) = \frac{(x-1)T_N'(x)}{(t_j-1)(x-t_j)T_N''(t_j)}, \quad 1 \leq j \leq N-1,$$

and

$$\psi_N^{(N-1)}(x) = \frac{T_N'(x)}{T_N'(t_N)}.$$

Hence

$$\phi_j^{(N)}(x) = \frac{1}{(t_j-1)T_N''(t_j)} \int_{-1}^{x} \frac{(z-1)T_N'(z)}{(z-t_j)}\, dz, \quad 1 \leq j \leq N-1, \tag{3.6}$$

and

$$\phi_N^{(N)}(x) = \frac{1}{T_N'(t_N)} \int_{-1}^{x} T_N'(z)dz = \frac{T_N(x) - T_N(t_0)}{T_N'(t_N)}. \tag{3.7}$$

With the Chebyshev zeros nodal distribution,

$$\psi_j^{(N-1)}(x) = \frac{(x-1)T_{N-1}(x)}{(s_j-1)(x-s_j)T_{N-1}'(s_j)}, \quad 1 \leq j \leq N-1,$$

and

$$\psi_N^{(N-1)}(x) = \frac{T_{N-1}(x)}{T_{N-1}(s_N)}\,.$$

Hence

$$\phi_j^{(N)}(x) = \frac{1}{(s_j-1)T'_{N-1}(s_j)} \int_{-1}^{x} \frac{(z-1)T_{N-1}(z)}{(z-s_j)}\,dz, \quad 1 \le j \le N-1, \tag{3.8}$$

and

$$\begin{aligned}\phi_N^{(N)}(x) &= \frac{1}{T_{N-1}(s_N)} \int_{-1}^{x} T_{N-1}(z)dz \\ &= \int_{-1}^{x} T_{N-1}(z)dz \\ &= \frac{1}{2}\left[\frac{T_N(x)}{N} - \frac{T_{N-2}(x)}{N-2}\right] + \frac{(-1)^N}{N(N-2)}\,. \end{aligned} \tag{3.9}$$

The first-order inverse matrices $(\tilde{D}_N^{(1)})^{-1}$ have a row sum property which is the analogue of the condition (2.21) that applies to $(\tilde{D}_N^{(2)})^{-1}$. If $\phi^{(N)}(x) = \sum_{j=1}^N \phi_j^{(N)}(x)$ and $\psi^{(N-1)}(x) = \sum_{j=1}^N \psi_j^{(N-1)}(x)$, then

$$\left.\begin{aligned} &\frac{d\phi^{(N)}}{dx}\,(x) = \psi^{(N-1)}(x), \\ &\text{with} \quad \phi^{(N)}(-1) = 0. \end{aligned}\right\} \tag{3.10}$$

Here $\phi^{(N)} \in P_N, \psi^{(N-1)} \in P_{N-1}$, and $\psi^{(N-1)}(x_i) = 1$ for $i = 1, 2, \ldots, N$. Clearly $\psi^{(N-1)}(x) \equiv 1$ and

$$\phi^{(N)}(x) = x + 1. \tag{3.11}$$

It follows that

$$\phi^{(N)}(x_i) = \sum_{j=1}^{N} (\tilde{D}_N^{(1)})^{-1}(i,j) = x_i + 1, \quad 1 \le i \le N,$$

and

$$\max_{1 \le i \le N} \phi^{(N)}(x_i) = \phi^N(x_N) = 2.$$

It may be shown (see the appendix) that, for each of the Chebyshev nodal distributions, the elements $\phi_j^{(N)}(x_N)$ have a constant sign for $1 \le j \le N$, from which it follows that $\sum_{j=1}^N |(\tilde{D}_N^{(1)})^{-1}(N,j)| = 2$. However, numerical experiments indicate that elements in $(\tilde{D}_N^{(1)})^{-1}$ do not have a uniform sign, and, as a result of cancellation, it may not be the case that

$$\|(\tilde{D}_N^{(1)})^{-1}\|_\infty \le 2. \tag{3.12}$$

TABLE 4
$\|(\tilde{D}_N^{(1)})^{-1}\|_\infty$ *for Chebyshev zeros nodes.*

N	$\|(\tilde{D}_N^{(1)})^{-1}\|_\infty$
4	2.3769
8	2.0924
16	2.0211
32	2.0050
64	2.0012
128	2.0003
256	2.0001
257	2.0001

$\|(\tilde{D}_N^{(1)})^{-1}\|_\infty$ was computed for each of the two Chebyshev nodal distributions over a range of values of N. In the case of Chebyshev extrema nodes, the norm was found to have the value 2 for a selection of values of N between 4 and 257. In the case of Chebyshev zeros nodes, the computed values of $\|(\tilde{D}_N^{(1)})^{-1}\|_\infty$ for several values of N are given in Table 4. The computed results suggest that (3.12) holds for nodal distribution (2.15(a)), and that $\|(\tilde{D}_N^{(1)})^{-1}\|_\infty$ tends to 2 from above as N increases for distribution (2.15(b)). Further comments on the structure of $\|(\tilde{D}_N^{(1)})^{-1}\|_\infty$ are given in the appendix.

4. Comments. Expressions have been given for the elements of the inverses of first- and second-order pseudospectral differentiation matrices. Simple upper bounds have been presented for the maximum norms of the second-order differentiation matrices when Chebyshev collocation points are used. The analytic expressions for the matrix elements may be useful in studying properties of the inverse matrices, and the bounds can be utilized in obtaining error bounds for the computed solutions of differential equations by the Chebyshev pseudospectral method. The bounds on $\|(\tilde{D}_N^{(2)})^{-1}\|_\infty$ for the Chebyshev distributions of nodes indicate that, in the maximum norm, the growth rate of the condition number of $\tilde{D}_N^{(2)}$ with N is similar to the growth of $\|D_N^{(2)}\|_\infty$: it is readily shown numerically that this is $O(N^4)$ as N increases. It would be of interest to investigate analogous properties of differentiation matrices associated with other sets of nodal distributions that have quadratic clustering—minimum node spacing decreasing like $O(1/N^2)$.

It may also be of interest to note that, for both Chebyshev distributions, $\|(\tilde{D}_N^{(2)})^{-1}\|_1$ and $\|(\tilde{D}_N^{(2)})^{-1}\|_2$ tend to 0.5708 and 0.4076, respectively, as N increases. In the case of the first-order differentiation matrix, $\|(\tilde{D}_N^{(1)})^{-1}\|_1$ and $\|(\tilde{D}_N^{(1)})^{-1}\|_2$ tend to 1.8 and 1.4, respectively, as N increases: the convergence rate is slower for the inverses of the first-order matrices. With evenly spaced nodes, the 1-norm and the 2-norm increase rapidly with N. These results are obtained by direct computation.

5. Appendix.

5.1. Negativity property of $(\tilde{D}_N^{(2)})^{-1}$.

Signs of $\frac{d\phi_j^{(N)}}{dx}$ at $x = \pm 1$. In the appendix we confine our analysis to the Chebyshev extrema distribution of nodes. Results may be obtained for the Chebyshev zeros case using an analogous approach.

LEMMA 5.1. $I_j^{(N)} := \int_{-1}^{1} \psi_j^{(N-2)}(x)dx > 0$ *for* $j = 1, 2, \ldots, N-1$, *with*

$$\psi_j^{(N-2)}(x) = \frac{T_N'(x)}{(x - t_j)T_N''(t_j)}$$

and $t_j = -\cos(\frac{\pi j}{N})$.

Note that $\{t_j\}_{j=1}^{N-1}$ are the zeros of $U_{N-1}(x)$, the Chebyshev polynomial of the second kind of degree $(N-1)$ [13]. Using the orthogonality properties of $\{U_k\}_{k=0}^{N-2}$, we may write

$$\psi_j^{(N-2)}(x) = \sum_{k=0}^{N-2} c_{j,k}\, U_k(x), \tag{5.1}$$

where $c_{j,k} = \frac{2}{\pi} \int_{-1}^{1} (1-x^2)^{\frac{1}{2}} \psi_j^{(N-2)}(x) U_k(x)dx$. For $0 \le k \le N-2$, $\psi_j^{(N-2)}\, U_k \in P_{2N-4}$, and we may therefore use the Gaussian quadrature rule (see [5])

(5.2)

$$\int_{-1}^{1} (1-x^2)^{\frac{1}{2}} f(x)dx = \sum_{i=1}^{N-1} w_i\, f(t_i) + \frac{\pi}{2^{2N-1}(2N-2)!}\, f^{(2N-2)}(\xi), \quad -1 < \xi < 1,$$

with zero error term ($f \equiv \psi_j^{(N-2)}\, U_k$), where $w_i = \frac{\pi}{N} \sin^2(\frac{\pi i}{N})$. This enables us to write

$$\begin{aligned} c_{j,k} &= \frac{2}{\pi} \sum_{l=1}^{N-1} w_l\, \psi_j^{(N-2)}(t_l) U_k(t_l) \\ &= \frac{2}{\pi}\, w_j\, U_k(t_j) \\ &= \frac{2}{N}\, (-1)^k \sin(\eta_j^{(N)}) \sin[(k+1)\eta_j^{(N)}], \end{aligned}$$

where $\eta_j^{(N)} = \frac{\pi j}{N}$. Hence

$$\begin{aligned} I_j^{(N)} &= \sum_{k=0}^{N-2} c_{j,k} \int_{-1}^{1} U_k(x)dx \\ &= \sum_{k=0}^{N-2} c_{j,k}\, \nu_k, \quad \text{where } \nu_k = \begin{cases} 0, & k \text{ odd}, \\ \frac{2}{k+1}, & k \text{ even}, \end{cases} \\ &= \sum_{m=0}^{q} c_{j,2m}\, \nu_{2m}, \quad \text{where } 2q = \begin{cases} N-2, & N \text{ even}, \\ N-3, & N \text{ odd}, \end{cases} \\ &= \frac{4}{N} \sin(\eta_j^{(N)}) \sum_{m=0}^{q} \frac{1}{(2m+1)} \sin[(2m+1)\eta_j^{(N)}]. \end{aligned}$$

This may be written as

$$I_j^{(N)} = \frac{2}{N} \left\{ 1 - \sum_{m=1}^{q} b_m \cos(2m\eta_j^{(N)}) - \frac{1}{2q+1} \cos[2(q+1)\eta_j^{(N)}] \right\},$$

where $b_m = \frac{2}{(2m-1)(2m+1)}$. Clearly,

$$I_j^{(N)} > \frac{2}{N}\left\{1 - \left(1 - \frac{1}{3}\right) - \left(\frac{1}{3} - \frac{1}{5}\right) - \cdots - \left(\frac{1}{2q-1} - \frac{1}{2q+1}\right) - \frac{1}{2q+1}\right\}$$
$$> 0.$$

This establishes the inequality given as Lemma 5.1.

The inequality in Lemma 5.1 may be strengthened to $I_j^{(N)} > 4/N^2$, and this stronger result is required later. It is clear that $I_j^{(N)} = I_{N-j}^{(N)}$ for $j = 1, 2, \ldots, N-1$, so we need only consider $j = 1, 2, \ldots, jmax$, where $jmax = N/2$ or $jmax = (N-1)/2$, depending on whether N is even or odd. Numerical experiments indicate that the quantity

$$F_j^{(N)} = \frac{N}{4}(I_j^{(N)} - I_1^{(N)}) > 0$$

for $j = 2, 3, \ldots, jmax$: the minimum value of $I_j^{(N)}$ occurs at $j = 1$, and it is therefore sufficient to show that $I_1^{(N)} > 4/N^2$.

The positivity of $F_j^{(N)}$ is more readily appreciated if we write $F_j^{(N)}$ in the form

$$F_j^{(N)} = \sum_{m=1}^{q} b_m \sin(m\eta_{j+1}^{(N)}) \sin(m\eta_{j-1}^{(N)}) + \frac{1}{2q+1} \sin[(q+1)\eta_{j+1}^{(N)}] \sin[(q+1)\eta_{j-1}^{(N)}].$$

For $j = 2, 3, \ldots, jmax$, the initial terms in this summation are positive, and the sum of the initial set of positive terms exceeds the sum of the moduli of the remaining terms.

LEMMA 5.2.

$$I_j^{(N)} > 4/N^2 \quad \textit{for} \quad j = 1, 2, \ldots, jmax.$$

Given that $I_j^{(N)} \geq I_1^{(N)}$ for $j = 1, 2, \ldots, jmax$, it is sufficient to establish the inequality for $j = 1$.

If N is even ($2q = N - 2$),

$$\frac{N}{2} I_1^{(N)} = 1 + \frac{1}{N-1} - \sum_{m=1}^{q} b_m \cos(2m\eta_1^{(N)})$$
$$> \frac{N}{N-1} - \left\{\left(1 - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{5}\right) + \cdots + \left(\frac{1}{2q-1} - \frac{1}{2q+1}\right)\right\}$$
$$= \frac{N}{N-1} - \left(1 - \frac{1}{N-1}\right) = \frac{2}{N-1}$$
$$> \frac{2}{N}.$$

Hence $I_1^{(N)} > 4/N^2$ if N is even.

If N is odd ($2q = N-3$),

$$\begin{aligned}\frac{N}{2}\, I_1^{(N)} &= 1 + \frac{1}{N-2}\,\cos\left(\frac{\pi}{N}\right) - \sum_{m=1}^{q} b_m\,\cos(2m\eta_1^{(N)}) \\ &> 1 + \frac{1}{N-2}\,\cos\left(\frac{\pi}{N}\right) - \left(1 - \frac{1}{N-2}\right) \\ &= \frac{1}{N-2}\,\left(1 + \cos\left(\frac{\pi}{N}\right)\right) \\ &> \frac{1}{N-2}\,\left(2 - \frac{\pi^2}{2N^2}\right) \quad \text{if} \quad N \geq 3, \\ &> \frac{1}{N-2}\,\left(2 - \frac{4}{N}\right) \\ &= \frac{2}{N}.\end{aligned}$$

Hence $I_1^{(N)} > 4/N^2$ if N is odd. This establishes the inequality given as Lemma 5.2.

LEMMA 5.3.

$$J_j^{(N)} := \int_{-1}^{1} (x-1)\psi_j^{(N-2)}(x)dx < 0 \quad \textit{for} \quad j = 1, 2, \ldots, N-1,$$

where $\psi_j^{(N-2)}$ *is as defined in Lemma* 5.1.

In this case we write

$$(x-1)\psi_j^{(N-2)}(x) = \sum_{k=0}^{N-1} d_{j,k}\, U_k(x), \tag{5.3}$$

where $d_{j,k} = \frac{2}{\pi}\int_{-1}^{1}(1-x^2)^{\frac{1}{2}}(x-1)\psi_j^{(N-2)}(x)\, U_k(x)dx$. For $k = 0, 1, \ldots, N-2$ we evaluate this integral as in Lemma 5.1 above, since $(x-1)\psi_j^{(N-2)}(x)U_k(x)$ is a polynomial of degree at most $(2N-3)$. It is readily shown that

$$d_{j,k} = (t_j - 1)c_{j,k} \quad \text{for} \quad k = 0, 1, \ldots, N-2.$$

The coefficient $d_{j,N-1}$ may be obtained by equating coefficients of x^{N-1} in (5.2). This identification gives $d_{j,N-1} = N/T_N''(t_j)$. From (5.1) and (5.2) we now obtain

$$\begin{aligned}J_j^{(N)} &= (t_j - 1)I_j^{(N)} + d_{j,N-1}\int_{-1}^{1} U_{N-1}(x)dx \\ &= (t_j - 1)I_j^{(N)} + \frac{1}{N}\, d_{j,N-1}(1 - (-1)^N) \\ &= (t_j - 1)I_j^{(N)} + \frac{(-1)^j}{N^2}\,\sin^2(\eta_j^{(N)})(1 - (-1)^N).\end{aligned}$$

It follows that

$$J_j^{(N)} = \begin{cases} (t_j - 1)I_j^{(N)}, & N \text{ even}, \\ (t_j - 1)\left[I_j^{(N)} - \dfrac{2(-1)^j}{N^2}\,(t_j + 1)\right], & N \text{ odd}. \end{cases} \tag{5.4}$$

Since $(t_j - 1) < 0$ and $I_j^{(N)} > 0$, then $J_j^{(N)} < 0$ if N is even. Furthermore, since $I_j^{(N)} > 4/N^2$, then $J_j^{(N)} < 0$ if N is odd. Lemma 5.3 is now established for any positive integer N and $j = 1, 2, \ldots, N-1$.

LEMMA 5.4.

$$K_j^{(N)} := \int_{-1}^{1} (x+1)\psi_j^{(N-2)}(x)dx > 0 \quad \textit{for} \quad j = 1, 2, \ldots, N-1,$$

where $\psi_j^{(N-2)}$ *is as defined in Lemma* 5.1.

Note that

$$K_j^{(N)} = J_j^{(N)} + 2I_j^{(N)} = \begin{cases} (t_j+1)I_j^{(N)}, & N \text{ even}, \\ (t_j+1)\left[I_j^{N)} - \dfrac{2(-1)^j}{N^2}(t_j - 1)\right], & N \text{ odd}. \end{cases} \tag{5.5}$$

Since $(t_j+1) > 0$ and $I_j^{(N)} > 4/N^2$, the proof of Lemma 5.4 follows immediately from (5.4).

From equation (2.13), we see that $\frac{d\phi_j^{(N)}}{dx}(-1) = \frac{1}{2} J_j^{(N)}$ and $\frac{d\phi_j^{(N)}}{dx}(+1) = \frac{1}{2} K_j^{(N)}$. Lemmas 5.3 and 5.4 show that for the Chebyshev extrema distribution of nodes, the gradients of $\phi_j^{(N)}$ at the boundaries satisfy $\frac{d\phi_j^{(N)}}{dx}(-1) < 0$ and $\frac{d\phi_j^{(N)}}{dx}(+1) > 0$. Similar results can be derived for the Chebyshev zeros distribution of nodes: in this latter case, the expansions (5.1) and (5.2) are effected in terms of Chebyshev polynomials of the first kind.

Sign of $\boldsymbol{\phi_j^{(N)}(x)}$ in $\boldsymbol{(-1, 1)}$. For the Chebyshev extrema distribution of nodes, (2.9) in $\phi_j^{(N)}$ takes the form

$$T_N''(t_j)(x - t_j)\frac{d^2\phi_j^{(N)}}{dx^2}(x) = T_N'(x). \tag{5.6}$$

A first integration of (5.5), making use of boundary conditions at $x = -1$, is

$$T_N''(t_j)\left[(x - t_j)\frac{d\phi_j^{(N)}}{dx}(x) - \phi_j^{(N)}(x)\right] = T_N(x) + A,$$

where $A = -T_N''(t_j)(1 + t_j)\frac{d\phi_j^{(N)}}{dx}(-1) - T_N(-1)$. The solution may now be written as

$$\phi_j^{(N)}(x) = -(x - t_j)\int_x^1 \frac{f(z)}{(z - t_j)^2}\, dz, \tag{5.7}$$

where $f(x) = (T_N(x) + A)/T_N''(t_j)$. Alternatively, if the first integral incorporates the boundary conditions at $x = +1$, the solution is written as

$$\phi_j^{(N)}(x) = (x - t_j)\int_{-1}^x \frac{g(z)}{(z - t_j)^2}\, dz, \tag{5.8}$$

where $g(x) = (T_N(x) + B)/T_N''(t_j)$ and $B = T_N''(t_j)(1 - t_j)\frac{d\phi^{(N)}}{dx}(+1) - T_N(+1)$. To obtain approximations to $\phi_j^{(N)}(x)$ by means of (5.6) and (5.7) we first show that

the dominant terms in $f(x)$ and $g(x)$ are, respectively, $-(1+t_j)\,\frac{d\phi_j^{(N)}}{dx}\,(-1)$ and $(1-t_j)\,\frac{d\phi_j^{(N)}}{dx}\,(+1)$. To establish this, note that

$$\begin{aligned}\left|\frac{T_N(x)-T_N(\pm1)}{T_N''(t_j)}\right| &\le \frac{2}{|T_N''(t_j)|}\\ &= \frac{2}{N^2}\,(1-t_j^2)\\ &< (1-t_j^2)\,\frac{I_j^{(N)}}{2}\\ &\le \frac{1}{2}\max\left(-(1+t_j)J_j^{(N)},(1-t_j)K_j^{(N)}\right)\\ &= \max\left(-(1+t_j)\frac{d\phi_j^{(N)}}{dx}\,(-1),(1-t_j)\frac{d\phi_j^{(N)}}{dx}\,(+1)\right).\end{aligned}$$

If $f(x)$ and $g(x)$ are replaced by their respective dominant terms, then approximations to $\phi_j^{(N)}(x)$ may now be obtained from (5.6) and (5.7) as

$$\phi_j^{(N)}(x) \sim (1+t_j)\,\frac{d\phi_j^{(N)}}{dx}\,(-1)\frac{(1-x)}{(1-t_j)} \tag{5.9}$$

and

$$\phi_j^{(N)}(x) \sim -(1-t_j)\,\frac{d\phi_j^{(N)}}{dx}\,(+1)\frac{(x+1)}{(1+t_j)}. \tag{5.10}$$

If N is even, then

$$-(1+t_j)\frac{d\phi_j^{(N)}}{dx}\,(-1) = (1-t_j)\frac{d\phi_j^{(N)}}{dx}\,(+1) = \frac{1}{2}\,(1-t_j^2)I_j^{(N)},$$

and $\phi_j^{(N)}(x)$ may be represented by either (5.8) or (5.9). If N is odd, then

$$(1-t_j)\frac{d\phi_j^{(N)}}{dx}\,(+1) + (1+t_j)\frac{d\phi_j^{(N)}}{dx}\,(-1) = \frac{2(-1)^j}{N^2}\,(1-t_j^2),$$

and $\phi_j^{(N)}(x)$ may be approximated by (5.9) and (5.10), depending on whether j is odd or even. In either case, the dominant terms in the solution indicate that $\phi_j^{(N)}(x) < 0$ in $(-1,1)$.

5.2. Comments on the structure of $(\tilde{D}_N^{(1)})^{-1}$. The Nth row of $(\tilde{D}_N^{(1)})^{-1}$ is $[\phi_1^{(N)}(t_N),\phi_2^{(N)}(t_N),\ldots,\phi_N^{(N)}(t_N)]$. For the Chebyshev extrema distribution of nodes, we see from (3.7) that $\phi_N^{(N)}(t_N)$ takes values 0 or $2/N^2$ for N even or odd. For $j=1,2,\ldots,N-1$, we see from (3.6) that

$$\begin{aligned}\phi_j^{(N)}(t_N) &= \frac{1}{(t_j-1)T_N''(t_j)}\int_{-1}^{1}\frac{(z-1)T_N'(z)}{(z-t_j)}\,dz\\ &= \frac{1}{(t_j-1)}\int_{-1}^{1}(z-1)\psi_j^{(N-2)}(z)dz,\end{aligned}$$

where $\psi_j^{(N-2)}$ is the interpolation polynomial for extrema nodes used in section 5.1. It follows that $\phi_j^{(N)}(t_N) = J_j^{(N)}/(t_j - 1)$, which is strictly positive since $J_j^{(N)} < 0$ and $(t_j - 1) < 0$. Thus $\phi_j^{(N)}(t_N) \geq 0$ for any positive integer N and $j = 1, 2, \ldots, N$. It has been shown in section 3.2 that the sum of these nonnegative elements in the Nth row of $(\tilde{D}_N^{(1)})^{-1}$ is 2. Direct computation shows that

$$\sum_{j=1}^{N} \left| (\tilde{D}_N^{(1)})^{-1}(i,j) \right| < 2 \quad \text{for} \quad i = 1, 2, \ldots, N-1.$$

The maximum row sum is

$$\sum_{j=1}^{N} \left| (\tilde{D}_N^{(1)})^{-1}(N,j) \right| = \sum_{j=1}^{N} (\tilde{D}_N^{(1)})^{-1}(N,j) = 2.$$

For the Chebyshev zeros distribution of nodes, an argument similar to that used above (assuming that the appropriate $J_j^{(N)}$ is negative) shows that $\phi_j^{(N)}(s_N) \geq 0$ for any positive integer N and $j = 1, 2, \ldots, N$. The sum of these nonnegative elements in the Nth row of $(\tilde{D}_N^{(1)})^{-1}$ is 2, as shown in section 3.2. In this case, direct computation shows that the maximum value of $\sum_{j=1}^{N} |(\tilde{D}_N^{(1)})^{-1}(i,j)|$ occurs at $i = N-2$, and it is this value that gives rise to the entries in Table 4. The value of $\sum_{j=1}^{N} |(\tilde{D}_N^{(1)})^{-1}(i,j)|$ is strictly less than 2 if i differs from $N-2$ or N. Furthermore, $\sum_{j=1}^{N} |(\tilde{D}_N^{(1)})^{-1}(i,j)|$ is strictly monotonic increasing with i for $i = 1, 2, \ldots, N-2$.

The structure of the inverses of the first-order Chebyshev pseudospectral differentiation matrices is worthy of further investigation.

Acknowledgments. I wish to thank Andre Weideman for suggesting this problem. The work was initiated while I was visiting the University of Stellenbosch, South Africa, in November and December, 2001. I am also grateful to a reviewer for valuable comments that led to some improvements in the presentation of this work.

REFERENCES

[1] A. L. Andrew, *Centrosymmetric matrices*, SIAM Rev., 40 (1998), pp. 697–698.
[2] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, New York, 2000.
[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988.
[4] P. J. Davis, *Interpolation and Approximation*, Dover, New York, 1975.
[5] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed., Academic Press, New York, 1984.
[6] W. S. Don and A. Solomonoff, *Pseudo-Pack 2.3β*, Center for Fluid Mechanics, Brown University, 1998; available online at http://www.cfm.brown.edu/people/wsdon/home.html.
[7] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge University Press, Cambridge, UK, 1996.
[8] B. Fornberg and D. M. Sloan, *A review of pseudospectral methods for solving partial differential equations*, Acta Numer., 3 (1994), pp. 203–267.
[9] D. Funaro, *Polynomial Approximation of Differential Equations*, Springer-Verlag, Berlin, 1992.
[10] D. Funaro, *FORTRAN Routines for Spectral Methods*, Pubblicazioni 891, Instituto di Analisi Numerica del Consiglio Nazionale delle Ricerche, Pavia, Italy, 1993.
[11] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 26, SIAM, Philadelphia, 1977.
[12] W. Huang and D. M. Sloan, *The pseudospectral method for third-order differential equations*, SIAM J. Numer. Anal., 29 (1992), pp. 1626–1647.

[13] T. J. Rivlin, *The Chebyshev Polynomials*, John Wiley & Sons, New York, 1974.
[14] L. N. Trefethen, *Spectral Methods in MATLAB*, Software Environ. Tools 10, SIAM, Philadelphia, 2000.
[15] J. A. C. Weideman and S. C. Reddy, *A MATLAB differentiation matrix suite*, ACM Trans. Math. Software, 26 (2000), pp. 465–519; available online at http://ucs.orst.edu/∼weidemaj/differ.html.
[16] B. D. Welfert, *Generation of pseudospectral differentiation matrices* I, SIAM J. Numer. Anal., 34 (1997), pp. 1640–1657.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 49–74

FINITE ELEMENT APPROXIMATIONS TO THE DISCRETE SPECTRUM OF THE SCHRÖDINGER OPERATOR WITH THE COULOMB POTENTIAL*

WEIYING ZHENG† AND LUNG-AN YING†

Abstract. In the present paper, the authors consider the Schrödinger operator H with the Coulomb potential defined in R^{3m}, where m is a positive integer. Both bounded domain approximations to multielectron systems and finite element approximations to the helium system are analyzed. The spectrum of H becomes completely discrete when confined to bounded domains. The error estimate of the bounded domain approximation to the discrete spectrum of H is obtained. Since numerical solution is difficult for a higher-dimensional problem of dimension more than three, the finite element analyses in this paper are restricted to the S-state of the helium atom. The authors transform the six-dimensional Schrödinger equation of the helium S-state into a three-dimensional form. Optimal error estimates for the finite element approximation to the three-dimensional equation, for all eigenvalues and eigenfunctions of the three-dimensional equation, are obtained by means of local regularization. Numerical results are shown in the last section.

Key words. spectrum approximation, Schrödinger equation, weighted norm, local regularization, finite element method

AMS subject classifications. 65N30, 65N25, 81Q05

DOI. 10.1137/S0036142902403474

1. Introduction.

The multielectron Coulomb problem in quantum mechanics cannot be solved in a finite form. Nevertheless it challenges and stimulates many mathematicians and physicists to devote themselves to developing efficient methods for solving the system.

Several successful approximation techniques in quantum physics/chemistry have been developed for this problem. They include the Hartree–Fock method [15], the finite difference method [19], [35], the correlation-function hyperspherical-harmonic method [18], [24], and various variational approximations. For the Hartree–Fock method, every electron is considered independently to be in a central electric field formed by the nucleus and other electrons. The finite difference method needs a rectangular domain in R^N and uniform grids. The double and triple basis set methods (which are variational methods indeed) are very powerful for the eigenvalue problem of the helium atom. Kono and Hattori (see [21], [22]) used two sets of basis functions $r_1^i r_2^j r_{12}^k e^{-\xi r_1 - \eta r_2} A$ ("ξ terms") and $r_1^i r_2^j r_{12}^k e^{-\zeta(r_1+r_2)} A$ ("ζ terms") to calculate the energy levels for the S, P, and D states of the helium atom. (A is an appropriate angular factor.) The former set of functions is expected to describe the whole wave function roughly, while the latter is expected to describe the short- and middle-range correlation effects. Their calculations yield 9–10 significant digits for S states. Kleindienst, Lüchow, and Merckens [20] and Drake and Yan [12] applied the double basis set method to S-states of helium. Their basis functions are $r_1^i r_2^j r_{12}^k e^{-\xi_1 r_1 - \eta_1 r_2}$ and $r_1^i r_2^j r_{12}^k e^{-\xi_2 r_1 - \eta_2 r_2}$. Drake and Yan employed truncations to ensure numerical stability and convergence. By complete optimization of the exponential scale factors ξ_1,

*Received by the editors March 5, 2002; accepted for publication (in revised form) June 13, 2003; published electronically January 6, 2004. This research was supported by the China State Major Key project for Basic Researches and the Science Fund of the Ministry of Education of China.
http://www.siam.org/journals/sinum/42-1/40347.html

†School of Mathematical Sciences, Peking University, Beijing, 100871 China (zhengweiying @yahoo.com, yingla@pku.edu.cn).

η_1, ξ_2, and η_2, they achieved more than 15 significant digits. Recently, Drake, Cassar, and Nistor [13] obtained 21 significant digits for the ground state of helium by the triple basis set method. Korobov [23] even obtained 25 significant digits for the ground state of helium. That work can be used as a benchmark for other approaches for three-body systems. All three of these excellent works in variational methods promote the development of few-body problems in quantum mechanics.

The finite element method (FEM) is used initially in elastic mechanics and fluid mechanics. It uses local interpolation functions to approximate the unknown function (see [8] and [36]), and thus can describe the local properties of wave functions. Therefore, we can expect to obtain good approximations to the energy. Important works on FEM applied to atomic and molecular problems first appeared in 1975 (see [3]). They were devoted to one- or two-dimensional problems [3], [4], [14], [16], [30]. In 1985, Levin and Schertzer [25] published the first work applying FEM to three-dimensional problems. They calculated the ground state of the helium atom. Most of the works applying FEM to three-body problems have appeared since 1990 (i.e., [1], [6], [31], [37]). All of the cited works obtained very good results.

To apply FEM to quantum mechanics, we should consider three aspects of the problem. The first is the spectrum approximation of the whole space Schrödinger operator by the operator defined on some bounded domain. The second is the error estimate for FEM approximation. The third is the real computation of approximate solutions. To the best of our knowledge, we have not found any work analyzing the first two aspects.

In this paper, we consider the first aspect for the system of an arbitrary atom. But as for the finite element aspect, since any problem of dimension more than three is a great challenge for both modern numerical methods and computers, we restrict the finite element analysis and computation to the S-state of the helium atom. In fact, we can transform the $3m$-dimensional Schrödinger equation (see [38] for $m = 2, 3$) into a $3(m-1)$-dimensional form rigorously, and theoretical analysis of the FEM applied to the simplified equation can be obtained similarly, in view of the argument of sections 3, 4, and 5 in the present paper. However, real computations are very difficult to carry out because of numerous degrees of freedom. Our numerical results on the lithium atom (the Schrödinger equation is nine-dimensional) will appear in another paper [39].

The present paper consists of three parts. First, we consider the bounded domain approximation of the $3m$-dimensional Schrödinger equation (m is the number of electrons in an ion). The spectrum of the Schrödinger operator H consists of the discrete spectrum included in $(-a, 0)$(for some $a > 0$) and the continuous spectrum $[0, +\infty)$. We show that the spectrum of H becomes completely discrete if it is restricted to bounded domains. In section 2, we show that for any eigenvalue of the whole space problem and for any $\epsilon > 0$, assuming the bounded domain large enough, there is an eigenpair of the bounded domain problem such that the errors of both the eigenvalue and the eigenfunction are smaller than ϵ. Secondly, we analyze the finite element approximation of the S-state of the helium atom. In section 3, the six-dimensional Schrödinger equation is transformed into a three-dimensional form, and some Hilbert spaces with weighted inner products and norms are defined. Because we cannot say that the solutions of both the three-dimensional equation and the six-dimensional one are continuous, the technique of local regularization [25] is used to prove the convergency of the finite element scheme. In section 4, we describe the three-dimensional local regularization operator in detail. In section 5, an equivalent variational equation of the three-dimensional equation and its FEM approximation are given for the

helium atom of the S-state. The optimal order error estimate of the finite element scheme is obtained. Thirdly, we have calculated approximate solutions by the finite element scheme. In section 6, we give the numerical results from two kinds of FEM approximations to the three-dimensional energy equation. The results are better than existing finite element results. Furthermore, from the figures we can see that our approximate wave functions coincide very well with many physical properties well known to physicists, and with many essential physical assumptions in quantum mechanics which are not added into our computations a priori.

The difficulties appear in three aspects: 1. the proof of the continuity and coercivity of the bilinear forms in the variational equations with the presence of the singularities in the Coulomb potential, 2. the proof of the convergency of the finite element scheme while the variational spaces are not standard Sobolev spaces, and 3. obtaining precise results in presence of numerous unknowns and singular integrals.

Through the paper, C represents the generic constant independent of minded parameters; the symbol "$\Longleftrightarrow$" means "be equivalent to." We use atomic units except where explicitly explained, i.e., Bohr radius a_0 for length, Rydberg (Hartree only in section 6) for energy. We consider the nonrelativistic and spin-independent case.

2. Discrete spectrum approximations of the Schrödinger operator in bounded domains. Let $m > 0$ be an integer, $N = 3m$. The Schrödinger equation of an m-electron ion is

$$H\psi = E\psi \quad \text{in } R^N, \tag{2.1}$$

where

$$\begin{aligned} H\psi &= -\Delta\psi + V\psi, \\ \Delta\psi &= \sum_{i=1}^{m} \left(\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \right), \\ V &= -\sum_{i=1}^{m} \frac{2Z}{r_i} + \sum_{1 \le i < j \le m} \frac{2}{r_{ij}}; \end{aligned}$$

(x_i, y_i, z_i) are the coordinates of the ith electron, $r_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$ is the distance between the ith electron and the nucleus, $1 \le i \le m$; and $r_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{1/2}$ is the distance between the ith electron and the jth electron, $1 \le i < j \le m$. Z is the charge number of the nucleus.

It is well known that H is self-adjoint and bounded below (Theorem 10.33 and the analysis on p. 323 of [34]). Its spectrum $\sigma(H)$ is included in R^1. Furthermore, the continuous spectrum of H is $[0, +\infty)$, and $\forall s \in \sigma(H) \bigcap (-\infty, 0)$, s is an eigenvalue of H (see Theorems 10.30 and 10.31 in [34]).

LEMMA 2.1. *Let $1 \le p < 2$, $3/2 < q < 2$, $1 \le q_1 < \min(3, \frac{Nq}{N-2q})$. Then there exists a constant C such that $\forall v \in H^1(R^N)$, $u \in W^{2,q}(R^N)$,*

$$\frac{u}{r_i}, \quad \frac{u}{r_{ij}} \in L^{q_1}(R^N), \tag{2.2}$$

and

$$\begin{cases} \displaystyle\int_{R^N} \frac{v^2}{r_i^p} dx, \quad \int_{R^N} \frac{v^2}{r_{ij}^p} dx \le C \|v\|_{1,R^N}^2, \\ \displaystyle\left\| \frac{u}{r_i} \right\|_{0,q_1,R^N}, \quad \left\| \frac{u}{r_{ij}} \right\|_{0,q_1,R^N} \le C \|u\|_{2,q,R^N}; \end{cases} \tag{2.3}$$

moreover, if v is compactly supported in R^N, then

$$\frac{v}{r_i}, \quad \frac{v}{r_{ij}} \in L^p(R^N), \tag{2.4}$$

and

$$\left\| \frac{v}{r_i} \right\|_{0,p,R^N}, \ \left\| \frac{v}{r_{ij}} \right\|_{0,p,R^N} \le C\|v\|_{1,R^N}, \tag{2.5}$$

where $1 \le i, j \le m$ and $i < j$.

Proof. (1) $\forall x \in R^N$, set $x = (x^{(1)}, \dots, x^{(m)})$, $x^{(i)} = (x_i, y_i, z_i)$, $\tilde{x}^{(i)} = (x^{(1)}, \dots, x^{(i-1)}, x^{(i+1)}, \dots, x^{(m)})$. Define $B_i = \{(x_i, y_i, z_i) \mid x_i^2 + y_i^2 + z_i^2 \le 1\}$, $i = 1, \dots, m$. Since $H^1(B_i) \hookrightarrow L^6(B_i)$, $W^{2,q}(B_i) \hookrightarrow C(B_i)$, $W^{2,q}(R^3) \hookrightarrow W^{1,q}(R^3) \hookrightarrow L^2(R^3)$, let $0 \le t < 2, 0 \le s < 3$; then $\forall v \in H^1(R^N)$, $u \in W^{2,q}(R^N)$,

$$\int_{R^3 \setminus B_i} \frac{v^2}{r_i^t} dx^{(i)} \le \|v\|_{0,R^3}^2 \le \|v\|_{1,R^3}^2, \quad \int_{R^3 \setminus B_i} \frac{u^2}{r_i^s} dx^{(i)} \le \|u\|_{0,R^3}^2 \le C\|u\|_{2,q,R^3}^2,$$

$$\int_{B_i} \frac{v^2}{r_i^t} dx^{(i)} \le \left(\int_{B_i} v^6 dx^{(i)} \right)^{\frac{1}{3}} \left(\int_{B_i} r_i^{-\frac{3t}{2}} dx^{(i)} \right)^{\frac{2}{3}} \le C\|v\|_{0,6,B_i}^2 \le C\|v\|_{1,B_i}^2,$$

$$\int_{B_i} \frac{u^2}{r_i^s} dx^{(i)} \le \|u\|_{0,\infty,B_i}^2 \int_{B_i} r_i^{-s} dx^{(i)} \le C\|u\|_{2,q,B_i}^2 \quad \text{for almost every } \tilde{x}^{(i)} \in R^{N-3}.$$

Hence

$$\int_{R^3} \frac{v^2}{r_i^t} dx^{(i)} \le C\|v\|_{1,R^3}^2, \quad \int_{R^3} \frac{u^2}{r_i^s} dx^{(i)} \le C\|u\|_{2,q,R^3}^2. \tag{2.6}$$

By Tonelli's theorem [34], integrating (2.6) with respect to the rest variables gives

$$\int_{R^N} \frac{v^2}{r_i^t} dx \le C\|v\|_{1,R^N}^2, \quad \int_{R^N} \frac{u^2}{r_i^s} dx \le C\|u\|_{2,q,R^N}^2. \tag{2.7}$$

Setting $s = \frac{N(q-2)+4q}{N(q-q_1)+2qq_1} q_1$, by (2.7) we have

$$\int_{R^N} \left(\frac{u}{r_i} \right)^{q_1} dx \le \left(\int_{R^N} \frac{u^2}{r_i^s} dx \right)^{\frac{q_1}{s}} \left(\int_{R^N} u^{\frac{q_1(s-2)}{s-q_1}} dx \right)^{\frac{s-q_1}{s}} \le C\|u\|_{2,q,R^N}^{q_1}.$$

If the support of $v \in H^1(R^N)$ is compact, set $R > 0$ large enough and define

$$B_i(0,R) = \left\{ (x_i, y_i, z_i) \in R^3 \ \middle| \ \sqrt{x_i^2 + y_i^2 + z_i^2} \le R \right\},$$

$$\Omega_R = \left\{ (x^{(1)}, \dots, x^{(m)}) \in R^N \ \middle| \ \sqrt{x_i^2 + y_i^2 + z_i^2} \le R, \quad i = 1, 2, \dots, m \right\}.$$

Assuming $\operatorname{supp} v \subset \Omega_R$, by Hölder's inequality there exists a positive constant C, depending on R and p, such that

$$\int_{R^3} \left(\frac{v}{r_i} \right)^p dx^{(i)} = \int_{B_i(0,R)} \left(\frac{v}{r_i} \right)^p dx^{(i)} \le C\|v\|_{0,6,B_i(0,R)}^p \le C\|v\|_{1,B_i(0,R)}^p. \tag{2.8}$$

Integrating (2.8) with respect to the rest of the variables produces the following:

$$\int_{R^N}\left(\frac{v}{r_i}\right)^p d\tilde{x}^{(i)} \le C\int_{R^{N-3}} \|v\|^p_{1,B_i(0,R)} d\tilde{x}^{(i)} \le C\|v\|^p_{1,R^N}.$$

(2) For any $1 \le i < j \le m$, let $\xi^{(i)} = x^{(i)} - x^{(j)}$, $\xi^{(i)} = (\xi_i, \eta_i, \zeta_i)$, $x^{(i)} = (x_i, y_i, z_i)$. For any $v = v(x^{(1)}, \dots, x^{(m)}) \in H^1(R^N)$, define

$$v_{ij}(x^{(1)}, \dots, \xi^{(i)}, \dots, x^{(m)}) = v(x^{(1)}, \dots, \xi^{(i)} + x^{(j)} \dots, x^{(m)}),$$

By Tonelli's theorem [34], we have

$$\begin{aligned}\int_{R^{N-3}}\int_{R^3} & v^2(x^{(1)}, \dots, x^{(i)}, \dots, x^{(m)}) dx^{(i)} d\tilde{x}^{(i)} \\ &= \int_{R^{N-3}}\int_{R^3} v_{ij}^2(x^{(1)}, \dots, \xi^{(i)}, \dots, x^{(m)}) d\xi^{(i)} d\tilde{x}^{(i)} \\ &= \int_{R^N} v_{ij}^2 dx \qquad \forall v \in H^1(R^N).\end{aligned}$$

Thus $v_{ij} \in L^2(R^N)$. In the same way, we have

$$\frac{\partial v_{ij}}{\partial x_k}, \quad \frac{\partial v_{ij}}{\partial y_k}, \quad \frac{\partial v_{ij}}{\partial z_k} \in L^2(R^N), \quad 1 \le k \le N, \quad k \ne i, j.$$

Since $\frac{\partial v_{ij}}{\partial \xi_i} = \frac{\partial v}{\partial x_i}$, $\frac{\partial v_{ij}}{\partial x_j} = \frac{\partial v}{\partial x_i} + \frac{\partial v}{\partial x_j}$, we have

$$\int_{R^{N-3}}\int_{R^3}\left|\frac{\partial v}{\partial \xi_i}\right|^2 d\xi^{(i)} d\tilde{x}^{(i)} = \int_{R^{N-3}}\int_{R^3}\left|\frac{\partial v}{\partial x_i}\right|^2 d\xi^{(i)} d\tilde{x}^{(i)} = \int_{R^{N-3}}\int_{R^3}\left|\frac{\partial v}{\partial x_i}\right|^2 dx^{(i)} d\tilde{x}^{(i)},$$

i.e., $\frac{\partial v_{ij}}{\partial \xi_i} \in L^2(R^N)$. Similarly, we have

$$\frac{\partial v_{ij}}{\partial \eta_i}, \quad \frac{\partial v_{ij}}{\partial \zeta_i}, \quad \frac{\partial v_{ij}}{\partial x_j}, \quad \frac{\partial v_{ij}}{\partial y_j}, \quad \frac{\partial v_{ij}}{\partial z_j} \in L^2(R^N).$$

Therefore $v_{ij} \in H^1(R^N)$. In the same way, if $u \in W^{2,q}(R^N)$, we have $u_{ij} \in W^{2,q}(R^N)$, $1 \le i, j, \le m$. By (1), $v/r_{ij} = v_{ij}/|\xi_{(i)}|$, $u/r_{ij} = u_{ij}/|\xi_{(i)}|$ satisfy (2.3)–(2.5). ∎

Let $\sigma > 0$ be a constant, $\hat{x} = (\hat{x}^{(1)}, \dots, \hat{x}^{(m)}) = \sigma x$, $\hat{x}^{(i)} = (\hat{x}_i, \hat{y}_i, \hat{z}_i)$, $\hat{\psi}(\hat{x}) = \psi(\sigma x)$ for any $x \in R^N$. Then (2.1) becomes

$$-\sigma^2\hat{\Delta}\hat{\psi} - \sigma\sum_{i=1}^m \frac{2Z\hat{\psi}}{\hat{r}_i} + \sigma\sum_{1\le i<j\le m}\frac{2\hat{\psi}}{\hat{r}_{ij}} = E\hat{\psi} \qquad \text{in } R^N, \tag{2.9}$$

where $\hat{\Delta}$ concerns the derivatives with respect to $\hat{x}$. Obviously, $\psi \in H^1(R^N) \Longleftrightarrow \hat{\psi} \in H^1(R^N)$. By Lemma 2.1, we can choose σ, K large enough such that

$$2\sigma\sum_{i=1}^m \int_{R^N}\frac{2Z\hat{\varphi}^2}{\hat{r}_i} d\hat{x} \le \sigma^2|\hat{\varphi}|^2_{1,R^N} + K\|\hat{\varphi}\|^2 \quad \forall \hat{\varphi} \in H^1(R^N). \tag{2.10}$$

Let $\lambda = E + K$; then the variational form of (2.9) is the following: Find $(\lambda, \hat{\psi}) \in R^1 \times H^1(R^N)$ and $\hat{\psi} \ne 0$ such that

$$\hat{a}(\hat{\psi}, \hat{\varphi}) = \lambda\int_{R^N}\hat{\psi}\hat{\varphi} d\hat{x} \quad \forall \hat{\varphi} \in H^1(R^N), \tag{2.11}$$

where

$$\hat{a}(\hat{\psi},\hat{\varphi}) = \sigma^2 \int_{R^N} \hat{\nabla}\hat{\psi}\cdot\hat{\nabla}\hat{\varphi}d\hat{x} + \int_{R^N}\left(-\sum_{i=1}^{m}\frac{2Z\sigma}{\hat{r}_i} + \sum_{1\le i<j\le m}\frac{2\sigma}{\hat{r}_{ij}} + K\right)\hat{\psi}\hat{\varphi}d\hat{x}.$$

It is easy to see that $\hat{a}(\cdot,\cdot)$ is continuous and coercive on $H^1(R^N)$ by Lemma 2.1 and (2.10).

The weak form of (2.1) is the following: Find $(\lambda,\psi)\in R^1\times H^1(R^N)$ and $\psi\neq 0$ such that

$$a(\psi,\varphi) = \lambda\int_{R^N}\psi\varphi dx \qquad \forall\varphi\in H^1(R^N), \tag{2.12}$$

where

$$a(\psi,\varphi) = \int_{R^N}\nabla\psi\cdot\nabla\varphi dx + \int_{R^N}\left(-\sum_{i=1}^{m}\frac{2Z}{r_i} + \sum_{1\le i<j\le m}\frac{2}{r_{ij}} + K\right)\psi\varphi dx.$$

By the transform $a(\psi,\varphi)=\sigma^{-N}\hat{a}(\hat{\psi},\hat{\varphi})$ and

$$\min\{\sigma^N,\sigma^{N-2}\}\|\psi\|^2_{1,R^N}\le\|\hat{\psi}\|^2_{1,R^N}\le\max\{\sigma^N,\sigma^{N-2}\}\|\psi\|^2_{1,R^N},$$

we know that $a(\cdot,\cdot)$ is continuous and coercive on $H^1(R^N)$. For the sake of simplicity in notation, we drop the continuity and coercivity constants and write $\|\cdot\|_{1,R^N}=\sqrt{a(\cdot,\cdot)}$ throughout this paper.

We consider the approximation of (2.12) in a bounded domain. Let $R>0$ be large enough and $B=B(0,R)\subset R^N$ be the ball with radius R and center at the origin. The approximation of (2.12) is defined as follows: Find $(\lambda_B,\psi_B)\in R^1\times H_0^1(B)$ and $\psi_B\neq 0$ such that

$$a(\psi_B,\phi_B) = \lambda\int_{R^N}\psi_B\phi_B dx \quad \forall\phi_B\in H_0^1(B). \tag{2.13}$$

For all $\phi_B\in H_0^1(B)$, we extend ϕ_B by zero to the exterior of B and still denote the extension as $\phi_B\in H^1(R^N)$. Thus (2.13) is the Galerkin approximation of (2.12). $a(\cdot,\cdot)$ is continuous and coercive on $H_0^1(B)$, and the continuity and coercivity constants are independent of the radius R.

THEOREM 2.2. *If the Schrödinger operator H is restricted to the bounded domain B, then its spectrum is discrete. It has the form*

$$0<\lambda_{B1}\le\lambda_{B2}\le\cdots\to+\infty, \tag{2.14}$$

where $\lambda_{Bi}=\lambda_{B,i+1}$ means that λ_{Bi} is multiple. If ψ_B is the eigenfunction in (2.13) *associated with λ_B, then $\psi_B\in H^2(B)\bigcap H_0^1(B)$.*

Proof. Since $H_0^1(B)\hookrightarrow\hookrightarrow L^2(B)$, by the Lax–Milgram theorem [9], we know that (2.13) has eigenvalues and eigenfunctions. Its spectrum is discrete and has the form of (2.14).

Let (λ_B,ψ_B) be an eigenpair of (2.13), i.e.,

$$\begin{cases} -\Delta\psi_B - \displaystyle\sum_{i=1}^{m}\frac{2Z\psi_B}{r_i} + \sum_{1\le i<j\le m}\frac{2\psi_B}{r_{ij}} = E\psi_B & \text{in } B,\\ \psi = 0 \quad \text{on } \partial B. \end{cases} \tag{2.15}$$

By Lemma 2.1, we have

$$\frac{\psi_B}{r_i}, \quad \frac{\psi_B}{r_{ij}} \in L^p(B), \quad 1 \le i < j \le m, \quad \forall\, 1 < p < 2.$$

By the L^p theory of elliptic equations [17], $\psi_B \in W^{2,p}(B)$ for any $1 < p < 2$. Then by Lemma 2.1,

$$\frac{\psi_B}{r_i}, \quad \frac{\psi_B}{r_{ij}} \in L^2(B), \quad 1 \le i < j \le m.$$

Thus $\psi_B \in H^2(B)$. □

THEOREM 2.3. *If (λ, ψ) is an eigenpair of (2.12) with $\|\psi\|_{1,R^N} = 1$, then for any $\epsilon > 0$ there exist $R > 0$ and an eigenpair (λ_B, ψ_B) of (2.13) such that*

$$|\lambda - \lambda_B| < C\epsilon^2, \tag{2.16}$$

$$\|\psi - \psi_B\|_{1,R^N} < C\epsilon, \tag{2.17}$$

where ψ_B is extended by zero to the exterior of B and C is a positive constant independent of R and ϵ.

Proof. By Theorem 10.33 in [34] and the coercivity of $a(\cdot,\cdot)$, we know that the discrete spectrum of (2.12) is

$$0 < \lambda_1 \le \lambda_2 \le \cdots < K,$$

where $\lambda_i = \lambda_{i+1}$ means that λ_i is multiple, and K is the unique accumulation. By the minmax theorem [10],

$$\lambda_i \le \lambda_{Bi}, \quad i = 1, 2, \ldots .$$

Define V_i, V_{Bi} as the eigenspaces associated with λ_i and λ_{Bi}, respectively. We assume $\psi \in V_i$. By the theory of abstract spectrum approximation (p. 699 of [10]), there exists $\psi_{Bi} \in V_{Bi}$ such that

$$|\lambda_i - \lambda_{Bi}| \le C(\varepsilon_B(\lambda_i))^2, \tag{2.18}$$

$$\|\psi - \psi_{Bi}\|_{1,R^N} \le C\varepsilon_B(\lambda_i), \tag{2.19}$$

where

$$\varepsilon_B(\lambda_i) = \sup_{\substack{u \in V_i, \\ \|u\|_{1,R^N}=1}} \inf_{v \in H_0^1(B)} \|u - v\|_{1,R^N}. \tag{2.20}$$

C is a constant depending on the continuity and coercivity constants of the bilinear form $a(\cdot,\cdot)$ and λ_i, but is independent of R.

Let $\{\psi_1, \ldots, \psi_l\}$ be an orthonormal basis of V_i with respect to the norm of $H^1(R^N)$, and let l be the multiplicity of λ_i. For any $\epsilon > 0$, since $\psi_k \in H^1(R^N)$, there exists a $\phi_k \in C_0^\infty(R^N)$ such that $\|\psi_k - \phi_k\|_{1,R^N} < \epsilon/l$. Set R large enough such that $\cup_{k=1}^l \operatorname{supp}\phi_k \subset B = B(0,R)$; then

$$\varepsilon_B(\lambda_i) \le \sum_{k=1}^{l} \|\psi_k - \phi_k\|_{1,R^N} < \epsilon. \tag{2.21}$$

In view of (2.18)–(2.21), we have (2.16) and (2.17). □

3. Weighted norms and Hilbert spaces. We consider the eigenvalue problem of the S-state of the helium atom, i.e., $Z = m = 2$. The eigenvalue equations of the Hamiltonian and the square of the angular momentum are

$$-\Delta_1\psi - \Delta_2\psi + \left(\frac{2}{r_{12}} - \frac{4}{r_1} - \frac{4}{r_2}\right)\psi = E\psi, \tag{3.1}$$

$$\begin{aligned}\Bigg\{&\left[\sum_{i=1}^{2}\left(y_i\frac{\partial}{\partial z_i} - z_i\frac{\partial}{\partial y_i}\right)\right]^2 + \left[\sum_{i=1}^{2}\left(z_i\frac{\partial}{\partial x_i} - x_i\frac{\partial}{\partial z_i}\right)\right]^2\\ &+ \left[\sum_{i=1}^{2}\left(x_i\frac{\partial}{\partial y_i} - y_i\frac{\partial}{\partial x_i}\right)\right]^2 + l(l+1)\Bigg\}\psi = 0,\end{aligned} \tag{3.2}$$

where $l = 0, 1, \ldots$. Let θ', ϕ, ϕ' be three Euler angles such that (r_1, θ', ϕ') are the spherical coordinates of the first electron in the fixed system $o - xyz$, ϕ is the interfractial angle between the $r_1 - z$ plane and the $r_1 - r_2$ plane, and θ is the interelectronic angle. We introduce the Hylleraas–Breit transform [7]:

$$\begin{cases} x_1 = r_1\sin\theta'\cos\phi', \\ y_1 = r_1\sin\theta'\sin\phi', \\ z_1 = r_1\cos\theta', \\ x_2 = r_2(\sin\theta\cos\theta'\cos\phi\cos\phi' - \sin\theta\sin\phi\sin\phi' + \cos\theta\sin\theta'\cos\phi'), \\ y_2 = r_2(\sin\theta\cos\phi\cos\theta'\sin\phi' + \sin\theta\sin\phi\cos\phi' + \cos\theta\sin\theta'\sin\phi'), \\ z_2 = r_2(\cos\theta\cos\theta' - \sin\theta\sin\theta'\cos\phi). \end{cases} \tag{3.3}$$

We can transform (3.1) and (3.2) into the following forms by (3.3):

$$L(\psi) - \frac{A_1(\psi)}{r_1^2} - \frac{A_2(\psi)}{r_2^2} = E\psi, \tag{3.4}$$

$$\left[\frac{\partial^2}{\partial\theta'^2} + \mathrm{ctg}\theta'\frac{\partial}{\partial\theta'} + \frac{1}{\sin^2\theta'}\left(\frac{\partial^2}{\partial\phi^2} + \frac{\partial^2}{\partial\phi'^2}\right) - \frac{2\cos\theta'}{\sin^2\theta'}\frac{\partial^2}{\partial\phi\partial\phi'} + l(l+1)\right]\psi = 0, \tag{3.5}$$

where

$$\begin{aligned} L(\psi) = &-\frac{1}{r_1^2}\frac{\partial}{\partial r_1}\left(r_1^2\frac{\partial\psi}{\partial r_1}\right) - \frac{1}{r_2^2}\frac{\partial}{\partial r_2}\left(r_2^2\frac{\partial\psi}{\partial r_2}\right)\\ &- \left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right)\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \left(\frac{2}{r_{12}} - \frac{4}{r_1} - \frac{4}{r_2}\right)\psi,\\ A_1(\psi) = &\frac{1}{\sin\theta'}\frac{\partial}{\partial\theta'}\left(\sin\theta'\frac{\partial\psi}{\partial\theta'}\right) + (\mathrm{ctg}^2\theta + \mathrm{ctg}^2\theta' + 2\mathrm{ctg}\theta\mathrm{ctg}\theta'\cos\phi)\frac{\partial^2\psi}{\partial\phi^2}\\ &+ \frac{1}{\sin\theta'}\frac{\partial^2\psi}{\partial\phi'^2} - 2\cos\phi\frac{\partial^2\psi}{\partial\theta\partial\theta'} - 2\frac{\sin\phi}{\sin\theta'}\frac{\partial^2\psi}{\partial\theta\partial\phi'} + 2\sin\phi\mathrm{ctg}\theta\frac{\partial^2\psi}{\partial\phi\partial\theta'}\\ &+ 2\mathrm{ctg}\theta'\sin\phi\frac{\partial^2\psi}{\partial\theta\partial\phi} - \frac{2}{\sin\theta'}(\mathrm{ctg}\theta' + \mathrm{ctg}\theta\cos\phi)\frac{\partial^2\psi}{\partial\phi\partial\phi'},\\ A_2(\psi) = &\frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}. \end{aligned}$$

For the S-state of the helium atom, $l = 0$; then (3.5) has only constant solutions. Thus any wave function u of the S-state depends on only three variables r_1, r_2, and θ. Therefore, we can transform (3.4) of the S-state into a three-dimensional form

$$L(u) = Eu. \tag{3.6}$$

Assume that $\Omega \subset X = [0, +\infty) \times [0, +\infty) \times [0, \pi]$ is a bounded domain. $u = u(r_1, r_2, \theta)$, $v = v(r_1, r_2, \theta)$. We define inner products, norms, and Hilbert spaces as follows:

$$(u, v)_0 = \int_\Omega uvr_1^2 r_2^2 \sin\theta dr_1 dr_2 d\theta, \quad \|u\|_{0,r,\Omega}^2 = \int_\Omega u^2 r_1^2 r_2^2 \sin\theta dr_1 dr_2 d\theta,$$

$$(u, v)_1 = (u, v)_0 + \int_\Omega \left[r_1^2 r_2^2 \frac{\partial u}{\partial r_1}\frac{\partial v}{\partial r_1} + r_1^2 r_2^2 \frac{\partial u}{\partial r_2}\frac{\partial v}{\partial r_2} + (r_1^2 + r_2^2)\frac{\partial u}{\partial \theta}\frac{\partial v}{\partial \theta}\right] \sin\theta dr_1 dr_2 d\theta,$$

$$|u|_{1,r,\Omega}^2 = \int_\Omega \left[r_1^2 r_2^2 \left(\frac{\partial u}{\partial r_1}\right)^2 + r_1^2 r_2^2 \left(\frac{\partial u}{\partial r_2}\right)^2 + (r_1^2 + r_2^2)\left(\frac{\partial u}{\partial \theta}\right)^2\right] \sin\theta dr_1 dr_2 d\theta,$$

$$|u|_{2,r,\Omega}^2 = \left|\frac{\partial u}{\partial r_1}\right|_{1,r,\Omega}^2 + \left|\frac{\partial u}{\partial r_2}\right|_{1,r,\Omega}^2 + \left|\frac{\partial u}{\partial \theta}\right|_{1,r,\Omega}^2,$$

$$\|u\|_{i+1,r,\Omega}^2 = \|u\|_{i,r,\Omega}^2 + |u|_{i+1,r,\Omega}^2, \quad i = 0, 1,$$

$$H_r^i(\Omega) = \left\{ v \mid \ \|v\|_{i,r,\Omega}^2 < +\infty \right\}, \quad i = 0, 1,$$

$$H_r^2(\Omega) = \left\{ v \| v\|_{2,r,\Omega}^2 < +\infty, \ \int_\Omega u^2 (r_1^2 + r_2^2) \sin\theta dr_1 dr_2 d\theta < \infty \right\}.$$

LEMMA 3.1. *Assume that $\Omega \subset X$ is a bounded domain; then*

$$H_r^1(\Omega) \hookrightarrow\hookrightarrow H_r^0(\Omega). \tag{3.7}$$

Furthermore, if there exists a constant $d_\Omega > 0$, such that for any $(r_1, r_2, \theta) \in \Omega$, $r_1, r_2 \geq d_\Omega$, then

$$H_r^2(\Omega) \hookrightarrow\hookrightarrow H_r^1(\Omega). \tag{3.8}$$

Proof. (1): Proof of (3.7). Let $x = (x_1, y_1, z_1, x_2, y_2, z_2) = \mathsf{H}(r_1, r_2, \theta, \theta', \phi, \phi')$ be the Hylleraas–Breit transform defined by (3.3), and let $\hat{\Omega} \subset R^6$ be a bounded domain defined by

$$\hat{\Omega} = \{x \mid x = \mathsf{H}(r_1, r_2, \theta, \theta', \phi, \phi'), \ (r_1, r_2, \theta) \in \Omega, \ 0 \leq \theta' \leq \pi, \ 0 \leq \phi, \phi' \leq 2\pi\}. \tag{3.9}$$

The Jacobian determinant of (3.3) is

$$\det\left(\frac{\partial(x_1, y_1, z_1, x_2, y_2, z_2)}{\partial(r_1, r_2, \theta, \theta', \phi, \phi')}\right) = r_1^2 r_2^2 \sin\theta \sin\theta'. \tag{3.10}$$

By direct calculation, we have $\|u\|_{i,r,\Omega}^2 = \frac{1}{8\pi^2}\|u\|_{i,\hat{\Omega}}^2$, where $\|\cdot\|_{i,\hat{\Omega}}$ is the norm of the standard Sobolev space $H^i(\hat{\Omega})$, $i = 0, 1$. It is easy to show that $H_r^0(\Omega)$ and $H_r^1(\Omega)$ are Hilbert spaces, and $H_r^1(\Omega) \hookrightarrow\hookrightarrow H_r^0(\Omega)$.

(2): Proof of (3.8). Let $\{v_n\}$ be a bounded sequence in $H_r^2(\Omega)$; then $\{v_n\}, \{\frac{\partial v_n}{\partial r_1}\}$, $\{\frac{\partial v_n}{\partial r_2}\}, \{\frac{\partial v_n}{\partial \theta}\}$ are bounded uniformly in $H_r^1(\Omega)$. By (1), we can choose (successively) a subsequence denoted as $\{v_n\}$ too, such that $\{v_n\}, \{\frac{\partial v_n}{\partial r_1}\}, \{\frac{\partial v_n}{\partial r_2}\}, \{\frac{\partial v_n}{\partial \theta}\}$ are Cauchy sequences in $H_r^0(\Omega)$. Thus $\{v_n\}$ is a Cauchy sequence in the measure

$$\left\{ \|v_n\|_{0,r,\Omega}^2 + \int_\Omega \left(\left| \frac{\partial v_n}{\partial r_1} \right|^2 + \left| \frac{\partial v_n}{\partial r_2} \right|^2 + \left| \frac{\partial v_n}{\partial \theta} \right|^2 \right) r_1^2 r_2^2 \sin\theta dr_1 dr_2 d\theta \right\}^{1/2}.$$

It is clear that

$$\int_\Omega \left| \frac{\partial v_n}{\partial \theta} \right|^2 (r_1^2 + r_2^2) \sin\theta dr_1 dr_2 d\theta \le \frac{2}{d_\Omega^2} \int_\Omega \left| \frac{\partial v_n}{\partial \theta} \right|^2 r_1^2 r_2^2 \sin\theta dr_1 dr_2 d\theta,$$

and so we have

$$\|v_n\|_{1,r,\Omega}^2 \le \max\left\{1, 2/d_\Omega^2\right\} \left(\|v_n\|_{0,r,\Omega}^2 + \left\| \frac{\partial v_n}{\partial r_1} \right\|_{0,r,\Omega}^2 + \left\| \frac{\partial v_n}{\partial r_2} \right\|_{0,r,\Omega}^2 + \left\| \frac{\partial v_n}{\partial \theta} \right\|_{0,r,\Omega}^2 \right).$$

Therefore $\{v_n\}$ is a Cauchy sequence in $H_r^1(\Omega)$ and converges. □

Remark 3.2. Suppose that Ω is bounded. For any integer $k > 0$, define

$$P_k(\Omega) = \left\{ p = \sum_{0 \le l+m+n \le k} \alpha_{lmn} r_1^l r_2^m \theta^n \;\middle|\; \alpha_{lmn} \in R^1,\ (r_1, r_2, \theta) \in \Omega \right\};$$

then $P_k(\Omega) \subset H^2(\Omega) \subset H_r^2(\Omega)$.

LEMMA 3.3. *Supposing $R > 0$ is a constant, there exists a constant C independent of R such that, for any $f \in H^1([0, R])$,*

$$\int_0^R f^2 dx \le C \max\{R^2, R^{-2}\} \left(\int_0^R x^2 f^2 dx + \int_0^R f'^2 dx \right). \tag{3.11}$$

Proof. For any $f \in C^1([R/2, R])$ there exists $\xi \in [R/2, R]$ such that

$$f(\xi) = \frac{2}{R} \int_{R/2}^R f(x) dx.$$

Then for any $x \in [R/2, R]$, by Hölder's inequality we have

$$\begin{aligned} |f(x)| &= \left| f(\xi) + \int_\xi^x f'(t) dt \right| \le \frac{2}{R} \int_{R/2}^R |f(t)| dt + \int_{R/2}^R |f'(t)| dt \\ &\le \left(\frac{2}{R} \int_{R/2}^R |f(t)|^2 dt \right)^{1/2} + \left(\frac{R}{2} \int_{R/2}^R |f'(t)|^2 dt \right)^{1/2}. \end{aligned}$$

Thus there exists a positive constant C independent of R such that

$$|f|_{0,\infty,[R/2,R]} \le C \max\{\sqrt{R}, 1/\sqrt{R}\} \|f\|_{1,[R/2,R]}. \tag{3.12}$$

Set $x_0 \in [R/2, R]$. By (3.12), we have, for any $x \in [0, R]$ and $f \in C^\infty([0, R])$,

$$f^2(x) \le 2\left\{f^2(x_0) + \left[\int_{x_0}^{x} f'(x)dx\right]^2\right\} \le C\left\{f^2(x_0) + R\int_0^R f'^2(x)dx\right\} \tag{3.13}$$
$$\le C\max\{R, R^{-3}\}\left(\int_{R/2}^R x^2 f^2(x)dx + \int_0^R f'^2(x)dx\right) \quad \forall x \in [0, R],$$

where C is a constant independent of R. We obtain (3.11) for all functions in $C^\infty([0, R])$ by integrating both sides of (3.13) over $[0, R]$. Therefore (3.11) is true for all functions in $H^1([0, R])$ by the density of $C^\infty([0, R])$ in $H^1([0, R])$. □

Now, we expand each function $u(r_1, r_2, \theta)$ in some Banach space defined on Ω to $R^3 \setminus \Omega$. Thus r_1, r_2, θ are not the distances and the interelectronic angle in the previous sense. For any $u = u(r_1, r_2, \theta)$, define

$$|||u|||^2_{0,\Omega} = \int_\Omega u^2 r_1^2 r_2^2 |\sin\theta| dr_1 dr_2 d\theta, \tag{3.14}$$

$$|||u|||^2_{1,\Omega} = \int_\Omega \left[\left(\frac{\partial u}{\partial r_1}\right)^2 + \left(\frac{\partial u}{\partial r_2}\right)^2 + \left(\frac{\partial u}{\partial \theta}\right)^2\right](r_1^2 + r_2^2)|\sin\theta| dr_1 dr_2 d\theta, \tag{3.15}$$

$$H(\Omega) = \{u \mid |||u|||^2_{0,\Omega} + |||u|||^2_{1,\Omega} < \infty\}.$$

THEOREM 3.4. *Suppose that $\Omega \subset X$ is a bounded open domain satisfying C^1-regularity* [2]. *There exists a linear operator*

$$\mathsf{E}: \quad H(\Omega) \to H(R^3)$$

such that, for any $u \in H(\Omega)$,

$$\mathsf{E}u(r_1, r_2, \theta) = u(r_1, r_2, \theta), \quad \textit{almost everywhere in } (r_1, r_2, \theta) \in \Omega, \tag{3.16}$$

$$|||\mathsf{E}u|||_{i,R^3} \le C|||u|||_{i,\Omega}, \quad i = 0, 1, \tag{3.17}$$

where C is a constant depending on Ω.

Proof. Let $B(0, 1)$ be the open unit ball in R^3. Since Ω is bounded and C^1-regular, there exist a finite number of bounded open sets $O_0, \ldots, O_M$ such that $O_0 \subset\subset \Omega$, $\partial\Omega \subset \cup_{i=1}^M O_i$, and $\Omega \subset \cup_{i=0}^M O_i$, and there exist $m + 1$ transforms $(\xi, \eta, \zeta) = \varphi_i(r_1, r_2, \theta)$ such that

$$\varphi_i(O_i) = B(0, 1), \quad \varphi_i(O_i \cap \partial\Omega) = \Sigma = B(0, 1) \cap \{\zeta = 0\},$$
$$\varphi_i(O_i \cap \Omega) = B^+(0, 1) = \{(\xi, \eta, \zeta) \in B(0, 1) \mid \zeta > 0\},$$
$$\varphi_i \in C^1(\overline{O_i \cap \partial\Omega}), \quad \varphi_i^{-1} \in C^1(\bar{\Sigma}).$$

For the convenience of notation, define $\mathbf{x} = (r_1, r_2, \theta)$ and $\hat{\mathbf{x}} = (\xi, \eta, \zeta)$. We choose a partition of unity $\{\alpha_i\}$ associated with $\{O_i\}$ satisfying

$$\alpha_i \in C_0^\infty(O_i),\ 0 \le \alpha_i \le 1,\ 0 \le i \le M, \qquad \sum_{i=0}^M \alpha_i(\mathbf{x}) = 1 \quad \forall \mathbf{x} \in \Omega.$$

(1): Proof for functions $u \in C^1(\bar{\Omega})$. Set $u_i = u\alpha_i$; then $u = \sum_{i=0}^M u_i$. Let

$$\begin{aligned} &\varrho(\mathbf{x}) = r_1^2 r_2^2 |\sin\theta|, && \varrho_i(\hat{\mathbf{x}}) = \varrho(\varphi_i^{-1}(\hat{\mathbf{x}})), \\ &\omega(\mathbf{x}) = (r_1^2 + r_2^2)|\sin\theta|, && \omega_i(\hat{\mathbf{x}}) = \omega(\varphi_i^{-1}(\hat{\mathbf{x}})), \\ &v_i(\hat{\mathbf{x}}) = u_i(\varphi_i^{-1}(\hat{\mathbf{x}})) = u_i(\mathbf{x}). \end{aligned} \tag{3.18}$$

For any $0 \le i \le M$, since $\frac{\partial v_i}{\partial \hat{\mathbf{x}}_k} = \sum_{l=1}^{3} \frac{\partial u_i}{\partial \mathbf{x}_l} \frac{\partial \mathbf{x}_l}{\partial \hat{\mathbf{x}}_k}$, $\frac{\partial u_i}{\partial \mathbf{x}_l} = \sum_{k=1}^{3} \frac{\partial v_i}{\partial \hat{\mathbf{x}}_k} \frac{\partial \hat{\mathbf{x}}_k}{\partial \mathbf{x}_l}$, there exists a constant C depending only on φ_i and φ_i^{-1} such that

$$\int_{B^+(0,1)} |v_i|^2 \varrho_i d\hat{\mathbf{x}} = \int_{\Omega \bigcap O_i} |u_i|^2 \varrho J_i d\mathbf{x} \le C \, |||u_i|||^2_{0,\Omega \bigcap O_i} , \tag{3.19}$$

$$\begin{aligned} \int_{B^+(0,1)} \left| \frac{\partial v_i}{\partial \hat{\mathbf{x}}_k} \right|^2 \omega_i d\hat{\mathbf{x}} &\le C \sum_{l=1}^{3} \int_{B^+(0,1)} \left| \frac{\partial u_i}{\partial \mathbf{x}_l} \right|^2 \left| \frac{\partial \mathbf{x}_l}{\partial \hat{\mathbf{x}}_k} \right|^2 \omega_i d\hat{\mathbf{x}} \\ &\le C \sum_{l=1}^{3} \left| \left| \left| \frac{\partial u_i}{\partial \mathbf{x}_l} \right| \right| \right|^2_{0,\Omega \bigcap O_i} , \end{aligned} \tag{3.20}$$

where $J_i = \det(\frac{\partial \varphi_i}{\partial \mathbf{x}})$ is the Jacobian determinant. Similarly, we have

$$|||u_i|||^2_{0,\Omega \bigcap O_i} \le C \int_{B^+(0,1)} |v_i|^2 \varrho_i d\hat{\mathbf{x}}, \tag{3.21}$$

$$\left| \left| \left| \frac{\partial u_i}{\partial \mathbf{x}_l} \right| \right| \right|^2_{0,\Omega \bigcap O_i} \le C \sum_{k=1}^{3} \int_{B^+(0,1)} \left| \frac{\partial v_i}{\partial \hat{\mathbf{x}}_k} \right|^2 \omega_i d\hat{\mathbf{x}}. \tag{3.22}$$

We expand u_0 by zero to the exterior of O_0 and denote the extension as $\tilde{u}_0$; then $\tilde{u}_0 \in C_0^1(R^3)$. We expand $v_i(\hat{\mathbf{x}})$ as follows:

$$\tilde{v}_i(\hat{\mathbf{x}}) = \begin{cases} v_i(\hat{\mathbf{x}}), & \hat{\mathbf{x}} \in B^+(0,1) \cup \Sigma, \\ 4v_i(\xi, \eta, -\frac{1}{2}\zeta) - 3v_i(\xi, \eta, -\zeta), & \hat{\mathbf{x}} \in B^-(0,1), \end{cases} \tag{3.23}$$

where $B^-(0,1) = \{\hat{\mathbf{x}} \in B(0,1) \,|\, \zeta < 0\}$. Obviously, $\tilde{v}_i \in C^1(B^+(0,1) \cup B^-(0,1))$. For any multiple index $\alpha \in \{(0,0,0), (1,0,0), (0,1,0), (0,0,1)\}$ and $\hat{\mathbf{x}}^0 \in \Sigma$, we have

$$\lim_{\substack{\hat{\mathbf{x}} \in B^+(0,1) \\ \hat{\mathbf{x}} \to \hat{\mathbf{x}}^0}} D^\alpha \tilde{v}_i(\hat{\mathbf{x}}) = D^\alpha v_i(\hat{\mathbf{x}}^0),$$

$$\begin{aligned} \lim_{\substack{\hat{\mathbf{x}} \in B^-(0,1) \\ \hat{\mathbf{x}} \to \hat{\mathbf{x}}^0}} D^\alpha \tilde{v}_i(\hat{\mathbf{x}}) &= \lim_{\substack{\hat{\mathbf{x}} \in B^-(0,1) \\ \hat{\mathbf{x}} \to \hat{\mathbf{x}}^0}} \left[4\left(-\frac{1}{2}\right)^{\alpha_3} D^\alpha v_i\left(\xi, \eta, -\frac{1}{2}\zeta\right) + 3(-1)^{\alpha_3+1} D^\alpha v_i(\xi, \eta, -\zeta) \right] \\ &= D^\alpha v_i(\hat{\mathbf{x}}^0). \end{aligned}$$

Thus $\tilde{u}_i(\mathbf{x}) = \tilde{v}_i(\varphi_i(\mathbf{x})) \in C_0^1(O_i)$. Expand $\tilde{u}_i$ by zero to the exterior of O_i and denote the extension by $\tilde{u}_i$ too; then $\tilde{u}_i \in C_0^1(R^3)$. Define $\mathsf{E}u = \sum_{i=0}^{M} \tilde{u}_i$; then $(\mathsf{E}u)(\mathbf{x}) = \sum_{i=0}^{M} u_i(\mathbf{x}) = u(\mathbf{x})$ $\forall \mathbf{x} \in \Omega$. Combing (3.19)–(3.23) yields (3.16) and (3.17).

(2): Proof for functions $u \in H(\Omega)$. Let $\hat{\Omega}$ be defined as in (3.9). In view of

$$\|u\|^2_{i,r,\Omega} = \frac{1}{8\pi^2} \|u\|^2_{i,\hat{\Omega}},$$

we have $H_r^i(\Omega) \hookrightarrow H^i(\hat{\Omega})$, $i = 0, 1$, in the sense of isomorphism. Since $C^1(\overline{\hat{\Omega}})$ is dense in $H^i(\hat{\Omega})$, for any $v(r_1, r_2, \theta) \in H_r^i(\Omega) \hookrightarrow H^i(\hat{\Omega})$ there exists $\{v_n(r_1, r_2, \theta)\} \subset C^1(\overline{\hat{\Omega}})$ such that $v_n(r_1, r_2, \theta)$ converge to $v(r_1, r_2, \theta)$ in $H^i(\hat{\Omega})$, and hence in $H_r^i(\Omega)$, $i = 0, 1$. Since $\frac{\partial v_n}{\partial \xi} = \sum_{i=1}^{6} \frac{\partial v_n}{\partial x_i} \frac{\partial x_i}{\partial \xi}$, $\frac{\partial v_n}{\partial x_i}$, and $\frac{\partial x_i}{\partial \xi}$ are continuous and bounded, we

have $v_n(r_1, r_2, \theta) \in C^1(\bar{\Omega})$, where $\xi = r_1, r_2, \theta, \theta', \phi, \phi'$. Thus $C^1(\bar{\Omega})$ is dense in $H_r^i(\Omega)$, $i = 0, 1$.

Since $\Omega \subset X$, $C^1(\bar{\Omega})$ is dense in $H(\Omega)$ in view of $H(\Omega) \hookrightarrow H_r^1(\Omega)$. There exists a sequence $\{u_n\} \subset C^1(\bar{\Omega})$ converging to u in $H(\Omega)$. By (3.18), $\{\mathsf{E}u_n\}$ is a Cauchy sequence in $H(R^3)$ and hence converges to some $w \in H(R^3)$. Set $\mathsf{E}u = w$. Since $|||u_n - \mathsf{E}u|||_{i,\Omega} = |||\mathsf{E}u_n - \mathsf{E}u|||_{i,\Omega} \to 0$, we have $\mathsf{E}u = u$, a.e. in Ω. Furthermore,

$$|||\mathsf{E}u|||_{i,R^3} = \lim_{n\to\infty} |||\mathsf{E}u_n|||_{i,R^3} \le C \lim_{n\to\infty} |||u_n|||_{i,\Omega} = C|||u|||_{i,\Omega}, \quad i = 0, 1.$$

The proof is complete. □

We define $\Omega_{r_1,\theta}$ as the projection of Ω onto the $r_1-\theta$ plane. For any $(r_1, \theta) \in \Omega_{r_1,\theta}$, define

$$\Omega_{r_2}(r_1, \theta) = \{r_2 \mid (r_1, r_2, \theta) \in \Omega\}.$$

$\Omega_{r_2,\theta}$ and $\Omega_{r_1}(r_2, \theta)$ are defined in the same way.

THEOREM 3.5. *Let $\Omega \subset X$ be a bounded domain satisfying one of the following two conditions:*

(a) *Ω is C^1-regular;*

(b) *The boundary $\partial\Omega$ is Lipschitz continuous, and there exists a constant $d > 0$ such that, for almost every $(r_2, \theta) \in \Omega_{r_2,\theta}$ (respectively, $(r_1, \theta) \in \Omega_{r_1,\theta}$), if $\Omega_r \subset \Omega_{r_1}(r_2, \theta)$ (respectively, $\Omega_{r_2}(r_1, \theta)$) is a maximal simply connected set, then meas$(\Omega_r) \ge d$.*

Furthermore, we assume that there are $f_1, f_2, \ldots, f_M \in (H_r^m(\Omega))'$ satisfying

$$\forall p \in P_{m-1}(\Omega), \quad \sum_{i=1}^{M} f_i(p) = 0 \Longleftrightarrow p = 0. \tag{3.24}$$

Then there exists a constant $C(\Omega)$ such that

$$\|v\|_{m,r,\Omega} \le C(\Omega) \left\{ |v|_{m,r,\Omega} + \left| \sum_{i=1}^{M} f_i(v) \right| \right\} \quad \forall v \in H_r^m(\Omega),\ m = 1, 2. \tag{3.25}$$

Proof. (1) In view of (3.7), we can prove (3.25) in the case of $m = 1$ by the argument of Theorem 3.1.1 in [9, p. 115].

(2) Proof of (3.25) in the case of $m = 2$. If (3.25) were false, then for any integer $n > 0$ there should exist $v_n \in H_r^2(\Omega)$ such that $\|v_n\|_{2,r,\Omega} = 1$ and

$$|v_n|_{2,r,\Omega} + \left| \sum_{i=1}^{M} f_i(v_n) \right| < \frac{1}{n}. \tag{3.26}$$

In view of Theorem 3.4, the definition of $\|\cdot\|_{2,r,\Omega}$, and (3.25) for $m = 1$, there exists a subsequence of $\{v_n\}$ (also denoted as $\{v_n\}$), which is a Cauchy sequence under the following measures:

$$\begin{cases} \|\cdot\|_{0,r,\Omega}, \quad \left\| \dfrac{\partial \cdot}{\partial\theta} \right\|_{0,r,\Omega}, \quad \left\| \dfrac{\partial \cdot}{\partial r_i} \right\|_{0,r,\Omega}, \quad i = 1, 2, \\ \left\{ \displaystyle\int_\Omega \left[\left(\frac{\partial^2 \cdot}{\partial r_1 \partial\theta} \right)^2 + \left(\frac{\partial^2 \cdot}{\partial r_2 \partial\theta} \right)^2 + \left(\frac{\partial^2 \cdot}{\partial\theta^2} \right)^2 \right] (r_1^2 + r_2^2) \sin\theta dr_1 dr_2 d\theta \right\}^{1/2}. \end{cases} \tag{3.27}$$

(i) Suppose that Ω satisfies the condition (a). We set $u_n = \frac{\partial v_n}{\partial \theta}$; then $u_n \in H(\Omega)$. Choose $R > 0$ to be sufficiently large such that $\Omega \subset [0, R] \times [0, R] \times [0, \pi]$. By Lemma 3.3 and Theorem 3.4, we have

$$\int_{\Omega_{r_2}(r_1,\theta)} \left(\frac{\partial v_n}{\partial \theta}\right)^2 r_1^2 \sin\theta dr_2 \le \int_0^R (\mathsf{E}u_n)^2 r_1^2 \sin\theta dr_2$$
$$\le C\left\{\int_0^R (\mathsf{E}u_n)^2 r_1^2 r_2^2 \sin\theta dr_2 + \int_0^R \left(\frac{\partial \mathsf{E}u_n}{\partial r_2}\right)^2 r_1^2 \sin\theta dr_2\right\} \quad \forall (r_1, \theta) \in \Omega_{r_1,\theta},$$

where C depends on R. By Lemma 3.3, integrating both sides of the above inequality over $\Omega_{r_1,\theta}$ produces

$$\int_\Omega \left(\frac{\partial v_n}{\partial \theta}\right)^2 r_1^2 \sin\theta dr_1 dr_2 d\theta \le C \sum_{i=0}^{1} |||\mathsf{E}u_n|||^2_{i,[0,R]\times[0,R]\times[0,\pi]} \le C \sum_{i=0}^{1} |||u_n|||^2_{i,\Omega}. \tag{3.28}$$

Thus $\{v_n\}$ is a Cauchy sequence under $[\int_\Omega (\frac{\partial \cdot}{\partial \theta})^2 r_1^2 \sin\theta dr_1 dr_2 d\theta]^{1/2}$. Similarly, $\{v_n\}$ is a Cauchy sequence under $[\int_\Omega (\frac{\partial \cdot}{\partial \theta})^2 r_2^2 \sin\theta dr_1 dr_2 d\theta]^{1/2}$. Consequently, $\{v_n\}$ is a Cauchy sequence in $H_r^1(\Omega)$ by (3.27), and so in $H_r^2(\Omega)$ by (3.26). Suppose $v_n \to v \in H_r^2(\Omega)$; then $|v|_{2,r,\Omega} = 0$ and $\sum_{i=1}^M f_i(v) = 0$ by (3.26). Thus $v \in P_1(\Omega)$, and $v = 0$ by (3.24). Therefore, $v = 0$ contradicts the following identities: $\|v\|_{2,r,\Omega} = \lim_{n\to\infty} \|v_n\|_{2,r,\Omega} = 1$. Thus (3.25) is true for $m = 2$.

(ii) Suppose that Ω satisfies the condition (b). Without loss of generality, we may suppose that $\Omega_{r_2}(r_1, \theta)$ is simply connected for any $(r_1, \theta) \in \Omega_{r_1,\theta}$. By Lemma 3.3, we have

$$\int_{\Omega_{r_2}(r_1,\theta)} \left(\frac{\partial v_n}{\partial \theta}\right)^2 r_1^2 \sin\theta dr_2 \le C \int_{\Omega_{r_2}(r_1,\theta)} \left[\left(\frac{\partial v_n}{\partial \theta}\right)^2 r_2^2 + \left(\frac{\partial^2 v_n}{\partial r_2 \partial \theta}\right)^2\right] r_1^2 \sin\theta dr_2$$

$\forall (r_1, \theta) \in \Omega_{r_1,\theta}$, where C depends on d but is independent of (r_1, θ). Thus we can get (3.25) by the argument of (i). □

Remark 3.6. Assume that Ω satisfies the conditions in Theorem 3.5. Define

$$|\Omega| = \int_\Omega r_1^2 r_2^2 \sin\theta dr_1 dr_2 d\theta, \quad f(v) = \frac{1}{|\Omega|} \int_\Omega v r_1^2 r_2^2 \sin\theta dr_1 dr_2 d\theta.$$

By Hölder's inequality, we have $|f(v)| \le |\Omega|^{-2} \|v\|_{0,r,\Omega}$. Thus $f \in (H_r^1(\Omega))'$. Now (3.26) implies

$$\|v - f(v)\|_{1,r,\Omega} \le |v|_{1,r,\Omega} \quad \forall v \in H_r^1(\Omega). \tag{3.29}$$

Remark 3.7. Assume $R > 0$ and $\Omega = [0, R] \times [0, R] \times [0, \pi]$. Define

$$H_{0r}^1(\Omega) = \{v \in H_r^1(\Omega) |\ v|_{r_1=R} = v|_{r_2=R} = 0\}.$$

Then we have $H_{0r}^1(\Omega) \hookrightarrow H_0^1(\hat{\Omega})$ in the sense of isomorphism, and, by Poincáre's inequality,

$$\|v\|_{1,r,\Omega} = \frac{\sqrt{2}}{4\pi} \|v\|_{1,\hat{\Omega}} \le C(\Omega)|v|_{1,\hat{\Omega}} = \sqrt{8}\pi C(\Omega)|v|_{1,r,\Omega}. \tag{3.30}$$

Equations (3.29) and (3.30) are the so-called Friedrichs-type inequality and Poincáre-type inequality, respectively.

THEOREM 3.8. *Let $\Omega \subset X$ be the domain in Theorem 3.5 and $T: H_r^k(\Omega) \to H_r^m(\Omega)$ be a linear and continuous mapping satisfying*

$$Tp = p \quad \forall p \in P_{k-1}(\Omega). \tag{3.31}$$

Then there exists a constant $C(\Omega)$ such that

$$\|u - Tu\|_{m,r,\Omega} \le C(\Omega)|u|_{k,r,\Omega} \quad \forall u \in H_r^k(\Omega), \tag{3.32}$$

where $0 \le m \le 2$, $1 \le k \le 2$, $m \le k$.

Proof. (3.32) can be proved by virtue of Theorem 3.5 and the argument of Theorem 3.1.4 in [9, p. 121]. □

4. Local regularization operator. Because we cannot say that the solutions of (2.13) and (5.1) are continuous, difficulties appear in proving the convergency of the finite element scheme. The technique of local regularization (Clément's interpolation [11]) will be used. Thus we are in the position of describing the construction of the three-dimensional local regularization operator.

Set $R > 0$ be large enough, and define

$$\begin{aligned}
&\Omega = [0,R] \times [0,R] \times [0,\pi], && \partial\Omega = \{(r_1,r_2,\theta) \in \Omega |\ r_1 = R \text{ or } r_2 = R\},\\
&\Gamma_1 = \{(r_1,r_2,\theta) \in \partial\Omega \,|\, r_1 = 0\}, && \Gamma_2 = \{(r_1,r_2,\theta) \in \partial\Omega \,|\, r_2 = 0\},\\
&\Gamma_3 = \{(r_1,r_2,\theta) \in \partial\Omega \,|\, \theta = 0\}, && \Gamma_4 = \{(r_1,r_2,\theta) \in \partial\Omega \,|\, \theta = \pi\}.
\end{aligned}$$

Suppose that $\mathcal{T}_h$ is a regular subdivision of Ω. Each element in $\mathcal{T}_h$ is a cuboid. (We can also obtain similar results for regular hexahedrons, but the analysis is very tedious.) h is the maximal diameter of all elements. The regularity of K means that there exists a constant σ independent of K such that $\forall K \in \mathcal{T}_h$, $h_K \le \sigma|e|$; here h_K is the diameter of K, e is any edge of K. Denote all nodes of $\bar{\Omega}$ by $A_1, A_2, \ldots, A_I$, and define $\triangle_i = \cup_{K \in \mathcal{T}_h, A_i \in K} K$ as the macro element associated with the node A_i.

$$Q_k(K) = \left\{ p \,|\, p = \sum_{l,m,n=0}^{k} \alpha_{lmn} r_1^l r_2^m \theta^n,\ (r_1,r_2,\theta) \in K \right\}.$$

Since the behaviors of the weights (see (3.18)) on an inner element differ from those on a boundary element, different kinds of elements or macro elements must be affine equivalent to different reference elements or macro elements, respectively. Each macro element must be one of the following four cases:

1. An element in $\mathcal{T}_h$. Let $l_{10} = [0,1]$, $l_{11} = [1,2]$; its affine equivalent reference macro element must be one of $l_{1i} \times l_{1j} \times l_{1k}$, $0 \le i,j,k \le 1$.
2. The combination of two elements with a common face. Let $l_{20} = [0,2]$, $l_{21} = [1,3]$; its affine equivalent reference macro element must be one of $l_{2i} \times l_{1j} \times l_{1k}$, $l_{1i} \times l_{2j} \times l_{1k}$, or $l_{1i} \times l_{1j} \times l_{2k}$, $0 \le i,j,k \le 1$.
3. The combination of four elements with a common edge. Its affine equivalent reference macro element must be one of $l_{1i} \times l_{2j} \times l_{2k}$, $l_{2i} \times l_{1j} \times l_{2k}$, or $l_{2i} \times l_{2j} \times l_{1k}$, $0 \le i,j,k \le 1$.
4. The combination of eight elements with a common vertex. Its affine equivalent reference macro element must be one of $l_{2i} \times l_{2j} \times l_{2k}$, $0 \le i,j,k \le 1$.

Each reference element is a cube with unit volume and is included in some reference macro element. Clearly, the total number of reference elements and reference macro elements is finite.

Suppose $h_\Delta < \sigma h_K$, for any $K \subset \Delta$, where h_Δ is the diameter of Δ. For any $K \in \mathcal{T}_h$, $F_K : \hat{K} \to K$ is the affine transform from some reference element $\hat{K}$ to K. For any macro element Δ, assume that $\hat{\Delta} = \cup_{K\subset\Delta} F_K^{-1}(K)$ is a macro reference element defined in cases $1, 2, 3$, or 4. Define F_Δ: $F_\Delta|_K = F_K$ and F_Δ^{-1}: $F_\Delta^{-1}|_{\hat{K}} = F_K^{-1}$, with $F_K(\hat{K}) = K$.

If v is a function defined on Δ and $\hat{u}$ is defined on $\hat{\Delta}$, denote $\hat{v} := v \circ F_\Delta$ and $u := \hat{u} \circ F_\Delta^{-1}$, respectively. Without ambiguity, we also use piecewise-defined norms on macro elements:

$$|\hat{v}|^2_{m,r,\hat{\Delta}} = \sum_{\hat{K}\subset\hat{\Delta}} |v \circ F_\Delta|^2_{m,r,\hat{K}}, \qquad |u|^2_{m,r,\Delta} = \sum_{K\subset\Delta} |\hat{u} \circ F_\Delta^{-1}|^2_{m,r,K}. \tag{4.1}$$

In view of (4.1), it is easy to prove $v \in H_r^m(\Delta) \Longleftrightarrow \hat{v} \in H_r^m(\hat{\Delta})$, $m = 0, 1, 2$.

We define the H_r^0-projection $\mathcal{P}_{\hat{\Delta}} : H_r^0(\hat{\Delta}) \to P_k(\hat{\Delta})$ as follows: $\forall v \in H_r^0(\hat{\Delta})$,

$$(\mathcal{P}_{\hat{\Delta}} v, p)_0 = (v, p)_0 \quad \forall p \in P_k(\hat{\Delta}). \tag{4.2}$$

Since the weights vanish on some boundary elements $K \cap (\cup_{i=1}^4 \Gamma_i) \neq \emptyset$, we need to deal with their transformations under the affine transforms by detailed analysis. To do so, we first need the following estimate for the transformation of $\sin\theta$. Define

$$\Lambda(\theta, h) = \begin{cases} 1, & \pi/2 - h \le \theta \le \pi/2, \\ \sin\theta, & \theta \ge \pi/2, \\ \sin(\theta + h), & \theta \le \pi/2 - h. \end{cases} \tag{4.3}$$

LEMMA 4.1. *Let $\sigma > 0$ be a constant, $h_1, h_2 \le h \le \sigma \min\{h_1, h_2\}$, $\theta \ge h_1$, and $\theta + h_1 + h_2 + h/\sigma \le \pi$. When h is sufficiently small, there exists a constant C independent of h and θ such that*

$$\Lambda(\theta, h_1) \cdot \max\left\{\frac{1}{\sin\theta}, \frac{1}{\sin(\theta + h_1)}, \frac{1}{\sin(\theta + h_1 + h_2)}\right\} \le C. \tag{4.4}$$

Proof. Let $M = \max\{1/\sin\theta, 1/\sin(\theta + h_1), 1/\sin(\theta + h_1 + h_2)\}$. We consider (4.4) in three cases:

1. When $\pi/2 - h_1 - h_2 \le \theta \le \pi/2$, $\Lambda \cdot M \le M$, since h is small enough, (4.4) is true obviously.
2. When $\theta \ge \pi/2$, it is clear that

$$\Lambda M \le \frac{\sin\theta}{\sin(\theta + h_1 + h_2)} = \frac{\sin(\pi - \theta)}{\sin(\pi - \theta - h_1 - h_2)}.$$

   If $\theta \le \pi - 2(h_1 + h_2)$, then

$$\Lambda M \le \frac{\sin(\pi - \theta)}{\sin((\pi - \theta)/2)} \le 2.$$

   If $\theta \ge \pi - 2(h_1 + h_2)$, we need only to choose h such that $2\sin(h/\sigma) \ge h/\sigma$; then

$$\Lambda M \le \frac{\sin 2(h_1 + h_2)}{\sin(h/\sigma)} \le \frac{4h}{h/(2\sigma)} \le \frac{8}{\sigma}.$$

3. When $\theta \le \pi/2 - h_1 - h_2$, $\Lambda M \le \frac{\sin(\theta + h_1)}{\sin\theta} \le \frac{\sin 2\theta}{\sin\theta} \le 2$.

The proof is complete. ∎

THEOREM 4.2. *Suppose that $\Delta = F_\Delta(\hat{\Delta})$ is a macro element associated with some node of $\mathcal{T}_h$, $u \in H_r^m(\Delta)$. Define $\mathcal{P}_\Delta u = (\mathcal{P}_{\hat{\Delta}}\hat{u}) \circ F_\Delta^{-1}$. Then*

$$|u - \mathcal{P}_\Delta u|_{l,r,\Delta} \le C h_\Delta^{m-l} |u|_{m,r,\Delta}, \quad 0 \le l \le m \le 2, \tag{4.5}$$

where C is a constant independent of h.

Proof. For the sake of simplicity, without loss of generality, we may suppose $\Delta = \cup_{i=1}^4 K_i$ and analyze K_i in two representative cases.

(1) $K_1, \dots, K_4$ are boundary elements where the weights degenerate. Suppose h_Δ is small enough and

$$\begin{aligned} K_1 &= [0,h_1]\times[0,h_2]\times[0,h_3], & K_2 &= [0,h_1]\times[0,h_2]\times[h_3,h_3+h_3^{K_2}],\\ K_3 &= [0,h_1]\times[h_2,h_2+h_2^{K_3}]\times[0,h_3], & K_4 &= [0,h_1]\times[h_2,h_2+h_2^{K_3}]\times[h_3,h_3+h_3^{K_2}]. \end{aligned}$$

The reference macro element and reference elements are defined as

$$\hat{\Delta} = [0,1]\times[0,2]\times[0,2], \quad \hat{K}_1 = [0,1]\times[0,1]\times[0,1],$$
$$\hat{K}_2 = [0,1]\times[0,1]\times[1,2], \quad \hat{K}_3 = [0,1]\times[1,2]\times[0,1], \quad \hat{K}_4 = [0,1]\times[1,2]\times[1,2].$$

On any finite-dimensional space, all norms are equivalent, and so we have

$$\|\mathcal{P}_{\hat{\Delta}}\hat{u}\|_{i,r,\hat{\Delta}}^2 \le C\|\mathcal{P}_{\hat{\Delta}}\hat{u}\|_{0,r,\hat{\Delta}}^2 \le C\|\hat{u}\|_{0,r,\hat{\Delta}}^2. \tag{4.6}$$

Hence the projection $\mathcal{P}_{\hat{\Delta}}$ is stable on $\|\cdot\|_{i,r,\hat{\Delta}}$, $i = 1, 2$. By (4.2) and Theorem 3.8, we have

$$\begin{aligned} \|u - \mathcal{P}_\Delta u\|_{0,r,K_1}^2 &\le h_1^3 h_2^3 h_3 \int_{\hat{K}_1} |\hat{u} - \mathcal{P}_{\hat{\Delta}}\hat{u}|^2 \xi^2 \eta^2 \sin(h_3\zeta) d\xi d\eta d\zeta \\ &\le \max\{2, 1/\sin\zeta_0\} h_1^3 h_2^3 h_3^2 \|\hat{u} - \mathcal{P}_{\hat{\Delta}}\hat{u}\|_{0,r,\hat{K}_1}^2 \\ &\le C h_\Delta^8 |\hat{u}|_{m,r,\hat{\Delta}}^2, \qquad m = 1, 2, \end{aligned} \tag{4.7}$$

where $\zeta_0 \in (0,1)$ satisfies $\zeta_0 \le 2\sin\zeta_0$. Similarly, we have

$$\begin{aligned} \|u - \mathcal{P}_\Delta u\|_{0,r,K_4}^2 &\le h_1^3 h_2^{K_3} h_3^{K_2} (h_2 + h_2^{K_3})^2 \Lambda(h_3, h_3^{K_2}) \int_{\hat{K}_4} |\hat{u} - \mathcal{P}_{\hat{\Delta}}\hat{u}|^2 \xi^2 d\xi d\eta d\zeta \\ &\le C h_1^3 h_2^{K_3} h_3^{K_2} (h_2 + h_2^{K_3})^2 \Lambda(h_3, h_3^{K_2}) \|\hat{u} - \mathcal{P}_{\hat{\Delta}}\hat{u}\|_{0,r,\hat{K}_1}^2 \\ &\le C h_\Delta^8 |\hat{u}|_{m,r,\hat{\Delta}}^2, \quad m = 1, 2; \end{aligned} \tag{4.8}$$

$$\|u - \mathcal{P}_\Delta u\|_{0,r,K_2\cup K_3}^2 \le C h_\Delta^8 |\hat{u}|_{m,r,\hat{\Delta}}^2, \quad m = 1, 2, \tag{4.9}$$

$$|u - \mathcal{P}_\Delta u|_{1,r,\Delta}^2 \le C h_\Delta^6 |\hat{u}|_{m,r,\hat{\Delta}}^2, \quad m = 1, 2. \tag{4.10}$$

Set h_Δ small enough such that $h_\Delta < \sin(2h_\Delta)$; by detailed analysis similar to (4.7) and (4.8), we have

$$|\hat{u}|_{m,r,\hat{\Delta}}^2 \le C h_\Delta^{2m-8} |u|_{m,r,\Delta}^2, \quad m = 1, 2, \tag{4.11}$$

where C depends only on R and σ. Combining (4.7)–(4.11) yields (4.5).

(2) $K_1, \ldots, K_4$ are inner elements where the weights are strictly positive:

$$
\begin{aligned}
K_1 &= [r_{10}, r_{10}+h_1] \times [r_{20}, r_{20}+h_2] \times [\theta_0, \theta_0+h_3],\\
K_2 &= [r_{10}, r_{10}+h_1] \times [r_{20}+h_2, r_{20}+h_2+h_2^{K_2}] \times [\theta_0, \theta_0+h_3],\\
K_3 &= [r_{10}, r_{10}+h_1] \times [r_{20}, r_{20}+h_2] \times [\theta_0+h_3, \theta_0+h_3+h_3^{K_3}],\\
K_4 &= [r_{10}, r_{10}+h_1] \times [r_{20}+h_2, r_{20}+h_2+h_2^{K_2}] \times [\theta_0+h_3, \theta_0+h_3+h_3^{K_3}],
\end{aligned}
$$

where $r_{10}, r_{20}, \theta_0 \geq \sigma h_\Delta$. The reference macro element and reference elements are defined as

$$
\hat{\Delta} = [1,2] \times [1,3] \times [1,3], \quad \hat{K}_1 = [1,2] \times [1,2] \times [1,2],
$$
$$
\hat{K}_2 = [1,2] \times [2,3] \times [1,2], \quad \hat{K}_3 = [1,2] \times [1,2] \times [2,3], \quad \hat{K}_4 = [1,2] \times [2,3] \times [2,3].
$$

Then for $m = 1, 2$, by affine transforms, there exists a generic constant C independent of h, such that

$$
\begin{aligned}
\|u - \mathcal{P}_\Delta u\|^2_{0,r,K_1} &\leq C h_1 h_2 h_3 \Lambda(\theta_0, h_3)(r_{10}+h_1)^2 (r_{20}+h_2)^2 \|\hat{u} - \mathcal{P}_{\hat{\Delta}} \hat{u}\|^2_{0,\hat{K}_1}\\
&\leq C h_1 h_2 h_3 \Lambda(\theta_0, h_3)(r_{10}+h_1)^2 (r_{20}+h_2)^2 \|\hat{u} - \mathcal{P}_{\hat{\Delta}} \hat{u}\|^2_{0,r,\hat{K}_1}\\
&\leq C h_\Delta^3 (r_{10}+h_\Delta)^2 (r_{20}+h_\Delta)^2 \Lambda(\theta_0, h_3) |\hat{u}|^2_{m,r,\hat{\Delta}}.
\end{aligned} \tag{4.12}
$$

Similarly, for $i = 2, 3, 4$ we have

$$
\begin{aligned}
\|u - \mathcal{P}_\Delta u\|^2_{0,r,K_i} &\leq C h_\Delta^3 (r_{10}+h_\Delta)^2 (r_{20}+h_\Delta)^2\\
&\quad \times [\Lambda(\theta_0, h_3) + \Lambda(\theta_0+h_3, h_3+h_3^{K_3})] |\hat{u}|^2_{m,r,\hat{\Delta}},
\end{aligned} \tag{4.13}
$$

$$
\begin{aligned}
|u - \mathcal{P}_\Delta u|^2_{1,r,\Delta} &\leq C h_\Delta (r_{10}+h_\Delta)^2 (r_{20}+h_\Delta)^2\\
&\quad \times [\Lambda(\theta_0, h_3) + \Lambda(\theta_0+h_3, h_3+h_3^{K_3})] |\hat{u}|^2_{m,r,\hat{\Delta}}.
\end{aligned} \tag{4.14}
$$

Set h_Δ small enough such that $h_\Delta < \sin(2h_\Delta)$; by affine transforms and detailed analysis similar to (4.13), we have

$$
\begin{aligned}
|\hat{u}|^2_{m,r,\hat{\Delta}} &\leq C \max\left\{1/\sin\theta_0, 1/\sin(\theta_0+h_3), 1/\sin(\theta_0+h_3+h_3^{K_3})\right\}\\
&\quad \times r_{10}^{-2} (r_{20}+h_2)^{-2} h_\Delta^{2m-3} |u|^2_{m,r,\Delta}.
\end{aligned} \tag{4.15}
$$

Combining (4.12)–(4.15), we obtain (4.5) by Lemma 4.1.

We can prove (4.5) for other macro elements similarly. □

5. Finite element approximations. The equivalent weak form of (3.6) is the following: Find $(\lambda, u) \in R^1 \times H^1_{0r}(\Omega)$ and $u \neq 0$ such that

$$
a_r(u, v) = \lambda\,(u, v)_0 \quad \forall v \in H^1_{0r}(\Omega), \tag{5.1}
$$

where $\lambda = K + E$, K is the constant in (2.10), and

$$
\begin{aligned}
a_r(u, v) = \int_\Omega \Bigg[& r_1^2 r_2^2 \left(\frac{\partial u}{\partial r_1}\frac{\partial v}{\partial r_1} + \frac{\partial u}{\partial r_2}\frac{\partial v}{\partial r_2} \right) + (r_1^2 + r_2^2) \frac{\partial u}{\partial \theta}\frac{\partial v}{\partial \theta}\\
& + \left(\frac{2 r_1^2 r_2^2}{\sqrt{r_1^2 + r_2^2 - 2 r_1 r_2 \cos\theta}} - 4 r_1^2 r_2 - 4 r_1 r_2^2 \right) uv \Bigg] \sin\theta\, dr_1 dr_2 d\theta.
\end{aligned}
$$

Since $a(\cdot,\cdot)$ is continuous and coercive on $H_0^1(\hat{\Omega})$, where $\hat{\Omega}$ is defined as that in (3.9), we know that $a_r(\cdot,\cdot)$ is continuous and coercive on $H_{0r}^1(\Omega)$ by the proof of Lemma 3.1. We define $\|\cdot\|_{1,r,\Omega} = \sqrt{a_r(\cdot,\cdot)}$ for the sake of simplicity in notation.

We consider the Lagrangian finite element approximation to (5.1). For any $K \in \mathcal{T}_h$, denote the set of nodes in K as

$$\mathcal{V}(K) = \{8 \text{ vertices and } (k+1)^3 - 8 \ k\text{-section points of } K\}.$$

$\cup_{K\in\mathcal{T}_h}\mathcal{V}(K)$ is the set of nodes of $\mathcal{T}_h$. Define the finite element space as

$$V_h = \{v(r_1, r_2, \theta) \in C^0(\Omega) \mid v|_{\partial\Omega} = 0,\ v|_K \in Q_k(K)\ \forall K \in \Im_h\}.$$

The discrete approximation of (5.1) is the following: Find $(\lambda_h, u_h) \in R^1 \times V_h$ and $u_h \neq 0$ such that

$$a_r(u_h, v_h) = \lambda_h\,(u_h, v_h)_0 \quad \forall v_h \in V_h. \tag{5.2}$$

Obviously, $V_h \subset C^0(\Omega) \cap H_{0r}^1(\Omega)$, and so (5.2) is the Galerkin approximation of (5.1). We drop the subscript "B" in (2.13) (or (5.1)) and suppose that $0 < \lambda_1 \leq \lambda_2 \leq \cdots$ are the eigenvalues of (5.1), $0 < \lambda_{h1} \leq \lambda_{h2} \leq \cdots < \lambda_{N_h}$ are the eigenvalues of (5.2), $N_h = \dim(V_h)$. Denote the eigenspaces associated with λ_i and λ_{hj} as V_i and V_{hj}, respectively, $1 \leq i \leq \cdots$, $1 \leq j \leq N_h$. By the minmax theorem [10], $\lambda_i \leq \lambda_{hi}$, $1 \leq i \leq N_h$.

Let $\hat{\Delta}$ be a reference macro element and $\hat{K} \subset \hat{\Delta}$ be a reference element. Define

$$\mathcal{V}(\hat{K}) = \{a_i^{\hat{K}} \mid 1 \leq i \leq (k+1)^3\}$$

associated with a basis $\{\hat{\varphi}_i^{\hat{K}} \mid 1 \leq i \leq (k+1)^3\}$ of $Q_k(\hat{K})$. Assume

$$\hat{\varphi}_i^{\hat{K}}(a_j^{\hat{K}}) = \delta_{ij}, \quad 1 \leq i, j \leq (k+1)^3,$$

where $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$. For any $\hat{v} \in H_r^0(\hat{K})$, we define the finite element interpolation operator by means of the local regularization as follows:

$$\pi_{\hat{K}}\hat{v} = \sum_{i=1}^{(k+1)^3} (\mathcal{P}_{\hat{\Delta}}\hat{v})(a_i^{\hat{K}})\hat{\varphi}_i^{\hat{K}}. \tag{5.3}$$

Suppose $v \in H_r^0(\Omega)$. For any $K = F_K(\hat{K}) \in \mathcal{T}_h$, define the finite element interpolation operator on $\hat{\Delta}$, K, and Ω as follows: $\pi_{\hat{\Delta}}\hat{v}|_{\hat{K}} = \pi_{\hat{K}}\hat{v}$, $\pi_K v = (\pi_{\hat{K}}\hat{v}) \circ F_K^{-1}$, $\pi v|_K = \pi_K v$, where $\hat{K} \subset \hat{\Delta}$ is some reference element. On any space of finite dimension, all norms are equivalent. There exists a constant C such that

$$\|\pi_{\hat{K}}\hat{v}\|_{m,r,\hat{K}} \leq C\|\mathcal{P}_{\hat{\Delta}}\hat{v}\|_{0,\infty,\hat{K}} \leq C\|\mathcal{P}_{\hat{\Delta}}\hat{v}\|_{0,r,\hat{K}} \leq \|\hat{u}\|_{0,r,\hat{\Delta}}. \tag{5.4}$$

Thus the operator $\pi_{\hat{K}} : H_r^0(\hat{\Delta}) \to H_r^m(\hat{K})$ is linear and continuous, $m = 0, 1$.

THEOREM 5.1. *There exists a constant C independent of h such that for any $v \in H_r^2(\Omega)$,*

$$\|v - \pi v\|_{m,r,\Omega} \leq Ch^{2-m}|v|_{2,r,\Omega}, \quad m = 0, 1. \tag{5.5}$$

Proof. Suppose that $\hat{\Delta}$ is a reference macro element and $\hat{K} \subset \hat{\Delta}$ is a reference element. By the definition of the operators $\pi_{\hat{K}}$, $\pi_{\hat{\Delta}}$, and $\mathcal{P}_{\hat{\Delta}}$, we have

$$\pi_{\hat{\Delta}} p = p \quad \forall p \in P_k(\hat{\Delta}). \tag{5.6}$$

By (5.4), (5.6), and Theorem 3.8, for any $\hat{v} \in H_r^2(\hat{\Delta})$ we have

$$\|\hat{v} - \pi_{\hat{\Delta}} \hat{v}\|_{m,r,\hat{\Delta}} \leq C|\hat{v}|_{2,r,\hat{\Delta}}, \quad m = 0, 1. \tag{5.7}$$

For any macro element Δ, by affine transforms, (5.7), and the argument in the proof of Theorem 4.2, we know that there exists a constant C independent of h such that

$$\|v - \pi v\|_{m,r,\Delta} \leq C h_{\Delta}^{2-m} |v|_{2,r,\Delta}, \quad m = 0, 1. \tag{5.8}$$

Summing up each side of (5.8) over all macro elements, we get (5.5). □

THEOREM 5.2. *For any $1 \leq i \leq N_h$, suppose (λ_i, u_i) is an eigenpair of* (5.1) *with $\|u_i\|_{0,r,\Omega} = 1$. There exist a constant C independent of h and an eigenfunction $u_{hi} \in V_{hi}$ with $\|u_{hi}\|_{0,r,\Omega} = 1$ such that*

$$|\lambda_i - \lambda_{hi}| < Ch^2, \tag{5.9}$$

$$\|u_i - u_{hi}\|_{m,r,\Omega} < Ch^{2-m}, \quad m = 0, 1. \tag{5.10}$$

Proof. By the theory of abstract spectrum approximation (p. 699 of [10]), we know that, for any $1 \leq i \leq N_h$, there exists a constant C independent of h such that

$$|\lambda_i - \lambda_{hi}| < C\varepsilon(\lambda_i)^2, \tag{5.11}$$

where

$$\varepsilon(\lambda_i) = \sup_{v \in V_i, \|v\|_{1,r,\Omega}=1} \inf_{v_h \in V_h} \|v - v_h\|_{1,r,\Omega}. \tag{5.12}$$

Let $\{u_{ij},\ 1 \leq j \leq N_i\}$ be a basis of V_i, $\|u_{ij}\|_{1,r,\Omega} = 1$, $N_i = \dim V_i$. By Theorem 5.1, we have

$$\varepsilon(\lambda_i) \leq \sum_{j=1}^{N_i} \|u_{ij} - \pi u_{ij}\|_{1,r,\Omega} \leq Ch \sum_{j=1}^{N_i} |u_{ij}|_{2,r,\Omega} \leq Ch. \tag{5.13}$$

Thus (5.9) is true.

We can prove (5.10) by Theorem 5.2 and the argument of Theorem 6.2 in [32, p. 235], but we do not give the tedious description here. □

Remark 5.3. In real computation, we have introduced a variational equation equivalent to (5.1). Define $\mu = \cos\theta$; then $\Omega = [0, R] \times [0, R] \times [-1, 1]$, and the

corresponding bilinear form, inner products, and norms are

$$(u,v)_0 = \int_\Omega uvr_1^2r_2^2dr_1dr_2d\mu, \qquad \|u\|_{0,r,\Omega}^2 = \int_\Omega u^2r_1^2r_2^2dr_1dr_2d\mu,$$

$$(u,v)_1 = (u,v)_0 + \int_\Omega \left[r_1^2r_2^2\frac{\partial u}{\partial r_1}\frac{\partial v}{\partial r_1} + r_1^2r_2^2\frac{\partial u}{\partial r_2}\frac{\partial v}{\partial r_2} + (r_1^2+r_2^2)(1-\mu^2)\frac{\partial u}{\partial \mu}\frac{\partial v}{\partial \mu}\right] dr_1dr_2d\mu,$$

$$|u|_{1,r,\Omega}^2 = \int_\Omega \left[r_1^2r_2^2\left(\frac{\partial u}{\partial r_1}\right)^2 + r_1^2r_2^2\left(\frac{\partial u}{\partial r_2}\right)^2 + (r_1^2+r_2^2)(1-\mu^2)\left(\frac{\partial u}{\partial \mu}\right)^2\right] dr_1dr_2d\mu,$$

$$\|u\|_{1,r,\Omega}^2 = \|u\|_{0,r,\Omega}^2 + |u|_{1,r,\Omega}^2, \qquad H_r^i(\Omega) = \{v|\ \ \|v\|_{i,r,\Omega}^2 < +\infty\}, \quad i = 0,1,$$

$$a_r(u,v) = \int_\Omega \left[r_1^2r_2^2\left(\frac{\partial u}{\partial r_1}\frac{\partial v}{\partial r_1} + \frac{\partial u}{\partial r_2}\frac{\partial v}{\partial r_2}\right) + (r_1^2+r_2^2)(1-\mu^2)\frac{\partial u}{\partial \mu}\frac{\partial v}{\partial \mu} + \left(\frac{2r_1^2r_2^2}{\sqrt{r_1^2+r_2^2-2r_1r_2\mu}} - 4r_1^2r_2 - 4r_1r_2^2\right)uv\right] dr_1dr_2d\mu.$$

We define $H_{0r}^1(\Omega)$ as in Remark 3.7. A variational equation equivalent to (5.1) is the following: Find $(\lambda, u) \in R^1 \times H_{0r}^1(\Omega)$ and $u \neq 0$ such that

$$a_r(u,v) = \lambda\,(u,v)_0 \qquad \forall v \in H_{0r}^1(\Omega). \tag{5.14}$$

The partitions $\mathcal{T}_h$ for $\Omega = [0,R] \times [0,R] \times [-1,1]$ are similar to those in section 4. $\forall K \in \mathcal{T}_h$ define

$$Q_k(K) = \left\{ p \,\middle|\ p = \sum_{l,m,n=0}^{k} \alpha_{lmn} r_1^l r_2^m \mu^n,\ (r_1,r_2,\mu) \in K \right\}.$$

The finite element approximation to (5.14) is: Find $(\lambda_h, u_h) \in R^1 \times V_h$ and $u_h \neq 0$ such that

$$a_r(u_h,v_h) = \lambda_h\,(u_h,v_h)_0 \qquad \forall v \in V_h, \tag{5.15}$$

where

$$V_h = \{v(r_1,r_2,\mu) \in C(\Omega)|\ v|_K \in Q_k(K)\ \forall K \in \mathcal{T}_h;\ v|_{\partial\Omega} = 0\}$$

is the finite element space.

Comparing (5.15) with (5.2) in real computation, we have found that (5.15) gives more precise results with the same number of unknowns. The analysis for (5.15) will be the subject of our future research.

6. Numerical results. Since V_h is a finite-dimensional space, define $N = \dim(V_h)$. We can choose a basis $\{\Phi_1, \ldots, \Phi_N\}$ of V_h such that

$$\text{supp}\Phi_i = \cup_{K \in \Im_h, a_i \in \Sigma_K} K,$$

where a_i is a node of $\mathcal{T}_h$. Let $u_h = \sum_{i=1}^N \alpha_i\Phi_i$ and $v_h = \Phi_i$, $1 \le i \le N$, in (5.2) and (5.15). Then we obtain an equivalent generalized eigenvalue problem:

$$A\,X = \lambda_h\,M\,X, \tag{6.1}$$

where $X = (\alpha_1, \ldots, \alpha_N)^T$, $A = (a(\Phi_i, \Phi_j))_{N\times N}$, $M = ((\Phi_i, \Phi_j))_{N\times N}$.

We use the inverse iteration method [5] to solve the generalized eigenvalue problem (6.1). This method is convenient for computing the smallest (real) eigenvalue of an (unsymmetric) generalized eigenvalue problem with large and sparse matrices. In each step of iteration, the main computational cost is the solution of the following system of equations for Y:

$$AY = F. \tag{6.2}$$

However, in fact we need only solve (6.2) in the first step if using LU-factorization of A, since we can store the inverse matrix of A for all following steps.

The computational cost of (6.2) is of order $O(N^3)$ for a dense matrix. Since the finite element matrices are banded, and their band widths are bounded by some positive integer $M \ll N$, the cost of (6.2) is not more than $2MN^2$. Thus the first iteration of our eigenvalue solver needs $O(N^2)$ floating point operations, but each of the following iterations needs only $O(N)$ floating point operations. For the solution of large sparse generalized eigenvalue problems, improvements of this method have been developed rapidly. They devote themselves to reducing the cost of the first iteration; i.e., they solve (6.2) by efficient iterative methods instead of LU-factorization. Each iterative step of their eigenvalue solvers (such as *preconditioned inverse iteration* [26]) needs only $O(N)$ operations. For more detailed analyses, we refer to Neymeyr's excellent work [26], or to the journal articles [27], [28], [29] and references therein. We consider the improvement of our eigenvalue solver as future work.

We carried out our computation on a personal computer: Intel PIII750 with 1G SDRAM. The experiment shows that 1. the energy errors decrease with R or the number of nodes increasing; 2. with the considered state becoming more highly exited, R should be larger, and more nodes far from the nucleus are needed; 3. very large R makes no remarkable improvement in the precision.

The main error concerns the potential $V = -\frac{2}{r_1} - \frac{2}{r_2} + \frac{1}{r_{12}}$. For the triplet, the wave function is antisymmetric with respect to the two electrons, so they cannot be very close to each other. That is to say, when r_{12} is very small, the wave function u tends to zero. When we calculate $\int_\Omega \frac{u^2}{r_{12}} dr_1 dr_2 d\mu$ with a Gaussian integration formula [33], the error for the triplet is much smaller than that for the singlet with the same number of Gaussian points. Furthermore, from the figures below, we can see that the wave function $|u|$ of the ground state is much larger than that of excited states in the domain where r_{12} is small and in the neighborhood of the nucleus containing the singularities. Thus we have used more and more Gaussian points and grid points along the μ-direction, when the state varies from the triplet, the singlet to the ground state.

All matrix elements are computed by the standard Gaussian integration formula. With the number of Gaussian points increasing, the computing time becomes longer. Let N_e be the number of elements associated with some partition of Ω, N_g be the number of Gaussian points along one direction, and T_e be the CPU time to compute a pair of element matrices by the one-point Gaussian formula. The CPU time to obtain the global stiffness matrix and mass matrix is

$$T \approx N_e \times T_e \times N_g^3. \tag{6.3}$$

The number of degrees of freedom (DOF), number of Gaussian points (GPs), and the computational time T are listed in Table 6.1.

TABLE 6.1
Computational efforts for the S-states.

State	$1s1s\ ^1S$	$1s2s\ ^1S$	$1s2s\ ^3S$
Number of DOF	57472	65320	68243
Number of GPs	$27 \times 27 \times 27$	$21 \times 21 \times 21$	$9 \times 9 \times 9$
CPU time T (hours)	44.3	20.6	1.62

TABLE 6.2
FEM results for the helium atom (all values in a.u.).

State	$1s1s\ ^1S$	$1s2s\ ^1S$	$1s2s\ ^3S$
Highly precise results [12], [23]	$-2.903724377034119\ldots$	$-2.1459740460544\ldots$	$-2.1752293782367\ldots$
Results given by (5.15)	-2.903724106	-2.1459740042	-2.1752293277
Results given by (5.2)	-2.903715597	-2.1459703835	-2.1752288326
FEM [6]	-2.9036118	-2.145960	-2.1752214
FEM [25]	-2.90326		
FEM [31]	-2.90324		

We place grid points symmetrically along r_1 and r_2 for all states. The grid points are (for r_1, r_2, μ)

1. $1s1s\ ^1S$:
   - 0.0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0, 1.2, 1.6, 2.0, 2.6, 3.2, 4.2, 6.0, 9.0, 15.0;
   - -1.0, -0.6, -0.2, 0.2, 0.6, 1.0;
2. $1s2s\ ^1S$:
   - 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 3.0, 3.4, 3.8, 4.2, 4.8, 5.6, 8.0, 11.5, 15.0, 20.0;
   - -1.0, -0.5, 0.5, 1.0;
3. $1s2s\ ^3S$:
   - 0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2, 3.4, 3.6, 4.0, 4.5, 5.5, 7.0, 10.0, 13.0, 16.0, 20.0, 25.0;
   - -1.0, 0.0, 1.0.

The relative error of our approximate eigenvalue for the ground ($1s1s$-) state is 10^{-7}a.u., and those for the $1s2s$-states are 10^{-8}a.u. by (5.15). The precisions of existing finite element eigenvalues are generally $10^{-4} - 10^{-6}$a.u. (see Table 6.2).

From the graphs of approximate wave functions (see Figures 6.1–6.2), we can get the following properties. 1. Although we add no physical assumptions to our computations a priori, such as the symmetry (for the singlets) and the antisymmetry (for the triplet), our approximate wave functions coincide with these properties very well. 2. Wave functions oscillate heavily in the neighborhood of the nucleus where the singularity of the Coulomb potential is very strange. This is well known by physicists and chemists. 3. In a sufficiently small neighborhood of the nucleus, absolute values $|u_h|$ of wave functions are very small. This implies that electrons seldom visit there. 4. With the distance between each electron and the nucleus increasing, wave functions decrease quickly. Thus it is reasonable to solve the Schrödinger equation in bounded domains.

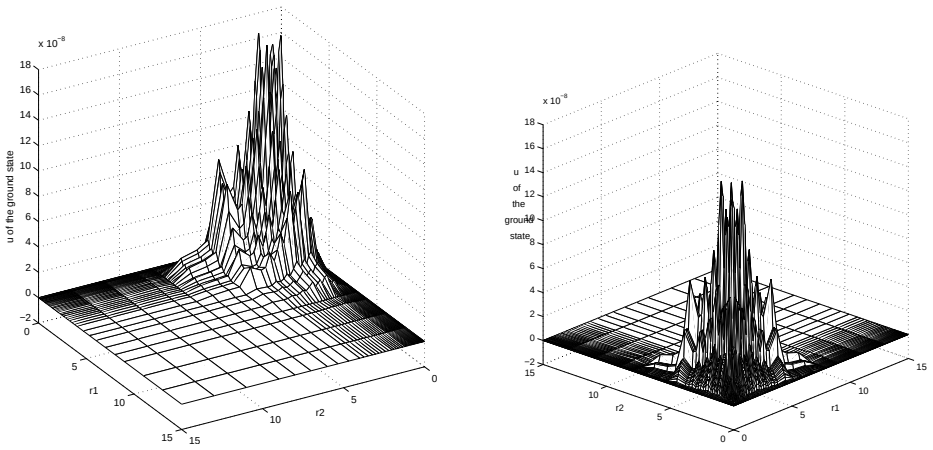


FIG. 6.1. *Wave function of the* $1s1s\ {}^1S$*-state.*

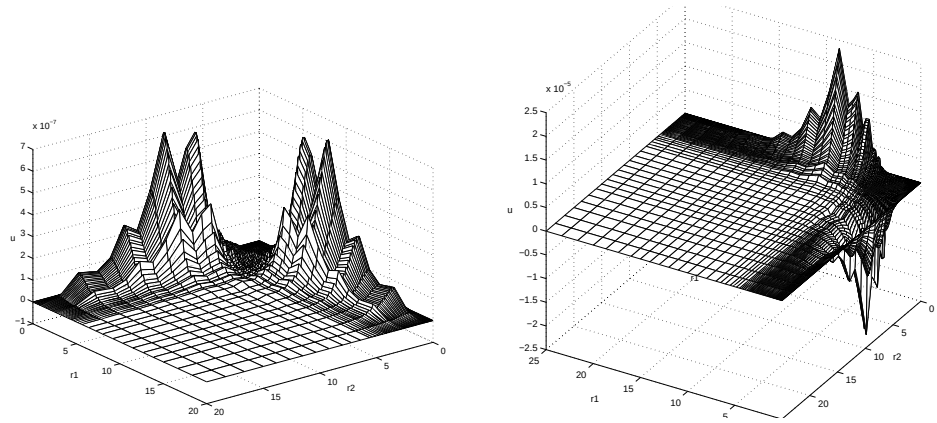


FIG. 6.2. *Wave functions of the* $1s2s\ {}^1S$*-state (left) and the* $1s2s\ {}^3S$*-state (right).*

Acknowledgment. The authors would like to thank professor Peizhu Ding of Jilin University for his valuable suggestions and discussions.

REFERENCES

[1] J. ACKERMANN, *Finite-element expectation values for correlated two-electron wave functions*, Phys. Rev. A, 52 (1995), pp. 1968–1975.
[2] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1978.
[3] A. ASKAR, *Finite element method for bound state calculations in quantum mechanics*, J. Chem. Phys., 62 (1975), pp. 732–734.
[4] A. ASKAR AND A. S. CAKMAK, *Finite element methods for reactive scattering*, Chem. Phys., 33 (1978), pp. 267–286.
[5] K. BATHE AND E. WILSON, *Numerical Methods in Finite Element Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1973.
[6] M. BRAUN, W. SCHWEIZER, AND H. HEROLD, *Finite-element calculations for the S states of helium*, Phys. Rev. A, 48 (1993), pp. 1916–1920.
[7] G. BREIT, *Separation of angles in the two-electron problem*, Phys. Rev., 35 (1930), pp. 569–578.

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, Berlin, 1998.
[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, New York, Oxford, 1978.
[10] P. G. Ciarlet and J. L. Lions, *Handbook of Numerical Analysis* II, in Finite Element Methods, North–Holland, Amsterdam, 1989.
[11] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 9 (1975), pp. 77–84.
[12] G. W. F. Drake, and Z.-C. Yan, *Variational eigenvalues for the S states of helium*, Chem. Phys. Letters, 229 (1994), pp. 486–490.
[13] G. W. F. Drake, M. M. Cassar, and R. A. Nistor, *Ground-state energies for helium, H^-, and Ps^-*, Phys. Rev. A, 65 (2002), paper 054501.
[14] M. Duff, H. Rabitz, A. Askar, A. Cakmak, and M. Ablowitz, *A comparison between finite element methods and spectral methods as applied to bound state problems*, J. Chem. Phys., 73 (1980), pp. 1543–1559.
[15] Ch. Froese–Fischer, *The Hartree–Fock Method for Atoms*, Wiley-Interscience, New York, 1977.
[16] M. Fridman, Y. Rosenfeld, A. Rabinovitch, and R. Thieberger, *Finite element method for solving the two-dimensional Schrödinger equation*, J. Comput. Phys., 26 (1978), pp. 169–180.
[17] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1983.
[18] M. I. Haftel and V. B. Mandelzweig, *Precise nonvariation calculations on the helium atom*, Phys. Rev. A, 38 (1988), pp. 5995–5999.
[19] I. L. Hawk and D. L. Hardcastle, *Finite-difference solution to the Schrödinger equation for the helium isoelectronic sequence*, Comput. Phys. Commun., 16 (1979), pp. 159–166.
[20] H. Kliendienst, A. Lüchow, and H.-P.Merckens, *Accurate upper and lower bounds for some excited S states of the He atom*, Chem. Phys. Lett., 218 (1994), pp. 441–444.
[21] A. Kono and S. Hattori, *Variational calculations for excited states in He* I: *Improved estimation of the ionization energy from accurate energies for the $n\,^3S$, $n\,^1D$, $n\,^3D$ series*, Phys. Rev. A, 31 (1985), pp. 1199–1202.
[22] A. Kono and S. Hattori, *Energy levels for S, P, D states in He through precision variational calculations*, Phys. Rev. A, 34 (1986), 1727–1735.
[23] V. I. Korobov, *Nonrelativistic ionization energy for the helium atom*, Phys. Rev. A, 66 (2002), paper 024501.
[24] R. Krivec, M. I. Haftel, and V. B. Mandelzweig, *Precise nonvariation calculation of excited states of helium with the correlation-function hyperspherical-harmonic method*, Phys. Rev. A, 44 (1991), pp. 7158–7164.
[25] F. S. Levin and J. Shertzer, *Finite-element solution of the Schrödinger equation for the helium ground state*, Phys. Rev. A, 32 (1985), pp. 3285–3290.
[26] K. Neymeyr, *A Hierarchy of Preconditioned Eigensolvers for Elliptic Differential Operators*, research report, Mathematisches Institüt, Universität Tübingen, Tübingen, Germany, 2001; available online http://na.uni-tuebingen.de/klaus/papers.shtml.
[27] K. Neymeyr, *A geometric theory for preconditioned inverse iteration,* I: *Extrema of the Rayleigh quotient*, Linear Algebra Appl., 332 (2001), pp. 61–85.
[28] K. Neymeyr, *A geometric theory for preconditioned inverse iteration,* II: *Convergence estimates*, Linear Algebra Appl., 332 (2001), pp. 87–104.
[29] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration,* III: *A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.
[30] S. Nordholm and G. Bacsky, *Generalized finite element method applied to bound state calculation*, Chem. Phys. Lett., 42 (1976), pp. 259–263.
[31] A. Scrinzi, *A 3-dimensional finite elements procedure for quantum mechanical applications*, Comput. Phys. Commun., 86 (1995), pp. 67–80.
[32] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.
[33] A. H. Stroud, *Approximate Calculation of Multipole Integrals*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
[34] J. Weidmann, *Linear Operators in Hilbert Spaces*, Springer-Verlag, New York, 1980.
[35] N. W. Winter, A. Laferriere, and V. McKoy, *Numerical solution of the two-electron Schrödinger equation*, Phys. Rev. A, 2 (1970), pp. 49–59.
[36] O. C. Zienkiewwicz and R. L. Taylor, *The Finite Element Method*, 4th ed., Vol. 1, McGraw-Hill, London, 1989.

[37] W. Zheng and L. Ying, *Finite element calculation for helium atom*, Internat. J. Quantum Chem., (2004), to appear.
[38] W. Zheng, *The Finite Element Method for Atomic and Molecular Problems*, Ph. D. thesis, Peking University, Beijing, P. R. China, 2002.
[39] W. Zheng, Lung-An Ying, and Peizhu Ding, *Numerical solutions of the Schrödinger equation for the ground lithium by the finite element method*, Appl. Math. Comput., to appear.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 75–89

TIME-SPACE DISCRETIZATION OF THE NONLINEAR HYPERBOLIC SYSTEM $u_{tt} = \operatorname{div}(\sigma(Du) + Du_t)$*

CARSTEN CARSTENSEN† AND GEORG DOLZMANN‡

Abstract. The numerical treatment of the hyperbolic system of nonlinear wave equations with linear viscosity, $u_{tt} = \operatorname{div}(\sigma(Du) + Du_t)$, is studied for a large class of globally Lipschitz continuous functions σ, including nonmonotone stress-strain relations. The analyzed method combines an implicit Euler scheme in time with Courant (continuous and piecewise affine) finite elements in space for a class of varying time steps with varying meshes. Explicit a priori error bounds in $L^\infty(L^2)$, $L^2(W^{1,2})$, and $W^{1,2}(L^2)$ are established for the solutions of the fully discrete scheme.

Key words. finite elements, a priori error estimates, nonlinear wave equations

AMS subject classifications. 65N12, 65N15, 35G25, 73G25

DOI. 10.1137/S0036142901393413

1. Introduction.

In this paper we study the numerical treatment of the nonlinear hyperbolic system

$$u_{tt} = \operatorname{div}(\sigma(Du) + Du_t) \text{ in } \Omega \times (0,T) \tag{1.1}$$

subject to the boundary and initial conditions

$$\begin{aligned} u &= 0 \quad \text{on } \partial\Omega\times(0,T),\\ u &= u_0 \text{ in } \quad \Omega\times\{0\},\\ u_t &= v_0 \text{ in } \quad \Omega\times\{0\}. \end{aligned}$$

Here, u is a vector-valued mapping from $\Omega \subset \mathbb{R}^n$ into $\mathbb{R}^m$, and the initial data satisfy $u_0 \in W_0^{1,2}(\Omega;\mathbb{R}^m)$ and $v_0 \in L^2(\Omega;\mathbb{R}^m)$.

The physical interest in this equation lies in the fact that it describes for $m = n$ the evolution of a viscoelastic body with reference configuration Ω. Nonmonotone stress-strain relations, modeled by $\sigma = D\Phi$ for nonconvex energy density functions Φ, are of main interest in simulations of solid-solid phase transitions; see, e.g., [1, 9, 10, 4, 5]. This equation arises also in the two-dimensional scalar case ($n = 2$ and $m = 1$) for the out-of-plane displacement field of an antiplane shear deformation [11]. Numerical experiments in [7] employed a discontinuous stabilized approximation because of difficulties with Q_1 finite elements. Our analysis shows that a P_1 finite element discretization leads to a convergent scheme provided that the solution is sufficiently regular.

Inspired by the uniqueness proof for Lipschitz continuous stresses σ in [10, 6], we present in this paper the a priori error analysis for an approximating scheme for the

*Received by the editors August 6, 2001; accepted for publication (in revised form) June 13, 2003; published electronically January 6, 2004.
http://www.siam.org/journals/sinum/42-1/39341.html

†Institute for Applied Mathematics and Numerical Analysis, Vienna University of Technology, Wiedner Hauptstraße 8-10 A-1040 Vienna, Austria (Carsten.Carstensen@tuwien.ac.at). The research of this author was partially supported by the Powell Foundation at the California Institute of Technology and by the Max Planck Society.

‡Department of Mathematics, University of Maryland, College Park, MD 20742-4015 (dolzmann@math.umd.edu). The research of this author was partially supported by the California Institute of Technology through grants AFOSR/MURI (F 49602-98-1-0433), by the Max Planck Society, and by the NSF through grant DMS0104118.

system (1.1) that combines continuous finite elements in spaces with a discontinuous Galerkin approximation in time. We obtain estimates for the approximation error of the deformation u, the deformation gradient Du, and the velocity field $v = u_t$ under very general assumptions. In particular, the time step size k_j has only to satisfy the condition $k_j \le Qk_{j+1}$ for a global constant $Q > 0$, and the spatial triangulations are only assumed to be quasiuniform at each time step with a typical diameter h_j of the elements. The spatial L^2-approximation error is balanced with the time step error through an interesting coupling of k_j and h_j. The quotient h_j^2/k_j should not become too large; i.e., the time steps should not be too small compared to the square of the spatial discretization parameter h_j. Our general result, Theorem 4.1, immediately implies the following convergence estimates (see section 2 for the precise definitions).

THEOREM 1.1. *Let Ω be a bounded polygonal domain in $\mathbb{R}^n$. Fix $T > 0$ and define discrete times $0 = t_0 < t_1 < \cdots < t_N = T$. Suppose that $\mathcal{T}_j$ is a regular triangulation of Ω for $j = 0, \dots, N$ such that $\mathcal{T}_j$ is a refinement of $\mathcal{T}_{j-1}$. Assume that $\mathcal{S}_{0,j}(\mathcal{T}_j)$ is the space of all continuous functions that vanish on $\partial\Omega$ and are affine on the elements in $\mathcal{T}_j$. Let $U_j \in \mathcal{S}_{0,j}$ be the solution of the implicit Euler scheme defined in section 2.3 and let*

$$k_j = t_j - t_{j-1}, \quad k = \max_{j=1,\dots,N} k_j, \quad Q = \max_{j=2,\dots,N} \frac{k_j}{k_{j-1}}, \quad h = \max_{j=0,\dots,N} h_j.$$

Assume, furthermore, that σ is globally Lipschitz continuous and that the solution u of the system (1.1) belongs to $W^{1,\infty}(W^{2,2}) \cap W^{2,2}(W^{1,2})$. Finally, define the discretization errors e_j and δ_j at the time step t_j by

$$e_j = u(t_j) - U_j, \quad \delta_j = v(t_j) - \frac{1}{k_j}\big(U_j - U_{j-1}\big).$$

Then there exist constants c_1 and c_2 such that the following holds: If $c_1 k < 1$, then

$$\max_{\nu=1,\dots,N} \|e_\nu\|^2 + \sum_{\nu=1}^{N} k_\nu \big(\|\delta_\nu\|^2 + \|De_\nu\|^2\big) \le c_2(T + T^2 + h^4 + h^8) \exp\big(c_1 T\big) \|u\|_{h,k}^2,$$

where

$$\|u\|_{h,k}^2 = k^2 \| \, |u_{tt}| + |Du_{tt}| + |Du_t| \, \|_{L^2(L^2)}^2 + h^2 \left(1 + \max_{j=1,\dots,N} \frac{h_j^2}{k_j}\right) \|u\|_{W^{1,\infty}(W^{2,2})}^2.$$

The constants c_1 and c_2 depend only on the Lipschitz constant of σ, on the shape of the triangles in the triangulations $\mathcal{T}_j$, and on k_j via Q, but neither on h_j nor on u.

Remarks. (1) The statement of the theorem assumes tacitly that the initial data can be approximated sufficiently well; see estimate (2.7) below.

(2) The assumption $c_1 k \le 1$ implies that $2k_j \operatorname{Lip}(\sigma) < 1$ for $j = 1, \dots, N$. This condition ensures that the discrete scheme has a unique solution; see Theorem 2.2.

(3) The constants c_1 and c_2 depend on $\operatorname{Lip}(\sigma)$. The geometry of the mesh enters via the quotient of the diameters of the largest ball contained in an element and the smallest ball that contains the element. The definition of Q shows that the constants do not depend on k_j if the time steps are fixed or decreasing in j. The condition that time step k_{j+1} should not be much bigger than the previous time step k_j is related to a discrete integration by parts formula in Proposition 3.5.

(4) The regularity assumptions in Theorem 4.1 (and thus in Theorem 1.1) are stronger than the regularity properties guaranteed by the known existence results in

Theorem 2.1. However, the class of equations covered by our general convergence analysis includes, for example, the (smooth) solutions of the system

$$u_{tt} = \Delta u + \Delta u_t \tag{1.2}$$

with smooth and consistent initial values. The regularity of solutions of (1.2) is intermediate between that of the heat equation and the wave equation. This can be seen by using the transformation $v(t) = \exp(t)u(t)$ in (1.2), which, after integration in time, leads to the equation

$$v_t = \Delta v + F(v) \tag{1.3}$$

for v and $F(v)(t) = f + 2v(t) - \int_0^t v(\tau)\mathrm{d}\tau$. The regularity of solutions to the perturbed heat equation (1.3) follows from a successive application of the regularity theory for the heat equation.

(5) The existence result in Theorem 2.1 is obtained by analyzing a time-discrete problem and the convergence of its solutions. A second natural technique is the Galerkin method based on a spatial discretization and the solution of a family of ordinary differential equations in time. It is an open question whether the Galerkin method converges of not. This suggests that a (mild) constraint on the discretization parameters in space and time might be necessary to guarantee convergence.

(6) In general, solutions of (1.1) may fail to be smooth, and hence it remains an open problem whether or not the proposed numerical scheme converges under the regularity of the solutions guaranteed by Theorem 2.1. The coupling of the discretization parameters discussed in remark (5) is one key point which requires higher regularity.

The paper is organized as follows. We define the discrete scheme in section 2 and prove existence and uniqueness of the discrete solution. Section 3 contains a series of estimates for the solutions of the approximating scheme which are used in section 4 to prove the general convergence result in Theorem 4.1 which contains Theorem 1.1 as a special case.

2. The discrete scheme. In this section, we introduce the relevant notation and define the discrete scheme. Then we prove existence and uniqueness of the discrete solutions and derive an identity in the spirit of Galerkin orthogonality. This relation replaces in our convergence analysis the identity

$$(u - \overline{u})_{tt} = \operatorname{div}\big(\sigma(Du) - \sigma(D\overline{u}) + D(u - \overline{u})_t\big)$$

for the difference of two solutions u and $\overline{u}$ of the system (1.1) and from which one easily deduces uniqueness of solutions for Lipschitz continuous σ; see section 2.2.

2.1. Notation. We assume that $\Omega \subset \mathbb{R}^n$ is a polygonal domain with boundary $\Gamma = \partial\Omega$ and exterior normal ν to Γ. We use the standard notation for the Lebesgue spaces $L^p(\Omega;\mathbb{R}^m)$ with norm $\|\cdot\|_p$, and we write $(\cdot,\cdot)$ for the inner product in L^2. The Sobolev spaces $W^{k,p}(\Omega;\mathbb{R}^m)$ are equipped with the standard norm $\|\cdot\|_{k,p}$ and the seminorm $|\cdot|_{k,p}$, respectively. We frequently abbreviate $X(0,T;Y(\Omega))$ by $X(Y)$ if the corresponding domain Ω, the time interval $(0,T)$, and the range of the functions is clear from the context. Thus $L^2(L^2)$ denotes, for example, both $L^2(0,T;L^2(\Omega))$ and $L^2(0,T;L^2(\Omega;\mathbb{R}^m))$. The space of all real $m \times n$ matrices $\mathbb{M}^{m\times n}$ is equipped with the Frobenius norm, $|A|^2 = \operatorname{tr}(A^T A)$, where A^T denotes the transpose of the matrix A, and with the inner product $F : G$ that is induced by the scalar product in $\mathbb{R}^{mn}$.

Let $0 = t_0 < t_1 < \cdots < t_N = T$ be a partition of the time interval $[\,0,T\,]$ into N subintervals $I_j = (t_{j-1}, t_j)$ of length $k_j = t_j - t_{j-1}$, $j = 1, \ldots, N$. Suppose that $(\mathcal{T}_j)_{j=0,\ldots,N}$ is a family of regular triangulations in the sense of [2] with maximal mesh-size h_j and that the union of all elements in $\mathcal{T}_j$ is equal to $\overline{\Omega}$. We denote by $\mathcal{S}_{0,j} = \mathcal{S}_{0,j}(\mathcal{T}_j)$ the finite element space of continuous functions $u_h : \Omega \to \mathbb{R}^m$ that have zero boundary values on Γ and are affine on the elements in $\mathcal{T}_j$. We use the interpolation operator Π_j onto $\mathcal{S}_{0,j}$, due to Scott and Zhang [13], which satisfies the projection property

$$\Pi_j v = v \quad \text{for all } v \in \mathcal{S}_{0,j}, \tag{2.1}$$

the stability estimate

$$\|D\Pi_j u\| \le c_S \|Du\| \quad \text{for all } u \in W_0^{1,2}(\Omega), \tag{2.2}$$

and the approximation estimate

$$\|u - \Pi_j u\| + h_j \|Du - D\Pi_j u\| \le c_A h_j^2 \|D^2 u\| \quad \text{for all } u \in W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega). \tag{2.3}$$

Throughout the paper, u denotes the unique solution of the system (1.1) guaranteed by Theorem 2.1 below. We define the discrete solution $U_j \in \mathcal{S}_{0,j}$ at the time step t_j in section 2.3, and we use $V_j = (U_j - U_{j-1})/k_j$ for $j = 1, \ldots, N$ as an approximation for the discrete velocities.

The goal of our analysis is to estimate the errors in u and in u_t. To simplify the notation, we set $v = u_t$, $u_j = u(t_j)$, $v_j = v(t_j)$, and

$$e_j = u_j - U_j = u(t_j) - U_j, \quad \delta_j = v_j - V_j = v(t_j) - V_j \quad \text{for } j = 1, \ldots, N.$$

2.2. Existence and uniqueness for Lipschitz continuous stress functions. Our analysis relies on the existence result [6] for the system (1.1) which requires that $\sigma(F) = \partial\Phi(F)/\partial F$, where the stored energy function Φ has the following three properties (H1), (H2), and (H3).

(H1) $\Phi \in C^2(\mathbb{M}^{m\times n})$.

(H2) There exist constants $\overline{c}$, $\overline{C} > 0$, and $p \ge 2$ such that

$$\overline{c}|F|^p - \overline{C} \le \Phi(F) \le \overline{C}\big(|F|^p + 1\big), \quad |\sigma(F)| \le \overline{C}(|F|^{p-1} + 1)$$

for all $F \in \mathbb{M}^{m\times n}$.

(H3) There exists a constant $K > 0$ such that

$$-K|F - G|^2 \le \big(\sigma(F) - \sigma(G)\big) : (F - G) \quad \text{for all } F, G \in \mathbb{M}^{m\times n}.$$

Hypothesis (H3) follows, for example, from monotonicity or global Lipschitz continuity of σ. In the latter case one can choose $K = \mathrm{Lip}(\sigma)$. In this situation, the following existence result holds (see [4, 5] for related results).

THEOREM 2.1 (see [6, Theorem 4.1]). *Under the foregoing assumptions, the system* (1.1) *has a weak solution*

$$\begin{aligned} u \in\ & L^\infty(0,\infty; W_0^{1,2}(\Omega;\mathbb{R}^m)) \cap W^{1,\infty}(0,\infty; L^2(\Omega;\mathbb{R}^m)) \\ & \cap W^{1,2}_{loc}([0,\infty); W^{1,2}(\Omega;\mathbb{R}^m)) \cap W^{2,2}_{loc}([0,\infty); W^{-1,2}(\Omega;\mathbb{R}^m)); \end{aligned}$$

i.e., $u(\cdot,0)=u_0$, $u_t(\cdot,0)=v_0$, *and, for all* $\zeta\in C_0^\infty(\Omega\times(0,\infty);\mathbb{R}^m)$,

$$\int_0^\infty\int_\Omega \big((\sigma(Du)+Du_t):D\zeta-u_t\cdot\zeta_t\big)\mathrm{d}x\,\mathrm{d}t=0. \tag{2.4}$$

Moreover, u *satisfies the dissipation inequality*

$$E[u(t),u_t(t)]-E[u_0,v_0]\le-\int_0^t\int_\Omega|Du_t|^2\mathrm{d}x\,\mathrm{d}s$$

for almost every $t>0$, *where the total energy is given by*

$$E[u,v]=\int_\Omega\left(\Phi(Du)+\frac{1}{2}|v|^2\right)\mathrm{d}x\,.$$

If σ is globally Lipschitz continuous, then the following inequalities imply uniqueness of the weak solution u (see [6, 10]). Suppose that u and $\overline{u}$ are solutions with the same initial and boundary conditions. If we test the difference of the two equations

$$u_{tt}=\operatorname{div}\big(\sigma(Du)+Du_t\big),\quad \overline{u}_{tt}=\operatorname{div}\big(\sigma(D\overline{u})+D\overline{u}_t\big)$$

by $u-\overline{u}$ and integrate in space and time, then we obtain

$$\begin{aligned}\partial_T\frac{1}{2}\left(\int_0^T\int_\Omega|Du-D\overline{u}|^2\mathrm{d}x\,\mathrm{d}t+\int_\Omega|u(T)-\overline{u}(T)|^2\mathrm{d}x\right)\\ \le\operatorname{Lip}(\sigma)\int_0^T\int_\Omega\big(|Du-D\overline{u}|^2+|u_t-\overline{u}_t|^2\big)\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{2.5}$$

Similarly, if we use $u_t-\overline{u}_t$ as a test function, we get

$$\partial_T\frac{1}{2}\int_0^T\int_\Omega|u_t-\overline{u}_t|^2\mathrm{d}x\,\mathrm{d}t\le\frac{1}{4}\operatorname{Lip}^2(\sigma)\int_0^T\int_\Omega|Du-D\overline{u}|^2\mathrm{d}x\,\mathrm{d}t. \tag{2.6}$$

The asserted uniqueness follows by applying Gronwall's inequality (see, e.g., [12]) to the sum of the two inequalities. A discrete version of this Gronwall argument is used in section 3 as the key ingredient in Theorem 4.1.

2.3. Definition of the implicit scheme. In order to define the discrete scheme, let U_0, $V_0\in\mathcal{S}_0$ denote given approximations to u_0 and v_0. We assume that

$$\|e_0\|=\|u_0-U_0\|\le c_Ah_0\|Du_0\|,$$

and additionally for $u_0\in W^{2,2}(\Omega;\mathbb{R}^m)$ and $v_0\in W^{1,2}(\Omega;\mathbb{R}^m)$, that

$$\|e_0\|+h_0\|De_0\|\le c_Ah_0^2\|D^2u_0\|,\quad \|\delta_0\|=\|v_0-V_0\|\le c_Ah_0\|Dv_0\|. \tag{2.7}$$

We then define successively the discrete solution U_j at time t_j by minimizing the variational integral (2.8) below. Since we allow a variable step-size in the time discretization, we cannot discretize the second derivatives with a second difference quotient, and we use a backward difference quotient for the discrete velocities instead.

THEOREM 2.2. *Suppose that* $\sigma=D\Phi$ *is Lipschitz continuous with Lipschitz constant* $\operatorname{Lip}(\sigma)$ *and that* k *is small enough such that* $2k\operatorname{Lip}(\sigma)\le1$. *Then there exists for* $j=1,\dots,N$ *a unique solution* U_j *of the variational problem: Minimize*

$$\int_\Omega\left(\Phi(DU)+\frac{1}{2k_j}|DU-DU_{j-1}|^2+\frac{1}{2}\left|\frac{1}{k_j}(U-U_{j-1})-V_{j-1}\right|^2\right)\mathrm{d}x \tag{2.8}$$

among all functions $U \in \mathcal{S}_{0,j}$. The minimizer U_j is a solution of the corresponding Euler–Lagrange system in weak form, i.e., a solution of

$$\int_\Omega \big(\big(k_j\sigma(DU_j) + D(U_j - U_{j-1})\big) : DW_j + (V_j - V_{j-1})\cdot W_j\big)\mathrm{d}x = 0 \tag{2.9}$$

for all $W_j \in \mathcal{S}_{0,j}$.

Proof. We need only to show that the variational integral has a convex integrand. Existence and uniqueness of solutions follow then from the direct method in the calculus of variations (see, e.g., [3]). By assumption,

$$0 \le \frac{1}{k_j}|A-B|^2 - \mathrm{Lip}(\sigma)|A-B|^2 \le \left(\sigma(A) + \frac{1}{k_j}A - \left(\sigma(B) + \frac{1}{k_j}B\right)\right) : (A-B);$$

that is, $\sigma(F) + k_j^{-1}F$ is monotone, and hence $\Phi(F) + |F|^2/(2k_j)$ is convex. ∎

Remark. The structural assumption $\sigma = D\Phi$ guarantees the existence of the solution u and U_j of the continuous and the discretized system. The error analysis below is entirely based on the Galerkin orthogonality (2.8) and does not rely on this assumption.

2.4. Discrete orthogonality. The following version of the Galerkin orthogonality is an important ingredient in the proof of Theorem 4.1.

PROPOSITION 2.3. *Suppose that u is the unique solution of the system* (1.1) *and that $\{U_j\}$ is the unique approximation constructed in* (2.8). *Then*

$$(\delta_j - \delta_{j-1}, W_j) + \int_{I_j} \big(\sigma(Du) - \sigma(DU_j), DW_j\big)\mathrm{d}t + (De_j - De_{j-1}, DW_j) = 0 \tag{2.10}$$

for $j = 1, \dots, N$ and for all $W_j \in \mathcal{S}_{0,j}$.

Proof. The idea is to test the weak formulation (2.4) by $\chi_j W_j$ in order to get an analogue of (2.9); χ_j denotes the characteristic function of the time interval I_j. Let $W_j^\ell \in C_0^\infty(\Omega;\mathbb{R}^m)$ be a sequence of smooth functions with $\|W_j - W_j^\ell\|_{W^{1,2}(\Omega)} \to 0$ as $\ell \to \infty$, and choose $\psi_\mu \in C_0^\infty(I_j)$ with $\psi_\mu \equiv 1$ on $(t_{j-1}+\mu, t_j - \mu)$. Then

$$\int_0^\infty \int_\Omega \left((\sigma(Du) + Du_t) : DW_j^\ell(x)\psi_\mu(t) - u_t \cdot W_j^\ell(x)\frac{\partial}{\partial t}\psi_\mu(t)\right)\mathrm{d}x\,\mathrm{d}t = 0$$

for $\mu \in (0, k_j/4)$ and $\ell \in \mathbb{N}$. This expression converges for $\ell \to \infty$ and $\mu \to 0$ to

$$\int_{I_j} \big(\sigma(Du(t,x)) + Du_t(t,x), DW_j\big)\,\mathrm{d}t + (v_j - v_{j-1}, W_j) = 0.$$

The assertion of the proposition follows by subtracting this equation from the Euler–Lagrange equation (2.9). ∎

2.5. A discrete Gronwall inequality. Our convergence result is based on the following discrete Gronwall inequality.

LEMMA 2.4 (see [8, Lemma 1.4.2]). *Suppose that $a > 0$ and that $\{b_\nu\}$, $\{\tau_\nu\}$ are sequences of nonnegative real numbers. Assume that the sequence $\{\varphi_\nu\}$ satisfies*

$$\varphi_0 \le a \quad \text{and} \quad \varphi_\nu \le a + \sum_{j=1}^{\nu} b_j + \sum_{j=0}^{\nu-1} \tau_j\varphi_j$$

for $n \geq 1$. Then

$$\varphi_\nu \leq \left(a + \sum_{j=1}^{\nu} b_j\right) \exp\left(\sum_{k=0}^{\nu-1} \tau_k\right).$$

3. Estimates for the solution of the discrete system. In our estimates, we will frequently express a difference $(e_j - e_{j-1})/k_j$ of spatial errors as an error in velocities, δ_j. The resulting correction term K_j is characterized in the following lemma.

LEMMA 3.1. *Let K_j, $j = 1, \ldots, N$, be given by*

$$K_j = \frac{1}{k_j} \int_{t_{j-1}}^{t_j} (s - t_{j-1}) u_{tt}(s) \mathrm{d}s.$$

Then, for $j = 1, \ldots, N$,

$$\frac{1}{k_j}(e_j - e_{j-1}) = \delta_j + K_j \tag{3.1}$$

and

$$\|K_j\|^2 \leq \frac{k_j}{3} \int_{t_{j-1}}^{t_j} \int_\Omega u_{tt}^2 \mathrm{d}x\, \mathrm{d}t, \quad \|DK_j\|^2 \leq \frac{k_j}{3} \int_{t_{j-1}}^{t_j} \int_\Omega |Du_{tt}|^2 \mathrm{d}x\, \mathrm{d}t. \tag{3.2}$$

Proof. It follows from Taylor's formula that

$$u(t_j, x) = u(t_{j-1}, x) + k_j u_t(t_j, x) - \int_{t_{j-1}}^{t_j} (s - t_{j-1}) u_{tt}(s, x) \mathrm{d}s.$$

This allows us to estimate

$$\begin{aligned} \frac{1}{k_j}(e_j - e_{j-1}) &= \frac{U_j - U_{j-1}}{k_j} - \frac{u(t_j) - u(t_{j-1})}{k_j} \\ &= V_j - v(t_j) + K_j = \delta_j + K_j. \end{aligned}$$

Finally, by Hölder's inequality,

$$\begin{aligned} \|K_j\|^2 &= \int_\Omega \left(\frac{1}{k_j} \int_{t_{j-1}}^{t_j} (s - t_j) u_{tt}(s) \mathrm{d}s \right)^2 \mathrm{d}x \\ &\leq \frac{1}{k_j^2} \left(\int_{t_{j-1}}^{t_j} (s - t_{j-1})^2 \mathrm{d}s \right) \left(\int_{t_{j-1}}^{t_j} \int_\Omega u_{tt}^2 \mathrm{d}x\, \mathrm{d}t \right) \\ &= \frac{k_j}{3} \int_{t_{j-1}}^{t_j} \int_\Omega u_{tt}^2 \mathrm{d}x\, \mathrm{d}t. \end{aligned}$$

This concludes the proof of the lemma. ∎

The next proposition is a discrete analogue of (2.6). The idea is to use the orthogonality (2.10) with $\Pi_j \delta_j = \delta_j + (\Pi_j \delta_j - \delta_j)$ and to recover the structure of (2.6) plus approximation errors.

PROPOSITION 3.2. *The following estimate holds for $j = 1, \dots, N$:*

$$\begin{aligned}
&\frac{1}{2}\|\delta_j\|^2 - \frac{1}{2}\|\delta_{j-1}\|^2 + \frac{1}{4}\|\delta_j - \delta_{j-1}\|^2 \\
&\quad \le 2c_S^2 \operatorname{Lip}^2(\sigma) k_j \|De_j\|^2 - \frac{k_j}{2}\|D\delta_j\|^2 + \|\delta_j - \Pi_j\delta_j\|^2 \\
&\qquad + \frac{5k_j}{2}\|D\delta_j - D\Pi_j\delta_j\|^2 + \frac{5k_j}{2}\|DK_j\|^2 + c_S^2 \operatorname{Lip}^2(\sigma)\, k_j^2 \|Du_t\|^2_{L^2(I_j;L^2)}.
\end{aligned}$$

Proof. It follows from (2.10) with $W_j = \Pi_j\delta_j$ that

$$\begin{aligned}
\frac{1}{2}\|\delta_j\|^2 - \frac{1}{2}\|\delta_{j-1}\|^2 + \frac{1}{2}\|\delta_j - \delta_{j-1}\|^2 &= (\delta_j, \delta_j - \delta_{j-1}) \\
&= (\delta_j - \delta_{j-1}, \delta_j - \Pi_j\delta_j) - (De_j - De_{j-1}, D\Pi_j\delta_j) \\
&\quad - \int_{I_j} \big(\sigma(Du) - \sigma(DU_j), D\Pi_j\delta_j\big)\mathrm{d}t.
\end{aligned}$$

We denote the three terms on the right-hand side by T_1, T_2, and T_3. Then

$$T_1 \le \frac{1}{4}\|\delta_j - \delta_{j-1}\|^2 + \|\delta_j - \Pi_j\delta_j\|^2,$$

and by (3.1)

$$\begin{aligned}
T_2 &= -k_j\big(D\delta_j + DK_j, D\delta_j - D\delta_j + D\Pi_j\delta_j\big) \\
&\le k_j\bigg(-\frac{3}{4}\|D\delta_j\|^2 + \frac{5}{2}\|DK_j\|^2 + \frac{5}{2}\|D\delta_j - D\Pi_j\delta_j\|^2\bigg).
\end{aligned}$$

We have by Hölder's inequality, Young's inequality, and the stability estimate (2.2) that

$$\begin{aligned}
|T_3| &\le \int_{I_j} \|\sigma(Du) - \sigma(DU_j)\|\, \|D\Pi_j\delta_j\|\mathrm{d}t \\
&\le \sqrt{k_j}\|\sigma(Du) - \sigma(DU_j)\|_{L^2(I_j;L^2)}\, \|D\Pi_j\delta_j\| \\
&\le c_S^2 \operatorname{Lip}^2(\sigma) \int_{I_j}\int_\Omega |Du - DU_j|^2 \mathrm{d}x\,\mathrm{d}t + \frac{k_j}{4c_S^2}\|D\Pi_j\delta_j\|^2 \\
&\le 2c_S^2 \operatorname{Lip}^2(\sigma) \int_{I_j}\int_\Omega \big(|Du - Du_j|^2 + |Du_j - DU_j|^2\big)\mathrm{d}x\,\mathrm{d}t + \frac{k_j}{4}\|D\delta_j\|^2.
\end{aligned}$$

Since

$$|Du - Du_j|^2 = \bigg|\int_t^{t_j} Du_t(s)\mathrm{d}s\bigg|^2 \le (t_j - t)\int_t^{t_j} |Du_t(s)|^2\mathrm{d}s$$

and

$$\int_{t_{j-1}}^{t_j} (t_j - t)\int_t^{t_j} |Du_t|^2\mathrm{d}s\,\mathrm{d}t \le \int_{t_{j-1}}^{t_j} (t_j - t)\mathrm{d}t \int_{t_{j-1}}^{t_j} |Du_t|^2\mathrm{d}s = \frac{k_j^2}{2}\int_{t_{j-1}}^{t_j} |Du_t|^2\mathrm{d}t,$$

we obtain

$$\int_{I_j}\int_\Omega |Du - Du_j|^2\mathrm{d}x\,\mathrm{d}t \le \frac{k_j^2}{2}\int_{I_j}\int_\Omega |Du_t|^2\mathrm{d}x\,\mathrm{d}t.$$

The assertion of the proposition follows easily from the foregoing estimates. ∎

In the next proposition we derive the analogue of the estimate (2.5) for the discrete scheme.

PROPOSITION 3.3. *The following estimate holds for $j = 1, \dots, N$:*

$$\begin{aligned}
&\frac{1}{2}\|De_j\|^2 - \frac{1}{2}\|De_{j-1}\|^2 + \frac{1}{4}\|De_j - De_{j-1}\|^2 \\
&\quad \le \left(c_S^2 \operatorname{Lip}^2(\sigma) + \frac{1}{2}\right) k_j \|De_j\|^2 - (\delta_j - \delta_{j-1}, e_j) + \frac{k_j}{2}\|D\delta_j\|^2 \\
&\qquad + \frac{1}{4}\|\delta_j - \delta_{j-1}\|^2 + \|e_j - \Pi_j e_j\|^2 + k_j \|De_j - D\Pi_j e_j\|^2 \\
&\qquad + \frac{k_j}{2}\|DK_j\|^2 + \frac{1}{2}\, c_S^2 \operatorname{Lip}^2(\sigma) k_j^2 \, \|Du_t\|^2_{L^2(I_j;L^2)}.
\end{aligned}$$

Proof. We obtain from (2.10) with $W_j = \Pi_j e_j$ that

$$\begin{aligned}
\frac{1}{2}\|De_j\|^2 - \frac{1}{2}\|De_{j-1}\|^2 + \frac{1}{2}\|De_j - De_{j-1}\|^2 &= \big(De_j - De_{j-1}, De_j - D\Pi_j e_j\big) \\
- (\delta_j - \delta_{j-1}, \Pi_j e_j) - \int_{I_j} \big(\sigma(Du) - \sigma(DU_j), D\Pi_j e_j\big) \mathrm{d}t.&
\end{aligned}$$

We denote the three terms on the right-hand side by T_1, T_2, and T_3. By Lemma 3.1,

$$\begin{aligned}
T_1 &= k_j \big(D\delta_j + DK_j, De_j - D\Pi_j e_j\big) \\
&\le k_j \left(\frac{1}{2}\|D\delta_j\|^2 + \frac{1}{2}\|DK_j\|^2 + \|De_j - D\Pi_j e_j\|^2\right),
\end{aligned}$$

and by Young's inequality

$$\begin{aligned}
T_2 &= (\delta_j - \delta_{j-1}, e_j - \Pi_j e_j) - (\delta_j - \delta_{j-1}, e_j) \\
&\le -(\delta_j - \delta_{j-1}, e_j) + \frac{1}{4}\|\delta_j - \delta_{j-1}\|^2 + \|e_j - \Pi_j e_j\|^2.
\end{aligned}$$

Finally, T_3 can be estimates as in the proof of Proposition 3.2 by

$$\begin{aligned}
T_3 &\le \sqrt{k_j} \operatorname{Lip}(\sigma) \|Du - DU_j\|_{L^2(I_j;L^2)} \|D\Pi_j e_j\| \\
&\le \frac{1}{2}\, c_S^2 \operatorname{Lip}^2(\sigma) \, \|Du - DU_j\|^2_{L^2(I_j;L^2)} + \frac{k_j}{2c_S^2}\|D\Pi_j e_j\|^2 \\
&\le \frac{1}{2}\, c_S^2 \operatorname{Lip}^2(\sigma) k_j^2 \, \|Du_t\|^2_{L^2(I_j;L^2)} + \left(c_S^2 \operatorname{Lip}^2(\sigma) + \frac{1}{2}\right) k_j \|De_j\|^2.
\end{aligned}$$

The assertion of the proposition follows from the foregoing inequalities. ∎

We now combine the estimates in Propositions 3.2 and 3.3 to obtain an estimate on the time interval I_j.

COROLLARY 3.4. *Let $\mathcal{A}_j$ denote the approximation errors,*

$$\mathcal{A}_j = \|\delta_j - \Pi_j \delta_j\|^2 + \frac{5}{2}\, k_j \|D\delta_j - D\Pi_j \delta_j\|^2 + \|e_j - \Pi_j e_j\|^2 + k_j \|De_j - D\Pi_j e_j\|^2,$$

and $\mathcal{R}_j$ the terms depending on the regularity of u,

$$\mathcal{R}_j = \frac{k_j}{2}\, \|K_j\|^2 + 3k_j \|DK_j\|^2 + \frac{3}{2}\, c_S^2 \operatorname{Lip}^2(\sigma) k_j^2 \|Du_t\|^2_{L^2(I_j;L^2)}.$$

Then the following estimate holds for $j = 1, \dots, N$:

$$\begin{aligned}(3.3)\qquad &\frac{1}{2}\|\delta_j\|^2 - \frac{1}{2}\|\delta_{j-1}\|^2 + \frac{1}{2}\|De_j\|^2 - \frac{1}{2}\|De_{j-1}\|^2 + \frac{1}{4}\|De_j - De_{j-1}\|^2 \\ &\le \left(3c_S^2 \operatorname{Lip}^2(\sigma) + \frac{1}{2}\right) k_j \|De_j\|^2 - (\delta_j - \delta_{j-1}, e_j) + \mathcal{A}_j + \mathcal{R}_j - \frac{k_j}{2}\|K_j\|^2.\end{aligned}$$

Remark. The discrete Gronwall inequality will be applied to the sum of (3.3) in j. The sum of the terms $-(\delta_j - \delta_{j-1}, e_j)$ on the right-hand side is estimated by a discrete summation by parts which leads to the (discrete) time integral of $\|\delta_j\|^2$.

PROPOSITION 3.5. *Let*

$$(3.4)\qquad q_j = \frac{k_j}{k_{j-1}}, \quad Q_\nu = \max_{j=2,\dots,\nu} q_j, \quad c_3 = \max\left\{3c_S^2 \operatorname{Lip}^2(\sigma) + \frac{1}{2}, Q_N + \frac{1}{2}\right\}.$$

Then the following estimate holds for $\nu = 1, \dots, N$:

$$\begin{aligned}\frac{1}{2}\|\delta_\nu\|^2 + \frac{1}{2}\|De_\nu\|^2 \le c_3 \sum_{j=1}^{\nu} k_j\big(\|\delta_j\|^2 + \|De_j\|^2\big) - (\delta_\nu, e_\nu) + (\delta_0, e_1) \\ + \sum_{j=1}^{\nu}(\mathcal{A}_j + \mathcal{R}_j) + \frac{1}{2}\|\delta_0\|^2 + \frac{1}{2}\|De_0\|^2.\end{aligned}$$

Proof. We take the sum of the inequality (3.3) for $j = 1, \dots, \nu$ and obtain

$$\begin{aligned}\frac{1}{2}\|\delta_\nu\|^2 + \frac{1}{2}\|De_\nu\|^2 \le \left(3c_S^2 \operatorname{Lip}^2(\sigma) + \frac{1}{2}\right) \sum_{j=1}^{\nu} k_j \|De_j\|^2 - \sum_{j=1}^{\nu} \big(\delta_j - \delta_{j-1}, e_j\big) \\ + \sum_{j=1}^{\nu} \left(\mathcal{A}_j + \mathcal{R}_j - \frac{k_j}{2}\|K_j\|^2\right) + \frac{1}{2}\|\delta_0\|^2 + \frac{1}{2}\|De_0\|^2.\end{aligned}$$

With a discrete summation by parts in the second term on the right-hand side, we deduce that

$$\begin{aligned}-\sum_{j=1}^{\nu} \big(\delta_j - \delta_{j-1}, e_j\big) &= -(\delta_\nu, e_\nu) + (\delta_0, e_1) + \sum_{j=1}^{\nu-1} \big(\delta_j, e_{j+1} - e_j\big) \\ &= -(\delta_\nu, e_\nu) + (\delta_0, e_1) + \sum_{j=1}^{\nu-1} k_{j+1}\big(\delta_j, \delta_{j+1} + K_{j+1}\big) \\ &\le -(\delta_\nu, e_\nu) + (\delta_0, e_1) + \sum_{j=1}^{\nu-1} k_{j+1}\left(\|\delta_j\|^2 + \frac{1}{2}\|\delta_{j+1}\|^2 + \frac{1}{2}\|K_{j+1}\|^2\right).\end{aligned}$$

The assertion follows from this inequality since $k_{j+1}\|\delta_j\|^2 = q_{j+1}k_j\|\delta_j\|^2$. ∎

In order to apply Gronwall's inequality, we need a further summation of the inequality in Proposition 3.5. The term $-(\delta_\nu, e_\nu)$ corresponds to the spatial integral of $u_t u = \partial_t |u|^2/2$ which fits naturally in the formulation of Gronwall's inequality in section 2.2. This is not the case in the implicit time discretization used here.

PROPOSITION 3.6. *Let*

$$(3.5)\qquad \varphi_\nu = \|e_\nu\|^2 + \sum_{j=1}^{\nu} k_j\big(\|\delta_j\|^2 + \|De_j\|^2\big).$$

Then

$$\frac{1}{2}\varphi_N \le \left(c_3+\frac{1}{2}\right)\sum_{\nu=1}^{N} k_\nu\varphi_\nu + T(\delta_0,e_1) + \frac{1}{2}\|e_0\|^2 + \frac{T}{2}\big(\|\delta_0\|^2+\|De_0\|^2\big) + 2\sum_{\nu=1}^{N} k_\nu \sum_{j=1}^{\nu}\big(\mathcal{A}_j+\mathcal{R}_j\big),$$

and

$$\big|T(\delta_0,e_1)\big| \le k_1\sum_{\nu=1}^{N} k_\nu\varphi_\nu + 2k_1T\mathcal{R}_1 + \frac{3T}{4}\,\|\delta_0\|^2 + T\|e_0\|^2. \tag{3.6}$$

Proof. We multiply the inequality in the assertion of Proposition 3.5 by k_ν and take the sum of the resulting inequalities from $\nu=1$ to N. This leads to

$$\frac{1}{2}\sum_{\nu=1}^{N} k_\nu\big(\|\delta_\nu\|^2+\|De_\nu\|^2\big) \le c_3\sum_{\nu=1}^{N} k_\nu\sum_{j=1}^{\nu} k_j\big(\|\delta_j\|^2+\|De_j\|^2\big)$$
$$-\sum_{\nu=1}^{N} k_\nu(\delta_\nu,e_\nu) + T(\delta_0,e_1) + \sum_{\nu=1}^{N} k_\nu\sum_{j=1}^{\nu}\big(\mathcal{A}_j+\mathcal{R}_j\big) + \frac{T}{2}\big(\|\delta_0\|^2+\|De_0\|^2\big).$$

In view of Lemma 3.1,

$$\begin{aligned}-\sum_{\nu=1}^{N} k_\nu(\delta_\nu,e_\nu) &= -\sum_{\nu=1}^{N}(e_\nu-e_{\nu-1}-k_\nu K_\nu,e_\nu)\\ &= -\frac{1}{2}\sum_{\nu=1}^{N}\big(\|e_\nu\|^2-\|e_{\nu-1}\|^2+\|e_\nu-e_{\nu-1}\|^2\big)+\sum_{\nu=1}^{N} k_\nu(K_\nu,e_\nu)\\ &\le -\frac{1}{2}\,\|e_N\|^2+\frac{1}{2}\,\|e_0\|^2+\frac{1}{2}\sum_{\nu=1}^{N} k_\nu\big(\|e_\nu\|^2+\|K_\nu\|^2\big).\end{aligned}$$

This implies

$$\begin{aligned}&\frac{1}{2}\left\{\|e_N\|^2+\sum_{\nu=1}^{N} k_\nu\big(\|\delta_\nu\|^2+\|De_\nu\|^2\big)\right\}\\ &\le \left(c_3+\frac{1}{2}\right)\sum_{\nu=1}^{N} k_\nu\left\{\|e_\nu\|^2+\sum_{j=1}^{\nu} k_j\big(\|\delta_j\|^2+\|De_j\|^2\big)\right\}\\ &\quad + T(\delta_0,e_1)+\frac{1}{2}\|e_0\|^2+\frac{T}{2}\big(\|\delta_0\|^2+\|De_0\|^2\big)+2\sum_{\nu=1}^{N} k_\nu\sum_{j=1}^{\nu}\big(\mathcal{A}_j+\mathcal{R}_j\big).\end{aligned}$$

It remains only to estimate the term $T(\delta_0,e_1)$ on the right-hand side. This is accomplished by showing that

$$\big|T(\delta_0,e_1)\big| \le k_1\sum_{\nu=1}^{N} k_\nu\varphi_\nu + 2k_1T\mathcal{R}_1 + \frac{3T}{4}\,\|\delta_0\|^2 + T\|e_0\|^2.$$

Indeed, it follows from (3.1) that

$$\begin{aligned}\big|(\delta_0,e_1)\big| &\le \big|(\delta_0,e_1-e_0)\big| + \big|(\delta_0,e_0)\big| \le \big|k_1(\delta_0,\delta_1+K_1)\big| + \big|(\delta_0,e_0)\big| \\ &\le \frac{1}{4}\|\delta_0\|^2 + k_1^2\|\delta_1\|^2 + \frac{1}{4}\|\delta_0\|^2 + k_1^2\|K_1\|^2 + \frac{1}{4}\|\delta_0\|^2 + \|e_0\|^2.\end{aligned}$$

By definition, $k_1\|K_1\|^2 \le 2\mathcal{R}_1$ and $k_1\|\delta_1\|^2 \le \varphi_\nu$ for $\nu = 1,\dots,N$. Thus

$$T\big|(\delta_0,e_1)\big| \le \frac{3T}{4}\,\|\delta_0\|^2 + T\|e_0\|^2 + k_1\sum_{\nu=1}^{N} k_\nu\varphi_\nu + 2k_1\sum_{\nu=1}^{N} k_\nu\mathcal{R}_1.$$

This implies the assertion of the proposition. □

4. General convergence result. We are now in a position to state and prove the convergence result for the approximation of solutions of the system (1.1) by the implicit Euler scheme defined in section 2.3. Recall that $\mathcal{A}_j$, $\mathcal{R}_j$, and φ_ν have been defined in Proposition 3.4 and Corollary 3.6, respectively. Moreover, we set

$$\mathcal{A}(T) = 8\big(1+k_1\big)\sum_{\nu=1}^{N} k_\nu \sum_{j=1}^{\nu}\mathcal{A}_j, \quad \mathcal{R}(T) = 8\big(1+k_1\big)\sum_{\nu=1}^{N} k_\nu \sum_{j=1}^{\nu}\mathcal{R}_j,$$

and

$$a = 2(2T+1)\|e_0\|^2 + 2T\|De_0\|^2 + 5T\|\delta_0\|^2. \tag{4.1}$$

Finally, we define $c_4 = c_4(\mathrm{Lip}(\sigma), Q_N)$ and $c_5 = c_5(\mathrm{Lip}(\sigma), Q_N)$ by

$$c_4 = 4\bigg(c_3 + \frac{1}{2} + k_1\bigg), \quad c_5 = \max\bigg\{\frac{3}{2}\,c_S^2\,\mathrm{Lip}^2(\sigma), 1\bigg\}, \tag{4.2}$$

respectively.

THEOREM 4.1. *Suppose that Ω, $\mathcal{T}_j$, and $\mathcal{S}_{0,j}$ satisfy the assumptions in section* 2.1 *and that there exists a $Q_N > 0$ such that $k_{j-1} \le Q_N k_j$ for $j = 2,\dots,N$. Assume, furthermore, that σ is globally Lipschitz continuous and that the unique solution of the system* (1.1) *belongs to $W^{1,2}(W^{1,2})$. Suppose that $k < 1/c_4$. Then*

$$\max_{\nu=1,\dots,N}\|e_\nu\|^2 + \sum_{\nu=1}^{N} k_\nu\big(\|\delta_\nu\|^2 + \|De_\nu\|^2\big) \le \big(a + \mathcal{A}(T) + \mathcal{R}(T)\big)\exp\big(c_4T\big).$$

Moreover, if $u \in W^{1,\infty}(W^{2,2}) \cap W^{2,2}(W^{1,2})$, then

$$\mathcal{R}(T) \le 12c_5T\,k^2\,\Big(\|u_{tt}\|^2_{L^2(L^2)} + \|Du_{tt}\|^2_{L^2(L^2)} + \|Du_t\|^2_{L^2(L^2)}\Big),$$

and

$$\mathcal{A}(T) \le 60c_A^2T^2h^2\max_{j=1,\dots,N}\bigg(\frac{h_j^2}{k_j}+1\bigg)\|u\|^2_{W^{1,\infty}(W^{2,2})} + \mathcal{C}(T),$$

where $\mathcal{C}(T)$ is the coarsening error,

$$\mathcal{C}(T) = 12T\sum_{j=1}^{N}\bigg\{\frac{2}{k_j^2}\big\|U_{j-1} - \Pi_jU_{j-1}\big\| + \frac{5}{k_j}\big|U_{j-1} - \Pi_jU_{j-1}\big|^2_{1,2}\bigg\},$$

which vanishes if $\mathcal{T}_j$ is a refinement of $\mathcal{T}_{j-1}$. Finally, if the approximation estimate (2.7) *holds, then*

$$a \le c_A^2 h_0^2\big((4T+2)h_0^2 + 4T\big)\,\big(\|D^2 u_0\|^2 + \|Dv_0\|^2\big).$$

Remark. Theorem 4.1 constitutes the best a priori estimates known for the system (1.1) provided that the solution is sufficiently regular. Under the assumption that the triangulation is only refined in time but not coarsened (i.e., that $\mathcal{C}(T) = 0$), we obtain an estimate of order $\exp(T)(h^2 + k^2)$; cf. Theorem 1.1.

Proof of Theorem 4.1. Based on the estimates of section 3, we first show that φ_ν satisfies the assumptions in the discrete Gronwall inequality of Lemma 2.4. It follows from Proposition 3.6 and (3.6) that

$$\begin{aligned}\frac{1}{2}\varphi_N \le \left(c_3 + \frac{1}{2} + k_1\right)\sum_{\nu=1}^{N} k_\nu \varphi_\nu + 2\big(1+k_1\big)\sum_{\nu=1}^{N} k_\nu \sum_{j=1}^{\nu}\big(\mathcal{A}_j + \mathcal{R}_j\big)\\ + \left(T + \frac{1}{2}\right)\|e_0\|^2 + \frac{T}{2}\,\|De_0\|^2 + \frac{5T}{4}\|\delta_0\|^2.\end{aligned}$$

By assumption, $k_N \le k$ and thus $c_4 k_N = 4\big(c_3 + 1/2 + k_1\big)k_N \le 1$. This allows us to absorb the term $\big(c_3 + \frac{1}{2} + k_1\big)k_N\varphi_N \le \varphi_N/4$ on the left-hand side. We obtain

$$\begin{aligned}\varphi_N \le c_4 \sum_{\nu=1}^{N-1} k_\nu \varphi_\nu + 8(1+k_1)\sum_{\nu=1}^{N} k_\nu \sum_{j=1}^{\nu}\big(\mathcal{A}_j + \mathcal{R}_j\big)\\ + 2(2T+1)\|e_0\|^2 + 2T\|De_0\|^2 + 5T\|\delta_0\|^2.\end{aligned}$$

It follows that the assumptions in the discrete Gronwall inequality in Lemma 2.4 are satisfied with a as defined in (4.1) and

$$b_\nu = 8(1+k_1)k_\nu \sum_{j=1}^{\nu}\big(\mathcal{A}_j + \mathcal{R}_j\big), \quad \tau_\nu = c_4 k_\nu, \quad \nu = 1, \dots, N.$$

Thus

$$\varphi_N \le \left(a + \sum_{\nu=1}^{N} b_\nu\right)\exp(c_4 T) = \big(a + \mathcal{A}(T) + \mathcal{R}(T)\big)\,\exp(c_4 T).$$

We now estimate $\mathcal{A}(T)$ and $\mathcal{R}(T)$. By definition of $\mathcal{R}_j$ in Corollary 3.4 and by (3.2) we infer

$$\mathcal{R}_j \le c_5 k_j^2\,(\|u_{tt}\|^2_{L^2(I_j;L^2)} + \|Du_{tt}\|^2_{L^2(I_j;L^2)} + \|Du_t\|^2_{L^2(I_j;L^2)}).$$

This implies

$$\sum_{\nu=1}^{N} k_\nu \sum_{j=1}^{\nu}\mathcal{R}_j \le c_5 T\,k^2\,\big(\|u_{tt}\|^2_{L^2(L^2)} + \|Du_{tt}\|^2_{L^2(L^2)} + \|Du_t\|^2_{L^2(L^2)}\big).$$

Since $kc_4 \le 1$ implies $k_1 \le 1/2$, we obtain

$$\mathcal{R}(T) \le 12 c_5 T\,k^2\,(\|u_{tt}\|^2_{L^2(L^2)} + \|Du_t\|^2_{L^2(L^2)} + \|Du_t\|^2_{L^2(L^2)}).$$

It remains to estimate the approximation error. By definition of e_j and (2.1),

$$e_j - \Pi_j e_j = u(t_j) - U_j - \Pi_j\big(u(t_j) - U_j\big) = u(t_j) - \Pi_j u(t_j).$$

The approximation estimate (2.3) implies

$$\|e_j - \Pi_j e_j\|^2 + k_j\|De_j - D\Pi_j e_j\|^2 \le c_A^2 h_j^2\big(h_j^2 + k_j\big)\|D^2 u(t_j)\|^2.$$

Similarly,

$$\delta_j - \Pi_j\delta_j = v(t_j) - \Pi_j v(t_j) + \frac{1}{k_j}\big(U_{j-1} - \Pi_j U_{j-1}\big).$$

This estimate and (2.2)–(2.3) yield

$$\begin{aligned}\|\delta_j - \Pi_j\delta_j\|^2 + \frac{5}{2}k_j\|D\delta_j - D\Pi_j\delta_j\|^2 &\le 2c_A^2 h_j^2\left(h_j^2 + \frac{5}{2}k_j\right)\|D^2 u_t(t_j)\|^2\\ &\quad + \frac{2}{k_j^2}\|U_{j-1} - \Pi_j U_{j-1}\|^2 + \frac{5}{k_j}|U_{j-1} - \Pi_j U_{j-1}|_{1,2}^2.\end{aligned}$$

We conclude from the foregoing estimates that

$$\begin{aligned}\sum_{j=1}^N \mathcal{A}_j &\le 5\sum_{j=1}^N h_j^2 c_A^2\big(h_j^2 + k_j\big)\big(\|D^2 u(t_j)\|^2 + \|D^2 u_t(t_j)\|^2\big) + \frac{1}{12T}\,\mathcal{C}(T)\\ &\le 5c_A^2 T h^2 \max_{j=1,\dots,N}\left(\frac{h_j^2}{k_j} + 1\right)\|u\|_{W^{1,\infty}(W^{2,2})}^2 + \frac{1}{12T}\,\mathcal{C}(T).\end{aligned}$$

Hence

$$\mathcal{A}(T) \le 12\,T\sum_{j=1}^N \mathcal{A}_j \le 60c_A^2 T^2 h^2 \max_{j=1,\dots,N}\left(1 + \frac{h_j^2}{k_j}\right)\|u\|_{W^{1,\infty}(W^{2,2})}^2 + \mathcal{C}(T).$$

Clearly, if $\mathcal{T}_j$ is a refinement of $\mathcal{T}_{j-1}$, then, by (2.1), $\mathcal{C}(T) = 0$. Finally, if (2.7) holds, then

$$a \le c_A^2\big(2(2T+1)h_0^4 + 2Th_0^2\big)\|D^2 u_0\|^2 + 4Th_0^2\|Dv_0\|^2.$$

This proves the assertion of the theorem. □

Proof of Theorem 1.1. This is a special case of Theorem 4.1. □

Acknowledgment. It is our pleasure to thank Constantine Dafermos for discussions on the regularity of hyperbolic equations and remarks (4) and (5) in the introduction.

REFERENCES

[1] G. Andrews and J. M. Ball, *Asymptotic behaviour and changes of phase in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 44 (1982), pp. 306–341.

[2] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[3] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York, 1989.

[4] S. Demoulini, *Young measure solutions for nonlinear evolutionary systems of mixed type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 143–162.
[5] S. Demoulini, *Weak solutions for a class of nonlinear systems of viscoelasticity*, Arch. Ration. Mech. Anal., 155 (2000), pp. 299–334.
[6] G. Friesecke and G. Dolzmann, *Implicit time discretization and global existence for a quasilinear evolution equation with nonconvex energy*, SIAM J. Math. Anal., 28 (1997), pp. 363–380.
[7] P. Klouček and M. Luskin, *The computation of the dynamics of the martensitic transformation*, Contin. Mech. Thermodyn., 6 (1994), pp. 209–240.
[8] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.
[9] R. L. Pego, *Phase transitions in one-dimensional nonlinear viscoelasticity: Admissibility and stability*, Arch. Ration. Mech. Anal., 97 (1987), pp. 353–394.
[10] P. Rybka, *Dynamical modelling of phase transitions by means of viscoelasticity in many dimensions*, Proc. Roy. Soc. Edinburgh Sect. A, 121 (1992), pp. 101–138.
[11] P. J. Swart and P. J. Holmes, *Energy minimization and the formation of microstructure in dynamic anti-plane shear*, Arch. Ration. Mech. Anal., 121 (1992), pp. 37–85.
[12] W. Walter, *Ordinary Differential Equations*, Springer-Verlag, New York, 1998.
[13] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 90–108

ANALYSIS OF A MULTIGRID ALGORITHM FOR TIME HARMONIC MAXWELL EQUATIONS*

JAYADEEP GOPALAKRISHNAN†, JOSEPH E. PASCIAK‡, AND LESZEK F. DEMKOWICZ§

Abstract. This paper considers a multigrid algorithm suitable for efficient solution of indefinite linear systems arising from finite element discretization of time harmonic Maxwell equations. In particular, a "backslash" multigrid cycle is proven to converge at rates independent of refinement level if certain indefinite block smoothers are used. The method of analysis involves comparing the multigrid error reduction operator with that of a related positive definite multigrid operator. This idea has previously been used in multigrid analysis of indefinite second order elliptic problems. However, the Maxwell application involves a nonelliptic indefinite operator. With the help of a few new estimates, the earlier ideas can still be applied. Some numerical experiments with lowest order Nedelec elements are also reported.

Key words. multigrid, indefinite, Maxwell equations, preconditioner, Nedelec space, Jacobi, Gauss–Seidel, Poincaré inequality, finite element

AMS subject classifications. 65F10, 65N55, 65N30

DOI. 10.1137/S003614290139490X

1. Introduction.

The purpose of this paper is to study certain multigrid methods for the solution of the discrete equations which result from time harmonic Maxwell equations. Since the introduction of Nedelec elements [22], finite element methods using these **curl**-conforming elements have become a popular choice for discretization of Maxwell equations. An analysis of the finite element method in the time harmonic case and lossless media was provided in [20]. However, the efficient solution of the resulting linear systems has remained a challenge, mainly for two reasons: the linear systems are indefinite, and the differential operator **curl** has a large null space.

For the time harmonic problem, although a multigrid analysis has been lacking, numerical experiments indicating the suitability of certain two-level and multilevel algorithms can be found in literature [3, 4, 23]. Numerical results for parallel preconditioners based on Schwarz overlapping techniques were reported in [23]. Computational experiments with a multigrid V-cycle have been reported [3, 4]. More recently, an analysis for an additive overlapping preconditioner and a two-level multiplicative variant were given in [16].

Two works that made recent advances related to the development of preconditioners for Maxwell equations, [1] and [17], deserve special mention. Both provided smoothers for use in a multigrid V-cycle for the positive definite bilinear form $\mathbf{\Lambda}(\cdot,\cdot)$ defined later in (2.1). These smoothers are based on two different subspace decompositions of the Nedelec space. Our smoothers for the indefinite problem are constructed

*Received by the editors September 6, 2001; accepted for publication (in revised form) April 1, 2003; published electronically January 6, 2004. This work was supported in part by Air Force Contract F49620-98-1-0255 and NSF grant DMS-9973328. The authors also acknowledge support from the Texas Institute for Computational and Applied Mathematics, Austin, TX, and the Supercomputing Institute of the University of Minnesota.

http://www.siam.org/journals/sinum/42-1/39490.html

†Department of Mathematics, University of Florida, Gainesville, FL 32611 (jayg@math.ufl.edu).

‡Department of Mathematics, Texas A&M University, College Station, TX 77843 (pasciak@math.tamu.edu).

§Department of Aerospace Engineering and Engineering Mechanics, University of Texas, Austin, TX 78712 (leszek@ticam.utexas.edu).

based on the same decompositions, and our analysis makes use of the results in [1] and [17].

The current paper provides an analysis for a multilevel algorithm. Specifically, we prove that the so-called *backslash cycle* gives a convergent linear iterative method with a convergence rate independent of mesh size, provided the coarse grid is sufficiently fine. The latter restriction stems from the indefiniteness and seems unavoidable both in theory and practice. Fundamentally different solution methods may be needed to overcome this. Nonetheless, in spite of this restriction there are many practical applications (of moderate frequencies) where a multigrid iteration using a relatively fine coarse grid can reduce computational effort significantly.

The analysis we will provide is based on [16] and an earlier paper on multigrid applied to elliptic nonsymmetric and indefinite problems [6] (see also [8]). In [6], a perturbation technique to analyze a multigrid algorithm for indefinite or nonsymmetric operators was developed. This involves comparing the error propagation operator of the multigrid algorithm with that of a multigrid algorithm for a corresponding positive definite operator. The difference between these operators was then proved small for elliptic problems that may be nonsymmetric or indefinite. However, our application involves a nonelliptic operator. We will show that techniques in [6] can still be applied. In [16], some fundamental estimates were developed concerning the approximation properties of the discrete solution operator corresponding to the time harmonic Maxwell approximation. These estimates will play an important role in the analysis given here.

The outline of the remainder of the paper is as follows. In section 2, we define the problem and give the multigrid algorithm. Smoothers are defined and analyzed in section 3. Convergence estimates for the multigrid algorithm are given in section 4. Finally, the results of numerical experiments are given in section 5.

2. The problem and multigrid algorithm. We set up a model problem arising from time harmonic Maxwell equations and a simple multigrid algorithm in this section. First we establish notation for some spaces and their norms. Let Ω be an open bounded connected polyhedral domain in $\mathbb{R}^3$, and let $L^2(\Omega)$ denote the space of square integrable functions on Ω. We will use $(\cdot,\cdot)_\Omega$ and $\|\cdot\|_{0,\Omega}$ to denote the innerproduct and norm, respectively, in $L^2(\Omega)$ or $L^2(\Omega)^3$. The latter will often be abbreviated to $\|\cdot\|$. In the space of vector functions in $L^2(\Omega)^3$ with square integrable $\mathbf{curl}$, tangential traces $\boldsymbol{n}\times\boldsymbol{u}$ on the boundary $\partial\Omega$ are well defined [15], and we define

$$H_0(\mathbf{curl}\,;\Omega) = \{\boldsymbol{u}\in (L^2(\Omega))^3 : \mathbf{curl}\,\boldsymbol{u}\in (L^2(\Omega))^3,\ \boldsymbol{n}\times\boldsymbol{u}=0 \text{ on } \partial\Omega\}.$$

Here $\boldsymbol{n}$ is the unit outward normal on the boundary $\partial\Omega$. This space is normed with $\|\cdot\|_{\boldsymbol{\Lambda},\Omega} = \boldsymbol{\Lambda}(\cdot,\cdot)^{1/2}$, where

$$\boldsymbol{\Lambda}(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u},\boldsymbol{v})_\Omega + (\mathbf{curl}\,\boldsymbol{u},\mathbf{curl}\,\boldsymbol{v})_\Omega. \tag{2.1}$$

Analogous definitions hold for $\|\cdot\|_{0,D}$, $(\cdot,\cdot)_D$, and $\|\cdot\|_{\boldsymbol{\Lambda},D}$ in domains D different from Ω. In the notation for function spaces and their norms, when the domain is absent, it is to be taken as Ω; for example, $H_0(\mathbf{curl}) \equiv H_0(\mathbf{curl}\,;\Omega)$.

We restrict our attention to the time harmonic Maxwell equations in a homogeneous lossless media occupying Ω and also assume that the boundary of Ω is adjacent to a perfect conductor. The following equation is a variational system for the electric field $\boldsymbol{U}\in H_0(\mathbf{curl}\,;\Omega)$ given by Maxwell equations [11, 20] in the simple case of unit

material properties:

$$\boldsymbol{A}(\boldsymbol{U},\boldsymbol{v}) = (\boldsymbol{F},\boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in H_0(\mathbf{curl}\,;\Omega), \tag{2.2}$$

where

$$\boldsymbol{A}(\boldsymbol{U},\boldsymbol{v}) = (\mathbf{curl}\,\boldsymbol{U},\mathbf{curl}\,\boldsymbol{v}) - \omega^2(\boldsymbol{U},\boldsymbol{v}).$$

The vector $\boldsymbol{F}$, being a constant multiple of electric current, has zero divergence, and consequently $\operatorname{div}\boldsymbol{U} = 0$. In (2.2), ω is a real number denoting frequency of propagation. Note that there is a countable set of real values for ω for which (2.2) does not have a unique solution [19]. Throughout this paper we assume that ω is not one of these values and so (2.2) is uniquely solvable.

In our arguments later, we will need the solutions to (2.2) to be regular, and hence we assume that Ω is convex. It is well known [14, 20] that $\boldsymbol{U}, \mathbf{curl}\,\boldsymbol{U} \in (H^1(\Omega))^3$ and there is a constant C_Ω depending only on Ω such that

$$\|\boldsymbol{U}\|_{H^1} + \|\mathbf{curl}\,\boldsymbol{U}\|_{H^1} \le C_\Omega \|\boldsymbol{F}\|. \tag{2.3}$$

In (2.3), $\|\cdot\|_{H^1}$ denotes the norm in $(H^1(\Omega))^3$ and $H^1(\Omega) = \{u \in L^2(\Omega) : \mathbf{grad}\ u \in (L^2(\Omega))^3\}$. For later use, let us also denote $H^1_0(\Omega)$ to be the set of functions in $H^1(\Omega)$ which vanish on $\partial\Omega$.

The preconditioner which we shall consider is developed in terms of multilevel approximation subspaces of $H_0(\mathbf{curl})$. We start with a coarse partitioning of Ω into (nonoverlapping) tetrahedra $\mathcal{T}_1 = \{\tau_1^i : \ i = 1, \dots, N_0\}$. This forms a quasi-uniform mesh of mesh size d_1. A nested sequence of shape regular meshes $\mathcal{T}_k$, $k = 2, 3, \dots$, can be obtained by successively refining $\mathcal{T}_1$, using, e.g., techniques given in [2]. For a given tetrahedron τ, let h_τ denote the radius of the largest ball contained in τ, and let H_τ denote the diameter of τ. By uniformity, we assume that there is a constant ζ not depending on $\mathcal{T}_i$ satisfying

$$\zeta h_\tau \ge H_\tau \quad \text{for all } \tau \in \mathcal{T}_i, \quad i = 1, \dots, j. \tag{2.4}$$

Our goal is to solve the problem associated with the finest mesh $\mathcal{T}_j$ for some integer $j > 1$. The mesh size of $\mathcal{T}_1$ will be denoted by d_1 and can be taken to be the diameter of the largest tetrahedron. The mesh size of $\mathcal{T}_k$ is essentially $2^{1-k}d_1$.

For theoretical and practical purposes, the coarsest grid in the multilevel algorithm must be sufficiently fine. For $k = 1, \dots, J$, let M_k denote the lowest order Nedelec finite element subspaces [22] of $H_0(\mathbf{curl})$ (of the first kind) based on $\mathcal{T}_{k+L}$ for some $L \ge 0$. The coarsest approximation subspace M_1 can be made sufficiently accurate by increasing L. Since the meshes are nested, it follows that

$$M_1 \subset M_2 \subset \cdots \subset M_J.$$

The space M_k has a mesh size of $h_k = 2^{1-L-k}d_1 = 2^{1-k}h_1$. Also let W_k be the subspace of continuous scalar functions which are linear in every element of $\mathcal{T}_{k+L}$. In the appendix, we show how our results can be generalized to higher order Nedelec elements.

It was shown in [20] (see also [21]) that the discrete problem of finding $\boldsymbol{U}_k \in M_k$ satisfying

$$\boldsymbol{A}(\boldsymbol{U}_k,\boldsymbol{v}) = (\boldsymbol{F},\boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in M_k \tag{2.5}$$

has a unique solution provided h_k is small enough. We will assume that h_1 is small enough (or, equivalently, L is large enough) so that (2.5) is uniquely solvable for $k = 1, 2, \ldots, J$.

In our analysis, we shall use the projector $\boldsymbol{P}_k : H_0(\mathbf{curl}) \mapsto M_k$ defined by

$$\boldsymbol{A}(\boldsymbol{P}_k\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{A}(\boldsymbol{u}, \boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in M_k$$

and the orthogonal L^2-projector $\boldsymbol{Q}_k : (L^2(\Omega))^3 \mapsto M_k$ defined by

$$(\boldsymbol{Q}_k\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}, \boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in M_k.$$

That $\boldsymbol{P}_k$ is well defined (for $k = 1, 2, \ldots, J$) follows from the unique solvability of (2.5). Let us also introduce, for each k, an operator $\boldsymbol{A}_k : M_k \to M_k$ defined by

$$(\boldsymbol{A}_k\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{A}(\boldsymbol{u}, \boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in M_k.$$

Problem (2.5), on level J, can be rewritten in the above notation as

$$\boldsymbol{A}_J\boldsymbol{U}_J = \boldsymbol{Q}_J\boldsymbol{F}. \tag{2.6}$$

We describe a simple multigrid algorithm for iteratively computing the solution $\boldsymbol{U}_J$ of (2.6). Given an initial iterate $\boldsymbol{u}_0 \in M_J$, we define a sequence approximating $\boldsymbol{U}_J$ by

$$\boldsymbol{u}_{i+1} = \boldsymbol{Mg}_J(\boldsymbol{u}_i, \boldsymbol{Q}_J\boldsymbol{F}). \tag{2.7}$$

Here $\boldsymbol{Mg}_J(\cdot, \cdot)$ is the map of $M_J \times M_J$ into M_J defined by the following algorithm.

ALGORITHM 2.1. *Set* $\boldsymbol{Mg}_1(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{A}_1^{-1}\boldsymbol{w}$. *Let* $k > 1$ *and* $\boldsymbol{v}, \boldsymbol{w} \in M_k$. *Assuming that* $\boldsymbol{Mg}_{k-1}(\cdot, \cdot)$ *has been defined, we define* $\boldsymbol{Mg}_k(\boldsymbol{v}, \boldsymbol{w})$ *as follows:*

(1) *Set* $\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{R}_k(\boldsymbol{w} - \boldsymbol{A}_k\boldsymbol{v})$.

(2) $\boldsymbol{Mg}_k(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{x} + \boldsymbol{Mg}_{k-1}(\boldsymbol{0}, \boldsymbol{Q}_{k-1}(\boldsymbol{w} - \boldsymbol{A}_k\boldsymbol{x}))$.

Here $\boldsymbol{R}_k : M_k \mapsto M_k$ is a linear smoothing operator. Note that in this multigrid algorithm (often called a "backslash cycle") we smooth only as we proceed to coarser grids. Our smoothing operators will always be based on a generalized block Jacobi or block Gauss–Seidel iteration. In this case, the Gramm matrix inversions associated with $\boldsymbol{Q}_k$, $k = 2, \ldots, J$, are avoided (see [5] or [24]). The smoother $\boldsymbol{R}_k$ will be defined in section 3.

$\boldsymbol{Mg}_J(\cdot, \cdot)$ is a linear map from $M_J \times M_J$ into M_J. Moreover, the scheme is consistent in the sense that $\boldsymbol{v} = \boldsymbol{Mg}_J(\boldsymbol{v}, \boldsymbol{A}_J\boldsymbol{v})$ for all $\boldsymbol{v} \in M_J$. It easily follows that the linear operator $\boldsymbol{E} = \boldsymbol{Mg}_J(\cdot, 0)$ is the error reduction operator for (2.7), that is,

$$\boldsymbol{u} - \boldsymbol{u}_{i+1} = \boldsymbol{E}(\boldsymbol{u} - \boldsymbol{u}_i).$$

Error reduction operators for variational multigrid algorithms generally have a product representation (see, e.g., [7]). Let $\boldsymbol{T}_k = \boldsymbol{R}_k\boldsymbol{A}_k\boldsymbol{P}_k$ for $k > 1$ and set $\boldsymbol{T}_1 = \boldsymbol{P}_1$. Let $\boldsymbol{E}_k\boldsymbol{u} = \boldsymbol{u} - \boldsymbol{Mg}_k(\boldsymbol{0}, \boldsymbol{A}_k\boldsymbol{P}_k\boldsymbol{u})$ and $\boldsymbol{E}_0 \equiv \boldsymbol{I}$, the identity operator. Then

$$\boldsymbol{E}_k = \boldsymbol{E}_{k-1}(\boldsymbol{I} - \boldsymbol{T}_k)$$

and

$$\boldsymbol{E} = (\boldsymbol{I} - \boldsymbol{T}_1)(\boldsymbol{I} - \boldsymbol{T}_2)\cdots(\boldsymbol{I} - \boldsymbol{T}_J). \tag{2.8}$$

The product representation of the error operator given above will be a fundamental ingredient in the convergence analysis presented in section 4.

The above algorithm is a special case of more general multigrid algorithms in that we use only presmoothing. Alternatively, we could define an algorithm with just postsmoothing or both pre- and postsmoothing. The analysis of these algorithms is similar to that above and will not be presented. Algorithms with more than one smoothing are not generally advised since the smoothing iteration may be unstable.

Our multigrid analysis is based on perturbation and the estimates for the positive definite case. We define $\widetilde{\boldsymbol{P}}_k$, $\boldsymbol{\Lambda}_k$, and $\widetilde{\boldsymbol{T}}_k$ analogously to $\boldsymbol{P}_k$, $\boldsymbol{A}_k$, and $\boldsymbol{T}_k$ using the form $\boldsymbol{\Lambda}$ instead of $\boldsymbol{A}$.

3. Smoothers. In this section, we consider some smoothers appropriate for the multigrid algorithm (Algorithm 2.1). These smoothers are generalized Jacobi or Gauss–Seidel iterations, based on subspace decompositions of [1] and [17].

First, let us review the decomposition of [1]. For any $k \in \{2, 3, \dots, J\}$, let $x_{k,i}$, $i = 1, \dots, N_k^{\mathrm{I}}$, denote the interior vertices of the mesh $\mathcal{T}_{k+L}$. Let $\Omega_{k,i}^{\mathrm{I}}$ denote the interior of the union of the closures of the elements of $\mathcal{T}_{k+L}$ whose boundary contains $x_{k,i}$. Let $M_{k,i}^{\mathrm{I}}$ (resp., $W_{k,i}^{\mathrm{I}}$) denote the functions in M_k (resp., W_k) whose support is contained in $\overline{\Omega}_{k,i}^{\mathrm{I}}$. Then M_k admits the decomposition

$$M_k = \sum_{i=0}^{N_k^{\mathrm{I}}} M_{k,i}^{\mathrm{I}}.$$

Next, consider the decomposition of [17]. Let $\{\boldsymbol{\phi}_{k,i} : i = 1, \dots, n_k^M\}$ and $\{\psi_{k,i} : i = 1, \dots, n_k^W\}$ denote the usual nodal bases of M_k and W_k, respectively. Then this decomposition is given by

$$M_k = \sum_{i=0}^{N_k^{\mathrm{II}}} M_{k,i}^{\mathrm{II}},$$

where $M_{k,i}^{\mathrm{II}}$ equals the span of $\boldsymbol{\phi}_{k,i}$ for $i = 1, \dots, n_k^M$, while for $i = n_k^M + j$, $j = 1, \dots, n_k^W$, it equals the span of $\mathbf{grad}\ \psi_{k,j}$, and $N_k^{\mathrm{II}} = n_k^M + n_k^W$. Also let $\Omega_{k,i}^{\mathrm{II}}$ be such that $\overline{\Omega}_{k,i}^{\mathrm{II}}$ equals the support of nonzero functions in $M_{k,i}^{\mathrm{II}}$,

$$\mathring{M}_{k,i}^{\mathrm{I}} = \{\boldsymbol{u} \in M_{k,i}^{\mathrm{I}} \ :\ (\boldsymbol{u}, \mathbf{grad}\ \theta)_{\Omega_{k,i}^{\mathrm{I}}} = 0 \text{ for all } \theta \in W_{k,i}^{\mathrm{I}}\} \quad \text{for } i = 1, \dots, N_k^{\mathrm{I}},$$
$$\mathring{M}_{k,i}^{\mathrm{II}} = M_{k,i}^{\mathrm{II}} \quad \text{for } i = 1, \dots, n_k^M,$$

and let $\mathring{M}_{k,i}^{\mathrm{II}}$ for $i = n_k^M + 1, \dots, N_k^{\mathrm{II}}$ be empty.

Our smoothers for the indefinite form are based on the above decompositions. Let $d \in \{\mathrm{I}, \mathrm{II}\}$. Operators $\boldsymbol{Q}_{k,i}^d$, $\boldsymbol{\Lambda}_{k,i}^d$, and $\boldsymbol{A}_{k,i}^d$ are defined analogously to $\boldsymbol{Q}_k$, $\boldsymbol{\Lambda}_k$, and $\boldsymbol{A}_k$ by replacing M_k with $M_{k,i}^d$. The smoothing operators involve local solves on $M_{k,i}^d$, so before we define them we must ensure that the operators $\{\boldsymbol{A}_{k,i}^d\}$ are invertible. That this is the case if h_1 is taken sufficiently small is a consequence of the Poincaré–Friedrichs-type inequality of the next lemma. This inequality will also be important for a subsequent perturbation analysis.

In the remainder of the paper, we adopt the convention of denoting by C or c a generic constant independent of all mesh sizes $\{h_k\}$ and the number of levels J. It

will be explicitly stated when such an independence holds only in a range $0 < h_k < H$ for some H (i.e., only for small enough mesh sizes).

LEMMA 3.1. *For any* $\boldsymbol{q} \in \mathring{M}^d_{k,i}$, $d \in \{\mathrm{I}, \mathrm{II}\}$, $k = 2, \ldots, J$,

$$\|\boldsymbol{q}\| \le C h_k \|\mathbf{curl}\, \boldsymbol{q}\|. \tag{3.1}$$

Remark 3.1. Note that for discretely divergence free functions on a convex domain, such an inequality is proved in [15]. However, $\Omega^d_{k,i}$ may be nonconvex and we need the constant in the inequality to be independent of the shape of the mesh patches. The proof does not follow from a simple scaling argument as the discrete divergence free condition does not carry over under linear mapping unless the transformation is unitary.

Remark 3.2. In the case $d = \mathrm{II}$ and lowest order elements, this inequality is well known [17] and a simple proof can be given by a scaling argument.

Proof. Consider first a tetrahedron τ with a face f contained in the x–y plane with the origin at the barycenter of the face. A function $\boldsymbol{\phi}$ in the lowest order Nedelec edge space on τ with vanishing tangential components on f has the form

$$\boldsymbol{\phi} = (0, 0, \eta) + (\alpha_1, \alpha_2, 0) \times (x, y, z). \tag{3.2}$$

Here η, α_1, and α_2 are constants. Moreover, $\mathbf{curl}\, \boldsymbol{\phi} = 2\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, 0)$. Also note that if $\boldsymbol{a}$ is the vertex of τ not in f and $\boldsymbol{c}$ is any vertex of f, then the tangential component of $\boldsymbol{\phi}$ along the edge connecting $\boldsymbol{a}$ to $\boldsymbol{c}$ is given by

$$(\eta\, a_3 + (\boldsymbol{\alpha} \times \boldsymbol{c}) \cdot \boldsymbol{a})/|\boldsymbol{a} - \boldsymbol{c}|, \tag{3.3}$$

where a_3 is the z-component of $\boldsymbol{a}$. We will now prove the lemma for decompositions I and II separately.

Case $d = \mathrm{I}$. Let D be the domain formed by a collection of unit sized tetrahedra τ_j, $j = 0, 1, \ldots, N$, meeting at vertex $\boldsymbol{a}$, and let the corresponding approximation spaces (of $H_0(\mathbf{curl}\,; D)$ and $H^1_0(D)$, resp.) be denoted by M'_D and W'_D. Furthermore, let $\mathring{M}'_D = \{\boldsymbol{v} \in M'_D : (\boldsymbol{v}, \mathbf{grad}\, \theta)_D = 0 \text{ for all } \theta \in W'_D\}$. If we show that

$$\|\boldsymbol{v}\| \le C \|\mathbf{curl}\, \boldsymbol{v}\| \quad \text{for all } \boldsymbol{v} \in \mathring{M}'_D, \tag{3.4}$$

the required result follows easily by dilation.

Let $\boldsymbol{\phi} \in M'_D$, let f_j denote the face of τ_j not containing $\boldsymbol{a}$, and let $\boldsymbol{c}$ be a vertex of f_j. Let $\boldsymbol{a}$ and $\boldsymbol{c}$ have local coordinate triples $\boldsymbol{a}_j \equiv (a_{j,1}, a_{j,2}, a_{j,3})$ and $\boldsymbol{c}_j$, respectively, in the coordinate system on each tetrahedron which has f_j in the x–y plane, and the origin at its barycenter. Then, by (3.2), $\boldsymbol{\phi}$ has the form $\boldsymbol{\phi} = (0, 0, \eta_j) + \boldsymbol{\alpha}_j \times (x, y, z)$. By (3.3), the tangential component of $\boldsymbol{\phi}$ along the edge connecting $\boldsymbol{a}$ to $\boldsymbol{c}$ is given by $(\eta_j\, a_{j,3} + (\boldsymbol{\alpha}_j \times \boldsymbol{c}_j) \cdot \boldsymbol{a}_j)/|\boldsymbol{a}_j - \boldsymbol{c}_j|$. If τ_l is another tetrahedron in D sharing the vertex $\boldsymbol{c}$, then the same quantity is also given by $(\eta_l\, a_{l,3} + (\boldsymbol{\alpha}_l \times \boldsymbol{c}_l) \cdot \boldsymbol{a}_l)/|\boldsymbol{a}_l - \boldsymbol{c}_l|$. Here subscripts l indicate coordinates in the τ_l system. Thus

$$\eta_l = \frac{\eta_j\, a_{j,3} + (\boldsymbol{\alpha}_j \times \boldsymbol{c}_j) \cdot \boldsymbol{a}_j - (\boldsymbol{\alpha}_l \times \boldsymbol{c}_l) \cdot \boldsymbol{a}_l}{a_{l,3}}. \tag{3.5}$$

Let $\boldsymbol{v} \in \mathring{M}'_D$. We will construct a function $\boldsymbol{\phi}$ in M'_D which satisfies

$$\mathbf{curl}\, \boldsymbol{\phi} = \mathbf{curl}\, \boldsymbol{v} \quad \text{and} \quad \|\boldsymbol{\phi}\| \le C \|\mathbf{curl}\, \boldsymbol{v}\|, \tag{3.6}$$

with C depending only on the quasi uniformity condition. Note that $\boldsymbol{v} - \boldsymbol{\phi}$ is a gradient of a function in W_D', so

$$\|\boldsymbol{v}\| \le \|\boldsymbol{\phi}\| \le C\|\mathbf{curl}\,\boldsymbol{v}\|;$$

i.e., (3.4) follows if we construct $\boldsymbol{\phi}$ satisfying (3.6).

We define $\boldsymbol{\phi} = \boldsymbol{v} - \mu\mathbf{grad}\,\psi_{\boldsymbol{a}}$, where $\psi_{\boldsymbol{a}}$ is the nodal function in W_D' which is one on $\boldsymbol{a}$ and μ is to be determined. Clearly, $\mathbf{grad}\,\psi_{\boldsymbol{a}}$ has a local representation of the form

$$\mathbf{grad}\,\psi_{\boldsymbol{a}} = (0, 0, \zeta_j)$$

on τ_j with $\zeta_j \neq 0$. We choose μ so that $\eta_0 = 0$ in the above representation of $\boldsymbol{\phi}$. All of the remaining η_j's in the representation of $\boldsymbol{\phi}$ can be determined from the $\boldsymbol{\alpha}_j$'s by (3.5). By quasi uniformity, $\{a_{l,3}\}$ are uniformly bounded away from zero, so magnitudes of the η_l's can be bounded in terms of the $\boldsymbol{\alpha}$'s. Now (3.6) follows by quasi uniformity and the fact that the $\boldsymbol{\alpha}$'s can be bounded in terms of $\|\mathbf{curl}\,\boldsymbol{v}\|$.

Case $d = \mathrm{II}$. Let τ, f, $\boldsymbol{a}$, and $\boldsymbol{c}$ be as in the beginning of this proof. Then, in the coordinate system there, the nodal basis function $\boldsymbol{\phi}$ of the edge connecting $\boldsymbol{a}$ to $\boldsymbol{c}$ has the representation (3.2). Moreover, if $\boldsymbol{b}$ is an alternate vertex of f, then $\eta = -(\boldsymbol{\alpha} \times \boldsymbol{b}) \cdot \boldsymbol{a}/a_3$. Since η can be bounded by $\boldsymbol{\alpha} = \mathbf{curl}\,\boldsymbol{\phi}/2$, the proof can be finished in the same way as before. ∎

PROPOSITION 3.1. *There exists an* $H > 0$ *such that whenever* $h_1 \le H$, *any solution* $\boldsymbol{p}^d_{k,i} \in M^d_{k,i}$ *of the square system*

$$\boldsymbol{A}(\boldsymbol{p}^d_{k,i}, \boldsymbol{v}_{k,i}) = \boldsymbol{A}(\boldsymbol{u}, \boldsymbol{v}_{k,i}) \quad \textit{for all } \boldsymbol{v}_{k,i} \in M^d_{k,i}$$

satisfies

$$\left\|\boldsymbol{p}^d_{k,i}\right\|_{\boldsymbol{\Lambda},\Omega^d_{k,i}} \le C\,\|\boldsymbol{u}\|_{\boldsymbol{\Lambda},\Omega^d_{k,i}} \tag{3.7}$$

for $\boldsymbol{u} \in M_k$ *and for all* $i = 1, \ldots, N_k$ *and* $d \in \{\mathrm{I}, \mathrm{II}\}$. *It follows that* $\boldsymbol{A}^d_{k,i}$ *is nonsingular.*

Proof. In the case of decomposition I, the proof proceeds exactly as an analogous result in [16, Lemma 4.2], and we omit it.

In the case $d = \mathrm{II}$, for $i = 1, \ldots, n^M_k$, (3.7) follows for sufficiently small h_k from

$$\|\mathbf{curl}\,\boldsymbol{p}^d_{k,i}\|^2 - \omega^2\|\boldsymbol{p}^d_{k,i}\|^2 = (\mathbf{curl}\,\boldsymbol{u}, \mathbf{curl}\,\boldsymbol{p}^d_{k,i}) - \omega^2(\boldsymbol{u}, \boldsymbol{p}^d_{k,i})$$

by applying the Cauchy–Schwarz inequality on the right-hand side and Lemma 3.1 on the left-hand side. For the remaining i, since $\omega > 0$,

$$\|\boldsymbol{p}^d_{k,i}\|^2 = (\boldsymbol{u}, \boldsymbol{p}^d_{k,i}),$$

so (3.7) follows. ∎

Not only does Proposition 3.1 yield the invertibility of $\boldsymbol{A}^d_{k,i}$, but it also implies that the projection operator, $\boldsymbol{P}^d_{k,i} : M_k \mapsto M_{k,i}$, given by

$$\boldsymbol{A}(\boldsymbol{P}^d_{k,i}\boldsymbol{u}, \boldsymbol{v}_{k,i}) = \boldsymbol{A}(\boldsymbol{u}, \boldsymbol{v}_{k,i}) \quad \text{for all } \boldsymbol{u} \in M_k, \boldsymbol{v}_{k,i} \in M^d_{k,i}, \quad d \in \{\mathrm{I}, \mathrm{II}\}, \tag{3.8}$$

is well defined. Moreover, (3.7) implies

$$\left\|\boldsymbol{P}^d_{k,i}\boldsymbol{u}\right\|_{\boldsymbol{\Lambda},\Omega^d_{k,i}} \le C\,\|\boldsymbol{u}\|_{\boldsymbol{\Lambda},\Omega^d_{k,i}} \tag{3.9}$$

for all $\boldsymbol{u} \in M_k$. Also define $\widetilde{\boldsymbol{P}}_{k,i}^d$ analogously to $\boldsymbol{P}_{k,i}^d$ by replacing $\boldsymbol{A}$ with $\boldsymbol{\Lambda}$ in (3.8).

Now that we have proven the invertibility of $\boldsymbol{A}_{k,i}^d$, we can define the smoothers for the indefinite problem. Jacobi-type smoothers $\boldsymbol{J}_k^{\mathrm{I}}$ and $\boldsymbol{J}_k^{\mathrm{II}}$ are given by

$$\boldsymbol{J}_k^d = \gamma \sum_{i=0}^{N_k^d} (\boldsymbol{A}_{k,i}^d)^{-1} \boldsymbol{Q}_{k,i}, \qquad d \in \{\mathrm{I}, \mathrm{II}\}, \tag{3.10}$$

where γ is a scaling factor. Gauss–Seidel-type smoothers $\boldsymbol{G}_k^d$ for $d \in \{\mathrm{I}, \mathrm{II}\}$ are defined by the following algorithm.

ALGORITHM 3.1 (indefinite Gauss–Seidel). *Let $\boldsymbol{f}$ be in M_k. We define $\boldsymbol{G}_k^d$ by the following:*

(1) *Set $\boldsymbol{v}_0 = 0 \in M_k$.*

(2) *Define $\boldsymbol{v}_i$, for $i = 1, \ldots, N_k^d$, by*

$$\boldsymbol{v}_i = \boldsymbol{v}_{i-1} + (\boldsymbol{A}_{k,i}^d)^{-1} \boldsymbol{Q}_{k,i}^d (\boldsymbol{f} - \boldsymbol{A}_k \boldsymbol{v}_{i-1}).$$

(3) *Set $\boldsymbol{G}_k^d \boldsymbol{f} = \boldsymbol{v}_{N_k^d}$.*

The analogous Jacobi and Gauss–Seidel smoothers were given in [1] and [17] for the positive definite operators $\boldsymbol{\Lambda}_k$. These are denoted here by $\widetilde{\boldsymbol{J}}_k^d$ and $\widetilde{\boldsymbol{G}}_k^d$ and are again defined by (3.10) and Algorithm 3.1, respectively, but with $\boldsymbol{\Lambda}$ in place of $\boldsymbol{A}$. The scaling factor γ in (3.10) is chosen such that the $\boldsymbol{\Lambda}$-norm of $\boldsymbol{I} - \widetilde{\boldsymbol{J}}_k^d \boldsymbol{\Lambda}_k$ is less than or equal to one for $k = 2, \ldots, J$. Such a γ can be chosen independent of J by the limited overlap property of the subspaces.

Remark 3.3. In implementation, the application of the operator $(\boldsymbol{A}_{k,i}^d)^{-1} \boldsymbol{Q}_{k,i}^d$ reduces to solving a linear system involving the stiffness matrix associated with the indefinite form $\boldsymbol{A}(\cdot, \cdot)$, and the Gramm matrix inversion corresponding to $\boldsymbol{Q}_{k,i}^d$ is avoided.

4. Analysis of the multigrid iteration. In this section we provide an analysis of the multigrid iteration of section 2. This analysis is based on the product representation of the error operator (2.8). As done in [6] for second order elliptic problems, our analysis is based on perturbation from the uniform multigrid convergence estimates for a related symmetric positive definite problem.

We start with the estimate for the positive definite problem. For operators on M_k, $k = 1, \ldots, J$, we will use $\|\cdot\|_{\boldsymbol{\Lambda}}$ to denote the operator norm induced by the vector norm $\boldsymbol{\Lambda}(\cdot, \cdot)^{1/2}$. Set $\widetilde{\boldsymbol{R}}_k$ to be any one of $\widetilde{\boldsymbol{J}}_k^{\mathrm{I}}, \widetilde{\boldsymbol{J}}_k^{\mathrm{II}}, \widetilde{\boldsymbol{G}}_k^{\mathrm{I}}$, and $\widetilde{\boldsymbol{G}}_k^{\mathrm{II}}$. Let $\widetilde{\boldsymbol{T}}_k = \widetilde{\boldsymbol{R}}_k \boldsymbol{\Lambda}_k \widetilde{\boldsymbol{P}}_k$ for $k > 1$ and $\widetilde{\boldsymbol{T}}_1 = \widetilde{\boldsymbol{P}}_1$. Consider Algorithm 2.1 with $\boldsymbol{\Lambda}_k$ in place of $\boldsymbol{A}_k$ and $\widetilde{\boldsymbol{R}}_k$ in place of $\boldsymbol{R}_k$. Its error reduction operator is

$$\widetilde{\boldsymbol{E}} = (\boldsymbol{I} - \widetilde{\boldsymbol{T}}_1)(\boldsymbol{I} - \widetilde{\boldsymbol{T}}_2) \cdots (\boldsymbol{I} - \widetilde{\boldsymbol{T}}_J). \tag{4.1}$$

The following result is contained in [1, Theorems 3.1 and 4.2], [17, Theorem 3.1], and [18, Theorem 5.4].

THEOREM 4.1. *The multigrid error reduction operator in the case of the positive definite problem satisfies*

$$\boldsymbol{\Lambda}(\widetilde{\boldsymbol{E}}\boldsymbol{u}, \widetilde{\boldsymbol{E}}\boldsymbol{u}) \leq \hat{\delta}^2 \boldsymbol{\Lambda}(\boldsymbol{u}, \boldsymbol{u}) \quad \text{for all } \boldsymbol{u} \in M_J, \tag{4.2}$$

with $0 < \hat{\delta} < 1$ independent of J.

Remark 4.1. Although the results in [1] are formulated only for symmetric smoothers, let us verify that (4.2) holds for the nonsymmetric Gauss–Seidel smoother

$\widetilde{\boldsymbol{G}}_k^{\mathrm{I}}$ as well, as stated in Theorem 4.1. Indeed, we can be more general and consider instead the smoothing operator $\widetilde{\boldsymbol{R}}_k$ of a block successive overrelaxation iteration (SOR(α)) with a relaxation parameter $0 < \alpha < 2$ (with the blocks based on $\{M_{k,i}^d\}$, $d \in \{\mathrm{I}, \mathrm{II}\}$). We appeal to [9, Lemma 2.2], which shows that (4.2) holds for the $\widetilde{\boldsymbol{E}}$ obtained by any $\widetilde{\boldsymbol{R}}_k$, provided

$$\|\boldsymbol{I} - \widetilde{\boldsymbol{R}}_k \boldsymbol{\Lambda}_k\|_{\boldsymbol{\Lambda}} \le 1 \quad \text{and} \tag{4.3}$$

$$(\underline{\widetilde{\boldsymbol{R}}}_k^{-1} \boldsymbol{u}, \boldsymbol{u}) \le C\boldsymbol{\Lambda}(\boldsymbol{u}, \boldsymbol{u}) \quad \text{for all } \boldsymbol{u} \in (\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{k-1}) M_k, \tag{4.4}$$

where $\underline{\widetilde{\boldsymbol{R}}}_k = \widetilde{\boldsymbol{R}}_k + \widetilde{\boldsymbol{R}}_k^t - \widetilde{\boldsymbol{R}}_k^t \boldsymbol{\Lambda}_k \widetilde{\boldsymbol{R}}_k$. Here $\widetilde{\boldsymbol{R}}_k^t$ is the L^2-adjoint of $\widetilde{\boldsymbol{R}}_k$. That inequality (4.3) holds for the $\widetilde{\boldsymbol{R}}_k$ of SOR(α) follows immediately from the product representation,

$$\boldsymbol{I} - \widetilde{\boldsymbol{R}}_k \boldsymbol{\Lambda}_k = (\boldsymbol{I} - \alpha \widetilde{\boldsymbol{P}}_{k,N_k}^{\mathrm{I}}) \cdots (\boldsymbol{I} - \alpha \widetilde{\boldsymbol{P}}_{k,1}^{\mathrm{I}}).$$

It remains to see that (4.4) holds for this smoother. Techniques in [1] can be used to prove

$$\inf_{\{\boldsymbol{u}_i\}} \sum_{i=1}^{N_k^d} \boldsymbol{\Lambda}(\boldsymbol{u}_i, \boldsymbol{u}_i) \le C\boldsymbol{\Lambda}(\boldsymbol{u}, \boldsymbol{u}) \quad \text{for all } \boldsymbol{u} \in (\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{k-1}) M_k,$$

where the infimum is taken over all decompositions $\boldsymbol{u} = \sum_{i=1}^{N_k^d} \boldsymbol{u}_i$ such that $\boldsymbol{u}_i \in M_{k,i}^d$. It can be shown as in [7, Theorem 2.2] that for the $\widetilde{\boldsymbol{R}}_k$ of SOR(α),

$$(\underline{\widetilde{\boldsymbol{R}}}_k^{-1} \boldsymbol{u}, \boldsymbol{u}) \le \frac{(1 + c\alpha)^2}{2 - \alpha} \inf_{\{\boldsymbol{u}_i\}} \sum_{i=1}^{N_k^d} \boldsymbol{\Lambda}(\boldsymbol{u}_i, \boldsymbol{u}_i) \quad \text{for all } \boldsymbol{u} \in M_k.$$

Thus, Theorem 4.1 holds for the SOR(α) smoother.

We will analyze the multigrid algorithm by examining the difference between $\boldsymbol{E}$ and $\widetilde{\boldsymbol{E}}$. Let $\boldsymbol{Z}_k = \boldsymbol{T}_k - \widetilde{\boldsymbol{T}}_k$, and suppose we have

$$\|\boldsymbol{Z}_1\|_{\boldsymbol{\Lambda}} \le \epsilon \qquad \text{and} \tag{4.5}$$

$$\|\boldsymbol{Z}_k\|_{\boldsymbol{\Lambda}} \le C_1 h_k \quad \text{for } k = 2, \ldots, J. \tag{4.6}$$

Then, it can be shown that the difference $\boldsymbol{E}_k - \widetilde{\boldsymbol{E}}_k$ is small by an argument of [6] (see also [8, Lemma 11.1]). We include the argument here for the sake of completeness: First, note that by the triangle inequality, the $\boldsymbol{\Lambda}$-norm of $(\boldsymbol{I} - \boldsymbol{T}_k) = (\boldsymbol{I} - \widetilde{\boldsymbol{T}}_k - \boldsymbol{Z}_k)$ is less than or equal to $1 + ch_k$. Therefore,

$$\|\boldsymbol{E}_k\|_{\boldsymbol{\Lambda},\Omega} \le (1 + c\epsilon) \prod_{i=2}^{k} (1 + ch_i),$$

which can be bounded by a convergent infinite product. Thus $\|\boldsymbol{E}_k\|_{\boldsymbol{\Lambda},\Omega} \le C$.

To continue, we observe the following recursion:

$$\boldsymbol{E}_k - \widetilde{\boldsymbol{E}}_k = (\boldsymbol{E}_{k-1} - \widetilde{\boldsymbol{E}}_{k-1})(\boldsymbol{I} - \widetilde{\boldsymbol{T}}_k) - \boldsymbol{E}_{k-1} \boldsymbol{Z}_k, \tag{4.7}$$

which implies that for $k > 1$,

$$\begin{aligned} \|\boldsymbol{E}_k - \widetilde{\boldsymbol{E}}_k\|_{\boldsymbol{\Lambda},\Omega} &\le \|\boldsymbol{E}_{k-1} - \widetilde{\boldsymbol{E}}_{k-1}\|_{\boldsymbol{\Lambda},\Omega} \|\boldsymbol{I} - \widetilde{\boldsymbol{T}}_k\|_{\boldsymbol{\Lambda},\Omega} + \|\boldsymbol{E}_{k-1}\|_{\boldsymbol{\Lambda},\Omega} \|\boldsymbol{Z}_k\|_{\boldsymbol{\Lambda},\Omega} \\ &\le \|\boldsymbol{E}_{k-1} - \widetilde{\boldsymbol{E}}_{k-1}\|_{\boldsymbol{\Lambda},\Omega} + Ch_k. \end{aligned}$$

Repeated application of this inequality shows that the difference $\boldsymbol{E}_k - \widetilde{\boldsymbol{E}}_k$ is small:

$$\|\boldsymbol{E}_J - \widetilde{\boldsymbol{E}}_J\|_{\boldsymbol{\Lambda},\Omega} \le c(h_1 + \epsilon).$$

Thus, we have proven the following theorem.

THEOREM 4.2. *Let* $\boldsymbol{E}$ *satisfy* (2.8) *and* $\widetilde{\boldsymbol{E}}$ *satisfy* (4.1). *Assume that* (4.5) *and* (4.6) *hold. Then there are positive constants* C, $\hat{h}_1$, *and* $\hat{\epsilon}$ *depending only on* C_1 *above such that if* $h_1 \le \hat{h}_1$ *and* $\epsilon \le \hat{\epsilon}$,

$$\|\boldsymbol{E}\|_{\boldsymbol{\Lambda}} \le \|\widetilde{\boldsymbol{E}}\|_{\boldsymbol{\Lambda}} + C(h_1 + \epsilon).$$

In (4.5) and (4.6), the operator norm of $\boldsymbol{Z}_k$ can be taken to be that of $\boldsymbol{Z}_k : M_J \mapsto M_k$ or $\boldsymbol{Z}_k : M_k \mapsto M_k$, as both norms are equal. The proofs of our main results proceed by verifying (4.5) and (4.6). In the verification of (4.5), the nature of the subspace decompositions is immaterial, and a coarse grid estimate of [16] is critical, as seen in the following lemma.

LEMMA 4.3. *There exists* $H > 0$ *such that if* $h_1 \le H$, *then* (4.5) *holds with* $\epsilon = ch_1$.

Proof. For $\boldsymbol{u}, \boldsymbol{v} \in M_J$, the following identity holds:

$$\begin{aligned} \boldsymbol{\Lambda}(\boldsymbol{Z}_1\boldsymbol{u}, \boldsymbol{v}) &= \boldsymbol{\Lambda}(\boldsymbol{P}_1\boldsymbol{u} - \boldsymbol{u}, \widetilde{\boldsymbol{P}}_1\boldsymbol{v}) \\ &= \boldsymbol{A}(\boldsymbol{P}_1\boldsymbol{u} - \boldsymbol{u}, \widetilde{\boldsymbol{P}}_1\boldsymbol{v}) + (\omega^2 + 1)(\boldsymbol{P}_1\boldsymbol{u} - \boldsymbol{u}, \widetilde{\boldsymbol{P}}_1\boldsymbol{v}) \\ &= (\omega^2 + 1)(\boldsymbol{P}_1\boldsymbol{u} - \boldsymbol{u}, \widetilde{\boldsymbol{P}}_1\boldsymbol{v}). \end{aligned} \tag{4.8}$$

It is shown in [16], using a duality argument utilizing the regularity assumption, that there exists $H > 0$ such that if $h_1 \le H$, then

$$(\boldsymbol{u} - \boldsymbol{P}_1\boldsymbol{u}, \boldsymbol{w}) \le Ch_1 \|\boldsymbol{u} - \boldsymbol{P}_1\boldsymbol{u}\|_{\boldsymbol{\Lambda}} \|\boldsymbol{w}\|_{\boldsymbol{\Lambda}}$$

for all $\boldsymbol{u} \in M_J$ and $\boldsymbol{w} \in M_1$. Thus, the lemma follows. ∎

While verifying (4.6) for specific smoothers, it will be useful to have bounds for the perturbation operators $\boldsymbol{Z}^d_{k,i} : M_J \mapsto M^d_{k,i}$, $d \in \{\mathrm{I}, \mathrm{II}\}$, defined by

$$\boldsymbol{Z}^d_{k,i} = \boldsymbol{P}^d_{k,i} - \widetilde{\boldsymbol{P}}^d_{k,i}.$$

Note that in the case of subspaces of gradients of decomposition II,

$$\boldsymbol{Z}^{\mathrm{II}}_{k,i} = 0 \quad \text{for } i = n^M_k + 1, \ldots, N^{\mathrm{II}}_k.$$

An identity similar to (4.8) can be obtained for $\boldsymbol{Z}^d_{k,i}$:

$$\boldsymbol{\Lambda}(\boldsymbol{Z}^d_{k,i}\boldsymbol{u}, \boldsymbol{v}) = -(\omega^2 + 1)(\boldsymbol{u} - \boldsymbol{P}^d_{k,i}\boldsymbol{u}, \widetilde{\boldsymbol{P}}^d_{k,i}\boldsymbol{v}). \tag{4.9}$$

LEMMA 4.4. *There exists* $H > 0$ *such that if* $h_1 \le H$,

$$(\boldsymbol{u} - \boldsymbol{P}^d_{k,i}\boldsymbol{u}, \boldsymbol{v}_{k,i}) \le Ch_k \|\boldsymbol{u} - \boldsymbol{P}^d_{k,i}\boldsymbol{u}\|_{0,\Omega^d_{k,i}} \|\mathbf{curl}\, \boldsymbol{v}_{k,i}\|_{0,\Omega^d_{k,i}}$$

for all $\boldsymbol{u} \in M_J$ *and* $\boldsymbol{v}_{k,i} \in M^d_{k,i}$, $d \in \{\mathrm{I}, \mathrm{II}\}$, $k = 2, \ldots, J$.

Proof. In the case $d = \mathrm{I}$, observe that for any $\boldsymbol{u} \in M_J$, $\boldsymbol{u} - \boldsymbol{P}^{\mathrm{I}}_{k,i}\boldsymbol{u}$ is L^2-orthogonal to functions of the form $\mathbf{grad}\, w$ for any $w \in W^{\mathrm{I}}_{k,i}$. Decomposing $\boldsymbol{v}_{k,i} = \mathbf{grad}\, w + \boldsymbol{x}$, where $w \in W^{\mathrm{I}}_{k,i}$ and $\boldsymbol{x} \in \mathring{M}^{\mathrm{I}}_{k,i}$, and applying Lemma 3.1 give

$$\begin{aligned} (\boldsymbol{u} - \boldsymbol{P}_{k,i}\boldsymbol{u}, \boldsymbol{v}_{k,i}) = (\boldsymbol{u} - \boldsymbol{P}_{k,i}\boldsymbol{u}, \boldsymbol{x}) &\le Ch_k \|\boldsymbol{u} - \boldsymbol{P}_{k,i}\boldsymbol{u}\|_{0,\Omega_{k,i}} \|\mathbf{curl}\, \boldsymbol{x}\|_{0,\Omega_{k,i}} \\ &= Ch_k \|\boldsymbol{u} - \boldsymbol{P}_{k,i}\boldsymbol{u}\|_{0,\Omega_{k,i}} \|\mathbf{curl}\, \boldsymbol{v}_{k,i}\|_{0,\Omega_{k,i}}. \end{aligned}$$

In the case $d = \mathrm{II}$, the result immediately follows from Cauchy–Schwarz inequality and Lemma 3.1 for $\boldsymbol{v}_{k,i} \in \mathring{M}^{\mathrm{II}}_{k,i}$. For the remaining $\boldsymbol{v}_{k,i} \in M^{\mathrm{II}}_{k,i}$, both sides of the inequality of the lemma are zero. ∎

The following theorem is our main result.

THEOREM 4.5. *In Algorithm* 2.1*, set* $\boldsymbol{R}_k$ *to any of the smoothers* $\boldsymbol{J}^{\mathrm{I}}_k, \boldsymbol{G}^{\mathrm{I}}_k, \boldsymbol{J}^{\mathrm{II}}_k$*, and* $\boldsymbol{G}^{\mathrm{II}}_k$ *defined earlier. Then there exists an* $H > 0$ *such that whenever* $h_1 \le H$*,*

$$\boldsymbol{\Lambda}(\boldsymbol{E}\boldsymbol{u}, \boldsymbol{E}\boldsymbol{u}) \le \delta^2 \boldsymbol{\Lambda}(\boldsymbol{u}, \boldsymbol{u}) \quad \text{for all } \boldsymbol{u} \in M_J,$$

for $\delta = \hat{\delta} + ch_1$*. Here* $\hat{\delta}$ *is less than one (and independent of* J*) and is given by Theorem* 4.1 *applied to the corresponding smoother* $\widetilde{\boldsymbol{R}}_k$*. In addition,* c *is independent of* h_1*.*

Proof. We apply Theorem 4.2. By Lemma 4.3, we need only verify (4.6) for each of the smoothers $\boldsymbol{J}^{\mathrm{I}}_k, \boldsymbol{G}^{\mathrm{I}}_k, \boldsymbol{J}^{\mathrm{II}}_k$, and $\boldsymbol{G}^{\mathrm{II}}_k$. Since the proof for the case of the latter two smoothers is completely analogous to that for the case of the smoothers based on decomposition I, we give only the proof for $\boldsymbol{J}^{\mathrm{I}}_k$ and $\boldsymbol{G}^{\mathrm{I}}_k$.

In the case of $\boldsymbol{R}_k = \boldsymbol{J}^{\mathrm{I}}_k$, the perturbation operator $\boldsymbol{Z}_k$, $k > 1$, satisfies

$$\boldsymbol{Z}_k \boldsymbol{u} = \gamma \sum_{i=1}^{N_k} (\boldsymbol{P}^{\mathrm{I}}_{k,i} - \widetilde{\boldsymbol{P}}^{\mathrm{I}}_{k,i})\boldsymbol{u} = \gamma \sum_{i=1}^{N_k} \boldsymbol{Z}^{\mathrm{I}}_{k,i} \boldsymbol{u}$$

for any $\boldsymbol{u} \in M_k$. By (4.9), Lemma 4.4, and (3.9),

$$\boldsymbol{\Lambda}(\boldsymbol{Z}^{\mathrm{I}}_{k,i}\boldsymbol{u}, \boldsymbol{v}) = (\omega^2 + 1)(\boldsymbol{P}^{\mathrm{I}}_{k,i}\boldsymbol{u} - \boldsymbol{u}, \widetilde{\boldsymbol{P}}^{\mathrm{I}}_{k,i}\boldsymbol{v}) \le ch_k \, \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}, \Omega^{\mathrm{I}}_{k,i}} \, \|\boldsymbol{v}\|_{\boldsymbol{\Lambda}, \Omega^{\mathrm{I}}_{k,i}} \tag{4.10}$$

for any $\boldsymbol{u}, \boldsymbol{v} \in M_k$. Hence,

$$\boldsymbol{\Lambda}(\boldsymbol{Z}_k \boldsymbol{u}, \boldsymbol{v}) \le ch_k \sum_{i=1}^{N_k} \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}, \Omega^{\mathrm{I}}_{k,i}} \, \|\boldsymbol{v}\|_{\boldsymbol{\Lambda}, \Omega^{\mathrm{I}}_{k,i}}.$$

The inequality (4.6) now easily follows using the limited overlap properties of the domains $\Omega^{\mathrm{I}}_{k,i}$. This completes the proof of the theorem when $\boldsymbol{R}_k = \boldsymbol{J}^{\mathrm{I}}_k$.

Now consider the case $\boldsymbol{R}_k = \boldsymbol{G}^{\mathrm{I}}_k$. As before, it suffices to verify (4.6). Define $\widetilde{\boldsymbol{\mathcal{E}}}_i$ and $\boldsymbol{\mathcal{E}}_i$ by

$$\begin{aligned} \widetilde{\boldsymbol{\mathcal{E}}}_i &= (\boldsymbol{I} - \widetilde{\boldsymbol{P}}^{\mathrm{I}}_{k,i})(\boldsymbol{I} - \widetilde{\boldsymbol{P}}^{\mathrm{I}}_{k,i-1}) \cdots (\boldsymbol{I} - \widetilde{\boldsymbol{P}}^{\mathrm{I}}_{k,1}) \text{ and} \\ \boldsymbol{\mathcal{E}}_i &= (\boldsymbol{I} - \boldsymbol{P}^{\mathrm{I}}_{k,i})(\boldsymbol{I} - \boldsymbol{P}^{\mathrm{I}}_{k,i-1}) \cdots (\boldsymbol{I} - \boldsymbol{P}^{\mathrm{I}}_{k,1}), \end{aligned}$$

and let $\widetilde{\boldsymbol{\mathcal{E}}}_0 = \boldsymbol{\mathcal{E}}_0 = \boldsymbol{I}$. Then the perturbation operator $\boldsymbol{Z}_k : M_k \mapsto M_k$ for this example is

$$\boldsymbol{Z}_k = \boldsymbol{T}_k - \widetilde{\boldsymbol{T}}_k = \widetilde{\boldsymbol{\mathcal{E}}}_{N_k} - \boldsymbol{\mathcal{E}}_{N_k}.$$

We clearly have that

$$\widetilde{\boldsymbol{\mathcal{E}}}_i - \boldsymbol{\mathcal{E}}_i = (\boldsymbol{I} - \widetilde{\boldsymbol{P}}^{\mathrm{I}}_{k,i})(\widetilde{\boldsymbol{\mathcal{E}}}_{i-1} - \boldsymbol{\mathcal{E}}_{i-1}) - \boldsymbol{Z}^{\mathrm{I}}_{k,i} \boldsymbol{\mathcal{E}}_{i-1}.$$

Since the terms on the right are orthogonal with respect to $\boldsymbol{\Lambda}(\cdot, \cdot)$,

$$\|(\widetilde{\boldsymbol{\mathcal{E}}}_i - \boldsymbol{\mathcal{E}}_i)\boldsymbol{u}\|^2_{\boldsymbol{\Lambda}, \Omega} = \|(\boldsymbol{I} - \widetilde{\boldsymbol{P}}^{\mathrm{I}}_{k,i})(\widetilde{\boldsymbol{\mathcal{E}}}_{i-1} - \boldsymbol{\mathcal{E}}_{i-1})\boldsymbol{u}\|^2_{\boldsymbol{\Lambda}, \Omega} + \|\boldsymbol{Z}^{\mathrm{I}}_{k,i} \boldsymbol{\mathcal{E}}_{i-1} \boldsymbol{u}\|^2_{\boldsymbol{\Lambda}, \Omega}.$$

It follows from (4.10) that $\|\boldsymbol{Z}^{\mathrm{I}}_{k,i}\boldsymbol{v}\|_{\boldsymbol{\Lambda},\Omega} \le Ch_k\|\boldsymbol{v}\|_{\boldsymbol{\Lambda},\Omega^{\mathrm{I}}_{k,i}}$. This and the fact that the $\boldsymbol{\Lambda}$-operator norm of $(\boldsymbol{I}-\widetilde{\boldsymbol{P}}^{\mathrm{I}}_{k,i})$ is bounded by one imply that

$$\|(\widetilde{\boldsymbol{\mathcal{E}}}_i-\boldsymbol{\mathcal{E}}_i)\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} \le \|(\widetilde{\boldsymbol{\mathcal{E}}}_{i-1}-\boldsymbol{\mathcal{E}}_{i-1})\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} + Ch_k^2\|\boldsymbol{\mathcal{E}}_{i-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega_{k,i}}.$$

Summing over i and obvious manipulations give

$$\|(\widetilde{\boldsymbol{\mathcal{E}}}_{N_k}-\boldsymbol{\mathcal{E}}_{N_k})\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} \le Ch_k^2\sum_{i=1}^{N_k}\|\boldsymbol{\mathcal{E}}_{i-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega_{k,i}}. \tag{4.11}$$

We shall now show that for sufficiently small h_1,

$$\sum_{i=1}^{N_k}\|\boldsymbol{\mathcal{E}}_{i-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega_{k,i}} \le C\|\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega}. \tag{4.12}$$

We first note the identity

$$\boldsymbol{I}-\boldsymbol{\mathcal{E}}_i = \sum_{m=1}^{i}\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}.$$

Thus, by the arithmetic-geometric mean inequality, the definition of $\boldsymbol{\mathcal{E}}_i$, and the limited interaction property, it follows that

$$\begin{aligned}
\sum_{i=1}^{N_k}\|\boldsymbol{\mathcal{E}}_{i-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega^{\mathrm{I}}_{k,i}} &\le 2\sum_{i=1}^{N_k}\|\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega^{\mathrm{I}}_{k,i}} + 2\sum_{i=1}^{N_k}\|\boldsymbol{u}-\boldsymbol{\mathcal{E}}_{i-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega^{\mathrm{I}}_{k,i}}\\
&\le C\|\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} + 2\sum_{i=1}^{N_k}\left\|\sum_{m=1}^{i-1}\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\right\|^2_{\boldsymbol{\Lambda},\Omega^{\mathrm{I}}_{k,i}}\\
&\le C\left(\|\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} + \sum_{m=1}^{N_k}\sum_{i=1}^{N_k}\|\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega^{\mathrm{I}}_{k,i}}\right)\\
&\le C\left(\|\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} + \sum_{m=1}^{N_k}\|\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega}\right).
\end{aligned} \tag{4.13}$$

In order to estimate the last term on the right of (4.13), we write

$$\begin{aligned}
\|\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} &= \|\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} - \|\boldsymbol{\mathcal{E}}_m\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega}\\
&\quad - 2\boldsymbol{\Lambda}(\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u},(\boldsymbol{I}-\boldsymbol{P}^{\mathrm{I}}_{k,m})\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}).
\end{aligned} \tag{4.14}$$

Now by (4.9),

$$\boldsymbol{\Lambda}(\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u},(\boldsymbol{I}-\boldsymbol{P}^{\mathrm{I}}_{k,m})\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}) = (1+\omega^2)(\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u},(\boldsymbol{I}-\boldsymbol{P}^{\mathrm{I}}_{k,m})\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}),$$

so by Lemma 4.4 we have

$$\|\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} \le C(\|\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} - \|\boldsymbol{\mathcal{E}}_m\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega}) + Ch_k^2\|\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega^{\mathrm{I}}_{k,m}}.$$

Summing over m, we conclude that

$$\sum_{m=1}^{N_k} \|\boldsymbol{P}^{\mathrm{I}}_{k,m}\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} \le C\left(\|\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega} + h_k^2 \sum_{m=1}^{N_k} \|\boldsymbol{\mathcal{E}}_{m-1}\boldsymbol{u}\|^2_{\boldsymbol{\Lambda},\Omega^{\mathrm{I}}_{k,m}}\right). \tag{4.15}$$

Clearly (4.15) and (4.13) yield (4.12) for small enough h_1.

Finally, we obtain from (4.12) and (4.11) that for $k > 1$,

$$\|\boldsymbol{Z}_k\|_{\boldsymbol{\Lambda},\Omega} \le Ch_k.$$

The theorem follows from Lemma 4.3 and Theorem 4.2. ∎

Remark 4.2. The same analysis could be used for the SOR(α) iteration considered in Remark 4.1. In that case,

$$\boldsymbol{\mathcal{E}}_l = (\boldsymbol{I} - \alpha\boldsymbol{P}^d_{k,l})(\boldsymbol{I} - \alpha\boldsymbol{P}^d_{k,l-1})\cdots(\boldsymbol{I} - \alpha\boldsymbol{P}^d_{k,1}).$$

Also, by Remark 4.1, Theorem 4.1 holds with the SOR(α) smoother.

5. Numerical results. Numerical experiments were conducted using lowest order Nedelec elements on cubes. We report results of some of these experiments in this section. First, let us note that not only can Algorithm 2.1 be used as a linear solver for (2.6), but it can also be used to develop a preconditioner. Specifically, the operator $\boldsymbol{B}_J : M_J \mapsto M_J$ defined by $\boldsymbol{B}_J\boldsymbol{g} = \boldsymbol{Mg}_J(\boldsymbol{0}, \boldsymbol{g})$ is a preconditioner for $\boldsymbol{A}_J$ in the sense that the inequalities

$$\begin{aligned} &(1-\delta)\,\boldsymbol{\Lambda}(\boldsymbol{u},\boldsymbol{u}) \le \boldsymbol{\Lambda}(\boldsymbol{B}_J\boldsymbol{A}_J\boldsymbol{u},\boldsymbol{u}) \quad \text{and} \\ &\boldsymbol{\Lambda}(\boldsymbol{B}_J\boldsymbol{A}_J\boldsymbol{u},\boldsymbol{v}) \le (1+\delta)\boldsymbol{\Lambda}(\boldsymbol{u},\boldsymbol{u})^{1/2}\boldsymbol{\Lambda}(\boldsymbol{v},\boldsymbol{v})^{1/2} \end{aligned} \tag{5.1}$$

hold for all $\boldsymbol{u}, \boldsymbol{v} \in M_J$, for sufficiently small coarse mesh sizes. These bounds easily follow from Theorem 4.5, and δ is as in the theorem. They imply that when GMRES in the $\boldsymbol{\Lambda}(\cdot,\cdot)$ innerproduct is used to solve (2.6) with $\boldsymbol{B}_J$ as preconditioner, the number of iterations remains bounded independently of refinement level [12, 16]. In this section we will investigate the performance of $\boldsymbol{B}_J$ as a preconditioner for use in GMRES as well as that of the linear solver $\boldsymbol{Mg}_J(\cdot,\cdot)$ given by Algorithm 2.1.

In all experiments, our computational domain was $\Omega = (0,1)^3$. We investigate only the multigrid algorithm with the smoother $\boldsymbol{G}^{\mathrm{I}}_k$ based on decomposition I. The domain $(0,1)^3$ was meshed by a hierarchy of multilevel uniform cubic meshes. Each mesh is obtained by breaking up every cubic element of a coarser mesh into eight congruent cubes, the coarsest mesh being just $\{\Omega\}$. Clearly, our analysis holds in this situation. (In particular, a Poincaré–Friedrichs inequality like that of Lemma 3.1 is obvious for uniform cubic meshes.)

The linear system (2.6) is solved on a fine ($k = J$) mesh of mesh size h using one of the two above-mentioned iterative methods. The coarse solves of the multigrid algorithm are done on a coarse ($k = 1$) mesh of mesh size H. All coarse solves were done by direct methods of UMFPACK2.2 [10]. The right-hand side of (2.6) was chosen so that the true solution equals the interpolant of $\boldsymbol{U}(x,y,z) = [y(1-y)z(1-z), yx(1-x)z(1-z), x(1-x)y(1-y)]$. We report iteration counts for a set of combinations of h and H. The starting iterate was always zero. When the linear multigrid solver was used, the stopping criterion was that the $\boldsymbol{\Lambda}$-norm of error be reduced by a factor of 10^{-6}. The stopping criterion for GMRES was that the $\boldsymbol{\Lambda}$-norm of the residual (premultiplied by $\boldsymbol{B}_J$) was reduced by a factor of 10^{-6}. GMRES was set to restart after 50 iterations.

TABLE 5.1
Preconditioned GMRES iteration counts for the $\omega = 1$ case. Degrees of freedom at each refinement level are also shown in the last column.

		H					Degrees of
		1/2	1/4	1/8	1/16	1/32	of freedom
	1/4	6	–	–	–	–	108
	1/8	7	7	–	–	–	1176
h	1/16	9	9	8	–	–	10800
	1/32	10	10	9	7	–	92256
	1/64	11	10	10	8	7	762048
	1/128	11	11	10	9	8	6193536

TABLE 5.2
Linear multigrid iteration counts with $\omega = 1$.

		H				
		1/2	1/4	1/8	1/16	1/32
	1/4	6	–	–	–	–
	1/8	7	7	–	–	–
h	1/16	9	9	8	–	–
	1/32	10	11	10	8	–
	1/64	11	11	10	10	8
	1/128	12	11	10	10	10

TABLE 5.3
Linear multigrid iteration counts for $\omega = 7$. An entry "$\star$" indicates that the $\mathbf{\Lambda}$-norm of iterates became larger than 10^{99}, and iterations were stopped.

		H				
		1/2	1/4	1/8	1/16	1/32
	1/4	$\star$	–	–	–	–
	1/8	$\star$	35	–	–	–
h	1/16	$\star$	110	23	–	–
	1/32	$\star$	208	48	10	–
	1/64	$\star$	266	62	15	8
	1/128	$\star$	285	67	16	10

We start with the case $\omega = 1$. GMRES iteration counts are reported in Table 5.1. The preconditioner appears to be uniform, as iteration counts never exceeded 11 for all combinations of h and H we considered. For comparison, the case $h = 1/128$ without preconditioner did not converge even after 5000 iterations.

Iteration counts obtained using the linear multigrid solver are reported in Table 5.2, and these are in accordance with Theorem 4.5. Although, in the case $\omega = 1$, the algorithm gives uniform iteration counts for all choices of H considered, this is no longer the case for a higher wave number, as seen in Table 5.3. This is again in accordance with Theorem 4.5, as its conclusion holds only whenever the coarse mesh is sufficiently fine.

We have also considered the performance of $\boldsymbol{B}_J$ as a preconditioner in GMRES for the case of a higher wave number $\omega = 10$. It is a good preconditioner only for smaller coarse mesh sizes, as Table 5.4 shows. In other (unreported) experiments, the linear multigrid algorithm for this wave number converged only for one of the combinations of h and H considered. Theoretically, (5.1) guarantees that $\boldsymbol{B}_J$ is a good preconditioner only when $\boldsymbol{Mg}_J(\cdot,\cdot)$ is a good contraction. Nonetheless, our experiments indicate that the coarse mesh size at which $\boldsymbol{B}_J$ becomes a good preconditioner is larger than that required for $\boldsymbol{Mg}_J(\cdot,\cdot)$ to be a good contraction. Similar observations have been

TABLE 5.4

Preconditioned GMRES iteration counts for $\omega = 10$ case. An entry of the form $n^{\times}$ indicates that although the residual of the nth GMRES iterate met the stopping criterion, the iterate differed from the true solution by more than 10^{-3} in the $\mathbf{\Lambda}$-norm.

		H				
		1/2	1/4	1/8	1/16	1/32
	1/4	$3^{\times}$	–	–	–	–
	1/8	$2^{\times}$	37	–	–	–
h	1/16	$3^{\times}$	48	18	–	–
	1/32	$2^{\times}$	$78^{\times}$	22	16	–
	1/64	$2^{\times}$	$78^{\times}$	21	17	9
	1/128	$2^{\times}$	$79^{\times}$	21	16	10

TABLE 5.5

Numerical convergence rates for the linear multigrid iteration.

		H			
		1/2	1/4	1/8	1/16
	1/4	0.32	—	—	—
	1/8	0.40	0.40	—	—
h	1/16	0.42	0.42	0.42	—
	1/32	0.42	0.42	0.42	0.42
	1/64	0.42	0.42	0.42	0.42

$\omega = 1$

		H			
		1/2	1/4	1/8	1/16
	1/4	7.93	—	—	—
	1/8	9.67	0.92	—	—
	1/16	10.01	0.65	0.60	—
	1/32	10.06	0.58	0.44	0.43
	1/64	10.07	0.58	0.43	0.43

$\omega = 5$

made in studies of multigrid algorithms for the Helmholtz equation [13]. This may be an argument in favor of using GMRES preconditioned with multigrid as a solution strategy, rather than the linear multigrid solver. However, we must also keep in mind that if too large a mesh size is used, GMRES may find the residual too small and stop, even though the iterate is far from the true solution (see entries $n^{\times}$).

We conclude by providing numerical convergence rates for the linear multigrid iteration which also confirm our theoretical results. Entries of Table 5.5 provide estimates for $\|\boldsymbol{I} - \boldsymbol{B}_J \boldsymbol{A}_J\|_{\mathbf{\Lambda}}$ obtained by means of the power method in the cases $\omega = 1$ and $\omega = 5$ for a few combinations of h and H. We see that the only difference between the two cases is that larger ω requires a smaller coarse grid size. Note, though, that once the coarse grid is small enough, both cases give rise to approximately the same reduction rates.

Appendix. Here we will indicate how the main result of this paper can be generalized to higher order Nedelec spaces (of the first kind). Let M_k be defined with rth order Nedelec spaces on each tetrahedron, and let W_k be the corresponding conforming approximation space with polynomials of degree at most $r + 1$. The algorithms and definitions of subspace decompositions and smoothers generalize in an obvious way for the case of decomposition I. As we shall see, Case II can also be generalized provided a suitable choice of nodal basis is made.

Case I. First, note that Theorem 4.1 holds with the higher order spaces, as shown in [1]. The only proof in the previous sections that depended on the order of the spaces is that of Lemma 3.1. We will now prove that the inequality of the lemma holds for higher order spaces as well.

We start by considering the set S of all possible quasi-uniform tetrahedral meshes contained in the unit ball with at least one vertex on the unit sphere, every element having the origin as a vertex and the origin being an interior point of the mesh. Note that this implies that the resulting mesh domains are simply connected with only one

connected boundary component.

Each element in S is represented by a list of vertices and a list of tetrahedra (a tetrahedron number to vertex number list). We can assign labels to the members of S so that two members have the same label if and only if they have the same tetrahedron to vertex list. Quasi uniformity implies that the number of labels can be bounded in terms of ζ appearing in (2.4). Let R_l be the subset of elements of S with the lth label.

Any subdomain $\Omega^{\mathrm{I}}_{k,i}$ can be dilated and translated to an element of S. Thus, it suffices to prove that (3.4) holds for each D in R_l with constant independent of D. The general result holds, taking the minimum of these constants over $\{R_l\}$.

Clearly, each domain $D \in R_l$ has the same number of vertices, say, m. We can define a distance on R_l by using any norm on the vertex set, e.g., the Euclidean norm on $\mathbb{R}^{3m}$. It follows from quasi uniformity that R_l is a closed and bounded set in this norm and hence compact.

Let D be in R_l. Denote the corresponding approximation spaces (of $H_0(\mathbf{curl}\,; D)$ and $H^1_0(D)$, resp.) by M'_D and W'_D, and set $\mathring{M}'_D = \{\boldsymbol{q} \in M'_D : (\boldsymbol{q}, \mathbf{grad}\ \theta)_D = 0, \text{ for all } \theta \in W'_D\}$. Let

$$\mathcal{I}(D) = \inf_{\boldsymbol{q} \in \mathring{M}'_D} \frac{\|\mathbf{curl}\,\boldsymbol{q}\|_{0,D}}{\|\boldsymbol{q}\|_{0,D}}. \tag{A.1}$$

Note that since D is simply connected with a connected boundary, if $\mathbf{curl}\,\boldsymbol{q} = 0$ and $\boldsymbol{q} \in M'_D$, then $\boldsymbol{q}$ is a gradient of a function in W'_D. It follows that $\mathcal{I}(D) > 0$ for any $D \in R_l$. Thus, to prove that (3.4) holds uniformly for $D \in R_l$, it suffices to show that $\mathcal{I}(D)$ is continuous.

Suppose p and q are vertex sets of two meshes in R_l, with corresponding domains D_p and D_q, respectively. Let $\epsilon > 0$ be given. For $s \in \{p, q\}$, let $\{\boldsymbol{e}^s_i\}_{i=1}^{n_l}$ denote a nodal basis for M'_{D_s}. We identify functions in the above spaces with their extension by zero to the unit ball B. Let $\boldsymbol{z} \in \mathring{M}'_{D_p}$ be a function with $\|\boldsymbol{z}\|_{0,B} = 1$ for which the infimum in (A.1) is attained, and let

$$\boldsymbol{z} = \sum_{i=1}^{n_l} c_i \boldsymbol{e}^p_i \quad \text{and} \quad \boldsymbol{z}' = \sum_{i=1}^{n_l} c_i \boldsymbol{e}^q_i.$$

By quasi uniformity, it is easy to see that if $|p - q|$ is small enough (depending on ϵ), $\|\boldsymbol{z} - \boldsymbol{z}'\|_{\boldsymbol{\Lambda},B} \le \epsilon$. Note that $\boldsymbol{z}'$ is, in general, not in $\mathring{M}'_{D_q}$. Define $\psi' \in W'_{D_q}$ by

$$(\mathbf{grad}\ \psi', \mathbf{grad}\ \phi)_B = (\boldsymbol{z}', \mathbf{grad}\ \phi)_B \quad \text{for all } \phi \in W'_{D_q}.$$

Then $\boldsymbol{z}'' = \boldsymbol{z}' - \mathbf{grad}\ \psi'$ is in $\mathring{M}'_{D_q}$. Moreover, if $|p - q|$ is small enough, it can easily be shown that $\|\boldsymbol{z}'' - \boldsymbol{z}'\|_{\boldsymbol{\Lambda},B} \le \epsilon$, so

$$\|\boldsymbol{z} - \boldsymbol{z}''\|_{\boldsymbol{\Lambda},B} \le 2\epsilon.$$

Consequently,

$$\mathcal{I}(q) - \mathcal{I}(p) \le \frac{\|\mathbf{curl}\,\boldsymbol{z}''\|_{0,D_q}}{\|\boldsymbol{z}''\|_{0,D_q}} - \mathcal{I}(p) \le C\epsilon.$$

Interchanging the roles of p and q in the above argument, we also get that $\mathcal{I}(p) - \mathcal{I}(q) \le C\epsilon$. Thus, $\mathcal{I}(p)$ is continuous on R_l. This finishes the proof of Lemma 3.1 when $d = \mathrm{I}$.

Case II. Smoothing algorithms of this type can be generalized to higher order spaces provided a suitable choice of nodal basis is made. Note that there are choices of nodal basis for which Lemma 3.1 does not hold. We will provide one example of a nodal basis for which our analysis generalizes.

Once a set of degrees of freedom for M_k is defined, a corresponding nodal basis immediately follows. The particular choice of the degrees of freedom we have in mind consists of edge, face, and tetrahedral moments. For any domain D, let $P_l(D)$ denote the set of polynomials of degree at most l, and let $\mathcal{P}_l(D)$ denote any basis for $P_l(D)$. For every interior edge e and interior face f of the kth level mesh, define the edge and face moments

$$\alpha_e^p(\boldsymbol{u}) = \int_e p(\boldsymbol{u}\cdot\boldsymbol{t})\,dt, \quad \alpha_f^{\boldsymbol{q}}(\boldsymbol{u}) = \int_f \boldsymbol{q}\cdot(\boldsymbol{u}\times\boldsymbol{n})\,ds$$

for $p \in \mathcal{P}_{r-1}(e)$ and $\boldsymbol{q} \in (\mathcal{P}_{r-2}(f))^2$. The tetrahedral moments are defined by mapping to the reference tetrahedron $\hat{\tau}$ bounded by the planes $x = 0$, $y = 0$, $z = 0$, and $x + y + z = 1$. Let $\mathcal{R}_{r-3}$ be the set of all vector polynomials that are monomials of degree at most $r-3$ in one coordinate direction and zero in others; e.g., $\boldsymbol{r} = (x^i y^j z^k, 0, 0)$ is in $\mathcal{R}_{r-3}$. For every tetrahedron τ in the kth level mesh, define the tetrahedral moments

$$\alpha_\tau^{\boldsymbol{r}}(\boldsymbol{u}) = \int_{\hat{\tau}} \boldsymbol{r}\cdot\widehat{\boldsymbol{u}}\,dx,$$

where $\boldsymbol{r} \in \mathcal{R}_{r-3}$, $\widehat{\boldsymbol{u}}(\widehat{\boldsymbol{x}}) = B^t\boldsymbol{u}(\boldsymbol{x})$, and B is the matrix in the affine correspondence $\widehat{\tau} \overset{B\widehat{\boldsymbol{x}}+b}{\mapsto} \tau$. The edge, face, and tetrahedral moments defined above form a set of degrees of freedom for M_k and define a corresponding nodal basis $\mathcal{B}$ for M_k.

The basis $\mathcal{B}$ is divided into edge basis functions, face basis functions, and interior basis functions. An edge basis function $\boldsymbol{\phi}_e^p$ corresponding to an interior edge e has all of the above-defined degrees of freedom equal to zero except $\alpha_e^p(\boldsymbol{\phi}_e^p) = 1$ for some polynomial $p \in \mathcal{P}_{r-1}(e)$. Similarly, a face basis function $\boldsymbol{\phi}_f^{\boldsymbol{q}}$ has all its moments zero except $\alpha_f^{\boldsymbol{q}}(\boldsymbol{\phi}_f^{\boldsymbol{q}}) = 1$ for some interior face f and some $\boldsymbol{q} \in \mathcal{P}_{r-2}(f)^2$. Finally, we have interior basis functions $\boldsymbol{\phi}_\tau^{\boldsymbol{r}}$ supported on τ such that all its moments are zero except $\alpha_\tau^{\boldsymbol{r}}(\boldsymbol{\phi}_\tau^{\boldsymbol{r}}) = 1$ for some $\boldsymbol{r} \in \mathcal{R}_{r-3}$. Thus,

$$\begin{aligned}\mathcal{B} =& \{\boldsymbol{\phi}_e^p : \text{ for all interior mesh edges } e \text{ and } p \in \mathcal{P}_{r-1}(e)\}\\ &\cup\{\boldsymbol{\phi}_f^{\boldsymbol{q}} : \text{ for all interior mesh faces } f \text{ and } \boldsymbol{q} \in (\mathcal{P}_{r-2}(f))^2\}\\ &\cup\{\boldsymbol{\phi}_\tau^{\boldsymbol{r}} : \text{ for all } \boldsymbol{r} \in \mathcal{R}_{r-3} \text{ and all mesh tetrahedra } \tau\}.\end{aligned}$$

Our analysis generalizes to the case when M_k is decomposed as

$$M_k \;=\; \sum_{\boldsymbol{\phi}\in\mathcal{B}} \operatorname{span}(\boldsymbol{\phi}) \;\oplus\; \sum_i \operatorname{span}(\mathbf{grad}\ \psi_{k,i}),$$

where $\{\psi_{k,i}\}$ is a local nodal basis for W_k. To show this, we first note that Theorem 4.1 holds for this decomposition, as can be seen by following the arguments of [1]. The only other ingredient in our analysis that requires generalization is Lemma 3.1. We now show that $\|\boldsymbol{\phi}\|_{0,\Omega} \le Ch_k\|\mathbf{curl}\,\boldsymbol{\phi}\|_{0,\Omega}$ for all $\boldsymbol{\phi} \in \mathcal{B}$.

It suffices to prove that there is a $\hat{C} > 0$ such that

$$\|\widehat{\boldsymbol{\phi}}\|_{0,\hat{\tau}} \le \hat{C}\|\mathbf{curl}\,\widehat{\boldsymbol{\phi}}\|_{0,\hat{\tau}} \tag{A.2}$$

for all $\boldsymbol{\phi} \in \mathcal{B}$ with $\hat{C}$ independent of τ. Here, as before, $\widehat{\boldsymbol{\phi}}(\widehat{\boldsymbol{x}}) = B^t\,\boldsymbol{\phi}(\boldsymbol{x})$ for $\boldsymbol{x} \in \tau$ for some τ on which $\boldsymbol{\phi}$ is nonzero. Clearly, (3.1) follows from (A.2) by quasi uniformity and standard affine equivalence arguments since $\|\boldsymbol{\phi}\|_{0,\tau} \le Ch_k^{1/2}\|\widehat{\boldsymbol{\phi}}\|_{0,\hat{\tau}}$ and $\|\mathbf{curl}\,\widehat{\boldsymbol{\phi}}\|_{0,\hat{\tau}} \le Ch_k^{1/2}\|\mathbf{curl}\,\boldsymbol{\phi}\|_{0,\tau}$.

We prove (A.2) for each type of basis function. First, we consider $\widehat{\boldsymbol{\phi}}_{\widehat{e}}^{\,p}$. Let $L_{\widehat{e}}$ denote the space of functions $\boldsymbol{v}$ in the rth order Nedelec space on $\widehat{\tau}$ for which all edge, face, and tetrahedral moments are zero except those associated to the edge $\widehat{e}$, which is the image of e. Clearly, $\widehat{\boldsymbol{\phi}}_{\widehat{e}}^{\,p}$ is in $L_{\widehat{e}}$. For any nonzero function $\widehat{\boldsymbol{\phi}} \in L_{\widehat{e}}$, there exists a $p \in P_{r-1}(\widehat{f})$ on a face $\widehat{f}$ adjacent to $\widehat{e}$, such that

$$0 \neq (\widehat{\boldsymbol{\phi}} \cdot \boldsymbol{t}, p)_{\partial\widehat{f}} = (p, \mathbf{curl}\,\widehat{\boldsymbol{\phi}} \cdot \boldsymbol{n})_{\widehat{f}} - (\mathbf{grad}\; p \times \boldsymbol{n}, \widehat{\boldsymbol{\phi}})_{\widehat{f}} = (p, \mathbf{curl}\,\widehat{\boldsymbol{\phi}} \cdot \boldsymbol{n})_{\widehat{f}},$$

where $\boldsymbol{n}$ is the outward unit normal on $\widehat{f}$ and $\boldsymbol{t}$ is a unit tangent vector on $\partial\widehat{f}$ (appropriately oriented). Since the left-hand side is nonzero, $\mathbf{curl}\,\widehat{\boldsymbol{\phi}} \neq 0$. Thus by the finite dimensionality of $L_{\widehat{e}}$, (A.2) holds for all $\widehat{\boldsymbol{\phi}} \in L_{\widehat{e}}$ and hence holds for $\widehat{\boldsymbol{\phi}}_{\widehat{e}}^{\,p}$.

Next, let us show (A.2) for a mapped face basis function $\widehat{\boldsymbol{\phi}}_{\widehat{f}}^{\,\boldsymbol{q}}$. Let $L_{\widehat{f}}$ denote the subspace of the rth order Nedelec space on $\widehat{\tau}$ for which all edge, face, and tetrahedral moments are zero, except for moments on face $\widehat{f}$. Clearly, $\widehat{\boldsymbol{\phi}}_{\widehat{f}}^{\,\boldsymbol{q}}$ is in $L_{\widehat{f}}$. For any nonzero $\widehat{\boldsymbol{\phi}} \in L_{\widehat{f}}$, there is a $\boldsymbol{q} \in P_{r-2}(\widehat{\tau})^3$ such that

$$0 \neq (\widehat{\boldsymbol{\phi}} \times \boldsymbol{n}, \boldsymbol{q})_{\partial\widehat{\tau}} = (\mathbf{curl}\,\widehat{\boldsymbol{\phi}}, \boldsymbol{q})_{\widehat{\tau}} - (\widehat{\boldsymbol{\phi}}, \mathbf{curl}\,\boldsymbol{q})_{\widehat{\tau}} = (\mathbf{curl}\,\widehat{\boldsymbol{\phi}}, \boldsymbol{q})_{\widehat{\tau}}.$$

Thus, $\mathbf{curl}\,\widehat{\boldsymbol{\phi}} \neq 0$ for all $\widehat{\boldsymbol{\phi}} \in L_{\widehat{f}}$ and (A.2) follows for $\widehat{\boldsymbol{\phi}}_{\widehat{f}}^{\,\boldsymbol{q}}$.

Finally, consider an interior basis function $\boldsymbol{\phi}_{\tau}^{\boldsymbol{r}}$. Obviously, all face and edge moments of $\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}}$ are zero. We will now show only that for $\boldsymbol{r} = (x^i y^j z^k, 0, 0)$, $\mathbf{curl}\,\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}} \neq 0$, as the argument is similar for other $\boldsymbol{r} \in \mathcal{R}_{r-3}$. We argue by contradiction. If $\mathbf{curl}\,\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}} = 0$, then $\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}} = \mathbf{grad}\;\psi$ for some $\psi \in P_r(\widehat{\tau})$. Moreover, since face and edge moments of $\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}}$ are zero, ψ can be chosen such that $\psi|_{\partial\widehat{\tau}} = 0$. Therefore,

$$1 = (\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}}, \boldsymbol{r})_{\widehat{\tau}} = -(\psi, \operatorname{div}\boldsymbol{r})_{\widehat{\tau}}.$$

If $i = 0$, i.e., $\boldsymbol{r} = (y^j z^k, 0, 0)$, then $\operatorname{div}\boldsymbol{r} = 0$, which is a contradiction. If $i \ge 1$, then by the definition of $\boldsymbol{\phi}_{\tau}^{\boldsymbol{r}}$, $(\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}}, \widetilde{\boldsymbol{r}}) = 0$ for $\widetilde{\boldsymbol{r}} = (0, ix^{i-1}y^{j+1}z^k/(j+1), 0)$. However, $(\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}}, \widetilde{\boldsymbol{r}})_{\widehat{\tau}} = (\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}}, \boldsymbol{r})_{\widehat{\tau}}$, which is a contradiction. Therefore, $\mathbf{curl}\,\widehat{\boldsymbol{\phi}}_{\tau}^{\boldsymbol{r}} \neq 0$, and (A.2) follows for the interior basis functions as well. Thus, we have shown that Lemma 3.1 holds for the nodal basis functions of $\mathcal{B}$.

It is easy to see that there are various other choices of nodal bases for which the lemma does not hold. For instance, in the case $r \ge 4$ the function $\mathbf{grad}\;(\lambda_1\lambda_2\lambda_3\lambda_4)$ (where λ_i are the barycentric coordinates of a tetrahedron τ) is an example of an interior basis function for which Lemma 3.1 does not hold. Another example is the function $\lambda_i\mathbf{grad}\;\lambda_j + \lambda_j\mathbf{grad}\;\lambda_i$ in the case $r = 2$. This function has only one nonzero edge moment so may be a candidate for an edge basis function. However, it has nonzero face moments. Our analysis does not hold for decompositions based on such basis functions, and it is not clear if the associated indefinite multigrid method is convergent.

REFERENCES

[1] D. N. Arnold, R. S. Falk, and R. Winther, *Multigrid in* $\boldsymbol{H}(\mathrm{div})$ *and* $\boldsymbol{H}(\mathbf{curl})$, Numer. Math., 85 (2000), pp. 197–217.

[2] D. N. Arnold and A. Mukherjee, *Tetrahedral bisection and adaptive finite elements*, in Grid Generation and Adaptive Algorithms, Springer-Verlag, New York, 1999, pp. 29–42.

[3] R. Beck, P. Deuflhard, R. Hiptmair, R. H. W. Hoppe, and B. Wohlmuth, *Adaptive multilevel methods for edge element discretizations of Maxwell's equations*, Surveys Math. Indust., 8 (1999), pp. 271–312.

[4] R. Beck and R. Hiptmair, *Multilevel solution of the time-harmonic Maxwell's equations based on edge elements*, Internat J. Numer. Methods Engrg., 45 (1999), pp. 901–920.

[5] J. H. Bramble, *Multigrid Methods*, Pitman Res. Notes Math. Ser. 294, Longman Scientific & Technical, Harlow, UK, 1993.

[6] J. H. Bramble, D. Y. Kwak, and J. E. Pasciak, *Uniform convergence of multigrid V-cycle iterations for indefinite and nonsymmetric problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1746–1763.

[7] J. H. Bramble, J. E. Pasciak, J. P. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition*, Math. Comp., 57 (1991), pp. 1–21.

[8] J. H. Bramble and X. Zhang, *The analysis of multigrid methods*, in Handbook of Numerical Analysis Vol. VII, Handb. Numer. Anal. 7, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 173–415.

[9] J. H. Bramble and X. Zhang, *Uniform convergence of the multigrid V-cycle for an anisotropic problem*, Math. Comp., 70 (2001), pp. 453–470.

[10] T. A. Davis and I. S. Duff, *A combined unifrontal/multifrontal method for unsymmetric sparse matrices*, ACM Trans. Math. Software, 25 (1999), pp. 1–20.

[11] L. Demkowicz and L. Vardapetyan, *Modeling of electromagnetic absorption/scattering problems using hp-adaptive finite elements*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 103–124.

[12] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.

[13] H. C. Elman, O. G. Ernst, and D. P. O'Leary, *A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations*, SIAM J. Sci. Comput., 23 (2001), pp. 1291–1315.

[14] V. Girault, *Incompressible finite element methods for Navier-Stokes equations with nonstandard boundary conditions in* $\mathbf{R}^3$, Math. Comp., 51 (1988), pp. 55–74.

[15] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer Ser. Comput. Math. 5, Springer-Verlag, New York, 1986.

[16] J. Gopalakrishnan and J. E. Pasciak, *Overlapping Schwarz preconditioners for indefinite time harmonic Maxwell equations*, Math. Comp., 72 (2003), pp. 1–15.

[17] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.

[18] R. Hiptmair and A. Toselli, *Overlapping and multilevel Schwarz methods for vector valued elliptic problems in three dimensions*, in Parallel Solution of Partial Differential Equations, IMA Vol. Math. Appl. 120, Springer-Verlag, New York, 2000, pp. 181–208.

[19] R. Leis, *Exterior boundary-value problems in mathematical physics*, in Trends in Applications of Pure Mathematics to Mechanics, Volume II, Monographs Stud. Math. 5, H. Zorski, ed., Pitman, London, 1979, pp. 187–203.

[20] P. Monk, *A finite element method for aproximating the time-harmonic Maxwell equations*, Numer. Math., 63 (1992), pp. 243–261.

[21] P. Monk and L. Demkowicz, *Discrete compactness and approximation of Maxwell's equations in* $\mathbb{R}^3$, Math. Comp., 70 (2001), pp. 507–523.

[22] J. C. Nedelec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[23] W. Rachowicz, L. Demkowicz, A. Bajer, and T. Walsh, *A two-grid iterative solver for stationary Maxwell's equations*, in Iterative Methods in Scientific Computation II, D. Kincaid et al., eds., IMACS, New Brunswick, NJ, 1999, pp. 421–453.

[24] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition. Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 109–129

A MULTIGRID METHOD FOR VISCOELASTIC FLUID FLOW*

HYESUK LEE†

Abstract. We study a multigrid finite element method for a viscoelastic fluid flow obeying an Oldroyd-B–type constitutive law. The multigrid method is a time-saving method in which the full nonlinear system is solved on a coarse grid, and subsequent approximations are generated on a succession of refined grids by solving a linearized problem. We show that the linearized problem has an approximate solution and present an error bound for a two-grid method. We also numerically demonstrate that the multigrid method is significantly more efficient than the standard one-grid finite element method.

Key words. multigrid method, viscoelastic fluid, finite elements

AMS subject classifications. 65N30, 34K25

DOI. 10.1137/S0036142902415924

1. Introduction.

In this paper we consider a multigrid method for a viscoelastic fluid flow obeying an Oldroyd-B–type constitutive law. Over the last decades, significant progress has been made in developing finite element algorithms for viscoelastic fluids, resulting in commercial software packages available today. The difficulty in approximating a solution to the viscoelastic flow model arises from the hyperbolic nature of the constitutive equation, which requires stabilization in computation. Streamline upwinding was first considered in [9] for viscoelastic flow, and in [4] the discontinuous Galerkin method was implemented. In [1], a preconditioner for the GMRES iterative solver was specifically tailored to the characteristics of the discrete elastic-viscous split-stress method (DEVSS) algorithm. More mathematical studies on finite element methods for the solution of viscoelastic fluid flow are found in [2] and [12]. In [2] a discontinuous Galerkin method was studied for approximating the stress, and in [12] the streamline upwinding Petrov–Galerkin (SUPG) method for continuous approximation of the stress was analyzed.

When a discontinuous Galerkin method is used for stabilizing the approximation, the generated linear system is considerably larger than that resulting from a standard Galerkin method. In this sense, a multigrid method can give more savings in computing time when used together with the discontinuous finite element method. In [8] we studied a multigrid method derived from the decoupled algorithm presented in [10]. The advantage of that method is in the decoupling of a momentum equation from a constitutive equation on a finer mesh. After solving a full nonlinear equation on a coarse grid using an appropriate iteration method, one needs to solve two smaller linear systems on a fine grid instead of one large system. However, convergence of the decoupled algorithm is slow and mesh-dependent, so it is necessary to use several iterations on a finer grid.

The multilevel method considered in this paper is based on Newton linearization. A similar multigrid method for the Navier–Stokes equation was studied in [6] and [7].

*Received by the editors October 9, 2002; accepted for publication (in revised form) May 12, 2003; published electronically January 6, 2004. This work was partially supported by the ERC Program of the NSF under award ERC-9731680.

http://www.siam.org/journals/sinum/42-1/41592.html

†Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975 (hklee@clemson.edu).

For the viscoelastic flow problem, we solve a full nonlinear system on a coarse grid and one linear system on a fine grid with appropriately chosen coarse and fine grids.

An outline of this paper is as follows. In the remainder of this section, we introduce the model equation, some notation that will be used throughout the paper, and a weak formulation of the model equation. In section 2, a finite element approximation for a standard one-grid method is described and the error estimate proved in [2] is presented. In section 3, we present a multigrid algorithm and prove the existence of a solution to the discrete linearized problem. An error bound of the multigrid method is proved in section 4. Then, in the last section we present some numerical results demonstrating efficiency of the method.

Let Ω be a bounded domain in $\mathbb{R}^2$ with the Lipschitz continuous boundary Γ, and let $\mathbf{n}$ be the unit outward normal to the boundary Γ. Consider the model problem

$$\boldsymbol{\sigma} + \lambda(\mathbf{u}\cdot\nabla)\boldsymbol{\sigma} + \lambda g_a(\boldsymbol{\sigma},\nabla\mathbf{u}) - 2\,\alpha\, d(\mathbf{u}) = \mathbf{h} \quad \text{in } \Omega\,, \tag{1.1}$$

$$-\nabla\cdot\boldsymbol{\sigma} - 2(1-\alpha)\,\nabla\cdot d(\mathbf{u}) + \nabla p = \mathbf{f} \quad \text{in } \Omega\,, \tag{1.2}$$

$$\operatorname{div}\mathbf{u} = g \quad \text{in } \Omega\,, \tag{1.3}$$

$$\mathbf{u} = \mathbf{u}_\Gamma \quad \text{on } \Gamma\,, \tag{1.4}$$

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_{\Gamma_-} \quad \text{on } \Gamma_-\,, \tag{1.5}$$

where $\boldsymbol{\sigma}$ denotes the stress tensor, $\mathbf{u}$ the velocity vector, p the pressure of fluid, and λ the Weissenberg number. Assume that p has zero mean value over Ω. In (1.1) and (1.2), $d(\mathbf{u}) := (\nabla\mathbf{u} + \nabla\mathbf{u}^T)/2$ is the rate of the strain tensor, and α is a number such that $0 < \alpha < 1$, which may be considered as the fraction of viscoelastic viscosity. In (1.1), $g_a(\boldsymbol{\sigma},\nabla\mathbf{u})$ is defined by

$$g_a(\boldsymbol{\sigma},\nabla\mathbf{u}) := \frac{1-a}{2}(\boldsymbol{\sigma}\nabla\mathbf{u} + \nabla\mathbf{u}^T\boldsymbol{\sigma}) - \frac{1+a}{2}(\nabla\mathbf{u}\,\boldsymbol{\sigma} + \boldsymbol{\sigma}\nabla\mathbf{u}^T) \tag{1.6}$$

for $a \in [-1,1]$. In (1.5), Γ_- denotes the inflow boundary, where $\Gamma_- = \{\mathbf{x}\in\Gamma | \mathbf{u}\cdot\mathbf{n} < 0\}$. Note that if $\mathbf{u}_\Gamma = \mathbf{0}$, then there is no inflow boundary, so the boundary condition for the stress (1.5) is not necessary. The right-hand-side functions $\mathbf{h}$ and g are usually set to be zero in viscoelastic model equations by a constitutive law and the incompressibility condition.

We use the Sobolev spaces $W^{m,p}(D)$ with norms $\|\cdot\|_{m,p,D}$ if $p < \infty$ and $\|\cdot\|_{m,\infty,D}$ if $p = \infty$. We denote the Sobolev space $W^{m,2}$ by H^m, with the norm $\|\cdot\|_m$. Let $\mathbf{H}^m$ denote the corresponding spaces of vector-valued and tensor-valued functions. We will denote H^0 by L^2 and the standard L^2 inner product by $(\cdot,\cdot)_D$. If $D = \Omega$, D is omitted; i.e., $(\cdot,\cdot) = (\cdot,\cdot)_\Omega$ and $\|\cdot\| = \|\cdot\|_\Omega$.

An existence result for the problem (1.1)–(1.5) has been documented by Renardy [11] for the case that $\mathbf{h} = \mathbf{0}$, $g = 0$, and $\mathbf{u}_\Gamma = \mathbf{0}$, with the small data condition. He showed that if $\mathbf{f}\in\mathbf{H}^2(\Omega)$ and $\|\mathbf{f}\|_2$ is sufficiently small, the problem (1.1)–(1.5) admits a unique bounded solution $(\mathbf{u},\boldsymbol{\sigma},p)\in\mathbf{H}^3(\Omega)\times\mathbf{H}^2(\Omega)\times H^2(\Omega)$. In order to simplify our analysis, we make the assumption that $\mathbf{h} = \mathbf{0}$, $g = 0$, and $\mathbf{u}_\Gamma = \mathbf{0}$. However, it will be shown in section 5 that these conditions are not necessary for computations by our multigrid method.

Next, we define the function spaces for the velocity $\mathbf{u}$, the pressure p, and the

stress $\boldsymbol{\sigma}$, respectively:

$$\mathbf{H}_0^1(\Omega) := \{\mathbf{v} \in \mathbf{H}^1(\Omega) : \ \mathbf{v} = \mathbf{0} \quad \text{on } \Gamma\},$$
$$L_0^2(\Omega) := \{q \in L^2(\Omega) : \ \textstyle\int_\Omega q \, d\Omega = 0\},$$
$$\boldsymbol{\Sigma} := (L^2(\Omega))^{2\times 2} \cap \{\boldsymbol{\tau} = (\tau_{ij}) : \tau_{ij} = \tau_{ji}, \ \mathbf{u} \cdot \nabla \boldsymbol{\tau} \in (L^2(\Omega))^{2\times 2} \quad \forall \mathbf{u} \in \mathbf{H}_0^1(\Omega)\}.$$

Under the assumption that $\mathbf{h} = \mathbf{0}$, $g = 0$, and $\mathbf{u}_\Gamma = \mathbf{0}$, the corresponding variational form of (1.1)–(1.5) is obtained in the standard manner. Taking the inner product of (1.1)–(1.3) with stress, velocity, and pressure test functions, respectively, we obtain

$$(\boldsymbol{\sigma}, \boldsymbol{\tau}) + \lambda\left((\mathbf{u} \cdot \nabla)\boldsymbol{\sigma}, \boldsymbol{\tau}\right) + (g_a(\boldsymbol{\sigma}, \nabla \mathbf{u}), \boldsymbol{\tau}) - 2\alpha(d(\mathbf{u}), \boldsymbol{\tau}) = 0 \quad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma}, \tag{1.7}$$

$$(\boldsymbol{\sigma}, d(\mathbf{v})) + 2(1-\alpha)(d(\mathbf{u}), d(\mathbf{v})) + (p, \nabla \cdot \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega), \tag{1.8}$$

$$(q, \nabla \cdot \mathbf{u}) = 0 \quad \forall q \in L_0^2(\Omega). \tag{1.9}$$

We define the weak divergence free space

$$\mathbf{V} := \{\mathbf{v} \in \mathbf{H}_0^1(\Omega) : (q, \nabla \cdot \mathbf{v}) = 0 \quad \forall q \in L_0^2(\Omega)\} \tag{1.10}$$

to be used for the analysis. Note that, using (1.10), the weak formulation (1.7)–(1.9) is equivalent to:

$$(\boldsymbol{\sigma}, \boldsymbol{\tau}) + \lambda\left((\mathbf{u} \cdot \nabla)\boldsymbol{\sigma}, \boldsymbol{\tau}\right) + (g_a(\boldsymbol{\sigma}, \nabla \mathbf{u}), \boldsymbol{\tau}) - 2\alpha(d(\mathbf{u}), \boldsymbol{\tau}) = \mathbf{0} \quad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma}, \tag{1.11}$$

$$(\boldsymbol{\sigma}, d(\mathbf{v})) + 2(1-\alpha)(d(\mathbf{u}), d(\mathbf{v})) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{V}. \tag{1.12}$$

In the following finite element analysis we will use the bilinear form A defined on $\boldsymbol{\Sigma} \times \mathbf{V}$ by

$$\begin{aligned} &A((\boldsymbol{\sigma}, \mathbf{u}), (\boldsymbol{\tau}, \mathbf{v})) \\ &\quad := (\boldsymbol{\sigma}, \boldsymbol{\tau}) - 2\alpha(d(\mathbf{u}), \boldsymbol{\tau}) + 2\alpha(\boldsymbol{\sigma}, d(\mathbf{v})) + 4\alpha(1-\alpha)\,(d(\mathbf{u}), d(\mathbf{v})). \end{aligned} \tag{1.13}$$

It is easily shown that A is continuous and coercive on $\boldsymbol{\Sigma} \times \mathbf{V}$, i.e.,

$$\begin{aligned} A((\boldsymbol{\sigma}, \mathbf{u}), (\boldsymbol{\tau}, \mathbf{v})) &\le \|\boldsymbol{\sigma}\|_0 \|\boldsymbol{\tau}\|_0 + 2\alpha \|\nabla \mathbf{u}\|_0 \|\boldsymbol{\tau}\|_0 + 2\alpha \|\boldsymbol{\sigma}\|_0 \|\nabla \mathbf{v}\|_0 \\ &\quad + \ 4\alpha(1-\alpha)\|\nabla \mathbf{u}\|_0 \|\nabla \mathbf{v}\|_0 \\ &\le C(\|\boldsymbol{\sigma}\|_0 + \|\nabla \mathbf{u}\|_0)(\|\boldsymbol{\tau}\|_0 + \|\nabla \mathbf{v}\|_0) \\ &\le C\sqrt{\|\boldsymbol{\sigma}\|_0^2 + \|\nabla \mathbf{u}\|_0^2}\sqrt{\|\boldsymbol{\tau}\|_0^2 + \|\nabla \mathbf{v}\|_0^2} \\ &\le C\|(\boldsymbol{\sigma}, \mathbf{u})\|_{(L^2(\Omega))^{2\times 2} \times \mathbf{H}_0^1(\Omega)} \|(\boldsymbol{\tau}, \mathbf{v})\|_{(L^2(\Omega))^{2\times 2} \times \mathbf{H}_0^1(\Omega)}, \end{aligned} \tag{1.14}$$

and

$$\begin{aligned} A((\boldsymbol{\sigma}, \mathbf{u}), (\boldsymbol{\sigma}, \mathbf{u})) &= \|\boldsymbol{\sigma}\|_0^2 + 4\alpha(1-\alpha)\|d(\mathbf{u})\|_0^2 \\ &\ge C\|(\boldsymbol{\sigma}, \mathbf{u})\|^2_{(L^2(\Omega))^{2\times 2} \times \mathbf{H}_0^1(\Omega)} \end{aligned} \tag{1.15}$$

by the second Korn inequality. These properties of A will be used to show the existence of the solution of a linear fine-grid problem and to obtain an error estimation.

2. Finite element approximation. Suppose Ω is a polygonal domain and T_h is a triangulation of Ω such that $\overline{\Omega} = \{\cup K : K \in T_h\}$. Assume that there exist positive constants c_1, c_2 such that

$$c_1 h \le h_K \le c_2 \rho_K,$$

where h_K is the diameter of K, ρ_K is the diameter of the greatest ball included in K, and $h = \max_{K\in T_h} h_K$.

Let $P_k(K)$ denote the space of polynomials of degree less than or equal to k on $K \in T_h$. Then we define finite element spaces for the approximate of $(\mathbf{u}, p)$:

$$\begin{aligned}
\mathbf{X}^h &:= \{\mathbf{v} \in \mathbf{H}_0^1(\Omega) \cap (C^0(\overline{\Omega}))^2 : \mathbf{v}|_K \in P_2(K)^2 \ \forall K \in T_h\}, \\
S^h &:= \{q \in L_0^2(\Omega) \cap C^0(\overline{\Omega}) : q|_K \in P_1(K) \ \forall K \in T_h\}, \\
\mathbf{V}^h &:= \{\mathbf{v} \in \mathbf{X}^h : (q, \nabla \cdot \mathbf{v}) = 0 \ \forall q \in S^h\}.
\end{aligned}$$

The stress $\boldsymbol{\sigma}$ is approximated in the discontinuous finite element space of piecewise linears:

$$\boldsymbol{\Sigma}^h := \{\boldsymbol{\tau} \in \boldsymbol{\Sigma} : \boldsymbol{\tau}|_K \in P_1(K)^{2\times 2} \ \forall K \in T_h\}.$$

The finite element spaces defined above satisfy the standard approximation properties (see [3] or [5]), i.e., there exist an integer k and a constant C such that

$$\inf_{\mathbf{v}^h \in \mathbf{X}^h} \|\mathbf{v} - \mathbf{v}^h\|_1 \le C h^2 \|\mathbf{v}\|_3 \quad \forall\, \mathbf{v} \in \mathbf{H}^3(\Omega), \tag{2.1}$$

$$\inf_{q^h \in S^h} \|q - q^h\|_0 \le C h^2 \|q\|_2 \quad \forall\, q \in H^2(\Omega), \tag{2.2}$$

and

$$\inf_{\boldsymbol{\tau}^h \in \Sigma^h} \|\boldsymbol{\tau} - \boldsymbol{\tau}^h\|_0 \le C h^2 \|\boldsymbol{\tau}\|_2 \quad \forall\, \boldsymbol{\tau} \in \mathbf{H}^2(\Omega). \tag{2.3}$$

It is also well known that the Taylor–Hood pair $(\mathbf{X}^h, S^h)$ satisfies the inf-sup (or LBB) condition,

$$\inf_{0\neq q^h \in S^h} \sup_{0\neq \mathbf{v}^h \in \mathbf{X}^h} \frac{(q^h, \nabla \cdot \mathbf{v}^h)}{\|\mathbf{v}^h\|_1 \|q^h\|_0} \ge C, \tag{2.4}$$

where C is a positive constant independent of h.

Below we introduce some notation used in [2] in order to analyze an approximate solution by the discontinuous Galerkin method. We define

$$\partial K^-(\mathbf{u}) := \{\mathbf{x} \in \partial K, \ \mathbf{u} \cdot \mathbf{n} < 0\},$$

where ∂K is the boundary of K and $\mathbf{n}$ is the outward unit normal to ∂K,

$$\Gamma^h = \{\cup \partial K, K \in T_h\} \setminus \Gamma,$$

and

$$\boldsymbol{\tau}^{\pm}(\mathbf{u}) := \lim_{\epsilon \to 0^{\pm}} \boldsymbol{\tau}(\mathbf{x} + \epsilon \mathbf{u}(\mathbf{x})).$$

We also define

$$(\boldsymbol{\sigma},\boldsymbol{\tau})_h := \sum_{K\in T_h} (\boldsymbol{\sigma},\boldsymbol{\tau})_K,$$

$$\langle \boldsymbol{\sigma}^{\pm},\boldsymbol{\tau}^{\pm}\rangle_{h,\mathbf{u}} := \sum_{K\in T_h} \int_{\partial K^-(\mathbf{u})} (\boldsymbol{\sigma}^{\pm}(\mathbf{u}) : \boldsymbol{\tau}^{\pm}(\mathbf{u}))|\mathbf{n}\cdot\mathbf{u}|\, ds,$$

$$\|\boldsymbol{\tau}\|_{0,\Gamma^h} := \left(\sum_{K\in T_h} |\boldsymbol{\tau}|^2_{0,\partial K}\right)^{1/2}$$

for $\boldsymbol{\sigma},\boldsymbol{\tau} \in \prod_{K\in T_h}(L^2(K))^{2\times 2}$ and

$$\|\boldsymbol{\xi}\|_{m,p,h} := \left(\sum_{K\in T_h} |\boldsymbol{\xi}|^p_{m,p,K}\right)^{1/p}$$

for $\boldsymbol{\xi} \in \prod_{K\in T_h}(W^{m,p}(K))^{2\times 2}$ if $p<\infty$.

We introduce the operator B^h on $\mathbf{X}^h \times \boldsymbol{\Sigma}^h \times \boldsymbol{\Sigma}^h$ defined by

$$B^h(\mathbf{u},\boldsymbol{\sigma},\boldsymbol{\tau}) := ((\mathbf{u}\cdot\nabla)\boldsymbol{\sigma},\boldsymbol{\tau})_h + \frac{1}{2}(\nabla\cdot\mathbf{u}\,\boldsymbol{\sigma},\boldsymbol{\tau}) + \langle \boldsymbol{\sigma}^+ - \boldsymbol{\sigma}^-, \boldsymbol{\tau}^+\rangle_{h,\mathbf{u}}. \tag{2.5}$$

Note that the second term vanishes when $\nabla\cdot\mathbf{u}^h = 0$. This extra term is used to obtain coercivity of B^h. Using integration by parts, B^h may be written as

$$B^h(\mathbf{u},\boldsymbol{\sigma},\boldsymbol{\tau}) = -((\mathbf{u}\cdot\nabla)\boldsymbol{\tau},\boldsymbol{\sigma})_h - \frac{1}{2}(\nabla\cdot\mathbf{u}\,\boldsymbol{\tau},\boldsymbol{\sigma}) + \langle \boldsymbol{\sigma}^-, \boldsymbol{\tau}^- - \boldsymbol{\tau}^+\rangle_{h,\mathbf{u}}. \tag{2.6}$$

Hence, combining (2.5) and (2.6), we obtain

$$B^h(\mathbf{u}^h,\boldsymbol{\sigma}^h,\boldsymbol{\sigma}^h) = \frac{1}{2}\langle \boldsymbol{\sigma}^{h+} - \boldsymbol{\sigma}^{h-}, \boldsymbol{\sigma}^{h+} - \boldsymbol{\sigma}^{h-}\rangle_{h,\mathbf{u}^h} \geq 0. \tag{2.7}$$

The discontinuous Galerkin finite element approximation of (1.7)–(1.9) is then as follows: *find* $\mathbf{u}^h \in \mathbf{X}^h$, $p^h \in S^h$, $\boldsymbol{\sigma}^h \in \boldsymbol{\Sigma}^h$ *such that*

$$\begin{aligned}(\boldsymbol{\sigma}^h,\boldsymbol{\tau}^h) + \lambda B^h(\mathbf{u}^h,\boldsymbol{\sigma}^h,\boldsymbol{\tau}^h) + \lambda(g_a(\boldsymbol{\sigma}^h,\nabla\mathbf{u}^h),\boldsymbol{\tau}^h)& \\ - 2\alpha(d(\mathbf{u}^h),\boldsymbol{\tau}^h) = 0 \quad \forall \boldsymbol{\tau}^h \in \boldsymbol{\Sigma}^h,&\end{aligned} \tag{2.8}$$

$$(\boldsymbol{\sigma}^h, d(\mathbf{v}^h)) + 2(1-\alpha)\,(d(\mathbf{u}^h), d(\mathbf{v}^h)) - (p^h, \nabla\cdot\mathbf{v}^h) = (\mathbf{f},\mathbf{v}^h) \quad \forall \mathbf{v}^h \in \mathbf{X}^h, \tag{2.9}$$

$$(q^h, \nabla\cdot\mathbf{u}^h) = 0 \quad \forall q^h \in S^h. \tag{2.10}$$

Existence of a solution to the discrete problem (2.8)–(2.10) is proved in [2], under the assumption that there exists a bounded exact solution $(\mathbf{u},\boldsymbol{\sigma},p) \in \mathbf{H}^3(\Omega)\times\mathbf{H}^2(\Omega)\times H^2(\Omega)$. The following error estimations are also derived there:

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\|_0 + \|\nabla(\mathbf{u}-\mathbf{u}^h)\|_0 \leq N\, h^{3/2}\,, \tag{2.11}$$

$$\|p - p^h\|_0 \leq N\, h^{3/2} \tag{2.12}$$

for some $N > 0$. It is then clear that

$$\|\boldsymbol{\sigma}^h\|_0 \le M + N\,h^{3/2}\,, \tag{2.13}$$

$$\|\nabla \mathbf{u}^h\|_0 \le M + N\,h^{3/2}\,, \tag{2.14}$$

where $M = \max\{\|\mathbf{u}\|_3, \|\boldsymbol{\sigma}\|_2\}$. Note that, in view of (2.4), (2.8)–(2.10) are equivalent to the following: *find* $\mathbf{u}^h \in \mathbf{V}^h$ *and* $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}^h$ *such that*

$$\begin{aligned}(\boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) + \lambda B^h(\mathbf{u}^h, \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) + \lambda(g_a(\boldsymbol{\sigma}^h, \nabla \mathbf{u}^h), \boldsymbol{\tau}^h) \\ - 2\alpha(d(\mathbf{u}^h), \boldsymbol{\tau}^h) = 0 \quad \forall \boldsymbol{\tau}^h \in \boldsymbol{\Sigma}^h,\end{aligned} \tag{2.15}$$

$$(\boldsymbol{\sigma}^h, d(\mathbf{v}^h)) + 2(1-\alpha)\,(d(\mathbf{u}^h), d(\mathbf{v}^h)) = (\mathbf{f}, \mathbf{v}^h) \quad \forall \mathbf{v}^h \in \mathbf{V}^h. \tag{2.16}$$

Using the bilinear form A defined by (1.13), (2.15)–(2.16) can equivalently be written as

$$\begin{aligned}A((\boldsymbol{\sigma}^h, \mathbf{u}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h)) + \lambda B^h(\mathbf{u}^h, \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) \\ + \lambda(g_a(\boldsymbol{\sigma}^h, \nabla \mathbf{u}^h), \boldsymbol{\tau}^h) = 2\alpha(\mathbf{f}, \mathbf{v}^h) \quad \forall (\boldsymbol{\tau}^h, \mathbf{v}^h) \in \boldsymbol{\Sigma}^h \times \mathbf{V}^h.\end{aligned} \tag{2.17}$$

3. Two-grid problem. We consider two finite element meshes with $H > h$ and finite element spaces (X^H, S^H, Σ^H), (X^h, S^h, Σ^h). The two-grid method for approximating the solution of (1.7)–(1.9) is as follows.

ALGORITHM 3.1 (two-grid method).

Step 1. *Solve the nonlinear problem on the coarse mesh* H: *find* $\mathbf{u}^H \in \mathbf{X}^H$, $p^H \in S^H$, $\boldsymbol{\sigma}^H \in \boldsymbol{\Sigma}^H$ *such that*

$$\begin{aligned}&(\boldsymbol{\sigma}^H, \boldsymbol{\tau}^H) + \lambda B^H(\mathbf{u}^H, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^H) + \lambda(g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^H), \boldsymbol{\tau}^H) \\ &\qquad - 2\alpha(d(\mathbf{u}^H), \boldsymbol{\tau}^H) = 0 \quad \forall \boldsymbol{\tau}^H \in \boldsymbol{\Sigma}^H, \\ &(\boldsymbol{\sigma}^H, d(\mathbf{v}^H)) + 2(1-\alpha)\,(d(\mathbf{u}^H), d(\mathbf{v}^H)) - (p^H, \nabla \cdot \mathbf{v}^H) \\ &\qquad = (\mathbf{f}, \mathbf{v}^H) \quad \forall \mathbf{v}^H \in \mathbf{X}^H, \\ &(q^H, \nabla \cdot \mathbf{u}^H) = 0 \quad \forall q^H \in S^H.\end{aligned}$$

Step 2. *Solve the linearized problem on the fine mesh* h: *find* $\mathbf{u}^h \in \mathbf{X}^h$, $p^h \in S^h$, $\boldsymbol{\sigma}^h \in \boldsymbol{\Sigma}^h$ *such that*

$$\begin{aligned}&(\boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) + \lambda\left[B^h(\mathbf{u}^H, \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) + B^h(\mathbf{u}^h, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + \ (g_a(\boldsymbol{\sigma}^h, \nabla \mathbf{u}^H), \boldsymbol{\tau}^h)\right. \\ &\qquad \left. + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^h), \boldsymbol{\tau}^h)\right] - 2\alpha(d(\mathbf{u}^h), \boldsymbol{\tau}^h) \\ &\qquad\quad = \lambda\left[B^h(\mathbf{u}^H, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^H), \boldsymbol{\tau}^h)\right] \quad \forall \boldsymbol{\tau}^h \in \boldsymbol{\Sigma}^h, \\ &(\boldsymbol{\sigma}^h, d(\mathbf{v}^h)) + 2(1-\alpha)(d(\mathbf{u}^h), d(\mathbf{v}^h)) - (p^h, \nabla \cdot \mathbf{v}^h) = (\mathbf{f}, \mathbf{v}^h) \quad \forall \mathbf{v}^h \in \mathbf{X}^h, \\ &(q^h, \nabla \cdot \mathbf{u}^h) = 0 \quad \forall q^h \in S^h.\end{aligned}$$

In the second step of the above algorithm $\mathbf{u}^H$, $\boldsymbol{\sigma}^H$ are interpolates of $\mathbf{u}^H$, $\boldsymbol{\sigma}^H$ in $\mathbf{V}^h$ and $\boldsymbol{\Sigma}^h$, respectively. In the remainder of this section, we prove existence of a solution to the linear problem in Step 2 under the assumption that λ is sufficiently small.

Using the discrete divfree space $\mathbf{V}^h$, the linear problem in Step 2 may be written in an equivalent form: *find* $\mathbf{u}^h \in \mathbf{V}^h$, $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}^h$ *such that*

$$\begin{aligned}(\boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) + \lambda \big[B^h(\mathbf{u}^h, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + B^h(\mathbf{u}^H, \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^h, \nabla \mathbf{u}^H), \boldsymbol{\tau}^h) \\ + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^h), \boldsymbol{\tau}^h)\big] - 2\alpha(d(\mathbf{u}^h), \boldsymbol{\tau}^h) \\ = \lambda \big(B^h(\mathbf{u}^H, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^H), \boldsymbol{\tau}^h)\big) \quad \forall \boldsymbol{\tau}^h \in \boldsymbol{\Sigma}^h, \end{aligned} \tag{3.1}$$

$$(\boldsymbol{\sigma}^h, d(\mathbf{v}^h)) + 2(1-\alpha)(d(\mathbf{u}^h), d(\mathbf{v}^h)) = (\mathbf{f}, \mathbf{v}^h) \quad \forall \mathbf{v}^h \in \mathbf{V}^h. \tag{3.2}$$

Using the bilinear form defined by (1.13), (3.1)–(3.2) may be written as

$$\begin{aligned} A\big((\boldsymbol{\sigma}^h, \mathbf{u}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h)\big) + \lambda \big[B^h(\mathbf{u}^h, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + B^h(\mathbf{u}^H, \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) \\ + (g_a(\boldsymbol{\sigma}^h, \nabla \mathbf{u}^H), \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^h), \boldsymbol{\tau}^h)\big] \\ = \lambda \big[B^h(\mathbf{u}^H, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^H), \boldsymbol{\tau}^h)\big] \\ + 2\alpha(\mathbf{f}, \mathbf{v}^h) \quad \forall (\boldsymbol{\tau}^h, \mathbf{v}^h) \in \boldsymbol{\Sigma}^h \times \mathbf{V}^h. \end{aligned} \tag{3.3}$$

In order to simplify our notation, we introduce the bilinear operator Φ defined on $\boldsymbol{\Sigma}^h \times \mathbf{V}^h \times \boldsymbol{\Sigma}^h \times \mathbf{V}^h$ by

$$\begin{aligned}\Phi((\boldsymbol{\sigma}^h, \mathbf{u}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h)) := A\big((\boldsymbol{\sigma}^h, \mathbf{u}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h)\big) + \lambda \big[B^h(\mathbf{u}^h, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) \\ + B^h(\mathbf{u}^H, \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^h, \nabla \mathbf{u}^H), \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^h), \boldsymbol{\tau}^h)\big]. \end{aligned} \tag{3.4}$$

Note that

$$\begin{aligned}\Phi((\boldsymbol{\sigma}^h, \mathbf{u}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h)) \\ = \lambda \big[B^h(\mathbf{u}^H, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^H), \boldsymbol{\tau}^h)\big] + 2\alpha(\mathbf{f}, \mathbf{v}^h) \end{aligned} \tag{3.5}$$

if $(\boldsymbol{\sigma}^h, \mathbf{u}^h)$ satisfies (3.3).

Next, we present some inverse estimates (see [3] or [5]), which will be used to prove theorems in this paper: for $\mathbf{u}^h \in \mathbf{V}^h$ and $\boldsymbol{\sigma}^h \in \boldsymbol{\Sigma}^h$

$$\|\mathbf{u}^h\|_\infty \le Ch^{-1/2}\|\mathbf{u}^h\|_{0,4}\,, \tag{3.6}$$

$$\|\mathbf{u}^h\|_\infty \le Ch^{-1/2}\|\nabla \mathbf{u}^h\|_0\,, \tag{3.7}$$

$$\|\boldsymbol{\sigma}^h\|_\infty \le Ch^{-1}\|\boldsymbol{\sigma}^h\|_0\,, \tag{3.8}$$

$$\|\nabla \boldsymbol{\sigma}^h\|_{0,4,h} \le Ch^{-3/2}\|\boldsymbol{\sigma}^h\|_0\,. \tag{3.9}$$

The local inverse inequality [13, sec. 4.6.1]

$$\|\boldsymbol{\sigma}\|^2_{0,\partial K} \;\le\; C\frac{1}{h_K}\|\boldsymbol{\sigma}\|^2_{0,K} \tag{3.10}$$

will be also used to bound the jump term of B^h, where h_K denotes the local mesh parameter. The local inverse inequality then implies that

$$\begin{aligned}\langle \boldsymbol{\sigma}^{h^+} - \boldsymbol{\sigma}^{h^-}, \boldsymbol{\tau}^{h^+}\rangle_{h,\mathbf{u}} &\le \|\mathbf{u}\|_\infty \|\boldsymbol{\sigma}^h\|_{0,\Gamma^h}\|\boldsymbol{\tau}^h\|_{0,\Gamma^h} \\ &\le C\|\mathbf{u}\|_\infty (h^{-1/2}\|\boldsymbol{\sigma}^h\|_{0,\Omega})(h^{-1/2}\|\boldsymbol{\tau}^h\|_{0,\Omega}) \end{aligned} \tag{3.11}$$

for $\boldsymbol{\sigma}^h, \boldsymbol{\tau}^h \in \boldsymbol{\Sigma}^h$.

THEOREM 3.2. *The linear problem in Step* 2 *of the algorithm admits a unique solution* $(\mathbf{u}^h, \boldsymbol{\sigma}^h, p^h)$ *if* λ *is sufficiently small.*

Proof. We will use the Lax–Milgram theorem to prove that (3.4) has a unique solution. In the proof we use C^h to denote a constant depending on h.

(i) *If* λ *is sufficiently small, then* Φ *is coercive and* $\|(\boldsymbol{\sigma}^h, \mathbf{u}^h)\|_{(L^2(\Omega))^{2\times 2}\times \mathbf{H}^1_0(\Omega)}$ *is bounded.* Using the imbedding of H^1 in L^4, (2.13), (3.7)–(3.9), (3.11), the Poincaré–Friedrichs inequality, and Young's inequality with $\epsilon_1 > 0$, we have

$$
\begin{aligned}
&B^h(\mathbf{u}^h, \boldsymbol{\sigma}^H, \boldsymbol{\sigma}^h)\\
&\quad = ((\mathbf{u}^h \cdot \nabla)\boldsymbol{\sigma}^H, \boldsymbol{\sigma}^h)_h + \frac{1}{2}(\nabla \cdot \mathbf{u}^h\, \boldsymbol{\sigma}^H, \boldsymbol{\sigma}^h) + \langle \boldsymbol{\sigma}^{H^+} - \boldsymbol{\sigma}^{H^-}, \boldsymbol{\sigma}^{h^-}\rangle_{h,\mathbf{u}^h}\\
&\quad \le C_1 \left[\|\mathbf{u}^h\|_{0,4}\, \|\nabla \boldsymbol{\sigma}^H\|_{0,4,h}\, \|\boldsymbol{\sigma}^h\|_0 + \|\nabla \mathbf{u}^h\|_0\, \|\boldsymbol{\sigma}^H\|_\infty\, \|\boldsymbol{\sigma}^h\|_0\right]\\
&\qquad\qquad + C_1^h\, \|\boldsymbol{\sigma}^H\|_0\, \|\boldsymbol{\sigma}^h\|_0\, \|\mathbf{u}^h\|_\infty\\
&\quad \le C_2 \left[\|\nabla \mathbf{u}^h\|_0\, (h^{-3/2}\, \|\boldsymbol{\sigma}^H\|_0)\, \|\boldsymbol{\sigma}^h\|_0 + \|\nabla \mathbf{u}^h\|_0\, (h^{-1}\, \|\boldsymbol{\sigma}^H\|_0)\, \|\boldsymbol{\sigma}^h\|_0\right]\\
&\qquad\qquad + C_2^h\, \|\boldsymbol{\sigma}^H\|_0\, \|\boldsymbol{\sigma}^h\|_0\, (h^{-1/2}\|\nabla \mathbf{u}^h\|_0)\\
&\quad = \|\boldsymbol{\sigma}^H\|_0 \left[C_2 h^{-3/2} + C_2 h^{-1} + C_2^h h^{-1/2}\right]\, \|\boldsymbol{\sigma}^h\|_0\, \|\nabla \mathbf{u}^h\|_0\\
&\quad \le (M + NH^{3/2}) \left[\frac{1}{4\epsilon_1}\|\boldsymbol{\sigma}^h\|_0^2 + \epsilon_1\, (C_2 h^{-3/2} + C_2 h^{-1} + C_2^h h^{-1/2})^2\, \|\nabla \mathbf{u}^h\|_0^2\right]. \qquad (3.12)
\end{aligned}
$$

Also, by (2.13)–(2.14) and (3.8),

$$
\begin{aligned}
&(g_a(\boldsymbol{\sigma}^h, \nabla \mathbf{u}^H), \boldsymbol{\sigma}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^h), \boldsymbol{\sigma}^h)\\
&\qquad \le C_3 \left[\|\boldsymbol{\sigma}^h\|_0\, \|\nabla \mathbf{u}^H\|_\infty\, \|\boldsymbol{\sigma}^h\|_0 + \|\boldsymbol{\sigma}^H\|_\infty\, \|\nabla \mathbf{u}^h\|_0\, \|\boldsymbol{\sigma}^h\|_0\right]\\
&\qquad \le C_4 \left[h^{-1}\, \|\nabla \mathbf{u}^H\|_0\, \|\boldsymbol{\sigma}^h\|_0^2 + h^{-1}\, \|\boldsymbol{\sigma}^H\|_0\, \|\nabla \mathbf{u}^h\|_0\, \|\boldsymbol{\sigma}^h\|_0\right]\\
&\qquad \le C_4\, h^{-1}\, (M + NH^{3/2}) \left[\left(1 + \frac{1}{4\epsilon_2}\right)\, \|\boldsymbol{\sigma}^h\|_0^2 + \epsilon_2\, \|\nabla \mathbf{u}^h\|_0^2\right], \qquad (3.13)
\end{aligned}
$$

where $\epsilon_2 > 0$. Therefore, (1.15), (2.7), (3.12), and (3.13) imply that

$$
\begin{aligned}
&\Phi((\boldsymbol{\sigma}^h, \mathbf{u}^h), (\boldsymbol{\sigma}^h, \mathbf{u}^h))\\
&\quad \ge A((\boldsymbol{\sigma}^h, \mathbf{u}^h), (\boldsymbol{\sigma}^h, \mathbf{u}^h)) + \lambda B^h(\mathbf{u}^H, \boldsymbol{\sigma}^h, \boldsymbol{\sigma}^h) - \lambda \mid B^h(\mathbf{u}^h, \boldsymbol{\sigma}^H, \boldsymbol{\sigma}^h)\\
&\qquad\qquad + (g_a(\boldsymbol{\sigma}^h, \nabla \mathbf{u}^H), \boldsymbol{\sigma}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla \mathbf{u}^h), \boldsymbol{\sigma}^h) \mid\\
&\quad \ge \left[1 - \lambda(M + NH^{3/2}) \left(\frac{1}{4\epsilon_1} + C_4 h^{-1} \left(1 + \frac{1}{4\epsilon_2}\right)\right)\right] \|\boldsymbol{\sigma}^h\|_0^2\\
&\qquad\qquad + \Big[C_5 \alpha(1-\alpha) - \lambda(M + NH^{3/2})\\
&\qquad\qquad \cdot \Big(\epsilon_1 (C_2 h^{-3/2} + C_2 h^{-1} + C_2^h h^{-1/2})^2 + \epsilon_2 C_4 h^{-1}\Big)\Big]\, \|\nabla \mathbf{u}^h\|_0^2.
\end{aligned}
$$

Choosing ϵ_1, ϵ_2 appropriately and using the Poincaré–Friedrichs inequality, we have

$$
\Phi((\boldsymbol{\sigma}^h, \mathbf{u}^h), (\boldsymbol{\sigma}^h, \mathbf{u}^h)) \ge C^h(\|\boldsymbol{\sigma}^h\|_0^2 + \|\nabla \mathbf{u}^h\|_0^2) \ge C^h(\|\boldsymbol{\sigma}^h\|_0^2 + \|\mathbf{u}^h\|_1^2) \qquad (3.14)
$$

if λ is sufficiently small.

On the other hand, using (3.11), the inverse estimates, and the imbedding of H^1 in L^4,

$$
\begin{aligned}
&B^h(\mathbf{u}^H,\boldsymbol{\sigma}^H,\boldsymbol{\tau}^h)+(g_a(\boldsymbol{\sigma}^H,\nabla\mathbf{u}^H),\boldsymbol{\tau}^h)\\
&\quad=((\mathbf{u}^H\cdot\nabla)\boldsymbol{\sigma}^H,\boldsymbol{\tau}^h)_h+\frac{1}{2}(\nabla\cdot\mathbf{u}^H\,\boldsymbol{\sigma}^H,\boldsymbol{\tau}^h)\ +\langle\boldsymbol{\sigma}^{H^+}-\boldsymbol{\sigma}^{H^-},\boldsymbol{\tau}^{h^-}\rangle_{h,\mathbf{u}^H}\\
&\qquad\qquad+(g_a(\boldsymbol{\sigma}^H,\nabla\mathbf{u}^H),\boldsymbol{\tau}^h)\\
&\quad\le C_6\left[\|\mathbf{u}^H\|_{0,4}\,\|\nabla\boldsymbol{\sigma}^H\|_{0,4,h}\|\boldsymbol{\tau}^h\|_0+\|\nabla\mathbf{u}^H\|_0\,\|\boldsymbol{\sigma}^H\|_\infty\|\boldsymbol{\tau}^h\|_0\right]\\
&\qquad\qquad+C_3^h\,\|\boldsymbol{\sigma}^H\|_0\,\|\mathbf{u}^H\|_\infty\,\|\boldsymbol{\tau}^h\|_0+C_7\|\boldsymbol{\sigma}^H\|_0\,\|\nabla\mathbf{u}^H\|_\infty\,\|\boldsymbol{\tau}^h\|_0\\
&\quad\le\Big[C_8\left(\|\nabla\mathbf{u}^H\|_0\,(h^{-3/2}\|\boldsymbol{\sigma}^H\|_0)+\|\nabla\mathbf{u}^H\|_0\,(h^{-1}\|\boldsymbol{\sigma}^H\|_0)\right.\\
&\qquad\qquad\left.+\,\|\boldsymbol{\sigma}^H\|_0(h^{-1}\,\|\nabla\mathbf{u}^H\|_0)\right)+C_4^h\,\|\boldsymbol{\sigma}^H\|_0\,(h^{-1/2}\|\nabla\mathbf{u}^H\|_0)\Big]\,\|\boldsymbol{\tau}^h\|_0\\
&\quad=\left[C_8h^{-3/2}+2C_8h^{-1}+C_4^hh^{-1/2}\right]\,\|\nabla\mathbf{u}^H\|_0\,\|\boldsymbol{\sigma}^H\|_0\,\|\boldsymbol{\tau}^h\|_0\,. \qquad (3.15)
\end{aligned}
$$

Hence, using (3.5), (3.14), and (3.15), we have

$$
\|(\boldsymbol{\sigma}^h,\mathbf{u}^h)\|_{(L^2(\Omega))^{2\times2}\times\mathbf{H}_0^1(\Omega)}\le C^h(\|\nabla\mathbf{u}^H\|_0\|\boldsymbol{\sigma}^H\|_0+\|\mathbf{f}\|_{-1})\,.
$$

(ii) Φ *is continuous.* Using the imbedding of H^1 in L^4, (2.13)–(2.14), (3.7)–(3.9), (3.11), and the Poincaré–Friedrichs inequality, we have

$$
\begin{aligned}
&B^h(\mathbf{u}^H,\boldsymbol{\sigma}^h,\boldsymbol{\tau}^h)+B^h(\mathbf{u}^h,\boldsymbol{\sigma}^H,\boldsymbol{\tau}^h)\\
&\quad=((\mathbf{u}^H\cdot\nabla)\boldsymbol{\sigma}^h,\boldsymbol{\tau}^h)_h+\frac{1}{2}(\nabla\cdot\mathbf{u}^H\,\boldsymbol{\sigma}^h,\boldsymbol{\tau}^h)+\langle\boldsymbol{\sigma}^{h^+}-\boldsymbol{\sigma}^{h^-},\boldsymbol{\tau}^{h^-}\rangle_{h,\mathbf{u}^H}\\
&\qquad+((\mathbf{u}^h\cdot\nabla)\boldsymbol{\sigma}^H,\boldsymbol{\tau}^h)_h+\frac{1}{2}(\nabla\cdot\mathbf{u}^h\,\boldsymbol{\sigma}^H,\boldsymbol{\tau}^h)+\langle\boldsymbol{\sigma}^{H^+}-\boldsymbol{\sigma}^{H^-},\boldsymbol{\tau}^{h^-}\rangle_{h,\mathbf{u}^h}\\
&\quad\le\big[C_9\left(\,\|\mathbf{u}^H\|_{0,4}\,\|\nabla\boldsymbol{\sigma}^h\|_{0,4,h}+\|\nabla\mathbf{u}^H\|_0\,\|\boldsymbol{\sigma}^h\|_\infty\right)+C_5^h\,\|\boldsymbol{\sigma}^h\|_0\,\|\mathbf{u}^H\|_\infty\\
&\qquad+C_{10}\,\left(\|\mathbf{u}^h\|_{0,4}\,\|\nabla\boldsymbol{\sigma}^H\|_{0,4,h}+\|\nabla\mathbf{u}^h\|_0\,\|\boldsymbol{\sigma}^H\|_\infty\right)\\
&\qquad+C_6^h\,\|\boldsymbol{\sigma}^H\|_0\,\|\mathbf{u}^h\|_\infty\big]\,\|\boldsymbol{\tau}^h\|_0\\
&\quad\le\Big[C_{11}\,\Big(\|\nabla\mathbf{u}^H\|_0\,(h^{-3/2}\|\boldsymbol{\sigma}^h\|_0)+\|\nabla\mathbf{u}^H\|_0\,(h^{-1}\|\boldsymbol{\sigma}^h\|_0)\\
&\qquad+\|\nabla\mathbf{u}^h\|_0\,(h^{-3/2}\|\boldsymbol{\sigma}^H\|_0)+\|\nabla\mathbf{u}^h\|_0\,(h^{-1}\|\boldsymbol{\sigma}^H\|_0)\Big)\\
&\qquad+C_7^h\,\Big(\|\boldsymbol{\sigma}^h\|_0\,(h^{-1/2}\|\nabla\mathbf{u}^H\|_0)+\|\boldsymbol{\sigma}^H\|_0\,(h^{-1/2}\|\nabla\mathbf{u}^h\|_0)\Big)\Big]\,\|\boldsymbol{\tau}^h\|_0\\
&\quad\le(M+NH^{3/2})(C_{11}h^{-3/2}+C_{11}h^{-1}+C_7^hh^{-1/2})\,\left[\|\boldsymbol{\sigma}^h\|_0+\|\nabla\mathbf{u}^h\|_0\right]\,\|\boldsymbol{\tau}^h\|_0\\
&\quad\le C_8^h\,(\|\boldsymbol{\sigma}^h\|_0+\|\nabla\mathbf{u}^h\|_0)\,\|\boldsymbol{\tau}^h\|_0\\
&\quad\le C_9^h\,\sqrt{\|\boldsymbol{\sigma}^h\|_0^2+\|\nabla\mathbf{u}^h\|_0^2}\,\sqrt{\|\boldsymbol{\tau}^h\|_0^2+\|\nabla\mathbf{v}^h\|_0^2}\,. \qquad (3.16)
\end{aligned}
$$

By (2.13)–(2.14) and (3.8),

$$
\begin{aligned}
&(g_a(\boldsymbol{\sigma}^h,\nabla\mathbf{u}^H),\boldsymbol{\tau}^h)+(g_a(\boldsymbol{\sigma}^H,\nabla\mathbf{u}^h),\boldsymbol{\tau}^h)\\
&\qquad\le C_{12}\left[\,\|\boldsymbol{\sigma}^h\|_0\,\|\nabla\mathbf{u}^H\|_\infty+\|\boldsymbol{\sigma}^H\|_\infty\,\|\nabla\mathbf{u}^h\|_0\,\right]\,\|\boldsymbol{\tau}^h\|_0\\
&\qquad\le C_{13}\left[\|\boldsymbol{\sigma}^h\|_0\,(h^{-1}\,\|\nabla\mathbf{u}^H\|_0)+h^{-1}\,\|\boldsymbol{\sigma}^H\|_0\,\|\nabla\mathbf{u}^h\|_0\,\right]\,\|\boldsymbol{\tau}^h\|_0\\
&\qquad\le C_{13}\,(M+NH^{3/2})\,h^{-1}\left[\,\|\boldsymbol{\sigma}^h\|_0+\|\nabla\mathbf{u}^h\|_0\,\right]\,\|\boldsymbol{\tau}^h\|_0\\
&\qquad\le C_{10}^h\sqrt{\|\boldsymbol{\sigma}^h\|_0^2+\|\nabla\mathbf{u}^h\|_0^2}\,\sqrt{\|\boldsymbol{\tau}^h\|_0^2+\|\nabla\mathbf{v}^h\|_0^2}\,. \qquad (3.17)
\end{aligned}
$$

Therefore, using (1.14) and (3.16)–(3.17), we obtain

$$
\Phi((\boldsymbol{\sigma}^h,\mathbf{u}^h),(\boldsymbol{\tau}^h,\mathbf{v}^h))\le C^h\sqrt{\|\boldsymbol{\sigma}^h\|_0^2+\|\nabla\mathbf{u}^h\|_0^2}\,\sqrt{\|\boldsymbol{\tau}^h\|_0^2+\|\nabla\mathbf{v}^h\|_0^2}\,. \qquad (3.18)
$$

(iii) By the Lax–Milgram theorem, (3.4) has a unique solution $(\boldsymbol{\sigma}^h, \mathbf{u}^h) \in \boldsymbol{\Sigma}^h \times \mathbf{V}^h$. Since $(\mathbf{X}^h, S^h)$ satisfies the LBB condition (2.4), there exists $p^h \in S^h$ such that

$$(\boldsymbol{\sigma}^h, d(\mathbf{v}^h)) + 2(1-\alpha)\,(d(\mathbf{u}^h), d(\mathbf{v}^h)) - (p^h, \nabla\cdot\mathbf{v}^h) = (\mathbf{f}, \mathbf{v}^h) \quad \forall \mathbf{v}^h \in \mathbf{X}^h\,. \qquad \square$$

REMARK 3.3. *The above theorem establishes existence and uniqueness of the solution $(\mathbf{u}^h, \boldsymbol{\sigma}^h, p^h)$ in the fixed finite element space (X^h, S^h, Σ^h). (Note the dependence of the continuity and coercivity constants on h.) In the error estimates below, more involved analysis is used to establish bounds in which the constants are independent of h.*

REMARK 3.4. *The two-grid method is easily extended to a multigrid method as follows. Consider a sequence of mesh spacings, h_i, $i = 1, 2, \ldots, K$, such that $H > h_1 > h_2 > \cdots > h_K$, and let $(\mathbf{X}^{h_i}, S^{h_i}, \boldsymbol{\Sigma}^{h_i})$ be the finite element space corresponding to each mesh size h_i. Then in Step 2, we solve the linear problems.*

- *Set $(\boldsymbol{\sigma}^{h_0}, \mathbf{u}^{h_0}) = (\boldsymbol{\sigma}^H, \mathbf{u}^H)$.*
- *For $i = 1, 2, \ldots, K$, find $\mathbf{u}^{h_i} \in \mathbf{X}^{h_i}$, $p^{h_i} \in S^{h_i}$, $\boldsymbol{\sigma}^{h_i} \in \boldsymbol{\Sigma}^{h_i}$ such that*

$$\begin{aligned}
&(\boldsymbol{\sigma}^{h_i}, \boldsymbol{\tau}^{h_i}) + \lambda\left[B^h(\mathbf{u}^{h_{i-1}}, \boldsymbol{\sigma}^{h_i}, \boldsymbol{\tau}^{h_i}) + B^h(\mathbf{u}^{h_i}, \boldsymbol{\sigma}^{h_{i-1}}, \boldsymbol{\tau}^{h_i})\right.\\
&\quad \left.+(g_a(\boldsymbol{\sigma}^{h_i}, \nabla\mathbf{u}^{h_{i-1}}), \boldsymbol{\tau}^{h_i}) + (g_a(\boldsymbol{\sigma}^{h_{i-1}}, \nabla\mathbf{u}^{h_i}), \boldsymbol{\tau}^{h_i})\right] - 2\alpha(d(\mathbf{u}^{h_i}), \boldsymbol{\tau}^{h_i})\\
&\quad = \lambda\left[B^h(\mathbf{u}^{h_{i-1}}, \boldsymbol{\sigma}^{h_{i-1}}, \boldsymbol{\tau}^{h_i}) + (g_a(\boldsymbol{\sigma}^{h_{i-1}}, \nabla\mathbf{u}^{h_{i-1}}), \boldsymbol{\tau}^{h_i})\right] \quad \forall \boldsymbol{\tau}^{h_i} \in \boldsymbol{\Sigma}^{h_i}\,,\\
&(\boldsymbol{\sigma}^{h_i}, d(\mathbf{v}^{h_i})) + 2(1-\alpha)(d(\mathbf{u}^{h_i}), d(\mathbf{v}^{h_i})) - (p^{h_i}, \nabla\cdot\mathbf{v}^{h_i})\\
&\qquad\qquad = (\mathbf{f}, \mathbf{v}^{h_i}) \quad \forall \mathbf{v}^{h_i} \in \mathbf{X}^{h_i}\,,\\
&(q^{h_i}, \nabla\cdot\mathbf{u}^{h_i}) = 0 \quad \forall q^{h_i} \in S^{h_i}\,.
\end{aligned}$$

Extending the analysis in the proof of Theorem 3.2 to the multigrid method is straightforward. Numerical tests by the multigrid method will be discussed in section 5.

4. Error estimate. In this section we derive an error estimation for the two-grid method.

THEOREM 4.1. *The unique solution $(\mathbf{u}^h, \boldsymbol{\sigma}^h, p^h)$ to the linear problem in Step 2 satisfies*

$$\begin{aligned}
&\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\|_0 + \|\nabla(\mathbf{u} - \mathbf{u}^h)\|_0\\
&\qquad \le\ \overline{C}\,h^2 + C\,\lambda^{1/2}\,\Big[M^2\,h + Mh^{1/2}(\,\|\nabla(\mathbf{u} - \mathbf{u}^H)\|_0 + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^H\|_0\,)\\
&\qquad\quad +h^{-3/2}\,\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^H\|_0\,\|\nabla(\mathbf{u} - \mathbf{u}^H)\|_0\Big]
\end{aligned}$$

if λ is sufficiently small, where $\overline{C}$, C are constants independent of H and h.

Proof. If $(\mathbf{u}, \boldsymbol{\sigma})$ is an exact solution of the problem (1.1)–(1.4) with $\mathbf{h} = \mathbf{0}$, $g = 0$, and $\mathbf{u}_\Gamma = \mathbf{0}$, it also satisfies (1.11)–(1.12). Subtracting (3.1) and (3.2) from (1.11) and (1.12), respectively, we have

$$\begin{aligned}
&(\boldsymbol{\sigma} - \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) - 2\alpha(d(\mathbf{u} - \mathbf{u}^h), \boldsymbol{\tau}^h)\\
&\quad +\lambda\left[((\mathbf{u}, \nabla)\boldsymbol{\sigma}, \boldsymbol{\tau}^h) - B^h(\mathbf{u}^H, \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) - B^h(\mathbf{u}^h, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h)\right.\\
&\quad \left.+(g_a(\boldsymbol{\sigma}, \nabla\mathbf{u}), \boldsymbol{\tau}^h) - (g_a(\boldsymbol{\sigma}^h, \nabla\mathbf{u}^H), \boldsymbol{\tau}^h) - (g_a(\boldsymbol{\sigma}^H, \nabla\mathbf{u}^h), \boldsymbol{\tau}^h)\right]\\
&\qquad = -\lambda\left[B^h(\mathbf{u}^H, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla\mathbf{u}^H), \boldsymbol{\tau}^h)\right] \quad \forall \boldsymbol{\tau}^h \in \Sigma^h, \qquad (4.1)
\end{aligned}$$

$$(\boldsymbol{\sigma} - \boldsymbol{\sigma}^h, d(\mathbf{v}^h)) + 2(1-\alpha)(d(\mathbf{u} - \mathbf{u}^h), d(\mathbf{v}^h)) = 0 \quad \forall \mathbf{v}^h \in \mathbf{V}^h. \qquad (4.2)$$

Note that the exact solution $(\boldsymbol{\sigma}, \mathbf{u})$ satisfies the relation

$$((\mathbf{u}\cdot\nabla)\boldsymbol{\sigma}, \boldsymbol{\tau}^h) = B^h(\mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\tau}^h)$$

since $\boldsymbol{\sigma}$ is continuous and $\nabla\cdot\mathbf{u} = 0$. Multiplying (4.2) by 2α and adding to (4.1), we get

$$\begin{aligned} &A((\boldsymbol{\sigma}-\boldsymbol{\sigma}^h, \mathbf{u}-\mathbf{u}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h)) \\ &\quad +\lambda\left[B^h(\mathbf{u}, \boldsymbol{\sigma}, \boldsymbol{\tau}^h) - B^h(\mathbf{u}^H, \boldsymbol{\sigma}^h, \boldsymbol{\tau}^h) - B^h(\mathbf{u}^h, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h)\right. \\ &\quad \left.+(g_a(\boldsymbol{\sigma}, \nabla\mathbf{u}), \boldsymbol{\tau}^h) - (g_a(\boldsymbol{\sigma}^h, \nabla\mathbf{u}^H), \boldsymbol{\tau}^h) - (g_a(\boldsymbol{\sigma}^H, \nabla\mathbf{u}^h), \boldsymbol{\tau}^h)\right] \\ &\qquad\qquad = -\lambda\left[B^h(\mathbf{u}^H, \boldsymbol{\sigma}^H, \boldsymbol{\tau}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla\mathbf{u}^H), \boldsymbol{\tau}^h)\right]. \end{aligned} \tag{4.3}$$

Let $\tilde{\boldsymbol{\sigma}}^h$ be the L^2 projection of $\boldsymbol{\sigma}$ in $\boldsymbol{\Sigma}^h$, and let $\tilde{\mathbf{u}}^h \in \mathbf{V}^h$ be defined by

$$(\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h), \nabla\mathbf{v}^h) = 0 \qquad \forall \mathbf{v}^h \in \mathbf{V}^h.$$

We have then the following standard results from [10]: for $\mathbf{u} \in (\mathbf{H}^3(\Omega))$ and $\boldsymbol{\sigma} \in (\mathbf{H}^2(\Omega))$

$$\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0 \le C\,h^2\|\mathbf{u}\|_3\,, \tag{4.4}$$

$$\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0 + h\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_1 \le Ch^2\|\boldsymbol{\sigma}\|_2\,, \tag{4.5}$$

$$\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_{0,\Gamma^h} \le Ch^{3/2}\|\boldsymbol{\sigma}\|_2\,. \tag{4.6}$$

Note that $A((\boldsymbol{\sigma}-\boldsymbol{\sigma}^h, \mathbf{u}-\mathbf{u}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h))$ in (4.3) can be written as $A((\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h, \mathbf{u}-\tilde{\mathbf{u}}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h)) + A((\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h, \tilde{\mathbf{u}}^h-\mathbf{u}^h), (\boldsymbol{\tau}^h, \mathbf{v}^h))$ by the definition of A. Adding and subtracting $\tilde{\mathbf{u}}^h$, $\tilde{\boldsymbol{\sigma}}^h$ in (4.3) and letting $\boldsymbol{\tau}^h = \tilde{\boldsymbol{\sigma}}^h - \boldsymbol{\sigma}^h$, $\mathbf{v}^h = \tilde{\mathbf{u}}^h - \mathbf{u}^h$, we have

$$\begin{aligned} &A((\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h, \tilde{\mathbf{u}}^h-\mathbf{u}^h), (\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h, \tilde{\mathbf{u}}^h-\mathbf{u}^h)) + \lambda B^h(\mathbf{u}^H, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) \\ &= -A((\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h, \mathbf{u}-\tilde{\mathbf{u}}^h), (\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h, \tilde{\mathbf{u}}^h-\mathbf{u}^h)) - \lambda\Big[B^h(\mathbf{u}-\tilde{\mathbf{u}}^h, \boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) \\ &\quad + B^h(\mathbf{u}-\tilde{\mathbf{u}}^h, \boldsymbol{\sigma}^H-\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) + B^h(\tilde{\mathbf{u}}^h-\mathbf{u}^h, \boldsymbol{\sigma}^H, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) \\ &\quad + B^h(\mathbf{u}^H, \boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) + B^h(\mathbf{u}-\mathbf{u}^H, \boldsymbol{\sigma}-\boldsymbol{\sigma}^H, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) \\ &\quad + (g_a(\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h, \nabla\mathbf{u}^H), \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) + (g_a(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h, \nabla\mathbf{u}^H), \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) \\ &\quad + (g_a(\boldsymbol{\sigma}^H, \nabla(\mathbf{u}-\tilde{\mathbf{u}}^h), \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) + (g_a(\boldsymbol{\sigma}^H, \nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h), \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) \\ &\quad + (g_a(\boldsymbol{\sigma}-\boldsymbol{\sigma}^H, \nabla(\mathbf{u}-\mathbf{u}^H), \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\Big]. \end{aligned} \tag{4.7}$$

To prove the theorem, we will get a bound for the right-hand side of (4.7) in terms of h, H, $\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0$, $\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0$, $\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0$, and $\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0$. For simplicity we assume that $h < H < 1$ throughout this proof.

(i) *A term.* From the definition of A, and using (4.4), (4.5), and Young's inequality, we have

$$\begin{aligned} &A((\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h, \mathbf{u}-\tilde{\mathbf{u}}^h), (\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h, \tilde{\mathbf{u}}^h-\mathbf{u}^h)) \\ &\quad\le \|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0 + 2\alpha\,\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0 \\ &\qquad + 2\alpha\,\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0 + 4\alpha(1-\alpha)\,\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0 \\ &\quad\le \frac{1}{4\epsilon_1}\,\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0^2 + \epsilon_1\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2 + \frac{(2\alpha)^2}{4\epsilon_2}\,\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0^2 + \epsilon_2\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2 \end{aligned}$$

$$
\begin{aligned}
&+ \frac{(2\alpha)^2}{4\epsilon_3}\,\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0^2 + \epsilon_3\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2 + \frac{(4\alpha(1-\alpha))^2}{4\epsilon_4}\,\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0^2\\
&+ \epsilon_4\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\\
&= \left(\frac{1}{4\epsilon_1}+\frac{\alpha^2}{\epsilon_3}\right)\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0^2 + \left(\frac{\alpha^2}{\epsilon_2}+\frac{4\alpha^2(1-\alpha)^2}{\epsilon_4}\right)\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0^2\\
&\qquad + (\epsilon_1+\epsilon_2)\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2 + (\epsilon_3+\epsilon_4)\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\\
&\le C_1\,M^2\left(\frac{1}{4\epsilon_1}+\frac{\alpha^2}{\epsilon_3}+\frac{\alpha^2}{\epsilon_2}+\frac{4\alpha^2(1-\alpha)^2}{\epsilon_4}\right)h^4 \qquad (4.8)\\
&\qquad + (\epsilon_1+\epsilon_2)\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2 + (\epsilon_3+\epsilon_4)\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\,,
\end{aligned}
$$

where ϵ_i, $i=1,2,3,4$, are positive constants.

(ii) B^h *terms.* First, note that

$$\langle \boldsymbol{\sigma}^+-\boldsymbol{\sigma}^-, {\boldsymbol{\tau}^h}^-\rangle_{h,\mathbf{u}-\tilde{\mathbf{u}}^h} = 0 \qquad \forall \boldsymbol{\tau}^h\in\boldsymbol{\Sigma}^h,$$

since, by continuity, the jump of $\boldsymbol{\sigma}$ across any element boundary is zero. Using (2.5), the embedding of H^1 in L^4, H^2 in L^∞, $W^{2,2}$ in $W^{1,4}$, (2.13), (3.7)–(3.9), (3.11), and (4.4), we obtain a bound for the first and third B^h terms:

$$
\begin{aligned}
&B^h(\mathbf{u}-\tilde{\mathbf{u}}^h,\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) + B^h(\tilde{\mathbf{u}}^h-\mathbf{u}^h,\boldsymbol{\sigma}^H,\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\\
&\quad\le C_2\|\mathbf{u}-\tilde{\mathbf{u}}^h\|_{0,4}\,\|\nabla\boldsymbol{\sigma}\|_{0,4}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0 + \|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,\|\boldsymbol{\sigma}\|_\infty\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\qquad + \|\tilde{\mathbf{u}}^h-\mathbf{u}^h\|_{0,4}\,\|\nabla\boldsymbol{\sigma}^H\|_{0,4,h}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\qquad + \|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\,\|\boldsymbol{\sigma}^H\|_\infty\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\qquad + h^{-1}\,\|\boldsymbol{\sigma}^H\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,\|\tilde{\mathbf{u}}^h-\mathbf{u}^h\|_\infty\\
&\quad\le C_3\Big[\,\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,\|\boldsymbol{\sigma}\|_2 + \|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\,(h^{-3/2}\,\|\boldsymbol{\sigma}^H\|_0)\\
&\qquad + \|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\,(h^{-1}\,\|\boldsymbol{\sigma}^H\|_0)\\
&\qquad + h^{-1}\,\|\boldsymbol{\sigma}^H\|_0\,(h^{-1/2}\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\,\|_0)\Big]\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,,\\
&\quad\le C_4\left[M^2h^2 + h^{-3/2}\,(M+NH^{3/2})\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\right]\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,. \qquad (4.9)
\end{aligned}
$$

The boundedness of the remaining B^h terms can be shown using interpolates of $\mathbf{u}$. We establish the following two estimates. Let $\hat{\mathbf{u}}$ be the P_2 interpolate of $\mathbf{u}$ on V^h. Then, by interpolation properties and (3.7) [3],

$$
\begin{aligned}
\|\mathbf{u}-\tilde{\mathbf{u}}^h\|_\infty &\le \|\mathbf{u}-\hat{\mathbf{u}}\|_\infty + \|\hat{\mathbf{u}}-\tilde{\mathbf{u}}^h\|_\infty\\
&\le C_5\,(\,h^2\,\|\mathbf{u}\|_3 + h^{-1/2}\,\|\nabla(\hat{\mathbf{u}}-\tilde{\mathbf{u}}^h)\|_0)\\
&\le C_5\,(\,M\,h^2 + h^{-1/2}\,\|\nabla(\hat{\mathbf{u}}-\mathbf{u})\|_0 + h^{-1/2}\,\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,)\\
&\le C_6\,M\,(\,h^2+h^{3/2}\,)\\
&\le C_7\,M\,h^{3/2}\,. \qquad (4.10)
\end{aligned}
$$

Also, by (3.10) and (4.5)–(4.6),

$$
\begin{aligned}
\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_{0,\Gamma^h} &\le \|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_{0,\Gamma^h} + \|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^H\|_{0,\Gamma^h}\\
&\le C_8\left[M\,h^{3/2} + h^{-1/2}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^H\|_0\right]\\
&\le C_8\left[M\,h^{3/2} + h^{-1/2}\,(\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}\|_0 + \|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\,)\right]\\
&\le C_9\left[M\,h^{3/2} + h^{-1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\right]. \qquad (4.11)
\end{aligned}
$$

Then, using (2.6), (3.7)–(3.10), (4.4), (4.10), and (4.11), we obtain a bound for the second B^h term:

$$\begin{aligned}
&B^h(\mathbf{u}-\tilde{\mathbf{u}}^h,\boldsymbol{\sigma}^H-\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\\
&\quad=-(((\mathbf{u}-\tilde{\mathbf{u}}^h)\cdot\nabla)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}^H-\boldsymbol{\sigma})_h-\frac{1}{2}(\nabla\cdot(\mathbf{u}-\tilde{\mathbf{u}}^h)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}^H-\boldsymbol{\sigma})\\
&\qquad+\langle(\boldsymbol{\sigma}^H-\boldsymbol{\sigma})^-,(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)^--(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)^+\rangle_{h,\mathbf{u}-\tilde{\mathbf{u}}^h}\\
&\qquad\le C_{10}\Big[\|\mathbf{u}-\tilde{\mathbf{u}}^h\|_{0,4}\,\|\nabla(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\|_{0,4,h}\,\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_0\\
&\qquad\quad+\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_\infty\,\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_0\\
&\qquad\quad+\|\mathbf{u}-\tilde{\mathbf{u}}^h\|_\infty\,(\,h^{-1/2}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,)\,\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_{0,\Gamma^h}\Big]\\
&\qquad\le C_{11}\Big[\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_0\,(\,h^{-3/2}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,)\\
&\qquad\quad+\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_0\,(\,h^{-1}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,)\\
&\qquad\quad+M\,h^{-1/2}\,(h^{3/2}\,)\,(\,M\,h^{3/2}+h^{-1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\,)\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\Big]\\
&\qquad\le C_{12}\Big[M\,h^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0+M^2\,h^{5/2}\Big]\;\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,. \qquad (4.12)
\end{aligned}$$

Consider the fourth B^h term,

$$\begin{aligned}
&B^h(\mathbf{u}^H,\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h,\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\\
&\qquad=-\,((\mathbf{u}^H\cdot\nabla)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h)_h-\frac{1}{2}(\nabla\cdot\mathbf{u}^H(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h) \qquad (4.13)\\
&\qquad\quad+\langle(\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h)^-,(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)^--(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)^+\rangle_{h,\mathbf{u}^H}
\end{aligned}$$

by (2.6). Since $\nabla\cdot\mathbf{u}=0$, the second term can be written as $\frac{1}{2}(\nabla\cdot(\mathbf{u}^H-\mathbf{u})(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h)$, and using (3.8) and (4.5),

$$\begin{aligned}
\frac{1}{2}(\nabla\cdot(\mathbf{u}^H-\mathbf{u})(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h)\;&\le\;C_{13}\,\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_\infty\,\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0\\
&\le C_{14}\,\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0\,(h^{-1}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0)\,(M\,h^2)\\
&\le C_{14}M\,h\,\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,. \qquad (4.14)
\end{aligned}$$

For the first term in (4.13), let $\check{\mathbf{u}}$ be the P_1 continuous interpolate of $\mathbf{u}$ on $\mathbf{V}^h$. Then, by the imbedding of H^1 in L^4, the Poincaré–Friedrichs inequality, and interpolation properties,

$$\begin{aligned}
\|\mathbf{u}^H-\check{\mathbf{u}}\|_{0,4}\le C_{15}\,\|\nabla(\mathbf{u}^H-\check{\mathbf{u}})\|_0&\le C_{15}(\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0+\|\nabla(\mathbf{u}-\check{\mathbf{u}})\|_0)\\
&\le C_{16}\,(\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0+h\,M)\,. \qquad (4.15)
\end{aligned}$$

Since $\nabla(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)$ is in P_0 on each K, then $(\check{\mathbf{u}}\cdot\nabla)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)$ is in P_1, and thus

$$((\check{\mathbf{u}}\cdot\nabla)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h)=0 \qquad (4.16)$$

since $\tilde{\boldsymbol{\sigma}}^h$ is the L^2 projection of $\boldsymbol{\sigma}$ in $\boldsymbol{\Sigma}^h$. Now, using (3.9), (4.5), (4.15), and (4.16), we have

$$\begin{aligned}
((\mathbf{u}^H\cdot\nabla)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h)_h\;&=\;(((\mathbf{u}^H-\check{\mathbf{u}})\cdot\nabla)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h),\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h)_h\\
&\le C_{17}\|\mathbf{u}^H-\check{\mathbf{u}}\|_{0,4}\,\|\nabla(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\|_{0,4,h}\,\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0\\
&\le C_{18}(\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0+M\,h)\,(h^{-3/2}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0)\,(M\,h^2)\\
&\le C_{18}M\,h^{1/2}(\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0+M\,h)\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,. \qquad (4.17)
\end{aligned}$$

For the third term in (4.13), we use the following result obtained by the imbedding theorem of $W^{1,4}$ in L^∞, (3.6), and (4.4):

$$\begin{aligned}
\|\mathbf{u}^H\|_\infty &\le \|\mathbf{u}\|_\infty + \|\mathbf{u}-\tilde{\mathbf{u}}^h\|_\infty + \|\tilde{\mathbf{u}}^h-\mathbf{u}^H\|_\infty \\
&\le C_{19}\left[M + \|\mathbf{u}-\tilde{\mathbf{u}}^h\|_{1,4} + h^{-1/2}\,\|\tilde{\mathbf{u}}^h-\mathbf{u}^H\|_{0,4}\right] \\
&\le C_{20}\left[M + h\,\|\mathbf{u}\|_{2,4} + h^{-1/2}\,(\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u})\|_0 + \|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0)\right] \\
&\le C_{21}\left[M + M\,h + M\,h^{3/2} + h^{-1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\right] \\
&\le C_{22}\left[M + h^{-1/2}\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\right]. \qquad (4.18)
\end{aligned}$$

Using (4.6) and (4.18), the third term in (4.13) becomes

$$\begin{aligned}
&\langle(\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h)^-, (\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)^- - (\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)^+\rangle_{h,\mathbf{u}^H} \\
&\qquad \le C_{23}\,\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_{0,\Gamma^h}\,(h^{-1/2}\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0)\,\|\mathbf{u}^H\|_\infty \\
&\qquad \le C_{24}\,M\,h\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,\left(M + h^{-1/2}\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\right). \qquad (4.19)
\end{aligned}$$

Therefore, by (4.5), (4.14), (4.17), and (4.19), we obtain

$$\begin{aligned}
&B^h(\mathbf{u}^H, \boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) \\
&\qquad \le C_{25}\,\Big[M\,h^{1/2}\,\left(\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0 + M\,h\right) + M\,h\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0 \\
&\qquad\qquad +M\,h\,\left(M + h^{-1/2}\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\right)\Big]\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,. \qquad (4.20)
\end{aligned}$$

We now estimate the last B^h term. Using $\hat{\mathbf{u}}$, the P_2 interpolate of $\mathbf{u}$ on $\mathbf{V}^h$,

$$\begin{aligned}
\|\mathbf{u}-\mathbf{u}^H\|_\infty &\le \|\mathbf{u}-\hat{\mathbf{u}}\|_\infty + \|\hat{\mathbf{u}}-\mathbf{u}^H\|_\infty \\
&\le C_{26}\,\left(h^2\,\|\mathbf{u}\|_3 + h^{-1/2}\,\|\nabla(\hat{\mathbf{u}}-\mathbf{u}^H)\|_0\right) \\
&\le C_{26}\,\left(M\,h^2 + h^{-1/2}\,\|\nabla(\hat{\mathbf{u}}-\mathbf{u})\|_0 + h^{-1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\right) \\
&\le C_{27}\,\left(M\,h^2 + h^{-1/2}\,(h^2\,M) + h^{-1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\right) \\
&\le C_{28}\,\left(M\,h^{3/2} + h^{-1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\right). \qquad (4.21)
\end{aligned}$$

Then, by (3.7)–(3.9), (4.11), and (4.21),

$$\begin{aligned}
&B^h(\mathbf{u}-\mathbf{u}^H, \boldsymbol{\sigma}-\boldsymbol{\sigma}^H, \tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h) \\
&= -\,((((\mathbf{u}-\mathbf{u}^H)\cdot\nabla)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)), \boldsymbol{\sigma}-\boldsymbol{\sigma}^H)_h - \frac{1}{2}(\nabla\cdot(\mathbf{u}-\mathbf{u}^H)(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h), \boldsymbol{\sigma}-\boldsymbol{\sigma}^H) \\
&\qquad\qquad +\langle(\boldsymbol{\sigma}-\boldsymbol{\sigma}^H)^-, (\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)^- - (\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)^+\rangle_{h,\mathbf{u}-\mathbf{u}^H} \\
&\le C_{29}\,\Big[\|\mathbf{u}-\mathbf{u}^H\|_{0,4}\,\|\nabla(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\|_{0,4,h}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0 \\
&\qquad\qquad + \|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_\infty\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0 \\
&\qquad\qquad + \|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_{0,\Gamma^h}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_{0,\Gamma^h}\,\|\mathbf{u}-\mathbf{u}^H\|_\infty\Big] \\
&\le C_{30}\,\Big[\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,(\,h^{-3/2}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,)\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0
\end{aligned}$$

$$
\begin{aligned}
&+ \|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0 \|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0 \,(h^{-1}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0)\\
&+ (\,M\,h^{3/2}+h^{-1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\,)\,(\,h^{-1/2}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,)\\
&\Big(\,M\,h^{3/2}+h^{-1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,\Big)\Big],
\end{aligned}
$$

which implies

$$
\begin{aligned}
&B^h(\mathbf{u}-\mathbf{u}^H,\boldsymbol{\sigma}-\boldsymbol{\sigma}^H,\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\\
&\quad\le C_{31}\Big[h^{-3/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\\
&\quad\quad+(M\,h+h^{-1}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0)\,(\,M\,h^{3/2}+h^{-1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,)\Big]\;\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,.
\end{aligned}
\tag{4.22}
$$

Therefore, by (4.9), (4.12), (4.20), and (4.22), we have

$$
\begin{aligned}
&\text{Sum of } B^h \text{ terms}\\
&\le C_{32}\Big[M^2\,h^2+h^{-3/2}\,(M+NH^{3/2})\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\\
&\quad+M\,h^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0+M^2\,h^{5/2}\\
&\quad+M\,h^{1/2}\,(\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0+M\,h)+M\,h\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\\
&\quad+M\,h\,(M+h^{-1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0)+h^{-3/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\\
&\quad+(M\,h+h^{-1}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0)\,(M\,h^{3/2}+h^{-1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0)\Big]\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&=C_{32}\Big[M^2\,h^2+M^2\,h^{5/2}+M^2\,h^{3/2}+M^2\,h+M^2\,h^{5/2}\\
&\quad+M\,(3\,h^{1/2}+h)\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0+2\,M\,h^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\\
&\quad+2\,h^{-3/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\Big]\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\quad+h^{-3/2}\,(M+NH^{3/2})\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\le C_{33}\Big[M^2\,h+M\,h^{1/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0+M\,h^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\\
&\quad+h^{-3/2}\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\Big]\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\quad+h^{-3/2}\,(M+NH^{3/2})\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,.
\end{aligned}
$$

We obtain, by Young's inequality, that

$$
\begin{aligned}
\text{Sum of } B^h \text{ terms } \le\; & C_{33}\big[\epsilon_5\,\mathcal{R}^2+\frac{1}{4\epsilon_5}\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2+\frac{1}{4\epsilon_6}\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\\
&+C_{34}\,\epsilon_6\,h^{-3}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2\big],
\end{aligned}
\tag{4.23}
$$

where

$$
\begin{aligned}
\mathcal{R}\le\; & M^2h+Mh^{1/2}(\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0+\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0)\\
&+h^{-3/2}\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\,.
\end{aligned}
$$

(iii) g_a *terms.* Using (2.13)–(2.14), (3.8), and (4.4)–(4.5), we have

$$
\begin{aligned}
&\text{Sum of } g_a \text{ terms}\\
&\le C_{35}\Big[\|\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}^h\|_0\,\|\nabla\mathbf{u}^H\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_\infty+\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,\|\nabla\mathbf{u}^H\|_\infty\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\quad+\|\boldsymbol{\sigma}^H\|_0\,\|\nabla(\mathbf{u}-\tilde{\mathbf{u}}^h)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_\infty+\|\boldsymbol{\sigma}^H\|_\infty\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0
\end{aligned}
$$

$$
\begin{aligned}
&\qquad + \|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_\infty\Big]\\
&\le C_{36}\Big[(M\,h^2)\,(\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0+\|\nabla\mathbf{u}\|_0)\,(h^{-1}\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0)\\
&\qquad + (h^{-1}\|\nabla\mathbf{u}^H\|_0)\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2\\
&\qquad + (\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_0+\|\boldsymbol{\sigma}\|_0)\,(M\,h^2)\,(h^{-1}\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0)\\
&\qquad + (h^{-1}\|\boldsymbol{\sigma}^H\|_0)\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\qquad + \|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,(h^{-1}\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0)\Big]\\
&\le C_{37}\Big[(M+NH^{3/2})\,h^{-1}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2\\
&\qquad + \big(\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0+\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_0+2\,M\big)\,M\,h\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\qquad + h^{-1}\,(M+NH^{3/2})\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\qquad + h^{-1}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\Big]\\
&= C_{37}\Big[h^{-1}\,(M+NH^{3/2})\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2\\
&\qquad + \big(2\,M^2h+M\,h\,(\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0+\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_0)\\
&\qquad + h^{-1}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\,\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\big)\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\\
&\qquad + h^{-1}\,(M+NH^{3/2})\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0\Big]\,.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\text{Sum of } g_a \text{ terms } \le\; & C_{37}\Big[\epsilon_7\mathcal{U}^2+\frac{1}{4\epsilon_7}\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2+\frac{1}{4\epsilon_8}\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\\
&+ C_{38}\,\epsilon_8\,h^{-2}\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2+h^{-1}\,(M+NH^{3/2})\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2\Big]\,,
\end{aligned}
\tag{4.24}
$$

where

$$
\mathcal{U}\le 2\,M^2h+M\,h(\|\nabla(\mathbf{u}^H-\mathbf{u})\|_0+\|\boldsymbol{\sigma}^H-\boldsymbol{\sigma}\|_0)+h^{-1}\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^H\|_0\|\nabla(\mathbf{u}-\mathbf{u}^H)\|_0\,.
$$

(iv) Since $B^h(\mathbf{u}^H,\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h,\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h)\ge 0$, (4.7) and the estimates (4.8), (4.23), and (4.24) imply that

$$
\begin{aligned}
&A((\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h,\tilde{\mathbf{u}}^h-\mathbf{u}^h),(\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h,\tilde{\mathbf{u}}^h-\mathbf{u}^h))\,.\\
&\quad\le C_1\,M^2\,\left(\frac{1}{4\epsilon_1}+\frac{\alpha^2}{\epsilon_3}+\frac{\alpha^2}{\epsilon_2}+\frac{4\alpha^2(1-\alpha)^2}{\epsilon_4}\right)h^4\\
&\qquad + (\epsilon_1+\epsilon_2)\,\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2+(\epsilon_3+\epsilon_4)\,\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\\
&\qquad + \lambda C_{39}\left[\epsilon_5\,\mathcal{R}^2+\left(\frac{1}{4\epsilon_5}+C_{34}\,\epsilon_6\,h^{-3}\right)\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2+\frac{1}{4\epsilon_6}\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\right.\\
&\qquad + \epsilon_7\,\mathcal{U}^2+\left(\frac{1}{4\epsilon_7}+C_{38}\,\epsilon_8\,h^{-2}+h^{-1}\,(M+NH^{3/2})\right)\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2\\
&\qquad \left.+\ \frac{1}{4\epsilon_8}\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\right]\,.
\end{aligned}
$$

Choosing sufficiently small λ and ϵ_i appropriately for $i=1,2,\dots,8$ and using the coercivity of A, we have

$$
\|\tilde{\boldsymbol{\sigma}}^h-\boldsymbol{\sigma}^h\|_0^2+\|\nabla(\tilde{\mathbf{u}}^h-\mathbf{u}^h)\|_0^2\le\overline{C}\,h^4+C\lambda(\mathcal{R}^2+\mathcal{U}^2)\,.
$$

Hence

$$\begin{aligned}\|\tilde{\boldsymbol{\sigma}}^h - \boldsymbol{\sigma}^h\|_0 + \|\nabla(\tilde{\mathbf{u}}^h - \mathbf{u}^h)\|_0 \\ \leq \overline{C}\, h^2 + C\lambda^{1/2} \Big[M^2 h + M\, h^{1/2}(\|\nabla(\mathbf{u} - \mathbf{u}^H)\|_0 + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^H\|_0) \\ + h^{-3/2}\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^H\|_0 \|\nabla(\mathbf{u} - \mathbf{u}^H)\|_0 \Big],\end{aligned}$$

and we conclude by (4.4)–(4.5) that

$$\begin{aligned}\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\|_0 + \|\nabla(\mathbf{u} - \mathbf{u}^h)\|_0 \\ \leq \overline{C}\, h^2 + C\lambda^{1/2} \Big[M^2 h + M\, h^{1/2}(\|\nabla(\mathbf{u} - \mathbf{u}^H)\|_0 + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^H\|_0) \\ + h^{-3/2}\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^H\|_0 \|\nabla(\mathbf{u} - \mathbf{u}^H)\|_0 \Big]. \qquad \square\end{aligned}$$

REMARK 4.2. *The result in Theorem* 4.1 *is a rough error estimate. It suggests the optimal scaling* $H = h^{5/6}$ *for linear convergence. However, the actual optimal scaling and convergence rate obtained by numerical experiments with* $\lambda = 1$ *are much better than the theoretical ones.*

5. Numerical results. In this section we present the numerical results obtained by multigrid algorithms. The domain is taken to be the unit square $\Omega = [0,1] \times [0,1]$, and the parameters α, a in the model equations are chosen as 0.5 and 0, respectively. Although we assumed that $\mathbf{h} = \mathbf{0}$, $g = 0$, and $\mathbf{u}_\Gamma = \mathbf{0}$ for simplicity of our analysis, it turned out by various precalculations that these conditions are not necessary for computations by the multigrid method. Hence, the right-hand side functions in (1.1)–(1.3) are appropriately given so that the exact solution is

$$\begin{cases} \mathbf{u} = \begin{pmatrix} \sin(\pi x) y(y-1) \\ \sin(x)\,(x-1) y \cos(\pi y/2)) \end{pmatrix}, \\ p = \cos(2\pi x)\, y(y-1), \\ \boldsymbol{\sigma} = 2\alpha\, d(\mathbf{u}). \end{cases}$$

The example we choose for the numerical test satisfies homogeneous boundary condition for $\mathbf{u}$, but $\operatorname{div} \mathbf{u} = g \neq 0$. Note that the example satisfies the compatibility condition

$$\int_\Omega \operatorname{div} \mathbf{u}\, d\Omega = \int_\Gamma \mathbf{u}_\Gamma \cdot \mathbf{n}\, d\Gamma = 0.$$

We first compute the solution to (2.8)–(2.10) with the replacement $(q^h, \nabla \cdot \mathbf{u}^h) = (g, q^h)$ and nonzero right-hand side $(\mathbf{h}, \boldsymbol{\tau}^h)$ in the discrete constitutive equation using a standard one-grid method. For $\lambda = 0.1$, 1, and 2, the nonlinear equation is solved by the Newton iteration with the initial guess $(\mathbf{u}, p, \boldsymbol{\sigma}) = (\mathbf{0}, 0, \mathbf{0})$. We use the stopping criterion defined by

$$\|\text{change in successive values of } \mathbf{u} \text{ and } \boldsymbol{\sigma}\|^2 < 10^{-9}$$

for each mesh size chosen with the maximum number of iteration set at 15. Since we have an exact solution, we compute errors and the experimental rate of convergence computed by comparing the errors on two grids. All computations were performed on a SUN Enterprise 4000 with six processors.

TABLE 5.1
Results using a one-level method.

λ	h	No. of iter.	L^2 error of $\mathbf{u}$	L^2 rate of $\mathbf{u}$	H^1 error of $\mathbf{u}$	H^1 rate of $\mathbf{u}$	L^2 error of $\boldsymbol{\sigma}$	L^2 rate of $\boldsymbol{\sigma}$
0.1	$\frac{1}{2}$	6	$1.124 \cdot 10^{-2}$		$1.773 \cdot 10^{-1}$		$1.325 \cdot 10^{-1}$	
	$\frac{1}{4}$	5	$1.262 \cdot 10^{-3}$	3.2	$3.858 \cdot 10^{-2}$	2.2	$2.895 \cdot 10^{-2}$	2.2
	$\frac{1}{8}$	5	$1.360 \cdot 10^{-4}$	3.2	$9.341 \cdot 10^{-3}$	2.1	$6.722 \cdot 10^{-3}$	2.1
	$\frac{1}{16}$	5	$1.661 \cdot 10^{-5}$	3.0	$2.358 \cdot 10^{-3}$	2.0	$1.600 \cdot 10^{-3}$	2.1
1.0	$\frac{1}{2}$	8	$1.141 \cdot 10^{-2}$		$1.846 \cdot 10^{-1}$		$1.339 \cdot 10^{-1}$	
	$\frac{1}{4}$	7	$1.589 \cdot 10^{-3}$	2.8	$4.520 \cdot 10^{-2}$	2.0	$2.914 \cdot 10^{-2}$	2.2
	$\frac{1}{8}$	6	$1.630 \cdot 10^{-4}$	3.3	$1.062 \cdot 10^{-2}$	2.0	$6.852 \cdot 10^{-3}$	2.1
	$\frac{1}{16}$	6	$1.969 \cdot 10^{-5}$	3.1	$2.661 \cdot 10^{-3}$	2.0	$1.738 \cdot 10^{-3}$	2.0
2.0	$\frac{1}{2}$	9	$1.244 \cdot 10^{-2}$		$1.971 \cdot 10^{-1}$		$1.670 \cdot 10^{-1}$	
	$\frac{1}{4}$	8	$1.831 \cdot 10^{-3}$	2.8	$5.053 \cdot 10^{-2}$	2.0	$3.691 \cdot 10^{-2}$	2.2
	$\frac{1}{8}$	7	$1.817 \cdot 10^{-4}$	3.3	$1.115 \cdot 10^{-2}$	2.2	$8.353 \cdot 10^{-3}$	2.1
	$\frac{1}{16}$	7	$2.300 \cdot 10^{-5}$	3.0	$2.753 \cdot 10^{-3}$	2.0	$2.032 \cdot 10^{-3}$	2.0

TABLE 5.2
Results using a two-level method.

H	h	CPU time	L^2 error of $\mathbf{u}$	L^2 rate of $\mathbf{u}$	H^1 error of $\mathbf{u}$	H^1 rate of $\mathbf{u}$	L^2 error of $\boldsymbol{\sigma}$	L^2 rate of $\boldsymbol{\sigma}$
$\frac{1}{2}$	$\frac{1}{4}$	8.2	$1.643 \cdot 10^{-3}$		$4.614 \cdot 10^{-2}$		$3.116 \cdot 10^{-2}$	
$\frac{1}{2}$	$\frac{1}{8}$	81.4	$3.946 \cdot 10^{-4}$	2.06	$1.455 \cdot 10^{-2}$	1.67	$2.183 \cdot 10^{-2}$	0.51
$\frac{1}{3}$	$\frac{1}{4}$	20.3	$1.574 \cdot 10^{-3}$		$4.489 \cdot 10^{-2}$		$2.925 \cdot 10^{-2}$	
$\frac{1}{3}$	$\frac{1}{8}$	93.5	$1.703 \cdot 10^{-4}$	3.21	$1.077 \cdot 10^{-2}$	2.06	$7.560 \cdot 10^{-3}$	1.95
$\frac{1}{3}$	$\frac{1}{12}$	406.4	$6.991 \cdot 10^{-5}$	2.20	$5.037 \cdot 10^{-3}$	1.87	$4.785 \cdot 10^{-3}$	1.13
$\frac{1}{4}$	$\frac{1}{8}$	117.4	$1.639 \cdot 10^{-4}$		$1.064 \cdot 10^{-2}$		$6.896 \cdot 10^{-3}$	
$\frac{1}{4}$	$\frac{1}{12}$	429.7	$4.871 \cdot 10^{-5}$	2.99	$4.738 \cdot 10^{-3}$	2.00	$3.220 \cdot 10^{-3}$	1.88
$\frac{1}{4}$	$\frac{1}{16}$	1290.4	$2.192 \cdot 10^{-5}$	2.78	$2.697 \cdot 10^{-3}$	1.96	$1.998 \cdot 10^{-3}$	1.66
$\frac{1}{5}$	$\frac{1}{8}$	168.4	$1.632 \cdot 10^{-4}$		$1.062 \cdot 10^{-2}$		$6.850 \cdot 10^{-3}$	
$\frac{1}{5}$	$\frac{1}{12}$	482.3	$4.696 \cdot 10^{-5}$	3.07	$4.709 \cdot 10^{-3}$	2.01	$3.071 \cdot 10^{-3}$	2.00
$\frac{1}{5}$	$\frac{1}{16}$	1339.8	$1.992 \cdot 10^{-5}$	2.98	$2.666 \cdot 10^{-3}$	1.98	$1.771 \cdot 10^{-3}$	1.91
$\frac{1}{6}$	$\frac{1}{12}$	546.1	$4.688 \cdot 10^{-5}$		$4.709 \cdot 10^{-3}$		$3.064 \cdot 10^{-3}$	
$\frac{1}{6}$	$\frac{1}{16}$	1405.8	$1.979 \cdot 10^{-5}$	3.00	$2.662 \cdot 10^{-3}$	1.98	$1.771 \cdot 10^{-3}$	1.91
$\frac{1}{8}$	$\frac{1}{16}$	1723.1	$1.973 \cdot 10^{-5}$		$2.662 \cdot 10^{-3}$		$1.740 \cdot 10^{-3}$	

In Table 5.1 we present the results obtained using the one-grid method. It is observed that the number of Newton iterations increases a little for a larger value of λ. The experimental rates of $\mathbf{u}$ and $\boldsymbol{\sigma}$ are very close to theoretical rates when P_2 and P_1 elements are used for an elliptic problem, respectively. In [10], a decoupled algorithm for the same model equation is presented, and the theoretical H^1 and L^2 rates of $\mathbf{u}$ and $\boldsymbol{\sigma}$, respectively, are proved to be 1.5. In fact, the experimental rates

TABLE 5.3
Results using three- and four-level methods.

H	h_1	h_2	h_3	CPU time	L^2 error of $\mathbf{u}$	H^1 error of $\mathbf{u}$	L^2 error of $\boldsymbol{\sigma}$
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$		87.2	$1.637 \cdot 10^{-4}$		$4.659 \cdot 10^{-2}$
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{16}$		1278.2	$2.196 \cdot 10^{-5}$	$2.672 \cdot 10^{-3}$	$1.774 \cdot 10^{-3}$
$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{16}$		1288.7	$1.978 \cdot 10^{-5}$	$2.662 \cdot 10^{-3}$	$1.744 \cdot 10^{-3}$
$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$		1381.2	$1.973 \cdot 10^{-5}$	$2.662 \cdot 10^{-3}$	$1.740 \cdot 10^{-3}$
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	1337.7	$1.973 \cdot 10^{-5}$	$2.662 \cdot 10^{-3}$	$1.740 \cdot 10^{-3}$

TABLE 5.4
Comparison of multilevel methods to a one-level method.

Grids	CPU time	L^2 error of $\mathbf{u}$	H^1 error of $\mathbf{u}$	L^2 error of $\boldsymbol{\sigma}$
$\frac{1}{4}$	38.4	$1.589 \cdot 10^{-3}$	$4.520 \cdot 10^{-2}$	$2.914 \cdot 10^{-2}$
$\frac{1}{4}$, $\frac{1}{3}$	20.34	$1.574 \cdot 10^{-3}$	$4.448 \cdot 10^{-2}$	$2.925 \cdot 10^{-2}$
$\frac{1}{4}$, $\frac{1}{2}$	8.2	$1.643 \cdot 10^{-3}$	$4.614 \cdot 10^{-2}$	$3.166 \cdot 10^{-2}$
$\frac{1}{8}$	469.7	$1.630 \cdot 10^{-4}$	$1.062 \cdot 10^{-2}$	$6.852 \cdot 10^{-3}$
$\frac{1}{8}$, $\frac{1}{5}$	168.4	$1.632 \cdot 10^{-4}$	$1.062 \cdot 10^{-2}$	$6.850 \cdot 10^{-3}$
$\frac{1}{8}$, $\frac{1}{4}$	117.4	$1.639 \cdot 10^{-4}$	$1.064 \cdot 10^{-2}$	$6.896 \cdot 10^{-3}$
$\frac{1}{8}$, $\frac{1}{3}$	93.5	$1.703 \cdot 10^{-4}$	$1.077 \cdot 10^{-2}$	$7.560 \cdot 10^{-3}$
$\frac{1}{8}$, $\frac{1}{2}$	81.4	$3.946 \cdot 10^{-4}$	$1.455 \cdot 10^{-2}$	$2.183 \cdot 10^{-2}$
$\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$	87.2	$1.637 \cdot 10^{-4}$	$1.064 \cdot 10^{-2}$	$6.882 \cdot 10^{-3}$
$\frac{1}{12}$	2342.2	$4.672 \cdot 10^{-5}$	$4.706 \cdot 10^{-3}$	$3.056 \cdot 10^{-3}$
$\frac{1}{12}$, $\frac{1}{6}$	546.1	$4.688 \cdot 10^{-5}$	$4.709 \cdot 10^{-3}$	$3.064 \cdot 10^{-3}$
$\frac{1}{12}$, $\frac{1}{5}$	482.3	$4.696 \cdot 10^{-5}$	$4.709 \cdot 10^{-3}$	$3.071 \cdot 10^{-3}$
$\frac{1}{12}$, $\frac{1}{4}$	429.7	$4.871 \cdot 10^{-5}$	$4.738 \cdot 10^{-3}$	$3.220 \cdot 10^{-3}$
$\frac{1}{12}$, $\frac{1}{3}$	406.4	$6.991 \cdot 10^{-5}$	$5.037 \cdot 10^{-3}$	$4.785 \cdot 10^{-3}$
$\frac{1}{16}$	7480.9	$1.969 \cdot 10^{-5}$	$2.661 \cdot 10^{-3}$	$1.738 \cdot 10^{-3}$
$\frac{1}{16}$, $\frac{1}{8}$	1723.1	$1.973 \cdot 10^{-5}$	$2.662 \cdot 10^{-3}$	$1.740 \cdot 10^{-3}$
$\frac{1}{16}$, $\frac{1}{6}$	1405.8	$1.979 \cdot 10^{-5}$	$2.662 \cdot 10^{-3}$	$1.771 \cdot 10^{-3}$
$\frac{1}{16}$, $\frac{1}{5}$	1339.8	$1.992 \cdot 10^{-5}$	$2.666 \cdot 10^{-3}$	$1.771 \cdot 10^{-3}$
$\frac{1}{16}$, $\frac{1}{4}$	1290.4	$2.192 \cdot 10^{-5}$	$2.697 \cdot 10^{-3}$	$1.998 \cdot 10^{-3}$
$\frac{1}{16}$, $\frac{1}{6}$, $\frac{1}{2}$	1278.2	$2.196 \cdot 10^{-5}$	$2.672 \cdot 10^{-3}$	$1.774 \cdot 10^{-3}$
$\frac{1}{16}$, $\frac{1}{6}$, $\frac{1}{3}$	1288.7	$1.978 \cdot 10^{-5}$	$2.662 \cdot 10^{-3}$	$1.744 \cdot 10^{-3}$
$\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$	1381.2	$1.973 \cdot 10^{-5}$	$2.662 \cdot 10^{-3}$	$1.740 \cdot 10^{-3}$
$\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$	1337.7	$1.973 \cdot 10^{-5}$	$2.662 \cdot 10^{-3}$	$1.740 \cdot 10^{-3}$

of convergence obtained by the decoupled algorithm are larger than 1.5 but less than the rates presented in Table 5.1. See [8].

In Table 5.2 we present results for the same example using a two-grid method

with $\lambda = 1$, and in Table 5.3 we present results using three- and four-grid methods for the same value of λ. In presenting these computations, we fix the coarse grid H and then vary the fine grid. The full nonlinear problem is solved on the coarse mesh, and then one linearized Newton step is performed on the fine meshes. The rates of convergence in Table 5.2 were obtained by comparing the errors on two consecutive fine grids which used the same initial coarse grid. It is experimentally observed that if a fine grid is fixed, the coarse grid H should be as small as $H \approx h^{2/3}$. For example, when $h = 1/8$, $h^{2/3} = 1/4$, so the two-level method with $H = 1/4$ works as accurately as the one-grid method. However, a choice of $H = 1/3$ yields larger errors, as seen in Table 5.2. Similarly, if $h = 1/12$ or $1/16$, the optimal scaling is $(1/12)^{2/3} \approx 1/5$ or $(1/16)^{2/3} \approx 1/6$.

To compare the accuracy and relative efficiency of the standard one-grid method with the multigrid method, we compare the results generated with a fixed value of h for the one-grid method with the results obtained using a multigrid method where the finest grid has the same mesh spacing h. Table 5.4 summarizes these results by combining the results from Tables 5.1–5.3. It can be easily seen that calculation using the multigrid method gave considerable savings in computational time for similar accuracy. For example, when $h = \frac{1}{16}$, six Newton iterations are needed by the one-grid method, requiring CPU time of 7480.9. The two-grid method with $H = \frac{1}{8}$ and $h = \frac{1}{16}$ required six iterations on H to solve the nonlinear problem and one iteration on h, requiring CPU time of 1723.1. Using the four-grid method with $H = \frac{1}{2}$, $h_1 = \frac{1}{4}$, $h_2 = \frac{1}{8}$, and $h_3 = \frac{1}{16}$, the computational time is reduced to 1337.7, which is 18% of CPU time required for the one-grid method. Note that the finer grid calculations gave more savings in time, and most of the savings are already obtained by using the two-grid method. More savings in time can be obtained by using three- or four-grid methods.

Acknowledgment. The author thanks V. Ervin for his helpful discussions and suggestions.

REFERENCES

[1] F. Baaijens, *An iterative solver for the DEVSS/DG method with application to smooth and non-smooth flows of the upper convected Maxwell fluid*, J. Non-Newtonian Fluid Mech., 75 (1998), pp. 119–138.

[2] J. Baranger and D. Sandri, *Finite element approximation of viscoelastic fluid flow: Existence of approximate solutions and error bounds*, Numer. Math., 63 (1992), pp. 13–27.

[3] S. Brenner and L. Scott, *The Mathematical Theory of Finite Element Methods*, Springer–Verlag, New York, 1994.

[4] M. Fortin and A. Fortin, *A new finite approach for the F.E.M. simulation of viscoelastic flows*, J. Non-Newtonian Fluid Mech., 32 (1989), pp. 295–310.

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

[6] W. Layton, *A two-level discretization method for the Navier-Stokes equations*, Comput. Math. Appl., 26 (1993), pp. 33–38.

[7] W. Layton, H. K. Lee, and J. Peterson, *Numerical solution of the stationary Navier–Stokes equations using a multilevel finite element method*, SIAM J. Sci. Comput., 20 (1998), pp. 1–12.

[8] H. Lee and A. Liakos, *Two-level finite element discretization of viscoelastic fluid flow*, Comput. Methods Appl. Mech. Engrg., to appear.

[9] J. M. Marchal and M. J. Crochet, *A new finite element for calculating viscoelastic flow*, J. Non-Newtonian Fluid Mech., 26 (1987), pp. 77–114.

[10] K. Najib and D. Sandri, *On a decoupled algorithm for solving a finite element problem for the approximation of viscoelastic fluid flow*, Numer. Math., 72 (1995), pp. 223–238.

[11] M. Renardy, *Existence of slow steady flows of viscoelastic fluids with differential constitutive equations*, ZAMM Z. Angew. Math. Mech., 65 (1985), pp. 449–451.
[12] D. Sandri, *Finite element approximation of viscoelastic fluid flow: Existence of approximate solutions and error bounds. Continuous approximation of the stress*, SIAM J. Numer. Anal., 31 (1994), pp. 362–377.
[13] C. Schwab, *p- and hp-Finite Element Methods. Theory and Applications to Solid and Fluid Mechanics*, Oxford University Press, New York, 1998.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 130–152

ASYNCHRONOUS FAST ADAPTIVE COMPOSITE-GRID METHODS FOR ELLIPTIC PROBLEMS: THEORETICAL FOUNDATIONS*

BARRY LEE†¶, STEPHEN F. MCCORMICK‡, BOBBY PHILIP§¶, AND DANIEL J. QUINLAN†¶

Abstract. Accurate numerical modeling of complex physical, chemical, and biological systems requires numerical simulation capability over a large range of length scales, with the ability to capture rapidly varying phenomena localized in space and/or time. Adaptive mesh refinement (AMR) is a numerical process for dynamically introducing local fine resolution on computational grids during the solution process, in response to unresolved error in a computation. Fast adaptive composite-grid (FAC) methods are a class of algorithms that exploit the multilevel structure of AMR grids to solve elliptic problems efficiently. This paper develops a theoretical foundation for AFACx, an asynchronous FAC method. A new multilevel condition number estimate establishes that the convergence rate of the AFACx algorithm does not degrade as the number of refinement levels in the AMR hierarchy increases.

Key words. adaptive mesh refinement, asynchronous, fast adaptive composite-grid, elliptic solvers, FAC, AFAC, AFACx

AMS subject classifications. 65F10, 65N22, 65N50, 65N55

DOI. 10.1137/S0036142902400767

1. Introduction.

Adaptive mesh refinement (AMR) is a numerical process for dynamically introducing local fine resolution on computational grids during the solution process in response to unresolved error in a computation. Local fine resolution is achieved by dynamically adapting the existing computational grid based on additional grid points (point-based AMR) or finer local grids (block-structured AMR). AMR approaches are attractive because they often achieve orders of improvement in computational efficiency and memory usage. AMR techniques were first introduced by Brandt [19] in the early 1970s for general problems in a multilevel context and by Berger and Oliger [6] in the 1980s for hyperbolic problems. Since then, AMR research has been pursued by several groups (cf. [1, 2, 5, 29, 42, 43]).

For elliptic problems, when numerical simulations involve a large number of refinement levels and are extremely large, effective parallel methods for AMR must be considered. It is then desirable to develop elliptic solvers that asynchronously process all grids, or at least asynchronously process grids at a fixed refinement level. In addition, as the number of refinement levels increases, the convergence rate should not degrade as a function of the number of refinement levels.

*Received by the editors January 9, 2002; accepted for publication (in revised form) June 16, 2003; published electronically January 6, 2004. This work was performed by an employee of the U.S. Government. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.
http://www.siam.org/journals/sinum/42-1/40076.html

†Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550 (lee123@llnl.gov, dquinlan@llnl.gov).

‡Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO 80309 (stevem@newton.colorado.edu).

§Current Address: CCS-3, Modeling, Algorithms, and Informatics Group, Los Alamos National Laboratory, Los Alamos, NM 87545 (bphilip@lanl.gov).

¶The work of these authors was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

The fast adaptive composite-grid (FAC) method was developed in the 1980s [32, 33, 34, 35] to provide more robust discretization and solution methods for elliptic problems on AMR grids. Its strength lies in its ability to use existing single grid solvers on uniform meshes for different refinement levels, with the combined effect of solving a nonuniform composite-grid problem. Though FAC allows for asynchronous processing of disjoint grids at a given refinement level and its convergence rate is bounded independently of the number of refinement levels, the multiplicative way it treats the various refinement levels imposes sequentialness in its processing. For large-scale parallel AMR applications, this sequential nature of FAC, like that of other AMR techniques, represents a serious bottleneck to full scalability.

This difficulty led to the development of the asynchronous version of FAC, called AFAC [28, 33, 36, 37]. AFAC, like FAC, is blessed with level-independent convergence bounds and the convenience of enabling uniform grid solvers. But it has the added advantage of allowing asynchronous processing of all refinement levels. This important asynchronous feature is obtained at the cost of only a modest fixed decrease in convergence rates [33].

Further research into improving computational efficiency associated with the uniform grid solvers on local grid patches led to the development of AFACx [41]. AFACx is very inexpensive because it uses only simple relaxation methods on all but the coarsest grid. Numerical results [39, 41] show that the attendant reduction in computational and communication costs of AFACx comes with no significant degradation in convergence rates compared to AFAC based on multigrid solvers.

Convergence bounds for FAC were established in [32, 38] under certain regularity assumptions. Widlund and Dryja proposed and analyzed variants of FAC [45, 23]. Reusken and Ferket [24] compared FAC with the local defect correction (LDC) method [27] introduced by Hackbusch. AFAC was introduced by Hart and McCormick in [28]. Optimality in the multilevel case for AFAC applied to a model problem was shown in [31]. This was followed by the development of AFACx [41]. Cheng [21, 22] established optimal bounds on the condition number of the multilevel AFAC iteration operator with exact solvers. Moe [39, 7] presented performance results for FAC and AFAC on parallel machines. Quinlan, in his thesis [41], presented a two-level convergence analysis for AFACx assuming a sufficient number of smoothing steps at each level and showed that it is closely related to the convergence rate of AFAC. Shapira [44] compared the performance of AFAC and AFACx. However, a multilevel theory for AFACx remained a gap in the theory of multilevel FAC-type methods.

Closely related to AFACx are the additive preconditioners of Bramble, Pasciak, Xu [16] and Bramble, Pasciak, and Vassilevski [18]. The theoretical framework developed by Bramble, Pasciak, Xu, Wang, Oswald, Griebel, and others [16, 15, 12, 14, 13, 17, 11, 47, 48, 50, 9, 8, 49, 26] presents a powerful tool for analyzing multilevel methods. Relying heavily on this modern multilevel framework for multilevel methods and some of the assumptions therein, we present in this paper a new multilevel condition number estimate for the AFACx operator.

The new theoretical results presented in this paper are strongly backed by numerical evidence [41, 40] and performance results [41]. Recent numerical work [40, 30] shows that AFACx can be applied successfully to elliptic PDE systems arising from first-order system least squares (FOSLS) formulations on adaptively refined curvilinear AMR grids. As increasingly complex PDE systems are simulated and the need for AMR is increasingly crucial, theoretical and computational analyses of fast, parallelizable, and efficient multilevel solvers and preconditioners such as AFACx and BPX [16] become increasingly important for validating the results of complex simulations.

This paper develops new multilevel estimates establishing that the condition number of the AFACx operator, like that for AFAC, is bounded independently of the number of refinement levels. We start by introducing the model problem and necessary preliminaries in section 2. In section 3, we introduce the various FAC-type methods, and finally, in section 4, we establish the theory.

2. Model problem. Consider a linear, self-adjoint, second-order elliptic boundary value problem in $\mathbb{R}^n$, $n = 2, 3$, of the form

$$\left\{ \begin{array}{rl} Lu \equiv -\sum\limits_{i,j=1}^{n} \frac{\partial}{\partial x_i}(a_{ij}(x)\frac{\partial u}{\partial x_j}) = & f \quad \text{in } \Omega, \\ u = & 0 \quad \text{on } \partial\Omega, \end{array} \right. \tag{2.1}$$

where u is the unknown, $f \in L^2(\Omega)$ is the source term, and a_{ij} are appropriate coefficients. Assume that

- domain $\Omega \subset \mathbb{R}^n$ is convex polygonal;
- coefficients $a_{ij}(x) \in C^0(\overline{\Omega}), 1 \le i, j \le n$;
- matrix $[a_{ij}(x)]_{1 \le i,j \le n}$ is symmetric almost everywhere in Ω; and
- operator L is uniformly elliptic in the sense that there exists a constant $\theta > 0$ such that $\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \ge \theta|\xi|^2$ for almost all x in Ω and all ξ in $\mathbb{R}^n$, where $|\cdot|$ is the Euclidean norm.

This section is concerned with the numerical solution of the algebraic equations that arise from discretizing problem (2.1) on adaptively refined curvilinear grids. We focus on the plane $\mathbb{R}^2$ for simplicity.

2.1. Variational formulation. Under the above assumptions, the natural linear space in which to seek a weak solution of (2.1) is $V := H_0^1(\Omega)$, and the variational problem is Find $u \in V$ such that

$$a(u, v) = f(v) \quad \forall v \in V, \tag{2.2}$$

where the respective bilinear and linear forms are

$$a(u, v) = \int_\Omega \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} \mathrm{d}\Omega, \tag{2.3}$$

$$f(v) = \int_\Omega f v \mathrm{d}\Omega. \tag{2.4}$$

It is known [25] that (2.2) has a unique solution, $u \in V$. Moreover, $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ is symmetric and continuous [25], so uniform ellipticity of L and Poincaré's inequality (cf. [10]) imply that $a(\cdot, \cdot)$ is coercive on V: there exists a constant $\gamma > 0$ such that

$$a(u, u) \ge \gamma \|u\|_V^2 \quad \forall u \in V. \tag{2.5}$$

Coercivity, in turn, implies that $a(\cdot, \cdot)$ defines an equivalent inner product over space V. Furthermore, by the Riesz representation theorem (cf. [10]), $a(\cdot, \cdot)$ induces a bounded linear operator $\mathcal{A} : V \to V$ uniquely determined by

$$a(u, v) = (\mathcal{A}u, v) \quad \forall u, v \in V. \tag{2.6}$$

2.2. Partially refined meshes. To discretize (2.2) on partially refined meshes, we introduce the following notation. Let $\Omega_J \subseteq \Omega_{J-1} \subseteq \cdots \subseteq \Omega_1 \equiv \Omega$ be a nested sequence of nonempty bounded open polygonal Lipschitz domains. Subdomains $\Omega_k,\ k = 2, 3, \ldots, J$, can be viewed as regions where the solution may vary on increasingly finer scales and, hence, regions where local refinement patches are generated during the AMR process. Let $\mathcal{T}_1^c = \{\tau_i^1\}_{i=1}^{N_1}$ be a triangulation of $\Omega_1,\ N_1 \geq 4$, meaning that they cover Ω_1 and do not overlap in the sense that the intersection of any two triangles in the triangulation is either empty, a common vertex, or a common edge. Assume that $\mathcal{T}_1^c$ is quasi-uniform. We assume also that the boundaries of Ω_2 align with the edges of elements in $\mathcal{T}_1^c$, and at least one edge of $\mathcal{T}_1^c$ is contained in Ω_2. Triangulation $\mathcal{T}_k^c = \{\tau_i^k\}_{i=1}^{N_k},\ k = 2, 3, \ldots, J$, of Ω, is obtained from $\mathcal{T}_{k-1}^c$ in the following manner. Since $\Omega_k \subseteq \Omega_{k-1}$ and its boundary aligns with elements of $\mathcal{T}_{k-1}^c$, then there exists a local "coarse" triangulation, $\mathcal{T}_k^{h_{k-1}} = \{\tau_{i_j}^{k-1}\}_{j=1}^{M_k}, M_k \leq N_{k-1}$, consisting of elements of $\mathcal{T}_{k-1}^c$ that cover Ω_k, where h_{k-1} is the length of the longest edge of triangles in $\mathcal{T}_k^{h_{k-1}}$. $\mathcal{T}_k^{h_{k-1}}$ is then a quasi-uniform triangulation of Ω_k. Now we uniformly refine elements of $\mathcal{T}_k^{h_{k-1}}$ by subdividing each triangle into four triangles by connecting the midpoints of the edges. This yields a "fine" local triangulation $\mathcal{T}_k^{h_k}$ of Ω_k, which is regular in the sense of Bank, Dupont, and Yserentant [3]. Elements of $\mathcal{T}_{k-1}^c$ that lie in the complement of Ω_k and the elements of $\mathcal{T}_k^{h_k}$ together form the elements of $\mathcal{T}_k^c = (\mathcal{T}_{k-1}^c \backslash \mathcal{T}_k^{h_{k-1}}) \cup \mathcal{T}_k^{h_k}$. This process leads to a series of nested triangulations $\{\mathcal{T}_k^c\}_{k=1}^J$ of Ω that form partially refined locally quasi-uniform meshes.

2.3. Finite element spaces. Henceforth, we assume that conforming piecewise linear finite elements are used, although our results will clearly apply to more general cases. We thus define $V_k^c \subset H_0^1(\Omega), k = 1, 2, \ldots J$, to be the space spanned by standard piecewise linear nodal basis functions with local support about the nodes of triangulation $\mathcal{T}_k^c$. Because of the conformity of the finite elements, note that there are no degrees of freedom associated with fine nodes that lie on boundary $\partial\Omega_k$. Continuity implies that these "slave" nodes are evaluated simply by interpolation from adjacent coarse nodes. Now, the "fine" local finite element space defined in the interior of domain Ω_k is $V_k^{h_k} = V_k^c \cap H_0^1(\Omega_k)$. By our use of $H_0^1(\Omega_k)$ here, we mean that functions in $V_k^{h_k}$ have support only in the interior of Ω_k. Similarly, we define "coarse" local finite element spaces by $V_k^{h_{k-1}} = V_{k-1}^c \cap H_0^1(\Omega_k), V_k^{h_{k-1}} \subset V_k^{h_k},\ k = 2, \ldots, J$. Note that the local spaces are nested: $V_1^c \subseteq V_2^c \subseteq \cdots \subseteq V_J^c \subset H_0^1(\Omega)$. However, the coarse local spaces are generally nonnested because they typically correspond to increasingly smaller local subdomains.

2.4. The discrete variational problem. Having chosen finite-dimensional composite-grid space V_J^c, the discrete variational problem is Find $u^c \in V_J^c$ such that

$$a(u^c, v) = f(v) \ \ \forall v \in V_J^c. \tag{2.7}$$

This problem is equivalent to solving the linear system

$$A^c u^c = f^c, \tag{2.8}$$

where A^c is a symmetric positive-definite matrix induced by the linear operator $\mathcal{A}^c$ defined over composite-grid space V_J^c. Note that V_J^c is a finite-dimensional subspace of $H_0^1(\Omega)$. For notational convenience, denote spaces V_k^c by $V_k,\ k = 1, 2, \ldots, J$; operator $\mathcal{A}^c$ by A; A-inner product $a(\cdot, \cdot)$ on V_J by $A(\cdot, \cdot)$; and the induced A-norm by $|||\cdot|||$. The L^2 inner product on V_J is denoted by $(\cdot, \cdot)$ and its induced norm by $\|\cdot\|$.

2.5. Stationary linear iteration. A consistent stationary linear iterative process for linear system

$$Au = f$$

can be written in the form

$$u^{n+1} = u^n + B(f - Au^n), \tag{2.9}$$

where B is an approximate inverse of A. Here we are thinking of iteration (2.9) as one of our FAC-type algorithms defined below. If B is symmetric with respect to the A inner product, then the process is said to be symmetric. Letting $e^n = u - u^n$ denote the error in the nth iterate, then (2.9) implies that

$$e^{n+1} = (I - BA)e^n. \tag{2.10}$$

Therefore, for the iteration to converge in general, we must have $\rho(I - BA) < 1$, where $\rho(\cdot)$ denotes the spectral radius. For common multiplicative-type algorithms, it is often easy to establish this condition. However, for additive-type multilevel solvers, typically all that can be shown is that $\kappa(BA)$, the condition number of the operator BA, is independent of the number of levels. Such a result implies that the "damped" linear iteration

$$u^{n+1} = u^n + \omega B(f - Au^n) \tag{2.11}$$

converges for sufficiently small ω. It is this type of a result that we establish for the AFACx algorithm defined below.

To describe the FAC algorithms, we need to define operators that approximate (2.7) at the different refinement levels, projection operators that transfer data from fine to coarse spaces, and smoothing operators on the different spaces, all in terms of the discrete inner products $(\cdot, \cdot)$ and $A(\cdot, \cdot)$.

2.6. Approximating composite-grid operators on coarser levels.

DEFINITION 1. *For $k = 1, 2, \ldots, J$, define operator $A_k : V_k \longrightarrow V_k$ by*

$$(A_k w, \phi) = A(w, \phi) \quad \forall \phi \in V_k.$$

Note that operator A_k is symmetric and positive-definite in inner products $A(\cdot, \cdot)$ and $(\cdot, \cdot)$.

2.7. Projection operators. We introduce the following projection operators typically used in multilevel theory.

DEFINITION 2. *For $k = 1, 2, \ldots, J$, define "elliptic projection" operator $P_k : V_J \longrightarrow V_k$ by*

$$A(P_k w, \phi) = A(w, \phi) \quad \forall \phi \in V_k.$$

DEFINITION 3. *For $k = 1, 2, \ldots, J$, define "L^2 projection" operator $Q_k : V_J \longrightarrow V_k$ by*

$$(Q_k w, \phi) = (w, \phi) \quad \forall \phi \in V_k.$$

It can be shown that P_k and Q_k are orthogonal projection operators satisfying the following basic properties:

- $P_k P_l = P_l$, $P_l P_k = P_l$, $Q_k Q_l = Q_l$, $Q_l Q_k = Q_l$ for $l \leq k$.

Additionally, P_k and Q_k are related according to

$$A_k P_k = Q_k A. \tag{2.12}$$

2.8. Smoothing operators. We can write one step of a general stationary linear smoothing procedure applied to

$$A_k u_k = f_k \tag{2.13}$$

in the form

$$u_k^{n+1} = u_k^n + R_k(f_k - A_k u_k^n), \tag{2.14}$$

where $R_k : V_k \longrightarrow V_k$. Note that (2.14) is of the same form as (2.9), but we use R here and below (possibly with subscripts and hats) to signify smoothing. Now the error, $e_k^n = u_k - u_k^n$, obeys the following propagation equation:

$$e_k^{n+1} = (I - R_k A_k)e_k^n \tag{2.15}$$

or

$$e_k^{n+1} = (I - T_k)e_k^n, \tag{2.16}$$

where $T_k : V_J \longrightarrow V_k$ is defined by $T_k = R_k A_k P_k$. For simplicity, we assume that $R_1 = A_1^{-1}$ and that $R_k,\ k = 2, 3, \ldots, J$, are symmetric with respect to the L^2 inner product. Consider the special case $R_k = \hat{R}_k \equiv \frac{1}{\lambda_k} I$, where λ_k is the spectral radius of A_k. The smoothing process is then just Richardson's iteration defined by

$$u_k^{n+1} = u_k^n + \frac{I}{\lambda_k}(f_k - A_k u_k^n). \tag{2.17}$$

Corresponding to $\hat{R}_k$, we define $\hat{T}_k = \hat{R}_k A_k P_k$.

To further quantify the properties that a simple smoother must satisfy, we make the following assumptions commonly made in modern multilevel analyses. While we do briefly comment on the motivation for each assumption and the conditions under which they hold, we refer the reader to [11, 46, 15, 16, 12, 13, 48, 17, 47, 50] for further details. It suffices to state that the assumptions are valid (cf. [11]) for our model problem with partially refined locally quasi-uniform meshes and simple smoothers like Richardson, damped Jacobi, and symmetric Gauss–Seidel.

The first assumption concerns the Richardson operator, $\hat{R}_k$.

A.1. There exist constants $\epsilon \in (0, 1)$ and $\gamma > 0$ such that

$$A(\hat{T}_k w, w) \le (\gamma \epsilon^{k-l})^2 A(w, w) \quad \forall w \in V_l,\ l \le k, \quad k = 1, 2, \ldots, J. \tag{2.18}$$

Roughly speaking, assumption A.1 asserts that the smoother attenuates "smooth" error components slowly; i.e., energy reduction in "smooth" components (represented by components in subspaces $V_l, l < k$) is small compared to energy reduction in the "oscillatory" components. This assumption is a generalization of the strengthened Cauchy–Schwarz inequalities first introduced by Yserentant [49] for hierarchical bases and used extensively in multilevel theory [47, 14, 11, 48]. Constant γ depends on the ellipticity of the boundary value problem and the variation of the coefficients in (2.1). For our model boundary value problem (2.1) discretized with piecewise linears on simplices, (2.18) has been shown to hold (cf. [14, 48]). However, it is apparently not known whether assumption A.1 holds when the coefficients in (2.1) are not very smooth, e.g., when they are only bounded and measurable (cf. [11]).

The next two assumptions allow for more general smoothers, R_k.

A.2. There exist constants $a_0 \in (0,1)$ and $a_1 > 1$ such that

$$a_0 \frac{\|u\|^2}{\lambda_k} \leq (R_k u, u) \leq a_1 \frac{\|u\|^2}{\lambda_k} \quad \forall u \in V_k, 2 \leq k \leq J. \tag{2.19}$$

Assumption A.2 can also be written in the form

$$a_0(\hat{R}_k u, u) \leq (R_k u, u) \leq a_1(\hat{R}_k u, u) \quad \forall u \in V_k, 2 \leq k \leq J, \tag{2.20}$$

which implies that smoothing operator R_k, $k = 2, 3, \ldots, J$, is spectrally equivalent to the Richardson smoothing operator $\hat{R}_k$. It is easy to see that A.2 implies that

$$a_0\ A(\hat{T}_k u, u) \leq A(T_k u, u) \leq a_1\ A(\hat{T}_k u, u) \quad \forall u \in V_J,\ k = 2, 3, \ldots, J. \tag{2.21}$$

Spectral equivalence of symmetric Gauss–Seidel to the Richardson smoother is shown in [48]. The upper inequality in (2.21) holds in general for point-smoothers (cf. [11]).

A.3. There exists a constant $\theta \in (0, 2)$ such that

$$A(T_k v, T_k v) \leq \theta A(T_k v, v) \quad \forall v \in V_k,\ k = 1, 2, \ldots, J. \tag{2.22}$$

Assumption A.3 is a natural consequence of assuming that operators $I - T_k$ are contractive in the energy norm, i.e.,

$$|||I - T_k||| < 1, \quad k = 1, 2, \ldots, J. \tag{2.23}$$

Note that $\theta = 1$ for Richardson, $\theta < 1$ for under-damped Richardson, and $\theta = 1$ for suitably scaled Jacobi and block Jacobi smoothers. In [13], (2.22) is shown to hold for various line and point-based Jacobi and Gauss–Seidel smoothers.

Assumption A.3 can be derived from assumption A.1 under a suitable assumption on γ and spectral equivalence of the smoothers to Richardson iteration. However, in general, the assumption on γ cannot be established without special scaling of the smoothers, so we choose to state both assumptions separately.

In addition to assumptions A.1–A.3 on smoothers R_k and $\hat{R}_k$, a "weak regularity" assumption is required. This condition replaces the standard full regularity and approximation assumption (cf. [12]) with a weaker assumption on operator A and smoothers $\hat{R}_k, k = 2, 3, \ldots, J$.

A.4. There exists a constant $\eta > 0$ such that

$$A(v, v) \leq \eta \sum_{k=1}^{J} A(\hat{T}_k v, v) \quad \forall v \in V_J. \tag{2.24}$$

Assumption A.4 is shown to hold for our model problem discretization in [11]. However, it is also noted in [11] that, in the application to second-order elliptic equations for coefficients with large jumps, A.4 is not known to hold independent of the size of the jumps.

3. Algorithms. We now describe the FAC, AFAC, and AFACx algorithms and complete the section with a discussion of existing theory.

3.1. FAC. Let $u_c^n \in V_J$ denote the current approximation to the solution of composite-grid equation (2.8).

ALGORITHM 1. *One iteration of the basic FAC algorithm consists of the following steps.*

For $k = 1, 2, \ldots, J$, *do:*
  find $w_k \in V_k^{h_k}$ *such that*
    $a(u_c^{n+(k-1)/J} + w_k, v) = f(v) \quad \forall v \in V_k^{h_k}$;
  set $u_c^{n+k/J} = u_c^{n+(k-1)/J} + w_k$.

As can be seen from this pseudolanguage, FAC involves the solution of the residual equation on all refinement levels. The correction on a coarse level is computed before the correction on the next finer level, thus providing boundary conditions for the finer level equations. FAC is multiplicative, since it can be represented as a product of linear operators. Multiplicative algorithms are inherently sequential because each operation depends on its predecessor, making them less attractive in a parallel environment.

3.2. AFAC. Processing on each level in FAC attempts to resolve all components of the solution to the composite-grid residual equation that are represented on a refinement level and coarser levels. On the other hand, processing of each level by AFAC [33, 36, 37] attempts only to resolve components that can be represented on that refinement level. This objective is not dependent on resolving components of the solution to the residual equation that are represented on coarser or finer levels, so it provides for independent level processing. The principal step in AFAC is resolving solution components on each composite-grid level. Let $u_c^n \in V$ denote the current approximation to the solution of composite-grid equation (2.8).

ALGORITHM 2. *One iteration of the AFAC algorithm consists of the following steps.*

For $k = 1, 2, \ldots, J$, *do:*
  find $w_k^f \in V_k^{h_k}$ *such that*
    $a(u_c^n + w_k^f, v) = f(v) \quad \forall v \in V_k^{h_k}$;
  if $(k > 1)$, *then*
    find $w_k^r \in V_k^{h_{k-1}}$ *such that*
      $a(u_c^n + w_k^r, z) = f(z) \quad \forall z \in V_k^{h_{k-1}}$;
set $w_1^r = 0$;
set $u_c^{n+1} = u_c^n + \sum_{k=1}^{J} (w_k^f - w_k^r)$.

AFAC appears to have optimal or near-optimal complexity in a parallel computing environment because it allows for simultaneous processing of all refinement levels. This is important because the solution process on each grid, even with the most efficient solvers, dominates computational complexity. This is especially true for systems where the solution process is significantly more computationally intensive than the evaluation of the residuals. Coupled with multigrid processing on each level and nested iteration [32] on the composite-grids, the computational cost of AFAC is proportional to the cost of a global-grid solve alone (see Hart and McCormick [28] and McCormick [33] for further details).

The following two-grid result is proved in [32].

THEOREM 3.1. *Suppose* A^c *is positive-definite. Then the spectral radii of the two-level exact solver forms of* $AFAC^c$ *and* FAC^c *satisfy*

$$\rho(AFAC^c) = \rho^{\frac{1}{2}}(FAC^c). \tag{3.1}$$

Here, FAC^c *and* $AFAC^c$ *denote the respective FAC and AFAC error propagation op-*

erators on the composite-grid space, and $\rho(\cdot)$ denotes spectral radius. The convergence factor for one iteration of the two-level exact solver form of AFAC satisfies

$$|||AFAC^c|||| \leq \left(\frac{\delta}{1+\delta}\right)^{\frac{1}{4}}, \tag{3.2}$$

where constant $\delta > 0$ is independent of h but depends on the regularity of (2.1) *and the approximation properties of its discretization (see* [32] *for further details).*

Now, assume for each k, $1 \leq k \leq J$, that there exists a bounded Lipschitz polyhedral region $\hat{\Omega}_k$ such that $\Omega_k \subset \hat{\Omega}_k$, $(\hat{\Omega}_k \setminus \Omega_k) \cap \Omega = \emptyset$, and $\partial\hat{\Omega}_k \cap \Omega_{k+1} = \emptyset$, and that the Lipschitz constants of $\hat{\Omega}_k \setminus \Omega_{k+1}$ are uniformly bounded. In addition, assume there exist constants $\gamma_1 \geq \gamma_0 > 0$ and $q \in (0,1)$ such that $\gamma_0 q^k \leq h_k \leq \gamma_1 q^k$, $k = 1, 2, \ldots, J$. Under these assumptions, the following theorem was proved in [21].

THEOREM 3.2. *The AFAC operator has a condition number that is bounded independent of the number of refinement levels and the number of degrees of freedom.*

It is important to note that the results hold when the exact solvers on each level in FAC and AFAC are replaced by approximate solvers (e.g., multigrid solvers), provided that they give a fixed local error reduction (see [32] and [21]). Note also that in [32], the two-level results do not depend on the refinement ratios (h_{k+1}/h_k).

3.3. AFACx. AFAC removes the sequential nature inherent in the FAC algorithm. However, it is possible to further reduce the computational effort on each level by carefully replacing the local solvers in AFAC with smoothers. AFACx is exactly such an algorithm.

To define this scheme, we introduce auxiliary bilinear forms $b_k^r(\cdot,\cdot) : V_k^{h_{k-1}} \times V_k^{h_{k-1}} \longrightarrow \mathbb{R}$ and $b_k^f(\cdot,\cdot) : V_k^{h_k} \times V_k^{h_k} \longrightarrow \mathbb{R}$. These forms correspond to symmetric positive-definite operators $B_k^r : V_k^{h_{k-1}} \longrightarrow V_k^{h_{k-1}}$ and $B_k^f : V_k^{h_k} \longrightarrow V_k^{h_k}$ that represent the action of smoothers on the "restricted" local coarse grid space $V_k^{h_{k-1}}$ and the local "fine" grid space $V_k^{h_k}$. Let $u_c^n \in V$ denote the current approximation to the solution of composite-grid equation (2.8).

ALGORITHM 3. *One iteration of the AFACx algorithm consists of the following steps.*

For $k = 1, 2, \ldots, J$, do:
  if ($k = 1$), then
    find $u_1^f \in V_1^{h_1}$ such that
      $a(u_c^n + u_1^f, v) = f(v) \ \ \forall v \in V_1^{h_1}$;
  else
    find $w_k^r \in V_k^{h_{k-1}}$ such that
      $b_k^r(w_k^r, z) = f(z) - a(u_c^n, z) \ \ \forall z \in V_k^{h_{k-1}}$;
    find $u_k^f \in V_k^{h_k}$ such that
      $b_k^f(w_k^r + u_k^f, v) = f(v) - a(u_c^n, v) \ \ \forall v \in V_k^{h_k}$;
set $u_c^{n+1} = u_c^n + \sum_{k=1}^J u_k^f$.

The above pseudolanguage shows that AFACx replaces the solves on the local restricted coarse and fine levels in AFAC on all but the coarsest level by smoothing steps. Smoothing is performed on the restricted coarse level to obtain the correction w_k^r on each level k. Smoothing on the fine level with initial guess w_k^r then yields u_k^f, which approximates the component of the correction that is representable only on level k.

AFACx is generally more efficient than AFAC because the various uniform grids (the local fine and restricted coarse refinement levels) are processed only by smoothing, instead of the somewhat more expensive multigrid solvers used in AFAC. This reduction in cost apparently comes with no significant degradation in convergence rates. The following two-level result is due to Quinlan [41].

THEOREM 3.3. *Consider the two-level AFACx algorithm that involves one smoothing step on the fine grid patch and n smoothing steps on the restricted coarse grid. Then, for sufficiently large n, the spectral radius of the AFACx error propagation operator is bounded uniformly by a constant less than one, assuming only that this is true for the AFAC error operator.*

3.4. Symmetric AFACx. The operator corresponding to the AFACx algorithm described above is not symmetric with respect to the A inner product. To facilitate condition number estimates, we work instead with a symmetrized form of AFACx developed as follows. Let $u_c^n \in V$ denote the current approximation to the solution of composite-grid equation (2.8).

ALGORITHM 4. *One iteration of the symmetrized AFACx algorithm consists of the following steps.*

For $k = 1, 2, \ldots, J$, *do:*

if $(k = 1)$, *then*

find $u_1^f \in V_1^{h_1}$ *such that*

$a(u_c^n + u_1^f, v) = f(v) \ \ \forall v \in V_1^{h_1}$;

else

find $w_{k,0}^r \in V_k^{h_{k-1}}$ *such that*

$b_k^r(w_{k,0}^r, z) = f(z) - a(u_c^n, z) \ \ \forall z \in V_k^{h_{k-1}}$;

find $w_{k,0}^f \in V_k^{h_k}$ *such that*

$b_k^f(w_{k,0}^r + w_{k,0}^f, v) = f(v) - a(u_c^n, v) \ \ \forall v \in V_k^{h_k}$;

find $w_{k,1}^f \in V_k^{h_k}$ *such that*

$b_k^f(w_{k,1}^f, v) = f(v) - a(u_c^n, v) \ \ \forall v \in V_k^{h_k}$;

find $w_{k,1}^r \in V_k^{h_{k-1}}$ *such that*

$b_k^r(w_{k,1}^r, z) = f(z) - a(u_c^n + w_{k,1}^f, z) \ \ \forall z \in V_k^{h_{k-1}}$;

set $w_{k,1}^f = w_{k,1}^f + w_{k,1}^r$;

set $w_{k,2}^f = w_{k,1}^f - w_{k,0}^r$;

set $u_k^f = (w_{k,0}^f + w_{k,2}^f)/2$;

set $u_c^{n+1} = u_c^n + \sum_{k=1}^J u_k^f$.

The pseudolanguage above principally involves computing two approximations $w_{k,0}^f$ and $w_{k,2}^f$ and averaging them to form u_k^f at each level k. u_k^f then approximates the component of the composite-grid correction that can be represented at level k. $w_{k,0}^f$ is obtained in the following manner: smooth on the coarse-grid residual equation and interpolate to the local fine level to obtain $w_{k,0}^r$, then smooth on the fine level with initial guess $w_{k,0}^r$ to obtain $w_{k,0}^f$. To compute $w_{k,2}^f$, we first compute $w_{k,1}^f$ by applying a two-level correction scheme (as described in [20, p. 33]) on the local fine and restricted levels. $w_{k,2}^f$ is then set to be the difference $w_{k,1}^f - w_{k,0}^r$. In practice, the unsymmetric form of AFACx is used, while the symmetric form is useful for theoretical analysis.

4. Condition number estimates for AFACx. In this section, a new condition number estimate is developed for the multilevel AFACx algorithm. We show that the condition number of the symmetrized AFACx operator (Algorithm 4) is bounded independently of the number of refinement levels.

The following lemma, which is a generalization of the standard Cauchy–Schwarz inequality, is used extensively in the proofs that follow.

LEMMA 4.1 (see [47]). *Let $T \in \mathcal{L}(W)$ be a nonnegative self-adjoint operator with respect to $\langle \cdot, \cdot \rangle$, where $(W, \langle \cdot, \cdot \rangle)$ is a finite-dimensional inner product space and $\mathcal{L}(W)$ is the space of linear operators that map W into itself. Then*

$$|\langle Tu, v\rangle| \;\leq\; \langle Tu, u\rangle^{\frac{1}{2}} \langle Tv, v\rangle^{\frac{1}{2}} \quad \forall u, v \in W. \tag{4.1}$$

It is easy to show that the following lemma holds for operator $T_k = R_k A_k P_k$.

LEMMA 4.2 (see [11]). *Operator $T_k : V_J \longrightarrow V_J$, $k = 1, 2, \ldots, J$, is nonnegative and self-adjoint with respect to the A inner product on V_J.*

4.1. Full refinement. Consider first the case of full refinement: $\Omega_1 = \Omega_2 = \cdots = \Omega_J$. Note that the "restricted coarse" grid is the entire global coarse grid, so that $V_k^{h_{k-1}} = V_{k-1},\ k = 2, 3, \ldots, J$, and the "local fine" grid is the entire global fine grid, so that $V_k^{h_k} = V_k,\ k = 1, 2, \ldots, J$. Define $R_0 = 0,\ P_0 = 0$, and $Q_0 = 0$. Then the operator corresponding to one iteration of AFACx (Algorithm 3) with a single smoothing step each on the fine grid and the restricted coarse grid can be expressed as

$$B^a = \sum_{k=1}^{J} (R_k Q_k - R_k A_k R_{k-1} Q_{k-1}) A. \tag{4.2}$$

To avoid theoretical complications in satisfying assumption A.3 for $\theta \in (1, 2)$ for general smoothers, we work instead with the operator

$$B^a = \sum_{k=1}^{J} \left(R_k Q_k - \frac{1}{2} R_k A_k R_{k-1} Q_{k-1} \right) A, \tag{4.3}$$

which corresponds to damping the restricted coarse grid smoothing by an additional factor of $\frac{1}{2}$. Using relation (2.12), B^a may be rewritten as

$$B^a = \sum_{k=1}^{J} T_k \left(I - \frac{T_{k-1}}{2} \right), \tag{4.4}$$

where T_0 is identically zero. Expressing P_k as the telescoping series $\sum_{l=1}^{k} (P_l - P_{l-1})$, we can then write

$$B^a = \sum_{k=1}^{J} T_k (P_k - P_{k-1}) + \sum_{k=1}^{J} \sum_{l=1}^{k-1} T_k \left(I - \frac{T_{k-1}}{2} \right) (P_l - P_{l-1}). \tag{4.5}$$

Interchanging the order of summation in the second term in (4.5) allows us to rewrite B^a as

$$B^a = \sum_{l=1}^{J} T_l (P_l - P_{l-1}) + \sum_{l=1}^{J-1} \sum_{k=l+1}^{J} T_k \left(I - \frac{T_{k-1}}{2} \right) (P_l - P_{l-1}). \tag{4.6}$$

4.2. Richardson smoothing. First, consider the case when Richardson iteration is used as the smoother. We then have $R_k = \hat{R}_k$ and $T_k = \hat{T}_k$. Let $P_0 \equiv 0$ and define $w_l = (P_l - P_{l-1})v$, $l = 1, 2, \ldots, J$, for a given $v \in V_J$.

Our next lemma establishes a simple but important approximation property on each level.

LEMMA 4.3. *Let T_k, $k = 2, 3, \ldots, J$, satisfy bound* (2.22). *Then*

$$A\left(T_k\left(I - \frac{T_{k-1}}{2}\right)w_l, \left(I - \frac{T_{k-1}}{2}\right)w_l\right) \leq \left(1 + \frac{\gamma\sqrt{\theta}}{2}\right)^2 (\gamma\epsilon^{k-l})^2 A(w_l, w_l), \quad k = 2, 3, \ldots, J,\ l < k. \tag{4.7}$$

Proof. First, note that

$$A\left(T_k\left(I - \frac{T_{k-1}}{2}\right)w_l, \left(I - \frac{T_{k-1}}{2}\right)w_l\right) \leq A(T_k w_l, w_l) + |A(T_k w_l, T_{k-1} w_l)| + \frac{1}{4}A(T_k T_{k-1} w_l, T_{k-1} w_l). \tag{4.8}$$

Setting $T = T_k$, $u = w_l$, and $v = T_{k-1}w_l$ in Cauchy–Schwarz inequality (4.1), we have

$$|A(T_k w_l, T_{k-1} w_l)| \leq A(T_k w_l, w_l)^{\frac{1}{2}} A(T_k T_{k-1} w_l, T_{k-1} w_l)^{\frac{1}{2}}. \tag{4.9}$$

From (4.8) and (4.9), we thus have

$$A\left(T_k\left(I - \frac{T_{k-1}}{2}\right)w_l, \left(I - \frac{T_{k-1}}{2}\right)w_l\right) \leq \left(A(T_k w_l, w_l)^{\frac{1}{2}} + \frac{1}{2}A(T_k T_{k-1} w_l, T_{k-1} w_l)^{\frac{1}{2}}\right)^2. \tag{4.10}$$

Using assumption A.1 with $w = T_{k-1}w_l$ and applying assumption A.3, we see that the last term in (4.10) is bounded according to

$$A(T_k T_{k-1} w_l, T_{k-1} w_l) \leq (\gamma\epsilon)^2 \theta A(T_{k-1} w_l, w_l). \tag{4.11}$$

Applying assumption A.1 again, we have

$$A(T_{k-1} w_l, w_l) \leq (\gamma\epsilon^{k-l-1})^2 A(w_l, w_l). \tag{4.12}$$

Combining (4.11) and (4.12) yields

$$A(T_k T_{k-1} w_l, T_{k-1} w_l) \leq \gamma^2 \theta (\gamma\epsilon^{k-l})^2 A(w_l, w_l). \tag{4.13}$$

Now, using assumption A.1 to bound the first term on the right-hand side of (4.10) and (4.13) to bound the second term, we have

$$A\left(T_k\left(I - \frac{T_{k-1}}{2}\right)w_l, \left(I - \frac{T_{k-1}}{2}\right)w_l\right) \leq \left(1 + \frac{\gamma\sqrt{\theta}}{2}\right)^2 (\gamma\epsilon^{k-l})^2 A(w_l, w_l). \tag{4.14}$$

∎

Before we prove the main results of this section, we state the following useful identity (cf. [11, 47]).

LEMMA 4.4. *Let* $w_l = (P_l - P_{l-1})v,\ v \in V_J,\ l = 1, 2, \ldots, J$, *with* $P_0 = 0$. *Then*

$$\sum_{l=1}^{J} A(w_l, w_l) = A(v, v). \tag{4.15}$$

The next few lemmas are used to show that a symmetrized version of B^a has a uniformly bounded condition number. We first show that B^a is bounded uniformly in the A-norm.

LEMMA 4.5. *There exists constant* $C_1 > 0$, *independent of the number of levels* J, *such that*

$$A(B^a v, v) \le C_1 A(v, v) \quad \forall v \in V_J.$$

Proof. From (4.6), we have

$$\begin{aligned} A(B^a v, v) &= \sum_{l=1}^{J-1} \sum_{k=l+1}^{J} A\left(T_k \left(I - \frac{T_{k-1}}{2}\right) w_l, v\right) + \sum_{l=1}^{J} A(T_l w_l, v) \\ &= \sum_{l=1}^{J-1} \sum_{k=l+1}^{J} A\left(T_k \left(I - \frac{T_{k-1}}{2}\right) w_l, P_k v\right) + \sum_{l=1}^{J} A(T_l w_l, P_l v). \end{aligned}$$

Expressing P_k as a telescoping series, $P_k = \sum_{j=1}^{k} (P_j - P_{j-1})$, we have

$$\begin{aligned} A(B^a v, v) &= \sum_{l=1}^{J-1} \sum_{k=l+1}^{J} A\left(T_k \left(I - \frac{T_{k-1}}{2}\right) w_l, \sum_{j=1}^{k} (P_j - P_{j-1}) v\right) + \sum_{l=1}^{J} A(T_l w_l, P_l v) \\ &= \sum_{l=1}^{J-1} \sum_{k=l+1}^{J} \sum_{j=1}^{k} A\left(T_k \left(I - \frac{T_{k-1}}{2}\right) w_l, w_j\right) + \sum_{l=1}^{J} A(T_l w_l, P_l v). \end{aligned} \tag{4.16}$$

Now, applying Cauchy–Schwarz inequality (4.1) with $T = T_k,\ u = (I - T_{k-1}) w_l$, and $v = w_j$ yields

$$\begin{aligned} &A\left(T_k \left(I - \frac{T_{k-1}}{2}\right) w_l, w_j\right) \\ &\qquad \le A\left(T_k \left(I - \frac{T_{k-1}}{2}\right) w_l, \left(I - \frac{T_{k-1}}{2}\right) w_l\right)^{\frac{1}{2}} A(T_k w_j, w_j)^{\frac{1}{2}}. \end{aligned} \tag{4.17}$$

Bounding the first factor on the right-hand side of (4.17) using Lemma 4.3 and the second factor using assumption A.1 yields

$$A\left(T_k \left(I - \frac{T_{k-1}}{2}\right) w_l, w_j\right) \le \left(1 + \frac{\gamma\sqrt{\theta}}{2}\right) (\gamma \epsilon^{k-l}) A(w_l, w_l)^{\frac{1}{2}} (\gamma \epsilon^{k-j}) A(w_j, w_j)^{\frac{1}{2}}. \tag{4.18}$$

Finally, applying the arithmetic-geometric mean inequality in (4.18) yields

$$A\left(T_k \left(I - \frac{T_{k-1}}{2}\right) w_l, w_j\right) \le \frac{(1 + \frac{\gamma\sqrt{\theta}}{2}) \gamma^2 \epsilon^{2k-l-j}}{2} (A(w_l, w_l) + A(w_j, w_j)). \tag{4.19}$$

Using bound (4.19) for the individual terms in (4.16), we have

$$A(B^a v, v) \leq \left(\frac{(1+\frac{\gamma\sqrt{\theta}}{2})\gamma^2}{2}\right)\left(\sum_{l=1}^{J-1}\sum_{k=l+1}^{J}\sum_{j=1}^{k}\epsilon^{2k-l-j}(A(w_l,w_l)+A(w_j,w_j))\right)$$

$$+\sum_{l=1}^{J}A(T_l w_l, P_l v). \tag{4.20}$$

We now bound each of the three terms on the right-hand side of (4.20) individually.

Term I. First write

$$S_1 \equiv \sum_{l=1}^{J-1}\sum_{k=l+1}^{J}\sum_{j=1}^{k}\epsilon^{2k-l-j}A(w_l,w_l) = \sum_{l=1}^{J-1}\sum_{k=l+1}^{J}\epsilon^{k-l}A(w_l,w_l)\left(\sum_{j=1}^{k}\epsilon^{k-j}\right).$$

Hence,

$$S_1 \leq \left(\frac{1}{1-\epsilon}\right)\sum_{l=1}^{J-1}A(w_l,w_l)\left(\sum_{k=l+1}^{J}\epsilon^{k-l}\right).$$

Again, bounding terms involving powers of ϵ and applying Lemma 4.4, we have

$$S_1 \leq \frac{\epsilon}{(1-\epsilon)^2}\sum_{l=1}^{J}A(w_l,w_l) = \frac{\epsilon}{(1-\epsilon)^2}A(v,v). \tag{4.21}$$

Term II. Let

$$\begin{aligned} S_2 &\equiv \sum_{l=1}^{J-1}\sum_{k=l+1}^{J}\sum_{j=1}^{k}\epsilon^{2k-l-j}A(w_j,w_j) \\ &= \sum_{l=1}^{J-1}\sum_{k=l+1}^{J}\epsilon^{k-l}\left(\sum_{j=1}^{k}\epsilon^{k-j}A(w_j,w_j)\right). \end{aligned} \tag{4.22}$$

Now, let $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_J)^t$, where $\alpha_j = A(w_j, w_j)$, $j = 1, 2, \ldots, J$. Also, let $\mathcal{E} = (\mathcal{E}_{ij})_{1\leq i,j\leq J}$ denote the $J \times J$ lower triangular matrix with entries given by

$$\mathcal{E}_{ij} = \begin{cases} \epsilon^{i-j} & : \quad i \geq j, \\ 0 & : \quad i < j. \end{cases}$$

Finally, let $\underline{\beta}_l = (\beta_{1l}, \beta_{2l}, \ldots, \beta_{Jl})^t$, $l = 1, 2, \ldots, J$, denote column vectors with entries given by

$$\beta_{il} = \begin{cases} 0 & : \quad i \leq l, \\ \epsilon^{i-l} & : \quad i > l, \end{cases}$$

and define $\tilde{\beta}^t = (1, 1, \ldots, 1)$. Then, (4.22) can be written as

$$S_2 = \sum_{l=1}^{J-1}\sum_{k=1}^{J}\beta_{kl}\left(\sum_{j=1}^{J}\mathcal{E}_{kj}\alpha_j\right) = \sum_{l=1}^{J-1}\underline{\beta}_l^t\,\mathcal{E}\,\underline{\alpha} = \left(\sum_{l=1}^{J-1}\underline{\beta}_l^t\right)\mathcal{E}\,\underline{\alpha}. \tag{4.23}$$

Each entry of the column vector $\sum_{l=1}^{J-1} \underline{\beta}_l^t$ can be bounded by $\frac{\epsilon}{1-\epsilon}$. Since all quantities are nonnegative,

$$S_2 \le \left(\frac{\epsilon}{1-\epsilon}\right) \underline{\tilde{\beta}}^t \mathcal{E}\, \underline{\alpha} = \left(\frac{\epsilon}{1-\epsilon}\right) \sum_{k=1}^{J} \sum_{j=1}^{k} \mathcal{E}_{kj}\alpha_j. \tag{4.24}$$

Interchanging the order of summation and noting that $\mathcal{E}$ is lower triangular, we thus have

$$\begin{aligned} S_2 &\le \left(\frac{\epsilon}{1-\epsilon}\right) \sum_{j=1}^{J} \sum_{k=j}^{J} \mathcal{E}_{kj}\alpha_j \\ &= \left(\frac{\epsilon}{1-\epsilon}\right) \sum_{j=1}^{J} \sum_{k=j}^{J} \epsilon^{k-j} A(w_j, w_j) \\ &= \left(\frac{\epsilon}{1-\epsilon}\right) \sum_{j=1}^{J} A(w_j, w_j) \left(\sum_{k=j}^{J} \epsilon^{k-j}\right) \\ &\le \frac{\epsilon}{(1-\epsilon)^2} \sum_{j=1}^{J} A(w_j, w_j). \end{aligned} \tag{4.25}$$

Applying Lemma 4.4 in (4.25), we therefore have

$$S_2 \le \frac{\epsilon}{(1-\epsilon)^2} A(v, v). \tag{4.26}$$

Term III. Again using the telescoping series $P_k = \sum_{l=1}^{k} (P_l - P_{l-1})$, we have

$$\begin{aligned} S_3 \equiv \sum_{k=1}^{J} A(T_k w_k, P_k v) &= A(P_1 v, P_1 v) + \sum_{k=2}^{J} A(T_k w_k, P_k v) \\ &= A(P_1 v, P_1 v) + \sum_{k=2}^{J} A\left(T_k w_k, \sum_{l=1}^{k} (P_l - P_{l-1}) v\right) \\ &= A(P_1 v, P_1 v) + \sum_{k=2}^{J} \sum_{l=1}^{k} A(T_k w_k, w_l). \end{aligned}$$

Let $\hat{S}_3 = \sum_{k=2}^{J} \sum_{l=1}^{k} A(T_k w_k, w_l)$. Then applying Cauchy–Schwarz inequality (4.1) followed by assumption A.1 yields

$$\begin{aligned} \hat{S}_3 = \sum_{k=2}^{J} \sum_{l=1}^{k} A(T_k w_k, w_l) &\le \sum_{k=2}^{J} \sum_{l=1}^{k} A(T_k w_k, w_k)^{\frac{1}{2}} A(T_k w_l, w_l)^{\frac{1}{2}} \\ &\le \sum_{k=2}^{J} \sum_{l=1}^{k} \gamma A(w_k, w_k)^{\frac{1}{2}} (\gamma \epsilon^{k-l}) A(w_l, w_l)^{\frac{1}{2}} \\ &\le \gamma^2 \sum_{k=1}^{J} \sum_{l=1}^{J} A(w_k, w_k)^{\frac{1}{2}} (\epsilon^{|k-l|}) A(w_l, w_l)^{\frac{1}{2}}. \end{aligned} \tag{4.27}$$

Set $\alpha_k = A(w_k, w_k)^{\frac{1}{2}}$. Then

$$\hat{S}_3 \leq \gamma^2 \sum_{k=1}^{J} \sum_{l=1}^{J} (\epsilon^{|k-l|}) \alpha_k \alpha_l = \gamma^2 \langle\langle \hat{\mathcal{E}} \vec{\alpha}, \vec{\alpha} \rangle\rangle,$$

where $\hat{\mathcal{E}}$ is the $J \times J$ symmetric positive-definite matrix with entries $\epsilon^{|k-l|}$, $\vec{\alpha}$ is the $J \times 1$ column vector with entries α_k, $k = 1, 2, \ldots, J$, and $\langle\langle \cdot, \cdot \rangle\rangle$ is the Euclidean inner product. The largest eigenvalue of $\hat{\mathcal{E}}$ is bounded by its maximal row sum, which in turn is bounded by $\frac{2}{1-\epsilon}$. Therefore,

$$\begin{aligned} \hat{S}_3 &\leq \gamma^2 \left(\frac{2}{1-\epsilon} \right) \langle\langle \vec{\alpha}, \vec{\alpha} \rangle\rangle \\ &= \left(\frac{2\gamma^2}{1-\epsilon} \right) \left(\sum_{k=1}^{J} \alpha_k^2 \right) \\ &= \left(\frac{2\gamma^2}{1-\epsilon} \right) \left(\sum_{k=1}^{J} A(w_k, w_k) \right). \end{aligned} \tag{4.28}$$

From Lemma 4.4 and (4.28), we thus have

$$\begin{aligned} S_3 &= A(P_1 v, P_1 v) + \hat{S}_3 \\ &\leq A(v, v) + \left(\frac{2\gamma^2}{1-\epsilon} \right) A(v, v) \\ &= \left(1 + \frac{2\gamma^2}{1-\epsilon} \right) A(v, v). \end{aligned} \tag{4.29}$$

Substituting (4.21), (4.26), and (4.29) into (4.20), we therefore conclude that

$$A(B^a v, v) \leq \left(\frac{(1 + \frac{\gamma \theta^{\frac{1}{2}}}{2}) \gamma^2 \epsilon}{(1-\epsilon)^2} + 1 + \frac{2\gamma^2}{1-\epsilon} \right) A(v, v), \tag{4.30}$$

which proves the lemma with

$$C_1 = \left(\frac{\left(1 + \frac{\gamma \theta^{\frac{1}{2}}}{2} \right) \gamma^2 \epsilon}{(1-\epsilon)^2} + 1 + \frac{2\gamma^2}{1-\epsilon} \right). \qquad \square$$

The next lemma shows that B^a is coercive in the A inner product.

LEMMA 4.6. *Under assumptions* A.3 *and* A.4, *there exists constant* $C_0 > 0$, *independent of the number of levels* J, *such that*

$$A(B^a v, v) \geq C_0 A(v, v) \quad \forall v \in V_J.$$

Proof. From (4.4), we have that

$$A(B^a v, v) = \sum_{l=1}^{J} A\left(T_l\left(I - \frac{T_{l-1}}{2}\right)v, v\right)$$

$$= \sum_{l=1}^{J} A(T_l v, v) - \sum_{l=1}^{J} \frac{1}{2} A(T_l T_{l-1} v, v) \tag{4.31}$$

$$= \sum_{l=1}^{J} A(T_l v, v) - \sum_{l=1}^{J} \frac{1}{2} A(T_{l-1} v, T_l v). \tag{4.32}$$

Applying the standard Cauchy–Schwarz inequality yields

$$|A(T_{l-1}v, T_l v)| \le A(T_{l-1}v, T_{l-1}v)^{\frac{1}{2}} A(T_l v, T_l v)^{\frac{1}{2}}. \tag{4.33}$$

Applying assumption A.3 to (4.33), we get

$$|A(T_{l-1}v, T_l v)| \le \theta A(T_{l-1}v, v)^{\frac{1}{2}} A(T_l v, v)^{\frac{1}{2}}. \tag{4.34}$$

Hence, using the standard Cauchy–Schwarz inequality and nonnegativeness of operator T_J with respect to the A inner product we have

$$\sum_{l=1}^{J} A(T_l T_{l-1} v, v) \le \theta \sum_{l=1}^{J} A(T_{l-1}v, v)^{\frac{1}{2}} A(T_l v, v)^{\frac{1}{2}}$$

$$\le \theta \left(\sum_{l=1}^{J} A(T_{l-1}v, v)\right)^{\frac{1}{2}} \left(\sum_{l=1}^{J} A(T_l v, v)\right)^{\frac{1}{2}}$$

$$\le \theta \left(\sum_{l=1}^{J} A(T_l v, v)\right). \tag{4.35}$$

Relations (4.32) and (4.35) and assumption A.4 combine to show that

$$A(B^a v, v) \ge \left(1 - \frac{\theta}{2}\right)\left(\sum_{l=1}^{J} A(T_l v, v)\right)$$

$$\ge \frac{1}{2\eta}(2-\theta) A(v, v) \quad \forall v \in V_J. \tag{4.36}$$ ∎

The following theorem is a direct consequence of Lemmas 4.5 and 4.6.

THEOREM 4.7. *There exists constant $C > 0$, independent of the number of levels J, such that*

$$\kappa(B^s) \le C,$$

where $B^s = \frac{1}{2}(B^a + (B^a)^*)$ is the operator corresponding to symmetrized AFACx (Algorithm 4), with $(B^a)^*$ denoting the adjoint of B^a with respect to the A inner product.

4.3. More general smoothers. The estimates established above apply when the smoother used on each level is a Richardson iteration. In practice, simple but more robust smoothers such as damped Jacobi or Gauss–Seidel are usually employed. Assumption A.2 now becomes important in establishing condition number estimates for AFACx with general symmetric smoothers on each level. Lemma 4.3 is restated as follows for general $R_k \neq \hat{R}_k$ that is symmetric in the L^2 inner product.

LEMMA 4.8. *Let T_k satisfy* (2.21). *Then*

$$\begin{aligned} &A\left(T_l\left(I-\frac{T_{l-1}}{2}\right)w_k,\left(I-\frac{T_{l-1}}{2}\right)w_k\right)\\ &\leq a_1\left(1+\frac{\gamma\sqrt{a_1\theta}}{2}\right)^2\left(\gamma\epsilon^{k-l}\right)^2 A(w_k,w_k),\quad l=2,3,\ldots,J,\ k<l. \end{aligned} \tag{4.37}$$

The proof is along the same lines as that for Lemma 4.3.

Also, Lemma 4.5 now reads as follows.

LEMMA 4.9. *We have*

$$A(B^a v,v)\leq C_1 A(v,v)\quad \forall v\in V_J,$$

where

$$C_1=\left(a_1\left(1+\frac{\gamma\sqrt{a_1\theta}}{2}\right)\gamma^2\frac{\epsilon}{(1-\epsilon)^2}+1+\frac{2a_1\gamma^2}{(1-\epsilon)}\right).$$

Lemma 4.6 becomes the following.

LEMMA 4.10. *We have*

$$A(B^a v,v)\geq C_0 A(v,v)\quad \forall v\in V_J,$$

where $C_0=\frac{a_0}{2\eta}(2-\theta)$.

For symmetric smoothers that are spectrally equivalent to Richardson iteration, the condition number of the symmetrized AFACx operator is therefore again bounded independently of the number of levels.

4.4. Partial refinement. For the case of partial refinement, local "restricted coarse" $V_k^{h_{k-1}}$ is a subspace of $V_k^{h_k}\cap V_{k-1}$, and local "fine" $V_k^{h_k}$ is a subspace of V_k, $k=2,3,\ldots,J$. However, spaces $V_k^{h_k}$, $k=2,3,\ldots,J$, need not be nested. To treat this more general setting, we need to define operators at the different levels, projection operators between levels, and smoothing operators. Note that, in what follows, superscripts of "f" and "r" denote linear operators mapping to local "fine" $V_k^{h_k}$ and "restricted coarse" $V_k^{h_{k-1}}$, respectively, for given level k.

DEFINITION 4. *For $k=2,\ldots,J$, define operator $A_k^f: V_k^{h_k}\longrightarrow V_k^{h_k}$ by*

$$(A_k^f w,\phi)=A(w,\phi)\quad \forall\phi\in V_k^{h_k},\ w\in V_k^{h_k}.$$

DEFINITION 5. *For $k=2,\ldots,J$, define operator $A_k^r: V_k^{h_{k-1}}\longrightarrow V_k^{h_{k-1}}$ by*

$$(A_k^r w,\phi)=A(w,\phi)\quad \forall\phi\in V_k^{h_{k-1}},\ w\in V_k^{h_{k-1}}.$$

Orthogonal "elliptic" projection operators P_k^f, $k=1,2,\ldots,J$, and P_k^r, $k=2,3,\ldots,J$, are defined as follows.

DEFINITION 6. $P_k^f : V_J \longrightarrow V_k^{h_k}$ *is defined by*

$$A(P_k^f w, \phi) = A(w, \phi) \quad \forall \phi \in V_k^{h_k},\ w \in V_J.$$

DEFINITION 7. $P_k^r : V_J \longrightarrow V_k^{h_{k-1}}$ *is defined by*

$$A(P_k^r w, \phi) = A(w, \phi) \quad \forall \phi \in V_k^{h_{k-1}},\ w \in V_J.$$

Orthogonal "L^2" projection operators Q_k^f, $k = 1, 2, \ldots, J$, and Q_k^r, $k = 2, 3, \ldots, J$, are defined as follows.

DEFINITION 8. $Q_k^f : V_J \longrightarrow V_k^{h_k}$ *is defined by*

$$(Q_k^f w, \phi) = (w, \phi) \quad \forall \phi \in V_k^{h_k},\ w \in V_J.$$

DEFINITION 9. $Q_k^r : V_J \longrightarrow V_k^{h_{k-1}}$ *is defined by*

$$(Q_k^r w, \phi) = (w, \phi) \quad \forall \phi \in V_k^{h_{k-1}},\ w \in V_J.$$

Symmetric positive-definite smoothing operators $R_k^f : V_k \longrightarrow V_k^{h_k}$ and $R_k^r : V_{k-1} \longrightarrow V_k^{h_{k-1}}$ are also assumed to be defined.

The following relationships hold between the various operators: $Q_k^f A = A_k^f P_k^f$, $Q_k^r A = A_k^r P_k^r$, $R_k^f = R_k^f Q_k^f$, and $R_k^r = R_k^r Q_k^r$, $k = 2, 3, \ldots, J$.

For the case of partial refinement, we present the proof for only the AFACx operator with Richardson iteration as the smoother. It is obvious by now that the case of more general symmetric smoothers that are spectrally equivalent to Richardson iteration is easily handled through conditions like assumption A.2. Henceforth, let $R_1^f = (A_1^f)^{-1}$, $R_k^f = \frac{1}{\lambda_k} I$, $k = 2, 3, \ldots, J$, and $R_k^r = \frac{1}{\lambda_{k-1}} I$, $k = 2, 3, \ldots, J$. Define $T_k^f = R_k^f A_k^f P_k^f$ and $T_k^r = R_k^r A_k^r P_k^r$.

The following lemma is needed for the case of partial refinement.

LEMMA 4.11. *We have*

$$A(T_k^r v, v) \le A(T_{k-1}^f v, v) \quad \forall v \in V_J,\ k = 2, 3, \ldots, J. \tag{4.38}$$

Proof. From the basic properties of the L^2 projection operators listed in section 2.7, we have

$$\begin{aligned}
\|(Q_{k-1}^f - Q_k^r)u\|^2 &= ((Q_{k-1}^f - Q_k^r)u,\ (Q_{k-1}^f - Q_k^r)u) \\
&= \|Q_{k-1}^f u\|^2 + \|Q_k^r u\|^2 - 2(Q_{k-1}^f u, Q_k^r u) \\
&= \|Q_{k-1}^f u\|^2 + \|Q_k^r u\|^2 - 2(Q_{k-1}^f u, (Q_k^r)^2 u) \\
&= \|Q_{k-1}^f u\|^2 + \|Q_k^r u\|^2 - 2(Q_k^r Q_{k-1}^f u, Q_k^r u) \\
&= \|Q_{k-1}^f u\|^2 + \|Q_k^r u\|^2 - 2(Q_k^r u, Q_k^r u) \\
&= \|Q_{k-1}^f u\|^2 - \|Q_k^r u\|^2 \qquad \forall u \in V_J,\ k = 2, 3, \ldots, J. \qquad (4.39)
\end{aligned}$$

Since $\|(Q_{k-1}^f - Q_k^r)u\|^2 \ge 0$, (4.39) implies that

$$\|Q_k^r u\|^2 \le \|Q_{k-1}^f u\|^2 \quad \forall u \in V_J,\ k = 2, 3, \ldots, J. \tag{4.40}$$

Let $u = Av$. Then

$$\begin{aligned}
&\|Q_k^r Av\|^2 \le \|Q_{k-1}^f Av\|^2 \\
\Rightarrow\ &\|A_k^r P_k^r v\|^2 \le \|A_{k-1}^f P_{k-1}^f v\|^2 \\
\Rightarrow\ &\frac{1}{\lambda_{k-1}}(A_k^r P_k^r v, A_k^r P_k^r v) \le \frac{1}{\lambda_{k-1}}(A_{k-1}^f P_{k-1}^f v, A_{k-1}^f P_{k-1}^f v) \\
\Rightarrow\ &(R_k^r A_k^r P_k^r v, A_k^r P_k^r v) \le (R_{k-1}^f A_{k-1}^f P_{k-1}^f v, A_{k-1}^f P_{k-1}^f v) \\
\Rightarrow\ &A(T_k^r v, v) \le A(T_{k-1}^f v, v). \qquad \square
\end{aligned}$$

Assumptions similar to A.1 and A.3 in the previous section are made for the case of partial refinement. We refer to [14] for proof that the assumptions made below are valid in the case of partial refinement.

A.5. There exist constants $\epsilon \in (0,1)$ and $\gamma > 0$ such that

$$A(T_k^f w, w) \le (\gamma \epsilon^{k-l})^2 A(w,w) \quad \forall w \in V_l^{h_l},\ l \le k, \quad k = 1, 2, \ldots, J. \tag{4.41}$$

Then, from Lemma 4.11 and assumption A.5, we have

$$A(T_k^r w, w) \le (\gamma \epsilon^{k-l-1})^2 A(w,w) \quad \forall w \in V_l^{h_l},\ l \le k-1, \quad k = 2, \ldots, J. \tag{4.42}$$

A.6. There exists constant $\theta \in (0,2)$ such that

$$A(T_k^f v, T_k^f v) \le \theta A(T_k^f v, v) \quad \forall v \in V_J \tag{4.43}$$

and

$$A(T_k^r v, T_k^r v) \le \theta A(T_k^r v, v) \quad \forall v \in V_J, \quad k = 1, 2, \ldots, J. \tag{4.44}$$

In addition to the assumptions on the smoothers, a weak regularity assumption analogous to A.4 is also needed.

A.7. There exists a constant $\eta > 0$ such that

$$A(v,v) \le \eta \sum_{k=1}^{J} A(T_k^f v, v) \quad \forall v \in V_J. \tag{4.45}$$

Finally, we make the following assumption.

A.8. $\text{Range}(P_k - P_{k-1}) \subseteq V_k^{h_k}, \quad k = 1, 2, \ldots, J.$

AFACx operator B^a for the case of partial refinement can be written as

$$\begin{aligned}
B^a &= \sum_{k=1}^{J} R_k^f A_k^f P_k^f \left(I - \frac{1}{2} R_k^r A_k^r P_k^r \right) \\
&= \sum_{k=1}^{J} T_k^f \left(I - \frac{T_k^r}{2} \right).
\end{aligned} \tag{4.46}$$

The proofs of the following lemmas are virtually the same as the proofs for Lemmas 4.5 and 4.6, respectively. Assumptions A.5–A.8 and Lemma 4.11 take the place of assumptions A.1, A.3, and A.4.

LEMMA 4.12. *Under assumptions* A.5, A.6, *and* A.8, *there exists a constant* $C_1 > 0$, *independent of the number of levels* J, *such that*

$$A(B^a v, v) \le C_1 A(v,v) \ \ \forall v \in V_J. \tag{4.47}$$

LEMMA 4.13. *Under assumptions* A.5–A.8, *there exists a constant* $C_0 > 0$, *independent of the number of levels* J, *such that*

$$A(B^a v, v) \geq C_0 A(v, v) \quad \forall v \in V_J. \tag{4.48}$$

The following theorem follows immediately from Lemmas 4.12 and 4.13.

THEOREM 4.14. *There exists constant* $C > 0$, *independent of the number of levels* J, *such that*

$$\kappa(B^s) \leq C,$$

where $B^s = \frac{1}{2}(B^a + (B^a)^*)$ *is the operator corresponding to symmetrized AFACx (Algorithm* 4*).*

5. Conclusions. In this paper, we have presented a new multilevel condition number estimate for the AFACx algorithm. This estimate shows that the condition number of the AFACx operator does not degrade as the number of refinement levels in the AMR hierarchy increases. Numerical results supporting these theoretical estimates are presented in a forthcoming paper.

REFERENCES

[1] APPLIED NUMERICAL ALGORITHMS GROUP, *Chombo Reference Manual*, available online at http://seesar.lbl.gov/anag/software/chombo.html.

[2] S. BADEN, N. CHRISOCHOIDES, D. GANNON, AND M. NORMAN, EDS., *Structured Adaptive Mesh Refinement (SAMR) Grid Methods*, IMA Vol. Math. Appl. 117, Springer, New York, 2000.

[3] R. E. BANK, T. F. DUPONT, AND H. YSERENTANT, *The hierarchical basis multigrid method*, Numer. Math., 52 (1988), pp. 427–458.

[4] M. BERGER, *Adaptive Mesh Refinement for Hyperbolic Partial Differential Equations*, Ph.D. thesis, Stanford University, Stanford, CA, 1982.

[5] M. BERGER AND R. LEVEQUE, *AMRCLAW software*, available online at http://www.amath.washington.edu/řjl/amrclaw.

[6] M. J. BERGER AND J. OLIGER, *Adaptive mesh refinement for hyperbolic partial differential equations*, J. Comput. Phys., 53 (1984), pp. 484–512.

[7] P. E. BJORSTAD, R. MOE, AND M. D. SKOGEN, *Parallel domain decomposition and iterative refinement algorithms*, in Parallel Algorithms for Partial Differential Equations (Proceedings of the Sixth GAMM-Seminar, Kiel, 1990), W. Hackbusch, ed., Notes Numer. Fluid Mech. 31, Vieweg Verlag, Wiesbaden, 1990, pp. 216–227.

[8] F. BORNEMANN AND H. YSERENTANT, *A basic norm equivalence for the theory of multilevel methods*, Numer. Math., 64 (1993), pp. 455–476.

[9] F. A. BORNEMANN, *Interpolation spaces and optimal multilevel preconditioners*, in Domain Decomposition Methods in Scientific and Engineering Computing, D. E. Keyes and J. Xu, eds., Contemp. Math. 180, AMS, Providence, RI, 1994, pp. 3–8.

[10] D. BRAESS, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 1997.

[11] J. BRAMBLE, *Multigrid Methods*, Pitman Research Notes in Mathematics Series 24, Longman Scientific and Technical, Harlow, UK, 1993.

[12] J. BRAMBLE AND J. PASCIAK, *New convergence estimates for multigrid algorithms*, Math. Comp., 49 (1987), pp. 311–329.

[13] J. BRAMBLE AND J. PASCIAK, *The analysis of smoothers for multigrid algorithms*, Math. Comp., 58 (1992), pp. 467–488.

[14] J. BRAMBLE AND J. PASCIAK, *New estimates for multilevel algorithms including the V-cycle*, Math. Comp., 60 (1993), pp. 447–471.

[15] J. BRAMBLE, J. PASCIAK, J. WANG, AND J. XU, *Convergence estimates for multigrid algorithms without regularity assumptions*, Math. Comp., 57 (1991), pp. 23–45.

[16] J. BRAMBLE, J. PASCIAK, AND J. XU, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.

[17] J. BRAMBLE AND J. XU, *Some estimates for a weighted* L^2 *projection*, Math. Comp., 56 (1991), pp. 463–476.

[18] J. H. Bramble, J. E. Pasciak, and P. S. Vassilevski, *Computational scales of Sobolev norms with application to preconditioning*, Math. Comp., 69 (1999), pp. 463–480.
[19] A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.
[20] W. L. Briggs, V. E. Henson, and S. F. McCormick, *A Multigrid Tutorial*, 2nd ed., SIAM, Philadelphia, 2000.
[21] H. Cheng, *Iterative Solution of Elliptic Finite Element Problems on Partially Refined Meshes and the Effect of Using Inexact Solvers*, Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University, New York, 1993.
[22] H. Cheng, *On the optimality of some FAC and AFAC methods for elliptic finite element problems*, Taiwanese J. Math., 2 (1998), pp. 405–426.
[23] M. Dryja and O. B. Widlund, *On the optimality of an additive iterative refinement method*, in Proceedings of the Fourth Copper Mountain Conference on Multigrid Methods, J. Mandel, S. F. McCormick, J. E. Dendy, C. Farhat, G. Lansdale, S. V. Porter, J. W. Ruge, and K. Stüben, eds., SIAM, Philadelphia, 1989, pp. 161–170.
[24] P. Ferket and A. Reusken, *Further analysis of the local defect correction method*, Computing, 56 (1996), pp. 117–139.
[25] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.
[26] M. Griebel and P. Oswald, *On the abstract theory of additive and multiplicative Schwarz algorithms*, Numer. Math., 70 (1995), pp. 163–180.
[27] W. Hackbusch, *Local defect correction method and domain decomposition techniques*, in Defect Correction Methods, K Böhner and H. J. Stetter, eds., Comput. Suppl. 5, Springer, Vienna, 1984, pp. 89–113.
[28] L. Hart and S. McCormick, *Asynchronous multilevel adaptive methods for solving partial differential equations: Basic ideas*, Parallel Comput., 12 (1989), pp. 131–144.
[29] S. Kohn, X. Garaizar, R. Hornung, and S. Smith, *SAMRAI*, available online at http://www.llnl.gov/CASC/SAMRAI/.
[30] B. Lee, S. F. McCormick, B. Philip, and D. J. Quinlan, *Asynchronous fast adaptive composite-grid methods: Numerical results*, SIAM J. Sci. Comput., 25 (2003), pp. 682–699.
[31] J. Mandel and S. McCormick, *Iterative solution of elliptic equations with refinement: The model multi-level case*, in Domain Decomposition Methods (Los Angeles), T. F. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1989, pp. 81–92.
[32] S. McCormick, *Fast adaptive composite grid (FAC) methods: Theory for the variational case*, in Defect Correction Methods: Theory and Applications, K. Böhmer and H. J. Stetter, eds., Comput. Suppl. 5, Springer, Vienna, 1984, pp. 115–122.
[33] S. F. McCormick, *Multilevel Adaptive Methods for Partial Differential Equations*, Frontiers Appl. Math. 6, SIAM, Philadelphia, 1989.
[34] S. F. McCormick, *Multilevel Projection Methods for Partial Differential Equations*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 62, SIAM, Philadelphia, 1992.
[35] S. McCormick, S. McKay, and J. Thomas, *Computational complexity of the fast adaptive composite grid (FAC) method*, Appl. Numer. Math., 6 (1989), pp. 315–327.
[36] S. McCormick and D. Quinlan, *Asynchronous multilevel adaptive methods for solving partial differential equations on multiprocessors: Performance results*, Parallel Comput., 12 (1989), pp. 145–156.
[37] S. McCormick and D. Quinlan, *Idealized analysis of asynchronous multilevel methods*, in Adaptive Multilevel and Hierarchical Computational Strategies, Appl. Mech. Div. Ser. 157, American Society of Mechanical Engineers, New York, 1992, pp. 1–8.
[38] S. McCormick and J. Thomas, *The fast adaptive composite-grid method for elliptic equations*, Math. Comp., 46 (1986), pp. 439–456.
[39] R. Moe, *Iterative Local Uniform Mesh Refinement Methods and Parallel Processing*, Ph.D. thesis, University of Bergen, Bergen, Norway, 1992.
[40] B. Philip, *Asynchronous Fast Adaptive Composite Grid Methods for Elliptic Problems on Adaptively Refined Curvilinear Grids*, Ph.D. thesis, University of Colorado at Boulder, Boulder, CO, 2001.
[41] D. Quinlan, *Adaptive Mesh Refinement for Distributed Parallel Architectures*, Ph.D. thesis, University of Colorado at Denver, Denver, CO, 1993.
[42] D. Quinlan, *AMR++ Manual*, Technical report LA-UR-97-4325, Los Alamos National Laboratory, Los Alamos, NM, 1997.
[43] D. Quinlan, *AMR++ Tutorial*, Technical report, Lawrence Livermore National Laboratory, Livermore, CA, 2000.

[44] Y. Shapira, *Multigrid for locally refined meshes*, SIAM J. Sci. Comput., 21 (1999), pp. 1168–1190.
[45] O. B. Widlund, *Optimal iterative refinement methods*, in Domain Decomposition Methods (Los Angeles), T. F. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1989, pp. 114–125.
[46] J. Xu, *Theory of Multilevel Methods*, Ph.D. thesis, Cornell University, Ithaca, NY, 1989.
[47] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[48] J. Xu, *An introduction to multigrid convergence theory*, in Iterative Methods in Scientific Computing, T. Chan, R. Chan, and G. Golub, eds., Springer-Verlag, Singapore, 1997, pp. 169–241.
[49] H. Yserentant, *Old and new convergence proofs for multigrid methods*, Acta Numer., (1993), pp. 285–326.
[50] X. Zhang, *Multilevel Schwarz methods*, Numer. Math., 63 (1992), pp. 521–539.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 153–178

THE HIGH ACCURATE BLOCK-GRID METHOD FOR SOLVING LAPLACE'S BOUNDARY VALUE PROBLEM WITH SINGULARITIES*

A. A. DOSIYEV†

Abstract. A high accurate difference-analytical method is introduced for the solution of the mixed boundary value problem for Laplace's equation on graduated polygons. The polygon can have broken sections and be multiply connected. The uniform estimate of the error of the approximate solution is of order $O(h^6)$, whereas it is of order $O(h^6/r_j^{p-\lambda_j})$ for the errors of p-order derivatives $(p = 1, 2, \dots)$ in a finite neighborhood of reentry vertices; here, h is the mesh step, r_j is the distance from the current point to the vertex in question, $\lambda_j = 1/(a\alpha_j)$, and $a = 1$ or 2 depending on the types of boundary conditions. Further, $\alpha_j\pi$ is the value of the interior angle at the considered vertex. Numerical experiments are illustrated in section 8 to support the analysis made.

Key words. singularity, artificial boundary, block-grid method, matching operator

AMS subject classifications. 35A35, 35A40, 35C15, 65N06, 65N15, 65N22, 65N99

DOI. 10.1137/S0036142900382715

1. Introduction. As is well known, the use of the classical finite difference method (FDM) or the finite element method (FEM) for the solution of elliptic boundary value problems with singularities proves ineffective. A special construction is usually needed for the numerical scheme near the singularities. Various forms of these methods are often used for both the FDM and the FEM. To start with, (a) the use of a mesh refinement method dealing with the singular points (Volkov [23], [24] and Dosiyev [9] for the FDM; Thatcher [20] for the FEM) and (b) the use of singular terms of the series expansion of the exact solution around the singular points (Motz [16], Fox and Sankar [11], Volkov [25], and Fryazinov [12] in the FDM; Fix [10], Wait and Mitchell [30], Barnhill and Whiteman [5], Blum and Dobrowolski [6], and Olson, Georgiou, and Schultz [17] in the FEM) are two such forms. In the mesh refinement approach, the number of nodes, and consequently the number of unknowns, increase up to $O(h^{-2}\ln h^{-1})$ [9], [23], [24]. The use of the singular term approach, on the other hand, results in a system of ill-conditioned algebraic equations. This is because singular terms with a nonlocal carrier are added to the basic piecewise-polynomial functions with a local carrier, resulting in basic total functions that are "almost" linearly dependent for small h. Other forms also exist: (c) the use of a modified difference approximation in a fixed neighborhood of corners (Babuska and Rosenzweig [3], Andreev [1], [2], and Zenger and Gietl [34]) results in error bounds for a fixed point in the interior of the domain that are of the same order as in the case of smooth solutions but not uniform; and (d) a number of useful combinations of different methods can be used to take into account the differential properties of the solution in different parts of the domain and achieve an effective realization of the resulting system of algebraic equations (see Li [14]).

To elaborate the methods further, Li, Mathon, and Sermer [13] (see also [14]) use piecewise expansions into particular solutions to approximate the boundary conditions

*Received by the editors December 18, 2000; accepted for publication (in revised form) July 16, 2003; published electronically January 28, 2004.
http://www.siam.org/journals/sinum/42-1/38271.html
†Department of Mathematics, Eastern Mediterranean University, Gazimagusa, Cyprus, Mersin 10, Turkey (adiguzel.dosiyev@emu.edu.tr).

in a least-square sense. In this boundary method, to get high accurate results a large number of particular solutions are needed: in application to the Motz problem, 34 solutions are needed to result in an absolute error in the maximum norm of order $5.47E-9$. This large number of particular solutions may cause serious difficulties due to the ill conditioning of the associated least-squares matrices. Here, the condition number is $3.97E+07$, but it can be decreased to 3617 by dividing the given domain in the Motz problem into three subdomains. Different numbers of particular solutions are then used for each subdomain. However, even when the best combination of these numbers is used, the accuracy is of order $E-06$ only.

The method of auxiliary mapping (MAM) is introduced by Babuska and Oh [4] in the context of the p-version of the FEM to deal with corner singularities. Several extensions of the MAM that include boundary singularities and application to the h-p version of the FEM are given by Lucas and Oh [15]. The best result due to [15] in application to the Motz problem gives an absolute error in the maximum norm as $2.22E-08$ with an optimal mesh refinement when $p=10$. In [15], the results achieved are compared with the extremely accurate results of [13].

A sequence of approximation to the exact boundary conditions at an artificial boundary is given by Wu and Han [33]; exact boundary conditions are obtained around each singular point as a series. The original problem is then reduced to a boundary value problem in a domain away from the singularities; the FEM is applied to the resulting problem. When this method is applied to the Motz problem (see [33]), it is observed that only a few coefficients in the series expansion of the exact solution around the singular points can be approximated. Further, these approximations are not accurate, and increasing the number of terms in the approximation of the exact boundary conditions at an artificial boundary yields no improvement.

Finally, we mention the block method (BM) given by Volkov [26] for solving boundary value problems for Laplace's equation as one that is not a difference method. In the BM, the analytic properties of harmonic functions are most successfully used to construct an approximate solution. The approximate solution is an approximation of the integral representation of the harmonic function in a fixed number of blocks (sectors, semicircles, and circles) covering the given domain. With boundary conditions given by algebraic polynomials or analytic functions, the problem of convergence of the approximate solution and its derivatives of any order is investigated. This method was developed for different problems (see Volkov [29] and references therein). The rates of convergence of the approximate solution and its derivatives of the BM are higher (they converge exponentially with respect to the number of quadrature nodes) than the results obtained in above-mentioned approaches, but the application of the BM becomes restricted when the boundary function on some sides of polygon is given as a nonanalytic function (see Volkov [28]).

In light of the above-reviewed methods, a new difference-analytical method, called the block-grid method (BGM), for the solution of the mixed boundary value problems for Laplace's equation on graduated polygons is introduced. Thus, the proposed method is a combination of two methods which takes only superiorities of each one of them: the BM, which finely takes into account the behavior of the exact solution near the vertices of interior angles $\neq \pi/2$ of the polygon (on the "singular" part), and the FDM, which has a simple structure and high accuracy on square grids of the rectangles covering the remainder, "nonsingular"part of the polygon. A gluing operator of sixth order of accuracy is constructed for gluing together the grids and blocks. Furthermore, the restriction on the boundary functions to be algebraic polynomials is required only

on the sides of interior angles $\neq \pi/2$. The system of finite difference equations on the union of all rectangles may be solved by the alternating method of Schwarz with the number of iterations $O(\ln \varepsilon^{-1})$, where ε is the prescribed accuracy, by solving standard 9-point difference equations of Laplace on the rectangular domain at each iteration. The approximate solution on blocks is defined as a harmonic function, and any order of derivatives can be found by simple differentiation. The uniform estimate of the error of the approximate solution is of order $O(h^6)$, whereas it is of order $O(h^6/r_j^{p-\lambda_j})$ for the errors of p-order derivatives ($p = 1, 2, \dots$) in the "singular" part of the polygon; here r_j is the distance from the current point to the vertex of the block in question, and λ_j depends on the magnitude of the angle at the vertex and the type of boundary conditions on the sides of the considered block. Furthermore, when the boundary conditions are either Dirichlet or mixed type on the sides of interior angles $\neq \pi/2$, then the error of approximate solution on the block sectors decreases as $r_j^{\lambda_j} h^6$, which gives the additional accuracy of this approach near the singular points, with respect to existing FDM or FEM modifications for the singular problems. Finally, we illustrate the effectiveness of the method in solving the problem in L-shaped polygons with the corner and boundary singularities and the well-known Motz problem.

The BGM in the case of the Dirichlet problem on graduated polygons was given in [7], [8].

2. Boundary value problem on polygons. Let G be an open simply connected polygon with sides parallel to the x and y axes, let γ_j, $j = 1, 2, \dots, N$, be its sides, including the ends, enumerated counterclockwise, and let $\alpha_j \pi$, $0 < \alpha_j \leq 2$, be the interior angle formed by its sides γ_{j-1} and γ_j ($\gamma_0 = \gamma_N$). Denote by A_j the vertex of the jth angle, by s the arclength, measured along the boundary of G in the positive direction, and by s_j the value of s at A_j. Let r_j, θ_j be a polar system of coordinates with pole in A_j, where the angle θ_j is taken couterclockwise from the side γ_j, and ν_j is a parameter taking the values 0 or 1; further, $\overline{\nu}_j = 1 - \nu_j$.

We consider the boundary value problem

$$\Delta u = 0 \quad \text{on } G, \tag{2.1}$$

$$\nu_j u + \overline{\nu}_j u_n^{(1)} = \nu_j \varphi_j + \overline{\nu}_j \psi_j \quad \text{on } \gamma_j, \quad j = 1, 2, \dots, N, \tag{2.2}$$

where $\Delta \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$, $u_n^{(1)}$ is the derivative along the inner normal, φ_j and ψ_j are given functions, and

$$1 \leq \nu_1 + \nu_2 + \cdots + \nu_n \leq N, \tag{2.3}$$

$$\nu_j \varphi_j + \overline{\nu}_j \psi_j \in C_{6,\lambda}(\gamma_j), \quad 0 < \lambda < 1, \quad 1 \leq j \leq N. \tag{2.4}$$

Furthermore, at the vertices A_j for $\alpha_j = 1/2$ the conjugation conditions

$$\nu_j \varphi_j^{(2q+\delta_{\tau-2})} + \overline{\nu}_j \psi_j^{(2q+\delta_\tau)} = (-1)^{q+\delta_\tau+\delta_{\tau-1}} \left(\nu_{j-1} \varphi_{j-1}^{(2q+\delta_{\tau-1})} + \overline{\nu}_{j-1} \psi_{j-1}^{(2q+\delta_\tau)} \right) \tag{2.5}$$

(except maybe for $q = 3$, when $\tau = 3$) are satisfied; $\tau = \nu_{j-1} + 2\nu_j$, $q = 0, 1, \dots, \overline{Q}$, $\overline{Q} = [(6 - \delta_{\tau-1} - \delta_{\tau-2})/2] - \delta_\tau$, and $\delta_w = 1$ when $w = 0$; $\delta_w = 0$ when $w \neq 0$. At the vertices A_j with $\alpha_j \neq 1/2$ no compatibility conditions are required to hold for the boundary conditions; in particular, the values of φ_{j-1} and φ_j at A_j might be different. In addition, we require that when $\alpha_j \neq 1/2$ the boundary functions on γ_{j-1} and γ_j be given as algebraic polynomials of arclength s measured along γ.

We represent the given boundary functions (algebraic polynomials) on γ_{j-1} and γ_j for $\alpha_j \neq 1/2$ in the form

$$\sum_{k=0}^{\tau_{j-1}} a_{jk}^0 r_j^k \quad \text{and} \quad \sum_{k=0}^{\tau_j} b_{jk}^0 r_j^k, \tag{2.6}$$

respectively; here a_{jk}^0 and b_{jk}^0 are numerical coefficients, and τ_{j-1} and τ_j are degrees of those polynomials.

Let $E = \{j : \alpha_j \neq 1/2, j = 1, 2, \dots, N\}$. In the neighborhood of A_j, $j \in E$, we construct two fixed block-sectors $T_j^i = T_j(r_{ji}) \subset G$, $i = 1, 2$, where $0 < r_{j2} < r_{j1} < \min\{s_{j+1} - s_j, s_j - s_{j-1}\}$, $T_j(r) = \{(r_j, \theta_j) : \ 0 < r_j < r, \ 0 < \theta_j < \alpha_j \pi\}$.

On the closed sector $\overline{T}_j^1$, $j \in E$, we consider a function $Q_j(r_j, \theta_j)$ with the following properties:

(i) $Q_j(r_j, \theta_j)$ is harmonic and bounded on the open sector T_j^1;

(ii) continuous on $\overline{T}_j^1$ everywhere, except for the point A_j (the vertex of the sector) for $\nu_j = \nu_{j-1} = 1$ and $a_{j0}^0 \neq b_{j0}^0$, where a_{j0}^0 and b_{j0}^0 are given numbers from (2.6), i.e., for boundary conditions discontinuous at A_j;

(iii) continuously differentiable on $\overline{T}_j^1 \setminus A_j$ and satisfies the boundary conditions (2.2) on $\gamma_{j-1} \cap \overline{T}_j^1$ and $\gamma_j \cap \overline{T}_j^1$, $j \in E$.

For definiteness we assume that $Q_j(r_j, \theta_j)$, with the above properties, has the form (3.2)–(3.9) in [29].

Remark 2.1. For the case of $\nu_{j-1} = \nu_j = 1$, we formally set the value of $Q_j(r_j, \theta_j)$ and the solution u of problem (2.1), (2.2) at the vertex A_j equal to $(a_{j0}^0 + b_{j0}^0)/2$.

We set (see Volkov [26])

$$R(m, m, r, \theta, \eta) = R(r, \theta, \eta) + (-1)^m R(r, \theta, -\eta), \tag{2.7}$$

$$R(1 - m, m, r, \theta, \eta) = R(m, m, r, \theta, \eta) - (-1)^m R(m, m, r, \theta, \pi - \eta), \tag{2.8}$$

where

$$R(r, \theta, \eta) = \frac{1 - r^2}{2\pi(1 - 2r\cos(\theta - \eta) + r^2)} \tag{2.9}$$

is the kernel of the Poisson integral for a unit circle. We specify the kernel

$$R_j(r_j, \theta_j, \eta) = \lambda_j R\left(\nu_{j-1}, \nu_j, \left(\frac{r_j}{r_{j2}}\right)^{\lambda_j}, \lambda_j \theta_j, \lambda_j \eta\right), \quad j \in E, \tag{2.10}$$

where

$$\lambda_j = \frac{1}{(2 - \nu_{j-1}\nu_j - \overline{\nu}_{j-1}\overline{\nu}_j)\alpha_j}. \tag{2.11}$$

The following lemma shows important properties of the solution, which will be used to construct the proposed method.

LEMMA 2.2. *The solution u of the boundary value problem* (2.1), (2.2) *can be represented on $\overline{T}_j^2 \setminus V_j$, $j \in E$ in the form*

$$u(r_j, \theta_j) = Q_j(r_j, \theta_j) + \int_0^{\alpha_j \pi} (u(r_{j2}, \eta) - Q_j(r_{j2}, \eta)) R_j(r_j, \theta_j, \eta) d\eta, \tag{2.12}$$

and the harmonic function $u_n(r_j,\theta_j)$, which is obtained as a result of approximation of the integral in (2.12) *by the composite rectangular formula with n equally spaced nodes, satisfies the inequality*

$$| u - u_n | \le c_j \exp(-d_j n) \quad \text{on } \overline{T}_j^3, \quad 0 < r_{j3} < r_{j2}, \tag{2.13}$$

where V_j is the curvilinear part of the boundary of the sector T_j^2, and c_j and d_j are positive constants independent on n.

The proof of Lemma 2.2 follows from Theorems 3.1 and 5.1 in [29].

Remark 2.3. The constants c_j and d_j in Lemma 2.2 depend on the radius r_{j3} of T_j^3, and the sequence $u_n(r_j,\theta_j)$, $n = 1,2,\dots$, is unbounded in the open sector T_j^2 and hence diverges in T_j^2 in the uniform metric. Therefore, the exponential convergence of $u_n(r_j,\theta_j)$, $n = 1,2,\dots$, is guaranteed only in the sector $\overline{T}_j^3$ with the radius $r_{j3} < r_{j2}$.

3. 9-point solution on rectangles. Let $\Pi = \{(x,y) : 0 < x < a, 0 < y < b\}$ be a rectangle, a/b be rational, γ_j $(j = 1,2,3,4)$ be the sides, including the ends, enumerated counterclockwise starting from the left side ($\gamma_0 \equiv \gamma_4$, $\gamma_5 \equiv \gamma_1$), and let $\gamma = \cup_{j=1}^4 \gamma_j$ be the boundary of Π, let ν_j be a parameter taking the values 0 or 1, and let $\overline{\nu}_j = 1 - \nu_j$.

We consider the boundary value problem (2.1)–(2.5) on $G \equiv \Pi$:

$$\Delta u = 0 \quad \text{on } \Pi, \tag{3.1}$$

$$\nu_j u + \overline{\nu}_j u_n^{(1)} = \nu_j \varphi_j + \overline{\nu}_j \psi_j \quad \text{on } \gamma_j, \quad j = 1,2,3,4. \tag{3.2}$$

Let $h > 0$, with $a/h \ge 2$, $b/h \ge 2$, be integers. We assign Π^h, a square net on Π, with step h, obtained with the lines x, $y = 0, h, 2h, \dots$. Let γ_j^h be a set of nodes on the interior of γ_j, and let

$$\dot{\gamma}_j^h = \gamma_j \cap \gamma_{j+1}, \quad \gamma^h = \bigcup (\gamma_j^h \cup \dot{\gamma}_j^h), \quad \overline{\Pi}^h = \Pi^h \cup \gamma^h.$$

We consider the system of finite difference equations (see [22])

$$u_h = B u_h \quad \text{on } \Pi^h, \tag{3.3}$$

$$u_h = \overline{\nu}_j B_j u_h + E_{jh}(\varphi_j, \psi_j) \quad \text{on } \gamma_j^h, \tag{3.4}$$

$$u_h = \overline{\nu}_j \overline{\nu}_{j+1} \dot{B}_j u_h + \dot{E}_{j,h}(\varphi_j, \varphi_{j+1,}, \psi_j, \psi_{j+1}) \text{ on } \dot{\gamma}_j^h, \quad j = 1,2,3,4, \tag{3.5}$$

where

$$\begin{aligned} Bu(x,y) \equiv {} & (u(x+h,y) + u(x,y+h) + u(x-h,y) \\ & + u(x,y-h))/5 + (u(x+h,y+h) + u(x-h,y+h) \\ & + u(x-h,y-h) + u(x+h,y-h))/20, \end{aligned} \tag{3.6}$$

the operators B_j, E_{jh}, $\dot{B}_j$, $\dot{E}_{jh}$ in the right coordinate system with the axis x_j, directed along γ_{j+1} and the axis y_j, directed along γ_j have the expressions

$$\begin{aligned} B_j u(0,y_j) \equiv {} & (2u(h,y_j) + u(0,y_j+h) + u(0,y_j-h))/5 \\ & + (u(h,y_j+h) + u(h,y_j-h))/10, \end{aligned} \tag{3.7}$$

$$E_{jh}(\varphi_j,\psi_j) \equiv \nu_j \varphi_j - \overline{\nu}_j \left(\frac{3h}{5} \psi_j - \frac{2h^5}{5!5} \psi_j^{(4)} - \frac{2h^7}{7!} \psi_j^{(6)} \right), \tag{3.8}$$

$$\dot{B}_j u(0,0) \equiv (2u(h,0) + 2u(0,h) + u(h,h))/5, \tag{3.9}$$

$$\begin{aligned}\dot{E}_{j,h}(\varphi_j, \varphi_{j+1}, \psi_j, \psi_{j+1}) \equiv \nu_j\varphi_j + \overline{\nu}_j\nu_{j+1}\varphi_{j+1} - \overline{\nu}_j\overline{\nu}_{j+1}\left(\frac{3h}{5}(\psi_j + \psi_{j+1})\right.\\ + \frac{h^2}{5}\psi_{j+1}^{(1)} - \frac{2h^5}{5!5}\left(\psi_j^{(4)} + \psi_{j+1}^{(4)}\right)\\ \left. + \frac{4h^6}{6!5}\left(\psi_j^{(5)} + \psi_{j+1}^{(5)}\right) - \frac{2h^7}{7!}\left(\psi_j^{(6)} + \psi_{j+1}^{(6)}\right)\right).\end{aligned} \tag{3.10}$$

The system of finite difference equations (3.3)–(3.5) which has nonnegative coefficients with the conditions (2.3) is uniquely solvable.

THEOREM 3.1. *Let u be the solution of problem* (3.1), (3.2). *Then*

$$\max_{\overline{\Pi}^h} \mid u_h - u \mid \leq ch^6, \tag{3.11}$$

where u_h is the solution of the system (3.3)–(3.5), *and c is a constant independent of h.*

Proof. Theorem 3.1 is proved in the same way as Theorem 1.1 of [24] except that instead of the function (1.14) of [24] we need to take the function

$$u_j(x,y) = (-1)^j\nu_j\nu_{j+1}\frac{2\left(\widetilde{\varphi}_j^{(6)}(s_{j+1}) + \widetilde{\varphi}_{j+1}^{(6)}(s_{j+1})\right)}{\pi 6!}\operatorname{Im}\{(z-z_j)^6\ln(z-z_j)\}$$

(making the corresponding changes), where $\widetilde{\varphi}_l = \nu_l\varphi_l$, $z = x + iy$, and $z_j = x_j + iy_j$ is the complex coordinate of the vertex $\gamma_j \cap \gamma_{j+1}$. □

4. Description of the block-grid method. Let us consider in addition to the sectors T_j^1, T_j^2, and T_j^3 (see section 2) in the neighborhood of each vertex A_j, $j \in E$, of the polygon G, a sector T_j^4, where $0 < r_{j4} < r_{j3} < r_{j2}$, and let $G^T = G \setminus (\cup_{j\in E}\overline{T}_j^4)$.

The construction of the proposed method can be divided into the following steps.

Step 1. We blockade all singular corners $A_j, j \in E$, by the double sectors $T_j^i = T_j(r_{ji})$, $i = 2,3$, with $r_{j3} < r_{j2}$, $T_k^2 \cap T_l^3 = \emptyset$, $k \neq l$, $k,l \in E$, and cover the given polygon by overlapping rectangles Π_k, $k = 1,2,\ldots,M$, and sectors $T_j^3, j \in E$, such that the distance from $\overline{\Pi}_k$ to the singular point A_j is not less than r_{j4} for all $k = 1,2,\ldots,M$ and $j \in E$ (see Figure 1, in the case when $M = 4$ and $E = \{1\}$; i.e., the singular point is only A_1).

Step 2. On each rectangle Π_k, we use the 9-point scheme for the approximation of Laplace's equation on square grids with the step size $h_k \leq h$, h being the parameter, and as an approximate solution on $\overline{T}_j^3, j \in E$ (in singular parts), we take the harmonic function $u_n(r_j, \theta_j)$, defined in Lemma 2.2 (see also Remark 2.3).

Step 3. We construct the sixth order matching operator to connect the subsystems.

Step 4. We use Schwarz's alternating procedure by solving standard difference equations of Laplace on the rectangular domain at each iteration.

Now we describe the procedure of obtaining the algebraic system of equations for the numerical solution of problem (2.1), (2.2).

Let $\Pi_k \subset G_T$, $k = 1,2,\ldots,M$ $(M < \infty)$ be certain fixed open rectangles with sides a_{1k} and a_{2k} parallel to the x and y axes, with a_{1k}/a_{2k} rational and

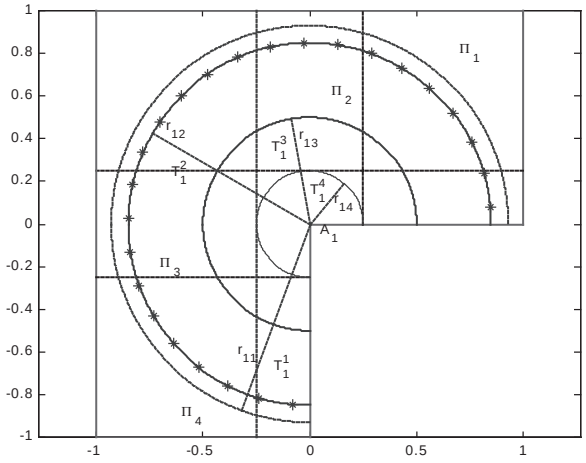


FIG. 1.

$G \subset (\cup_{k=1}^{M} \Pi_k) \cup (\cup_{j \in E} T_j^3) \subset G$. We denote by η_k the boundary of the rectangle Π_k and by V_j the curvilinear part of the boundary of the sector T_j^2, and let $t_j = (\cup_{k=1}^{M} \eta_k) \cap \overline{T}_j^3$. The following general requirement is imposed on the arrangement of the rectangles Π_k, $k = 1, 2, \ldots, M$: any point P lying on $\eta_k \cap G_T$, $1 \le k \le M$, or located on $V_j \cap G$, $j \in E$, falls inside at least one of the rectangles $\Pi_{k(p)}$, $1 \le k(p) \le M$, depending on P, the distance from P to $G_T \cap \eta_{k(p)}$ being not less than some constant κ_0 independent of P.

We will call the quantity κ_0 the *depth* of a gluing of the rectangles Π_k, $k = 1, 2, \ldots, M$. We introduce a parameter $h \in (0, \kappa_0/4]$ and define a square grid on Π_k, $1 \le k \le M$, with maximal possible step $h_k \le \min\{h, \min\{a_{1k}, a_{2k}\}/6\}$ such that the boundary η_k lies entirely on the grid lines. Let Π_k^h be the set of grid nodes on Π_k, let η_k^h be the set of nodes on η_k, and let $\overline{\Pi}_k^h = \Pi_k^h \cup \eta_k^h$. We denote the set of nodes in the closure of $\eta_k \cap G^T$ by η_{k0}^h, the set of nodes on t_j by t_j^h, and the set of remaining nodes on η_k by η_{k1}^h. We also specify a natural number $n \ge [\ln^{1+\varkappa} h^{-1}] + 1$, where $\varkappa > 0$ is a fixed number, and the quantities $n(j) = \max\{4, [\alpha_j n]\}$, $\beta_j = \alpha_j \pi / n(j)$, and $\theta_j^m = (m - 1/2)\beta_j$, $j \in E$, $1 \le m \le n(j)$. On the arc V_j, we choose the points (r_{j2}, θ_j^m), $1 \le m \le n(j)$, and denote the set of these points by V_j^n. Let

$$\omega^{h,n} = \left(\bigcup_{k=1}^{M} \eta_{k0}^h \right) \cup \left(\bigcup_{j \in E} V_j^n \right), \quad \overline{G}_*^{h,n} = \omega^{h,n} \cup \left(\bigcup_{k=1}^{M} \overline{\Pi}_k^h \right).$$

Let φ_j and ψ_j be given functions from the boundary conditions (2.2), and let $\varphi = \{\varphi_j\}_{j=1}^N$ and $\psi = \{\psi_j\}_{j=1}^N$. On the set $\omega^{h,n}$ we introduce the linear matching operator S^6. The value of $S^6(u_h, \varphi, \psi)$ at the point $P \in \omega^{h,n}$ is defined linearly in terms of the values of the function u_h at the nodes of the grid constructed on the rectangle $\Pi_{k(p)} \ni P$ and the assigned boundary values of $\varphi^{(m)}$, $m = 0, 1, \ldots, 5$ (or $\psi^{(q)}$, $q = 0, 1, \ldots, 4$) at a fixed number of points. If there is more than one rectangle containing P, we choose $\Pi_{k(p)}$ such that part of the boundary $\eta_{k(p),0}$ is the maximum distance away from P. The pattern of the operator S^6 lies in a neighborhood $O(h)$ of the point P, and in a uniform metric for $\varphi \equiv 0$ and $\psi \equiv 0$ its norm is not greater than one.

Moreover, $u - S^6(u,\varphi,\psi) = O(h^6)$ uniformly on $\omega^{h,n}$. An operator S^6 with these properties will be constructed in section 5 by developing the method given in [8] for the Dirichlet problem.

Let

$$R_j^{(q)}(r_j,\theta_j) = \frac{R_j(r_j,\theta_j,\theta_j^q)}{\max\left\{1, \beta_j \sum_{p=1}^{n(j)} R_j(r_j,\theta_j,\theta_j^p)\right\}}, \tag{4.1}$$

where $R_j(r,\theta,\eta)$ is the kernel defined by (2.10). It is easy to check that

$$0 \le R_j^{(q)}(r_j,\theta_j) \le R_j(r_j,\theta_j,\theta_j^q), \tag{4.2}$$

where $j \in E$, $0 \le q \le n(j)$. Furthermore, from the estimation (2.29) in [26] follows the existence of the positive constants n_0 and σ such that, for $n \ge n_0$,

$$\max_{(r_j,\theta_j)\in\overline{T}_j^3} \beta_j \sum_{q=1}^{n(j)} R_j(r_j,\theta_j,\theta_j^q) \le \sigma < 1 \tag{4.3}$$

when $\nu_{j-1} + \nu_j \ge 1$, and on the basis of (4.1) and (4.2)

$$0 \le \beta_j \sum_{q=1}^{n(j)} R_j^{(q)}(r_j,\theta_j) \le 1, \ j \in E \tag{4.4}$$

when $\nu_{j-1} = \nu_j = 0$.

Consider the system of linear algebraic equations

$$u_h = Bu_h \ \text{ on } \Pi_k^h, \quad u_h = \bar\nu_m B_m u_h + E_{mh}(\varphi,\psi) \ \text{ on } \eta_{k1}^h \cap \gamma_m, \tag{4.5}$$

$$u_h = \bar\nu_m \bar\nu_{m+1} \dot B_m u_h + \dot E_{mh}(\varphi_m,\varphi_{m+1},\psi_m,\psi_{m+1}) \quad \text{on } \eta_{k1}^h \cap \gamma_m \cap \gamma_{m+1}, \tag{4.6}$$

$$u_h(r_j,\theta_j) = Q_j(r_j,\theta_j) + \beta_j \sum_{q=1}^{n(j)} R_j^{(q)}(r_j,\theta_j)(u_h(r_{j2},\theta_j^q) - Q_j(r_{j2},\theta_j^q)), (r_j,\theta_j) \in t_j^h, \tag{4.7}$$

$$u_h = S^6(u_h,\varphi,\psi) \ \text{ on } \omega^{h,n}, \tag{4.8}$$

where $1 \le m \le N$, $1 \le k \le M$, $j \in E$; B, B_m, $\dot B_m$, E_{mh}, and $\dot E_{mh}$ are defined by the formulas (3.6)–(3.10), respectively; $Q_j(r_j,\theta_j)$ is the harmonic polynomial described in section 2.

Definition 4.1. *The solution of the system* (4.5)–(4.8) *is called a numerical solution of the problem* (2.1), (2.2) *on* $\overline{G}_*^{h,n}$.

Definition 4.2. *Let* u_h *be the solution of the system* (4.5)–(4.8). *The function*

$$U_h(r_j,\theta_j) = Q_j(r_j,\theta_j) + \beta_j \sum_{q=1}^{n(j)} R_j(r_j,\theta_j,\theta_j^q)(u_h(r_{j2},\theta_j^q) - Q_j(r_{j2},\theta_j^q)) \tag{4.9}$$

is called an approximate solution of the problem (2.1), (2.2) *on the closed block* $\overline{T}_j^3$, $j \in E$.

Definition 4.3. *The system* (4.5)–(4.9) *is called the block-grid equations.*

5. Construction of the sixth order matching operator. Let $\Omega = \{(x, y) : x^2 + y^2 < 1\}$ be a unit circle and $u \in C_{6,0}(\overline{\Omega})$ be a harmonic function on Ω. On the basis of Taylor's formula for any point $(x, y) \in \Omega$, we have

$$u(x,y) = \sum_{k=0}^{5} a_k \operatorname{Re}\ z^k + \sum_{k=1}^{5} b_k \operatorname{Im}\ z^k + O(r^6), \tag{5.1}$$

where $r = \sqrt{x^2 + y^2}$,

$$\begin{aligned} &a_0 = u(0,0), \quad a_1 = \frac{\partial u(0,0)}{\partial x}, \quad a_2 = \frac{1}{2}\frac{\partial^2 u(0,0)}{\partial x^2}, \quad a_3 = \frac{1}{3!}\frac{\partial^3 u(0,0)}{\partial x^3}, \\ &a_4 = \frac{1}{4!}\frac{\partial^4 u(0,0)}{\partial x^4}, \quad a_5 = \frac{1}{5!}\frac{\partial^5 u(0,0)}{\partial x^5}; \end{aligned} \tag{5.2}$$

$$\begin{aligned} &b_1 = \frac{\partial u(0,0)}{\partial y}, \quad b_2 = \frac{1}{2}\frac{\partial^2 u(0,0)}{\partial x \partial y}, \quad b_3 = \frac{1}{3!}\frac{\partial^3 u(0,0)}{\partial x^2 \partial y}, \\ &b_4 = \frac{1}{4!}\frac{\partial^4 u(0,0)}{\partial x^3 \partial y}, \quad b_5 = \frac{1}{5!}\frac{\partial^5 u(0,0)}{\partial y^5}. \end{aligned} \tag{5.3}$$

We denote

$$F_5(x,y) = \sum_{k=0}^{5} a_k \operatorname{Re} z^k + \sum_{k=1}^{5} b_k \operatorname{Im} z^k. \tag{5.4}$$

We construct the operator S^6 from the condition that the expression $S^6(F_5, \varphi, \psi)$ gives the exact value of any harmonic polynomial $F_5(x, y)$ defined by the formula (5.4) at each point $P \in \omega^{h,n}$. For simplicity, we will denote the step size of the square grid in the rectangle containing the point $P \in \omega^{h,n}$ by h.

Case 1. The point $P \in \omega^{h,n}$ lies on a grid line. We place the origin of the rectangular system of coordinates at the node P_0 and direct the positive axis of x along the grid line so that $P = P(\delta h, 0)$, $1 \le \delta \le 3/2$. We take points $P_1(2h, 0)$, $P_2(3h, 0)$, $P_3(2h, h)$, $P_4(h, h)$, $P_5(0, h)$. First, we find numerical coefficients λ', λ'_k, $k = 1, \dots, 5$, such that the representation

$$u_0 = \lambda' u + \lambda'_1 u_1 + \lambda'_2 u_2 + \lambda'_3 u_3 + \lambda'_4 u_4 + \lambda'_5 u_5 \tag{5.5}$$

is satisfied for the harmonic polynomials $\operatorname{Re} z^n, n = 0, 1, \dots, 5$, where $u = u(P)$, $u_k = u(P_k)$, $k = 0, 1, \dots, 5$, $z = x + iy$. We then have

$$\begin{aligned} \lambda' + \lambda'_1 + \lambda'_2 + \lambda'_3 + \lambda'_4 + \lambda'_5 &= 1, \\ \delta\lambda' + 2\lambda'_1 + 3\lambda'_2 + 2\lambda'_3 + \lambda'_4 &= 0, \\ \delta^2\lambda' + 4\lambda'_1 + 9\lambda'_2 + 3\lambda'_3 - \lambda'_5 &= 0, \\ \delta^3\lambda' + 8\lambda'_1 + 27\lambda'_2 + 2\lambda'_3 - 2\lambda'_4 &= 0, \\ \delta^4\lambda' + 16\lambda'_1 + 81\lambda'_2 - 7\lambda'_3 - 4\lambda'_4 + \lambda'_5 &= 0, \\ \delta^5\lambda' + 32\lambda'_1 + 243\lambda'_2 - 38\lambda'_3 - 4\lambda'_4 &= 0. \end{aligned} \tag{5.6}$$

Solving system (5.6), we obtain $\lambda' = 1/(1 - \mu_0), \lambda'_q = -\mu_q/(1 - \mu_0), q = 1, \dots, 5$,

where

$$\mu_1 = \delta^2(21 - 34\delta + 21\delta^2 - 4\delta^3)/20, \quad \mu_2 = \delta(\delta - 1)(3\delta^3 - 12\delta^2 + 13\delta - 2)/60,$$
$$\mu_3 = \delta(18 - 87\delta + 153\delta^2 - 87\delta^3 + 15\delta^4)/120, \quad \mu_4 = \delta(6 + \delta - 4\delta^2 + \delta^3)/10,$$
$$\mu_5 = \delta(30 - 49\delta + 31\delta^2 - 9\delta^3 + \delta^4)/40, \quad \mu_0 = \sum_{j=1}^{5} \mu_i.$$

It is easy to check that the inequalities $\lambda' > 0$, $\lambda'_1 < 0$, $\lambda'_2 \le 0$, $\lambda'_3 < 0$, $\lambda'_4 < 0$, $\lambda'_5 < 0$ hold for $1 \le \delta \le 3/2$, with $\lambda'_2 = 0$ only when $\delta = 1$.

Now we take the nodal points $P_6(2h, -h)$, $P_7(h, -h)$, and $P_8(0, -h)$, respectively, symmetric to the points P_3, P_4, and P_5 with respect to the x-axis. Since $\operatorname{Im} z^k = 0$, $k = 0, 1, \dots, 5$, for $y = 0$ and is odd with respect to y, and $\operatorname{Re} z^k$, $k = 0, 1, \dots, 5$, is even with respect to y, from (5.5) we obtain the expression

$$S^6 u \equiv \sum_{k=0}^{8} \lambda_k u_k, \tag{5.7}$$

which gives the exact value of the harmonic polynomial $F_5(x, y)$ at the point P, where

$$\lambda_0 = 1 - \mu_0, \quad \lambda_1 = \mu_1, \quad \lambda_2 = \mu_2, \quad \lambda_{q+3} = \lambda_q = \mu_q/2, \quad q = 3, 4, 5. \tag{5.8}$$

It is easy to check that

$$\lambda_2 \ge 0, \quad \lambda_q > 0, \quad q \ne 2; \quad \sum_{k=0}^{8} \lambda_k = 1, \tag{5.9}$$

and $S^6 u \equiv Bu$ for $\delta = 1$, where B is the operator of (3.6).

Case 2. The point $P \in \omega^{h,n}$ lies inside a grid cell and in the rectangular system of coordinates of Case 1 has coordinates $P = P(\delta h, \mu h)$, $1 \le \delta \le 3/2$, $0 < \mu \le 1/2$. On grid lines we take the additional points $P'_0 = (0, \mu h)$, $P'_1 = (2h, \mu h)$, $P'_2 = (3h, \mu h)$, $P'_3 = (2h, h + \mu h)$, $P'_4 = (h, h + \mu h)$, $P'_5 = (0, h + \mu h)$, $P'_6 = (2h, -h + \mu h)$, $P'_7 = (h, -h + \mu h)$, $P'_8 = (0, -h + \mu h)$. From the values of u at the points P'_k, $k = 0, 1, \dots, 8$, we form the expressions

$$S^6 u \equiv \sum_{k=0}^{8} \lambda_k u'_k,$$

where $u'_k = u(P'_k)$, and λ_k, $k = 0, \dots, 8$, are found from (5.8). Since all the points P'_k, $k = 0, 1, \dots, 8$, lie on grid lines, we express all the values of u'_k, $k = 0, 1, \dots, 8$, in terms of nodal values of the function u by formula (5.7) and finally obtain an expression for $S^6 u$ which gives the exact value of any harmonic polynomial of the fifth degree at the point $P(\delta h, \mu h)$:

$$S^6 u \equiv \sum_{k=0}^{30} \xi_k u_k, \tag{5.10}$$

where

$$\xi_k \ge 0, \quad \sum_{k=0}^{30} \xi_k = 1. \tag{5.11}$$

If any of the points P_k in (5.7) or (5.10) are outside the domain $\overline{G}$, there are two possibilities to consider.

Case 3. The points emerge through the side γ_τ of the boundary γ, where $\tau = j-1$ or $\tau = j$, $j \in E$. Since for $j \in E$ the function φ_τ or ψ_τ is given as an algebraic polynomial of the arclength s, the harmonic function $u - Q_j$, where Q_j is the function defined in section 2, is extendable as an odd function across γ_τ if the boundary condition (2.2) on γ_τ is the Dirichlet type and is extendable as an even function if the boundary condition is the Neumann type. Thus, an operator S^6 has been constructed, as in Cases 1 and 2, but only for the function $u - Q_j$, and $Q_j(P)$ must then be added to the expression $S^6(u - Q_j)$.

Case 4. The points emerge through the side γ_m when φ_m or ψ_m is given as a function of the class $C_{6,\lambda}(\gamma_m)$ only.

Subcase 4a. $\nu_m = 1$; i.e., $u = \varphi_m$ on γ_m and $\varphi_m \in C_{6,\lambda}(\gamma_m)$, $0 < \lambda < 1$. We position the origin of the rectangular system of coordinates on γ_m so that the point P lies on the positive y-axis, and the x-axis is in the direction of the vertex A_{m+1} along γ_m. Since

$$\sum_{k=1}^{5} b_k \operatorname{Im} z^k = 0 \quad \text{if } y = 0, \tag{5.12}$$

by representing the function $\varphi_m \in C_{6,\lambda}(\gamma_m)$ in the neighborhood of $x = 0$ using Taylor's formula, and using (5.2) for the solution of problem (2.1), (2.2) in the neighborhood $|z| \le 4h$, $z = x + iy$, of the origin, we find the coefficients a_k, $k = 0, 1, \dots, 5$, of (5.1) as

$$a_k = \frac{1}{k!} \frac{d^k \varphi_m(0)}{dx^k}.$$

We put

$$\widetilde{u}(x,y) \equiv u(x,y) - \sum_{k=0}^{5} a_k \operatorname{Re} z^k = \sum_{k=1}^{5} b_k \operatorname{Im} z^k + O(h^6)$$

for $y > 0$ and complete the definition with $\widetilde{u}(x,y) = -\widetilde{u}(x,-y)$ for $y < 0$. Obviously, in the given neighborhood $\widetilde{u}(x,y)$ is equal to the harmonic polynomial (5.12) with accuracy $O(h^6)$ by virtue of the fact that this polynomial is odd relative to the x-axis. We then form the expression for $S^6\widetilde{u}$ using (5.7) or (5.10), adding the quantity

$$\left(\sum_{k=0}^{5} a_k \operatorname{Re} z^k \right)(P).$$

Subcase 4b. $\nu_m = 0$; i.e., $u_n^{(1)} = \psi_m$ on γ_m. Let

$$\Pi_\tau = \{(x,y) : -a_{1\tau}/2 < x < a_{1\tau}/2, 0 < y < a_{2\tau}\}$$

be one of the rectangles chosen in section 4 for which $P \in \Pi_\tau$, $(\overline{\Pi}_\tau \cap \gamma) \subseteq \gamma_m$. Since $\psi_m \in C_{6,\lambda}(\gamma_m)$, the solution u of the problem (2.1), (2.2) in any rectangle

$$\Pi'_\tau = \{(x,y) : -a'_{1\tau}/2 < x < a'_{1\tau}/2, 0 < y < a'_{2\tau}\},$$

where $0 < a'_{i\tau} < a_{i\tau}$, $i = 1, 2$, and $P \in \Pi'_\tau$ is

$$u \in C_{7,\lambda}(\overline{\Pi}'_\tau). \tag{5.13}$$

Then in the neigborhood $|z| \le 4h$ of the origin by using Taylor's formula we have

$$\frac{\partial u(x,y)}{\partial y}\Big|_{y=0} = \frac{\partial u(0,0)}{\partial y} + x\frac{\partial^2 u(0,0)}{\partial x \partial y} + \frac{x^2}{2}\frac{\partial^3 u(0,0)}{\partial x^2 \partial y} + \frac{x^3}{3!}\frac{\partial^4 u(0,0)}{\partial x^3 \partial y} + \frac{x^4}{4!}\frac{\partial^5 u(0,0)}{\partial x^4 \partial y}$$
$$+ \frac{x^5}{5!}\frac{\partial^6 u(0,0)}{\partial x^5 \partial y} + O(h^6). \tag{5.14}$$

Furthermore, taking into account the boundary condition $u_n^{(1)} = \psi_m$ we have

$$\frac{\partial u(x,y)}{\partial y}\Big|_{y=0} \equiv \psi_m(x) = \psi_m(0) + x\frac{d\psi_m(0)}{dx} + \frac{x^2}{2}\frac{d^2\psi_m(0)}{dx^2} + \frac{x^3}{3!}\frac{d^3\psi_m(0)}{dx^3}$$
$$+ \frac{x^4}{4!}\frac{d^4\psi_m(0)}{dx^4} + \frac{x^5}{5!}\frac{d^5\psi_m(0)}{dx^5} + O(h^6). \tag{5.15}$$

On the basis of (5.14) and (5.15) we obtain

$$\frac{\partial u(x,y)}{\partial y} = \psi_m(0),\ \frac{\partial^2 u(0,0)}{\partial x \partial y} = \frac{d\psi_m(0)}{\partial x},\ \frac{\partial^3 u(0,0)}{\partial x^2 \partial y} = \frac{d^2\psi_m(0)}{\partial x^2},\ \frac{\partial^4 u(0,0)}{\partial x^3 \partial y} = \frac{d^3\psi_m(0)}{\partial x^3},$$

$$\frac{\partial^5 u(0,0)}{\partial x^4 \partial y} = \frac{d^4\psi_m(0)}{\partial x^4}, \quad \frac{\partial^6 u(0,0)}{\partial x^5 \partial y} = \frac{d^5\psi_m(0)}{\partial x^5}. \tag{5.16}$$

By (5.3) and (5.16), for the coefficients b_k, $k = 1, 2, \dots, 5$, in the representation (5.1), we have

$$b_k = \frac{1}{k!}\frac{d^{k-1}\psi_m(0)}{dx^{k-1}}, \quad k = 1, 2, \dots, 5.$$

We define, for $y > 0$, the function

$$\widetilde{\widetilde{u}}(x,y) \equiv u(x,y) - \sum_{k=1}^{5} b_k \operatorname{Im}\ z^k = \sum_{k=0}^{5} a_k \operatorname{Re}\ z^k + O(h^6)$$

and complete the definition for $y < 0$ with $\widetilde{\widetilde{u}}(x,y) = \widetilde{\widetilde{u}}(x,-y)$. It is evident that in the neighborhood $|z| \le 4h$ the function $\widetilde{\widetilde{u}}(x,y)$ coincides with the harmonic polynomial

$$\sum_{k=0}^{5} a_k \operatorname{Re} z^k$$

with an accuracy $O(h^6)$ by virtue of the fact that this polynomial is even relative to the x-axis. We then form the expression for $S^6\widetilde{\widetilde{u}}$ using (5.7) or (5.10), adding the quantity

$$\left(\sum_{k=1}^{5} b_k \operatorname{Im}\ z^k\right)(P).$$

Remark 5.1. In Cases 1 and 2, it is assumed that all points P_k as well as P are in $\overline{G}_*^{h,n}$. We denote the set of such points P by ω_1^h. Similarly, we denote the set of P in Cases 3 and 4 by ω_2^h and ω_3^h, respectively. Furthermore, let ω_{3a}^h and ω_{3b}^h be the subsets of ω_3^h, which are defined in Subcases 4a and 4b, respectively. It is obvious that $\omega_3^h = \omega_{3a}^h \cup \omega_{3b}^h$ and $\omega_{3a}^h \cap \omega_{3b}^h = \emptyset$. Then the matching operator S^6 can be expressed as follows:

$$S^6(u,\varphi,\psi) = \begin{cases} S^6 u & \text{on} \quad \omega_1^h, \\ S^6(u - Q_j) + Q_j(P) & \text{on} \quad \omega_2^h, \\ S^6\left(u - \sum\limits_{k=0}^{5} a_k \operatorname{Re}\ z^k\right) + \left(\sum\limits_{k=0}^{5} a_k \operatorname{Re}\ z^k\right)(P) & \text{on} \quad \omega_{3a}^h, \\ S^6\left(u - \sum\limits_{k=1}^{5} b_k \operatorname{Im}\ z^k\right) + \left(\sum\limits_{k=1}^{5} b_k \operatorname{Im}\ z^k\right)(P) & \text{on} \quad \omega_{3b}^h. \end{cases}$$

6. Analysis of the block-grid equations.

THEOREM 6.1. *There is a natural number n_0 such that for all $n \geq n_0$ the system of equations* (4.5)–(4.8) *has a unique solution.*

Proof. The proof is obtained on the basis of (4.3), (4.4), (5.7)–(5.11), and Remark 5.1 by analogy with [8]. □

Let

$$\varepsilon_h = u_h - u, \tag{6.1}$$

where u_h is a solution of system (4.5)–(4.8), and u is the trace on $\overline{G}_*^{h,n}$ of the solution of (2.1), (2.2). On the basis of (2.1), (2.2), (4.5)–(4.8), and (6.1) the error ε_h satisfies the system of difference equations

$$\begin{aligned} \varepsilon_h &= B\varepsilon_h + r_h^1 \ \text{ on } \Pi_k^h, \\ \varepsilon_h &= \overline{\nu}_m B_m \varepsilon_h + r_h^2 \ \text{ on } \eta_{k1}^h \cap \gamma_m, \\ \varepsilon_h &= \overline{\nu}_m \overline{\nu}_{m+1} \dot{B}_m \varepsilon_h + r_h^3 \ \text{ on } \eta_{k1}^h \cap \gamma_m \cap \gamma_{m+1}, \\ \varepsilon_h(r_j,\theta_j) &= \beta_j \sum_{q=1}^{n(j)} R_j^{(q)}(r_j,\theta_j)\varepsilon_h(r_{j2},\theta_j^q) + r_{jh}^4, (r_j,\theta_j) \in t_j^h, \\ \varepsilon_h &= S^6\varepsilon_h + r_h^5 \ \text{ on } \omega^{h,n}, \end{aligned} \tag{6.2}$$

where $1 \leq m \leq N$, $1 \leq k \leq M$, $j \in E$,

$$r_h^1 = Bu - u \ \text{ on } \bigcup_{k=1}^{M} \Pi_k^h, \tag{6.3}$$

$$r_h^2 = \overline{\nu}_m B_m u - u + E_{mh}(\varphi_m, \psi_m) \ \text{ on } \gamma_m \cap \left(\bigcup_{k=1}^{M} \eta_{k1}^h\right), \tag{6.4}$$

$$\begin{aligned} r_h^3 &= \overline{\nu}_m \overline{\nu}_{m+1} \dot{B}_m u - u + \dot{E}_{mh}(\varphi_m, \varphi_{m+1}, \psi_m, \psi_{m+1}) \\ &\text{on } \gamma_m \cap \gamma_{m+1} \cap \left(\bigcup_{k=1}^{M} \eta_{k1}^h\right), \end{aligned} \tag{6.5}$$

$$r_{jh}^4 = \beta_j \sum_{q=1}^{n(j)} R_j^{(q)}(r_j,\theta_j)(u(r_{j2},\theta_j^q) - Q_j(r_{j2},\theta_j^q))$$

$$- (u(r_j,\theta_j) - Q_j(r_j,\theta_j)) \ \text{ on } \bigcup_{j\in E} t_j^h, \tag{6.6}$$

$$r_h^5 = \begin{cases} S^6 u - u \ \text{ on } \omega_1^h, \\ S^6(u-Q_j) - (u-Q_j)(P) \ \text{ on } \omega_2^h, \\ S^6\left(u - \sum\limits_{k=0}^{5} a_k \operatorname{Re} z^k\right) - \left(u - \sum\limits_{k=0}^{5} a_k \operatorname{Re} z^k\right)(P) \ \text{ on } \omega_{3a}^h, \\ S^6\left(u - \sum\limits_{k=1}^{5} b_k \operatorname{Im} z^k\right) - \left(u - \sum\limits_{k=1}^{5} b_k \operatorname{Im} z^k\right)(P) \ \text{ on } \omega_{3b}^h. \end{cases} \tag{6.7}$$

In what follows and for simplicity, we will denote constants which are independent of h by c.

LEMMA 6.2. *There exists a natural number n_0 such that for all $n \geq \max\left\{n_0, \left[\ln^{1+\varkappa} h^{-1}\right] + 1\right\}$, $\varkappa > 0$ being a fixed number,*

$$\max_{j\in E} \left|r_{jh}^4\right| \leq ch^6. \tag{6.8}$$

Proof. On the basis of (6.6), Lemma 2.2, (4.1), Lemma 6.5 in [29], and by virtue of $t_j^h \subset \overline{T}_j^3$ and the boundedness of the difference $u(r_{j2},\theta_j^q) - Q_j(r_{j2},\theta_j^q)$, $1 \leq q \leq n(j)$, we obtain

$$\begin{aligned} \left|r_{jh}^4\right| \leq & \left| \beta_j \sum_{q=1}^{n(j)} R_j(r_j,\theta_j,\theta_j^q)(u(r_{j2},\theta_j^q) - Q_j(r_{j2},\theta_j^q)) \right. \\ & \left. - \int_0^{\alpha_j \pi} R_j(r_j,\theta_j,\eta)(u(r_{j2},\eta) - Q_j(r_{j2},\eta))d\eta \right| \\ & + \beta_j \sum_{q=1}^{n(j)} \left|R_j^{(q)}(r_j,\theta_j) - R_j(r_j,\theta_j,\theta_j^q)\right| \left|u(r_{j2},\theta_j^q) - Q_j(r_{j2},\theta_j^q)\right| \\ & \leq c_j^0 \exp\left\{-d_j^0 n\right\}, \quad j \in E, \end{aligned} \tag{6.9}$$

where c_j^0 and $d_j^0 > 0$ are constants independent of n. Putting $c^0 = \max_{j\in E}\left\{c_j^0\right\}$ and $d^0 = \min_{j\in E}\left\{d_j^0\right\}$ from (6.9) we have

$$\max_{j\in E} \left|r_{jh}^4\right| \leq c^0 \exp\left\{-d^0 n\right\}. \tag{6.10}$$

Let n_0 be a natural number to hold the inequality (4.3). Then, for all $n \geq \max\left\{n_0, \left[\ln^{1+\varkappa} h^{-1}\right] + 1\right\}$, where $\varkappa > 0$ is a fixed number, we have the inequality (6.8). □

Since the set of points $\omega^{h,n}$ located from the vertices of the polygon G at the distance exceeding some positive quantity independent of h, then by virtue of (2.4), (2.5), and estimation (4.64) in [21], from (6.7) we obtain

$$\max_{\omega^{h,n}} \left|r_h^5\right| \leq ch^6. \tag{6.11}$$

THEOREM 6.3. *Assume that conditions* (2.3)–(2.5) *hold. Then there exists a natural number* n_0 *such that for all* $n \geq \max\left\{n_0, \left[\ln^{1+\varkappa} h^{-1}\right] + 1\right\}$, $\varkappa > 0$ *being a fixed number,*

$$\max_{\overline{G}_*^{h,n}} |u_h - u| \leq ch^6.$$

Proof. Let us take an arbitrary rectangular grid $\Pi_{k^*}^h$, and let $t_{k^*j}^h = \overline{\Pi}_{k^*}^h \cap t_j^h$. Let $t_{k^*j}^h \neq \emptyset$, and let v_h be a solution of system (6.2) in the case when the discrepancies r_h^1, r_h^2, r_h^3, r_{jh}^4, and r_h^5 in $\overline{\Pi}_{k^*}^h$ are the same as in (6.3)–(6.7) but are zero in $\overline{G}_*^{h,n} \setminus \overline{\Pi}_{k^*}^h$. It is easy to show that

$$W = \max_{\overline{G}_*^{h,n}} |v_h| = \max_{\overline{\Pi}_{k^*}^h} |v_h| . \tag{6.12}$$

We represent the function v_h on $\overline{G}_*^{h,n}$ as

$$v_h = \sum_{\kappa=1}^{4} v_h^\kappa, \tag{6.13}$$

where the functions v_h^κ, $\kappa = 2, 3, 4$, are defined on $\overline{\Pi}_{k^*}^h$ as a solution of the system of equations

$$\begin{aligned}
v_h^\kappa &= Bv_h^\kappa + r^\kappa(h) \ \text{ on } \Pi_{k^*}^h, \\
v_h^\kappa &= \overline{\nu}_m B_m v_h^\kappa + r^\kappa(h) \ \text{ on } \eta_{k^*1}^h \cap \gamma_m, \\
v_h^\kappa &= \overline{\nu}_m \overline{\nu}_{m+1} \dot{B}_m v_h^\kappa + r^\kappa(h) \ \text{ on } \eta_{k^*1}^h \cap \gamma_m \cap \gamma_{m+1}, \\
v_h^\kappa(r_j, \theta) &= r_j^\kappa(h), \quad (r_j, \theta_j) \in t_{k^*j}^h, \\
v_h^\kappa &= r_5^\kappa(h) \ \text{ on } \omega^{h,n};
\end{aligned} \tag{6.14}$$

with

$$v_h^\kappa = 0, \quad \kappa = 2, 3, 4, \ \text{ on } \overline{G}_*^{h,n} \setminus \overline{\Pi}_{k^*}^h, \tag{6.15}$$

$r^\kappa(h) = 0$, $\kappa = 2, 3$; $r_j^\kappa(h) = 0$, $\kappa = 3, 4$, $r_j^2(h) = r_{jh}^4$; $r_5^\kappa(h) = 0$, $\kappa = 2, 4$, $r_5^3(h) = r_h^5$;

$$r^4(h) = \begin{cases} r_h^1 \ \text{ on } \Pi_{k^*}^h, \\ r_h^2 \ \text{ on } \eta_{k^*1}^h \cap \gamma_m, \\ r_h^3 \ \text{ on } \eta_{k^*1}^h \cap \gamma_m \cap \gamma_{m+1}. \end{cases}$$

Hence according to (6.13)–(6.15) the function v_h^1 satisfies the system of equations

$$\begin{aligned}
v_h^1 &= Bv_h^1 \ \text{ on } \Pi_k^h, \\
v_h^1 &= \overline{\nu}_m B_m v_h^1 \ \text{ on } \eta_{k1}^h \cap \gamma_m, \\
v_h^1 &= \overline{\nu}_m \overline{\nu}_{m+1} \dot{B}_m v_h^1 \ \text{ on } \eta_{k1}^h \cap \gamma_m \cap \gamma_{m+1}, \\
v_h^1(r_j, \theta_j) &= \beta_j \sum_{q=1}^{n(j)} R_j^{(q)}(r_j, \theta_j) \sum_{\kappa=1}^{4} v_h^\kappa(r_{j2}, \theta_j^q), (r_j, \theta_j) \in t_{kj}^h, \\
v_h^1 &= S^6 \left(\sum_{\kappa=1}^{4} v_h^\kappa \right) \ \text{ on } \omega^{h,n}, \quad 1 \leq m \leq N, \quad 1 \leq k \leq M, \quad j \in E,
\end{aligned} \tag{6.16}$$

where the functions v_h^κ, $\kappa = 2, 3, 4$, are assumed to be known.

Taking into account (6.8), (6.11), and (6.15), on the basis of the structure of operators B, B_m, and $\dot{B}_m$ and the principle of maximum, we have

$$W_2 = \max_{\overline{G}_*^{h,n}} \left|v_h^2\right| \le ch^6, \tag{6.17}$$

$$W_3 = \max_{\overline{G}_*^{h,n}} \left|v_h^3\right| \le ch^6. \tag{6.18}$$

The function v_h^4 being a solution of the system (6.14), when $\kappa = 4$, with (6.15) is the error function of the finite difference solution, with step $h_{k^*} \le h$, of the boundary value problem (3.1)–(3.2). By virtue of Theorem 3.1 we have

$$W_4 = \max_{\overline{G}_*^{h,n}} \left|v_h^4\right| = \max_{\overline{\Pi}_{k^*}^h} \left|v_h^4\right| \le ch^6. \tag{6.19}$$

We estimate the function v_h^1, which is, according to Theorem 6.1 , the unique solution of system (6.16). On the basis of (4.3), (4.4), (6.16), and the gluing condition of the figures Π_k, $k = 1, 2, \dots, M$, T_j^2, $j \in E$, there exists a real number λ^*, $0 < \lambda^* < 1$, independent of h, such that for all $n \ge \max\left\{n_0, \left[\ln^{1+\varkappa} h^{-1}\right] + 1\right\}$ we have

$$W_1 = \max_{\overline{G}_*^{h,n}} \left|v_h^1\right| \le \lambda^* W + \sum_{i=2}^{4} \max_{\overline{G}_*^{h,n}} \left|v_h^i\right|. \tag{6.20}$$

From (6.12), (6.13), and (6.17)–(6.20) we obtain

$$W = \lambda^* W + 2\sum_{i=2}^{4} W_i \le \lambda^* W + ch^6, \quad 0 < \lambda^* < 1,$$

i.e.,

$$W = \max_{\overline{G}_*^{h,n}} |v_h| \le ch^6. \tag{6.21}$$

In the case when $t_{k^*j}^h \equiv \emptyset$ the function $v_h^2 \equiv 0$ on $\overline{G}_*^{h,n}$ and the inequality (6.21) hold true. Since the number of grid rectangles in $\overline{G}_*^{h,n}$ is finite, for the solution of (6.2) we have

$$\max_{\overline{G}_*^{h,n}} |\varepsilon_h| \le ch^6. \qquad \square$$

Now we consider the question of convergence of function $U_h(r_j, \theta_j)$ defined by the formula (4.9). Taking into account the properties of functions $Q_j(r_j, \theta_j)$, $j \in E$, and the fact that the kernel $R_j(r_j, \theta_j, \eta)$ satisfies the homogeneous boundary condition defined by (2.2) on $(\gamma_{j-1} \cup \gamma_j) \cap \overline{T}_j^2$, the function $U_h(r_j, \theta_j)$ is bounded, harmonic on $T_j^* = T_j(r_j^*)$, $r_j^* = (r_{j2} + r_{j3})/2$, $j \in E$, and continuous up to the boundary of T_j^*, except for the vertex A_j when the specified boundary values are discontinuous at A_j. Moreover, on the rectilinear parts of the boundary of T_j^*, except, maybe, the vertex A_j, the function $U_h(r_j, \theta_j)$ satisfies the boundary conditions defined in (2.2).

THEOREM 6.4. *There is a natural number n_0 such that for all $n \ge \max\left\{n_0, \left[\ln^{1+\varkappa} h^{-1}\right] + 1\right\}$, $\varkappa > 0$ being a fixed number, the following inequalities are valid:*

$$\left|\frac{\partial^p}{\partial x^{p-q} \partial y^q}\left(U_h(r_j, \theta_j) - u(r_j, \theta_j)\right)\right| \le c_p h^6 \quad \text{on } \overline{T}_j^3, \tag{6.22}$$

first, for integer λ_j *and any* ν_{j-1} *and* ν_j *when* $p \geq \lambda_j$ *and, second, for* $\nu_{j-1} = \nu_j = 0$ *and any* λ_j *when* $p = 0$;

$$\left| \frac{\partial^p}{\partial x^{p-q} \partial y^q} (U_h(r_j, \theta_j) - u(r_j, \theta_j)) \right| \leq c_p h^6 / r_j^{p-\lambda_j} \quad \text{on } \overline{T}_j^3, \tag{6.23}$$

for any λ_j, *if* $\nu_{j-1} + \nu_j \geq 1$, $0 \leq p < \lambda_j$ *or* $\nu_{j-1} = \nu_j = 0$, $1 \leq p < \lambda_j$;

$$\left| \frac{\partial^p}{\partial x^{p-q} \partial y^q} (U_h(r_j, \theta_j) - u(r_j, \theta_j)) \right| \leq c_p h^6 / r_j^{p-\lambda_j} \quad \text{on } \overline{T}_j^3 \backslash A_j \tag{6.24}$$

for noninteger λ_j *and any* ν_{j-1} *and* ν_j *when* $p > \lambda_j$. *Everywhere,* $0 \leq q \leq p$, λ_j *is the quantity* (2.11), ν_{j-1} *and* ν_j *are parameters entering into the boundary conditions* (2.2), u *is a solution of the problem* (2.1), (2.2), $c_p, p = 0, 1, \dots$, *are constants independent of* $r_j = r_j(x, y)$, $\theta_j = \theta_j(x, y)$, *and* h.

Proof. On the bases of (4.9) and Lemma 2.2, on the closed block $\overline{T}_j^*$, $j \in E$, we have

$$\begin{aligned} U_h(r_j, \theta_j) - u(r_j, \theta_j) = {} & \beta_j \sum_{q=1}^{n(j)} R_j(r_j, \theta_j, \theta_j^q)(u(r_{j2}, \theta_j^q) - Q_j(r_j, \theta_j^q)) \\ & - \int_0^{\alpha_j \pi} R_j(r_j, \theta_j, \eta)(u(r_{j2}, \eta) - Q_j(r_{j2}, \eta)) d\eta \\ & + \beta_j \sum_{q=1}^{n(j)} R_j(r_j, \theta_j, \theta_j^q)(u_h(r_{j2}, \theta_j^q) - u(r_j, \theta_j^q)). \end{aligned} \tag{6.25}$$

Since $r_j^* = (r_{j2} + r_{j3})/2$ by analogy with the proof of Lemma 6.2 for $n \geq \left[\ln^{1+\varkappa} h^{-1}\right] + 1$, $\varkappa > 0$ being a fixed number, we have

$$\begin{aligned} & \left| \beta_j \sum_{q=1}^{n(j)} R_j(r_j, \theta_j, \theta_j^q)(u(r_{j2}, \theta_j^q) - Q_j(r_j, \theta_j^q)) \right. \\ & \left. - \int_0^{\alpha_j \pi} R_j(r_j, \theta_j, \eta)(u(r_{j2}, \eta) - Q_j(r_{j2}, \eta)) d\eta \right| \\ & \leq c h^6 \quad \text{on } \overline{T}_j^*, \quad j \in E. \end{aligned} \tag{6.26}$$

On the basis of Theorem 6.3 and the boundedness of $\beta_j \sum_{q=1}^{n(j)} R_j(r_j, \theta_j, \theta_j^q)$ for all $n \geq \max\left\{n_0, \left[\ln^{1+\varkappa} h^{-1}\right] + 1\right\}$ we obtain

$$\left| \beta_j \sum_{q=1}^{n(j)} R_j(r_j, \theta_j, \theta_j^q)(u_h(r_{j2}, \theta_j^q) - u(r_j, \theta_j^q)) \right| \leq c h^6 \quad \text{on } \overline{T}_j^*, \quad j \in E. \tag{6.27}$$

From (6.25)–(6.27) for all $n \geq \max\left\{n_0, \left[\ln^{1+\varkappa} h^{-1}\right] + 1\right\}$ we have

$$|U_h(r_j, \theta_j) - u(r_j, \theta_j)| \leq c h^6 \quad \text{on } \overline{T}_j^*, \quad j \in E. \tag{6.28}$$

Since $T_j^3 \subset \overline{T}_j^*$, $j \in E$, then from the inequality (6.28) the proof of (6.22) follows when $p = 0$.

To establish the validity of remainder inequalities of Theorem 6.4 we put

$$\varepsilon_h(r_j,\theta_j) = U_h(r_j,\theta_j) - u(r_j,\theta_j) \ \text{ on } \overline{T}_j^*, \quad j \in E. \tag{6.29}$$

From (4.9), (6.29), and Remark 2.1 it follows that the function $\varepsilon_h(r_j,\theta_j)$ is continuous on $\overline{T}_j^*$ and is a solution of the boundary value problem

$$\begin{aligned} \Delta\varepsilon_h &= 0 \ \text{ on } T_j^*, \\ \nu_m\varepsilon_h + \overline{\nu}_m(\varepsilon_h)'_n &= 0 \ \text{ on } \gamma_m \cap \overline{T}_j^*, \quad m = j-1, j, \\ \varepsilon_h(r_j^*,\theta_j) &= U_h(r_j^*,\theta_j) - u(r_j^*,\theta_j), \quad 0 \le \theta_j \le \alpha_j\pi. \end{aligned} \tag{6.30}$$

Taking into account (6.28)–(6.30), from Lemma 6.12 given by Volkov [29] follows all remainder inequalities of Theorem 6.4. □

Remark 6.5. From the error estimation formula (6.23) of Theorem 6.4 it follows that, when on the sides of interior angles $\neq \pi/2$, the boundary conditions are either Dirichlet or mixed type, then the error of approximate solution on the block sectors decreases as $r_j^{\lambda_j} h^6$, which gives an additional accuracy of the BGM near the singular points, with respect to existing FDM or FEM modifications for the singular problems.

7. The use of Schwarz's alternating method to solve the system of block-grid equations. According to Definitions 4.1–4.3, the approximate solution of problem (2.1), (2.2) must first be found in the domain $\overline{G}_*^{h,n}$ as the solution of the system of difference equations (4.5)–(4.8), and the solution itself and its derivatives of order p, $p = 1, 2, \dots$, at any point of $\overline{T}_j^3$, $j \in E$, except maybe the vertex A_j, can then be found using formula (4.9). Therefore, it is sufficient to justify the possibility of finding a solution of system (4.5)–(4.8) by Schwarz's alternating method.

We denote by γ^D the union of all sides of polygon G on which the boundary condition is Dirichlet type, i.e.,

$$\gamma^D = \bigcup_{j:\nu_j=1} \gamma_j.$$

From (2.3) it follows that $\gamma^D \neq \emptyset$. We define the following classes: Φ_τ, $\tau = 1, 2, \dots, \tau^*$, of rectangles Π_k, $k = 1, 2, \dots, M$ (see [8]). Class Φ_1 includes all rectangles whose intersection with γ^D contains a certain segment of positive length. Class Φ_2 contains all the rectangles which are not in class Φ_1 and whose intersection with rectangles of Φ_1 contains a segment of finite length, and so on. Let Π_{k0}^h be the set of nodes of the grid $\overline{\Pi}_k^h$ which are not less than $l_0 = \min\{\min_{1\le k\le M} \min\{a_{1k}, a_{2k}\}, \kappa_0\}/8$ from the set η_{k0}. Let

$$\Phi_{\tau 0}^h = \bigcup_{k:\Pi_k\in\Phi_\tau} \Pi_{k0}^h, \quad \tau = 1, 2, \dots, \tau^*; \quad G_{*0}^h = \bigcup_{\tau=1}^{\tau^*} \Phi_{\tau 0}^h.$$

Suppose we have a zero approximation $u_h^{(0)}$ to the exact solution u_h of (4.5)–(4.8). Finding $u_h^{(1)}$ on η_{k0} by the formula (4.8) and for all $j \in E$ on t_j^h by (4.7), we solve the system (4.5), (4.6) on each grid $\overline{\Pi}_k^h$ of rectangles, first from class Φ_1, then from class Φ_2, and so on. The next iteration is similar.

Consequently, we have the sequence $u_h^{(1)}, u_h^{(2)}, \dots$, defined as follows:

$$
\begin{aligned}
u_h^{(m)} &= S^6\left(u_h^{(m-1)}, \varphi, \psi\right) \quad \text{on } \omega^{h,n},\\
u_h^{(m)}(r_j,\theta_j) &= Q_j(r_j,\theta_j)\\
&\quad +\beta_j \sum_{q=1}^{n(j)} R_j^{(q)}(r_j,\theta_j)(u^{(m)}(r_{j2},\theta_j^q) - Q_j(r_{j2},\theta_j^q)) \quad \text{on } t_j^h, \qquad (7.1)\\
u_h^{(m)} &= Bu_h^{(m)} \quad \text{on } \Pi_k^h,\\
u_h^{(m)} &= \bar\nu_p B_p u_h^{(m)} + E_{ph}(\varphi,\psi) \quad \text{on } \eta_{k1}^h \cap \gamma_p,\\
u_h^{(m)} &= \bar\nu_p\bar\nu_{p+1}\dot B_p u_h^{(m)} + \dot E_{ph}(\varphi_p,\varphi_{p+1},\psi_p,\psi_{p+1}) \quad \text{on } \eta_{k1}^h \cap \gamma_p \cap \gamma_{p+1},
\end{aligned}
$$

where $1 \le k \le M$, $1 \le p \le N$, $j \in E$, $m = 1, 2, \dots$.

THEOREM 7.1. *For any $n \ge \max\left\{n_0, \left[\ln^{1+\varkappa} h^{-1}\right] + 1\right\}$ the system* (4.5)–(4.8) *can be solved by Schwarz's alternating method with any accuracy $\varepsilon > 0$ in a uniform metric with the number of iterations $O(\ln \varepsilon^{-1})$, independent of h and n, where n_0 and $\varkappa$ mean the same as in Theorem* 6.4.

Proof. Let

$$
\varepsilon^{(m)} = u_h^{(m)} - u_h \quad \text{on } \overline{G}_*^{h,n}, \qquad (7.2)
$$

where u_h is the exact solution of (4.5)–(4.8), and $u_h^{(m)}$ is the mth iteration defined by (7.1), $m = 1, 2, \dots$.

By virtue of (4.5)–(4.8), (7.1), (7.2), and Remark 5.1, for any m, we have

$$
\varepsilon_h^{(m)} = S^6\left(\varepsilon_h^{(m-1)}\right) \quad \text{on } \omega^{h,n}, \qquad (7.3)
$$

$$
\varepsilon_h^{(m)}(r_j,\theta_j) = \beta_j \sum_{q=1}^{n(j)} R_j^{(q)}(r_j,\theta_j)\varepsilon_h^{(m)}(r_{j2},\theta_j^q) \quad \text{on } t_j^h \cap \eta_k, \qquad (7.4)
$$

$$
\varepsilon_h^{(m)} = B\varepsilon_h^{(m)} \quad \text{on } \Pi_k^h, \qquad (7.5)
$$

$$
\varepsilon_h^{(m)} = \bar\nu_p B_p \varepsilon_h^{(m)} \quad \text{on } \eta_{k1}^h \cap \gamma_p, \qquad (7.6)
$$

$$
\varepsilon_h^{(m)} = \bar\nu_p\bar\nu_{p+1}\dot B_p \varepsilon_h^{(m)} \quad \text{on } \eta_{k1}^h \cap \gamma_p \cap \gamma_{p+1}, \qquad (7.7)
$$

where $1 \le k \le M$, $1 \le p \le N$, $j \in E$, $m = 1, 2, \dots$.

We denote

$$
W_h^{(m)} = \max_{\overline{G}_*^{h,n}} \left|\varepsilon_h^{(m)}\right|.
$$

On the basis of (4.3), (4.4), (5.9), (5.11), (7.3), and (7.4) for $n \ge n_0$ we have

$$
\max_{\cup_{k=1}^M \eta_{k,0}^h} \left|\varepsilon_h^{(m)}\right| \le \max_{G_{*0}^h} \left|\varepsilon_h^{(m-1)}\right| \le W_h^{(m-1)}, \qquad (7.8)
$$

$$
\max_{\cup_{j\in E} t_j^h} \left|\varepsilon_h^{(m)}(r_j,\theta_j)\right| \le \max_{G_{*0}^h} \left|\varepsilon_h^{(m-1)}\right| \le W_h^{(m-1)}, \quad m = 1, 2, \dots. \qquad (7.9)
$$

By virtue of (7.8), (7.9), and the maximum principle, from (7.5)–(7.7) we obtain

$$
W_h^{(0)} \ge W_h^{(1)} \ge W_h^{(2)} \ge \cdots, \qquad (7.10)
$$

$$
W_h^{(m)} \le \max_{G_{*0}^h} \left|\varepsilon_h^{(m-1)}\right|, \quad m = 1, 2, \dots. \qquad (7.11)
$$

Taking into account the sequence of calculation over classes Φ_τ, $\tau = 1, 2, \dots, \tau^*$, and the inequalities (7.10) and (7.11), by means of [8], we obtain

$$W_h^{(m+1)} \le \mu_{\tau^*}^m W_h^{(0)}, \quad m = 1, 2, \dots, \tag{7.12}$$

where $\mu_{\tau^*} < 1$ is independent of m and h.

From (7.12) follows the statement of Theorem 7.1. □

8. Numerical examples. We computed two numerical examples in order to test the effectiveness of the BGM. In Example 8.1, the polygon G is L-shaped (Figure 2), and the exact solution is known: it has both the boundary (discontinuity of the boundary functions) and the angle singularities at vertex A_1 of the interior angle ($\alpha_1\pi = 3\pi/2$). In Example 8.2 (Motz problem), the solution has singularities at the vertex A_1, with the interior angle $\alpha_1\pi = \pi$ (Figure 3), caused by abrupt changes in the type of boundary conditions. The exact solution of the Motz problem is unknown, and comparisons are made with the existing best results.

Let us describe the method of realization of Schwarz's iterations defined by (7.1): Let $\Pi = \{(x, y) : 0 < x < a, 0 < y < b\}$, where $a = 2^p h_0$, $b = 2^q h_0$, $h_0 > 0$ is a fixed number, and p and q are integers. We introduce a square grid with the lines $x = ih$, $y = jh$, $h = 2^{-m}h_0$, $m \ge 0$ being an integer, $i = 0, 1, \dots, 2^{p+m}$, $j = 0, 1, \dots, 2^{q+m}$. Let $\Pi_h = \{(x, y) : x = x_i = ih, 0 < i < 2^{p+m}, y = y_j = jh, 0 < j < 2^{q+m}\}$, Γ_h be a set of nodes on Γ (the boundary of Π), $\Gamma_{1h} = \{(x, y) : x = ih, 1 \le i \le 2^{p+m}, y = 0\}$.

We consider the finite difference problem

$$u_h = Bu_h \quad \text{on } \Pi_h, \tag{8.1}$$

$$u_h = \begin{cases} \varphi_h & \text{on } \Gamma_{1h}, \\ 0 & \text{on } \Gamma_h \backslash \Gamma_{1h}, \end{cases} \tag{8.2}$$

where φ_h is a given function on Γ_{1h}.

The solution of (8.1), (8.2) has the representation (see [31], [18])

$$u_h(x, y) = \sum_{n=1}^{2^{p+m}-1} b_n \frac{\sinh(\beta_n(1 - y/b))}{\sinh \beta_n} \sin \frac{n\pi x}{a}, \tag{8.3}$$

where

$$b_n = 2^{1-p-m} \sum_{k=1}^{2^{p+m}-1} \varphi_h(kh) \sin \frac{n\pi kh}{a}, \tag{8.4}$$

$$\beta_n = \frac{2b}{h} \sinh^{-1} \left(\frac{\sin \frac{n\pi h}{2a}}{\sqrt{1 - 2\sin^2(n\pi h/2a)/3}} \right). \tag{8.5}$$

If the boundary condition (8.2) is nonhomogeneous on the whole boundary Γ_h, then we subtract a second order algebraic polynomial which is a solution of (8.1) and with the given values of the boundary function at vertices of Π and then subdivide the given problem into four problems of the type (8.1), (8.2).

In Example 8.1, the four overlapping rectangles and in Example 8.2 three overlapping rectangles are taken, as shown in Figures 2 and 3, respectively. Furthermore,

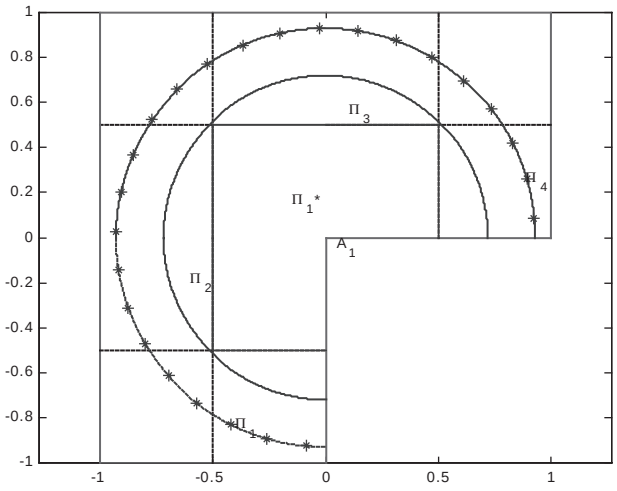


FIG. 2.

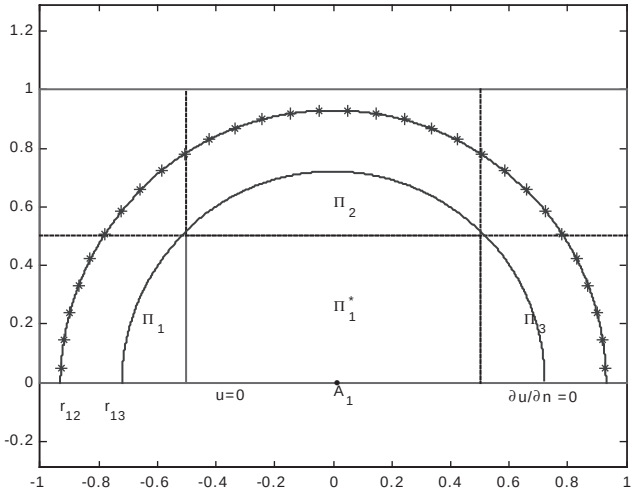


FIG. 3.

according to the boundary conditions on γ_0 and γ_1, $Q_1(r,\theta) = \theta$ is taken for Example 8.1, and $Q_1(r,\theta) = 0$ for Example 8.2. The results of realization of the iteration (7.1) by the formulas (8.3)–(8.5) for the solution of the Dirichlet problem in Example 8.1 are given in Tables 1 and 2. We use the same approach for the problem in Example 8.2 (Motz problem) after reducing the problem for each rectangle Π_k, $k = 1, 2, 3$ (see Figure 3), to the Dirichlet problem by extending the solution through the sides of Π_k as an even function when the homogeneous Neumann condition is given. Using the formulas (8.3)–(8.5) in (7.1) is effective, because the method of discrete fast Fourier transform is applicable for their realization. Some results for the solution of the Motz problem are given in Tables 3 and 4. In both problems, we request the maximum successive error on the sides of overlapping rectangles on G to be reduced by a factor of 10^{-14} as a convergence test for the Schwarz procedure, and all the computations are carried out in double precision. In all computations, we have used the starting

TABLE 1
$r_{12} = 0.87$.

(h^{-1}, n)	$\|\varepsilon_h\|_{C(G_{NS}^{h,n})}$	$\|\varepsilon_h\|_{C(G_S)}$	$\|\varepsilon_h\|_{C(G_S^{0.5})}$	$\|\varepsilon_h\|_{C(G_S^{0.125})}$
$(8, 50)$	$1.42D-6$	$2.17D-6$	$1.39D-7$	$2.82D-8$
$(8, 80)$	$1.40D-6$	$2.24D-8$	$1.85D-8$	$6.28D-9$
$(16, 60)$	$5.85D-8$	$1.32D-7$	$9.61D-9$	$2.88D-9$
$(16, 80)$	$2.21D-8$	$7.42D-10$	$3.68D-10$	$3.08D-11$
$(32, 60)$	$5.21D-8$	$1.25D-7$	$5.72D-9$	$1.84D-9$
$(32, 80)$	$3.66D-10$	$4.97D-10$	$2.00D-11$	$4.97D-12$

(h^{-1}, n)	$\|\varepsilon_h^{(1)}\|_{C(G_S)}$	$\|\varepsilon_h^{(1)}\|_{C(G_S^{0.125})}$	$\|\varepsilon_h^{(2)}\|_{C(G_S)}$	$\|\varepsilon_h^{(2)}\|_{C(G_S^{0.125})}$	*Iter.*
$(8, 80)$	$3.25D-7$	$1.53D-8$	$4.33D-6$	$5.21D-9$	31
$(16, 100)$	$4.41D-9$	$7.82D-11$	$7.50D-8$	$3.30D-10$	36
$(32, 60)$	$8.23D-6$	$5.30D-9$	$9.85D-5$	$7.67D-9$	38
$(32, 80)$	$4.35D-8$	$5.83D-12$	$2.51D-6$	$2.66D-11$	35
$(32, 120)$	$8.55D-11$	$7.02D-13$	$4.77D-10$	$1.79D-12$	37

TABLE 2
$r_{12} = 0.93$.

(h^{-1}, n)	$\|\varepsilon_h\|_{C(G_{NS}^{h,n})}$	$\|\varepsilon_h\|_{C(G_S)}$	$\|\varepsilon_h^{(1)}\|_{C(G_S)}$	$\|\varepsilon_h^{(2)}\|_{C(G_S)}$	*Iter.*
$(8, 35)$	$2.39D-6$	$5.89D-6$	$2.25D-4$	$5.54D-3$	25
$(8, 50)$	$1.41D-6$	$9.05D-8$	$1.32D-6$	$4.68D-5$	25
$(8, 60)$	$1.40D-6$	$9.96D-9$	$5.67D-8$	$1.22D-6$	26
$(16, 60)$	$2.17D-8$	$5.00D-10$	$4.07D-8$	$7.11D-7$	27
$(16, 80)$	$2.20D-8$	$2.43D-10$	$1.88D-9$	$2.77D-8$	27
$(32, 60)$	$3.65D-10$	$6.18D-10$	$4.17D-8$	$6.94D-7$	29
$(32, 80)$	$3.59D-10$	$1.05D-11$	$3.89D-11$	$2.49D-9$	30

value $u_h^{(0)} = 0$ for the Dirichlet problem on the L-shaped domain and $u_h^{(0)} = 30$ for the Motz problem. Furthermore, for both problems the radius r_{13} of sector T_1^3 is taken 0.72.

In all tables the following notation is used:

$\Pi_1^* = \overline{G} \backslash (\cup_{k=1}^{M} \overline{\Pi}_k)$, $G_{NS} = \overline{G} \backslash \Pi_1^*$ "nonsingular part" of G; $G_S = \overline{G} \cap \Pi_1^*$ "singular part" of G, $G_S^{\varrho} = G_S \cap \{r \leq \varrho\}$, $G_{NS}^{h,n} = G_{NS} \cap \overline{G}_*^{h,n}$, $\|w\|_{C(\Omega)} = \max_{\Omega} |w|$.

EXAMPLE 8.1. *Let G be L-shaped and defined as follows (see Figure 2):*

$$G = \{(x, y) : -1 < x < 1, -1 < y < 1\} \setminus G_1,$$

where $G_1 = \{(x, y) : 0 \leq x \leq 1, -1 \leq y \leq 0\}$. Let γ be the boundary of G. Consider the following problem:

$$\Delta u = 0 \quad on \quad G, \tag{8.6}$$

$$u = v(r, \theta) \quad on \quad \gamma, \tag{8.7}$$

where

$$v(r, \theta) = \theta + r^{\frac{2}{3}} \sin \frac{2\theta}{3} \tag{8.8}$$

is the exact solution of this problem.

In Tables 1 and 2, for the errors $\varepsilon_h = U_h - u$, $\varepsilon_h^{(1)} = r^{\frac{1}{3}}(\frac{\partial U_h}{\partial x} - \frac{\partial u}{\partial x})$, $\varepsilon_h^{(2)} = r^{\frac{4}{3}}(\frac{\partial^2 U_h}{\partial x^2} - \frac{\partial^2 u}{\partial x^2})$ in maximum norm between the block-grid solution U_h and the exact solution u of the problem in Example 8.1 are given.

TABLE 3
$r_{12} = 0.87$.

(h^{-1}, n)	$\|\varepsilon_h\|_{C(G_{NS}^{h,n})}$	$\|\varepsilon_h\|_{C(G_S)}$	$\|\varepsilon_h\|_{C(G_S^{0.5})}$	*Cond.*	*Iter.*
$(8, 40)$	$5.68D-5$	$5.57D-5$	$3.70D-6$	116	71
$(16, 50)$	$8.15D-7$	$9.58D-7$	$1.73D-7$	424	87
$(16, 60)$	$7.96D-7$	$1.14D-7$	$9.15D-8$	429	79
$(32, 60)$	$2.31D-8$	$4.25D-7$	$2.48D-9$	1630	80
$(32, 80)$	$2.10D-8$	$3.71D-9$	$1.06D-9$	1634	82

(h^{-1}, n)	$\|\varepsilon_h^{(1)}\|_{C(G_S)}$	$\|\varepsilon_h^{(1)}\|_{C(G_S^{0.5})}$	$\|\varepsilon_h^{(2)}\|_{C(G_S)}$	$\|\varepsilon_h^{(2)}\|_{C(G_S^{0.5})}$
$(16, 60)$	$1.17D-6$	$1.31D-7$	$2.00D-5$	$4.11D-7$
$(16, 80)$	$2.31D-7$	$1.40D-7$	$1.52D-6$	$4.46D-7$
$(32, 80)$	$7.24D-8$	$4.72D-9$	$1.32D-5$	$1.05D-8$
$(32, 100)$	$5.00D-9$	$3.40D-9$	$3.64D-8$	$7.33D-9$

TABLE 4
$r_{12} = 0.93$.

(h^{-1}, n)	$\|\varepsilon_h\|_{C(G_{NS}^{h,n})}$	$\|\varepsilon_h\|_{C(G_S)}$	$\|\varepsilon_h^{(1)}\|_{C(G_S)}$	$\|\varepsilon_h^{(2)}\|_{C(G_S)}$	*Cond.*	*Iter.*
$(8, 40)$	$5.12D-5$	$3.24D-6$	$1.18D-5$	$2.40D-4$	114	62
$(8, 50)$	$5.13D-5$	$3.03D-6$	$8.32D-6$	$4.75D-5$	117	64
$(16, 40)$	$8.35D-7$	$1.92D-7$	$1.38D-5$	$2.40D-4$	409	72
$(16, 60)$	$8.15D-7$	$6.89D-8$	$1.41D-7$	$8.41D-7$	419	66
$(32, 60)$	$2.18D-8$	$1.96D-9$	$4.03D-9$	$3.42D-8$	1583	73
$(32, 80)$	$2.10D-8$	$3.39D-9$	$3.91D-9$	$2.13D-8$	1593	73
$(32, 120)$	$2.22D-8$	$1.13D-9$	$3.05D-9$	$1.38D-8$		59

EXAMPLE 8.2. *(Motz problem). Let* $G = \{(x, y) : -1 < x < 1, 0 < y < 1\}$, *and let* γ *be its boundary (Figure* 3*). We consider the following problem:*

$$\begin{aligned} -\Delta u &= 0 \quad \textit{in } G, \\ u &= 0 \quad \textit{on } y = 0, \quad -1 \le x \le 0, \\ u &= 500 \quad \textit{on } x = 1, \\ \frac{\partial u}{\partial n} &= 0 \quad \textit{on the other boundary segments of } \gamma. \end{aligned}$$

The Motz problem is used [5], [13], [15], [16], [17], [19], [30], [32],[33] as a benchmark in many approaches for the singular problems. An extremely accurate result is obtained in [13], where piecewise expansions into particular solutions are used to approximate the boundary conditions in a least-square sense. To obtain high accurate results (for instance, the maximum error on $x = 1$ is $5.47E-9$) by this boundary method, a large number (34) of particular solutions are needed, and this may result in serious difficulties due to ill conditioning of the associated least-squares matrices (the condition number is $3.97E+07$). To decrease the condition number (down to 3617) in [13], different subdivisions of the given domain into three subdomains are considered. Different numbers of particular solutions are then used for each subdomain. However, even when the best combination of these numbers is used, the accuracy is of order $E-06$ only. Lucas and Oh [15] used the MAM in the context of the *h*-*p* version of the finite element method, and the best result (maximum error is $2.22E-08$) is obtained when $p = 10$. Comparisons in [15] were made with the extremely accurate results obtained in [13].

The results given in Tables 3 and 4 are obtained for the Motz problem. The errors

TABLE 5

$\|\varepsilon_h\|_{C(G)}$ in [13]	*Cond.*	$\|\varepsilon_h\|_{C(G)}$ in BGM	*Cond.*
$7.73D-5$	731	$5.68D-5$	116
$3.25D-6$	3617	$9.58D-7$	424

$\|\varepsilon_h\|_{C(x=1)} = 5.47D-9$, $\|\frac{\partial v}{\partial n}\|_{C(\Gamma_N)} \approx 10^{-7}$ in [13]	$cond = 3.97E+7$
$\|\varepsilon_h\|_{C(G)} = 2.10D-8$ in BGM	$cond = 1593$

TABLE 6

(h^{-1}, n)	$(8,40)$	$(8,50)$	$(16,40)$	$(16,60)$	$(32,60)$	$(32,80)$	$(32,120)$
iter	55	56	61	56	73	61	57

$\varepsilon_h = U_h - v$, $\varepsilon_h^{(1)} = r^{\frac{1}{2}}(\frac{\partial U_h}{\partial x} - \frac{\partial v}{\partial x})$, $\varepsilon_h^{(2)} = r^{\frac{3}{2}}(\frac{\partial^2 U_h}{\partial x^2} - \frac{\partial^2 v}{\partial x^2})$ are defined between the block-grid solution U_h and the extremely accurate solution v ($M = 34$) from [13] after correction of the 31st coefficient (dividing by 10), discovered by Lucas and Oh [15].

The results in Tables 1–4 show that when we decrease the step size h to improve an accuracy of the approximate solution, the number of nodes n of the quadrature formula must be increased (see Theorem 6.3); the results support the theoretical rate $O(h^6)$ for all of (h^{-1}, n). For the Motz problem (the exact solution is unknown), the comparisons are made with the results of Li, Mathon, and Sermer [13], and, as shown in Tables 3 and 4, the errors in nonsingular part are not better than $2.10E-8$ when $(h^{-1}, n) = (32, 80)$, and it is not obvious which approximate solution is better.

In Tables 3 and 4, for the Motz problem some condition numbers of the matrix of the block-grid equations (4.5)–(4.8) are given. They grow slower than $O(h^{-2})$, which is true for the usual FDM. In Table 5, the condition numbers together with the corresponding absolute errors in maximum norm for the BGM and boundary method in [13] are given.

Remark 8.3. As follows from Tables 1–4, the results in the "singular part" for $r_{j2} = 0.93$ are better than for $r_{j2} = 0.87$. Moreover, as shown in Tables 1 and 3, the results get better around the singular point, which support Remark 6.5. Therefore, in applications of the BGM the difference $r_{j2} - r_{j3}$ should be taken not less than some fixed number $k_0 > 0$. To emphasize this important computational aspect, we present the following results obtained when $r_{12} = 0.75$: (i) in the case of Example 8.1 for $(h^{-1}, n) = (32, 80)$, we get $\|\varepsilon_h\|_{C(G_{NS}^{h,n})} = 2.08D-3$, $\|\varepsilon_h\|_{C(G_S)} = 4.27D-3$; (ii) in the case of Example 8.2 for $(h^{-1}, n) = (16, 60)$, we get $\|\varepsilon_h\|_{C(G_{NS}^{h,n})} = 5.34D+1$, $\|\varepsilon_h\|_{C(G_S)} = 4.74D+1$. The explanation of this divergency follows from Remark 2.3, Definition 4.2, and Theorem 6.4. According to the construction of the matching operator (see section 5, Cases 3 and 4), the increase of the difference $r_{12} - r_{j3}$ does not cause any problem.

Remark 8.4. The number of iterations can be decreased if for every Schwarz iteration a few simple subiterations (without using the first and the second equations in (7.1)) are performed to improve the results on the interior boundaries of the rectangles Π_k, $k = 1, 2, \ldots, M$ outside of $\overline{\Pi}_1^*$. The results given in Table 6 are obtained for the Motz problem in the case of $r_{12} = 0.93$, where the five simple subiterations are used. In [14] for the solution of the algebraic system of equations obtained in the combination of the Ritz–Galerkin method and FDM the iterative subtracting method is proposed. This method requires a lower number of iterations.

9. Concluding remarks. In the proposed BGM, by making an artificial boundary, the problem with singularities is reduced to a domain without singularities. Exact boundary conditions on the artificial boundary are the integral representation of the solution in which the composite midpoint rule converges exponentially. To approximate the official boundary condition and Laplace's equation on the obtained domain, the sixth order finite difference schemes are used. On the singular parts (on blocks $\overline{T}_j^3$, $j \in E$), the approximate solution is defined by the formula (4.9), itself being a harmonic function which acquires completely all singularities of the derivatives of the required solution u via the function Q_j, appearing because of the consistency conditions at A_j, $j \in E$, not being fulfilled. The angle singularities are passed on sufficiently exactly to the approximate solution via kernel R_j. Furthermore, to connect the grids and the blocks, the sixth order matching operator is constructed. These properties are a prerequisite for a high rate of convergence established by Theorems 6.3 and 6.4 and by numerical results given in section 8. As it follows from Tables 1–4, the absolute error in the singular part (in G_S) of G is smaller than in the smoother part (in G_{NS}) of G, which agrees with the results obtained in Theorems 6.3 and 6.4 (see Remark 6.5).

If on the sides of the right interior angles of polygon G the boundary functions are given also as an algebraic polynomial of s, then, without conjugate conditions (2.5), the approximate solution in a neighborhood of vertices of these angles can be defined by the formula (4.9), and derivatives of any order can be found by its simple differentiation.

A parallelism of the sides of graduated polygon G to the x- and y-axis is assumed only for simplicity of presentation.

The method and results of this paper are valid for multiply connected graduated polygons.

The sixth order matching operator constructed in section 5 can also be used to build other highly accurate combined or domain decomposition methods.

Acknowledgments. The author thanks Professor E. A. Volkov for his valuable advice and the referee whose appropriate remarks contributed to improve the presentation of the paper. The author also thanks Mrs. S. Cival for her realization of some algorithms in section 8.

REFERENCES

[1] V. B. Andreev, *The mixed problem for Laplace's mesh equation in a half-plane*, Dokl. Akad. Nauk SSSR, 234 (1977), pp. 997–1000.

[2] V. B. Andreev, *Asymptotic solution of Laplace's grid equation in a corner*, Dokl. Akad. Nauk SSSR, 244 (1979), pp. 1289–1293.

[3] I. Babuska and M. B. Rosenzweig, *A finite element scheme for domains with corner*, Numer. Math., 20 (1972), pp. 1–21.

[4] I. Babuska and H. S. Oh, *The p-version of the finite element method for domains with corners and for infinite domains*, Numer. Methods Partial Differential Equations, 6 (1990), pp. 371–392.

[5] R. E. Barnhill and J. R. Whiteman, *Error analysis of Galerkin methods for Dirichlet problems containing boundary singularities*, J. Inst. Math. Appl., 15 (1975), pp. 121–125.

[6] H. Blum and M. M. Dobrowolski, *On finite element methods for elliptic equations on domains with corners*, Computing, 28 (1982), pp. 53–63.

[7] A. A. Dosiyev, *A block-grid method for increasing accuracy in the solution of the Laplace equation on polygons*, Russian Acad. Sci. Dokl. Math., 45 (1992), pp. 396–399.

[8] A. A. Dosiyev, *A block-grid method of increased accuracy for solving Dirichlet's problem for Laplace's equation on polygons*, Comput. Math. Math. Phys., 34 (1994), pp. 591–604.

[9] A. A. Dosiyev, *A fourth order accurate composite grids method for solving Laplace's boundary value problems with singularities*, Zh. Vychisl. Mat. Mat. Fiz. 42 (2002), pp. 867–884.
[10] G. Fix, *Higher order Rayleigh-Ritz approximations*, J. Math. Mech., 18 (1969), pp. 645–658.
[11] L. Fox and R. Sankar, *Boundary singularities in linear elliptic differential equations*, J. Inst. Math. Appl., 5 (1969), pp. 340–350.
[12] I. V. Fryazinov, *Difference schemes for Laplace's equation in graduated domains*, Zh. Vychisl. Mat. i Mat. Fiz., 18 (1978), pp. 1170–1185.
[13] Z.-C. Li, R. Mathon, and P. Sermer, *Boundary methods for solving elliptic problems with singularities and interfaces*, SIAM J. Numer. Anal., 24 (1987), pp. 487–498.
[14] Z.-C. Li, *Combined Methods for Elliptic Problems with Singularities, Interfaces and Infinities*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1998.
[15] T. R. Lucas and H. S. Oh, *The method of auxiliary mapping for the finite element solutions of elliptic problems containing singularities*, J. Comput. Phys., 108 (1993), pp. 327–342.
[16] H. Motz, *The treatment of singularities of partial differential equations by relaxation methods*, Quart. Appl. Math., 4 (1946), pp. 371–377.
[17] L. G. Olson, G. C. Georgiou, and W. W. Schultz, *An efficient finite element method for treating singularities in Laplace's equation*, J. Comput. Phys., 96 (1991), pp. 391–410.
[18] S. E. Romanova, *An efficient method for the approximate solution of the Laplace difference equation on rectangular domains*, Zh. Vychisl. Mat. i Mat. Fiz., 23 (1983), pp. 660–673.
[19] J. B. Rosser and N. Papamichael, *A Power Series Solution of a Harmonic Mixed Boundary Value Problem*, MRC Technical Summary Report 1405, University of Wisconsin, Madison, WI, 1975.
[20] R. W. Thatcher, *The use of infinite grid refinement at singularities in the solution of Laplace's equation*, Numer. Math., 25 (1976), pp. 163–178.
[21] E. A. Volkov, *Differentiability properties of solutions of boundary value problems for the Laplace's equation on polygons*, Tr. Mat. Inst. Akad. Nauk SSSR, 77 (1965), pp. 113–142.
[22] E. A. Volkov, *Effective error estimates for grid method solutions of boundary-value problems for Laplace's and Poisson's equations on rectangle and certain triangles*, Tr. Mat. Inst. Akad. Nauk SSSR, 74 (1966), pp. 55–85.
[23] E. A. Volkov, *Method of composite meshes for finite and infinite domain with a piecewise-smooth boundary*, Trudy Mat. Inst. Steklov., 96 (1968), pp. 117–148.
[24] E. A. Volkov, *On the method of composite meshes for Laplace's equation on polygons*,Trudy Mat. Inst. Steklov., 140 (1976), pp. 68–102.
[25] E. A. Volkov, *A difference-analytic method of calculating the potential field on polygons*, Soviet Math. Dokl., 18 (1977), pp. 1531–1535.
[26] E. A. Volkov, *An exponentially converging method for solving Laplace's equation on polygons*, Math. USSR Sb., 37 (1980), pp. 295–325.
[27] E. A. Volkov, *An asymptotically fast approximate method of finding a solution of the difference Laplace equation on mesh segments*, Proc. Steklov Inst. Math., 173 (1987), pp. 71–92.
[28] E. A. Volkov, *Approximate solution of the Laplace equation on polygons under nonanalytic boundary conditions by the block method*, Trudy Mat. Inst. Steklov., 194 (1992), pp. 63–88 (in Russian).
[29] E. A. Volkov, *Block Method for Solving the Laplace Equation and Constructing Conformal Mappings*, CRC Press, Boca Raton, FL, 1994.
[30] R. Wait and A. R. Mitchell, *Corner singularities in elliptic problems by finite element methods*, J. Comput. Phys., 8 (1971), pp. 45–52.
[31] W. Wasow, *On the truncation error in the solution of Laplace's equation by finite differences*, J. Research Nat. Bur. Standards, 48 (1952), pp. 345–348.
[32] N. M. Wigley, *An efficient method for subtracting off singularities at corners for Laplace's equation*, J. Comput. Phys., 78 (1988), pp. 369–377.
[33] X. Wu and H. Han, *A finite-element method for Laplace- and Helmholtz-type boundary value problems with singularities*, SIAM J. Numer. Anal., 34 (1997), pp. 1037–1050.
[34] C. Zenger and H. Gietl, *Improved difference schemes for the Dirichlet problem of Poisson's equation in the neighborhood of corners*, Numer. Math., 30 (1978), pp. 315–332.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 179–208

GODUNOV-TYPE METHODS FOR CONSERVATION LAWS WITH A FLUX FUNCTION DISCONTINUOUS IN SPACE*

ADIMURTHI†, JÉRÔME JAFFRÉ‡, AND G. D. VEERAPPA GOWDA†

This paper is dedicated to the memory of Jacques-Louis Lions

Abstract. Scalar conservation laws with a flux function discontinuous in space are approximated using a Godunov-type method for which a convergence theorem is proved. The case where the flux functions at the interface intersect is emphasized. A very simple formula is given for the interface flux. A numerical comparison between the Godunov numerical flux and the upstream mobility flux is presented for two-phase flow in porous media. A consequence of the convergence theorem is an existence theorem for the solution of the scalar conservation laws under consideration. Furthermore, for regular solutions, uniqueness has been shown.

Key words. conservation laws, discontinuous coefficients, finite difference, finite volume, flow in porous media

AMS subject classifications. 35F25, 35L65, 65M06, 65M12, 76S05, 76M12, 76M20

DOI. 10.1137/S003614290139562X

1. Introduction. Let f and g be continuous functions on an interval $I \subset \mathbb{R}$, and define the flux function $F(x,u) = H(x)f(u) + (1-H(x))g(u)$, where $H(x)$ is the Heaviside function. Let $u_0 \in L^\infty(\mathbb{R}, I)$, and consider the following scalar conservation law:

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} F(x,u) = 0 \quad \text{for} \quad x \in \mathbb{R}, \quad t > 0, \\ &u(x,0) = u_0(x), \quad x \in \mathbb{R}. \end{aligned} \tag{1.1}$$

This type of problem appears, for example, in modelling two-phase flow in a porous medium [8, 13], in sedimentation problems [7, 5], and in traffic flow [24].

It is well known that after a finite time (1.1) does not in general possess a continuous solution even if u_0 is sufficiently smooth. Hence by a solution of (1.1) we mean a solution in the weak sense. That is, $u \in L^\infty_{\text{loc}}(\mathbb{R} \times \mathbb{R}_+)$ such that for all $\varphi \in C_0^\infty(\mathbb{R} \times \overline{\mathbb{R}}_+)$

$$\int_{-\infty}^{\infty} \int_0^{\infty} \left(u \frac{\partial \varphi}{\partial t} + F(x,u) \frac{\partial \varphi}{\partial x} \right) dt dx + \int_{-\infty}^{\infty} u_0(x) \varphi(x,0) dx = 0\,. \tag{1.2}$$

Denoting $u_t = \frac{\partial u}{\partial t}$, $u_x = \frac{\partial u}{\partial x}$, then u satisfies (1.2) if and only if in the weak sense u satisfies

$$\begin{aligned} u_t + g(u)_x &= 0, \quad x < 0, \quad t > 0, \\ u_t + f(u)_x &= 0, \quad x > 0, \quad t > 0, \end{aligned} \tag{1.3}$$

and, at $x = 0$, u satisfies the Rankine–Hugoniot condition; namely, for almost all t,

$$f(u^+(t)) = g(u^-(t)), \tag{1.4}$$

where $u^+(t) = \lim_{x \to 0^+} u(x,t), \quad u^-(t) = \lim_{x \to 0^-} u(x,t)\,.$

*Received by the editors September 1, 2001; accepted for publication (in revised form) July 18, 2003; published electronically January 28, 2004.
http://www.siam.org/journals/sinum/42-1/39562.html
†TIFR Centre, P. B. 1234, Bangalore 560 012, India (aditi@math.tifrbng.res.in, gowda@math.tifrbng.res.in).
‡INRIA, BP 105, 78153 Le Chesnay Cedex, France (Jerome.Jaffre@inria.fr).

Because of the discontinuity of the flux F at $x = 0$, the Kruzkov method [19] does not guarantee a weak solution of (1.1), and, even if the solution exists, it may not be unique.

When there is no discontinuity of F at $x = 0$, that is, when $f \equiv g$, the problem has been studied and well understood. In this case, existence of a weak solution was obtained by Kruzkov [19] in the class of functions satisfying the Lax–Oleinik entropy condition [21, 25]. The solution thus obtained can be represented by an L^1-contractive semigroup [18, 19]. The method adopted in this case is that of vanishing viscosity. Furthermore, finite difference schemes are constructed using a numerical flux based on exact or approximate Riemann solvers such as Lax–Friedrich, Godunov, Engquist–Osher, upstream mobility, etc.... Convergence of these schemes is based on the following properties: conservation, consistency, monotonicity, and Lipschitz continuity. Using these properties, one obtains that the finite difference schemes are TVD (total variation diminishing) and satisfy the maximum principle and a numerical entropy condition. This allows one to pass to a limit to obtain a unique weak solution satisfying the Lax–Oleinik entropy condition.

When $f \neq g$, this problem was considered from the theoretical or numerical point of view in several papers [3, 20, 8, 13, 6, 5, 15, 1, 28, 29]. In general, the solution to (1.2) is not unique. To choose a correct solution, in [8] it was suggested to choose a solution which has $|u^+(t) - u^-(t)|$ minimum, but the problem of uniqueness was left open in the case of a general Cauchy problem. Nevertheless, this led to the construction of a numerical flux which was actually the same as the one used in [3, 14]. It turns out that the solution to the Riemann problem and the flux function given in [8] and the numerical scheme given in [3, 14, 28, 29] are correct when assuming that the flux functions f and g are not intersecting, even though this was not stated explicitly. Actually, they may intersect but in such a way that no undercompressive waves are produced, which is not the case when $f' > 0,\ g' < 0$ at the intersection point. It should be noted that in [28, 29] at the intersection points derivatives of fluxes f and g have the same sign. At an intersection point, if the derivative of g is negative and that of f is positive, then the problem becomes more difficult. Later, in [6, 7, 5] the problem was studied in the general case with a source term, and it was suggested to choose a solution with a minimal variation in the x-direction. For this purpose a condition called the Γ-condition was introduced, an explicit formula was given for a solution to the Riemann problem, and uniqueness was proved. Diehl's construction allows undercompressive waves; hence it is not clear that the solution thus obtained can be represented by an L^1-contractive semigroup.

In [15] it was shown that the solution to the Riemann problem with the numerical flux built upon it in [3, 8, 13] was not correct when the flux functions f and g intersect in the undercompressive case, and a correct solution was given for this case. Independently, in [1], the authors asked themselves the following question: "What is an appropriate condition on $x = 0$ so that the solution can be represented by an L^1-contractive semigroup?" Assuming that f and g are strictly convex with superlinear growth, using the Hamilton–Jacobi theory, they constructed an explicit weak solution satisfying an explicit interface entropy condition at $x = 0$, different from the Lax–Oleinik entropy condition satisfied for $x \neq 0$. This interface entropy condition means that it does not allow the undercompressive waves. Furthermore, it was shown that this solution is unique by proving that the solution can be represented by an L^1-contractive semigroup. The solution to the Riemann problem thus obtained is actually the same as in [15], though written in a more compact form. This leads to a very simple

way of calculating the interface flux and to derive its main properties: monotonicity and Lipschitz continuity. One should also note that the solution to the Riemann problem does not satisfy the maximum principle nor is it TVD.

Finally, we mention two papers which recently appeared and also investigate the problem of a nonlinear conservation law with a discontinuous flux function [17, 16].

In this paper, we consider the general case which includes the case where the flux functions are intersecting. Using the solution to the Riemann problem obtain in [15, 1] and the corresponding numerical flux, we study the resulting finite difference scheme and prove its convergence. In section 2 we present the continuous problem, defining in particular an interface entropy condition at $x = 0$, and we state an existence and uniqueness theorem for the solution to the continuous problem. In section 3 we present a Godunov-type method to calculate this solution, and in section 4, the core of this paper, we prove convergence of this numerical scheme. This scheme is conservative and monotone but not consistent in the usual sense. Due to the nonconsistency, it does not satisfy the maximum principle. In spite of this we show that the scheme is L^∞-bounded and L^1-stable. Furthermore, using the singular mapping technique introduced by Temple [27], we show that the scheme converges pointwise to weak solutions. These weak solutions satisfy the Lax–Oleinik entropy condition for $x \neq 0$ and the interface entropy condition at $x = 0$. This ensures the uniqueness of the limit solutions.

In section 5 we study the case of two-phase flow in porous media and introduce the alternative of the upstream mobility numerical flux [2]. One-dimensional numerical experiments are presented in section 6, and a comparison is made between these two numerical fluxes.

A consequence of the convergence theorem proved in section 4 is an existence theorem for the continuous problem for a larger class of functions f and g than the one studied in [1], where they were assumed to be convex. Uniqueness is shown in the appendix by proving that the solutions to the continuous problem form an L^1-contractive semigroup.

2. The continuous problem. Let $s < S$ denote the endpoints of the interval of the definition of f and g. In the following we will assume that f and g are smooth functions with the same endpoints and each one with one global minimum, reached at θ_f and θ_g, respectively, and with no other local minimum (see Figure 1).

Hypotheses. Assume that f, g are Lipschitz continuous functions on $[s, S]$ satisfying

(H_1) $f(s) = g(s), \quad f(S) = g(S),$

(H_2) f and g have one global minimum and no other local minimum in $[s, S]$.

Denote by Lip (f) and Lip (g) the Lipschitz constants of f and g. We will need also the constant

$$M = \max\{\text{Lip}\,(f), \text{Lip}\,(g)\}.$$

In order to state an existence and uniqueness theorem for the continuous problem we need to define regular solutions and entropy conditions. Since the flux function is not continuous, there are actually two different entropy conditions, one in the interior (which is the same as the usual Lax–Oleinik entropy condition) and the other at the interface which was introduced in [1].

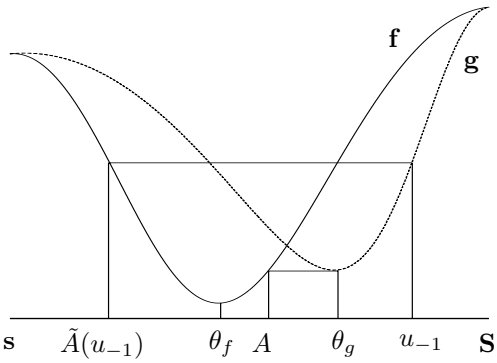


FIG. 1. *Flux functions f and g satisfying hypothesis* (H_2).

Entropy pairs. For $i = 1, 2, (\varphi_i, \psi_i)$ are said to be entropy pairs if φ_i is a convex function on $[s, S]$ and $(\psi_1'(\theta), \psi_2'(\theta)) = (\varphi_1'(\theta) f'(\theta), \varphi_2'(\theta) g'(\theta))$ for $\theta \in [s, S]$.

Let $u_0 \in L^\infty(\mathbb{R})$ be the initial data with $s \le u_0(x) \le S$ for all $x \in \mathbb{R}$, and let u be a weak solution of (1.2) with $s \le u(x,t) \le S$ for all $(x,t) \in \mathbb{R} \times \mathbb{R}_+$.

Interior entropy condition. With u_0 and u as above, u is said to satisfy an interior entropy condition if for any entropy pairs $(\varphi_i, \psi_i), i = 1, 2,$ u satisfies in the sense of distributions

$$\begin{aligned} \frac{\partial \varphi_1(u)}{\partial t} + \frac{\partial \psi_1(u)}{\partial x} &\le 0 \text{ in } x > 0, \quad t > 0, \\ \frac{\partial \varphi_2(u)}{\partial t} + \frac{\partial \psi_2(u)}{\partial x} &\le 0 \text{ in } x < 0, \quad t > 0. \end{aligned} \tag{2.1}$$

Interface entropy condition. With u_0 and u as above, assume that $u^+(t) = \lim_{x \to 0^+} u(x,t)$ and $u^-(t) = \lim_{x \to 0^-} u(x,t)$ exist for almost all $t > 0$, and define

$$L = \left\{ t > 0;\ u^+(t) \in (\theta_f, S], u^-(t) \in [s, \theta_g) \right\},$$
$$U = \left\{ t \in L;\ u^+(t) = u^-(t) = S \right\} \cup \left\{ t \in L;\ u^-(t) = u^+(t) = s \right\}.$$

Then u is said to satisfy the interface entropy condition if

$$\text{meas } \{L \setminus U\} = 0\,. \tag{2.2}$$

This means that the characteristics must connect back to the x-axis on at least one side of the jump in F; i.e., undercompressive waves are not allowed.

Regular solution. u is said to be a regular solution of (1.2) if the discontinuities of u form a discrete set of Lipschitz curves.

We need also an estimator $N(f, g, u_0)$ of the total variation of the flux function evaluated at u_0. This estimator will be defined precisely below at (3.5).

We can now state our existence and uniqueness theorem for the continuous problem.

THEOREM 2.1. *Let $u_0 \in L^\infty(\mathbb{R})$ such that $s \le u_0(x) \le S$ for all $x \in \mathbb{R}$ and $N_h(f, g, u_0) < \infty$. Then there exists a weak solution $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ of* (1.2) *satisfying the following:*

(i) *For almost all $t > 0$ and $x \in \mathbb{R}$, $u(x+, t), u(x-, t)$ exist.*

(ii) *u satisfies the interior entropy condition* (2.1).

(iii) *If u is regular, then it satisfies also the interface entropy condition* (2.2) *and it is unique. Moreover, if $f = g$, then u is the unique entropy solution for the initial value problem studied in* [19].

An existence and uniqueness theorem was proved in [1] for convex functions f and g using arguments from the Hamilton–Jacobi theory. However, functions satisfying hypotheses (H_1) and (H_2) are not necessarily convex, as shown Figure 1, and a consequence of the convergence theorem, Theorem 3.2, proved below is that existence in Theorem 2.1 is valid for such functions. Uniqueness follows by showing that the solutions form an L^1-contractive family, which is done in the appendix.

We remark that a similar analysis to what is done in this paper for f and g satisfying hypothesis (H_2) can be done for the case where f and g satisfy hypothesis (H_3) instead:

(H_3) f and g have one global maximum and no other local maximum in $[s, S]$,

as shown in Figure 2. θ_f and θ_g would denote the points at which the maxima of f and g are reached. In the analysis below only the case where f and g satisfy (H_2) will be considered.

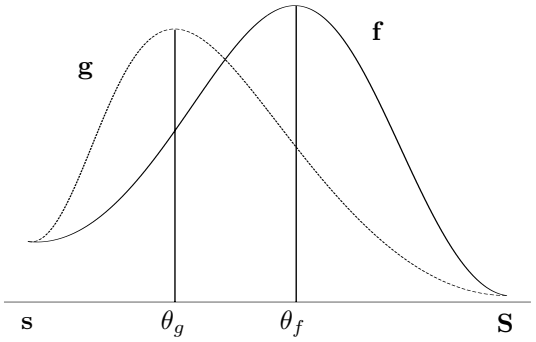


FIG. 2. *Flux functions f and g satisfying hypothesis* (H_3).

3. A Godunov-type finite volume method. Let F be the Godunov numerical flux with respect to f:

$$F(a,b) = \begin{cases} \min\limits_{\theta\in[a,b]} f(\theta) & \text{if} \quad a < b, \\ \max\limits_{\theta\in[b,a]} f(\theta) & \text{if} \quad a \geq b, \end{cases} \tag{3.1}$$

and similarly for the numerical flux G with respect to g.

Taking advantage of hypothesis (H_2), equivalent formulas can be used [1]:

$$\begin{aligned} F(a,b) &= \max\{F(a,S),\ F(s,b)\} = \max\{F(a,\theta_f),\ F(\theta_f,b)\} \\ &= \max\{f(\theta_f)(1-H(a_1)) + f(a)H(a_1), f(\theta_f)H(a_1) + f(b)(1-H(a_1))\} \\ &= \max\{f(\max\{a,\theta_f\}), f(\min\{\theta_f,b\})\}, \end{aligned}$$

where $a_1 = (a-\theta_f)$ and H is the Heaviside function. Note that the last two expressions are much simpler to use in calculations than formula (3.1).

In the case where f satisfies hypothesis (H_3) instead, the equivalent formulas are

$$\begin{aligned} F(a,b) &= \min\{F(a,s),\ F(S,b)\} = \min\{F(a,\theta_f),\ F(\theta_f,b)\} \\ &= \min\{f(\theta_f)H(a_1) + f(a)(1-H(a_1)), f(\theta_f)(1-H(a_1)) + f(b)H(a_1)\} \\ &= \min\{f(\min\{a,\theta_f\}), f(\max\{\theta_f,b\})\}. \end{aligned}$$

Interface flux $\overline{F}$. At the point $x = 0$ where the flux function changes we introduce the numerical flux $\overline{F}$ calculated by using the Riemann problem solution given in [1]:

$$\begin{aligned}\overline{F}(a,b) &= \max\{G(a,S),\ F(s,b)\} = \max\{G(a,\theta_g),\ F(\theta_f,b)\}\\ &= \max\{g(\theta_g)(1-H(a_1)) + g(a)H(a_1),\\ &\qquad\qquad f(\theta_f)H(b_1) + f(b)(1-H(b_1))\}\\ &= \max\{g(\max\{a,\theta_g\}), f(\min\{\theta_f,b\})\},\end{aligned} \tag{3.2}$$

where $a_1 = (a-\theta_g), b_1 = (b-\theta_f)$.

These four expressions of $\overline{F}$ are equivalent, but only the last two are useful for computational purposes. This flux $\overline{F}$ coincides with the one given in [15]. When f and g do not intersect this numerical flux reduces to the one given in [3, 9, 13, 6].

Remark 3.1. In the case where f and g satisfy hypothesis (H_3) the definition of the interface flux should be

$$\begin{aligned}\overline{F}(a,b) &= \min\{G(a,s),\ F(S,b)\} = \min\{G(a,\theta_g),\ F(\theta_f,b)\}\\ &= \min\{g(\theta_g)H(a_1) + g(a)(1-H(a_1)),\\ &\qquad\qquad f(\theta_f)(1-H(b_1)) + f(b)H(b_1)\}\\ &= \min\{g(\min\{a,\theta_g\}), f(\max\{\theta_f,b\})\},\end{aligned} \tag{3.3}$$

where θ_f and θ_g are now the maxima of f and g.

Let $h > 0$ and define the space grid points as follows:

$$x_{-1/2} = x_{1/2} = 0, \quad x_{j+1/2} = j\,h \quad \text{for } j \geq 0, \quad x_{j-1/2} = jh \quad \text{for } j \leq 0.$$

We will also use the midpoints of the intervals:

$$x_j = \left(\frac{2j-1}{2}\right)h \quad \text{for } j \geq 1, \quad x_j = \left(\frac{2j+1}{2}\right)h \quad \text{for } j \leq -1.$$

For time discretization the time step is $\Delta t > 0$, and let $t_n = n\Delta t,\ \lambda = \frac{\Delta t}{h}$.

For an initial data $u_0 \in L^\infty(\mathbb{R})$ we define

$$u^0_{j+1} = \frac{1}{h}\int_{x_{j+1/2}}^{x_{j+3/2}} u_0(x)dx \quad \text{if} \quad j \geq 0, \quad u^0_{j-1} = \frac{1}{h}\int_{x_{j-3/2}}^{x_{j-1/2}} u_0(x)dx \quad \text{if} \quad j \leq 0,$$

$$\begin{aligned}N_h(f,g,u_0) = &\sum_{i<-1} |G(u^0_i, u^0_{i+1}) - G(u^0_{i-1}, u^0_i)| + \sum_{i>1} |F(u^0_i, u^0_{i+1}) - F(u^0_{i-1}, u^0_i)|\\ &+|\overline{F}(u^0_{-1}, u^0_1) - G(u^0_{-2}, u^0_{-1})| + |F(u^0_1, u^0_2) - \overline{F}(u^0_{-1}, u^0_1)|,\end{aligned} \tag{3.4}$$

$$N(f,g,u_0) = \sup_{h>0} N_h(f,g,u_0). \tag{3.5}$$

It is easy to see that if $u_0 \in BV(\mathbb{R})$, then $N(f,g,u_0) \leq C||u_0||_{BV}$, where C is a constant depending only on the Lipschitz constants of f and g.

Now we can define the explicit finite volume scheme $\{u^n_i\}$ inductively as follows:

$$\begin{aligned}u^{n+1}_1 &= u^n_1 - \lambda(F(u^n_1, u^n_2) - \overline{F}(u^n_{-1}, u^n_1)),\\ u^{n+1}_i &= u^n_1 - \lambda(F(u^n_i, u^n_{i+1}) - F(u^n_{i-1}, u^n_i)) \quad \text{if } i > 1,\\ u^{n+1}_{-1} &= u^n_{-1} - \lambda(\overline{F}(u^n_{-1}, u^n_1) - G(u^n_{-2}, u^n_{-1})),\\ u^{n+1}_i &= u^n_i - \lambda(G(u^n_i, u^n_{i+1}) - G(u^n_{i-1}, u^n_i)) \quad \text{if} \quad i < -1.\end{aligned} \tag{3.6}$$

Observe that this is, a Godunov scheme for $i \neq \pm 1$, that is, away from $x = 0$, and that for $i = \pm 1$ the scheme is not consistent; that is, in general $\overline{F}(u, u)$ need not be equal to $f(u)$ or $g(u)$. Because of this, the maximum principle does not hold.

For $u_0 \in L^\infty(\mathbb{R})$ and grid length h and Δt with $\lambda = \frac{\Delta t}{h}$ fixed, define the function $u_h \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ associated with $\{u_i^n\}$ calculated by the scheme (3.6):

$$u_h(x,t) = u_i^n \quad \text{for } (x,t) \in [x_{i-1/2}, x_{i+1/2}) \times [n\Delta t,\ (n+1)\Delta t), \quad i \neq 0. \tag{3.7}$$

Now we can state the following convergence theorem.

THEOREM 3.2. *Assume that λ, M satisfies the CFL condition $\lambda M \leq 1$. Let $u_0 \in L^\infty(\mathbb{R})$ such that $s \leq u_0(x) \leq S$ for all $x \in \mathbb{R}$ and $N(f, g, u_0) < \infty$. For $h > 0$, let $\lambda = \frac{\Delta t}{h}$ and u_h be the corresponding calculated solution given by* (3.6), (3.7). *Then there exists a subsequence $h_k \to 0$ such that u_{h_k} converges a.e. to a weak solution u of* (1.2) *satisfying interior entropy condition* (2.1). *Suppose the discontinuities of every limit function u of $\{u_h\}$ is a discrete set of Lipschitz curves; then $u_h \to u$ in $L^\infty_{loc}(\mathbb{R}_+, L^1_{loc}(\mathbb{R}))$ as $h \to 0$, and u satisfies the interface entropy condition* (2.2).

The proof of this theorem is the object of the next section.

Remark 3.3. The CFL condition still reads $\lambda M \leq 1$ in the case of a discontinuous flux function.

4. Proof of the convergence theorem, Theorem 3.2.

4.1. Properties of the numerical flux. Before going into the details of the proof, we need to study the properties of the numerical flux F, G, and $\overline{F}$.

From definitions (3.1), (3.2), $F, G, \overline{F}$, are nondecreasing functions in the first variable and nonincreasing functions in the second variable. Furthermore, the functions F, G, and $\overline{F}$ satisfy for any $a, a_1, a_2, b, b_1, b_2 \in [s, S]$

$$\begin{array}{rcl} (|F(a_1,b) - F(a_2,b)|,\ |F(a,b_1) - F(a,b_2)|) & \leq & M(|a_1 - a_2|,\ |b_1 - b_2|), \\ (|G(a_1,b) - G(a_2,b)|,\ |G(a,b_1) - G(a,b_2)|) & \leq & M(|a_1 - a_2|,\ |b_1 - b_2|), \\ (|\overline{F}(a_1,b) - \overline{F}(a_2,b)|,\ |\overline{F}(a,b_1) - \overline{F}(a,b_2)|) & \leq & M(|a_1 - a_2|,\ |b_1 - b_2|). \end{array} \tag{4.1}$$

The following lemma is easy to prove.

LEMMA 4.1. *Let f and g satisfy* (H_1) *and* (H_2). *Then $\overline{F}$ satisfies*

$$\begin{array}{rcl} \overline{F}(s,s) & = & f(s) = g(s), \quad \overline{F}(S,S) = f(S) = g(S), \\ \overline{F}(a,b) & = & F(a,b) \quad \text{if} \quad f \equiv g\,. \end{array}$$

Now we define for $X, Y, Z \in [s, S]$

$$\begin{array}{rcl} H_{-2}(X,Y,Z) & = & Y - \lambda(\overline{F}(Y,Z) - G(X,Y)), \\ H_{-1}(X,Y,Z) & = & Y - \lambda(G(Y,Z) - G(X,Y)), \\ H_1(X,Y,Z) & = & Y - \lambda(F(Y,Z) - F(X,Y)), \\ H_2(X,Y,Z) & = & Y - \lambda(F(Y,Z) - \overline{F}(X,Y)). \end{array}$$

Then we have the following lemma.

LEMMA 4.2. *Let $\lambda M \leq 1$ and $a \in [s, S]$; then we have the following:*

(i) $H_{\pm_1}(a,a,a) = a$, *and* $H_{\pm_2}(s,s,s) = s, H_{\pm_2}(S,S,S) = S$.

(ii) H_i *is nondecreasing in each of its variables.*

(iii) *Let* $\{T_i\}_{i\in z\setminus\{0\}}$ *be a sequence in* $[s,S]$, *and define* $P_i=(T_{i-1},T_i,T_{i+1})$ *if* $|i|\geq 2$, $P_1=(T_{-1},T_1,T_2)$, *and* $P_{-1}=(T_{-2},T_{-1},T_1)$. *Then for* $|i|\geq 3$,

$$\begin{aligned}
\frac{\partial H_1}{\partial X}(P_{i+1}) &+ \frac{\partial H_1}{\partial Y}(P_i) &+ \frac{\partial H_1}{\partial Z}(P_{i-1}) &= 1,\\
\frac{\partial H_{-1}}{\partial X}(P_{i+1}) &+ \frac{\partial H_{-1}}{\partial Y}(P_i) &+ \frac{\partial H_{-1}}{\partial Z}(P_{i-1}) &= 1,\\
\frac{\partial H_1}{\partial X}(P_3) &+ \frac{\partial H_1}{\partial Y}(P_2) &+ \frac{\partial H_2}{\partial Z}(P_1) &= 1,\\
\frac{\partial H_{-2}}{\partial X}(P_{-1}) &+ \frac{\partial H_{-1}}{\partial Y}(P_{-2}) &+ \frac{\partial H_{-1}}{\partial Z}(P_{-3}) &= 1,\\
\frac{\partial H_2}{\partial X}(P_1) &+ \frac{\partial H_{-2}}{\partial Y}(P_{-1}) &+ \frac{\partial H_{-1}}{\partial Z}(P_{-2}) &= 1,\\
\frac{\partial H_1}{\partial X}(P_2) &+ \frac{\partial H_2}{\partial Y}(P_1) &+ \frac{\partial H_{-2}}{\partial Z}(P_{-1}) &= 1.
\end{aligned}$$

Proof. From Lemma 4.1, $\overline{F}(s,s)=f(s)=g(s)$, $\overline{F}(S,S)=f(S)=g(S)$, and for all $a\in[s,S]$, $F(a,a)=f(a), G(a,a)=g(a)$. Hence $H_{\pm1}(a,a,a)=a, H_{\pm2}(s,s,s)=s, H_{\pm2}(S,S,S)=S$. This proves (i). By symmetry it is enough to prove (ii) for H_2. Let (X,Y,Z), $X_1\leq X_2, Y_1\leq Y_2, Z_1\leq Z_2$, be given. Then

$$H_2(X_1,Y,Z)-H_2(X_2,Y,Z)=\lambda(\overline{F}(X_1,Y)-\overline{F}(X_2,Y))\leq 0.$$

Without loss of generality we can assume that $g(\theta_g)=\min g\geq \min f=f(\theta_f)$. For $X\geq\theta_g, Z\leq\theta_f$, define $\tilde{A}(X)\leq\theta_f$ and $\tilde{B}(Z)\geq\theta_f$ by $f(\tilde{A}(X))=g(X)$ and $f(Z)=f(\tilde{B}(Z))$. Then we have by direct calculations

$$I_1=F(Y_1,Z)-F(Y_2,Z)=\begin{cases}
f(\tilde{B}(Z))-f(Y_2) & \text{if } Z\leq\theta_f, \quad Y_1\leq\tilde{B}(Z)\leq Y_2,\\
f(Y_1)-f(Y_2) & \text{if } Z\leq\theta_f, \quad Y_1\geq\tilde{B}(Z)\\
 & \text{or } Z\geq\theta_f, \quad Y_1\geq\theta_f,\\
f(\theta_f)-f(Y_2) & \text{if } Z\geq\theta_f, \quad Y_1\leq\theta_f\leq Y_2,\\
0 & \text{otherwise,}
\end{cases}$$

$$I_2=\overline{F}(X,Y_1)-\overline{F}(X,Y_2)=\begin{cases}
f(Y_1)-f(\min(Y_2,\tilde{A}(X))) & \text{if } X\geq\theta_g, \quad Y_1\leq\tilde{A}(X),\\
f(Y_1)-f(\min(Y_2,\tilde{A}(\theta_g))) & \text{if } X\leq\theta_g, \quad Y_1\leq\tilde{A}(\theta_g),\\
0 & \text{otherwise.}
\end{cases}$$

Let $I=-\lambda(I_1-I_2)$. Then from the above calculation, $I=-\lambda I_1$ if $Y_1\geq\theta_f$ and $I=\lambda I_2$ if $Y_2\leq\theta_f$. In either case we have $|I|\leq\lambda M|Y_1-Y_2|$. Now suppose that $Y_1\leq\theta_f\leq Y_2$; then we have

$$|I|\leq(|I_1|+|I_2|)\leq\lambda M(|Y_2-\theta_f|+|Y_1-\tilde{A}(\theta_g)|)=\lambda M|Y_1-Y_2|.$$

Hence, since F and $\overline{F}$ are nondecreasing in the first variable and nonincreasing in the second variable, we obtain

$$\begin{aligned}
H_2(X,Y_1,Z)-H_2(X,Y_2,Z) &= Y_1-Y_2-\lambda(F(Y_1,Z)-F(Y_2,Z))\\
&\quad +\lambda(\overline{F}(X,Y_1)-\overline{F}(X,Y_2))\\
&= Y_1-Y_2-\lambda(I_1-I_2)\leq Y_1-Y_2+\lambda M|Y_1-Y_2|\\
&\leq (1-\lambda M)(Y_1-Y_2)\leq 0,\\
H_2(X,Y,Z_1)-H_2(X,Y,Z_2) &= -\lambda(F(Y,Z_1)-F(Y,Z_2))\ \leq\ 0.
\end{aligned}$$

This proves (ii).

Let $i \geq 3$; then

$$\frac{\partial H_1}{\partial X}(P_{i+1}) + \frac{\partial H_1}{\partial Y}(P_i) + \frac{\partial H_1}{\partial Z}(P_{i-1}) = \lambda \frac{\partial F}{\partial a}(T_i, T_{i+1}) + 1 - \lambda \left(\frac{\partial F}{\partial a}(T_i, T_{i+1}) - \frac{\partial F}{\partial b}(T_{i-1}, T_i) \right) - \lambda \frac{\partial F}{\partial b}(T_{i-1}, T_i) = 1.$$

This proves the first equality in (iii). The proof is similar for the second equality in (iii). For the third, fourth, fifth, and sixth equalities we have

$$\begin{aligned}
&\frac{\partial H_1}{\partial X}(P_3) + \frac{\partial H_1}{\partial Y}(P_2) + \frac{\partial H_2}{\partial Z}(P_1) = \lambda \frac{\partial F}{\partial a}(T_2, T_3) + 1 \\
&\qquad -\lambda \left(\frac{\partial F}{\partial a}(T_2, T_3) - \frac{\partial F}{\partial b}(T_1, T_2) \right) - \lambda \frac{\partial F}{\partial b}(T_1, T_2) \;=\; 1, \\
&\frac{\partial H_{-2}}{\partial X}(P_{-1}) + \frac{\partial H_{-1}}{\partial Y}(P_{-2}) + \frac{\partial H_{-1}}{\partial Z}(P_3) = \lambda \frac{\partial G}{\partial a}(T_{-2}, T_{-1}) + 1 \\
&\qquad -\lambda \left(\frac{\partial G}{\partial a}(T_{-2}, T_{-1}) - \frac{\partial G}{\partial b}(T_{-3}, T_{-2}) \right) - \lambda \frac{\partial G}{\partial b}(T_{-3}, T_{-2}) \;=\; 1, \\
&\frac{\partial H_2}{\partial X}(P_1) + \frac{\partial H_{-2}}{\partial Y}(P_{-1}) + \frac{\partial H_{-1}}{\partial Z}(P_{-2}) = \lambda \frac{\partial \overline{F}}{\partial a}(T_{-1}, T_1) + 1 \\
&\qquad -\lambda \left(\frac{\partial \overline{F}}{\partial a}(T_{-1}, T_1) - \frac{\partial G}{\partial b}(T_{-2}, T_{-1}) \right) - \lambda \frac{\partial G}{\partial b}(T_{-2}, T_{-1}) \;=\; 1, \\
&\frac{\partial H_1}{\partial X}(P_2) + \frac{\partial H_2}{\partial Y}(P_1) + \frac{\partial H_{-2}}{\partial Z}(P_{-1}) = \lambda \frac{\partial F}{\partial a}(T_1, T_2) + 1 \\
&\qquad -\lambda \left(\frac{\partial F}{\partial a}(T_1, T_2) - \frac{\partial \overline{F}}{\partial b}(T_{-1}, T_1) \right) - \lambda \frac{\partial \overline{F}}{\partial b}(T_{-1}, T_1) \;=\; 1.
\end{aligned}$$

This completes the proof of Lemma 4.2.

4.2. L^∞ and TV bounds. The next lemmas show that the scheme (3.6) is L^1-contractive and the idea of the proof is taken from [11].

LEMMA 4.3. *Let $u_0 \in L^\infty(\mathbb{R}, [s, S])$ be the initial data, and let $\{u_i^n\}$ be the corresponding solution calculated by the finite volume scheme* (3.6)*. When $\lambda M \leq 1$, then*

$$s \leq u_i^n \leq S \quad \forall i, n. \tag{4.2}$$

Proof. Since $s \leq u_0 \leq S$, hence for all i, $s \leq u_i^0 \leq S$. By induction, assume that (4.2) holds for n. Then from (i) and (ii) of Lemma 4.2 we have

$$\begin{aligned}
s &= H_{-1}(s,s,s) \leq H_{-1}(u_{i-1}^n, u_i^n, u_{i+1}^n) = u_1^{n+1} \leq H_{-1}(S,S,S) = S \text{ if } i \leq -2, \\
s &= H_1(s,s,s) \leq H_1(u_{i-1}^n, u_i^n, u_{i+1}^n) = u_i^{n+1} \leq H_1(S,S,S) = S \quad \text{if } i \geq 2, \\
s &= H_{-2}(s,s,s) \leq H_{-2}(u_{-2}^n, u_{-1}^n, u_1^n) = u_{-1}^{n+1} \leq H_{-2}(S,S,S) = S, \\
s &= H_2(s,s,s) \leq H_2(u_{-1}^n, u_1^n, u_2^n) = u_1^{n+1} \leq H_2(S,S,S) = S.
\end{aligned}$$

This proves (4.2).

LEMMA 4.4. *Let $u_0, v_0 \in L^\infty(\mathbb{R},\ [s, S])$ be initial datas, and let $\{u_i^n\}$ and $\{v_i^n\}$ be the corresponding solutions calculated by the finite volume scheme* (3.6)*. Let $\lambda M \leq 1$*

and $i_0 \le j_0$; then

$$\begin{aligned}
\sum_{\substack{i_0 \le i \le j_0 \\ i \neq 0}} |u_i^{n+1} - v_i^{n+1}| &\le \sum_{\substack{i_0 - 1 \le i \le j_0 + 1 \\ i \neq 0}} |u_i^n - v_i^n|, \\
\sum_{i \neq 0} |u_i^{n+1} - u_i^n| &\le \sum_{i \neq 0} |u_i^n - u_i^{n-1}|.
\end{aligned}$$

Lemma 4.4 is a localized version of the Crandall–Tartar lemma [4], which we will prove along the lines of [11].

Proof. The first inequality in Lemma 4.4 will be proved for $i_0 \le -1$ and $j_0 \ge 1$. The other cases follow in the same manner. For $\theta \in [0,1]$, let $p_i^n(\theta) = \theta u_i^n + (1-\theta) v_i^n$ and

$$P_i^n(\theta) = \begin{cases} (p_{i-1}^n(\theta),\ p_i^n(\theta), p_{i+1}^n(\theta)) & \text{if} \quad |i| \ge 2, \\ (p_{-2}^n(\theta),\ p_{-1}^n(\theta), p_1^n(\theta)) & \text{if} \quad i = -1, \\ (p_{-1}^n(\theta),\ p_1^n(\theta),\ p_2^n(\theta)) & \text{if} \quad i = 1. \end{cases}$$

From Lemma 4.3 we have $p_i^n(\theta) \in [s, S]$ for all i, n, and θ. From their definitions, the H_i's are uniformly continuous functions, and from (ii) in Lemma 4.2 a.e. (X, Y, Z), $\frac{\partial H_i}{\partial X} \ge 0$, $\frac{\partial H_i}{\partial Y} \ge 0$, $\frac{\partial H_i}{\partial Z} \ge 0$. Hence from the mean value theorem

$$\begin{aligned}
\sum_{i_0}^{-2} |u_i^{n+1} - v_i^{n+1}| &= \sum_{i_0}^{-2} |H_{-1}(u_{i-1}^n, u_i^n, u_{i+1}^n) - H_{-1}(v_{i-1}^n, v_i^n, v_{i+1}^n)| \\
&\le \sum_{i_0}^{-2} |u_{i-1}^n - v_{i-1}^n| \int_0^1 \frac{\partial H_{-1}}{\partial X}(P_i^n(\theta)) d\theta \\
&+ \sum_{i_0}^{-2} |u_i^n - v_i^n| \int_0^1 \frac{\partial H_{-1}}{\partial Y}(P_i^n(\theta)) d\theta + |u_{i+1}^n - v_{i+1}^n| \int_0^1 \frac{\partial H_{-1}}{\partial Z}(P_i^n(\theta)) d\theta \\
&= |u_{i_0-1}^n - v_{i_0-1}^n| \int_0^1 \frac{\partial H_{-1}}{\partial X}(P_{i_0}^n(\theta)) d\theta \\
&+ \sum_{i_0}^{-3} |u_i^n - v_i^n| \int_0^1 \left(\frac{\partial H_{-1}}{\partial X}(P_{i+1}^n(\theta)) + \frac{\partial H_{-1}}{\partial Y}(P_i^n(\theta)) + \frac{\partial H_{-1}}{\partial Z}(P_{i-1}^n(\theta)) \right) d\theta \\
&+ |u_{-2}^n - v_{-2}^n| \int_0^1 \left(\frac{\partial H_{-1}}{\partial Y}(P_{-2}^n(\theta)) + \frac{\partial H_{-1}}{\partial Z}(P_{-3}^n(\theta)) \right) d\theta \\
&+ |u_{-1}^n - v_{-1}^n| \int_0^1 \frac{\partial H_{-1}}{\partial Z}(P_{-2}^n(\theta)) d\theta.
\end{aligned}$$

Now $\frac{\partial H_{-1}}{\partial X}(X, Y, Z) = \lambda \frac{\partial G}{\partial a}(X, Y) \le \lambda M \le 1$, and from the second equality of (iii) in Lemma 4.2 we obtain

$$\begin{aligned}
\sum_{i_0}^{-2} |u_i^{n+1} - v_i^{n+1}| \le \sum_{i_0 - 1}^{-3} |u_i^n - v_i^n| &+ |u_{-1}^n - v_{-1}^n| \int_0^1 \frac{\partial H_{-1}}{\partial Z}(P_{-2}^n(\theta)) d\theta \\
&+ |u_{-2}^n - v_{-2}^n| \int_0^1 \left(\frac{\partial H_{-1}}{\partial Y}(P_{-2}^n(\theta)) + \frac{\partial H_{-1}}{\partial Z}(P_{-3}^n(\theta)) \right) d\theta.
\end{aligned}$$

Since $\frac{\partial H_1}{\partial Z} = -\lambda \frac{\partial F}{\partial b} \le \lambda M \le 1$, the following inequalities result from the first equality

of (iii) in Lemma 4.2:

$$\begin{aligned}\sum_{2}^{j_0}|u_i^{n+1}-v_i^{n+1}| &\leq \sum_{3}^{j_0+1}|u_i^n-v_i^n|+|u_1^n-v_1^n|\int_0^1\frac{\partial H_1}{\partial X}(P_2^n(\theta))d\theta\\ &\quad+|u_2^n-v_2^n|\int_0^1\left(\frac{\partial H_1}{\partial Y}(P_2^n(\theta))+\frac{\partial H_1}{\partial X}(P_3^n(\theta))\right)d\theta.\end{aligned}$$

Moreover,

$$\begin{aligned}&|u_{-1}^{n+1}-v_{-1}^{n+1}|+|u_1^{n+1}-v_1^{n+1}|\\ &=|H_{-2}(u_{-2}^n,u_{-1}^n,u_1^n)-H_{-2}(v_{-2}^n,v_{-1}^n,v_1^n)|+|H_2(u_{-1}^n,u_1^n,u_2^n)-H_2(v_{-1}^n,v_1^n,v_2^n)|\\ &\leq|u_{-2}^n-v_{-2}^n|\int_0^1\frac{\partial H_{-2}}{\partial X}(P_{-1}^n(\theta))d\theta+|u_{-1}^n-v_{-1}^n|\int_0^1\left(\frac{\partial H_2}{\partial X}(P_1^n(\theta))+\frac{\partial H_{-2}}{\partial Y}(P_{-1}^n(\theta))\right)d\theta\\ &\quad+|u_1^n-v_1^n|\int_0^1\left(\frac{\partial H_2}{\partial Y}(P_1^n(\theta))+\frac{\partial H_{-2}}{\partial Z}(P_{-1}^n(\theta))\right)d\theta+|u_2^n-v_2^n|\int_0^1\frac{\partial H_2}{\partial Z}(P_1^n(\theta))d\theta.\end{aligned}$$

Summing up all the above three inequalities and from the last four equalities of (iii) in Lemma 4.2 we obtain

$$\begin{aligned}\sum_{\substack{i_0\leq i\leq j_0\\ i\neq 0}}|u_i^{n+1}-v_i^{n+1}| &\leq \sum_{i_0-1}^{-3}|u_i^n-v_i^n|+\sum_{3}^{j_0+1}|u_i^n-v_i^n|\\ &\quad+|u_{-2}^n-v_{-2}^n|\int_0^1\left(\frac{\partial H_{-2}}{\partial X}(P_{-1}^n(\theta))+\frac{\partial H_{-1}}{\partial Y}(P_{-2}^n(\theta))+\frac{\partial H_{-1}}{\partial Z}(P_{-3}^n(\theta))\right)d\theta\\ &\quad+|u_{-1}^n-v_{-1}^n|\int_0^1\left(\frac{\partial H_2}{\partial X}(P_1^n(\theta))+\frac{\partial H_{-2}}{\partial Y}(P_{-1}^n(\theta))+\frac{\partial H_{-1}}{\partial Z}(P_{-2}^n(\theta))\right)d\theta\\ &\quad+|u_1^n-v_1^n|\int_0^1\left(\frac{\partial H_1}{\partial X}(P_2^n(\theta))+\frac{\partial H_2}{\partial Y}(P_1^n(\theta))+\frac{\partial H_{-2}}{\partial Z}(P_{-2}^n(\theta))\right)d\theta\\ &\quad+|u_2^n-v_2^n|\int_0^1\left(\frac{\partial H_1}{\partial X}(P_3^n(\theta))+\frac{\partial H_1}{\partial Y}(P_2^n(\theta))+\frac{\partial H_2}{\partial Z}(P_1^n(\theta))\right)d\theta\\ &=\sum_{\substack{i_0-1\leq i\leq j_0+1\\ i\neq 0}}|u_i^n-v_i^n|.\end{aligned}$$

Take the special choice of v_0 by $v_0(x)=u_i^1$ in $[x_{i-1/2},x_{i+1/2})$. Then it follows easily that $v_i^n=u_i^{n+1}$. Now substituting this in the first inequality of the lemma and taking $i_0=-\infty,\ j_0=\infty$ we obtain the second inequality. This completes the proof of Lemma 4.4.

Next we use the singular mapping technique introduced in [27, 23, 22, 28] to obtain TV bounds for the transformed scheme, and this allows us to pass to the limit as $h\to 0$.

Let $k:[s,S]\to\mathbb{R}$ be a Lipschitz continuous function satisfying (H_2), and let K be the corresponding numerical flux as in (3.1). Let θ_k denote the unique minima of k. For $A\in[s,S], a,b\in\mathbb{R},\ \{u_{j-1},u_j,u_{j+1},u_{j+2}\}\subset[s,S]$, define

$$\begin{aligned}\psi_{k,A}(u) &= \int_A^u|k'(\theta)|d\theta, & \chi_-(k'(u)) &= \begin{cases}0 & \text{if} \quad u\in(\theta_k,S],\\ 1 & \text{if} \quad u\in[s,\theta_k],\end{cases}\\ \chi(a,b) &= \begin{cases}1 & \text{if} \quad a\leq b,\\ 0 & \text{if} \quad a>b,\end{cases} & \chi_+(k'(u)) &= \begin{cases}1 & \text{if} \quad u\in[\theta_k,S],\\ 0 & \text{if} \quad u\in[s,\theta_k),\end{cases}\end{aligned}$$

and $H_{i+1/2}=K(u_i,u_{i+1})$ for $j-1\leq i\leq j+1$.

Apart from χ_-, χ_+ we will use the standard notation

$$a_+ = \max(a,0), \quad a_- = \min(a,0), \quad a = a_+ + a_-, \quad |a| = a_+ - a_-.$$

LEMMA 4.5. *With the above notation we have the following inequalities:*

$$-\chi(u_j,u_{j+1}) \int_{u_j}^{u_{j+1}} k'_-(\theta)d\theta \ \le \ \chi_-(k'(u_j))|H_{j+1/2} - H_{j-1/2}|, \tag{4.3}$$

$$\chi(u_j,u_{j+1}) \int_{u_j}^{u_{j+1}} k'_+(\theta)d\theta \le \chi_+(k'(u_{j+1}))|H_{j+3/2} - H_{j+1/2}| \tag{4.4}$$

$$-(\psi_{k,A}(u_j) - \psi_{k,A}(u_{j+1}))_- = \chi(u_j,u_{j+1}) \left\{ \int_{u_j}^{u_{j+1}} k'_+(\theta)d\theta - \int_{u_j}^{u_{j+1}} k'_-(\theta)d\theta \right\}$$

$$\le \chi_-(k'(u_j))|H_{j+1/2} - H_{j-1/2}| + \chi_+(k'(u_{j+1}))|H_{j+3/2} - H_{j+1/2}|. \tag{4.5}$$

The proof of this lemma can be found in [28, Lemma 3.3], just replacing the requirement of a single maximum by a single minimum.

Singular mappings. Let f, g satisfy the hypotheses (H_1) and (H_2). Let θ_f, θ_g be the respective minima for f and g. Define the singular mappings ψ_1, ψ_2 associated with f and g as follows.

Case 1. $f(\theta_f) \le g(\theta_g)$. Choose $A \ge \theta_f$ such that $f(A) = g(\theta_g)$ and for $u \in [s,S]$

$$\psi_1(u) = \psi_{g,\theta_g}(u) = \int_{\theta_g}^{u} |g'(\theta)|d\theta, \quad \psi_2(u) = \psi_{f,A}(u) = \int_{A}^{u} |f'(\theta)|d\theta.$$

Case 2. $f(\theta_f) \ge g(\theta_g)$. Choose $A \le \theta_g$ such that $f(\theta_f) = g(A)$ and for $u \in [s,S]$

$$\psi_1(u) = \psi_{g,A}(u) = \int_{A}^{u} |g'(\theta)|d\theta, \quad \psi_2(u) = \psi_{f,\theta_f}(u) = \int_{\theta_f}^{u} |f'(\theta)|d\theta.$$

In order to obtain TV bounds for the transformed sequence under the singular mappings, we have to estimate the error term E defined as below. This error estimate will be carried out in the next two lemmas.

For $\{u_{-2}, u_{-1}, u_1, u_2\} \subset [s,S]$, define $z_1 = \psi_2(u_1)$, $z_{-1} = \psi_1(u_{-1})$, $H_{3/2} = F(u_1,u_2)$, $H_{1/2} = H_{-1/2} = \overline{F}(u_{-1},u_1)$, $H_{-3/2} = G(u_{-2},u_{-1})$, and

$$E = -(z_{-1} - z_1)_- - \chi(u_1,u_2) \int_{u_1}^{u_2} f'_-(\theta)d\theta + \chi(u_{-2},u_{-1}) \int_{u_{-2}}^{u_{-1}} g'_+(\theta)d\theta$$
$$-|H_{-1/2} - H_{-3/2}| - |H_{3/2} - H_{1/2}|.$$

LEMMA 4.6. *With the above notation, for any sequence* $\{u_{-2}, u_{-1}, u_1, u_2\} \subset [s,S]$, *we have* $E \le 0$.

Proof. Without loss of generality we can assume that $\theta_f \le \theta_g$ and $f(\theta_f) \le g(\theta_g)$ (see Figure 1). Now $\psi_2(S) = f(S) - f(A) = g(S) - g(\theta_g) = \psi_1(S)$ and $\psi_2(s) = -(f(s) - f(\theta_f)) - (f(A) - f(\theta_f)) \le -(f(A) - f(s)) = (g(\theta_g) - g(s)) = \psi_1(s)$. Hence the range of ψ_1 is contained in the range of ψ_2. Therefore for each $u \in [s,S]$ there exists a unique $\rho(u) \in [s,S]$ such that $\psi_1(u) = \psi_2(\rho(u))$ and $u \mapsto \rho(u)$ is an increasing function since ψ_1, ψ_2 are increasing functions.

For $u_{-1} \ge \theta_g$, define $\tilde{A}(u_{-1}) \le \theta_f$ by $f(\tilde{A}(u_{-1})) = g(u_{-1})$ (see Figure 1).

Step 1. Let $u_{-1} \geq \theta_g$, $u_1 \geq \tilde{A}(u_{-1})$.
In this case it is easy to see that $H_{1/2} = H_{-1/2} = \overline{F}(u_{-1}, u_1) = g(u_{-1})$ and

$$\chi(u_{-2}, u_{-1}) \int_{u_{-2}}^{u_{-1}} g'_+(\theta) d\theta = \chi(u_{-2}, u_{-1})(g(u_{-1}) - g(\max(\theta_g, u_{-2}))),$$
$$|H_{-1/2} - H_{-3/2}| = |\overline{F}(u_{-1}, u_1) - G(u_{-2}, u_{-1})| = |g(u_{-1}) - g(\max(\theta_g, u_{-2}))|$$
$$\geq \chi(u_{-2}, u_{-1}) \int_{u_{-2}}^{u_{-1}} g'_+(\theta) d\theta.$$

Hence

$$\chi(u_{-2}, u_{-1}) \int_{u_{-2}}^{u_{-1}} g'_+(\theta) d\theta - |H_{-1/2} - H_{-3/2}| \leq 0. \tag{4.6}$$

Since $u_{-1} \geq \theta_g$ this implies that $0 \leq g(u_{-1}) - g(\theta_g) = \psi_1(u_{-1}) = \psi_2(\rho(u_{-1}))$. Hence $\rho(u_{-1}) \geq A$ and $f(\rho(u_{-1})) = g(u_{-1})$.

Now $(z_{-1} - z_1)_- \neq 0$ if and only if $\psi_2(\rho(u_{-1})) = \psi_1(u_{-1}) < \psi_2(u_1)$. This implies that $\rho(u_{-1}) < u_1$. Therefore for $A \leq \rho(u_{-1}) < u_1$ we have

$$-(z_{-1} - z_1)_- = \psi_2(u_1) - \psi_2(\rho(u_{-1})) = \int_{\rho(u_{-1})}^{u_1} |f'(\theta)| d\theta = f(u_1) - f(\rho(u_{-1})).$$

Hence

$$-(z_{-1} - z_1)_- = \begin{cases} f(u_1) - f(\rho(u_{-1})) & \text{if } A \leq \rho(u_{-1}) < u_1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.7}$$

Now for $\tilde{A}(u_{-1}) \leq u_1$, $0 \leq \overline{F}(u_{-1}, u_1) - g(\theta_g) = g(u_{-1}) - g(\theta_g) = \psi_1(u_{-1}) = \psi_2(\rho(u_{-1})) = f(\rho(u_{-1})) - f(A)$, and therefore $\overline{F}(u_{-1}, u_1) = f(\rho(u_{-1}))$. Hence either $u_1 \leq u_2$ or $\rho(u_{-1}) < u_1$, and for all u_2 we have

$$\begin{aligned} |H_{3/2} - H_{1/2}| &= |F(u_1, u_2) - \overline{F}(u_{-1}, u_1)| = |F(u_1, u_2) - f(\rho(u_{-1}))| \\ &\geq \begin{cases} |f(u_1) - f(\rho(u_{-1}))| & \text{if } u_1 \geq \theta_f, \\ |f(\min(u_2, \theta_f)) - f(\rho(u_{-1}))| & \text{if } u_1 \leq \theta_f, \end{cases} \end{aligned} \tag{4.8}$$

$$-\chi(u_1, u_2) \int_{u_1}^{u_2} f'_-(\theta) d\theta = \begin{cases} 0 & \text{if } u_1 \geq \theta_f, \\ f(u_1) - f(\min(u_2, \theta_f)) & \text{if } u_1 \leq \theta_f. \end{cases} \tag{4.9}$$

Let $E_1 = -(z_{-1} - z_1)_- - \chi(u_1, u_2) \int_{u_1}^{u_2} f'_-(\theta) d\theta - |H_{1/2} - H_{3/2}|$. Suppose $\rho(u_{-1}) < u_1$; then $u_1 \geq \theta_f$, and hence from (4.7), (4.8), and (4.9) we have

$$E_1 \leq f(u_1) - f(\rho(u_{-1})) - |f(u_1) - f(\rho(u_{-1}))| \leq 0.$$

Suppose $u_1 \leq \rho(u_{-1})$; then $(z_{-1} - z_1)_- = 0$. If $\theta_f \leq u_1$, then $E_1 = -|H_{1/2} - H_{3/2}| \leq 0$. Let $\theta_f > u_1 \geq \tilde{A}(u_{-1})$; then by the definition of $\tilde{A}(u_{-1})$ we have $f(\tilde{A}(u_{-1})) = g(u_{-1}) = f(\rho(u_{-1}))$. If $u_2 \leq u_1$, then clearly $E_1 = -|H_{1/2} - H_{3/2}| \leq 0$. Let $u_1 \leq u_2$. Now $f(\tilde{A}(u_{-1})) = g(u_{-1}) = f(\rho(u_{-1}))$, and hence $f(u_1) \leq f(\tilde{A}(u_{-1})) = f(\rho(u_{-1}))$. Hence from (4.7), (4.8), and (4.9),

$$E_1 \leq f(u_1) - f(\min(u_2, \theta_f)) - |f(\rho(u_{-1})) - f(\min(u_2, \theta_f))| \leq 0.$$

This together with (4.6) implies that $E \leq 0$.

Step 2. $u_{-1} \geq \theta_g,\ u_1 < \tilde{A}(u_{-1})$.
In this case $\overline{F}(u_{-1}, u_1) = f(u_1)$ and $g(u_{-1}) \leq f(u_1)$.

$$\begin{aligned} |H_{-1/2} - H_{-3/2}| &= |\overline{F}(u_{-1}, u_1) - G(u_{-2}, u_{-1})| = |f(u_1) - g(\max(u_{-2}, \theta_g))| \\ &\geq \chi(u_{-2}, u_{-1})|g(u_{-1}) - g(\max(u_{-2}, \theta_g))| \\ &\geq \chi(u_{-2}, u_{-1}) \int_{u_{-2}}^{u_{-1}} g'_+(\theta) d\theta, \end{aligned} \tag{4.10}$$

$$\begin{aligned} |H_{3/2} - H_{1/2}| &= |F(u_1, u_2) - \overline{F}(u_{-1}, u_1)| = |f(\min(u_2, \theta_f)) - f(u_1)| \\ &\geq -\chi(u_1, u_2) \int_{u_1}^{u_2} f'_-(\theta) d\theta. \end{aligned} \tag{4.11}$$

From Step 1, $(z_{-1} - z_1)_- \neq 0$ if and only if $A < \rho(u_{-1}) < u_1$. Hence $(z_{-1} - z_1)_- = 0$. Combining this with (4.10) and (4.11) gives $E \leq 0$.

Step 3. $u_{-1} < \theta_g, u_1 \geq \tilde{A}(\theta_g)$ (see Figure 1).

In this case $\overline{F}(u_{-1}, u_1) = f(A) = g(\theta_g)$. Since $u_1 \geq \tilde{A}(\theta_g)$ this implies that $f(u_1) \leq f(A)$ if $u_1 \leq \theta_f$. Let $u_1 \leq u_2$; then

$$-\chi(u_1, u_2) \int_{u_1}^{u_2} f'_-(\theta) - |H_{3/2} - H_{1/2}| \leq \begin{cases} -|f(u_1) - f(A)| & \text{if } u_1 \geq \theta_f, \\ f(u_1) - f(\min(u_2, \theta_f)), & \\ -|f(u_1) - f(\min(u_2, \theta_f))| & \text{if } u_1 \leq \theta_f. \end{cases} \tag{4.12}$$

Since $u_{-1} \leq \theta_g$, hence $\chi(u_{-2}, u_{-1}) \int_{u_{-2}}^{u_{-1}} g'_+(\theta) d\theta = 0$. Let $(z_{-1} - z_1)_- = 0$; then from (4.12) we have $E \leq 0$. Suppose $(z_{-1} - z_1)_- \neq 0$; then $\rho(u_{-1}) \leq u_1$ and

$$\begin{aligned} -(z_{-1} - z_1)_- &= \psi_2(u_1) - \psi_2(\rho(u_{-1})) = \int_{\rho(u_{-1})}^{u_1} |f'(\theta)| d\theta, \\ |H_{-1/2} - H_{-3/2}| &= G(u_{-2}, u_{-1}) - \overline{F}(u_{-1}, u_1) \\ &\geq g(u_{-1}) - g(\theta_g) = -\psi_1(u_{-1}) = -\psi_2(\rho(u_{-1})). \end{aligned}$$

Hence from (4.12) we have

$$E \leq \begin{cases} -(z_{-1} - z_1)_- - |H_{-1/2} - H_{-3/2}| \leq \displaystyle\int_{\rho(u_{-1})}^{u_1} |f'(\theta)| d\theta - \int_{\rho(u_{-1})}^{A} |f'(\theta)| d\theta \leq 0 & \\ \qquad\qquad \text{if } u_1 \leq A, & \\ \displaystyle\int_{\rho(u_{-1})}^{u_1} |f'(\theta)| d\theta - \int_{\rho(u_{-1})}^{A} |f'(\theta)| d\theta - \int_{A}^{u_1} |f'(\theta)| d\theta = 0 \qquad \text{if } u_1 \geq A. \end{cases}$$

Hence in all cases $E \leq 0$.

Step 4. Let $u_{-1} \leq \theta_g,\ u_1 \leq \tilde{A}(\theta_g)$. In this case $\overline{F}(u_{-1}, u_1) = f(u_1)$ and $\chi(u_{-2}, u_{-1}) \int_{u_{-2}}^{u_{-1}} g'_+(\theta) d\theta = 0$. Let $u_1 \leq u_2, u_{-2} \leq u_{-1}$; then

$$\begin{aligned} |H_{3/2} - H_{1/2}| &= |F(u_1, u_2) - \overline{F}(u_{-1}, u_1)| = |f(\min(u_2, \theta_f)) - f(u_1)| \\ &= -\chi(u_1, u_2) \int_{u_1}^{u_2} f'_-(\theta) d\theta. \end{aligned} \tag{4.13}$$

If $u_1 \leq \rho(u_{-1})$, then $(z_{-1} - z_1)_- = 0$, and therefore from (4.13) $E \leq 0$. Hence assume that $\rho(u_{-1}) < u_1$; then $f(\rho(u_{-1})) > f(u_1)$. Since $\psi_2(\rho(u_{-1})) = \psi(u_{-1})$ this

implies that

$$f(\rho(u_{-1})) - f(\theta_f) + f(A) - f(\theta_f) = \int_{\rho(u_{-1})}^{A} |f'(\theta)| d\theta = -\psi_2(\rho(u_{-1}))$$
$$= -\psi_1(u_{-1}) = \int_{u_{-1}}^{\theta_g} |g'(\theta)| d\theta = g(u_{-1}) - g(\theta_g).$$

Hence $f(\rho(u_{-1})) - g(u_{-1}) = 2(f(\theta_f) - f(A)) \leq 0$, and therefore $f(u_1) \leq f(\rho(u_{-1})) \leq g(u_{-1})$. This implies that

$$|H_{-3/2} - H_{-1/2}| = |G(u_{-2}, u_{-1}) - \overline{F}(u_{-1}, u_1)| \geq g(u_{-1})_- f(u_1).$$

Since $f(A) = f(\tilde{A}(\theta_g))$ we have

$$\begin{aligned} E &\leq -(z_{-1} - z_1)_- - |H_{-3/2} - H_{-1/2}| \leq \int_{\rho(u_{-1})}^{u_1} |f'(\theta) d\theta - |g(u_{-1}) - f(u_1)| \\ &= f(\rho(u_{-1})) - g(u_{-1}) \leq 0. \end{aligned}$$

This proves Lemma 4.6.

LEMMA 4.7. *Let $u_0 \in L^\infty(\mathbb{R})$ such that $s \leq u_0(x) \leq S$ for all $x \in \mathbb{R}$ and $N(f, g, u_0) < 0$. Let $\{u_i^n\}$ be the scheme defined as in* (3.6). *Let ψ_1 and ψ_2 be as in Lemma* 4.6. *We introduce the constant*

$$L = \max\{Lip(\psi_1), Lip(\psi_2), \|\psi_1\|_\infty, \|\psi_2\|_\infty\}$$

and define

$$\begin{aligned} z_i^n &= \begin{cases} \psi_2(u_i^n) & \text{if} \quad i \geq 1, \\ \psi_1(u_i^n) & \text{if} \quad i \leq -1, \end{cases} \\ TV(z^n) &= \sum_{i \neq 0, -1} |z_i^n - z_{i+1}^n| + |z_{-1}^n - z_1^n|. \end{aligned} \tag{4.14}$$

Then

$$TV(z^n) \leq 2/\lambda \sum_{i \neq 0} |u_i^{n+1} - u_i^n| \leq 2/\lambda \sum_{i \neq 0} |u_i^1 - u_i^0| = 2N_h(f, g, u_0), \tag{4.15}$$

$$\sum_{i \neq 0} |z_i^n - z_i^m| \leq \lambda L |n - m| N(f, g, P, u_0). \tag{4.16}$$

Proof.

Define $H_{1/2} = H_{-1/2} = \overline{F}(u_{-1}^n, u_1^n)$ and

$$H_{j+1/2} = \begin{cases} F(u_j^n, u_{j+1}^n) & \text{if } j \geq 1, \\ G(u_j^n, u_{j+1}^n) & \text{if } j \leq -2. \end{cases}$$

Since $0 = \sum_{i \neq 0, -1} (z_i^n - z_{i+1}^n) + (z_{-1}^n - z_1^n)$,

$$\begin{aligned} \frac{1}{2} TV(z^n) &= \frac{1}{2} \left(\sum_{i \neq 0, -1} |z_i^n - z_{i+1}^n| + |z_{-1}^n - z_1^n| \right) \\ &= -\left(\sum_{i \neq 0, -1} (z_i^n - z_{i+1}^n)_- + (z_{-1}^n - z_1^n)_- \right) = I_1 + I_2 + I_3, \end{aligned}$$

where

$$\begin{aligned} I_1 &= -\sum_{i\le -3}(z_i^n - z_{i+1}^n)_-, \quad I_2 = -\sum_{i\ge 2}(z_i^n - z_{i+1}^n)_-, \\ I_3 &= -(z_{-2}^n - z_{-1}^n)_- - (z_{-1}^n - z_1^n)_- - (z_1^n - z_2^n)_-. \end{aligned}$$

From (4.3) to (4.5) we have

$$\begin{aligned} I_1 &= -\sum_{i\le -3}(z_i^n - z_{i+1}^n)_- = -\sum_{i\le -3}(\psi_1(u_i^n) - \psi_1(u_{i+1}^n))_- \\ &\le \sum_{i\le -3}\chi_-(g'(u_i^n))|H_{i+1/2} - H_{i-1/2}| + \chi_+(g'(u_{i+1}^n))|H_{i+3/2} - H_{i+1/2}| \\ &\le \sum_{i\le -3}|H_{i+1/2} - H_{i-1/2}| + \chi_+(g'(u_{-2}^n))|H_{-3/2} - H_{-5/2}|, \\ I_2 &= -\sum_{i\ge 2}(z_i^n - z_{i+1}^n)_- = -\sum_{i\ge 2}(\psi_2(u_i^n) - \psi_2(u_{i+1}^n))_- \\ &\le \sum_{i\ge 2}\chi_-(f'(u_i^n))|H_{i+1/2} - H_{i-1/2}| + \chi_+(f'(u_{i+1}^n))|H_{i+3/2} - H_{i+1/2}| \\ &\le \sum_{i\ge 3}|H_{i+1/2} - H_{i-1/2}| + \chi_-(f'(u_2^n))|H_{5/2} - H_{3/2}|, \\ -(z_{-2}^n - z_{-1}^n)_- &= \chi(u_{-2}^n, u_{-1}^n)\left(\int_{u_{-2}^n}^{u_{-1}^n} g'_+(\theta)d\theta - \int_{u_{-2}^n}^{u_{-1}^n} g'_-(\theta)d\theta\right) \\ &\le \chi(u_{-2}^n, u_{-1}^n)\int_{u_{-2}^n}^{u_{-1}^n} g'_+(\theta)d\theta + \chi_-(g'(u_{-2}^n))|H_{-3/2} - H_{-5/2}|, \\ -(z_1^n - z_2^n)_- &= \chi(u_1, u_2)\left(\int_{u_1^n}^{u_2^n} f'_+(\theta)d\theta - \int_{u_1^n}^{u_2^n} f'_-(\theta)d\theta\right) \\ &\le \chi_+(f'(u_2^n))|H_{5/2} - H_{3/2}| - \chi(u_1^n, u_2^n)\int_{u_1^n}^{u_2^n} f'_-(\theta)d\theta. \end{aligned}$$

Combining all the above three inequalities we obtain

$$\begin{aligned} \frac{1}{2}TV(z^n) &\le \sum_{|i|\ge 2}|H_{i+1/2} - H_{i-1/2}| + \chi(u_{-2}^n, u_{-1}^n)\int_{u_{-2}^n}^{u_{-1}^n} g'_+(\theta)d\theta \\ &\quad -\chi(u_1^n, u_2^n)\int_{u_1^n}^{u_2^n} f'_-(\theta)d\theta - (z_{-1}^n - z_1^n)_- \\ &= \sum_{i=-\infty}^{\infty}|H_{i+1/2} - H_{i-1/2}| + E, \end{aligned}$$

where

$$\begin{aligned} E = -(z_{-1}^n - z_1^n)_- - \chi(u_1^n, u_2^n)\int_{u_1^n}^{u_2^n} f'_-(\theta)d\theta + \chi(u_{-2}^n, u_{-1}^n)\int_{u_{-2}^n}^{u_{-1}^n} g'_+(\theta)d\theta \\ - |H_{-1/2} - H_{-3/2}| - |H_{3/2} - H_{1/2}|. \end{aligned}$$

From Lemma 4.6, $E \le 0$; hence from Lemma 4.4

$$\begin{aligned} TV(z^n) &= \sum_{i\ne 0,-1}|z_i^n - z_{i+1}^n| + |z_{-1}^n - z_1^n| \le 2\sum|H_{i+1/2} - H_{i-1/2}| \\ &= \frac{2}{\lambda}\sum_{i\ne 0}|u_i^{n+1} - u_i^n| \le \frac{2}{\lambda}\sum_{i\ne 0}|u_i^1 - u_i^0| = 2N_h(f, g, u_0). \end{aligned}$$

This proves (4.15).

Without loss of generality assume that $n \geq m$; then from Lemma 4.4 we have

$$\begin{aligned}\sum_{i\neq 0}|z_i^n - z_i^m| &= \sum_{i\leq -1}|z_i^n - z_i^m| + \sum_{i\geq 1}|z_i^n - z_i^m| \leq L\sum_{i\neq 0}|u_i^n - u_i^m| \\ &\leq L\sum_{i\neq 0}\sum_{j=0}^{n-m+1}|u_i^{n-j} - u_i^{n-j-1}| \\ &\leq L|n-m|\sum_{i\neq 0}|u_i^1 - u_i^0| = \lambda L|n-m|N_h(f,g,u_0).\end{aligned}$$

This proves (4.16) and hence Lemma 4.7.

The following lemma is the analogue of Lemma 4.7 in terms of functions instead of point values.

LEMMA 4.8. *Let $u_0, v_0 \in L^\infty(\mathbb{R},[s,S])$ such that $N(f,g,u_0) < \infty, N(f,g,v_0) < \infty$ are initial datas, and let u_h and v_h be the corresponding solutions obtained by the finite volume scheme* (3.6) *and defined as in* (3.7). *Let $\{z_i^n\}$ defined as in* (4.14) *for u_0 and z_h be the corresponding function defined as in* (3.7). *Then*

$$s \leq u_h(x,t) \leq S \qquad \forall\ (x,t) \in \mathbb{R}\times\mathbb{R}_+, \tag{4.17}$$

$$||z_h||_\infty \leq L, \qquad TV(z_h(\cdot,t)) \leq 2N_h(f,g,u_0), \tag{4.18}$$

$$\int_{\mathbb{R}}|u_h(x,t) - u_h(x,\tau)|dx \leq N_h(f,g,u_0)(2\Delta t + |t-\tau|), \tag{4.19}$$

$$\int_{\mathbb{R}}|z_h(x,t) - z_h(x,\tau)|dx \leq LN_h(f,g,u_0)(2\Delta t + |t-\tau|). \tag{4.20}$$

Moreover, for $a \leq b$ and $\tau < t$,

$$\int_a^b |u_h(x,t) - v_h(x,t)|dx \leq \int_{a-\frac{1}{\lambda}(t-\tau)}^{b+\frac{1}{\lambda}(t-\tau)} |u_h(x,\tau) - v_h(x,\tau)|dx + 4(S-s)h. \tag{4.21}$$

Proof. Inequalities (4.17) and (4.18) follow from (4.2) and (4.15). For inequality (4.19) let $t_n \leq t < t_{n+1}$ and $t_m \leq \tau < t_{m+1}$ so that

$$|n-m|\Delta t = |t_n - t_m| \leq |t_n - t| + |t-\tau| + |\tau - t_m| \leq 2\Delta t + |t-\tau|.$$

Hence from Lemma 4.2 we obtain

$$\begin{aligned}\int_{\mathbb{R}}|u_h(x,t) - u_h(x,\tau)|dx &= h\sum_{i\neq 0}|u_i^n - u_i^m| \leq h\sum_{i\neq 0}\sum_{j=0}^{n-m+1}|u_i^{n-j} - u_i^{n-j-1}| \\ &\leq h|n-m|\sum_{i\neq 0}|u_i^1 - u_i^0| \leq \frac{\Delta t|n-m|}{\lambda}\sum_{i\neq 0}|u_i^1 - u_i^0| \\ &\leq (2\Delta t + |t-\tau|)N_h(f,g,u_0).\end{aligned}$$

The proof of (4.20) follows from (4.19):

$$\begin{aligned}\int_{\mathbb{R}}|z_h(x,t) - z_h(x,\tau)|dx &\leq h\sum_{i\neq 0}|z_i^n - z_i^m| \leq Lh\sum_{i\neq 0}|u_i^n - u_i^m| \\ &\leq LN_h(f,g,u_0)(2\Delta t + |t-\tau|).\end{aligned}$$

We prove inequality (4.21) for $a < 0,\ b > 0$. The proofs are similar for the other cases. Let

$$\begin{aligned}&x_{i_0-3/2} < a \leq x_{i_0-1/2}, \qquad && x_{j_0+1/2} \leq b < x_{j_0+3/2}, \\ &t_{n+1} \leq t < t_{n+2}, && t_{n-p+1} \leq \tau < t_{n-p+2};\end{aligned}$$

so we have $x_{i_0-p-3/2} \le a - ph \le x_{i_0-p-1/2}$, $x_{j_0+p+1/2} \le b + ph < x_{j_0+p+3/2}$, and $t - \Delta t \le \tau + p\Delta t \le t + \Delta t$. From (4.17) $|u_h - v_h| \le (S - s)$; hence

$$\begin{aligned}\int_a^b |u_h(x,t) - v_h(x,t)|dx &= \int_a^{x_{i_0-1/2}} |u_h(x,t) - v_h(x,t)|dx \\ &+ \int_{x_{i_0-1/2}}^{x_{j_0+1/2}} |u_h(x,t) - v_h(x,t)|dx + \int_{x_{j_0+1/2}}^b |u_h(x,t) - v_h(x,t)|dx \\ &\le 2(S-s)h + h \sum_{\substack{i_0 \le i \le j_0 \\ i \ne 0}} |u_i^{n+1} - v_i^{n+1}|.\end{aligned}$$

Using Lemma 4.4 it follows that

$$\begin{aligned}&\int_a^b |u_h(x,t) - v_h(x,t)|dx \\ &\qquad \le 2(S-s)h + h \sum_{\substack{i_0-p \le i \le j_0+p \\ i \ne 0}} |u_i^{n+1-p} - v_i^{n+1-p}| \\ &\qquad = 2(S-s)h + \int_{x_{i_0-p-1/2}}^{x_{j_0+p+1/2}} |u_h(x,\tau) - v_h(x,\tau)|dx \\ &\qquad = 2(S-s)h + \int_{a-ph}^{b+ph} |u_h(x,\tau) - v_h(x,\tau)|dx \\ &\qquad\quad - \int_{a-ph}^{x_{i_0-p-1/2}} |u_h(x,\tau) - v_h(x,\tau)|dx - \int_{x_{j_0+p+1/2}}^{b+ph} |u_h(x,\tau) - v_h(x,\tau)|dx \\ &\qquad \le 2(S-s)h + \int_{a-\frac{t-\tau}{\lambda}}^{b+\frac{t-\tau}{\lambda}} |u_h(x,\tau) - v_h(x,\tau)|dx \\ &\qquad\quad + \int_{a-\frac{t-\tau}{\lambda}}^{a-ph} |u_h(x,\tau) - v_h(x,\tau)|dx + \int_{b+ph}^{b+\frac{t-\tau}{\lambda}} |u_h(x,\tau) - v_h(x,\tau)|dx \\ &\qquad \le 2(S-s)h + 2|\tfrac{t-\tau}{\lambda} - ph|(S-s) + \int_{a-\frac{t-\tau}{\lambda}}^{b+\frac{t-\tau}{\lambda}} |u_h(x,\tau) - u_h(x,\tau)|dx \\ &\qquad \le 4(S-s)h + \int_{a-\frac{t-\tau}{\lambda}}^{b+\frac{t-\tau}{\lambda}} |u_h(x,\tau) - v_h(x,\tau)|dx\,.\end{aligned}$$

This completes the proof of Lemma 4.8.

4.3. Convergence of a subsequence to the weak solution. From hypotheses (H_1) and (H_2) we will construct a solution to the Riemann problem with under-compressive data which will enable us to prove that the solution satisfies the interface entropy condition (2.2). For $\alpha, \beta \in [s, S]$, let

$$v_0(x,\alpha,\beta) = \begin{cases} \alpha & \text{if} \quad x < 0, \\ \beta & \text{if} \quad x \ge 0\,. \end{cases}$$

Then we have the following lemma.

LEMMA 4.9. *Assume that f, g satisfy hypotheses (H_1) and (H_2). Let $\alpha, \beta \in [s, S]$ be such that $\alpha \le \theta_g$ and $\beta \ge \theta_f$. Let $v_h(x,t,\alpha,\beta)$ be the solution given by the finite volume scheme (3.6) with initial data $v_0(x,\alpha,\beta)$ and $\lambda M \le 1$. Assume that for a subsequence $h_k \to 0$, $v_{h_k}(x,t,\alpha,\beta) \to v(x,t,\alpha,\beta)$ on $L^\infty_{loc}(\mathbb{R}_+, L^1_{loc}(\mathbb{R}))$. Then*

$$\begin{aligned}\lim_{x\to 0^-} v(x,t,\alpha,\beta) &= \theta_g \quad \text{if} \quad \min\ g > \min f, \\ \lim_{x\to 0^+} v(x,t,\alpha,\beta) &= \theta_f \quad \text{if} \quad \min\ g \le \min f.\end{aligned} \tag{4.22}$$

Proof. We will prove the lemma for $\min g > \min f$. The other cases can be proved in a similar manner. Let $A \geq \theta_f$ be such that $f(A) = g(\theta_g)$ (see Figure 1). Since $\alpha \leq \theta_g,\ \beta \geq \theta_f$ it follows from (3.2) that

$$\overline{F}(\alpha, \beta) = \max\{g(\theta_g),\ f(\theta_f)\} = g(\theta_g). \tag{4.23}$$

Then if $\{v_i^n\}$ are the grid values corresponding to the initial data $v_0(x, \alpha, \beta)$,

$$v_i^1 = \begin{cases} \alpha - \lambda(g(\theta_g) - g(\alpha)) & \text{if} \quad i = -1, \\ \alpha & \text{if} \quad i \leq -2, \\ \beta - \lambda(f(\beta) - g(\theta_g)) & \text{if} \quad i = 1, \\ \beta & \text{if} \quad i \geq 2. \end{cases}$$

This implies that $v_{-1}^1 = \alpha - \lambda(g(\theta_g) - g(\alpha))$ and $v_{-1}^1 = \alpha + \lambda(g(\alpha) - g(\theta_g)) \leq \alpha + \lambda M|\alpha - \theta_g| < \alpha + \theta_g - \alpha = \theta_g$. Let $\beta \in [\theta_f, A]$; then $f(\beta) \leq f(A) = g(\theta_g)$, and hence $v_1^1 = \beta - \lambda(f(\beta) - g(\theta_g)) \geq \beta$ and $v_1^1 = \beta + \lambda(g(\theta_g) - f(\beta)) = \beta + \lambda(f(A) - f(\beta)) \leq \beta + (A - \beta) = A$. If $\beta \in [A, S]$, then $f(\beta) > f(A) = g(\theta_g)$, and hence $v_1^1 = \beta - \lambda(f(\beta) - g(\theta_g)) < \beta$ and $v_1^1 = \beta - \lambda(f(\beta) - f(A)) \geq \beta - \lambda M(\beta - A) \geq A$. Hence $\{v_i^1\}$ satisfies

$$\begin{aligned} \alpha &\leq v_{-1}^1 \leq \theta_g, \\ \beta &\leq v_1^1 \leq A \quad \text{if} \quad \beta \in [\theta_f, A], \\ A &\leq v_1^1 \leq \beta \quad \text{if} \quad \beta \in [A, S], \\ v_i^1 &= \begin{cases} \alpha & \text{if} \quad i \leq -2, \\ \beta & \text{if} \quad i \geq 2. \end{cases} \end{aligned} \tag{4.24}$$

Now we claim that $\{v_i^n\}$ satisfies

$$\begin{aligned} \alpha &\leq v_{-n}^n \leq v_{-n+1}^n \leq \cdots \leq v_{-1}^n \leq \theta_g, \\ \beta &\leq v_1^n \leq \cdots \leq v_n^n \leq A \quad \text{if} \quad \beta \in [\theta_f, A], \\ A &\leq v_1^n \leq \cdots \leq v_n^n \leq \beta \quad \text{if} \quad \beta \in [A, S], \\ v_i^n &= \begin{cases} \alpha & \text{if} \quad i \leq -n-1, \\ \beta & \text{if} \quad i \geq n+1. \end{cases} \end{aligned} \tag{4.25}$$

From (4.24) the claim is true for $n = 1$. Assume that it is true up to $n - 1$. Since $v_1^{n-1} \geq \theta_f$ and $v_{-1}^{n-1} \leq \theta_g$, hence as in (4.23) $\overline{F}(v_{-1}^n, v_1^n) = g(\theta_g)$. Hence by the same argument as in (4.24), it follows that $\alpha \leq v_{-1}^n \leq \theta_g,\ \beta \leq v_1^n \leq A$ if $\beta \in [\theta_f, A]$ and $A \leq v_1^n \leq \beta$ if $\beta \in [A, S]$. Now (4.25) follows since the scheme is monotone and consistent for $|i| \geq 2$. This proves (4.25).

From (4.25) it follows that $v(x, t, \alpha, \beta)$ satisfies

$$\begin{aligned} \alpha &\leq v^-(t, \alpha, \beta) = \lim_{x \to 0^-} v(x, t, \alpha_1, \beta) \leq \theta_g, \\ \beta &\leq v^+(t, \alpha, \beta) = \lim_{x \to 0^+} v(x, t, \alpha, \beta) \leq A \quad \text{if} \quad \beta \in [\theta_f, A], \\ A &\leq v^+(t, \alpha, \beta) = \lim_{x \to 0^+} v(x, t, \alpha, \beta) \leq \beta \quad \text{if} \quad \beta \in [A, S]. \end{aligned}$$

From (4.25) and hypothesis (H_2) on the shape of f and g we observe that $\{v_i^{(n)}\}_{i \leq -1}$ is independent of β as long as $\beta \geq \theta_f$. Hence $v^-(t, \alpha, \beta)$ is independent of β, and hence $v^-(t, \alpha, \beta) = v^-(t, \alpha, \theta_f)$. Since $v^+(t, \alpha, \theta_f) \leq A$ and $g(v^-(t, \alpha, \beta)) = f(v^+(t, \alpha, \beta))$, hence $v^-(t, \alpha, \beta) = \theta_g$. This completes the proof of Lemma 4.9.

Proof of Theorem 3.2. Let $\lambda = \frac{\Delta t}{h} \le \frac{1}{M}$ be fixed. Since $N(f,g,u_0) < \infty$, then from Lemma 4.8 and by a standard argument there exists a subsequence $h_k \to 0$ such that z_{h_k} converges to z in $L^\infty(0,T,L^1_{loc}(\mathbb{R}))$ and for almost all fixed t, $z_{h_k}(.,t) \to z(.,t)$ in $L^1_{loc}(\mathbb{R})$. Let

$$u(x,t) = \begin{cases} \psi_2^{-1}(z(x,t)) & \text{if} \quad x > 0, \quad t > 0, \\ \psi_1^{-1}(z(x,t)) & \text{if} \quad x < 0, \quad t > 0. \end{cases}$$

Now for $x > 0, u_{h_k}(x,t) = \psi_2^{-1}(z_{h_k}(x,t))$ and for $x < 0, u_{h_k}(x,t) = \psi_1^{-1}(z_{h_k}(x,t))$ and ψ_1 and ψ_2 are continuous, and therefore for almost all t, $u_{h_k}(.,t) \to u(.,t)$ a.e. in $\mathbb{R}$. From (4.18), for a.e. t, $z(.,t) \in BV(\mathbb{R})$, and hence $z(x+,t), z(x-,t)$ exist for all $x \in \mathbb{R}$. This implies that $u(x+,t), u(x-,t)$ exist for all $x \in \mathbb{R}$ and a.e. t. We will complete the proof of the theorem in two steps.

Step 1. Let us prove that u is a weak solution of (1.2) satisfying the interior entropy condition (2.1). Remember that the scheme is not consistent. However, the proof follows almost as in the Lax–Wendroff theorem [10, Theorem 1.1].

Let $\varphi \in C_0^1(\mathbb{R} \times \mathbb{R}_+)$, and let

$$\varphi_j^n = \varphi(x_j, t_n), \quad j \in Z \setminus (0), \quad n \ge 0\,.$$

Multiplying (3.6) by φ_j^n and summing over j and n we obtain

$$h\sum_{n=1}^{\infty}\sum_{i\neq 0} u_i^n(\varphi_i^{n-1} - \varphi_i^n) + \Delta t\sum_{n=0}^{\infty}\sum_{-\infty}^{-1} G(u_{i-1}^n, u_i^n)(\varphi_{i-1}^n - \varphi_i^n)$$
$$+\Delta t\sum_{n=0}^{\infty}\sum_{2}^{\infty} F(u_{i-1}^n, u_i^n)(\varphi_{i-1}^n - \varphi_i^n) + \Delta t\sum_{n=0}^{\infty}\overline{F}(u_{-1}^n, u_1^n)(\varphi_{-1}^n - \varphi_1^n) - h\sum_{i\neq 0} u_i^0\varphi_i^0 = 0.$$

Let

$$\begin{array}{lll} g_h(x,t) = G(u_{i-1}^n, u_i^n), & i \le -1, \quad x \in (x_{i-1}, x_i], & t \in [n\Delta t, (n+1)\Delta t), \\ f_h(x,t) = F(u_{i-1}^n, u_i^n), & i \ge 2, \quad x \in (x_{i-1}, x_i], & t \in [n\Delta t, (n+1)\Delta t), \\ \overline{F}_h(t) = \overline{F}(u_{-1}^n, u_1^n), & t \in [n\Delta t, (n+1)\Delta t), & \\ \overline{\varphi}_h(t) = \varphi(\frac{h}{2}, n\Delta t) - \varphi(-\frac{h}{2}, n\Delta t), & t \in [n\Delta t, (n+1)\Delta t); & \end{array}$$

then the above equalities read as

$$\int_{-\infty}^{\infty}\int_{\Delta t}^{\infty} u_h(x,t)\left(\frac{\varphi_h(x,t) - \varphi_h(x,t-\Delta t)}{\Delta t}\right)dtdx$$
$$+\int_{-\infty}^{x_{-1}}\int_{0}^{\infty} g_h(x,t)\left(\frac{\varphi_h(x+\frac{h}{2},t) - \varphi_h(x-\frac{h}{2},t)}{h}\right)dtdx$$
$$+\int_{x_1}^{\infty}\int_{0}^{\infty} f_h(x,t)\left(\frac{\varphi_h(x+\frac{h}{2},t) - \varphi_h(x-\frac{h}{2},t)}{h}\right)dtdx$$
$$+\int_0^{\infty}\overline{F}_h(t)\overline{\varphi}_h(t)dt + \int_{-\infty}^{\infty} u_h(x)\varphi_h(x)dx = 0\,.$$

Let $h = h_k$ in the above equation, and by going to a subsequence if necessary, by using by the fact that $|\overline{F}(t)| \le \|F\|_\infty + \|G\|_\infty$ and by the dominated convergence theorem, it follows that as $k \to \infty, u_{h_k} \to u$ in $L^\infty_{loc}(\mathbb{R}, L^1_{loc}(\mathbb{R}))$ and the above equation gives that

$$\int_{-\infty}^{\infty}\int_0^{\infty}\left[u\frac{\partial\varphi}{\partial t} + (H(x)f(u) + (1-H(x))g(u))\frac{\partial\varphi}{\partial x}\right]dtdx + \int_{-\infty}^{\infty} u_0(x)\varphi(x,0)dx = 0,$$

where $H(x)$ is the Heaviside function. This proves that u is a weak solution.

In order to prove the interior entropy condition, for $l \in \mathbb{R}$ define

$$A(a,b) = F(a \wedge l,\ b \wedge l) - F(a \vee l,\ b \vee l), \quad B(a,b) = G(a \wedge l,\ b \wedge l) - G(a \vee l,\ b \vee l),$$
$$A^n_{j+1/2} = A(u^n_j, u^n_{j+1}), \quad B^n_{j+1/2} = B(u^n_j,\ u^n_{j+1}). \tag{4.26}$$

Then as in [4, 10], for $|i| \geq 2$, u^n_i satisfies

$$|u^{n+1}_i - l| \leq |u^n_i - l| - \lambda(A^n_{i+1/2} - A^n_{i-1/2}) \quad \text{if } i \geq 2, \tag{4.27}$$

$$|u^{n+1}_i - l| \leq |u^n_i - l| - \lambda(B^n_{i+1/2} - B^n_{i-1/2}) \quad \text{if } i \leq -2. \tag{4.28}$$

Let $0 \leq \varphi \in C^1_0(\mathbb{R}_+ \times \mathbb{R}_+)$. Then there exists $\alpha > 0$ such that supp $(\varphi) \subset \{(x,t); x > \alpha,\ t > \alpha\}$. Hence for h_k small, $\varphi_{h_k}(x,t) = 0$ for $x \leq x_4,\ t \geq \Delta t$. Let $l \in \mathbb{R}$, A, and $A_{i+1/2}$ be defined as in (4.26). Let $A_h(x,t) = A^n_{i+1/2}$ for $x_i \leq x < x_{i+1},\ t_n \leq t < t_{n+1}$. Then multiplying (4.27) by φ^n_i and summing we obtain

$$\int_0^\infty \int_0^\infty |u_{h_k} - l| \left(\frac{\varphi_{h_k}(x,t) - \varphi_{h_k}(x, t - \Delta t}{\Delta t} \right) + \int_0^\infty \int_{x_3}^\infty A_{h_k}(x,t) \left(\frac{\varphi(x + \frac{h_k}{2}, t) - \varphi(x - \frac{h_k}{2}, t)}{h_k} \right) dx dt \geq 0\,.$$

Now letting $h_k \to 0$ yields $\int_0^\infty \int_0^\infty [\,|u-l| \frac{\partial \varphi}{\partial t} + (f(u) - f(l)) sign(u-l) \frac{\partial \varphi}{\partial x}]\, dx\, dt \geq 0$, and similarly for $x < 0$. Hence u satisfies the interior entropy condition (2.1), and this complete the proof of Step 1.

Step 2. We will now show that if u is the weak solution constructed in Step 1 for some $h_k \to 0$, and assuming that the set of discontinuities of u is a discrete set of Lipschitz curves $\{\Gamma_j\}$, then u satisfies the interface entropy condition (2.2), and the solution thus obtained is unique.

The main ingredient to prove this is the choice of the solution constructed in Lemma 4.9. Without loss of generality we can assume that $\min g \geq \min f$. Since $x \to z(x,t)$ is TV bounded, hence $z(0+,t)$ and $z(0-,t)$ exist. This implies that $u^+(t)$ and $u^-(t)$ exist. Suppose that u does not satisfy the interface entropy condition (2.2). Then $meas\,\{L \setminus U\} \neq 0$. Since for $t \in L \setminus U$ $u^+(t) > \theta_f, u^-(t) < \theta_g$, from hypothesis (H_2) we obtain, for almost all $t \in L \setminus U$ $u^-(t) < S, u^+(t) > s$ and

$$meas\left\{t \in L; u^-(t) < S, u^+(t) > s\right\} \neq 0.$$

Hence from the hypothesis on u and (H_2) we can choose $t_0 \in L \setminus U,\ \alpha, \beta, \varepsilon\ \in \mathbb{R}_+$ such that they satisfy

$$t_0 = n_k \Delta t,\ u_{h_k}(x,t_0) \to u(x,t_0) \text{ in } L^1_{loc}(\mathbb{R}) \text{ and } u^+(t_0) > \theta_f,\ u^-(t_0) < \theta_g, \tag{4.29}$$
$$u \text{ is continuous in } [-\beta, 0) \times [t_0 - \alpha, t_0 + \alpha] \text{ and } (0, \beta] \times [t_0 - \alpha, t_0 + \alpha],$$

$$\begin{array}{ll} u^-(t_0) - \varepsilon \leq u(x,t) \leq u^-(t_0) + \varepsilon < \theta_g & \text{in } \ [-\beta, 0) \times [t_0 - \alpha, t_0 + \alpha], \\ \theta_f < u^+(t_0) - \varepsilon \leq u(x,t) \leq u^+(t_0) + \varepsilon & \text{in } \ (0, \beta] \times [t_0 - \alpha,\ t_0 + \alpha]. \end{array} \tag{4.30}$$

On $\mathbb{R} \times \{t_0\}$, define the functions

$$\begin{aligned} V_k(x,t_0) &= u_{h_k}(x,t_0) \\ V_{k,\varepsilon}(x,t_0) &= \begin{cases} u_{h_k}(x,t_0) & \text{if } \ |x| \geq \beta, \\ \max(u_{h_k}(x,t_0), u^-(t_0) - \varepsilon) & \text{if } \ -\beta \leq x \leq 0, \\ \max\left(u_{h_k}(x,t_0), u^+(t_0) - \varepsilon\right) & \text{if } \ 0 \leq x \leq \beta, \end{cases} \\ V_0 &= \begin{cases} \max(s, u^-(t_0) - \varepsilon) & \text{if } \ x \leq 0, \\ u^+(t_0) - \varepsilon & \text{if } \ x > 0. \end{cases} \end{aligned} \tag{4.31}$$

From (4.29) and (4.30) it follows that as $k \to \infty$, for almost every $x \in \mathbb{R}$

$$V_k(x,t_0) \to u(x,t_0), \quad V_{k,\varepsilon}(x,t_0) \to u(x,t_0). \tag{4.32}$$

With t_0 as the initial time and $h_k,\ \lambda = \frac{\Delta t}{h_k}$, as the grid lengths, let $\tilde{V}_{h_k}, \tilde{V}_{h_{k,\varepsilon}}$, and W_{h_k} be the respective solutions calculated with the finite volume scheme (3.6) for $t \geq t_0$ and associated with $V_k, V_{k,\varepsilon}, V_0$ as initial data at $t = t_0$. Since $V_k, V_{k,\varepsilon}$, and V_0 are such that $N(f,g,V_k), N(f,g,V_{k,\varepsilon}), N(f,g,V_0)$ are bounded, one can extract a subsequence still denoted by h_k such that $\tilde{V}_{h_k}, \tilde{V}_{h_{k,\varepsilon}}, W_{h_k}$ converge to u (since $u_{h_k} = \tilde{V}_{h_k}$), v, w a.e., respectively. Letting $h_k \to 0$ in (4.21) for any $a > 0$, $t > t_0$, we have

$$\int_{-a}^{a} |u(x,t) - v(x,t)|dx \leq \int_{-a-t_0/\lambda}^{a+t_0/\lambda} |u(x,t_0) - v(x,t_0)|dx = 0\,.$$

Hence $u \equiv v$.

From (4.31), $V_0(x,t_0) \leq V_{k,\varepsilon}(x,t_0)$ for $x \in [-\beta,\beta]$. Hence by monotonicity of the scheme (see (ii) of Lemma 4.2), $W_{h_k}(x,t) \leq \tilde{V}_{k,\varepsilon}(x,t) + o(\Delta t)$ for $-\beta + \frac{t-t_0}{\lambda} \leq x \leq \beta - \frac{t-t_0}{\lambda}$, and hence for a.e. (x,t) with $t > t_0$, $-\beta + \frac{t-t_0}{\lambda} \leq x \leq \beta - \frac{t-t_0}{\lambda}$,

$$w(x,t) \leq u(x,t).$$

From this inequality, Lemma 4.9, and (4.30) we have for a.e. $t \in (t_0,\ \min(t_0 + \lambda\beta, t_0 - \alpha))$

$$\theta_g = w^-(t) \leq u^-(t) \leq u^-(t_0) + \varepsilon < \theta_g,$$

which is a contradiction. Hence u satisfies the interface entropy conditions.

Let u, v be two limit points of the scheme $\{u_h\}$ such that u and v have a discrete set of Lipschitz curves as discontinuities. From Steps 1 and 2, u and v satisfy the entropy conditions (2.1) and (2.2), and hence from Lemma A.1, for $b > \overline{M}t$,

$$\int_{-b+\overline{M}t,}^{b+\overline{M}t,} |u(x,t) - v(x,t)|dx \leq \int_{-b}^{b} |u(x,0) - v(x,0)|dx = 0\,.$$

Hence $u \equiv v$. This proves Step 2.

Furthermore, let u and v be the weak solutions of (1.2), constructed in Steps 1 and 2 for the initial data u_0 and v_0, respectively. Then by taking $a = -\infty$, $b = +\infty$, and letting $h \to 0$ in (4.21) we obtain

$$\int_{\mathbb{R}} |u(x,t) - v(x,t)|dx \leq \int_{\mathbb{R}} |u_0(x) - v_0(x)|dx\,.$$

Finally, if $f \equiv g$, then by Lemma 4.1 $\overline{F}(a,b) = F(a,b)$, and hence the scheme is Godunov's scheme. Now Theorem 3.2 follows from Steps 1 to 2.

Note that the scheme defined in (3.6) using the interface flux $\overline{F}$ gives a much stronger bound, i.e., $E \leq 0$. This helps us to extend the result for a flux $F(x,u)$ having more discontinuities in the space variable, as stated in the next remark.

Remark 4.10. The above analysis can be extended readily to the equation

$$u_t + f(k(x),u)_x = 0,$$

where $f(a,b) \in C^1(\mathbb{R} \times \mathbb{R})$ and k is a piecewise smooth function satisfying

(i) $f(a,s) = f(b,s), \quad f(a,S) = f(b,S)$ for all $a, b \in \mathbb{R}$,

(ii) for all a the function $u \longrightarrow f(a,u)$ satisfies (H_2).

Remark 4.11. In a forthcoming paper we will extend the above analysis to all E-schemes, including Engquist–Osher, Lax–Friedrich, etc. The case of the upstream mobility is considered in the next section, where it is compared with scheme (3.6).

5. Two-phase flow in porous media. Capillary-free two-phase incompressible flow in a porous medium with a rock type changing at $x = 0$ is modelled by (1.3), (1.4), where u is the saturation of one of the two phases, say phase 1. Equations (1.3) represent conservation of phase 1 inside each rock type, and (1.4) ensures conservation of the same phase at the interface between the two rock types. The functions f and g are the Darcy velocities (divided by the porosity) of phase 1 in each rock type, and they have the form

$$\begin{aligned} f = f_1 &= \frac{1}{\phi}\,\frac{\lambda_1}{\lambda_1+\lambda_2}[q+(c_1-c_2)\lambda_2] \quad \text{for } x>0,\\ g = g_1 &= \frac{1}{\phi}\,\frac{\mu_1}{\mu_1+\mu_2}[q+(c_1-c_2)\mu_2] \quad \text{for } x<0, \end{aligned} \tag{5.1}$$

where ϕ is the porosity of the rock and q, a constant in space, is the total Darcy velocity, that is, the sum of the Darcy velocities of the two phases, $q = \phi(f_1 + f_2) = \phi(g_1 + g_2)$. The Darcy velocities (divided by the porosity) of phase 2 denoted by f_2, g_2 are given by

$$f_2 = \frac{1}{\phi}\,\frac{\lambda_2}{\lambda_1+\lambda_2}[q+(c_2-c_1)\lambda_1], \quad g_2 = \frac{1}{\phi}\,\frac{\mu_2}{\mu_1+\mu_2}[q+(c_2-c_1)\mu_1].$$

The quantities λ_1, μ_1 and λ_2, μ_2 are the effective mobilities of the two phases. They are functions of u satisfying the following properties:

$$\begin{aligned} &\lambda_1, \mu_1 \text{ are increasing functions of } u, \quad \lambda_1(s) = \mu_1(s) = 0,\\ &\lambda_2, \mu_2 \text{ are decreasing functions of } u, \quad \lambda_2(S) = \mu_2(S) = 0. \end{aligned} \tag{5.2}$$

The gravity constants c_1, c_2 of the phases are proportional to their density.

In such a context the flux functions f and g satisfy hypotheses (H_1), (H_2), or (H_3), and Theorems 2.1, and 3.2 apply, provided that an appropriate CFL condition is satisfied. In numerical computations one can, of course, use the numerical fluxes F, G defined in (3.1) inside the rock types and $\overline{F}$, defined in (3.2) at the interface.

However, petroleum engineers have designed, from simple physical considerations, another numerical flux called the upstream mobility flux. It is an ad hoc flux for two-phase flow in porous media which corresponds to an approximate solution to the Riemann problem. It is given by the following formula:

$$\begin{aligned} &F^{UM}(a,b) = \frac{1}{\phi}\,\frac{\lambda_1^*}{\lambda_1^*+\lambda_2^*}[q+(c_1-c_2)\lambda_2^*],\\ &\lambda_\ell^* = \begin{cases} \lambda_\ell(a) & \text{if } q+(c_\ell-c_i)\lambda_\ell^* > 0, \quad i=1,2, \quad i\neq\ell,\\ \lambda_\ell(b) & \text{if } q+(c_\ell-c_i)\lambda_\ell^* \le 0, \quad i=1,2, \quad i\neq\ell, \end{cases} \qquad \ell = 1,2, \end{aligned} \tag{5.3}$$

and similarly for G^{UM} associated with g. As we can see, the flux is calculated using the mobilities of the phases which are upstream with respect to the flow of the phases. When the two phases are flowing in the same direction, the Godunov flux and the upstream mobility flux give the same answer and coincide with standard upstream weighting, but they differ when the phases are flowing in opposite directions. This flux has been shown to have all the desired properties for convergence of the associated finite difference scheme [26, 2] in the case of a flux function which does not vary with space (one rock type).

The generalization of the upstream mobility flux to the case of two rock types is straightforward, and at the interface the corresponding flux is

$$\begin{aligned} &\overline{F}^{UM}(a,b) = \frac{1}{\phi}\,\frac{\lambda_1^*}{\lambda_1^*+\lambda_2^*}[q+(c_1-c_2)\lambda_2^*], \\ &\lambda_\ell^* = \begin{cases} \mu_\ell(a) & \text{if } q+(c_\ell-c_i)\lambda_\ell^* > 0, \quad i=1,2, \quad i\neq\ell, \\ \lambda_\ell(b) & \text{if } q+(c_\ell-c_i)\lambda_\ell^* \le 0, \quad i=1,2, \quad i\neq\ell, \end{cases} \quad \ell = 1,2. \end{aligned} \tag{5.4}$$

This upstream mobility flux at the interface satisfies the consistency condition of Lemma 4.1:

$$\overline{F}^{UM}(s,s) = f(s) = g(s) = 0, \quad \overline{F}^{UM}(S,S) = f(S) = g(S) = \frac{q}{\phi}.$$

6. Numerical experiments. We consider an idealized experiment in which two phases of different densities are flowing in a vertical closed core. This core is made of two rock types, the top part being associated with the flux function g and the bottom part associated with the function f defined in (5.1). The data associated with the problem are as follows:

$$\begin{aligned} &\phi = 1, \quad q = 0, \quad c_1 = 2, \quad c_2 = 1, \\ &s = 0., \quad S = 1., \quad \lambda_1 = 10u^2, \quad \lambda_2 = 20(1-u)^2, \quad \mu_1 = 50u^2, \quad \mu_2 = 5(1-u)^2, \end{aligned}$$

which gives the flux function f and g represented in Figure 3. Note that we are in the case where f and g satisfy hypothesis (H_3). Phase 1 is the heavy phase, and it moves downwards while phase 2, the light phase, moves upward.

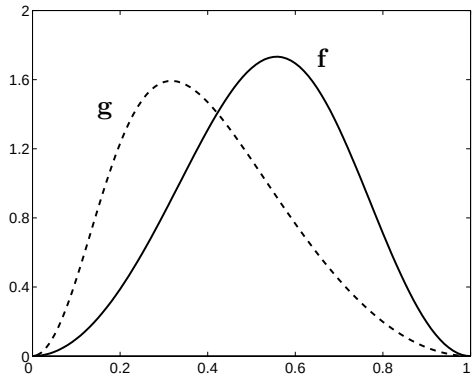


FIG. 3. *The flux functions at the interface for the numerical experiments.*

We present here two simulations which differ by the initial condition. In the first case we start with discontinuous data $u_0(x) = 1$ if $x < 0$, $u_0(x) = 0$ if $x > 0$; that is, at initial time the core is saturated with the heavy fluid (phase 1) in the upper half and with the light fluid (phase 2)in the lower half. The calculated solution is shown in Figure 4. In the second case we start with a constant initial data $u_0 = .5$ which corresponds to a situation where the two phases are "mixed." In this case the solution is shown in Figure 5.

In all figures the top part of the core is on the left of the picture and the bottom part is on the right. As expected, observe as time goes on the heavy fluid moving

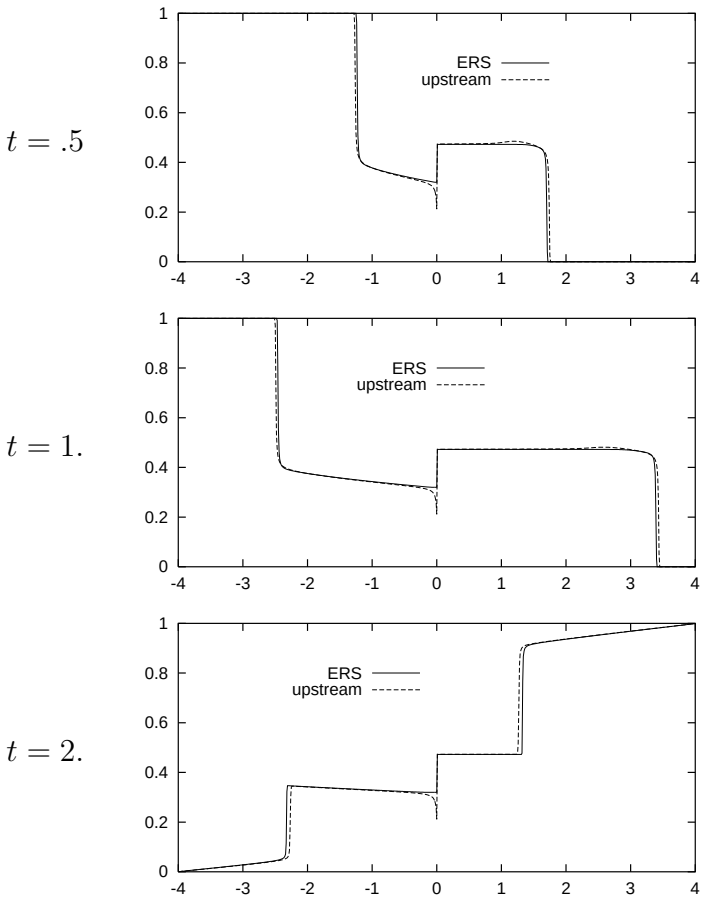


FIG. 4. *Finite difference solutions calculated with numerical flux* (3.1), (3.3) *(ERS) and with the upstream mobility flux* (5.3), (5.4) *at different times for a discontinuous initial data* ($h = 1/100$).

downward which is represented by its saturation u decreasing on the left and increasing on the right. Obviously, in the case of the continuous initial data we reach earlier the stationary state where the heavy phase occupies the bottom half of the core ($u(x) = 0$ if $x < 0, u(x) = 1$ if $x > 0$). However, one can observe the complexity of the solution, which presents several shocks.

In Figures 4 and 5 we compare the finite difference solutions calculated when using the numerical flux based on the exact Riemann solver (ERS) (3.1), (3.3) and the one calculated when using the upstream mobility flux (5.3), (5.4). We can observe that the latter is doing very well even in these complex situations. However, small differences can be seen. In particular, a small boundary layer appears on the left side of the interface. For these numerical examples these differences vanish when $h \to 0$ if they are measured in the L_1 norm.

Finally, in Figure 6 we present the solution given by the numerical flux which was presented in [3, 8, 13, 6] (ERS-NIF) and which is not valid when the flux functions intersect in the undercompressive case, which is our situation. The picture in Figure 6 is to be compared with the bottom picture in Figure 4. As expected, this numerical flux is not able to capture the complexity of the solution.

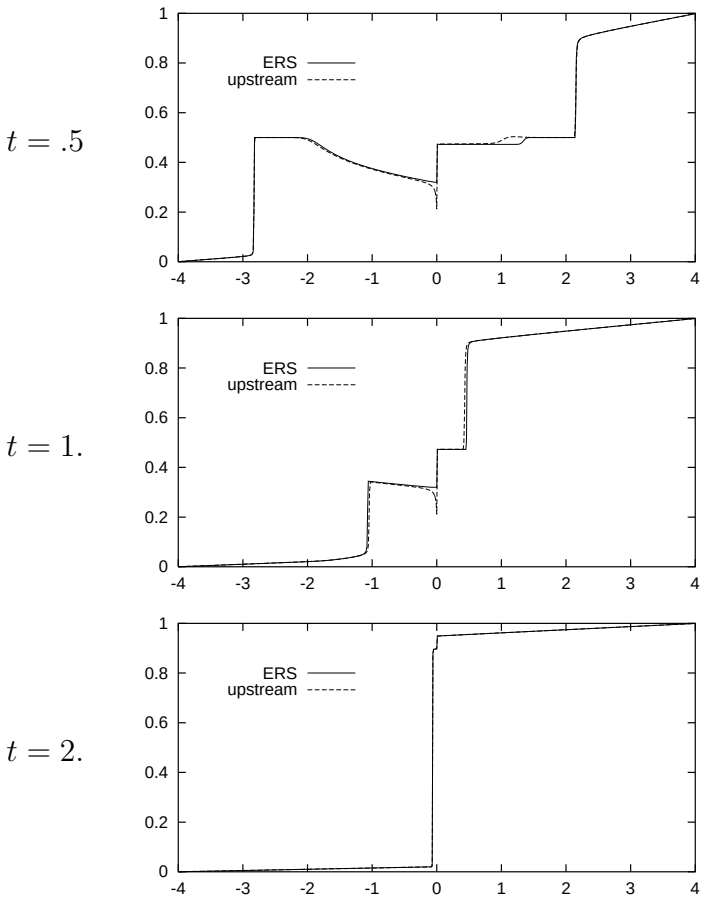


FIG. 5. *Finite difference solutions calculated with numerical flux* (3.1), (3.3) *(ERS) and with the upstream mobility flux* (5.3), (5.4) *at different times for a constant initial data* ($h = 1/100$).

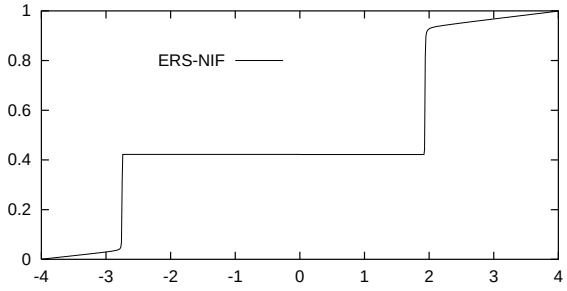


FIG. 6. *Finite difference solution calculated when using the numerical flux for nonintersecting fluxes.*

7. Conclusion. The calculation of the solutions of conservation laws with a flux function discontinuous in space needs appropriate numerical methods. We presented a Godunov method which uses an exact Riemann solver, and we proved convergence of the corresponding numerical scheme. We compared numerically with the upstream

mobility numerical flux used for multiphase flow in porous media, showing that the latter still works well in the case of a discontinuous flux function. A consequence of the proof of the convergence of the numerical scheme is an existence and uniqueness of the solution to the continuous problem.

Appendix A. End of the proof of existence and uniqueness theorem, Theorem 2.1. In this appendix we terminate the proof of Theorem 2.1 for nonconvex functions as in Figure 1. Existence was a consequence of the convergence theorem, Theorem 3.2, and to prove uniqueness we need to show that all solutions of (1.2) satisfying entropy conditions (2.1) and (2.2) can be represented by an L^1-contractive semigroup. The proof is as in [1], so we sketch only the proof. The main idea of this proof goes back to Kruzkov [19].

LEMMA A.1. *Let $u, v \in L^\infty(\mathbb{R}\times\mathbb{R}_+)$ with $s \le u, v \le S$ be two solutions of* (1.2) *with initial data $u_0, v_0 \in L^\infty(\mathbb{R})$, respectively. Assume the following:*

(i) *For almost every t, $u(x+,t), v(x+,t), u(x-,t)$, and $v(x-,t)$ exist.*

(ii) *The set of discontinuities of u and v is a discrete set $\{\Gamma_j\}_{j\in N}$ of Lipschitz curves.*

(iii) *u and v satisfies the entropy conditions* (2.1) *and* (2.2).

Then for any $\overline{M} \ge M$, $a < 0, b > 0$, $b - a \ge 2\overline{M}t$ the function

$$t \mapsto \int_{a+\overline{M}t}^{b-\overline{M}t} |u(x,t) - v(x,t)|dx$$

is nonincreasing.

Proof. The first three steps are exactly as in Kruzkov's proof (see [12, p. 24]), and the interface entropy condition (2.2) is used to prove Step 4.

Step 1. Let $l \in \mathbb{R}$, $\varphi_l(\theta) = |\theta - l|$, $\tilde{f}(\theta, l) = (f(\theta) - f(l))\ \mathrm{sign}\,(\theta - l)$. Let $0 \le \rho \in C_0^\infty(\mathbb{R}\times\mathbb{R}_+)$. Let $\Gamma_j^+ = \Gamma_j \cap \{(x,t) : x > 0, t > 0\}$ and $\nu^j = (\nu_1^j, \nu_2^j)$ be the a.e. normal to Γ_j^+. Then by integration by parts and using the interior entropy condition (2.1) we obtain

$$\begin{aligned}\int_0^\infty \int_0^\infty (\varphi_l(u(x,t))\frac{\partial\rho}{\partial t} + \tilde{f}\,(u(x,t),l)\,\frac{\partial\rho}{\partial x}dxdt \\ = \sum_{j=1}^\infty \int_{\Gamma_j^+} ([\varphi_l(u)]\nu_1^j + [\tilde{f}(u,l)]\nu_2^j)\rho d\sigma - \int_0^\infty \tilde{f}(u^+(t),l)\rho(0,t)dt \\ \ge -\int_0^\infty \tilde{f}(u^+(t),l)\rho(0,t)dt, \qquad \text{(A.1)}\end{aligned}$$

where $[\varphi_l(u)] = \varphi_l(u^-) - \varphi_l(u^+)$, the jump across of Γ_j^+, $[\tilde{f}(u,l)] = \tilde{f}(u^-,l) - \tilde{f}(u^+,l)$, the jump across of Γ_j^+, and $u^+(t) = u(0+,t)$.

Step 2. Let $A(x,t,y,s) = \frac{f(u(x,t))-f(v(y,s))}{u(x,t)-v(y,s)}$, $\alpha \in C_0^1((-1,0)\times(-1,0))$ with $\int_{\mathbb{R}^2}\alpha(z)dz = 1$ and $\beta \in C_0^1(\overline{\mathbb{R}}_+\times\mathbb{R}_+)$. Let $\varepsilon_1 > 0, \varepsilon_2 > 0$, and define

$$\rho_\varepsilon(x,t,y,\tau) = \frac{1}{\varepsilon_1\varepsilon_2}\alpha\left(\frac{x-y}{\varepsilon_1}, \frac{t-\tau}{\varepsilon_2}\right)\beta(y,s).$$

Now taking $l = v(y,\tau)$ and $\rho = \rho_\varepsilon(x,t,y,\tau)$ in (A.1) we integrate with respect to $(y,\tau) \in \mathbb{R}_+\times\mathbb{R}_+$. Then using symmetry and letting $\varepsilon_1 \to 0$, $\varepsilon_2 \to 0$ we obtain

$$\int_0^\infty\int_0^\infty |u(x,t)-v(x,t)|\left\{\frac{\partial\beta}{\partial t} + A(x,t,x,t)\frac{\partial\beta}{\partial x}\right\}dxdt \ge -\int_0^\infty \tilde{f}(u^+(t),v^+(t))\beta(0,t)dt.$$

Step 3. Let $b \geq 0$ and χ_ε be a decreasing smooth function in $(0,\infty)$ converging to $\chi_{[0,b]}$ as $\varepsilon \to 0$. Let $0 \leq \varphi \in C_0^1(\mathbb{R}_+)$, and let $\beta(x,t) = \chi_\varepsilon(|x| + \overline{M}t)\varphi(t)$ in the above equation. Letting $\varepsilon \to 0$ we can write

$$\int_0^\infty \varphi'(t) \int_0^{b-\overline{M}t} |u(x,t) - v(x,t)|dt \geq -\int_0^{b/\overline{M}} \tilde{f}(u^+(t), v^+(t))\varphi(t)dt\,.$$

Similarly for $x \leq 0$,

$$\int_0^\infty \varphi_1(t) \int_{a+\overline{M}t}^0 |u(x,t) - v(x,t)|dt \geq \int_0^{b/\overline{M}} \tilde{g}(u^-(t), v^-(t))\varphi(t)dt.$$

Adding both inequalities we obtain

$$\int_0^\infty \varphi'(t) \int_{a+\overline{M}t}^{b-\overline{M}t} |u(x,t) - v(x,t)|dx \geq \int_0^{b/\overline{M}} (\tilde{g}(u^-(t), v^-(t)) - \tilde{f}(u^+(t), v^+(t))\varphi(t))dt.$$

Step 4. So far, all the above steps are standard, and now we will make use of the interface entropy condition (2.2) to prove Lemma A.1. In order to prove the lemma it is sufficient to show that for almost all $t, I(t) \geq 0$, where

$$I(t) = \tilde{g}(u^-(t), v^-(t)) - \tilde{f}(u^+(t), v^+(t)) = |u^-(t) - v^-(t)|\frac{g(u^-(t)) - g(v^-(t))}{u^-(t) - v^-(t)} - |u^+(t) - v^+(t)|\frac{f(u^+(t)) - f(v^+(t))}{u^+(t) - v^+(t)}.$$

Without loss of generality, we can assume that $u^+(t) > v^+(t)$. If $f(u^+(t)) \leq f(v^+(t))$, then $I(t) \geq 0$. Hence let $f(u^+(t)) > f(v^+(t))$. Since $u^+(t) > v^+(t)$, from hypothesis (H_2) we have $u^+(t) \in (\theta_f, S]$. From the interface entropy condition (2.2) either $u^+(t) = u^-(t) = S$ or $u^-(t) \in (\theta_g, S]$. In the first case, $I(t) = 0$. In the latter case from the Rankine–Hugoniot condition, $g(u^-(t)) > g(v^-(t))$ and from hypothesis (H_2) $u^-(t) > v^-(t)$, and hence $I(t) = 0$. This completes the proof of (A.2) and of Lemma A.1.

Lemma A.1 implies that

$$\int_{a+\overline{M}t}^{b-\overline{M}t} |u(x,t) - v(x,t)|dx \leq \int_a^b |u(x,0) - v(x,0)|dx.$$

Letting $a \to -\infty$, $b \to +\infty$ we obtain the L^1 contractivity and terminate the proof of Theorem 2.1.

Acknowledgment. We would like to thank the referees for their thorough review of our paper and their useful constructive remarks.

REFERENCES

[1] Adimurthi and G. D. Veerappa Gowda, *Conservation law with discontinuous flux*, J. Math. Kyoto Univ., 43 (2003), to appear.

[2] Y. Brenier and J. Jaffré, *Upstream differencing for multiphase flow in reservoir simulation*, SIAM J. Numer. Anal., 28 (1991), pp. 685–696.

[3] G. Chavent, G. Cohen, and J. Jaffré, *A finite element simulator for incompressible two-phase flow*, Transp. Porous Media, 2 (1987), pp. 465–478.

[4] M. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[5] S. Diehl, *Conservation Laws with Applications to Continuous Sedimentation*, Doctoral Dissertation, Lund University, Lund, Sweden, 1995.

[6] S. Diehl, *On scalar conservation laws with point source and discontinuous flux functions*, SIAM J. Math. Anal, 26 (1995), pp. 1425–1451.

[7] S. Diehl, *A conservation law with point source and discontinuous flux function modelling continuous sedimentation*, SIAM J. Appl. Math., 56 (1996), pp. 388–419.

[8] T. Gimse and N. H. Risebro, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635–648.

[9] T. Gimse and N. Risebro, *Riemann problems with a discontinuous flux function*, in Third International Conference on Hyperbolic Problems. Theory, Numerical Methods and Applications, B. Engquist and B. Gustafsson, eds., Studentlitteratur, Chartwell-Bratt, Lund-Bromley, Sweden, 1991, pp. 488–502.

[10] E. Godlewski and P.-A. Raviart, *Hyperbolic Systems of Conservation Laws*, Mathematiques et Applications, Ellipses, Paris, 1991.

[11] A. Harten, J. M. Hyman, and P. Lax, *On finite difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math., 29 (1976), pp. 297–322.

[12] L. Hormander, *Lectures on Nonlinear Hyperbolic Differential Equations*, Math. Appl. 26, Springer, Berlin, 1997.

[13] J. Jaffré, *Flux calculation at the interface between two rock types for two-phase flow in porous media*, Transp. Porous Media, 21 (1995), pp. 195–207.

[14] J. Jaffré, *Numerical calculation of the flux across an interface between two rock types of a porous medium for a two-phase flow*, in Hyperbolic Problems: Theory, Numerics, Applications, J. Glimm, M. Graham, J. Grove, and B. Plohr, eds., World Scientific, Singapore, 1996, pp. 165–177.

[15] E. Kaasschieter, *Solving the Buckley-Leverett equation with gravity in a heterogeneous porous medium*, Comput. Geosci., 3 (1999), pp. 23–48.

[16] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *On a nonlinear degenerate parabolic transport-diffusion equations with a discontinuous coefficient*, Electron. J. Differential Equations, 93 (2002), pp. 1–23.

[17] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Upwind difference approximations for degenerate parabolic convection-diffusion equations with a discontinuous coefficient*, IMA J. Numer. Anal., 22 (2002), pp. 623–664.

[18] Q. B. Keyfitz, *Solutions with shocks: An example of an L_1-contractive semigroup*, Comm. Pure Appl. Math., 24 (1971), pp. 125–132.

[19] S. N. Kruzkov, *First order quasilinear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[20] H. Langtangen, A. Tveito, and R. Winther, *Instability of Buckley-Leverett flow in heterogeneous media*, Transp. Porous Media, 9 (1992), pp. 165–185.

[21] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[22] L. Lin, J. B. Temple, and J. Wang, *A comprison of convergence rates for Godunov's method and Glimm's method in resonant nonlinear systems of conservation laws*, SIAM J. Numer. Anal., 32 (1995), pp. 824–840.

[23] L. Longwei, B. Temple, and W. Jinghua, *Suppression of oscillations in Godunov's method for resonant non-strictly hyperbolic system*, SIAM J. Numer. Anal., 32 (1995), pp. 841–864.

[24] S. Mochen, *An analysis for the traffic on highways with changing surface conditions*, Math. Model., 9 (1987), pp. 1–11.

[25] O. A. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Uspekhi Mat. Nauk, 12 (1957), pp. 3–73.

[26] P. Sammon, *An analysis of upstream differencing*, SPE Reservoir Engrg., 3 (1988), pp. 1053–1056.

[27] B. Temple, *Global solution of the Cauchy problem for a class of* 2×2 *nonstrictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.
[28] J. D. Towers, *Convergence of a difference scheme for conservation laws with a discontinuous flux*, SIAM J. Numer. Anal., 38 (2000), pp. 681–698.
[29] J. D. Towers, *A difference scheme for conservation laws with a discontinuous flux: The nonconvex case*, SIAM J. Numer. Anal., 39 (2001), pp. 1197–1218.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 209–227

ASYMMETRIC CUBATURE FORMULAE WITH FEW POINTS IN HIGH DIMENSION FOR SYMMETRIC MEASURES*

NICOLAS VICTOIR†

Abstract. Let μ be a positive measure on $\mathbb{R}^d$ invariant under the group of reflections and permutations, and let m be a natural number. We describe a method to construct cubature formulae of degree m with respect to μ, with n positive weights and n points in the support of μ and such that n grows at most like d^m with the dimension d. We apply this method to classical measures to explicitly construct cubature formulae of degree 5 with the number of points growing at most like d^3.

Key words. cubature formulae, orthogonal array, t-design

AMS subject classifications. 65D32, 05B15, 51E05

DOI. 10.1137/S0036142902407952

1. Introduction.

We will denote by $\mathbb{R}[X_1, \dots, X_d]$ the space of polynomials in d variables with real coefficients and by $\mathbb{R}_m[X_1, \dots, X_d]$ its subspace made of polynomials of total degree less than or equal to m. Note that $\dim \mathbb{R}_m[X_1, \dots, X_d] = \binom{m+d}{d}$. We will write $\delta_{\mathbf{x}}$ for the Dirac probability at the point $\mathbf{x}$.

DEFINITION 1.1. *Let μ be a positive measure on $\mathbb{R}^d$, $d \geq 1$, and m be a positive integer. We say that the points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ and the weights $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ define a cubature formula of degree m, with respect to the measure μ, if*

$$\int_{\mathbb{R}^d} P(z)\mu(dz) = \sum_{k=1}^{n} \lambda_k P(\mathbf{x}_k) \tag{1}$$

holds for all polynomials P in $\mathbb{R}_m[X_1, \dots, X_d]$.

When $d = 1$ people use the term *quadrature* in place of *cubature*.

Remark 1.2. The existence of a cubature formula of degree m is equivalent to the existence of a finite measure ξ (a measure of the form $\xi = \sum_{i=1}^{n} \lambda_i \delta_{\mathbf{x}_i}$) such that, for all polynomials P of degree less than or equal to m, the integral of P with respect to ξ is equal to the integral of P with respect to μ.

Once we have a cubature formula, using the notation of the previous definition, we can approximate the integral of a smooth function f with respect to μ by

$$\sum_{i=1}^{n} \lambda_i f(\mathbf{x}_i).$$

The following theorem was first published by Tchakaloff (in the special case of a compactly supported measure). See [25], [28], [29] for its proof.

THEOREM 1.3. *Let d and m be positive integers and let μ be a positive measure on $\mathbb{R}^d$ with the property that $\int |P(z)|\mu(dz) < \infty$ for all $P \in \mathbb{R}_m[X_1, \dots, X_d]$. Then we can find n points $\mathbf{x}_1, \dots, \mathbf{x}_n$ in the support of μ and n positive real numbers $\lambda_1, \dots, \lambda_n$, with*

$$n \leq \dim \mathbb{R}_m[X_1, \dots, X_d],$$

*Received by the editors May 26, 2002; accepted for publication (in revised form) March 29, 2003; published electronically January 28, 2004.

http://www.siam.org/journals/sinum/42-1/40795.html

†Mathematical Institute, Oxford University, 24-29 St. Giles, Oxford OX1 3LB, UK (victoir@maths.ox.ac.uk).

such that the cubature relation (1) *holds for all* $P \in \mathbb{R}_m[X_1, \dots, X_d]$.

Unfortunately, this is only an existence theorem and does not provide any efficient method to obtain such cubature formulae. Hundreds of papers are devoted to the construction of cubature formulae. See the books [9], [18], [23], [28], the papers [5], [6], [7], [8], and the references therein. In [1], some Gaussian cubature formulae (i.e., cubature formulae with a minimum number of points) were constructed in high dimension. However, the measures considered there are rather artificial and exotic. It seems that no one has constructed explicit cubature formulae for classical measures of degree $m \geq 4$ in high dimension (i.e., in any given dimension) with no more than $\dim \mathbb{R}_m[X_1, \dots, X_d] = \binom{m+d}{m}$ points, despite Tchakaloff's theorem. Stroud, in the introduction of his celebrated book *Approximate Calculation of Multiple Integrals* [28], explains the need to construct such formulae. There exist some cubature formulae of degree 3 with few points, but the degree is not high enough to make the approximation accurate. Few points means, in this paper, that the number of points grows polynomially with the dimension.

Paraphrasing [18], an ideal cubature formula with respect to a positive measure should have points within the domain of integration, as few points as possible (in particular, fewer points than the dimension of the polynomial space $\mathbb{R}_m[X_1, \dots, X_d]$), and positive weights. Also, cubature formulae of higher degree will provide more accurate approximations of integrals. Nonetheless, the "space of cubature formulae is not totally ordered." By this, we mean that a cubature formula of degree 7 with some negative weights and another one of degree 5 with positive weights and with the same number of points cannot really be compared. Indeed, it is easy to find some functions whose integral will be better approximated by the first method and some other functions whose integral will be better approximated by the second method.

In this paper, we describe a method to construct cubature formulae with few points in any given dimension with respect to measures which are invariant under reflection and permutation of the axes. This method works particularly well for cubature formulae of degree 3 (where we get formulae with $O(d)$ points) and 5 (where we get formulae with $O(d^3)$ and even sometimes $O(d^2)$, which is well under the Tchakaloff bound), but we have not yet applied it for higher degrees.

In the next section, we will explain how combinatorics and coding theory help to construct cubature formulae with respect to the finite measure

$$\frac{1}{2^d} \sum_{\mathbf{g} \in \{-1,1\}^d} \delta_{\mathbf{g}}.$$

Formulae of degree $2m+1$ in d dimensions will require only $O(d^m)$ points. Some other combinatorial objects will allow us to construct cubature formulae with few points for measures of the form

$$\frac{1}{|\mathcal{S}_d.\mathbf{x}|} \sum_{\mathbf{y} \in \mathcal{S}_d.\mathbf{x}} \delta_{\mathbf{y}},$$

where we let $\mathcal{S}_d$, the group of permutation of order d, act naturally on $\mathbb{R}^d$. We also let $\mathcal{G}_d \simeq (\{-1,+1\}^d, *)$, the group of reflections of the axes, and $\mathcal{GS}_d$, the group of permutations and reflections of the axes, act naturally on $\mathbb{R}^d$. We will show that for a $\mathcal{GS}_d$-invariant measure μ, there exists a $\mathcal{GS}_d$-invariant cubature formula of degree m,

i.e., a formula of the form

$$\int P(z)\mu(dz) = \sum_{i=1}^{k} \lambda_i \frac{1}{|\mathcal{GS}_d.\mathbf{x}_i|} \sum_{\mathbf{y}\in\mathcal{GS}_d.\mathbf{x}_i} P(\mathbf{y}) \quad \forall P \text{ in } \mathbb{R}_m[X_1,\dots,X_d].$$

In this formula, we can find the points $\mathbf{x}_1,\dots,\mathbf{x}_k$ with the number of points k bounded by a term which depends only on m, the degree of the cubature formula. For example, for a cubature formula of degree 5, this upper bound is equal to 4, and it is possible to find two points that give the formula. Then, using the techniques described in the next two sections, we will construct cubature formulae with respect to $\frac{1}{|\mathcal{GS}_d.\mathbf{x}|}\sum_{\mathbf{y}\in\mathcal{GS}_d.\mathbf{x}}\delta_{\mathbf{y}}$. Those new approximations will give us a cubature formula of degree m with respect to μ and with few points. This explains why we first constructed some cubature formulae with respect to finite measures—that is, why we constructed finite measures approximating some other finite measures! The last section will deal with concrete examples of construction of such cubature formulae of degree 5 with respect to classical measures, such as the Gaussian measure and the Lebesgue measure on the unit hypercube, on the surface of the unit sphere, and on the whole unit sphere. We also include a table of formulae similar to those found in [7], [8].

2. Codes and orthogonal arrays.

2.1. Definitions and link with cubature formulae. In this section, we describe how we can find a cubature formula of degree $2m+1$ with respect to the measure of total mass one which puts equal mass at all of the vertices of a d-dimensional hypercube, i.e., with respect to

$$\delta_{\mathcal{G}_d} = \frac{1}{2^d}\sum_{\mathbf{g}\in\{-1,1\}^d}\delta_{\mathbf{g}}.$$

We will need to define orthogonal arrays. See [14] for a full description of these combinatorial objects.

DEFINITION 2.1. *An $N\times k$ array A with entries from a set S is said to be an orthogonal array with $|S|$ levels, strength t, and index λ (for some t in the range $0\le t\le k$) if every $N\times t$ subarray of A contains each t-uple based on S exactly λ times as a row.*

Such an array will be denoted by $OA(N,k,|S|,t)$. We do not put λ explicitly in this notation, as it is quite easy to see that $\lambda = N/|S|^t$.

THEOREM 2.2. *The points defined to be the rows of an orthogonal array with parameters $OA(N,d,2,2m+1)$ and with entries in $\{-1,1\}$, associated with the constant weight $1/N$, define a cubature formula of degree $2m+1$ with respect to $\delta_{\mathcal{G}_d}$.*

Proof. Let $\mathbf{x}_1,\dots,\mathbf{x}_N$ be the rows of an $OA(N,d,2,2m+1)$ with entries in $\{-1,1\}$. Let P_1 be the square of a monomial with leading coefficient 1. Then, using the particular form of P_1,

$$\sum_{k=1}^{N}\frac{1}{N}P_1(\mathbf{x}_k) = \sum_{k=1}^{N}\frac{1}{N}1 = \int P_1(z)\delta_{\mathcal{G}_d}(dz).$$

Now let $P_2 = Q^2R$, where $R = X_1^{\alpha_1}\dots X_d^{\alpha_d}$, $\alpha_i\in\{0,1\}$ with $1\le\sum_{i=1}^d\alpha_i\le 2m+1$, and Q is any monomial. Then

$$\sum_{k=1}^{N}\frac{1}{N}P_2(\mathbf{x}_k) = \sum_{k=1}^{N}\frac{1}{N}R(\mathbf{x}_k) = \sum_{k=1}^{N}\frac{1}{N}\prod_{i:\ \alpha_i=1}x_k^i.$$

If λ is the index of the orthogonal array $OA(N, d, 2, 2m+1)$, then for any functional Υ, $l \leq 2m+1$ and different indices $i_1, \dots, i_l$,

$$\sum_{k=1}^{N} \Upsilon(x_k^{i_1}, \dots, x_k^{i_l}) = \lambda \sum_{(y^1,\dots,y^l)\in\{-1,1\}^l} \Upsilon(y^1, \dots, y^l).$$

Hence, if $l = \sum_{i=1}^{d} \alpha_i$,

$$\begin{aligned}\sum_{k=1}^{N} \frac{1}{N} \prod_{i:\ \alpha_i=1} x_k^i &= \frac{\lambda}{N} \sum_{(y^1,\dots,y^l)\in\{-1,1\}^l} \prod_{j=1}^{l} y^j \\ &= 0 = \int P(z)\delta_{\mathcal{G}_d}(dz).\end{aligned}$$

We have proved that, for all monomials of degree less than or equal to $2m+1$,

$$\sum_{k=1}^{N} \frac{1}{N} P(\mathbf{x}_k) = \int P(z)\delta_{\mathcal{G}_d}(dz). \qquad \square$$

We are now going to link orthogonal arrays and codes. Let us first give the definition of a code.

DEFINITION 2.3. *Let S be a set of symbols of size s. A code is a collection C of vectors in S^k. These vectors are called codewords. The distance d_C of the code is defined by*

$$d_C = \min_{u,v\in C, u\neq v} card\,\{j \in \{1, \dots, k\}, u^j \neq v^j\}.$$

Such a code will be denoted by $(k, card\ C, d_C)_s$.

DEFINITION 2.4. *Let C be a $(k, N, d_C)_s$ code on S, and assume that S is a finite field. Define the dual code of C by*

$$C^{\perp} = \left\{ v \in S^k, \forall u \in C, uv^{\perp} = \sum_{i=1}^{k} u_i v_i = 0 \right\}.$$

Its distance $d_{C^{\perp}}$ will be called the dual distance of the code C.

The following is due to Delsarte [10].

THEOREM 2.5. *Let A be the $N \times k$ array such that its rows are the codewords of a $(k, N, d)_s$ code over a finite field S, with dual distance $d^{\perp}$. Then A is an $OA(N, k, s, d^{\perp} - 1)$.*

Thus we see that the codewords of a code with parameters $(k, N, d_C)_2$ over $\mathbb{Z}/2\mathbb{Z}$, with dual distance $2m+2$, give us a cubature formula of degree $2m+1$ with respect to $\delta_{\mathcal{G}_d}$. We will now give some examples of such orthogonal arrays.

2.2. Degree 3. We are going to describe the orthogonal arrays with parameters $OA(N, d, 2, 3)$, preferably with N as low as possible. This is closely related to Hadamard matrices. Write I_n for the $n \times n$ identity matrix and A^T for the transpose of a matrix A.

DEFINITION 2.6. *A Hadamard matrix of degree n is a matrix $H \in M_n(\{-1, 1\})$ such that $HH^T = nI_n$.*

PROPOSITION 2.7. *If there exists a Hadamard matrix of degree n, then $n \in \{1, 2\}$ or 4 divides n.*

We can construct an $OA(N,d,2,3)$ in the following way. Consider the least n greater than or equal to d, for which we have a Hadamard matrix A of degree n. Then consider the $2n \times n$ matrix B, such that the rows of B are made of the rows of A and of $-A$. Then delete $n-d$ columns of B. That gives us an $OA(2n,d,2,3)$. The Hadamard matrices have been heavily studied. Hadamard conjectured that if $n \in \{1,2\}$ or if 4 divides n, then a Hadamard matrix exists. It still has not been proved or disproved. Such matrices have been constructed for all $n \leq 1000$, except for $n = 428,668,716,764,892$. Moreover, there exists an easy way to construct them when n is a power of 2.

DEFINITION 2.8. *Let $A = (a_{ij})_{1\leq i,j\leq n} \in M_n(\mathbb{R})$ and $B \in M_m(\mathbb{R})$. Then we define the tensor product of A and B, $A \otimes B$ by the $nm \times nm$ matrix*

$$\begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \dots & a_{nn}B \end{pmatrix}.$$

PROPOSITION 2.9. *Let H_a and H_b be Hadamard matrices of degree a and b. Then $H_a \otimes H_b$ is a Hadamard matrix of degree ab.*

COROLLARY 2.10. *Let $H_1 = (1)$ and $H_2 = \left(\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right)$. Then define $H_{2^m} = H_1 \otimes (\bigotimes_{i=1}^{m} H_2)$. Then $H_{2^m} \in M_{2^m}(\{-1,1\})$ is a Hadamard matrix.*

These matrices H_{2^m} are called the Hadamard matrices of Sylvester type. There exists an extensive literature (e.g., [12], [14]) on techniques which deal with the construction of Hadamard matrices.

2.3. Degree 5. The arrays we are interested in are $OA(N,d,2,5)$. We saw that we can construct such an array by using the codewords of a code $(d,N,d_C)_2$ with dual distance 6. Lower-dimension constructions are easier.

2.3.1. $d = 5$, $N = 2^d = 32$. The rows of an $OA(N,d,2,5)$ correspond to the $N = 2^d$ points in $\{-1,1\}^d$.

2.3.2. $d = 6,7,8$, $N = 2^{d-1}$. Let A be the $OA(2^d,d,2,5)$ constructed above. Then define $A_{i,d} = \prod_{j=1}^{d-1} A_{i,j}$. That gives us an $OA(2^{d-1},d,2,5)$.

2.3.3. $d = 9$, $N = 128 = 2^{d-2}$. This is the first example of an orthogonal array constructed from a code, a cyclic code. We will describe in this particular case how to construct this array (or those codewords). Let G be the 7×9 matrix in $GF(2)$ (the Galois field with 2 elements, i.e., $\mathbb{Z}/2\mathbb{Z}$):

$$G = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

G is called a cyclic generator matrix. Define the set of points C in $GF(2)^9$:

$$C = \{xG, x \in GF(2)^7\},$$

the calculation being done in $GF(2)$. Let

$$\begin{aligned}\Phi_2 : GF(2) &\longrightarrow \{-1,1\},\\ 0 &\mapsto -1,\\ 1 &\mapsto 1.\end{aligned}$$

Now we can construct a matrix A with rows that are the points in $\Phi_2(C) \subset \{-1,1\}^9$. A is an $OA(128, 9, 2, 5)$.

2.3.4. $d = 10, \ldots, 16$, $N = 256$. The $OA(256, 16, 2, 5)$ is constructed using a Nordstrom–Robinson code. Define the generator matrix G in $M_{4,8}(\mathbb{Z}/4\mathbb{Z})$ by

$$G = \begin{pmatrix} 1 & 3 & 1 & 2 & 1 & 0 & 0 & 0\\ 1 & 0 & 3 & 1 & 2 & 1 & 0 & 0\\ 1 & 0 & 0 & 3 & 1 & 2 & 1 & 0\\ 1 & 0 & 0 & 0 & 3 & 1 & 2 & 1 \end{pmatrix}.$$

Define the set of points

$$C = \{xG, x \in (\mathbb{Z}/4\mathbb{Z})^4\},$$

the calculation being done in $\mathbb{Z}/4\mathbb{Z}$. Let us define the Gray map

$$\begin{aligned}\Phi_4 : \mathbb{Z}/4\mathbb{Z} &\longrightarrow \{-1,1\}^2,\\ 0 &\mapsto (-1,-1),\\ 1 &\mapsto (-1,1),\\ 2 &\mapsto (1,1),\\ 3 &\mapsto (1,-1).\end{aligned}$$

Now we can construct a matrix A with rows that are the elements of $\Phi_4(C) \subset \{-1,1\}^{16}$. A is an $OA(256, 16, 2, 5)$. To get an $OA(256, k, 2, 5)$ for $k = 10, \ldots, 15$, we just delete $16 - k$ columns of A.

2.3.5. Higher dimension. $OA(512, 20, 2, 5)$ and $OA(1024, 24, 2, 5)$ are described in [14]. Many more $OA(N, d, 2, 5)$ for higher d are known, in particular with $d = 2^m + 1$ and $N = 2^{2m+1}$ when $m \geq 5$. They come from BCH codes [14], [20]. The X4 construction [14] allows the construction of some $OA(2^{4m+1}, 2^{2m} + 2m, 2, 5)$ for $m \geq 2$. Finally, Kerdock codes ([20] for their original construction or [13] for a simpler one) gives us orthogonal arrays of the form $OA(4^{2m}, 4^m, 2, 5)$ for $m \geq 2$. Those codes allow us to write $N_5(d) = O(d^2)$.

2.4. A notation. We will denote by $\mathcal{G}_d^m$ the set of points of a cubature formula of degree m with respect to $\delta_{\mathcal{G}_d}$ (with points in $\{-1,+1\}^d$), such that we do not know another cubature formula with points in $\{-1,+1\}^d$, of the same degree with respect to the same measure but with fewer points. When $d \geq m$, $\mathcal{G}_d^m$ is described in term of an $OA(|\mathcal{G}_d^m|, d, 2, m)$. $\mathcal{G}_d^m$ can be also seen as a subset of the group $\mathcal{G}_d \simeq (\{-1,+1\}^d, *)$. When $d \leq m$, necessarily, we have $\mathcal{G}_d^m = \mathcal{G}_d$.

Hadamard matrices allow us to write $|\mathcal{G}_d^3| = 2d$ anytime that there exists a Hadamard matrix and $|\mathcal{G}_d^3| = O(d)$. Kerdock and BCH codes allow us to write $|\mathcal{G}_d^5| = O(d^2)$. For a general m, coding theory tells us that $|\mathcal{G}_d^{2m+1}| = O(d^m)$.

3. Permutation sets. Let us consider a point $\mathbf{x} = (x^1, \dots, x^d) \in \mathbb{R}^d$ and the measure of total mass one which puts equal mass at all the points of the form $\sigma.\mathbf{x} = (x^{\sigma(1)}, \dots, x^{\sigma(d)})$ for $\sigma \in \mathcal{S}_d$, i.e.,

$$\delta_{\mathcal{S}_d,\mathbf{x}} = \frac{1}{|\mathcal{S}_d.\mathbf{x}|} \sum_{\mathbf{y}\in\mathcal{S}_d.\mathbf{x}} \delta_{\mathbf{y}}$$

where $\mathcal{S}_d$ denotes the symmetric group. We want to describe a cubature formula of degree m with respect to $\delta_{\mathcal{S}_d,\mathbf{x}}$.

3.1. Block designs. We assume in this subsection that $x^1 = \cdots = x^k = \alpha$ and $x^{k+1} = \cdots = x^d = \beta$, where $\alpha \neq \beta$. Hence, we are looking for a cubature formula of degree t with respect to the measure on $\mathbb{R}^d$:

$$\delta_{\mathcal{S}_d,\mathbf{x}} = \frac{1}{\binom{d}{k}} \sum_{\mathbf{y}\in\mathcal{S}_d.\mathbf{x}} \delta_{\mathbf{y}}.$$

Knowledge of some t-designs will allow us to construct some cubature formulae of degree t.

DEFINITION 3.1. *Let V be a finite set of size v and let B be a collection of k-subsets of V, called blocks. Then (V, B) is a t-design with parameters t-(v, k, λ) if every t-subset of V is in exactly λ blocks.*

The incidence matrix A of a design $(V, B) = (\{1, \dots, v\}, \{B_1, \dots, B_b\})$ is defined, for $i = 1, \dots, b$ and $j = 1, \dots, v$, by

$$A_{i,j} = \mathbf{1}_{\{j\in B_i\}}.$$

It is easily seen that a t-design with parameters (v, k, λ) is a $(t-1)$-design with parameters $(v, k, \lambda\frac{v-(t-1)}{k-(t-1)})$ [2]. Also, every 0-subset (i.e., the empty set) is included in all the blocks, so a t-design with parameters (v, k, λ) is a 0-design with parameters (v, k, b), where $b = |B|$ is the number of blocks of the design. Hence, if a t-design has parameters (v, k, λ) and has b blocks, then

$$\lambda = b\frac{k(k-1)\dots(k-(t-1))}{v(v-1)\dots(v-(t-1))} = b\frac{\binom{k}{t}}{\binom{v}{t}}.$$

A 2-design is usually called a balanced incomplete block design or just a block design. For a block design, if the number of blocks b is equal to v, we say that the design is symmetric (or square for some authors).

Example 3.2. Let q be a prime power, $n \geq 2$ be an integer, and $V(n, q)$ be a n-dimensional vector space over $GF(q)$ (the Galois field of order q). $V(n, q)$ contains q^n vectors. A one-dimensional space is made of $q-1$ nonzero vectors (it is made of the vectors bx, where b ranges over $GF(q) - \{0\}$ and where x is a nonzero vector); hence there are $\frac{q^n-1}{q-1}$ one-dimensional spaces. If $x = (x_1, \dots, x_n)$ is a nonzero vector, then the set of vectors $y = (y_1, \dots, y_n)$ such that

$$x_1y_1 + \cdots + x_ny_n = 0$$

define a subspace of $V(n, q)$ of dimension $n-1$ (a hyperplane), and conversely, for any hyperplane, we can find a vector $x = (x_1, \dots, x_n)$ such that, for each vector $y = (y_1, \dots, y_n)$ of the hyperplane,

$$x_1y_1 + \cdots + x_ny_n = 0.$$

Thus we have $\frac{q^n-1}{q-1}$ hyperplanes. Each hyperplane contains $\frac{q^{n-1}-1}{q-1}$ one-dimensional spaces (itself being a vector space of dimension $n-1$), and the intersection of two hyperplanes is a subspace of $V(n,q)$ of dimension $n-2$, which contains $\frac{q^{n-2}-1}{q-1}$ one-dimensional spaces. Hence, the hyperplanes of $V(n,q)$ as blocks and the one-dimensional spaces of $V(n,q)$ as objects form a symmetric block design $(\frac{q^n-1}{q-1}, \frac{q^{n-1}-1}{q-1}, \frac{q^{n-2}-1}{q-1})$.

We refer to [2], [3], [12] for more details on designs, and for tables of known block designs.

Let $J_{b,d}$ be the $b \times d$ matrix with each entry set to be 1, and we remind the reader that $\mathbf{x}$ is defined in this subsection by $x^1 = \cdots = x^k = \alpha$ and $x^{k+1} = \cdots = x^d = \beta$, with $\alpha \neq \beta$.

THEOREM 3.3. *Let A be the incidence matrix of a t-design with parameters $t-(d,k,\lambda)$ and with b blocks. Let $\mathbf{x}_1, \ldots, \mathbf{x}_b \in \mathbb{R}^d$ be the rows of $(\alpha-\beta)A + \beta J_{b,d}$. Then those points associated to the constant weight $1/b$ (b is the number of blocks of the t-design) define a cubature formula of degree t with respect to $\delta_{\mathcal{S}_d,\mathbf{x}}$.*

Proof. First of all, note that, by considering the polynomials which are products of some $\frac{(X_j-\beta)^k}{(\alpha-\beta)^k}$ in place of the monomials, we see that we may take $\alpha = 1$ and $\beta = 0$. Since for all $k_1, \ldots, k_d \geq 0$, $\int P(z_1^{k_1}, \ldots, z_d^{k_d}) \delta_{\mathcal{S}_d,\mathbf{x}}(dz) = \int P(z_1, \ldots, z_d) \delta_{\mathcal{S}_d,\mathbf{x}}(dz)$, and since permuting the rows of the incidence matrix of a block design gives another block design with the same parameters, we have only to check that

$$\int P(z) \delta_{\mathcal{S}_d,\mathbf{x}}(dz) = \frac{1}{b} \sum_{i=1}^{b} P(\mathbf{x}_i)$$

for the polynomials X_1, X_1X_2, $\ldots$, $X_1 \ldots X_t$. Define $\lambda_t = \lambda$, and for $i < t$, $\lambda_i = \lambda_{i+1}(d-i)/(k-i)$, so that A is the incidence matrix of an $i-(d,k,\lambda_i)$ design, $i = 1, \ldots, t$. Let $P = X_1 \ldots X_i$. Then

$$\frac{1}{b} \sum_{j=1}^{b} P(\mathbf{x}_j) = \frac{\lambda_i}{b} = \frac{\binom{k}{i}}{\binom{d}{i}}.$$

The proof is then finished by noticing that

$$\int P(z) \delta_{\mathcal{S}_d,\mathbf{x}}(dz) = \binom{d-i}{k-i} \Big/ \binom{d}{k} = \frac{\binom{k}{i}}{\binom{d}{i}}. \qquad \square$$

For cubature formulae of degree 2, symmetric block designs will be very interesting, as they provide formulae with the minimum number of points.

3.2. k-homogeneous permutation sets. We now return to the more general case, i.e., we do not assume anything about the structure of the point $\mathbf{x} = (x^1, \ldots, x^d)$, and we recall that we are looking for a cubature formula of degree t with respect to the d-dimensional measure

$$\delta_{\mathcal{S}_d,\mathbf{x}} = \frac{1}{|\mathcal{S}_d.\mathbf{x}|} \sum_{\mathbf{y} \in \mathcal{S}_d.\mathbf{x}} \delta_{\mathbf{y}}.$$

The symmetric group $\mathcal{S}_d$ acts on the set $T = \{1, \ldots, d\}$ and acts naturally on $T^{\{k\}}$, the set of k-element subsets of T. For A, B in $T^{\{k\}}$, $[A;B]$ denotes the subset of $\mathcal{S}_d$

which consists of all permutations that move A to B. A nonempty subset Z_d of $\mathcal{S}_d$ is said to be k-homogeneous on T if the cardinality of $Z_d \cap [A;B]$ is independent of A and B in $T^{\{k\}}$. The cardinality of $Z_d \cap [A;B]$ is called the k-multiplicity of Z_d. When this cardinality is one, we say that Z_d is sharply k-homogeneous on T. If Z_d is k-homogeneous on T, then Z_d is $(k-1)$-homogeneous on T (when $2 \leq k \leq d/2$ [24]). See [4] for the construction of such sets. Particular cases of k-homogeneous sets are k-homogeneous groups [17].

Example 3.4. Let $K = GF(q)$ be the Galois field of order q, where q is a prime power. Let Γ be the group of mappings from K to $K : x \to ax + b$, where $b \in K$ and a is in the group of nonzero square of K. Then Γ is a sharply 2-homogeneous group on K. $|\Gamma| = q(q-1)/2$.

The link with our problem is explained in the following theorem.

THEOREM 3.5. *Let Z_d be a t-homogeneous set on $T = \{1, \ldots, d\}$. Then the set of points $\sigma.\mathbf{x}$, for $\sigma \in Z_d$, with constant weight $1/|Z_d.\mathbf{x}|$ defines a cubature formula of degree t with respect to $\delta_{\mathcal{S}_d,\mathbf{x}}$.*

Proof. It is straightforward to check the cubature relation for all monomials of degree less than or equal to t. ∎

3.3. A notation. We will denote by $\mathcal{S}_d^{m,\mathbf{x}}$ a subset of $\mathcal{S}_d$ such that the probability measure $\frac{1}{|\mathcal{S}_d^{m,x}.\mathbf{x}|}\sum_{\mathbf{y}\in\mathcal{S}_d^{m,x}.\mathbf{x}} \delta_{\mathbf{y}}$ defines a cubature formula of degree m with respect to $\delta_{\mathcal{S}_d,\mathbf{x}}$ (with points inside the support of $\delta_{\mathcal{S}_d,\mathbf{x}}$ and with positive weights) and such that we do not know another formula with fewer points (with points inside the support of $\delta_{\mathcal{S}_d,\mathbf{x}}$ and with positive weights).

4. Invariant cubature formulae. Now that we have constructed some cubature formulae with few points with respect to some discrete measures, we are going to see how these can be useful for the construction of cubature formulae with respect to a "symmetric" measure. First of all, we should remind the reader of the application to cubature of invariant theory. See [18] for a more detailed presentation.

4.1. Invariant theory. Let $\mathcal{G}$ be a group of bijective linear transformations on $\mathbb{R}^d$. A set Ω of $\mathbb{R}^d$ is said to be $\mathcal{G}$-invariant if, for all $\mathbf{g} \in \mathcal{G}$, $\mathbf{g}.\Omega = \Omega$. A function f on Ω is said to be $\mathcal{G}$-invariant if, for all $\mathbf{g} \in \mathcal{G}$, $f \circ \mathbf{g} = f$. Finally, a measure μ is said to be $\mathcal{G}$-invariant if its support is $\mathcal{G}$-invariant and if, for all measurable sets A and for all $\mathbf{g} \in \mathcal{G}$, $\mathbf{g}.A$ is measurable and $\mu(\mathbf{g}.A) = \mu(A)$. We will denote by $\mathbb{R}_m[X_1, \ldots, X_d](\mathcal{G})$ the space of all $\mathcal{G}$-invariant polynomials of maximum degree m.

DEFINITION 4.1. *A cubature formula with points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and weights $\lambda_1, \ldots, \lambda_n$ is said to be $\mathcal{G}$-invariant if $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is $\mathcal{G}$-invariant and $\mathbf{g}.\mathbf{x}_i = \mathbf{x}_j$ implies $\lambda_i = \lambda_j$. Equivalently, the cubature formula is $\mathcal{G}$-invariant if $\sum_{i=1}^n \lambda_i \delta_{\mathbf{x}_i}$ is a $\mathcal{G}$-invariant measure.*

$\mathcal{G}.\mathbf{x}_i = \{\mathbf{g}.\mathbf{x}_i, \mathbf{g} \in \mathcal{G}\}$ is called the orbit of the point $\mathbf{x}_i$. All the weights associated with the points inside the same orbit are equal.

A subset $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}\}$ is called a generator set of the above cubature formula if $\mathcal{G}.\mathbf{x}_{i_1}, \ldots, \mathcal{G}.\mathbf{x}_{i_k}$ forms a partition of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. We will also say that the collection of orbits $\mathcal{G}.\mathbf{x}_{i_j}$ and weights $\lambda_{i_j}|\mathcal{G}.\mathbf{x}_{i_j}|$, $j = 1, \ldots, k$, generate the above $\mathcal{G}$-invariant cubature formula.

The following theorem is due to Sobolev [26].

THEOREM 4.2. *Let μ be a $\mathcal{G}$-invariant measure. Then the orbits $\mathcal{G}.\mathbf{x}_1, \ldots, \mathcal{G}.\mathbf{x}_k$ and their weights $\lambda_1, \ldots, \lambda_k$ are the generators of a $\mathcal{G}$-invariant cubature formula of degree m with respect to μ if and only if for all $\mathcal{G}$-invariant polynomials P of degree*

less than or equal to m,

$$\int P(z)\mu(dz) = \sum_{i=1}^{k} \lambda_i P(\mathbf{x}_i).$$

From Tchakaloff's theorem and Sobolev's theorem, we obtain the following corollary.

COROLLARY 4.3. *Let d and m be positive integers and let μ be a positive $\mathcal{G}$-invariant measure on $\mathbb{R}^d$ with the property that $\int |P(z)|\mu(dz) < \infty$ for all $P \in \mathbb{R}_m[X_1, \ldots, X_d]$. Then we can find k orbits $\mathcal{G}.\mathbf{x}_1, \ldots, \mathcal{G}.\mathbf{x}_k$ (in the support of μ) and weights $\lambda_1, \ldots, \lambda_k$ that generate a $\mathcal{G}$-invariant cubature formula of degree m with respect to μ with*

$$k \leq \dim \mathbb{R}_m[X_1, \ldots, X_d](\mathcal{G}).$$

Proof. By Tchakaloff's theorem, we can find n points $\widetilde{\mathbf{x}_1}, \ldots, \widetilde{\mathbf{x}_n}$ in the support of μ together with their weights $\widetilde{\lambda_1}, \ldots, \widetilde{\lambda_n}$ that define a cubature formula of degree m with respect to μ. Then the points $\widetilde{\mathbf{x}_{\mathbf{g},i}} = \mathbf{g}.\widetilde{\mathbf{x}}_i$, $\mathbf{g} \in \mathcal{G}$, $i = 1, \ldots, n$, and the weights $\widetilde{\lambda_{\mathbf{g},i}} = \frac{\widetilde{\lambda}_i}{|\mathcal{G}|}$ define a $\mathcal{G}$-invariant cubature formula of degree m with respect to μ. Let $\mathcal{G}.\widehat{\mathbf{x}_1}, \ldots, \mathcal{G}.\widehat{\mathbf{x}_{k'}}$ be some orbits and $\widehat{\lambda_1}, \ldots, \widehat{\lambda_{k'}}$ some weights that generate this $\mathcal{G}$-invariant cubature formula. $\mathcal{G}.\widehat{\mathbf{x}_1}, \ldots, \mathcal{G}.\widehat{\mathbf{x}_{k'}}$ and $\widehat{\lambda_1}, \ldots, \widehat{\lambda_{k'}}$ generate a $\mathcal{G}$-invariant cubature formula of degree m with respect to μ. If $k^{'} > \dim \mathbb{R}_m[X_1, \ldots, X_d](\mathcal{G})$, using the same convex analysis argument as in Tchakaloff's theorem, it is possible to find $k \leq \dim \mathbb{R}_m[X_1, \ldots, X_d](\mathcal{G})$ orbits $\mathcal{G}.\mathbf{x}_1, \ldots, \mathcal{G}.\mathbf{x}_k \in \{\mathcal{G}.\widehat{\mathbf{x}_1}, \ldots, \mathcal{G}.\widehat{\mathbf{x}_{k'}}\}$ and some new weights $\lambda_1, \ldots, \lambda_k$ that generate a $\mathcal{G}$-invariant cubature formula of degree m with respect to μ. ∎

Assume that the orbits $\mathcal{G}.\mathbf{x}_1, \ldots, \mathcal{G}.\mathbf{x}_k$ together with their weights $\lambda_1, \ldots, \lambda_k$ generate a $\mathcal{G}$-invariant cubature formula of degree m with respect to μ. In other words, $\xi = \sum_{i=1}^{k} \lambda_i \xi_{\mathcal{G},\mathbf{x}_i}$, where $\xi_{\mathcal{G},\mathbf{x}_i} = \frac{1}{|\mathcal{G}.\mathbf{x}_i|} \sum_{\mathbf{y} \in \mathcal{G}.\mathbf{x}_i} \delta_{\mathbf{y}}$, is the measure associated with a cubature formula of degree m with respect to μ. Now assume that, for all i, we can find a cubature formula with n_i points and positive weights with respect to the (discrete) measure $\xi_{\mathcal{G},\mathbf{x}_i}$. Let $\widetilde{\xi_{\mathcal{G},\mathbf{x}_i}}$ denote the measures associated with these cubature formulae. Then $\widetilde{\xi} = \sum_{i=1}^{k} \lambda_i \widetilde{\xi_{\mathcal{G},\mathbf{x}_i}}$ is still a measure associated with a cubature formula of degree m with respect to μ. Note that this cubature formula is no longer $\mathcal{G}$-invariant. Its number of points is $\sum_{i=1}^{k} n_i$. We are going to show that, in the case where $\mathcal{G}$ is the group of permutations and reflections of the axes, $\dim \mathbb{R}_m[X_1, \ldots, X_d](\mathcal{G})$ does not depend on the dimension d and that we can find (and construct) cubature formulae with respect to $\xi_{\mathcal{G},\mathbf{x}_i}$ with the number of points bounded by the Tchakaloff bound. Hence, if we have some orbits and weights generating a $\mathcal{G}$-invariant cubature formula with respect to μ, it is possible to find a cubature formula with respect to μ which has a number of points which grows polynomially with the dimension (at least when $\mathcal{G}$ is the group of permutations and reflections of the axes).

4.2. Invariance under reflection. Let us consider a positive measure μ on $\mathbb{R}^d$ invariant with respect to the group of reflections of the axes $\mathcal{G}_d \simeq (\{-1, +1\}^d, *)$. The Lebesgue measure on the hypercube and on the unit sphere and the Gaussian measure on $\mathbb{R}^d$ are examples of such a measure.

$\mathcal{G}_d$ acts on $\mathbb{R}^d$ in a natural way: for a point $\mathbf{x} = (x^1, \ldots, x^d)$ and an element of the group $\mathbf{g} = (g^1, \ldots, g^d)$, we define $\mathbf{g}.\mathbf{x} = (g^1 x^1, \ldots, g^d x^d)$.

For $\mathbf{x} \in \mathbb{R}^d$, $\mathcal{G}_d.\mathbf{x}$ is of cardinality $2^{e(\mathbf{x})}$, where $e(\mathbf{x})$ is the number of nonzero coordinates of $\mathbf{x}$. That allows us to define the action of $\mathcal{G}_{e(\mathbf{x})}$ on $\mathbf{x}$: Assume that $i_1, \ldots, i_{e(\mathbf{x})}$ are the $e(\mathbf{x})$ nonzero coordinates of $\mathbf{x}$; then for $\mathbf{g} \in \mathcal{G}_{e(\mathbf{x})}$, we define the i_l coordinate of $\mathbf{g}.\mathbf{x}$ to be $g^l x^{i_l}$, while the coordinates which are zero remain zero under the action of $\mathbf{g}$. This just means that we consider only the reflection with respect to the ith axis when $x_i \neq 0$. Obviously, $\mathcal{G}_d.\mathbf{x} = \mathcal{G}_{e(\mathbf{x})}.\mathbf{x}$.

Corollary 4.3 tells us that there exists k (with $k \leq \dim \mathbb{R}_m[X_1, \ldots, X_d](\mathcal{G}_d)$) orbits $\mathcal{G}_d.\mathbf{x}_1, \ldots, \mathcal{G}_d.\mathbf{x}_k$ in the support of the measure μ and k weights $\lambda_1, \ldots, \lambda_k$ that generate a $\mathcal{G}_d$-invariant cubature formula of degree m with respect to μ. The measure $\sum_{i=1}^k \lambda_i \xi_{\mathcal{G}_d,\mathbf{x}_i}$, where $\xi_{\mathcal{G}_d,\mathbf{x}_i} = \frac{1}{|\mathcal{G}_d.\mathbf{x}_i|} \sum_{\mathbf{y} \in \mathcal{G}_d.\mathbf{x}_i} \delta_{\mathbf{y}}$, is then associated with a cubature formula of degree m with respect to μ. But the measure

$$\widetilde{\xi_{\mathcal{G}_d,\mathbf{x}_i}} = \frac{1}{2^{e(\mathbf{x}_i)}} \sum_{\mathbf{g} \in \mathcal{G}^m_{e(\mathbf{x}_i)}} \delta_{\mathbf{g}.\mathbf{x}_i}$$

defines a cubature formula of degree m with respect to $\xi_{\mathcal{G}_d,\mathbf{x}_i}$ ($\mathcal{G}^m_{e(\mathbf{x}_i)}$ has been defined in terms of orthogonal arrays in section 2). Hence, the finite measure $\sum_{i=1}^k \lambda_i \widetilde{\xi_{\mathcal{G}_d,\mathbf{x}_i}}$ defines a cubature formula of degree m with respect to μ.

The number of points in this cubature formula is $\sum_{i=1}^k |\mathcal{G}^m_{e(\mathbf{x}_i)}| \leq k|\mathcal{G}^m_d|$. We saw that $|\mathcal{G}^m_d| = O(d^{[m/2]})$. We therefore need to find a way to find k generators, with k growing polynomially with the dimension. First, let us write the problem in equivalent terms.

For $\mathbf{x} \in \mathbb{R}^d$, let us denote

$$\mathbf{x}^2 = \left(\left(x^1\right)^2, \ldots, \left(x^d\right)^2 \right)$$

and

$$\sqrt{\mathbf{x}} = \left(\sqrt{x^1}, \ldots, \sqrt{x^d} \right).$$

We associate with our measure μ the measure $\nu = \mu \circ \sqrt{.}$ with support included in $\mathbb{R}^d_+$, such that, for all integrable functions f,

$$\int f(z^2)\mu(dz) = \int f(z)\nu(dz).$$

Assume that the orbits $\mathcal{G}_d.\mathbf{x}_1, \ldots, \mathcal{G}_d.\mathbf{x}_k$ in the support of the measure μ and the weights $\lambda_1, \ldots, \lambda_k$ generate a $\mathcal{G}_d$-invariant cubature formula of degree m with respect to μ. Then the points $\mathbf{x}_1^2, \ldots, \mathbf{x}_k^2$ and the weights $\lambda_1, \ldots, \lambda_k$ define a $\mathcal{G}_d$-invariant cubature formula of degree $[m/2]$ with respect to ν. Indeed, let P be a polynomial of degree less than or equal to $[m/2]$, and let $Q = P(X^2)$. Then

$$\begin{aligned} \sum_{i=1}^k \lambda_i P(\mathbf{x}_i^2) = \sum_{i=1}^k \lambda_i Q(\mathbf{x}_i) &= \sum_{i=1}^k \lambda_i \frac{1}{2^{e(\mathbf{x}_i)}} \sum_{y \in \mathcal{G}_d.\mathbf{x}_i} Q(y) \\ &= \int Q(z)\mu(dz) = \int P(z)\nu(dz). \end{aligned}$$

It is straightforward to check that, reciprocally, if the points $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in the support of the measure ν and the weights $\lambda_1, \ldots, \lambda_k$ define a cubature formula of degree $[m/2]$

with respect to ν, then the orbits $\mathcal{G}_d.\sqrt{\mathbf{x}_1}, \dots, \mathcal{G}_d.\sqrt{\mathbf{x}_k}$ in the support of the measure μ and the weights $\lambda_1, \dots, \lambda_k$ generate a $\mathcal{G}_d$-invariant cubature formula of degree m with respect to μ.

Therefore, to find these k generators, $\mathcal{G}_d.\mathbf{x}_1, \dots, \mathcal{G}_d.\mathbf{x}_k$ is equivalent to finding a cubature formula of degree $[m/2]$ with respect to $\nu = \mu \circ \sqrt{.}$. We are now going to show that this problem is simpler when ν is invariant with respect to the group of permutation of the axes $\mathcal{S}_d$ (which is equivalent to the fact that μ is invariant with respect to the group of permutation of the axes $\mathcal{S}_d$).

4.3. Invariance under permutation. Let us consider a positive measure ρ on $\mathbb{R}^d$ invariant with respect to the group of permutation of the axes $\mathcal{S}_d$. The Lebesgue measure on the hypercube and on the unit sphere and the Gaussian measure on $\mathbb{R}^d$ are again examples of such a measure.

Corollary 4.3 tells us that there exist k (with $k \leq \dim \mathbb{R}_m[X_1, \dots, X_d](\mathcal{S}_d)$) orbits $\mathcal{S}_d.\mathbf{x}_1, \dots, \mathcal{S}_d.\mathbf{x}_k$ in the support of the measure ρ and k weights $\lambda_1, \dots, \lambda_k$ that generate a $\mathcal{S}_d$-invariant cubature formula of degree m with respect to ρ. Then, by definition, $\sum_{i=1}^k \lambda_i \delta_{\mathcal{S}_d.\mathbf{x}}$, (where $\delta_{\mathcal{S}_d.\mathbf{x}} = \frac{1}{|\mathcal{S}_d.\mathbf{x}|} \sum_{\mathbf{y} \in \mathcal{S}_d.\mathbf{x}} \delta_{\mathbf{y}}$) is the measure associated with a cubature formula of degree m with respect to ρ. But the measure

$$\widetilde{\delta_{\mathcal{S}_d.\mathbf{x}}} = \frac{1}{|\mathcal{S}_d^{m,\mathbf{x}}.\mathbf{x}_i|} \sum_{\sigma \in \mathcal{S}_d^{m,\mathbf{x}_i}} \sigma.\mathbf{x}_i$$

defines a cubature formula of degree m with respect to $\delta_{\mathcal{S}_d.\mathbf{x}}$ ($\mathcal{S}_d^{m,\mathbf{x}}$ has been defined in terms of designs and permutation sets in section 3). Hence, the finite measure $\sum_{i=1}^k \lambda_i \widetilde{\delta_{\mathcal{S}_d.\mathbf{x}}}$ defines a cubature formula of degree m with respect to ρ. This formula has $\sum_{i=1}^k |\mathcal{S}_d^{m,\mathbf{x}_i}|$ points.

$\mathbb{R}_m[X_1, \dots, X_d](\mathcal{GS}_d)$ is generated by $\sum_{\sigma \in \mathcal{S}_d} X_{\sigma(1)}^{p_1} \dots X_{\sigma(d)}^{p_d}$ for $(p_1, \dots, p_d)$ describing $\mathcal{P}_{m,d}$, where

$$\mathcal{P}_{m,d} = \{p = (p_1, \dots, p_d), \quad |p| \leq m, \quad m \geq p_1 \geq p_2 \geq \dots \geq p_d \geq 0\}.$$

For all $d \geq m$, $|\mathcal{P}_{m,d}| = |\mathcal{P}_{m,m}|$. This cardinality can be expressed in terms of number of Young tableaux [19]. The very important point here is that k, the number of generators, is bounded by a term ($|\mathcal{P}_{m,m}|$) which *does not depend on the dimension.* Thus, to get a cubature formula with respect to a symmetric measure with few points, one has to find k generators (and we know that we can do it with k bounded by $|\mathcal{P}_{m,m}|$) and "good" cubature formulae with respect to $\delta_{\mathcal{S}_d.\mathbf{x}}$ (this is a combinatorial problem; we gave some indications on how to do so in section 3).

4.4. Invariance under permutation and reflection. We now put the two previous sections together. Recall that $\mathcal{GS}_d$ is the group generated by all reflections and permutations of the axes. μ will now denote a positive measure on $\mathbb{R}^d$ which is $\mathcal{GS}_d$-invariant. The Lebesgue measure on the hypercube and on the unit sphere and the Gaussian measure on $\mathbb{R}^d$ are, once again, examples of such a measure.

We summarize our technique to find cubature formulae of degree m with respect to our $\mathcal{GS}_d$-invariant measure μ.

1. Find k orbits $\mathcal{GS}_d.\mathbf{x}_1, \dots, \mathcal{GS}_d.\mathbf{x}_k$ (with their elements in the support of μ) and k positive weights $\lambda_1, \dots, \lambda_k$ that generate a $\mathcal{GS}_d$-invariant cubature formula of degree m with respect to μ. This is equivalent to the fact that orbits $\mathcal{S}_d.\mathbf{x}_1^2, \dots, \mathcal{S}_d.\mathbf{x}_k^2$ and the weights $\lambda_1, \dots, \lambda_k$ that generate an $\mathcal{S}_d$-invariant

cubature formula of degree $[m/2]$ with respect to $\nu = \mu \circ \sqrt{.}$. One should be able to find these generators with $k \leq |\mathcal{P}_{[m/2],[m/2]}|$.

2. For all $i = 1, \dots, k$, construct, using the methods of section 3, a cubature formula of degree $[m/2]$ with respect to $\delta_{\mathcal{S}_d,\mathbf{x}}$. Those points are the points in $\mathcal{S}_d^{[m/2],x_i}.\mathbf{x}_i^2 = \{\mathbf{x}_{i,1}^2, \dots, \mathbf{x}_{i,n_i}^2\}$. The points $\mathbf{x}_{i,j}^2$, $i = 1, \dots, k$, $j = 1, \dots, n_i$, with the weights $\frac{\lambda_i}{n_i}$ now define a cubature formula of degree $[m/2]$ with respect to $\mu \circ \sqrt{.}$. Equivalently, the orbits $\mathcal{G}_d.\mathbf{x}_{i,j}$ and the weights $\frac{\lambda_i}{n_i}$ generate a $\mathcal{G}_d$-invariant cubature formula of degree m with respect to μ.
3. The points in $\mathbf{g}.\mathbf{x}_{i,j}$, $i = 1, \dots, k$, $j = 1, \dots, n_i$, $\mathbf{g} \in \mathcal{G}^m_{e(\mathbf{x}_i)}$, with the weights $\frac{\lambda_i}{n_i |\mathcal{G}^m_{e(\mathbf{x}_i)}|}$ define a cubature formula of degree m with respect to μ.

Example 4.4. Consider a $\mathcal{GS}_d$-invariant measure μ_d on $\mathbb{R}^d$ (for example, the Gaussian measure or the Lebesgue measure on the unit cube or on the unit sphere). Let V be the volume of μ_d, and assume that $\mathbf{x} = \sqrt{\frac{1}{V}\int z^2 \mu_d(dz)}$ belongs to the support of μ_d. Then the generator $\mathcal{G}_d.\mathbf{x}$ and its weight V generate a $\mathcal{G}_d$-invariant cubature formula of degree 3 with respect to μ_d (hence it provides a cubature formula with 2^d points). Using our construction, we see that the points $\mathbf{g}.\mathbf{x}$, $\mathbf{g} \in \mathcal{G}_d^3$, with their weight $V/|\mathcal{G}_d^3|$, define a cubature formula of degree 3 with respect to μ_d. It has $|\mathcal{G}_d^3| = O(d)$ points. Recall that $\mathcal{G}_d^3$ was defined in terms of Hadamard matrices and that whenever there exists a Hadamard matrix of degree d, $|\mathcal{G}_d^3| = 2d$ (which is the Möller lower bound [22]).

Let us now construct cubature formula of degree 5 with few points for some classical measures.

5. Application to some classical regions. In this section, we will describe cubature formula of degree 5 for the region C_d, $E_d^{r^2}$, U_d, S_d.

5.1. $E_d^{r^2}$, the Gaussian measure on $\mathbb{R}^d$. Our definition of $E_d^{r^2}$ differs very slightly from the one given by Stroud. We consider the measure on $\mathbb{R}^d$

$$\mu_d(d\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{x_1^2 + \cdots + x_d^2}{2}\right) dx_1 \dots dx_d.$$

We choose this definition of $E_d^{r^2}$ so that $\int f(z)\mu_d(z) = \mathbb{E}(f(N))$, where N is a d-dimensional normal random variable. First of all, we write $\nu_d = \mu_d \circ \sqrt{.}$. One can easily see that $\nu_d = \nu_1 \otimes \cdots \otimes \nu_1$ and that

$$\int 1\nu_1(dx) = 1, \quad \int x\nu_1(dx) = 1, \quad \int x^2\nu_1(dx) = 3.$$

We will provide two constructions.

5.1.1. First solution. When the dimension d is of the form $d = 3k - 2$, the orbits $\mathcal{GS}_d.x_0$ and $\mathcal{GS}_d.\mathbf{x}_1$, where

$$\mathbf{x}_0 = (0, \dots, 0), \quad \mathbf{x}_1 = (\sqrt{3}, \dots, \sqrt{3}, 0, \dots, 0)$$

(k coordinates of $\mathbf{x}_1$ are equal to $\sqrt{3}$, $2k-2$ to 0), with the weights

$$\omega_0 = \tfrac{2}{d+2}, \quad \omega_1 = \tfrac{d}{d+2},$$

generate a $\mathcal{GS}_d$-invariant cubature formula of degree 5 with respect to μ_d. Thus we need to find a cubature formula of degree 2 with respect to

$$\delta_{\mathcal{S}_d,\mathbf{x}} = \frac{1}{|\mathcal{S}_d.\mathbf{x}_1|} \sum_{\mathbf{y} \in \mathcal{S}_d.\mathbf{x}_1^2} \delta_{\mathbf{y}}.$$

According to section 3, this can be done by using a 2-design with parameters

$$2-(3k-2,k,\lambda)$$

for a given λ. A few symmetric block designs will give us good answers. There exist symmetric block designs with parameters $(7,3,1)$, $(16,6,2)$, $(25,9,3)$ [3], and $(70,24,8)$ [16] and more generally with parameters $(3^{k+2}-2,3^{k+1},3^k)$ [15], [21] and $(2.9^{k+1}-2,6.9^k,2.9^k)$ [27]. Symmetric design of the form $(9\lambda-2,3\lambda,\lambda)$ do not exist when $\lambda=4,5,6,7$. We give an example of the simplest of these block designs, described by its incidence matrix,

$$A_{7,3,1}=\begin{pmatrix}0&0&0&1&0&1&1\\1&0&0&0&1&0&1\\1&1&0&0&0&1&0\\0&1&1&0&0&0&1\\1&0&1&1&0&0&0\\0&1&0&1&1&0&0\\0&0&1&0&1&1&0\end{pmatrix}. \tag{2}$$

Therefore, assume that there exists a symmetric block design A with parameters $2-(d,(d+2)/3,(d+2)/9)$. Now define $\mathbf{x}_{1,i}$, $i=1,\ldots,d$, to be $\sqrt{3}$ times the ith row of A. Then $\mathbf{x}_{1,1}^2,\ldots,\mathbf{x}_{1,d}^2$ together with equal weights $1/d$ define a cubature formula of degree 2 with respect to $\delta_{\mathcal{S}_d.\mathbf{x}}$. Then the point $\mathbf{x}_0$ with its weight $\frac{2}{d+2}$ and the points $\mathbf{g}.\sqrt{\mathbf{x}_{1,i}}$ for $\mathbf{g}\in\mathcal{G}^5_{(d+2)/3}$, $i=1,\ldots,d$, with their weights equal to $\frac{1}{|\mathcal{G}^5_{(d+2)/3}|(d+2)}$, define a cubature formula of degree 5 with respect to the Gaussian measure μ_d. Thus, every time that there exists a design $(d=9\lambda-2,3\lambda,\lambda)$, we can construct a cubature formula of degree 5 with respect to the d-dimensional Gaussian measure, with the number of points being equal to $d|\mathcal{G}^5_{(d+2)/3}|+1=O(d^3)$. If we want to find a cubature formula for a given dimension for which the above method does not work (as we need the existence of some specific design), we choose the least d' greater than d such that the method works; then we project our points on a d-dimensional subspace of $\mathbb{R}^{d'}$ to get a cubature formula of degree 5 with respect to μ_d. It is interesting to note that, for $\lambda=1$ (hence $d=7$), that leads to a formula with $7|\mathcal{G}^5_3|+1=57$ points, which is exactly the Möller lower bound [22]. Considering that known cubature formulae of degree greater than 4 in dimension greater than 3 attaining the Möller lower bound are quite rare, it is surprising that this formula is actually the second cubature formula (with positive weights) for $E_d^{r^2}$ attaining the Möller lower bound. (See [28] for a description of the first one.)

5.1.2. Second solution. Let

$$\mathbf{x}_0=(r,0,\ldots,0),\quad \mathbf{x}_1=(s,\ldots,s),$$

where

$$r^2=\tfrac{d+2}{2},\quad s^2=\tfrac{d+2}{d-2}.$$

Then the orbits $\mathcal{GS}_d.\mathbf{x}_0$ and $\mathcal{GS}_d.\mathbf{x}_1$, with the weights

$$w_0=\tfrac{8d}{(d+2)^2},\quad w_1=\left(\tfrac{d-2}{d+2}\right)^2,$$

generate a $\mathcal{GS}_d$-invariant cubature formula of degree 5 with respect to μ_d. We denote by (a, b), for $a, b \in \{1, \ldots, d\}$, the permutation which swaps a and b and leaves invariant all the other elements.

Then the points $\pm(1, i).\mathbf{x}_0$, $i = 1, \ldots, d$, with their weights $\frac{w_0}{2d}$ and $\mathbf{g}.\mathbf{x}_1$, $\mathbf{g} \in \mathcal{G}_d^5$, with their weights $\frac{w_1}{|\mathcal{G}_d^5|}$, define a cubature formula of degree 5 with respect to the Gaussian measure μ_d. The formula has $|\mathcal{G}_d^5| + 2d = O(d^2)$ points. The orbits and the weights were obtained from the cubature formula $E_n^{r^2} : 5 - 3$ [28, p. 317].

5.2. C_d, the d-dimensional cube. We place on $[-1, 1]^d$ the measure

$$\mu_d(d\mathbf{x}) = \frac{1}{2^d} dx_1 \ldots dx_d.$$

Thus once again, we want to find a symmetric cubature formula of degree 2 with respect to the measure associated with μ_d. This measure ν_d is actually equal to

$$\nu_d(d\mathbf{x}) = \frac{1}{2^d} \frac{1}{\sqrt{x_1 \ldots x_d}} dx_1 \ldots dx_d.$$

Assume that the dimension d is odd and let

$$\alpha = \sqrt{\tfrac{1}{3} + \tfrac{2}{3\sqrt{5}}}, \quad \beta = \sqrt{\tfrac{1}{3} - \tfrac{2}{3\sqrt{5}}}.$$

Define $\mathbf{x}_0 = (\alpha, \ldots, \alpha)$ and $\mathbf{x}_1 \in \mathbb{R}^d$ such that $\mathbf{x}_1^i = \alpha$ if $i = 1, \ldots, (d-1)/2$ and $\mathbf{x}_1^i = \beta$ if $i = (d+1)/2, \ldots, d$. Then the orbits $\mathcal{GS}_d.x_0$ and $\mathcal{GS}_d.\mathbf{x}_1$, with the weights

$$w_0 = \tfrac{1}{d+1}, \quad w_1 = \tfrac{d}{d+1},$$

generate a $\mathcal{GS}_d$-invariant cubature formula of degree 5 with respect to μ_d. To find a cubature formula with respect to $\delta_{\mathcal{S}_d, \mathbf{x}}$, we need to find a $2 - (d, (d-1)/2, \lambda)$ design.

Symmetric block design with parameter $(4k - 1, 2k - 1, k)$ are called Hadamard designs, and they exist whenever a Hadamard matrix of degree $4k$ exists [2, Chapter I.9]. Indeed, let A be a Hadamard matrix of degree $4k$; we can always take its first column and first row to be full of 1's. Then replace the -1 by 0 and delete the first row and the first column. This gives the incidence matrix B of a 2-$(4k - 1, 2k - 1, k)$ design. Denote by $\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,d}$ the d rows of $(\alpha - \beta)B + \beta J_{d,d}$ ($x_{i,j} \in \mathbb{R}^d$). Then $\mathbf{x}_{1,1}^2, \ldots, \mathbf{x}_{1,d}^2$ and the equal weights $\frac{1}{d}$ define a cubature formula of degree 2 with respect to $\delta_{\mathcal{S}_d, \mathbf{x}}$.

This implies that $\mathbf{x}_0^2, \mathbf{x}_{1,1}^2, \ldots, \mathbf{x}_{1,d}^2$ with equal weights $1/(d+1)$ defines a cubature formula of degree 2 with respect to ν_d. To simplify the notation, let $\mathbf{x}_{1,0} = \mathbf{x}_0$. The points $\mathbf{g}.\mathbf{x}_{1,i}$, $i = 0, \ldots, d$, $\mathbf{g} \in \mathcal{G}_d^5$, with equal weights, define a cubature formula of degree 5 with respect to μ_d. This formula has $(d+1)|\mathcal{G}_d^5|$ points when there exists a Hadamard matrix of degree $d + 1$.

If we want to find a cubature formula for a given dimension which is not the degree of a Hadamard matrix minus one, once again we choose the lower d' greater than d such that we know a Hadamard matrix of degree $d + 1$; then we project our points on a d-dimensional subspace of $\mathbb{R}^{d'}$ to get a cubature formula of degree 5 with respect to μ_d. We get once again a formula with $O(d^3)$ points.

We could have described a formula of the same form as that described in section 5.1.2, but that would lead to points outside the hypercube.

5.3. U_d, the surface of the unit sphere. We consider the Lebesgue measure μ_d on the surface of the d-dimensional unit sphere, i.e., the set of points $\mathbf{x} = (x_1, \dots, x_d)$ such that $x_1^2 + \cdots + x_d^2 = 1$. We denote by $V = \int 1 \mu_d(dx)$ the surface of the unit sphere. We propose here a solution comparable to the one proposed in section 5.1.2. Indeed, let

$$\mathbf{x}_0 = (1, 0, \dots, 0), \quad \mathbf{x}_1 = \left(\sqrt{\tfrac{1}{d}}, \dots, \sqrt{\tfrac{1}{d}}\right)$$

and

$$w_0 = \tfrac{2V}{d+2}, \quad w_1 = \tfrac{dV}{d+2}.$$

The orbits $\mathcal{GS}_d.x_0$ and $\mathcal{GS}_d.\mathbf{x}_1$, with the weights w_0 and w_1, generate a $\mathcal{GS}_d$-invariant cubature formula of degree 5 with respect to μ_d. This leads to a cubature formula of degree 5 with respect to the Lebesgue measure on the unit sphere, with $|\mathcal{G}_d^5| + 2d = O(d^2)$ points. The orbits and the weights were obtained from the cubature formula $U_n : 5 - 2$ [28, p. 294].

5.4. S_d, the unit sphere. We consider the Lebesgue measure μ_d on the unit sphere in $\mathbb{R}^d$, i.e., the set of points $\mathbf{x} = (x_1, \dots, x_d)$ such that $x_1^2 + \cdots + x_d^2 \leq 1$. We denote by $V = \int 1 \mu_d(dx)$ the volume of the unit sphere. Let

$$\mathbf{x}_0 = (r, 0, \dots, 0), \quad \mathbf{x}_1 = (s, \dots, s)$$

and

$$w_0 = \tfrac{2dV}{(d+2)(d+4)r^4}, \quad w_1 = \tfrac{V}{(d+2)(d+4)s^4},$$

where

$$r^2 = 1 - \sqrt{\tfrac{2}{d+4}}, \quad s^2 = \tfrac{d(d+4)+2\sqrt{2(d+4)}}{(d^2+2d-4)(d+4)}.$$

The orbits $\mathcal{GS}_d.x_0$ and $\mathcal{GS}_d.\mathbf{x}_1$, with the weights w_0 and w_1, generate a $\mathcal{GS}_d$-invariant cubature formula of degree 5 with respect to μ_d. This leads, once again, to a cubature formula of degree 5 with respect to the Lebesgue measure on the unit sphere, with $|\mathcal{G}_d^5| + 2d = O(d^2)$ points. The orbits and the weights were obtained from the cubature formula $S_n : 5 - 3$ [28, p. 270].

5.5. Tables. In this section, we present some tables of the same type as those found in [7], [8]. PI means that the weights are positive and the points are inside the support of the measure. EI means that the weights are all equal (and positive) with the points inside the support of the measure. We saw that the number of points depends on the existence on some combinatorial objects and thus cannot easily be expressed as a function of the dimension. We make precise for which d our cubature

formulae are very close (or equal) to the Möller lower bound.

Region	Degree	Number of points	Quality	Ref.
$E_d^{r^2}$	3	$\lvert\mathcal{G}_d^3\rvert = O(d)$ $2d$ if $\exists$ Hadamard matrix of degree d	EI	Example 4.4
C_d	3	$\lvert\mathcal{G}_d^3\rvert = O(d)$ $2d$ if $\exists$ Hadamard matrix of degree d	EI	Example 4.4
S_d	3	$\lvert\mathcal{G}_d^3\rvert = O(d)$ $2d$ if $\exists$ Hadamard matrix of degree d	EI	Example 4.4
$E_d^{r^2}$	5	$\lvert\mathcal{G}_d^5\rvert + 2d = O(d^2)$ $d^2 + 2d$ if d is a power of 4	PI	section 5.1.2
$E_d^{r^2}$	5	$O(d^3)$ Möller lower bound for $d = 7$	PI	section 5.1.1
C_d	5	$O(d^3)$	EI	section 5.2
U_d	5	$\lvert\mathcal{G}_d^5\rvert + 2d = O(d^2)$ $d^2 + 2d$ if d is a power of 4	PI	section 5.3
S_d	5	$\lvert\mathcal{G}_d^5\rvert + 2d = O(d^2)$ $d^2 + 2d$ if d is a power of 4	PI	section 5.4

For $d \in \{3, 24\}$, we determine the exact number of points in these formulae, so that one can compare them with the ones in [7], [8]. No other formulae (of degree 5) with quality PI or EI have fewer points when $d \geq 8$.

dim	$E_d^{r^2}$ 5.1.1 degree 5	C_d degree 5	$U_d, S_d, E_d^{r^2}$ 5.1.2 degree 5	$E_d^{r^2}, C_d, S_d$ degree 3
3	19	32	14	8
4	25	128	24	8
5	35	256	42	16
6	41	256	44	16
7	57	512	78	16
8	149	1536	144	16
9	189	1536	146	24
10	225	3072	276	24
11	289	3072	278	24
12	321	4096	280	24
13	417	4096	282	32
14	481	4096	284	32
15	513	4096	286	32
16	513	5120	288	32
17	1027	10240	546	36
18	1185	10240	548	36
19	1473	10240	550	36
20	1761	12288	552	36
21	2049	24576	1066	40
22	2625	24576	1068	40
23	3009	24576	1070	40
24	3201	28672	1072	40

6. Conclusion. The main but basic idea of this paper is to find some measures $\xi_1, \ldots, \xi_k$ such that their sum provides a cubature formula with respect to a given measure μ and such that it is relatively easy to find cubature formulae with respect to the measures $\xi_1, \ldots, \xi_k$. Here, we have found these measures $\xi_1, \ldots, \xi_k$ using the invariance of μ with respect to some group of symmetries and the cubature formulae with respect to $\xi_1, \ldots, \xi_k$ using some (well-known) combinatorial objects. We believe that many new cubature formulae (with positive weights, points inside the support of μ, and with few points) can be found using this method.

Acknowledgments. I would like to thank Ben Hambly for his support and the two referees for pointing out many references and helping improve the messy presentation of the first draft.

REFERENCES

[1] H. Berens, H. J. Schmid, and Y. Xu, *Multivariate Gaussian cubature formulae*, Arch. Math. (Basel), 64 (1995), pp. 26–32.
[2] T. Beth, D. Jungnickel, and H. Lenz, *Design Theory*, Vol. I, 2nd ed., Encyclopedia Math. Appl. 78, Cambridge University Press, Cambridge, UK, 1999.
[3] T. Beth, D. Jungnickel, and H. Lenz, *Design Theory*, Vol. II, 2nd ed., Encyclopedia Math. Appl. 78, Cambridge University Press, Cambridge, UK, 1999.
[4] J. Bierbrauer, S. Black, and Y. Edel, *Some t-homogeneous sets of permutations*. Des. Codes Cryptogr., 9 (1996), pp. 29–38.
[5] R. Cools, *A survey of methods for constructing cubature formulae*, in Numerical Integration (Bergen, 1991), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 357, T. O. Espelid and A. Genz, eds., Kluwer, Dordrecht, 1992, pp. 1–24.
[6] R. Cools, *Constructing cubature formulae: The science behind the art*, in Acta Numerica 6, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 1997, pp. 1–54.
[7] R. Cools, *Monomial cubature rules since "Stroud": A compilation—part* 2, J. Comput. Appl. Math., 112 (1999), pp. 21–27.
[8] R. Cools and P. Rabinowitz, *Monomial cubature rules since "Stroud": A compilation*, J. Comput. Appl. Math., 48 (1993), pp. 309–336.
[9] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed., Comput. Sci. Appl. Math., Academic Press, Orlando, 1984.
[10] P. Delsarte, *An Algebraic Approach to the Association Schemes of Coding Theory*, Philips Res. Rep. Suppl. 10, 1973.
[11] L. N. Dobrodeev, *Cubature rules with equal coefficients for integrating functions with respect to symmetric domains*, Zh. Vychisl. Mat. Mat. Fiz, 18 (1978), pp. 846–852 (in Russian); USSR Comput. Math. Math. Phys., 18(4) (1979), pp. 27–34 (in English).
[12] M. Hall, Jr., *Combinatorial Theory*, 2nd ed., Wiley-Intersci. Ser. Discrete Math., John Wiley, New York, 1986.
[13] A. R. Hammons, Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé, *The Z_4-linearity of Kerdock, Preparata, Goethals, and related codes*, IEEE Trans. Inform. Theory, 40 (1994), pp. 301–319.
[14] A. S. Hedayat, N. J. A. Sloane, and J. Stufken, *Orthogonal Arrays. Theory and Applications*, Springer Ser. Statist., Springer-Verlag, New York, 1999.
[15] Y. J. Ionin, *Applying balanced generalized weighing matrices to construct block designs*, Electron. J. Combin., 8 (2001), no. 1, Research Paper 12, 15 pp.
[16] Z. Janki and T. Van Trung, *The existence of a symmetric block design for* $(70, 24, 8)$, Mitt. Math. Sem. Giessen, 165 (1984), pp. 17–18.
[17] W. M. Kantor, *k-homogeneous groups*, Math. Z., 124 (1972), pp. 261–265.
[18] A. R. Krommer and C. W. Ueberhuber, *Computational Integration*, SIAM, Philadelphia, 1998.
[19] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd ed., Oxford University Press, Oxford, UK, 1995.
[20] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. I, North-Holland Math. Library 16, North-Holland, Amsterdam, 1977.
[21] C. J. Mitchell, *An infinite family of symmetric designs*, Discrete Math., 26 (1979), pp. 247–250.
[22] H. M. Möller, *Lower bounds for the number of nodes in cubature formulae*, in Numerische Integration (Tagung, Math. Forschungsinst., Oberwolfach, 1978), Internat. Ser. Numer. Math. 45, Birkhäuser, Basel, 1979, pp. 221–230.
[23] I. P. Mysovskikh, *Interpolatory Cubature Formulas*, Nauka, Moscow, Leningrad, 1981 (in Russian).
[24] K. Nomura, *On t-homogeneous permutation sets*, Arch. Math. (Basel), 44 (1985), pp. 485–487.
[25] M. Putinar, *A note on Tchakaloff's theorem*, Proc. Amer. Math. Soc., 125 (1997), pp. 2409–2414.
[26] S. L. Sobolev, *Cubature formulas on the sphere invariant under finite groups of rotation*, Soviet Math. Dokl., 3 (1962), pp. 1307–1310.

[27] E. Spence, *A new family of symmetric* 2-(v, k, λ) *block designs*, European J. Combin., 14 (1993), pp. 131–136.
[28] A. H. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
[29] V. Tchakaloff, *Formules de cubatures mécaniques à coefficients non négatifs*, Bull. Sci. Math. (2), 81 (1957), pp. 123–134.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 228–251

CONVERGENCE OF FINITE VOLUME APPROXIMATIONS FOR A NONLINEAR ELLIPTIC-PARABOLIC PROBLEM: A "CONTINUOUS" APPROACH*

BORIS A. ANDREIANOV†, MICHAËL GUTNIC‡, AND PETRA WITTBOLD‡

Abstract. We study the approximation by finite volume methods of the model parabolic-elliptic problem $b(v)_t = \operatorname{div}(|Dv|^{p-2}Dv)$ on $(0,T)\times\Omega \subset \mathbb{R}\times\mathbb{R}^d$ with an initial condition and the homogeneous Dirichlet boundary condition. Because of the nonlinearity in the elliptic term, a careful choice of the gradient approximation is needed. We prove the convergence of discrete solutions to the solution of the continuous problem as the discretization step h tends to 0, under the main hypotheses that the approximation of the operator $\operatorname{div}(|Dv|^{p-2}Dv)$ provided by the finite volume scheme is still monotone and coercive, and that the gradient approximation is exact on the affine functions of $x \in \Omega$. An example of such a scheme is given for a class of two-dimensional meshes dual to triangular meshes, in particular for structured rectangular and hexagonal meshes. The proof uses the rewriting of the discrete problem under a "continuous" form. This permits us to directly apply the Alt–Luckhaus variational techniques which are known for the continuous case.

Key words. doubly nonlinear elliptic-parabolic equations, finite volume methods, convergence of approximate solutions, continuous approach

AMS subject classifications. 35J60, 35K55, 35K65, 35M10, 65M12, 76M12

DOI. 10.1137/S00361429014000062

1. Introduction.

Let Ω be an open bounded polygonal domain in $\mathbb{R}^d$, $d \geq 1$, and $T > 0$. We consider the initial boundary value problem for a system of nonlinear elliptic-parabolic equations:

$$\left\{\begin{array}{ll} b(v)_t = \operatorname{div} a_p(Dv) & \text{on } Q = (0,T)\times\Omega, \\ v = 0 & \text{on } \Sigma = (0,T)\times\partial\Omega, \\ b(v)(0,\cdot) = u^0 & \text{on } \Omega, \end{array}\right. \tag{1.1}$$

where $1 < p < \infty$ and $\operatorname{div} a_p(Dv) = \operatorname{div}(|Dv|^{p-2}Dv)$ is the N-dimensional p-Laplacian, i.e.,

$$a_p : \xi = (\xi_1,\dots,\xi_N) \in (\mathbb{R}^d)^N \mapsto |\xi|^{p-2}\xi = \left(\sum_{i,j}|\xi_i^j|^2\right)^{p/2-1}(\xi_1,\dots,\xi_N) \in (\mathbb{R}^d)^N.$$

We assume that

$$\left\{\begin{array}{l} b : \mathbb{R}^N \to \mathbb{R}^N \quad \text{is continuous cyclically monotone;} \quad \text{i.e.,} \\ \text{there exists a convex differentiable function } \Phi : \mathbb{R}^N \to \mathbb{R} \ \text{ s.t. } \ b = \nabla\Phi, \end{array}\right. \tag{1.2}$$

normalized by $b(0) = 0$ and $\Phi(0) = 0$. Moreover, we assume

$$u^0 \in L^1(\Omega)^N \quad \text{with} \quad \Psi(u_0) \in L^1(\Omega), \tag{1.3}$$

*Received by the editors December 21, 2001; accepted for publication (in revised form) May 30, 2003; published electronically January 28, 2004.
http://www.siam.org/journals/sinum/42-1/40006.html
†LATP, CMI, Université de Provence, Technopole de Château-Gombert, 39, rue Frédéric Joliot-Curie, 13453 Marseille Cedex 13, France (borisa@cmi.univ-mrs.fr).
‡IRMA, Université Louis Pasteur, 7, rue René Descartes, 67084 Strasbourg Cedex, France (gutnic@math.u-strasbg.fr, wittbold@math.u-strasbg.fr).

where Ψ is the Legendre transform of Φ given by

$$\Psi : z \in \mathbb{R}^N \mapsto \sup_{\sigma \in \mathbb{R}^N} \int_0^1 (z - b(s\sigma))\sigma \, ds = \sup_{\sigma \in \mathbb{R}^N} (\sigma z - \Phi(\sigma)).$$

Equations of elliptic-parabolic type (1.1) arise as models of the flow of fluids through porous media (cf., e.g., [6, 12]). They have already been studied extensively in the literature in the last decade from a theoretical point of view (cf., e.g., [1, 21, 22, 12, 7, 26, 8, 10, 2]). Existence of weak solutions of general systems of elliptic-parabolic equations has been proved in [1], using Galerkin approximations and time-discretization. Similar results have been obtained later by other authors using different methods (e.g., using a semigroup approach as in [7, 8] in the case $N = 1$).

In particular, it is known that in the case of the system (1.1), for any u_0 satisfying (1.3), there exists a weak solution of (1.1), where the weak solution is defined as follows. Denote by E the Banach space $L^p(0,T; W_0^{1,p}(\Omega))^N$ and by E' its dual; $E' = L^{p'}(0,T; W^{-1,p'}(\Omega))^N$, where $p' = p/(p-1)$ is the conjugate exponent of p. Denote by $\langle \cdot, \cdot \rangle_{E',E}$ the duality pairing between E' and E.

DEFINITION 1.1. *A function $v \in E$ is a weak solution of the problem (1.1) if $b(v) \in L^\infty(0,T; L^1(\Omega))^N$ and $b(v)_t \in \mathcal{D}'(Q)^N$ can be extended to a functional χ on E satisfying*

$$\langle \chi, \phi \rangle_{E',E} + \iint_Q a_p(Dv) \cdot D\phi = 0 \quad \textit{for all } \phi \in E, \tag{1.4}$$

$$\langle \chi, \xi \rangle_{E',E} = -\iint_Q b(v)\, \xi_t - \int_\Omega u_0(\cdot)\, \xi(0,\cdot) \qquad \begin{array}{l} \textit{for all } \xi \in E \textit{ with} \\ \xi_t \in L^\infty(Q)^N,\ \xi(T,\cdot) = 0. \end{array} \tag{1.5}$$

Note that if v is a weak solution of (1.1), then, by the "chain rule" lemma of [1], one has

$$\begin{array}{l} B(v) \in L^\infty(0,T; L^1(\Omega))^N, \qquad \text{where} \\ B : z \in \mathbb{R}^N \mapsto b(z)z - \Phi(z) \equiv \displaystyle\int_0^1 (b(z) - b(sz))z ds \equiv \Psi(b(z)) \in \mathbb{R}. \end{array} \tag{1.6}$$

From the results of [26, 10] it also follows that, in the scalar case $N = 1$, there is uniqueness of a weak solution of (1.1). To our knowledge, the question of uniqueness is open in the case $N \geq 2$.

In this paper we study the convergence of time-implicit approximations by finite volume numerical schemes for the model nonlinear elliptic-parabolic problem (1.1). Finite volume methods are well suited for numerical simulation of processes where extensive quantities are conserved, and they are popular methods among engineers in hydrology where equations of this type arise. Therefore justification of convergence of this numerical approximation process is of particular interest. In [17] the finite volume method has been studied and convergence of this approximation procedure has been proved for problem (1.1) in the particular case $p = 2$, $N = 1$. The same method has also been studied for this equation (i.e., $p = 2, N = 1$) in the presence of an additional convection term (cf. [18, 14]), and for a nonlinear diffusion problem in [16]. To our knowledge, in the case $p \neq 2$, only the convergence of finite element methods has been studied (cf. [19, 11, 5, 20] and their references).

Let us emphasize that our main object is not only to prove the convergence of some finite volume methods for (1.1), but also to develop a "continuous" approach for this proof. The main idea of this adaptation is to rewrite the discrete finite volume scheme under an equivalent continuous form and to apply known stability techniques for the continuous equation (cf. [1] and [2, Chap. V] for the version we use) in order to get convergence of discrete solutions to a solution of the continuous problem. The "continuous" approach and the convergence result have already been presented in [3].

In section 2, we describe the finite volume schemes and in particular the admissible flux approximations we use. We show the existence and uniqueness of the solution of a finite volume scheme and give some a priori estimates on discrete solutions. Then we state the convergence result. In section 3, we show in Proposition 3.3 that the solution of a finite volume scheme, originally satisfying a discrete system of algebraic equations, also verifies a "continuous" formulation similar to (1.4), (1.5). This representation makes clear in which sense finite volume schemes approximate the elliptic operator in (1.1); we prove that this approximation is consistent. In section 4 we prove the convergence theorem, passing to the limit in the "continuous" formulation of Proposition 3.3. In section 5, we analyze the two admissibility conditions imposed in section 2. For $d = 2$, we propose a scheme on meshes dual to triangular meshes that enters into our framework; in particular, we have the convergence result on structured rectangular and hexagonal meshes.

We consider the p-Laplacian as a prototype of a class of the so-called Leray–Lions-type operators; in [4], we discuss the extension of the techniques presented above to a particular case of the p-Laplacian operator with convection, studied in [12].

In order to simplify the notation, we restrict the exposition to the scalar equation ($N = 1$). The proofs of the auxiliary results used in section 4 can be found in [4].

2. The numerical method. In order to construct approximate solutions to the problem (1.1), we will use the implicit discretization in time and a finite volume scheme in space.

2.1. Finite volume meshes, discrete gradients and finite volume schemes for the problem (1.1). Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$. A finite volume mesh $\mathcal{T}$ of Ω is given by a family of open polygonal convex subsets of Ω with positive measure, called "control volumes," a family of subsets of $\overline{\Omega}$ contained in hyperplanes of $\mathbb{R}^d$, with positive $(d-1)$-measure (these are the interfaces between control volumes), and a family of points of $\overline{\Omega}$, one per control volume (these are the "centers" of the volumes). For a volume K with center $x_K \in \overline{K}$, the interfaces contained in $\partial\Omega$ are considered as additional "boundary" volumes, unless $x_K \in \partial\Omega$.

For the sake of simplicity, we shall denote by $\mathcal{T}$ the family $(K)_{K\in\mathcal{T}}$ of control volumes; $(x_K)_{K\in\mathcal{T}}$ denotes the family of their centers. The set of all volumes K such that $x_K \in \partial\Omega$ is denoted by $\mathcal{T}_{ext}$, and the set of all volumes K with $x_K \in \Omega$ is denoted by $\mathcal{T}_{int}$. The set of interfaces $K|L$ such that K or L or both belong to $\mathcal{T}_{int}$ is denoted by $\mathcal{E}$, and $K|L$ denotes the interface between two neighbors $K, L \in \mathcal{T}$. For all $K|L$, $\widehat{KL}$ denotes the "diamond" over $K|L$, i.e., the smallest convex set of $\mathbb{R}^d$ containing $K|L$, x_K and x_L. Whenever we use K, $K|L$, or n to index objects and make summations, we mean that $K \in \mathcal{T}$, $K|L \in \mathcal{E}$, and $n \in \{1, \dots, [T/k]+1\}$, where k is the time step of the scheme.

Following [15], we give the following definition.

DEFINITION 2.1. *We say that $\mathcal{T}$ is a finite volume mesh of Ω if the following hold:*

(2.1 i) *The closure of the union of all control volumes is $\overline{\Omega}$.*

(2.1 ii) *For any* $(K, L) \in (\mathcal{T})^2$ *with* $K \neq L$, *either the* $(d-1)$*-dimensional measure of* $\overline{K} \cap \overline{L}$ *is* 0 *or* $\overline{K} \cap \overline{L} = \overline{\sigma}$ *for some* $\sigma \in \mathcal{E}$ *(in which case we denote* $\sigma = K|L = L|K$*).*

(2.1 iii) *For any* $K \in \mathcal{T}$, *there exists a subset* $\mathcal{E}_K$ *of* $\mathcal{E}$ *such that* $\partial K = \cup_{\sigma \in \mathcal{E}_K} \overline{\sigma}$. *Furthermore,* $\mathcal{E} = \cup_{K \in \mathcal{T}} \mathcal{E}_K$. *We will denote by* $\mathcal{N}_K$ *the set of volumes adjacent to* K; *i.e.,* $\mathcal{N}_K = \{L \in \mathcal{T},\ K|L \in \mathcal{E}_K\}$.

(2.1 iv) *The family of points* $(x_K)_K$ *is such that* $x_K \in \overline{K}$ *for all* $K \in \mathcal{T}$, *and it is assumed that the straight line joining* x_K *and* x_L *is orthogonal to* $K|L$ *whenever* $L \in \mathcal{N}_K$.

Denote by $\mathfrak{m}(K)$ and $\mathfrak{d}(K)$ the d-dimensional measure and the diameter of $K \in \mathcal{T}$, respectively; and denote by $m(K|L)$ the $(d-1)$-dimensional measure of $K|L \in \mathcal{E}$. A mesh $\mathcal{T}$ is characterized, in particular, by the following numbers:

$$\text{size}(\mathcal{T}) = \max_K \mathfrak{d}(K), \qquad \zeta^*(\mathcal{T}) = \min_K \min_{\sigma \in \mathcal{E}_K} \frac{\text{dist}\,(x_K, \sigma)}{\mathfrak{d}(K)},$$

$$M(\mathcal{T}) = \max_K \text{card}(\mathcal{E}_K), \qquad \zeta_*(\mathcal{T}) = \frac{\min_K \min_{\sigma \in \mathcal{E}_K} \text{dist}\,(x_K, \sigma)}{\text{size}(\mathcal{T})}.$$

A finite volume method for (1.1) requires a family $\left((\mathcal{T}^h, k^h)\right)_h$ of meshes and corresponding time steps $k^h > 0$ such that both the size of the mesh and the time step go to zero. We will assume in our notation that the family is parametrized with h in some subset of $(0, 1)$ whose closure contains zero, and $\text{size}(\mathcal{T}^h) + k^h \leq h$. A couple $(\mathcal{T}^h, k^h)$ will be called a space-time grid.

In relation to a family $\left((\mathcal{T}^h, k^h)\right)_h$, we define the numbers

$$(2.1)\quad M = \sup_h M(\mathcal{T}^h) \in \overline{\mathbb{N}}, \quad \zeta^* = \inf_h \zeta^*(\mathcal{T}^h) \in \mathbb{R}^+, \quad \text{and} \quad \zeta_* = \inf_h \zeta_*(\mathcal{T}^h) \in \mathbb{R}^+.$$

DEFINITION 2.2. *We say that the family of meshes* $(\mathcal{T}^h)_h$ *is weakly proportional if* $M < \infty$ *and* $\zeta^* > 0$. *We say that the family of meshes* $(\mathcal{T}^h)_h$ *is strongly proportional if, in addition,* $\zeta_* > 0$.

Weak proportionality is standard (cf. [18]). Strong proportionality is a technical assumption which ensures that $(\mathcal{T}^h)_h$ has the interpolation property (cf. sections 2.5 and 5.2).

Given a grid $(\mathcal{T}^h, k^h)$, to each time-space volume $Q_K^n = I^n \times K$, $I^n = (k^h(n-1), k^h n)$ one associates an unknown value $v_K^n \in \mathbb{R}^N$. In order to obtain a finite volume scheme for (1.1), one "integrates" the equation in (1.1) over each grid volume Q_K^n. The time derivative in the left-hand side is approximated by the corresponding finite difference. On the right-hand side, one uses the Green formula and then needs to replace the flux on the lateral boundary of Q_K^n by some function of the unknowns $(v_K^n)_{K,n}$. For problem (1.1), this amounts to finding a substitution for Dv in the expression $\int_{I^n \times K|L} a_p(Dv) \cdot \nu_{K,L}$ (where $\nu_{K,L}$ is the unit normal vector to $K|L$ pointing from K into L). We will assume that this substitution is in L^p on each interface $I^n \times K|L$, typically constant in time and piecewise constant in space. We therefore consider "discrete gradient" operators $\mathcal{D}^h$ of the form

$$(2.2)\qquad \begin{cases} \mathcal{D}^h : (v_K^n)_{K,n} \mapsto (D_{K|L}^n)_{K|L,n}, \\ D_{K|L}^n \in L^p(I^n \times K|L) \quad \text{for all } K|L,\ n. \end{cases}$$

It seems natural, though not necessary, to require that $\mathcal{D}^h$ be a linear operator.

A finite volume scheme for (1.1) is defined by a grid $(\mathcal{T}^h, k^h)$ and a discrete gradient $\mathcal{D}^h$ associated with the grid. Finally, a finite volume method for (1.1) is

given by a family $\big((\mathcal{T}^h, k^h, \mathcal{D}^h)\big)_h$ of grids and associated discrete gradient operators $\mathcal{D}^h$. In sections 2.3 and 2.5 we state the admissibility conditions for such methods.

Now we are able to write the equations for a scheme $(\mathcal{T}^h, k^h, \mathcal{D}^h)$:

$$\mathfrak{m}(K)\frac{b(v_K^n)-b(v_K^{n-1})}{k^h} = \sum_{L\in\mathcal{N}_K}\int_{K|L} a_p(D^n_{K|L}(x))dx\cdot\nu_{K,L} \quad \text{for all } K\in\mathcal{T}^h_{int}, \text{ for all } n\in\{1,\dots,[T/k^h]+1\}. \tag{2.3}$$

The homogeneous Dirichlet boundary condition is taken into account by assigning

$$v_K^n = 0 \quad \text{for all } K\in\mathcal{T}^h_{ext}, \quad \text{for all } n\in\{1,\dots,[T/k^h]+1\}. \tag{2.4}$$

The initial condition is given by any values $v_K^0 \in b^{-1}(u_K^0)$, where

$$u_K^0 = \frac{1}{\mathfrak{m}(K)}\int_K u^0 \quad \text{for all } K\in\mathcal{T}^h_{int}. \tag{2.5}$$

We denote by u_0^h the piecewise constant initial function $\sum_K u_K^0 \mathbb{1}_K$, where $\mathbb{1}_K$ is the characteristic function of the set K. Other choices of u_K^0 are possible, provided one has $u_0^h \to u_0$ a.e. on Ω and $\Psi(u_0^h)\to\Psi(u_0)$ in $L^1(\Omega)$ as $h\to 0$, where Ψ is defined in the introduction. These properties hold for u_K^0 given by (2.5), due to the convexity of Ψ.

We denote by $(\mathcal{S}^h)$ the system (2.3), (2.4), (2.5) corresponding to a given finite volume scheme $(\mathcal{T}^h, k^h, \mathcal{D}^h)$.

2.2. Memento on notation. In this section we collect the most used notation related to the finite volume schemes.

$\mathcal{T}$:	a finite volume mesh;
$\mathcal{T}_{ext}, \mathcal{T}_{int}$:	the set of exterior,interior control volumes;
$\mathcal{E}$:	the set of interfaces between control volumes;
K, L:	control volumes of $\mathcal{T}$;
$K\|L$:	the interface between the two neighbors K and L;
$\mathcal{E}_K$:	the set of all interfaces surrounding K;
$\mathcal{N}_K$:	the set of all neighbors of K;
x_K:	the "center" of K;
$d_{K,L}$:	the distance between x_K and x_L, $d_{K,L} = \|x_K - x_L\|$;
$d_{K,K\|L}$:	the distance between x_K and $K\|L$; one has $d_{K,K\|L} + d_{L,K\|L} = d_{K,L}$;
$\nu_{K,L}$:	the unit normal vector to $K\|L$ pointing from K to L;
$\widehat{KL}$:	the smallest convex set of $\mathbb{R}^d$ containing $K\|L, x_K$, and x_L;
$\mathfrak{d}(K), \mathfrak{m}(K)$:	the diameter and the d-dimensional measure of K, respectively;
size$(\mathcal{T})$:	the size of the mesh $\mathcal{T}$, size$(\mathcal{T}) = \max_K \mathfrak{d}(K)$;
$m(K\|L)$:	the $(d-1)$-dimensional measure of $K\|L$;
$\|R\|$:	the $(d+1)$-dimensional measure of a set $R\subset\mathbb{R}^+\times\mathbb{R}^d$;
I^n:	the time interval, $I^n = ((n-1)k, nk)$;
Q_K^n:	the time-space grid element, $Q_K^n = I^n\times K$;
Σ_K^n:	the lateral boundary of Q_K^n, $\Sigma_K^n = I^n\times\partial K$;
$\Upsilon_\varsigma(K)$:	the union of all control volumes of $(\mathcal{T})$ that are separated from K by at most $(\varsigma-1)$ other control volumes;

$\mathbb{1}_A$: the characteristic function of a set A;
$(\mathcal{T}^h, k^h, \mathcal{D}^h)$: a finite volume scheme (mesh, time step, discrete gradient);
$(\mathcal{S}^h)$: the corresponding system of equations (2.3),(2.4),(2.5);
h: the discretization parameter, $h \geq \mathrm{size}(\mathcal{T}^h) + k^h$;
M, ζ^*: the weak proportionality bounds for $(\mathcal{T}^h)_h$,
$M = \sup_h \max_K \mathrm{card}(\mathcal{N}_K)$, and $\zeta^* = \inf_h \min_K \frac{\min_{L\in\mathcal{N}_K} d_{K,K|L}}{\mathfrak{d}(K)}$;
ζ_*: the strong proportionality bound for $(\mathcal{T}^h)_h$,
$\zeta_* = \inf_h \frac{\min_{K,L\in\mathcal{N}_K} d_{K,K|L}}{\mathrm{size}(\mathcal{T})}$;
v_K^n: the unknown of the scheme $(\mathcal{S}^h)$ corresponding to the volume Q_K^n;
$\overline{v}^h$: a discrete solution for the scheme $(\mathcal{T}^h, k^h, \mathcal{D}^h)$, $\overline{v}^h = \sum_{K,n} v_K^n \mathbb{1}_{Q_K^n}$;
u_0^h: the discrete initial data, $u_0^h = \sum_K u_K^0 \mathbb{1}_K$;
$D_{K|L}^n$: the discrete gradient values on $I^n \times K|L$, $D_{K|L}^n \in L^p(I^n \times K|L)$;
$\mathcal{D}^h$: the discrete gradient operator, $\mathcal{D}^h : (v_K^n)_{K,n} \mapsto (D_{K|L}^n)_{K|L,n}$;
$D_{\perp,K|L}^n$: the value $D_{\perp,K|L}^n = \frac{v_L^n - v_K^n}{d_{K,L}}$, featuring in the "discrete $L^p(0,T;W^{1,p}(\Omega))$ norm" of $\overline{v}^h$;
$\mathcal{D}_\perp^h$: the corresponding operator, $\mathcal{D}_\perp^h : (v_K^n)_{K,n} \mapsto (D_{\perp,K|L}^n)_{K|L,n}$.

It is convenient to extend $\mathcal{D}^h$ (as well as $\mathcal{D}_\perp^h$) to an operator acting from E into $L^p(Q)$. Let $\mathcal{P}^h$ be the operator from Ω to $\bigcup_{K|L}$ which projects $x \in K$ on ∂K along the ray joining x_K to x. We define the appropriate lifting operator $\mathcal{L}^h$ and averaging operator $\mathcal{M}^h$ by

$$\mathcal{L}^h\left[(D_{K|L}^n)_{K|L,n}\right](t,x) = \sum_{K|L,n} D_{K|L}^n(\mathcal{P}^h(x))\, \mathbb{1}_{I^n \times \widehat{KL}}(t,x),$$
$$\mathcal{M}^h : \eta \in L^1(Q) \mapsto \mathcal{M}^h[\eta] = (\eta_K^n)_{K,n} \subset \mathbb{R}^N, \quad \eta_K^n = \frac{1}{|Q_K^n|} \iint_{Q_K^n} \eta.$$

We will abusively write $\mathcal{D}^h$ for the operators $\mathcal{D}^h$, $\mathcal{L}^h \circ \mathcal{D}^h$, and $\mathcal{L}^h \circ \mathcal{D}^h \circ \mathcal{M}^h$; and the same for $\mathcal{D}_\perp^h$.

The following notations, specific to the "continuous" approach, are introduced in sections 2.5 and 3.1.

u^h: the continuous in t interpolation of $b(\overline{v}^h)$, affine on each time interval I^n;
v^h: an interpolated solution in E for $\overline{v}^h$ (cf. Definition 2.8);
$\mathcal{G}^h$: the interpolated gradient operator produced by $(\mathcal{T}^h, k^h, \mathcal{D}^h)$ (cf. Definition 3.2);
$\mathcal{A}$: the elliptic operator in (1.1), $\mathcal{A} : \eta \in E \mapsto -\mathrm{div}\, a_p(D\eta) \in E'$;
$\mathcal{A}^h$: the finite volume approximation of $\mathcal{A}$ produced by the scheme $(\mathcal{T}^h, k^h, \mathcal{D}^h)$, given by $\mathcal{A}^h : \eta \in E \mapsto -\mathrm{div}\, a_p(\mathcal{G}^h[\eta]) \in E'$.

2.3. Admissible flux approximations. For simplicity, we consider only the gradients that yield fully implicit schemes; in this case $\mathcal{D}^h, \mathcal{D}_\perp^h$ act independently on each set $\left(v_K^n\right)_K$, and the dependence on n does not matter for their definition.

Let us introduce the operator $\mathcal{D}_\perp^h$, which appears naturally in the a priori estimates of section 2.4:

$$\mathcal{D}_\perp^h : (v_K)_K \mapsto (D_{\perp,K|L})_{K|L}, \qquad D_{\perp,K|L} = \frac{v_L - v_K}{d_{K,L}} \in \mathbb{R}. \tag{2.6}$$

For $\varsigma \in \mathbb{N}$, denote by $\Upsilon_\varsigma(K)$ the union of all control volumes of $\mathcal{T}$ that are separated from K by at most $(\varsigma-1)$ other control volumes; for instance, $\Upsilon_1(K) = \bigcup_{L\in\mathcal{N}_K} L$. The choice of ς corresponds to the choice of control volumes that are really involved in the construction of $\mathcal{D}^h$ on ∂K.

Now we can make precise the assumptions on discrete gradient operators of the form (2.2).

DEFINITION 2.3. *Let $(\mathcal{T}^h, \mathcal{D}^h)_h$ be a family of finite volume meshes and corresponding discrete gradient operators. The gradient approximation provided by $\mathcal{D}^h$ is admissible if the following hold.*

(2.3 i) *$\mathcal{D}^h$ is linear and injective;*

(2.3 ii) *$\mathcal{D}^h$ provides a strictly monotone scheme; i.e., for all $(v_K)_K, (\widetilde{v}_K)_K \subset (\mathbb{R}^d)^N$ that do not coincide,*

$$\frac{1}{d}\sum_{K|L}\Big((v_L - v_K) - (\widetilde{v}_L - \widetilde{v}_K)\Big)\int_{K|L}\Big(a_p(D_{K|L}(x)) - a_p(\widetilde{D}_{K|L}(x))\Big)dx\cdot\nu_{K,L} > 0,$$

where $(D_{K|L})_{K|L} = \mathcal{D}^h\big[(v_K)_K\big]$, $(\widetilde{D}_{K|L})_{K|L} = \mathcal{D}^h\big[(\widetilde{v}_K)_K\big]$;

(2.3 iii) *$\mathcal{D}^h$ provides a scheme coercive at zero; i.e., there exists a constant $C_* > 0$, independent of h, such that for all $(v_K)_K \subset (\mathbb{R}^d)^N$ and $(D_{K|L})_{K|L} = \mathcal{D}^h\big[(v_K)_K\big]$, one has*

$$\frac{1}{d}\sum_{K|L}\Big(v_L - v_K\Big)\int_{K|L} a_p(D_{K|L}(x))dx\cdot\nu_{K,L} \geq C_*\Big\|\mathcal{D}^h_\perp[(v_K)_K]\Big\|^p_{L^p(\Omega)};$$

and there exists $\varsigma \in \mathbb{N}$, independent of h, such that the following hold.

(2.3 iv) *For each h, $\mathcal{D}^h$ is consistent with affine functions. More exactly, assume that, for $K \in \mathcal{T}^h$ given, there exists an affine function w on Ω such that $v_L = \frac{1}{\mathfrak{m}(L)}\int_L w$ whenever $L \subset \Upsilon_\varsigma(K)$. Then $D_{K|L}(x) = Dw = \text{const}$ for all $x \in K|L$ for all $L \in \mathcal{N}_K$.*

(2.3 v) *There exists a constant C^*, independent of h, such that, for all $\widetilde{K} \in \mathcal{T}^h$ and all sets of values $(v_K)_K$ of $\mathbb{R}^N$,*

$$\int_{\widetilde{K}}\Big|\mathcal{D}^h[(v_K)_K]\Big|^p \leq C^*\int_{\Upsilon_\varsigma(\widetilde{K})}\Big|\mathcal{D}^h_\perp[(v_K)_K]\Big|^p.$$

Conditions (2.3 ii) and (2.3 iv) imply strong restrictions on the gradient approximation. We provide some examples of methods with admissible gradient approximation in section 5.1.

2.4. Discrete solutions. Recall that we consider as unknowns the values v_K^n on $K \in \mathcal{T}_{int}$, assigning v_K^n to be zero in $K \in \mathcal{T}_{ext}$. We will repeatedly use the following "summation by parts" formula (cf., e.g., [15]).

Remark 2.4. Let $\mathcal{T}$ be a finite volume mesh of Ω in the sense of Definition 2.1. Let $(v_K)_{K\in\mathcal{T}} \subset \mathbb{R}^N$, $(F_{K,L})_{(K,L)\in\mathcal{T}^2} \subset \mathbb{R}^N$. Assume $v_K = 0$ for all $K \in \mathcal{T}_{ext}$ and $F_{K,L} = -F_{L,K}$ for all $K|L \in \mathcal{E}$. Then

$$\sum_K v_K \sum_{L\in\mathcal{N}_K} F_{K,L} = \sum_{K|L}(v_K - v_L)\,F_{K,L}.$$

If $(v_K^n)_{K,n}$ verifies $(\mathcal{S}^h)$, we say that the function $\overline{v}^h = \sum_{K,n} v_K^n \mathbb{1}_{Q_K^n}$ is the corresponding discrete solution. We prove the discrete version of the $L^p(0,T;W_0^{1,p}(\Omega))$- a priori estimate on $\overline{v}^h$ (which is exactly the estimate on $\mathcal{D}_\perp^h[\overline{v}^h]$ in L^p), and the discrete version of (1.6).

PROPOSITION 2.5. *Let $\big((\mathcal{T}^h, k^h)\big)_h$ be a family of finite volume grids and let $\big(\mathcal{D}^h\big)_h$ be a family of corresponding discrete gradient operators satisfying property* (2.3 iii) *of Definition* 2.3. *Then, for any solution $\overline{v}^h$ of the discrete problem $(\mathcal{S}^h)$, there exists a constant C which depends only on p, d, Ω, T, on C_* in* (2.3 iii), *and on $\|\Psi(u_0)\|_{L^1(\Omega)}$ such that*

$$\text{(i)} \qquad \left\|\mathcal{D}_\perp^h[\overline{v}^h]\right\|_{L^p(Q)}^p = \frac{1}{d}\sum_{K|L,n} m_{(K|L)} d_{K,L}\left|\frac{v_L^n - v_K^n}{d_{K,L}}\right|^p \le C;$$

$$\text{(ii)} \qquad \left\|B(\overline{v}^h)\right\|_{L^\infty(0,T;L^1(\Omega))} = \sup_{n\in\{1,\dots,[T/k^h]+1\}} \sum_K \mathfrak{m}(K)\, B(v_K^n) \le C.$$

Proof. Take $i \in \{1,\dots,[T/k^h]+1\}$ and multiply each term in (2.3) by v_K^i. By (2.3 iii), using Remark 2.4 and (2.4), one gets

$$\sum_K \mathfrak{m}(K)(b(v_K^i) - b(v_K^{i-1}))\, v_K^i + C_* k^h d \int_\Omega \left|\mathcal{D}_\perp^h[\overline{v}^h]\right|^p \le 0.$$

By the convexity of Φ, one has $(b(v_K^i) - b(v_K^{i-1}))\, v_K^i \ge B(v_K^i) - B(v_K^{i-1})$. Summing over i from 1 to $n \in \{1,\dots,[T/k^h]+1\}$ and taking into account the convexity of Ψ, we infer

$$\sum_K \mathfrak{m}(K) B(v_K^n) + C_* d \int_0^{nk^h}\int_\Omega \left|\mathcal{D}_\perp^h[\overline{v}^h]\right|^p \le \sum_K \mathfrak{m}(K)\Psi(u_K^0) = \sum_K \mathfrak{m}(K)\Psi\left(\frac{1}{\mathfrak{m}(K)}\int_K u^0\right) \le \int_\Omega \Psi(u^0). \qquad \square$$

Next, let us prove the discrete version of the Poincaré inequality and of the compact embedding of $W^{1,p}(\Omega)$ in $L^1(\Omega)$. Note that we do not need any proportionality assumptions on the mesh.

LEMMA 2.6. *Let $\Omega \subset \mathbb{R}^d$ be a polygonal domain of diameter $\mathfrak{d}(\Omega)$, and let $\mathcal{T}$ be a finite volume mesh of Ω. Let $\overline{v}^h = \sum_K v_K \mathbb{1}_K$ such that $(v_K)_{K\in\mathcal{T}} \subset \mathbb{R}$ and $v_K = 0$ for all $K \in \mathcal{T}_{ext}$. Then there exists a constant C which depends only on p and d such that*

$$\text{(i)} \qquad \|\overline{v}^h\|_{L^p(\Omega)} \le C\,\mathfrak{d}(\Omega)\left\|\mathcal{D}_\perp^h[\overline{v}^h]\right\|_{L^p(\Omega)};$$

$$\text{(ii)} \qquad \text{for all } \Delta > 0, \quad \sup_{|\Delta x|\le\Delta}\int_{\mathbb{R}^d} |\overline{v}^h(x+\Delta x) - \overline{v}^h(x)|\, dx \le \Delta \times \left\|\mathcal{D}_\perp^h[\overline{v}^h]\right\|_{L^1(\Omega)}.$$

Proof. (i) For $x \in \Omega$, set $\psi_{K|L}(x) = 1$ in the case that the orthogonal projection of $K|L$ on the hyperplane $\{x_1 = 0\}$ contains $(0, x_2, \dots, x_d)$, and set $\psi_{K|L}(x) = 0$ otherwise. One has

$$|\overline{v}^h(x)|^p \le \frac{1}{2}\sum_{K|L}\psi_{K|L}(x)\left||v_L|^p - |v_K|^p\right| \le C\sum_{K|L}\psi_{K|L}(x)\, d_{K,L}\frac{|v_L - v_K|}{d_{K,L}}\Big(|v_K|^{p-1} + |v_L|^{p-1}\Big).$$

Since $\int_\Omega \psi_{K|L}(x)\,dx \le m(K|L)\,\mathfrak{d}(\Omega)$, one has by the Hölder inequality

$$\int_\Omega |\overline{v}^h(x)|^p\,dx \le C\mathfrak{d}(\Omega)\left(\sum_{K|L}\frac{1}{d}\,m(K|L)\,d_{K,L}\Big|\frac{v_L-v_K}{d_{K,L}}\Big|^p\right)^{\frac{1}{p}}\left(\sum_{K|L}\frac{1}{d}\,m(K|L)\,d_{K,L}\big(|v_K|^p+|v_L|^p\big)\right)^{\frac{p-1}{p}}.$$

Denote $h=\mathrm{size}(\mathcal{T})$. Assertion (i) will follow by the Young inequality if we show that

$$\sum_{K|L}\frac{1}{d}\,m(K|L)\,d_{K,L}\big(|v_K|^p+|v_L|^p\big) \le (1+2^p)\sum_K \mathfrak{m}(K)\,|v_K|^p + 2(2h)^p\sum_{K|L}\frac{1}{d}\,m(K|L)\,d_{K,L}\Big|\frac{v_L-v_K}{d_{K,L}}\Big|^p, \tag{2.7}$$

since $h\le\mathfrak{d}(\Omega)$. Denote by R the left-hand side of (2.7). We have $d_{K,L}=d_{K,K|L}+d_{L,K|L}$; thus

$$R=\sum_K \mathfrak{m}(K)|v_K|^p+\sum_{K|L}\frac{1}{d}m(K|L)(|v_K|^p d_{L,K|L}+|v_L|^p d_{K,K|L}).$$

Note that

$$|v_K|^p d_{L,K|L}\le\begin{cases}2^p|v_L|^p d_{L,K|L} & \text{if } |v_K|\le 2|v_L|,\\ (2h)^p\Big|\frac{v_L-v_K}{d_{K,L}}\Big|^p d_{K,L} & \text{otherwise.}\end{cases}$$

Indeed, if $|v_K|>2|v_L|$, one has $|v_L-v_K|>\frac12|v_K|$ so that

$$|v_K|^p d_{L,K|L}\le|v_K|^p d_{K,L}\le 2^p|v_L-v_K|^p d_{K,L}\le 2^p h^p\Big|\frac{v_L-v_K}{d_{K,L}}\Big|^p d_{K,L}.$$

Using the same argument for $|v_L|^p d_{K,K|L}$, we obtain the desired estimate (2.7).

(ii) Now for $x\in\mathbb{R}^d$, set $\overline{\psi}_{K|L}(x)=1$ in the case where the segment $[x,x+\Delta x]$ crosses $K|L$, and set $\overline{\psi}_{K|L}(x)=0$ otherwise. Note that $\int_{\mathbb{R}^d}\overline{\psi}_{K|L}(x)\,dx\le m(K|L)\,\Delta$; hence (ii) follows, since

$$\int_{\mathbb{R}^d}|\overline{v}^h(x)-\overline{v}^h(x+\Delta x)|\,dx\le\int_{\mathbb{R}^d}\sum_{K|L}\overline{\psi}_{K|L}(x)|v_L-v_K|\,dx \le\Delta\sum_{K|L}m(K|L)d_{K,L}\Big|\frac{v_L-v_K}{d_{K,L}}\Big|.$$ ∎

Now we can state the result for existence and uniqueness of a discrete solution.

THEOREM 2.7. *Let $\mathcal{T}^h$ be a finite volume mesh of Ω, $k^h>0$, and let $\mathcal{D}^h$ be a discrete gradient associated to $\mathcal{T}^h$. If $\mathcal{D}^h$ satisfies* (2.3 iii)*, there exists a solution $(v_K^n)_{K,n}$ to the discrete problem $(\mathcal{S}^h)$. If $\mathcal{D}^h$ satisfies* (2.3 ii)*, the solution is unique.*

Proof of Theorem 2.7. Using Remark 2.4 and the coercivity of the scheme, we apply the Brouwer fixed point theorem and get existence. Uniqueness follows from the monotonicity of $b(\cdot)$ and the strict monotonicity of the scheme. See [3] for more detailed proofs. ∎

2.5. Interpolation property and main result. Consider a family of finite volume schemes $\big((\mathcal{T}^h, k^h, \mathcal{D}^h)\big)_h$ such that h tends to 0. Let $(v_K^n)_{K,n}$ be a solution to the scheme $(\mathcal{S}^h)$ and $\overline{v}^h$ the corresponding discrete solution.

We require the existence of what will be called "interpolated solutions" for $\overline{v}^h$, denoted by v^h, such that $v^h \in E$; these should be close to $\overline{v}^h$ (asymptotically as $h \to 0$) and satisfy the a priori estimate in E analogous to the estimate of Proposition 2.5(i) on $\overline{v}^h$. Moreover, the values $(v_K^n)_{K,n}$ should be recoverable from v^h. To this end, we require $\mathcal{M}^h[v^h] = (v_K^n)_{K,n}$.

DEFINITION 2.8. *A family of grids $\big(\mathcal{T}^h, k^h\big)_h$ has the interpolation property in E if, for any family $(\overline{v}^h)_h$ of functions such that $\overline{v}^h|_{Q_K^n} = v_K^n \equiv \text{const}$ for each $K \in \mathcal{T}^h$, $n \in \{1, \dots, [T/k^h]+1\}$, with $v_K^n = 0$ for $K \in \mathcal{T}_{ext}^h$ and with $\big\|\mathcal{D}_\perp^h[\overline{v}^h]\big\|_{L^p(Q)} \le C$ for all h, there exists a family $(v^h)_h \subset E$ such that*

$$\|v^h - \overline{v}^h\|_{L^p(Q)} \to 0 \quad \text{as} \quad h \to 0, \tag{2.8}$$

$$\mathcal{M}^h[v^h] = (v_K^n)_{K,n}, \tag{2.9}$$

$$\|v^h\|_E \le I(C) \quad \text{with some function } I : \mathbb{R}^+ \mapsto \mathbb{R}^+ \text{ independent of } h. \tag{2.10}$$

If $\overline{v}^h$ is a solution to a finite volume scheme, we say that v^h is an interpolated solution for $\overline{v}^h$.

The interpolation property is the main technical assumption required by the "continuous" approach. In section 5.2 we give two conditions ensuring this property. Now let us state the main result of this paper.

THEOREM 2.9. *Let $\big((\mathcal{T}^{h_m}, k^{h_m}, \mathcal{D}^{h_m})\big)_{m\in\mathbb{N}}$ be a sequence of finite volume schemes, where $k^{h_m} + \text{size}(\mathcal{T}^{h_m}) \le h_m \to 0$ as $m \to \infty$. Assume that the family of meshes is weakly proportional, the gradient approximation is admissible, and the interpolation property holds (cf. Definitions* 2.2, 2.3, and 2.8*).*

For $m \in \mathbb{N}$, let $\overline{v}^{h_m}$ be a discrete solution of (S^{h_m}). Then there exists a subsequence $(h_{m_l})_{l\in\mathbb{N}}$ such that $\overline{v}^{h_{m_l}} \rightharpoonup v$ in $L^p(Q)$ as $l \to \infty$, where v is a weak solution of the problem (1.1)*.*

Note that it suffices to strengthen slightly assumption (2.3 ii) of Definition 2.3 in order to get the strong convergence of $\overline{v}^{h_{m_l}}$ to v in $L^p(Q)$ (cf. [4, Corollary 1]). Moreover, in the case when $N = 1$ the whole sequence converges to the unique solution of (1.1). In this case error estimates can be proved (cf., e.g., [15] for the linear case), but this is not the purpose of the present paper.

In what follows, we write k instead of k^h and omit subscripts in sequences (h_m) and(h_{m_l}), simply writing that h tends to zero.

3. The "continuous" approach. Take the discrete solution $\overline{v}^h = \sum_{K,n} v_K^n \mathbb{1}_{Q_K^n}$ produced by the finite volume scheme $(\mathcal{S}^h)$. Let $v^h \in E$ be a corresponding interpolated solution. We will show that there exist functions $u^h \in L^1(Q)$ and $G^h \in L^p(Q)$ such that $u^h(0,\cdot) = u_0^h(\cdot)$ and $u^h{}_t = \operatorname{div} a_p(G^h)$ in the weak sense of Definition 1.1, and the functions u^h, G^h can be recovered from the interpolated solution. More exactly, we prove in Proposition 3.3 below that $u^h{}_t \in \mathcal{D}'$ can be extended to $\chi^h \in E'$ and

$$\langle \chi^h, \phi\rangle_{E',E} + \iint_Q a_p(\mathcal{G}^h[v^h]) \cdot D\phi = 0 \quad \text{for all} \quad \phi \in E, \tag{3.1}$$

$$\langle \chi^h, \xi \rangle_{E',E} = -\iint_Q u^h\, \xi_t - \int_\Omega u_0^h(\cdot)\, \xi(0,\cdot) \qquad \begin{array}{l}\text{for all } \xi \in E \text{ with} \\ \xi_t \in L^\infty(Q)^N,\ \xi(T,\cdot) = 0,\end{array} \tag{3.2}$$

with an operator $\mathcal{G}^h : E \mapsto L^p(Q)$ to be defined.

The analogy of (3.1), (3.2) with (1.4), (1.5) in Definition 1.1 plays the key role in the proof of the convergence result of Theorem 2.9.

3.1. Interpolated gradient and the "continuous" form of the scheme. First define u^h as the piecewise affine in t interpolation of $b(\overline{v}^h)$:

$$u^h = \sum_{K,n} \left(b(v_K^n) + \frac{t - kn}{k} (b(v_K^n) - b(v_K^{n-1})) \right) \mathbb{1}_{Q_K^n}. \tag{3.3}$$

Then (3.2) holds, since $u^h(0,\cdot) = u_0^h(\cdot)$ and the piecewise constant function ${u^h}_t$ extends to $\chi^h \in E'$ by

$$\langle \chi^h, \phi \rangle_{E',E} = \iint_Q {u^h}_t\, \phi \quad \text{for all } \phi \in E. \tag{3.4}$$

Next, note that in $(\mathcal{S}^h)$ the numerical flux is prescribed on the boundary of each control volume; we will extend it to Q as follows. For given $K \in \mathcal{T}^h_{int}$,$n$ and a function $F_K^n : \partial K \mapsto \mathbb{R}$, consider the following Neumann problem in the factor space $\mathcal{W}(K) = W^{1,p}(K)/\mathbb{R}$:

$$\begin{cases} \operatorname{div} a_p(Dw) = \dfrac{1}{\mathfrak{m}(K)} \displaystyle\sum_{K \in \mathcal{N}_K} \int_{K|L} F_K^n \quad \text{on } K, \\ a_p(Dw) \cdot \nu_K|_{\partial K} = F_K^n, \end{cases} \tag{3.5}$$

where ν_K is the exterior unit normal vector to ∂K. For $K \in \mathcal{T}^h_{ext}$ with $\mathfrak{m}(K) > 0$, we drop in (3.5) the Neumann boundary condition on $\partial K \cap \partial\Omega$ and seek $w \in W^{1,p}(K)$ with $w|_{\partial K \cap \partial\Omega} = 0$.

LEMMA 3.1. *Let $F_K^n \in L^{p'}(\partial K)$ ($F_K^n \in L^{p'}(\partial K \setminus \partial\Omega)$, if $K \in \mathcal{T}^h_{ext}$). Then* (3.5) *admits a unique solution.*

The proof is standard, using the coercivity and monotonicity argument [25, Chap. 2, Th. 2.1]. Now we can introduce the interpolated gradient operator.

DEFINITION 3.2. *The interpolated gradient operator $\mathcal{G}^h : E \mapsto L^p(Q)$ maps $\eta \in E$ into $\mathcal{G}^h[\eta]$ given by*

$$\begin{cases} \mathcal{G}^h[\eta] = \sum_{K,n} D\eta_K^n \mathbb{1}_{Q_K^n}, \quad \text{where} \quad \eta_K^n \in \mathcal{W}(K) \ \text{ solves} \\ -\displaystyle\int_K a_p(D\eta_K^n) \cdot D\varphi + \sum_{L \in \mathcal{N}_K} \int_{K|L} \varphi\, a_p(D^n_{K|L}) \cdot \nu_K = \frac{1}{\mathfrak{m}(K)} \int_K \varphi \sum_{L \in \mathcal{N}_K} \int_{K|L} a_p(D^n_{K|L}) \cdot \nu_K \\ \text{for all } \varphi \in W^{1,p}(K) \ (\text{for all } \varphi \in W^{1,p}(K) \text{ with } \varphi|_{\partial K \cap \partial\Omega} = 0, \ \text{in case } K \in \mathcal{T}^h_{ext}) \\ \text{and the values } D^n_{K|L}(x) \text{ are given by } (D^n_{K|L})_{K|L,n} = \mathcal{D}^h[\eta]. \end{cases}$$

If $\overline{v}^h$ solves $(\mathcal{S}^h)$, we set $G^h = \mathcal{G}^h[\overline{v}^h]$. We remark that $G^h = \mathcal{G}^h[v^h]$ by property (2.9) of interpolated solutions v^h. We show that (3.1) follows from (3.4) and the conservation of fluxes.

PROPOSITION 3.3. *Assume that $(v_K^n)_{K,n}$ is a solution of $(\mathcal{S}^h)$. Let v^h be a corresponding interpolated solution, let u^h and χ^h be defined by* (3.3) *and* (3.4)*, respectively,*

and let $\mathcal{G}^h$ be the interpolated gradient operator of Definition 3.2. *Then* (3.1), (3.2) *hold.*

Proof. It remains to check (3.1). By (3.3), for all $K \in \mathcal{T}^h$ and n, we have

$$u^h{}_t - \operatorname{div} a_p(G^h) = \frac{b(v_K^n) - b(v_K^{n-1})}{k} - \frac{1}{\mathfrak{m}(K)} \sum_{L \in \mathcal{N}_K} \int_{K|L} a_p(D^n_{K|L}(x))dx \cdot \nu_{K,L} = 0$$

everywhere on Q_K^n because of (2.3). Therefore, using (3.4) and integrating by parts in each Q_K^n, we have

$$\begin{aligned}\langle \chi^h, \phi\rangle_{E',E} + \int_Q a_p(\mathcal{G}^h[v^h]) \cdot D\phi \iint_Q u^h{}_t \phi + a_p(G^h) \cdot D\phi \\ = \sum_{K,n} \iint_{Q_K^n} (u^h{}_t - \operatorname{div} a_p(G^h))\phi + \sum_{K,n} \sum_{L \in \mathcal{N}_K} \iint_{I^n \times K|L} \phi\, a_p(D^n_{K|L}) \cdot \nu_{K,L} \\ = 0 + \sum_{K|L,n} \iint_{I^n \times K|L} \phi\, a_p(D^n_{K|L}) \cdot (\nu_{K,L} + \nu_{L,K}) = 0. \qquad \square\end{aligned}$$

3.2. Properties of the interpolated gradient and consistency. In view of (3.1) and (1.4), it is natural to compare the elliptic operator in (1.1),

$$\mathcal{A} : \eta \in E \mapsto -\operatorname{div} a_p(D\eta) \in E', \tag{3.6}$$

with the operators

$$\mathcal{A}^h : \eta \in E \mapsto -\operatorname{div} a_p(\mathcal{G}^h[\eta]) \in E'. \tag{3.7}$$

Indeed, $\mathcal{A}^h$ can be considered as the finite volume approximation of $\mathcal{A}$, whence the following definition.

DEFINITION 3.4. *Let $\big((\mathcal{T}^h, k^h, \mathcal{D}^h)\big)_h$ be a family of finite volume schemes for the problem* (1.1), *with* $\operatorname{size}(\mathcal{T}^h) + k^h \le h \to 0$. *We say that the approximation of* (1.1) *by these schemes is consistent if, for all $\eta \in E$, one has $\mathcal{A}^h[\eta] \to A[\eta]$ in E' as $h \to 0$.*

In this section we prove the following result.

THEOREM 3.5. *Let $\big((\mathcal{T}^h, k^h, \mathcal{D}^h)\big)_h$ be a family of finite volume schemes with a weakly proportional family of meshes and an admissible gradient approximation (cf. Definitions* 2.2 *and* 2.3*). Then it provides a consistent approximation of* (1.1), *in the sense of Definition* 3.4.

The proof of Theorem 3.5 is based upon the following properties of the interpolated gradient operator $\mathcal{G}^h$.

PROPOSITION 3.6. *Let $\big((\mathcal{T}^h, k^h, \mathcal{D}^h)\big)_h$ be a family of finite volume schemes with admissible gradient approximation and weakly proportional family of meshes.*

(i) *There exists a constant C such that for all $\eta \in E$ and $H \subset Q$ such that $H = \bigcup_{i=1}^m Q_{K_i}^{n_i}$,*

$$\iint_H \left|\mathcal{G}^h[\eta]\right|^p \le C \iint_{\Upsilon_{\varsigma+1}(H)} |D\eta|^p,$$

where $\Upsilon_{\varsigma+1}(H) = \bigcup_{i=1}^m I^{n_i} \times \Upsilon_{\varsigma+1}(K_i)$. In particular, $\big(\mathcal{G}^h\big)_h$ are uniformly bounded on E and

$$\left\|\mathcal{G}^h[\eta]\right\|_{L^p(Q)} \le C \|\eta\|_E. \tag{3.8}$$

(ii) *The operators $(\mathcal{G}^h)_h$ are locally equicontinuous on E. More exactly, there exists a constant $C(R)$ such that, whenever $\|\eta\|_E \le R$ and $\|\mu\|_E \le R$,*

$$\big\|\mathcal{G}^h[\eta] - \mathcal{G}^h[\mu]\big\|_{L^p(Q)} \le C(R)\big(\|\eta-\mu\|_E\big)^{\min\{p-1,\frac{1}{p-1}\}}, \tag{3.9}$$

$$\big\|a_p(\mathcal{G}^h[\eta]) - a_p(\mathcal{G}^h[\mu])\big\|_{L^{p'}(Q)} \le C(R)\big(\|\eta-\mu\|_E\big)^{\min\{(p-1)^2,\frac{1}{p-1}\}}. \tag{3.10}$$

In the statement above and in the rest of this section, C denotes a generic constant that depends only on p, d, Ω, on M, ζ^* of (2.1), and on C_*, C^*, ς of Definition 2.3, unless the additional dependence on R is specified. The proof uses the standard properties of the function $a_p(\cdot)$ (cf. [19, 12]): for all $y_1, y_2 \in \mathbb{R}^d$,

$$\begin{cases} |a_p(y_1)-a_p(y_2)|^{p'} \le C\,|y_1-y_2|^p, & 1<p\le 2;\\ |a_p(y_1)-a_p(y_2)|^{p'} \le C\,|y_1-y_2|^{p'}\Big(|y_1|^p+|y_2|^p\Big)^{\frac{p-2}{p-1}}, & p\ge 2; \end{cases} \tag{3.11}$$

$$\begin{cases} |y_1-y_2|^p \le C\left[(a_p(y_1)-a_p(y_2))\cdot(y_1-y_2)\right]^{\frac{p}{2}}\Big[|y_1|^p+|y_2|^p\Big]^{\frac{2-p}{2}}, & 1<p\le 2;\\ |y_1-y_2|^p \le C\,(a_p(y_1)-a_p(y_2))\cdot(y_1-y_2), & p\ge 2. \end{cases} \tag{3.12}$$

Before turning to the proofs of Proposition 3.6 and Theorem 3.5, note the following three lemmas.

LEMMA 3.7. *Let $K \subset \mathbb{R}^d$ be a bounded convex domain of $\mathbb{R}^d$ of diameter $\mathfrak{d}(K)$ and d-dimensional measure $\mathfrak{m}(K)$. Assume that K contains a ball of radius $\zeta^*\mathfrak{d}(K) > 0$. Then there exists a constant C such that, assigning $\overline{w} = \frac{1}{\mathfrak{m}(K)}\int_K w$, one has*

$$\int_{\partial K} |w-\overline{w}|^p \le C\,(\mathfrak{d}(K))^{p-1}\int_K |Dw|^p$$

for all $w \in W^{1,p}(K)$, where $w|_{\partial K}$ is understood in the sense of traces.

Proof. Applying, e.g., the proofs of [13, Theorems 59, 60, and 76] with $p=2$ replaced by a general $p \in (1,+\infty)$, we obtain the claim of the lemma with C depending on p, d, and the Lipschitz continuity of ∂K. Due to the convexity of K, C actually depends only on p, d, and ζ^*. ∎

LEMMA 3.8. *Let $(\mathcal{T}^h)_h$ be a weakly proportional family of meshes, and let $(\mathcal{D}^h_\perp)_h$ be the operators defined by* (2.6)*. Then there exists a constant C such that for all K, n for all $\eta \in E$,*

$$\iint_{Q_K^n}\Big|\mathcal{D}^h_\perp[\eta]\Big|^p \le C\iint_{I^n\times\Upsilon_1(K)}|D\eta|^p.$$

Proof. Let $\big(\eta_K^n\big)_{K,n} = \mathcal{M}^h[\eta]$ and $\eta^n_{K|L} = \frac{1}{km(K|L)}\int_{I^n\times K|L}\eta$ in the sense of traces. By definition,

$$\begin{aligned}\iint_{Q_K^n}\Big|\mathcal{D}^h_\perp[\eta]\Big|^p &= \sum_{K|L}\frac{1}{d}\,k\,m(K|L)\,d_{K,K|L}\left|\frac{\eta_L^n-\eta_K^n}{d_{K,L}}\right|^p\\ &\le C\sum_{K|L}\frac{1}{d}\,k\,m(K|L)\,d_{K,K|L}\left(\frac{\big|\eta^n_{K|L}-\eta^n_K\big|^p}{(d_{K,K|L})^p}+\frac{\big|\eta^n_{K|L}-\eta^n_L\big|^p}{d_{K,K|L}\,(d_{L,K|L})^{p-1}}\right)\\ &\le C\sum_{K|L}\frac{1}{d}\,k\,m(K|L)\,d_{K,K|L}\left|\frac{\eta^n_{K|L}-\eta^n_K}{d_{K,K|L}}\right|^p + C\sum_{K|L}\frac{1}{d}\,k\,m(K|L)\,d_{L,K|L}\left|\frac{\eta^n_{K|L}-\eta^n_L}{d_{L,K|L}}\right|^p.\end{aligned}$$

By the convexity of the function $z \mapsto |z|^p$ and Lemma 3.7,

$$k\, m(K|L)\, |\eta^n_{K|L} - \eta^n_K|^p \le \iint_{I^n \times K|L} |\eta - \eta^n_K|^p \le C\, \mathfrak{d}(K)^{p-1} \iint_{Q^n_K} |D\eta|^p,$$

and the same holds if K and L are exchanged. Hence by (2.1) we have

$$\iint_{Q^n_K} \left|\mathcal{D}^h_\perp[\eta]\right|^p \le C \sum_{L \in \mathcal{N}_K} \left(\iint_{Q^n_K} |D\eta|^p + \iint_{Q^n_L} |D\eta|^p \right) \le C \iint_{I^n \times \Upsilon_1(K)} |D\eta|^p. \qquad \square$$

LEMMA 3.9. *Let $\left((\mathcal{T}^h, k^h, \mathcal{D}^h)\right)_h$ be a family of finite volume schemes with a weakly proportional family of meshes and an admissible gradient approximation. Then the following hold:*

(i) *For all $R > 0$ there exists a constant $C(R)$ such that, whenever $\|\eta\|_E \le R$ and $\|\mu\|_E \le R$,*

$$\sum_{K,n} \mathfrak{d}(K) \iint_{\Sigma^n_K} \left| a_p(\mathcal{D}^h[\eta]) - a_p(\mathcal{D}^h[\mu]) \right|^{p'} \le C(R) \Big(\|\eta - \mu\|_E \Big)^{\min\{p,p'\}}.$$

(ii) *There exists a constant C such that for all $\eta \in E$ and $H, \Upsilon_{\varsigma+1}(H)$ as in Proposition 3.6 one has*

$$\sum_{i=1}^m \mathfrak{d}(K_i) \iint_{\Sigma^{n_i}_{K_i}} \left|\mathcal{D}^h[\eta]\right|^p \le C \iint_{\Upsilon_{\varsigma+1}(H)} |D\eta|^p.$$

Proof. First take K and consider $\varphi^K = \left|a_p(\mathcal{D}^h[\eta]) - a_p(\mathcal{D}^h[\mu])\right|^{p'} \in L^1(\partial K)$. Recall that the values of $\mathcal{D}^h$ have been extended from ∂K inside K by means of the projection operator $\mathcal{P}^h$ (cf. section 2.2). Hence by (2.1) we have

$$\begin{aligned} \mathfrak{d}(K) \int_{\partial K} |\varphi^K| &= \mathfrak{d}(K) \sum_{L \in \mathcal{N}_K} \int_{K|L} |\varphi^K| \\ &= d \sum_{L \in \mathcal{N}_K} \frac{\mathfrak{d}(K)}{d_{K,K|L}} \int_{\widehat{KL} \cap K} |\varphi^K \circ \mathcal{P}^h| \le \frac{d}{\zeta^*} \int_K |\varphi^K \circ \mathcal{P}^h|. \end{aligned} \tag{3.13}$$

If $1 < p \le 2$, (3.13) and (3.11) yield

$$\sum_{K,n} \mathfrak{d}(K) \iint_{\Sigma^n_K} \left| a_p(\mathcal{D}^h[\eta]) - a_p(\mathcal{D}^h[\mu]) \right|^{p'} \le C \sum_{K,n} \iint_{Q^n_K} \left| \mathcal{D}^h[\eta] - \mathcal{D}^h[\mu] \right|^p.$$

In turn, (2.3 i), (2.3 iv), Lemma 3.8, and (2.1) imply that

$$\begin{aligned} \sum_{K,n} \iint_{Q^n_K} \left|\mathcal{D}^h[\eta] - \mathcal{D}^h[\mu]\right|^p &\le C \sum_{K,n} \iint_{I^n \times \Upsilon_\varsigma(K)} \left|\mathcal{D}^h_\perp[\eta - \mu]\right|^p \\ &\le C \sum_{K,n} \iint_{I^n \times \Upsilon_{\varsigma+1}(K)} \left|D(\eta - \mu)\right|^p \le C \iint_Q |D\eta - D\mu|^p = C \left(\|\eta - \mu\|_E \right)^p, \end{aligned} \tag{3.14}$$

which was the claim of (i) for $1 < p \le 2$. Furthermore, we remark that (3.14) also holds for $p \ge 2$, in particular with $\eta = 0$ or $\mu = 0$. Therefore for $p \ge 2$, using (3.13), and then (3.11) and the Hölder inequality, we get

$$\sum_{K,n} \mathfrak{d}(K) \iint_{\Sigma^n_K} \left| a_p(\mathcal{D}^h[\eta]) - a_p(\mathcal{D}^h[\mu]) \right|^{p'} \le C \left(\sum_{K,n} \iint_{Q^n_K} \left| \mathcal{D}^h[\eta] - \mathcal{D}^h[\mu] \right|^p \right)^{\frac{p'}{p}} \times \Big(R^p \Big)^{\frac{p-2}{p-1}}.$$

Thus, in the case $p \geq 2$, (i) also follows from (3.14).

The proof of (ii) is similar, using the identity $|a_p(y)|^{p'} = |y|^p$ instead of inequalities (3.11). ∎

Proof of Proposition 3.6. Recalling Definition 3.2, for all $\kappa \in \mathcal{T}_{int}$ and n, denote by η_κ^n (respectively, by μ_κ^n) a function in $W^{1,p}(\kappa)$ that solves (3.5) with $F_\kappa^n(x) = a_p(D_{\kappa|L}^n(x))\cdot\nu_\kappa$ for $x \in \kappa|L \in \mathcal{E}_\kappa$, where $(D_{\kappa|L}^n)_{\kappa|L,n} = \mathcal{D}^h[\eta]$ (respectively, $(D_{\kappa|L}^n)_{\kappa|L,n} = \mathcal{D}^h[\mu]$). In other words, each of $\eta_\kappa^n, \mu_\kappa^n$ verifies the integral identity corresponding to (3.5) with all test functions in $W^{1,p}(\kappa)$. Taking for the test function $(\eta_\kappa^n - \mu_\kappa^n)$, subtracting the two identities, and integrating in $t \in \imath^n$, we obtain

$$\begin{aligned}\iint_{Q_\kappa^n} \Big(a_p(D\eta_\kappa^n) - a_p(D\mu_\kappa^n)\Big)\cdot\Big(D\eta_\kappa^n - D\mu_\kappa^n\Big) \\ = \int_{\imath^n}\int_{\partial\kappa} \Big(\eta_\kappa^n - \mu_\kappa^n - \overline{\eta_\kappa^n - \mu_\kappa^n}\Big)\Big(a_p(\mathcal{D}^h[\eta]) - a_p(\mathcal{D}^h[\mu])\Big)\cdot\nu_\kappa,\end{aligned} \tag{3.15}$$

where $\overline{\eta_\kappa^n - \mu_\kappa^n} = \frac{1}{\mathfrak{m}(\kappa)}\int_\kappa \eta_\kappa^n - \mu_\kappa^n$ for a.a. $t \in \imath^n$. Summing over κ, n, using the Hölder inequality and Lemmas 3.7 and 3.9(i), we have from (3.15)

$$\begin{aligned}&\iint_Q \Big(a_p(\mathcal{G}^h[\eta]) - a_p(\mathcal{G}^h[\mu])\Big)\cdot\Big(\mathcal{G}^h[\eta] - \mathcal{G}^h[\mu]\Big) \\ &\quad\leq \sum_{\kappa,n}\int_{\imath^n}\int_{\partial\kappa} \mathfrak{d}(\kappa)^{\frac{-1}{p'}}\,|\,\eta_\kappa^n - \mu_\kappa^n - \overline{\eta_\kappa^n - \mu_\kappa^n}\,| \\ &\qquad\qquad\times \mathfrak{d}(\kappa)^{\frac{1}{p'}}\Big|a_p(\mathcal{D}^h[\eta]) - a_p(\mathcal{D}^h[\mu])\Big| \\ &\quad\leq \left(\sum_{\kappa,n}\int_{\imath^n} \mathfrak{d}(\kappa)^{1-p}\int_{\partial\kappa} |\,\eta_\kappa^n - \mu_\kappa^n - \overline{\eta_\kappa^n - \mu_\kappa^n}\,|^p\right)^{\frac{1}{p}} \\ &\qquad\qquad\times\left(\sum_{\kappa,n}\mathfrak{d}(\kappa)\iint_{\Sigma_\kappa^n}\Big|a_p(\mathcal{D}^h[\eta]) - a_p(\mathcal{D}^h[\mu])\Big|^{p'}\right)^{\frac{1}{p'}} \\ &\quad\leq \left(\sum_{\kappa,n}\iint_{Q_\kappa^n}|D\eta_\kappa^n - D\mu_\kappa^n|^p\right)^{\frac{1}{p}}\|\eta-\mu\|_E^{\min\{p/p',1\}} \\ &\qquad= \Big\|\mathcal{G}^h[\eta] - \mathcal{G}^h[\mu]\Big\|_{L^p(Q)}\|\eta-\mu\|_E^{\min\{p/p',1\}}.\end{aligned} \tag{3.16}$$

In the same manner, taking $\mu = 0$ and using Lemma 3.9(ii), we get

$$\iint_H \Big|\mathcal{G}^h[\eta]\Big|^p C\,\left(\iint_H\Big|\mathcal{G}^h[\eta]\Big|^p\right)^{\frac{1}{p}}\left(\iint_{\Upsilon_{\varsigma+1}(H)}|D\eta|^p\right)^{\frac{1}{p'}},$$

which proves (i).

Now if $1 < p \leq 2$, (3.12), (3.16), and the Hölder inequality yield

$$\begin{aligned}&\Big\|\mathcal{G}^h[\eta] - \mathcal{G}^h[\mu]\Big\|_{L^p(Q)}^p \\ &\quad\leq C\left(\Big\|\mathcal{G}^h[\eta] - \mathcal{G}^h[\mu]\Big\|_{L^p(Q)}\|\eta-\mu\|_E^{\frac{p}{p'}}\right)^{\frac{p}{2}}\left(\Big\|\mathcal{G}^h[\eta]\Big\|_{L^p(Q)}^p + \Big\|\mathcal{G}^h[\mu]\Big\|_{L^p(Q)}^p\right)^{\frac{2-p}{2}}.\end{aligned}$$

Using (3.8), we obtain (3.9). Now (3.10) follows by (3.11).

If $p \geq 2$, (3.12) and (3.16) readily yield (3.9); hence (3.10) follows by (3.11). ∎

Proof of Theorem 3.5. We have to prove that $\left\|a_p(D\eta) - a_p(\mathcal{G}^h[\eta])\right\|_{L^{p'}(Q)} \to 0$ as $h \to 0$.

Let us first prove the theorem for the case of $\eta \in E$ that is piecewise constant in t and piecewise affine in x. Let $J \subset Q$ be the set of discontinuities of $D\eta$. Clearly, J is of finite d-dimensional Hausdorff measure $\mathcal{H}^d(J)$.

For ς given in Definition 2.3, let us introduce $H^h = \bigcup_{\{K,n \,|\, I^n \times \Upsilon_\varsigma(K) \cap J \neq \emptyset\}} Q_K^n$. Note that $|H^h| \le (\varsigma+1)h\,\mathcal{H}^d(J) \to 0$ as $h \to 0$; likewise, $|\Upsilon_{\varsigma+1}(H^h)| \to 0$ as $h \to 0$. Therefore, by Proposition 3.6(i), we have

$$\iint_{H^h} \left|a_p(D\eta) - a_p(\mathcal{G}^h[\eta])\right|^{p'} \le C \left(\iint_{H^h} |D\eta|^p + \iint_{\Upsilon_{\varsigma+1}(H^h)} |D\eta|^p \right) \to 0$$

as $h \to 0$. Moreover, for all Q_K^n such that $Q_K^n \cap H^h = \emptyset$, we have $\mathcal{G}^h[\eta] \equiv D\eta$ on Q_K^n. Indeed, we have $D[\eta] \equiv \text{const}$ on $\Upsilon_{\varsigma+1}(Q_K^n)$. Therefore $\mathcal{D}^h[\eta]|_{Q_K^n} \equiv D\eta = \text{const}$ by property (2.3 iv) of admissible gradient approximations. Hence $Dw = D\eta$ satisfies the boundary condition in (3.5); the equation is also satisfied, since $\operatorname{div} a_p(D\eta) \equiv 0$ on K and $\frac{1}{\mathfrak{m}(K)} \int_{\partial K} a_p(\mathcal{D}^h[\eta]) \cdot \nu_K = a_p(D\eta) \cdot \int_{\partial K} \nu_K = 0$.

It follows that $\left\|a_p(D\eta) - a_p(\mathcal{G}^h[\eta])\right\|_{L^{p'}(Q)} \to 0$ as $h \to 0$, which was our claim.

Now let us approximate an arbitrary function η in E by functions η_l that are piecewise constant in t and piecewise affine in x. Note that we can always choose this sequence η_l in E such that $\eta_l \to \eta$ in E and a.e. on Q as $l \to \infty$, and $|D\eta_l|^p$ are dominated by an $L^1(Q)$ function independent of l. We have

$$\begin{aligned} \left\|a_p(D\eta) - a_p(\mathcal{G}^h[\eta])\right\|_{L^{p'}(Q)} &\le \left\|a_p(D\eta) - a_p(D\eta_l)\right\|_{L^{p'}(Q)} \\ &+ \left\|a_p(D\eta_l) - a_p(\mathcal{G}^h[\eta_l])\right\|_{L^{p'}(Q)} + \left\|a_p(\mathcal{G}^h[\eta_l]) - a_p(\mathcal{G}^h[\eta])\right\|_{L^{p'}(Q)}. \end{aligned} \tag{3.17}$$

As $l \to 0$, the first term in the right-hand side of (3.17) converges to zero by the Lebesgue dominated convergence theorem, independently of h. The second one converges to zero as $h \to 0$ for all l fixed. Finally, by Proposition 3.6(ii), the third one converges to zero as $l \to \infty$ uniformly in h. Hence the result follows. ∎

4. Proof of Theorem 2.9. In the context of continuous dependence upon the data of weak solutions to "general" elliptic-parabolic problems (cf. [2, Chap.V]), the proof of convergence of weak solutions of approximating problems is based upon the three essential arguments (A), (B), and (C) below.

(A) A priori estimates, using (1.2) and the Alt–Luckhaus chain rule lemma (cf. [1, 26, 10]).

(B) Strong compactness in the parabolic term, using a variant of the Kruzhkov lemma (cf. [23]):

LEMMA 4.1 (cf. [4], [2, Chap. V]). *Let Ω be an open domain in $\mathbb{R}^d$, $Q = (0,T)\times\Omega$, and let the families of functions $(u^h)_h, (F_\alpha^h)_{h,\alpha}$ be bounded in $L^1(Q)$ and satisfy $\frac{\partial}{\partial t} u^h = \sum_{|\alpha| \le m} D^\alpha F_\alpha^h$ in $\mathcal{D}'(Q)$. Assume that u^h can be extended by zero outside Q, and one has*

$$\sup_{|\Delta x| \le \Delta} \iint_{\mathbb{R}^{d+1}} |u^h(t, x+\Delta x) - u^h(t,x)|\, dxdt \le \omega(\Delta), \quad \text{with } \lim_{\Delta \to 0} \omega(\Delta) = 0, \tag{4.1}$$

where $\omega(\cdot)$ does not depend on h. Then $(u^h)_h$ is relatively compact in $L^1(Q)$.

(C) Convergence in the elliptic term, using a variant of the Minty–Browder argument (cf., e.g., [25]).

LEMMA 4.2 (cf. [4], [2, Chap. V]). *Let E be a Banach space, E' its dual and $\langle\cdot,\cdot\rangle_{E',E}$ denote the duality product of elements of E' and E. Let $(v^h)_h \subset E$ and $v^h \rightharpoonup v$ as $h \to 0$. Let $\mathcal{A}^h$ be a sequence of monotone operators from E to E' such that $\mathcal{A}^h[v^h] \overset{*}{\rightharpoonup} -\chi$ for some $\chi \in E'$. Assume that $\mathcal{A}^h$ converge pointwise to some operator $\mathcal{A}$, and $\mathcal{A}$ is hemicontinuous (i.e., continuous in the weak-$*$ topology of E' along each direction). Assume that*

$$\liminf_{h\to 0} \langle \mathcal{A}^h[v^h], v^h\rangle_{E',E} \quad \leq \quad \langle -\chi, v\rangle_{E',E}. \tag{4.2}$$

Then $\chi + \mathcal{A}[v] = 0$, and (4.2) necessarily holds with equality.

Taking advantage of the "continuous" form (3.1), (3.2) of the discrete problem $(\mathcal{S}^h)$, we can prove the convergence of finite volume approximate solutions in the same way, using the discrete a priori estimates shown in Propositions 2.5 and 3.6(i), using next Lemma 4.1, and then using finally Lemma 4.2 together with the essential consistency result of Theorem 3.5.

Proof of Theorem 2.9. Let $\overline{v}^h$ be the solution of $(\mathcal{S}^h)$. Let v^h be a corresponding interpolated solution, and let $\mathcal{A}^h$ be the finite volume approximate of the operator $\mathcal{A}$ in (1.1) (cf. (3.6), (3.7)). Note that all the convergences we state below take place up to extraction of a subsequence.

(A) By Proposition 2.5(i), $\left\|\mathcal{D}^h_\perp[\overline{v}^h]\right\|_{L^p(Q)} \leq$ const uniformly in h so that the family $(v^h)_h$ is bounded in E, by (2.10). Hence there exists a function $v \in E$ such that $v^h \rightharpoonup v$ in E as $h \to 0$. By (2.8), one also has $\overline{v}^h \rightharpoonup v$ in $L^p(Q)$.

(B) We claim that the family $(u^h)_h$ given by (3.3) is relatively compact in $L^1(Q)$. Indeed, let us check the assumptions of Lemma 4.1. We have ${u^h}_t = \operatorname{div} a_p(\mathcal{G}^h[v^h])$ in $\mathcal{D}'(Q)$ by (3.1), (3.2), and the family $\left(a_p(\mathcal{G}^h[v^h])\right)_h$ is bounded in $L^{p'}(Q)$ by Proposition 2.5(i), equation (2.10), and Proposition 3.6(i) (note that $\left(\mathcal{A}^h[v^h]\right)_h$ is thus bounded in E'). Furthermore, (3.3) yields

$$\|u^h\|_{L^1(Q)} \leq 2 \iint_Q |b(\overline{v}^h)| + k^h \sum_K \mathfrak{m}(K)|u^0_K|,$$

and one has $|b(z)| \leq \delta B(z) + \sup_{|\zeta|\leq 1/\delta} |b(\zeta)|$ for all $\delta > 0$ (cf., e.g., [1]). By Proposition 2.5(ii) and since $u^h_0 = \sum_K \mathfrak{m}(K)|u^0_K| \to u_0$ in $L^1(\Omega)$ as $h \to 0$, it follows that $(u^h)_h$ is bounded in $L^1(Q)$.

Finally, by Proposition 2.5(i) and Lemma 2.6(ii), we obtain (4.1) with u^h replaced by $\overline{v}^h$. Hence the estimate (4.1) for u^h follows by (3.3), as in the continuous case (cf. [1]); see [4] for the detailed proof.

Thus the claim of (B) follows, and there exists a function $u \in L^1(Q)$ such that $u^h \to u$ in $L^1(Q)$ and a.e. on Q. In addition, we claim that $u = b(v)$, where v is the weak limit of v^h in E. It suffices to show that $\overline{v}^h \rightharpoonup v$ in $L^1(Q)$ and $b(\overline{v}^h) \to u$ in $L^1(Q)$, and then apply the monotonicity argument of [9]; see [4] for the detailed proof.

(C) By (A), we have $v^h \rightharpoonup v$ in E. We claim that v is a weak solution of (1.1).

By Proposition 3.3, $\chi^h + \mathcal{A}^h[v^h] = 0$ in E' and the initial condition (3.2) is verified for all h. The family $\left(\mathcal{A}^h[v^h]\right)_h$ is bounded in E' (cf. (B)), thus $(\chi^h)_h$ is weak-$*$ relatively compact in E'. By (3.4), (B), and Definition 1.1, we also have $\chi^h = {u^h}_t \to b(v)_t = \chi$ in $\mathcal{D}'(Q)$. Hence $\mathcal{A}^h[v^h] = -\chi^h \overset{*}{\rightharpoonup} -\chi$ in E'.

Moreover, passing to the limit in (3.2), using (B) and the convergence of u_0^h to u_0 in $L^1(\Omega)$, we get (1.5). Consequently, by the chain rule argument [1, Lemma 1.5] we have

$$\langle -\chi, v\rangle_{E',E} = -\int_\Omega \Psi(b(v(T,\cdot))) + \int_\Omega \Psi(u^0). \tag{4.3}$$

On the other hand, by (3.4), (3.3), (2.9), and the monotonicity of $b(\cdot)$, we have

$$\begin{aligned}
\langle -\chi^h, v^h\rangle_{E',E} &= -\frac{1}{k}\sum_{K,n}(b(v_K^n) - b(v_K^{n-1}))\iint_{Q_K^n} v^h \\
&= -\sum_{K,n} \mathfrak{m}(K)(b(v_K^n) - b(v_K^{n-1}))\, v_K^n \\
&\le -\sum_K \mathfrak{m}(K)\Psi(b(v_K^{[T/k^h]+1})) + \sum_K \mathfrak{m}(K)\Psi(u_K^0) \\
&= -\int_\Omega \Psi(b(\overline{v}^h(T,\cdot))) + \int_\Omega \Psi(u_0^h).
\end{aligned}$$

Recall that $\Psi(u_0^h) \to \Psi(u_0)$ in $L^1(\Omega)$. Without loss of generality, we can assume that $\overline{v}^h(T,\cdot) \to v(T,\cdot)$ a.e. on Ω; hence by the Fatou lemma and (4.3) we get (4.2).

Next, the operators $\mathcal{A}^h$ are monotone. Indeed, take $\varphi \in E$ and $(\varphi_K^n)_{K,n} = \mathcal{M}^h[\varphi]$. Arguing as in the proof of Proposition 3.3, integrating by parts in Q_K^n, and cancelling the boundary terms, we get

$$\begin{aligned}
\langle \mathcal{A}^h[\eta], \varphi\rangle_{E',E} &= \iint_Q a_p(\mathcal{G}^h[\eta])\cdot D\varphi = -\sum_{K,n}\iint_{Q_K^n} \varphi \operatorname{div} a_p(\mathcal{G}^h[v^h]) \\
&= -\sum_{K,n}\iint_{Q_K^n} \varphi \times \frac{1}{\mathfrak{m}(K)}\sum_{L\in\mathcal{N}_K}\int_{K|L} a_p(D_{K|L}^n(x))dx\cdot\nu_{K,L} \\
&= -k\sum_{K,n}\varphi_K^n \sum_{L\in\mathcal{N}_K}\int_{K|L} a_p(D_{K|L}^n(x))dx\cdot\nu_{K,L}.
\end{aligned} \tag{4.4}$$

Substituting (4.4) and applying Remark 2.4, we infer by property (2.3 ii) of Definition 2.3 that

$$\begin{aligned}
&\langle \mathcal{A}^h[\eta] - \mathcal{A}^h[\widetilde{\eta}], \eta - \widetilde{\eta}\rangle_{E',E} \\
&\quad = \frac{1}{d}\,k\sum_{K|L,n}\Big((\eta_L^n - \eta_K^n) - (\widetilde{\eta}_L^n - \widetilde{\eta}_K^n)\Big)\int_{K|L}\Big(a_p(D_{K|L}^n(x)) - a_p(\widetilde{D}_{K|L}^n(x))\Big)dx\cdot\nu_{K,L} \ge 0,
\end{aligned}$$

where $(\eta_K^n)_{K,n} = \mathcal{M}^h[\eta]$, $(D_{K|L}^n)_{K|L,n} = \mathcal{D}^h[\eta]$, and the same for $\widetilde{\eta}$.

Finally, by Theorem 3.5, $\mathcal{A}^h$ converge pointwise to the hemicontinuous operator $\mathcal{A}\cdot = -\operatorname{div} a_p(D\,\cdot)$. By Lemma 4.2 we conclude that $\chi + \mathcal{A}[v] = 0$ in E'. Thus (1.4) holds and v is a weak solution of (1.1). ∎

5. Examples of admissible methods. By an admissible method, we mean a method which provides an admissible gradient approximation and weakly proportional meshes satisfying the interpolation property. Recall that in this case, we have the convergence of the finite volume approximation (cf. Theorem 2.9). In this section, we prove that such admissible methods exist.

5.1. On discrete gradients. In this section we construct an admissible gradient for a family $(\mathcal{T}^h)_h$ of finite volume meshes of the Voronoï kind dual to a family $(\widehat{\mathcal{T}}^h)_h$ of triangular meshes.

Let us introduce some notation. We use $\widehat{o}$ to denote a triangle of the mesh $\widehat{\mathcal{T}}^h$; for all $\widehat{o} \in \widehat{\mathcal{T}}^h$, there exist $K, L, M \in \mathcal{T}^h$ such that $\widehat{o} = \Delta x_K x_L x_M$ (the triangle with the corners x_K, x_L, x_M). The three interfaces $K|L$, $L|M$, $M|K$ intersect at point $x_{\widehat{o}}$, which is the center of the circumscribed circle of triangle $\widehat{o}$. We require it to be inside $\widehat{o}$. Let us denote by $S_{\widehat{o}}, S_{K,L}, S_{L,M}$, and $S_{M,K}$ the surfaces of $\Delta x_K x_L x_M, \Delta x_{\widehat{o}} x_K x_L, \Delta x_{\widehat{o}} x_L x_M$, and $\Delta x_{\widehat{o}} x_M x_K$, respectively. One has $S_{\widehat{o}} = S_{K,L} + S_{L,M} + S_{M,K}$.

Recall that $\nu_{K,L} = \overrightarrow{x_K x_L}/d_{K,L}, \nu_{L,M} = \overrightarrow{x_L x_M}/d_{L,M}, \nu_{M,K} = \overrightarrow{x_M x_K}/d_{M,K}$. Note the following elementary lemma.

LEMMA 5.1. *Let $\widehat{o} = \Delta x_K x_L x_M$ be a triangle in $\mathbb{R}^2$, let $x_{\widehat{o}}$ be the center of its circumscribed circle, and let $x_{\widehat{o}} \in \Delta x_K x_L x_M$. With the above notation, for all r in $\mathbb{R}^2$, we have*

$$r = \frac{2}{S_{\widehat{o}}} \Big\{ S_{K,L}(r \cdot \nu_{K,L})\nu_{K,L} + S_{L,M}(r \cdot \nu_{L,M})\nu_{L,M} + S_{M,K}(r \cdot \nu_{M,K})\nu_{M,K} \Big\}.$$

This property can be generalized to any polygon in $\mathbb{R}^2$ which admits the circumscribed circle.

Furthermore, for $\widehat{o} \in \widehat{\mathcal{T}}^h$ such that $\widehat{o} = \Delta x_K x_L x_M$, let ${v^h}_{\widehat{o}}^0 : \mathbb{R}^2 \to \mathbb{R}^N$ be the affine function that takes the values v_K, v_L, v_M at the points x_K, x_L, x_M, respectively. The discrete gradient operator $\mathcal{D}^{h,0} = \mathcal{L}^h \circ \mathcal{D}^{h,0}$ is defined by

$$\mathcal{D}^{h,0} : (v_K)_K \mapsto \sum_{\widehat{o} \in \widehat{\mathcal{T}}^h} D{v^h}_{\widehat{o}}^0(x) \mathbb{1}_{\widehat{o}}(x).$$

In the case of structured hexagonal meshes, as well as that of structured rectangular ones, the family $(\mathcal{D}^{h,0})_h$ is admissible (this will be proved in Proposition 5.2, as a particular case). In general, this construction does not work. Indeed, if the points x_K are not the barycenters of $K \in \mathcal{T}^h$, property (2.3 iv) fails.

This can be overcome, for instance, in the following way. For all $K \in \mathcal{T}^h$, let y_K be the barycenter of K and set $\sigma_K = x_K - y_K$. For $\widehat{o} \in \widehat{\mathcal{T}}^h$ such that $\widehat{o} = \Delta x_K x_L x_M$, let $v^h_{\widehat{o}} : \mathbb{R}^2 \to \mathbb{R}^N$ be the affine function that takes the values v_K, v_L, v_M at the points y_K, y_L, y_M, respectively. The discrete gradient operator $\mathcal{D}^h = \mathcal{L}^h \circ \mathcal{D}^h$ is defined by

$$\mathcal{D}^h : (v_K)_K \mapsto \sum_{\widehat{o} \in \widehat{\mathcal{T}}^h} Dv^h_{\widehat{o}}(x) \mathbb{1}_{\widehat{o}}(x); \tag{5.1}$$

i.e., the affine interpolation over the triangle $\Delta y_K y_L y_M$ is actually used in the triangle $\Delta x_K x_L x_M$.

We will take advantage of considering $\mathcal{D}^h$ as a perturbation of $\mathcal{D}^{h,0}$. For all $\widehat{o} \in \widehat{\mathcal{T}}^h$, let us define the correction operators

$$R_{\widehat{o}} : r \in \mathbb{R}^2 \mapsto \frac{2}{S_{\widehat{o}}} \Bigg\{ S_{K,L} \left(r \cdot \frac{\sigma_L - \sigma_K}{d_{K,L}} \right) \nu_{K,L} + S_{L,M} \left(r \cdot \frac{\sigma_M - \sigma_L}{d_{L,M}} \right) \nu_{L,M} + S_{M,K} \left(r \cdot \frac{\sigma_K - \sigma_M}{d_{M,K}} \right) \nu_{M,K} \Bigg\},$$

with the notation introduced above. We need to guarantee that the Euclidiean norm of $R_{\widehat{O}}$ is less than $\min\{p-1, 1/(p-1)\}$ for all $\widehat{O} \in \mathcal{T}^h$.

PROPOSITION 5.2. *Assume that $(\mathcal{T}^h)_h$ is a family of meshes dual to a family of meshes $(\widehat{\mathcal{T}}^h)_h$ such that all $\widehat{O} \in \widehat{\mathcal{T}}^h$ are triangles with angles less than or equal to $\pi/2$.*

Assume that for all h, for all $\widehat{O} \in \widehat{\mathcal{T}}^h$,

$$\frac{2}{S_{\widehat{O}}}\left\{ S_{K,L}\frac{|\sigma_L - \sigma_K|}{d_{K,L}} + S_{L,M}\frac{|\sigma_M - \sigma_L|}{d_{L,M}} + S_{M,K}\frac{|\sigma_K - \sigma_M|}{d_{M,K}} \right\} < \min\{p-1, 1/(p-1)\},$$

where σ_K is the difference between the "center" x_K of the volume K and its barycenter, etc.

Then the family of discrete gradient operators $(\mathcal{D}^h)_h$ on $(\mathcal{T}^h)_h$ defined by (5.1) *is admissible in the sense of Definition* 2.3.

Proof. Since, for any affine function w on K, one has $\frac{1}{\mathfrak{m}(K)}\int_K w(x)dx = w(y_K)$, where y_K is the barycenter of K, property (2.3 iv) holds for $\mathcal{D}^h$ (with $\varsigma = 1$, by construction). Next, (2.3 i) is clear.

Let us establish the relation between $\mathcal{D}^{h,0}$ and $\mathcal{D}^h$. Denote by $D^0_{\widehat{O}}, D_{\widehat{O}}$ the values on $\widehat{O}$ of $\mathcal{D}^{h,0}[(v_K)_K]$ and $\mathcal{D}^h[(v_K)_K]$, respectively. Let us show that for all $\widehat{O} \in \widehat{\mathcal{T}}^h$,

$$D^0_{\widehat{O}} = (I - R_{\widehat{O}})D_{\widehat{O}}. \tag{5.2}$$

Indeed, if $\widehat{O} = \Delta x_K x_L x_M$, one has

$$\begin{aligned} D_{\widehat{O}} \cdot \nu_{K,L} &= \frac{v^h_{\widehat{O}}(x_L) - v^h_{\widehat{O}}(x_K)}{d_{K,L}} = \frac{(v_L + D_{\widehat{O}} \cdot \sigma_L) - (v_K + D_{\widehat{O}} \cdot \sigma_K)}{d_{K,L}} \\ &= \frac{v_L - v_K}{d_{K,L}} + D_{\widehat{O}} \cdot \frac{\sigma_L - \sigma_K}{d_{K,L}} = D^0_{\widehat{O}} \cdot \nu_{K,L} + D_{\widehat{O}} \cdot \frac{\sigma_L - \sigma_K}{d_{K,L}}. \end{aligned}$$

Writing the same relation for L, M and M, K, from Lemma 5.1, we get $D_{\widehat{O}} - D^0_{\widehat{O}} = R_{\widehat{O}} D_{\widehat{O}}$, whence (5.2) follows.

By Lemma 5.1 and the definition of $\mathcal{D}^h_\perp = \mathcal{L}^h \circ \mathcal{D}^h_\perp$ for all $\widehat{O} \in \widehat{\mathcal{T}}^h$ such that $\widehat{O} = \Delta x_K x_L x_M$ we have

$$\begin{aligned} \int_{\widehat{O}} \left|\mathcal{D}^{h,0}[(v_K)_K]\right|^p &= S_{\widehat{O}} \left|\mathcal{D}^{h,0}[(v_K)_K]\right|^p \\ &\le C^* \left\{ S_{K,L}\left|\frac{v_L - v_K}{d_{K,L}}\right|^p + S_{L,M}\left|\frac{v_M - v_L}{d_{L,M}}\right|^p + S_{M,K}\left|\frac{v_K - v_M}{d_{M,K}}\right|^p \right\} \\ &= C^* \int_{\widehat{O}} \left|\mathcal{D}^h_\perp[(v_K)_K]\right|^p \end{aligned} \tag{5.3}$$

with a constant C^* that depends only on p. Since for given $\widetilde{K} \in \mathcal{T}^h$ and $\widehat{O} \in \widehat{\mathcal{T}}^h$ we have $\widetilde{K} \cap \widehat{O} \neq \emptyset$ if and only if $\widehat{O} \in \Upsilon_1(\widetilde{K})$, it follows that property (2.3 v) holds for the discrete gradient $\mathcal{D}^{h,0}$, with $\varsigma = 1$. Now set $\theta_{\widehat{O}} = \|R_{\widehat{O}}\|$. We have $\theta_{\widehat{O}} < 1$. One has $|D_{\widehat{O}}| \le \|(I - R_{\widehat{O}})^{-1}\|\,|D^0_{\widehat{O}}| \le \frac{1}{1-\theta_{\widehat{O}}}|D^0_{\widehat{O}}|$; therefore (2.3 v) also holds for $\mathcal{D}^h$.

Next, each term in the sum in (2.3 iii) splits into two terms corresponding to the two parts of the interface $K|L$ included in different triangles $\widehat{O}_1, \widehat{O}_2 \in \widehat{\mathcal{T}}^h$. Let us write down all the terms corresponding to the same triangle $\widehat{O} \in \widehat{\mathcal{T}}^h$, $\widehat{O} = \Delta x_K x_L x_M$,

combine them using Lemma 5.1, and estimate using (5.2):

$$\begin{aligned}
&S_{K,L}(a_p(D_{\widehat{o}})\cdot\nu_{K,L})(D^0_{\widehat{o}}\cdot\nu_{K,L}) + S_{L,M}(a_p(D_{\widehat{o}})\cdot\nu_{L,M})(D^0_{\widehat{o}}\cdot\nu_{L,M})\\
&\quad + S_{M,K}(a_p(D_{\widehat{o}})\cdot\nu_{M,K})(D^0_{\widehat{o}}\cdot\nu_{M,K}) = \frac{S_{\widehat{o}}}{2}a_p(D_{\widehat{o}})\cdot D^0_{\widehat{o}}\\
&\qquad = \frac{S_{\widehat{o}}}{2}a_p(D_{\widehat{o}})\cdot(I-R_{\widehat{o}})D_{\widehat{o}} \geq \frac{1-\theta_{\widehat{o}}}{2}S_{\widehat{o}}|D_{\widehat{o}}|^p \geq \frac{1-\theta_{\widehat{o}}}{2(1+\theta_{\widehat{o}})^p}S_{\widehat{o}}|D^0_{\widehat{o}}|^p.
\end{aligned}$$

Property (2.3 iii) for $\mathcal{D}^h$ follows, because one has $|D^0| \geq |D^0_{\widehat{o}}\cdot\nu_{K,L}| = \frac{|v_L - v_K|}{d_{K,L}} = D^h_{\perp,K|L}$ so that

$$\sum_{\widehat{o}\in\widehat{\mathcal{T}}^h} S_{\widehat{o}}|D^0_{\widehat{o}}|^p = \left\|\mathcal{D}^{h,0}[(v_K)_K]\right\|^p_{L^p(\Omega)} \geq \left\|\mathcal{D}^h_\perp[(v_K)]\right\|^p_{L^p(\Omega)}.$$

The proof of (2.3 ii) is similar. Denoting the values $\widetilde{D}^0_{\widehat{o}}, \widetilde{D}_{\widehat{o}}$ of $\mathcal{D}^{h,0}[(\widetilde{v}_K)_K]$ and $\mathcal{D}^h[(\widetilde{v}_K)_K]$, respectively, on $\widehat{o}$, one can rewrite the sum in (2.3 ii) as

$$\begin{aligned}
\sum_{\widehat{o}\in\widehat{\mathcal{T}}^h} \Big\{ & S_{K,L}\Big((a_p(D_{\widehat{o}}) - a_p(\widetilde{D}_{\widehat{o}}))\cdot\nu_{K,L}\Big)\Big((D^0_{\widehat{o}} - \widetilde{D}^0_{\widehat{o}})\cdot\nu_{K,L}\Big)\\
&+ S_{L,M}\Big((a_p(D_{\widehat{o}}) - a_p(\widetilde{D}_{\widehat{o}}))\cdot\nu_{L,M}\Big)\Big((D^0_{\widehat{o}} - \widetilde{D}^0_{\widehat{o}})\cdot\nu_{L,M}\Big)\\
&+ S_{M,K}\Big((a_p(D_{\widehat{o}}) - a_p(\widetilde{D}_{\widehat{o}}))\cdot\nu_{M,K}\Big)\Big((D^0_{\widehat{o}} - \widetilde{D}^0_{\widehat{o}})\cdot\nu_{M,K}\Big)\Big\}\\
&= \frac{1}{2}\sum_{\widehat{o}\in\widehat{\mathcal{T}}^h} S_{\widehat{o}}(a_p(D_{\widehat{o}}) - a_p(\widetilde{D}_{\widehat{o}}))\cdot(D^0_{\widehat{o}} - \widetilde{D}^0_{\widehat{o}}).
\end{aligned}$$

Using (5.2) and denoting by H the Hessian matrix of the function $x \in \mathbb{R}^2 \mapsto \frac{1}{p}|x|^p$, we get

$$\begin{aligned}
&(a_p(D_{\widehat{o}}) - a_p(\widetilde{D}_{\widehat{o}}))\cdot(D^0_{\widehat{o}} - \widetilde{D}^0_{\widehat{o}})\\
&\qquad = (D_{\widehat{o}} - \widetilde{D}_{\widehat{o}})^t\left[\int_0^1 H(\widetilde{D}_{\widehat{o}} + \tau(D_{\widehat{o}} - \widetilde{D}_{\widehat{o}}))\,d\tau\,(I - R_{\widehat{o}})\right](D_{\widehat{o}} - \widetilde{D}_{\widehat{o}}).
\end{aligned}$$

For all $x \in \mathbb{R}^2$, $x \neq 0$, $H(x)$ is a symmetric matrix with positive eigenvalues λ_1, λ_2 such that $\lambda_1/\lambda_2 = p-1$. Thus the condition $\|R_{\widehat{o}}\| < \min\{p-1, 1/(p-1)\}$ ensures that, for all $\tau \in [0,1]$,

$$r^t\left[H(\widetilde{D}_{\widehat{o}} + \tau(D_{\widehat{o}} - \widetilde{D}_{\widehat{o}}))\,(I - R_{\widehat{o}})\right] r \geq a\, r^t\left[H(\widetilde{D}_{\widehat{o}} + \tau(D_{\widehat{o}} - \widetilde{D}_{\widehat{o}}))\right] r > 0$$

for all $r \in \mathbb{R}^2, r \neq 0$, with some constant $a > 0$. Now (2.3 ii) follows. ∎

5.2. On interpolated solutions. First note that it is sufficient to prove the interpolation property in $W^{1,p}_0(\Omega)$ if we require, in addition to the time-independent analogs of (2.8), (2.9) (referred to as (2.8′), (2.9′)), that

$$\|v^h\|_{W^{1,p}_0(\Omega)} \leq c \times \left\|\mathcal{D}^h_\perp[\overline{v}^h]\right\|_{L^p(\Omega)} \tag{2.10′}$$

with a constant c independent of h. We obtain the interpolation property in E with the function $I : C \mapsto c \times C$ by taking v^h constant on each $\imath^n$ and summing in $n \in \{1, \ldots, [T/k^h]+1\}$.

LEMMA 5.3. *Let $(\mathcal{T}^h)_h$ be a strongly proportional family of finite volume meshes of $\Omega \subset \mathbb{R}^d$. Then it has the interpolation property in $W_0^{1,p}(\Omega)$.*

In order to prove the lemma, we first show that the strong proportionality allows us to majorate the L^p norm of the translates of the discrete solutions $\overline{v}^h$ in Lemma 2.6(ii) by $\mathrm{const}\Delta\,(h+\Delta)^{p-1}$. Then we convolute $\overline{v}^h$ with the appropriate mollifier; finally, we restore the average over each mesh volume as in Lemma 5.4 below. The complete proof is given in [4].

Note that the interpolation property can fail on weakly proportional meshes, at least for $p > 2$.

Indeed, consider $\Omega = (0,1)^2$. For $s \geq 2$, let $\mathcal{T}^s$ be the finite volume mesh of Ω such that $\mathcal{T}^s_{int} = \{K^s, L^s\}$, where $K^s = \{(x,y) \in \Omega \,|\, x+y < 1/s\}$ with $x_{K^s} = (\frac{1}{4s}, \frac{1}{4s})$, and L^s is the interior of the complementary of K^s with $x_{L^s} = (\frac{1}{2}, \frac{1}{2})$. Take $\overline{v}^s$ such that $\overline{v}^s \equiv s^{1/p}$ on K^s and $\overline{v}^s \equiv 0$ on L^s. Then $\int_\Omega \left|\mathcal{D}^h_\perp[\overline{v}^s]\right|^p \leq \mathrm{const}$ uniformly in s. If there exist $v^s \in W_0^{1,p}(\Omega)$ interpolated solutions for $\overline{v}^s$, we have $\|v^s\|_{W_0^{1,p}(\Omega)} \leq \mathrm{const}$. Hence by the standard embedding theorem, v^s are uniformly bounded. This contradicts the fact that $\frac{1}{\mathfrak{m}(K^s)}\int_{K^s} v^s = s^{1/p} \to +\infty$ as $s \to +\infty$.

Nevertheless, we have the following result in the situation close to that of Proposition 5.2.

LEMMA 5.4. *Assume that $(\mathcal{T}^h)_h$ is a weakly proportional family of meshes of $\Omega \subset \mathbb{R}^2$ dual to a family of meshes $(\widehat{\mathcal{T}}^h)_h$ such that all $\widehat{o} \in \widehat{\mathcal{T}}^h$ are triangles with angles less than or equal to $\pi/2$. Then $(\mathcal{T}^h)_h$ has the interpolation property in $W_0^{1,p}(\Omega)$.*

Proof. Take discrete solutions $\overline{v}^h = \sum_K v_K \mathbb{1}_K$ on each of $\mathcal{T}^h$ such that, for all h $\left\|\mathcal{D}^h_\perp[\overline{v}^h]\right\|_{L^p(Q)} \leq C$. Denote by c the generic constant that depends only on p and ζ^*. Let $v^{h,0}$ be the continuous piecewise affine function on Ω that interpolates the values v_K, v_L, v_M at the points x_K, x_L, x_M over $\widehat{o}$ for all $\widehat{o} \in \widehat{\mathcal{T}}^h$ (we use the construction and notation of section 5.1). We have $v^{h,0} \in W_0^{1,p}(\Omega)$ and $Dv^{h,0} \equiv \mathcal{D}^{h,0}[\overline{v}^h]$ so that (5.3) yields $\|Dv^{h,0}\|_{L^p(\Omega)} \leq c \times C$. Note that for all $x \in K \in \mathcal{T}^h$,

$$|v^{h,0}(x) - \overline{v}^h(x)| = |v^{h,0}(x) - v^{h,0}(x_K)| \leq \mathfrak{d}(K)\,|Dv^{h,0}(x)|. \tag{5.4}$$

Hence $\|v^{h,0} - \overline{v}^h\|_{L^p(\Omega)} \leq h\,\|Dv^{h,0}\|_{L^p(\Omega)} \to 0$ as $h \to 0$. Now take a continuously differentiable function $\pi : \mathbb{R}^2 \to \mathbb{R}^+$ such that $\operatorname{supp}\pi = \{x \in \mathbb{R}^2 \mid |x| \leq 1\}$ and $\int_{\mathbb{R}^2} \pi = 1$. For all $K \in \mathcal{T}^h_{int}$ set $\varphi_K = \frac{\mathfrak{m}(K)}{(\zeta^*\mathfrak{d}(K))^2}\,\pi\left(\frac{x-x_K}{\zeta^*\mathfrak{d}(K)}\right)$ (for boundary volumes K of nonzero measure; i.e., if $x_K \in \partial\Omega$, an easy modification is needed in order to keep the trace on $\partial\Omega$ equal to zero). Set

$$v^h = v^{h,0} + \sum_K \alpha_K \varphi_K, \qquad \text{with} \quad \alpha_K = v_K - \frac{1}{\mathfrak{m}(K)}\int_K v^{h,0}.$$

Since $\operatorname{supp}\varphi_K \subset K$, by the choice of α_K, the family $(v^h)_h$ verifies (2.9′). Moreover, since $\frac{\mathfrak{m}(K)}{\mathfrak{d}(K)^2} \leq c$ for all $K \in \mathcal{T}^h$, for all h, by the Hölder inequality we get

$$\begin{aligned}
\int_\Omega |v^h - v^{h,0}|^p &= \sum_K |\alpha_K|^p \int_K |\varphi_K|^p \\
&\leq\ c \sum_K \frac{1}{\mathfrak{m}(K)^p}\left|\int_K \overline{v}^h - \int_K v^{h,0}\right|^p \mathfrak{m}(K)\Big(\frac{\mathfrak{m}(K)}{\mathfrak{d}(K)^2}\Big)^p \\
&\leq\ c \sum_K \frac{1}{\mathfrak{m}(K)^p}\mathfrak{m}(K)^{p'/p}\int_K |\overline{v}^h - v^{h,0}|^p \mathfrak{m}(K) = c\int_\Omega |\overline{v}^h - v^{h,0}|^p \to 0
\end{aligned}$$

as $h \to 0$. Thus $(v^h)_h$ satisfies (2.8′). In the same manner, using (5.4) we have

$$\begin{aligned}\int_\Omega |Dv^h - Dv^{h,0}|^p &= \sum_K |\alpha_K|^p \int_K |D\varphi_K|^p \\ &\le \sum_K c \sum_K \frac{1}{\mathfrak{m}(K)^p} \mathfrak{m}(K)^{p'/p} \int_K |\overline{v}^h - v^{h,0}|^p \times \mathfrak{m}(K) \Big(\frac{\mathfrak{m}(K)}{\mathfrak{d}(K)^3}\Big)^p \\ &\le c \sum_K \frac{1}{\mathfrak{d}(K)^p} \int_K |\overline{v}^h - v^{h,0}|^p \le c \sum_K \frac{1}{\mathfrak{d}(K)^p}\, \mathfrak{d}(K)^p \int_K |Dv^{h,0}|^p = c \int_\Omega |Dv^{h,0}|^p.\end{aligned}$$

Hence $\|v^h\|_{W_0^{1,p}(\Omega)} \le c \|v^{h,0}\|_{W_0^{1,p}(\Omega)} \le c \times C$, so $(v^h)_h$ satisfies (2.10′). Thus $(v^h)_h$ can be chosen as interpolated solutions for $(\overline{v}^h)_h$. ∎

REFERENCES

[1] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[2] B. Andreianov, *Quelques problèmes de la théorie de systèmes paraboliques dégénérés non-linéaires et de lois de conservation*, Ph.D. thesis, Laboratoire de Mathématiques, Université de Franche-Comté, Besançon, France, 2000.

[3] B. Andreianov, M. Gutnic, and P. Wittbold, *L'approche "continue" pour une méthode de volumes finis*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 477–482.

[4] B. Andreianov, M. Gutnic, and P. Wittbold, *Convergence of Finite Volume Approximations for a Nonlinear Elliptic-Parabolic Problem: A "Continuous" Approach*, prepublication 03036, Institut de Recherche Mathématique Avancée, Université Louis Pasteur, Strasbourg, France, 2003; also available online from http://www-irma.u-strasb.fr/irma/publications/2003/03036.ps.gz.

[5] J. W. Barrett and W. B. Liu, *Finite element approximation of the parabolic p-Laplacian*, SIAM J. Numer. Anal., 31 (1994), pp. 413–428.

[6] J. Bear, *Dynamics of Fluids in Porous Media*, Elsevier, New York, 1972.

[7] Ph. Bénilan and P. Wittbold, *On mild and weak solutions of elliptic-parabolic problems*, Adv. Differential Equations, 1 (1996), pp. 1053–1073.

[8] F. Bouhsiss, *Etude d'un problème parabolique par les semi-groupes non linéaires*, in Analyse Non Linéaire, Publ. Math. UFR Sci. Tech. Besançon 15, Univ. Franche-Comté, Besançon, France, 1995/1997, pp. 133–141.

[9] H. Brézis and W. A. Strauss, *Semi-linear second-order elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 565–590.

[10] J. Carrillo and P. Wittbold, *Uniqueness of renormalized solutions of degenerate elliptic-parabolic problems*, J. Differential Equations, 156 (1999), pp. 93–121.

[11] S.-S. Chow, *Finite element error estimates for nonlinear elliptic equations of monotone type*, Numer. Math., 54 (1989), pp. 373–393.

[12] J. I. Diaz and F. de Thélin, *On a nonlinear parabolic problem arising in some models related to turbulent flows*, SIAM J. Math. Anal., 25 (1994), pp. 1085–1111.

[13] Yu. V. Egorov and V. A. Kondratiev, *On Spectral Theory of Elliptic Operators*, Oper. Theory Adv. Appl. 89, Birkhäuser, Basel, 1996.

[14] R. Eymard, T. Gallouet, M. Gutnic, R. Herbin, and D. Hilhorst, *Approximation by the finite volume method of an elliptic-parabolic equation arising in environmental studies*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1505–1528.

[15] R. Eymard, T. Gallouët, and R. Herbin, *Finite volumes methods*, in Handbook of Numerical Analysis, P. G. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 715–1022.

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

[17] R. Eymard, M. Gutnic, and D. Hilhorst, *The finite volume method for an elliptic-parabolic equation*, Acta Math. Univ. Comenian. (N.S.), 67 (1998), pp. 181–195.

[18] R. Eymard, M. Gutnic, and D. Hilhorst, *The finite volume method for Richards equation*, Comput. Geosci., 3 (1999), pp. 259–294 (2000).

[19] R. Glowinski and A. Marrocco, *Sur l'approximation par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*, RAIRO Anal. Numér., 9 (1975), pp. 41–76.

[20] N. Ju, *Numerical analysis of the parabolic p-Laplacian: Approximation of trajectories*, SIAM J. Numer. Anal., 37 (2000), pp. 1861–1884.
[21] J. Kačur, *On a solution of degenerate elliptic-parabolic problems in Orlicz-Sobolev spaces.* I, Math. Z., 203 (1990), pp. 153–171.
[22] J. Kačur, *On a solution of degenerate elliptic-parabolic problems in Orlicz-Sobolev spaces.* II, Math. Z., 203 (1990), pp. 569–579.
[23] S. N. Kruzhkov, *Results on the nature of the continuity of solutions of parabolic equations and some of their applications*, Mat. Zametki, 6 (1969), pp. 97–108 (in Russian); English translation in Math. Notes, 6 (1969), pp. 517–523.
[24] J. Leray and J.-L. Lions, *Quelques résultats de Višik sur les problèmes elliptiques non linéaires par les méthodes de Minty-Browder*, Bull. Soc. Math. France, 93 (1965), pp. 97–107.
[25] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.
[26] F. Otto, L^1*-Contraction and uniqueness for quasilinear elliptic-parabolic problems*, J. Differential Equations, 131 (1996), pp. 20–38.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 252–264

ANALYSIS OF A NONSYMMETRIC DISCONTINUOUS GALERKIN METHOD FOR ELLIPTIC PROBLEMS: STABILITY AND ENERGY ERROR ESTIMATES*

MATS G. LARSON† AND A. JONAS NIKLASSON‡

Abstract. In this paper we analyze a nonsymmetric discontinuous Galerkin method for elliptic problems proposed by Oden, Babuška, and Baumann. Our main results are a complete inf-sup stability analysis and, as a consequence, error estimates in a mesh dependent energy norm allowing variable meshsize and order of polynomials. The analysis is carried out in two spatial dimensions on an unstructured triangulation.

Key words. discontinuous Galerkin method, error estimates, stability, energy norm, elliptic problem, nonsymmetric discontinuous Galerkin method

AMS subject classifications. 65N30, 65N15, 65N12

DOI. 10.1137/S0036142902413160

1. Introduction.

The discontinuous Galerkin (dG) method is a classical technique for numerical approximation of partial differential equations which have recently received new interest, motivated by some attractive features including a flexible discretization allowing easy implementation of *h-p* adaptivity, nonmatching grids, and a local conservation property. Of course there are disadvantages too; the number of degrees of freedom is larger (see [8]), and efficient iterative solvers are not yet developed.

In this paper we are concerned with the analytical and numerical study of the recent nonsymmetric dG method for elliptic problems proposed by Oden, Babuška, and Baumann in [12]. This method does not contain the stabilizing (penalty) term as the classical symmetric Nitsche method [11]. Plenty of numerical results were presented in [12], showing that a remarkable stability is hidden in the nonsymmetric form for polynomials of order higher than or equal to two in one and two spatial dimensions. The desire to analytically understand the stability properties of the nonsymmetric dG method is the motivation for the present paper. In an earlier paper [9] Larson and Niklasson showed complete stability estimates for a family of dG methods, including both the nonsymmetric method and the symmetric Nitsche method, in one spatial dimension. These results extended the analytical stability estimates presented by Babuška, Baumann, and Oden in [3] for polynomials of order three or higher in one spatial dimension. The analysis presented in this paper builds on the ideas in [9].

Our main result in this work is a complete discrete stability analysis, where we prove that the method is inf-sup stable with respect to a mesh dependent energy norm for quadratic and higher order polynomials on a general unstructured triangulation

*Received by the editors August 15, 2002; accepted for publication (in revised form) June 10, 2003; published electronically January 28, 2004. This research was supported by The Swedish Foundation for International Cooperation in Research and Higher Education.
http://www.siam.org/journals/sinum/42-1/41316.html

†Department of Mathematics, Chalmers University of Technology, Göteborg, SE-412 96, Sweden (mgl@math.chalmers.se). The research of this author was supported by the Swedish Council for Engineering Sciences.
‡Department of Applied Mechanics, Chalmers University of Technology, Göteborg, SE-412 96, Sweden (jonas.niklasson@me.chalmers.se).

in two spatial dimensions. We present numerical calculations of the inf-sup constant confirming our analytical estimates. Our analytical and numerical results confirm the numerical observations reported in [12]. The case of linear ploynomials is also investigated, and we show that the inf-sup constant either is zero or depends on the meshsize (depending on boundary conditions) if the mesh is of checkerboard type.

From the study of the discrete stability properties we immediately obtain optimal order a priori error estimates in the energy norm, in terms of local meshsize and local degree of polynomials. We present numerical results illustrating our error estimates. In two recent papers, Riviére, Wheeler, and Girault [13], [14], prove an a priori error estimate of the L^2 norm of the gradient of the error for the nonsymmetric dG method by relating it to a method where the discontinuities on each edge have average zero. However, no stability estimate for the nonsymmetric dG method is presented. We also mention the comprehensive overview and analysis of a large class of dG methods by Arnold, Brezzi, Cockburn, and Marini [2].

Key to our analysis is a splitting of the space of all discontinuous piecewise polynomials into a sum of a space of functions with constrained discontinuities, representing continuous scales, and a space of discontinuous functions with small spatial mean value. This splitting, properly constructed, leads to a triangular system which can be analyzed.

The remainder of this paper is organized as follows: in section 2 we introduce the nonsymmetric dG method and the necessary notation; in section 3 we present the splitting of the discontinuous piecewise polynomial space and the two-scale formulation of the dG method; and finally, in section 4 we show the stability estimate and the error estimate in the energy norm.

2. The model problem and the dG method.

2.1. A model problem. Let Ω be a polygonal domain in $\mathbf{R}^2$ with boundary Γ divided into two disjoint parts $\Gamma = \Gamma_N \cup \Gamma_D$. We consider the following linear elliptic model problem: find $u : \Omega \to \mathbf{R}$ such that

$$\begin{aligned} -\nabla \cdot \sigma(u) &= f \quad && \text{in } \Omega, \\ u &= g_D && \text{on } \Gamma_D, \\ \sigma_n(u) &= g_N && \text{on } \Gamma_N. \end{aligned} \tag{2.1}$$

Here the flux $\sigma(u)$ is defined by

$$\sigma(u) = A\nabla u, \tag{2.2}$$

with A a constant (or piecewise constant) symmetric positive definite matrix, and $\sigma_n(u)$ denotes the normal flux

$$\sigma_n(u) = n \cdot A\nabla u, \tag{2.3}$$

where n is the exterior unit normal of Γ. It is well known that there is a unique solution in $H^1(\Omega)$ for $f \in H^{-1}(\Omega)$, $g_D \in H^{1/2}(\Gamma_D)$, and $g_N \in H^{-1/2}(\Gamma_N)$ to (2.1) (see [6]), where $H^s(\omega)$ denote the standard Sobolev spaces on the set ω.

2.2. Discrete spaces. We let $\mathcal{K}$ be a triangulation of Ω into affine triangles K satisfying the minimal angle condition, implying that the trace inequality (2.17) and inverse inequality (2.19) below hold. We denote the set of all edges E by $\mathcal{E}$ and divide $\mathcal{E}$ into three disjoint sets

$$\mathcal{E} = \mathcal{E}_I \cup \mathcal{E}_D \cup \mathcal{E}_N, \tag{2.4}$$

where $\mathcal{E}_I$ is the set of all edges in the interior of Ω, $\mathcal{E}_D$ the edges on the Dirichlet part of the boundary Γ_D, and $\mathcal{E}_N$ the edges on the Neumann part Γ_N. We let $h : \Omega \to \mathbf{R}$ denote the mesh function such that $h|_K = h_K = \operatorname{diam}(K)$ and $h|_E = h_E = \operatorname{diam}(E)$, i.e., the length of the edge E. We let

$$\mathcal{V} = \bigoplus_{K \in \mathcal{K}} \mathcal{P}_p(K), \tag{2.5}$$

where $\mathcal{P}_p(K)$ is the space of all polynomials of degree less than or equal to p defined on K. The degree of polynomials, as well as the meshsize, may vary from element to element so that $p|_K = p_K$, and thus we allow h-p adaptivity.

2.3. The nonsymmetric dG method. In [12] Oden, Babuška, and Baumann proposed the following nonsymmetric dG method: find $u_h \in \mathcal{V}$ such that

$$a(u_h, v) = l(v) \quad \text{for all } v \in \mathcal{V}. \tag{2.6}$$

Here $a(\cdot, \cdot)$ is a bilinear form defined by

$$a(v, w) = a_{\mathcal{K}}(v, w) - a_{\mathcal{E}}(v, w) + a_{\mathcal{E}}(w, v), \tag{2.7}$$

where

$$a_{\mathcal{K}}(v, w) = \sum_{K \in \mathcal{K}} (\sigma(v), \nabla w)_K, \tag{2.8}$$

$$a_{\mathcal{E}}(v, w) = \sum_{E \in \mathcal{E}_I \cup \mathcal{E}_D} (\langle \sigma_n(v) \rangle, [w])_E, \tag{2.9}$$

and $l(\cdot)$ is a linear functional defined by

$$l(v) = (f, v) + \sum_{E \in \mathcal{E}_N} (g_N, v)_E + \sum_{E \in \mathcal{E}_D} (g_D, \langle \sigma_n(v) \rangle)_E. \tag{2.10}$$

We employed the notation

$$\langle v \rangle = \begin{cases} (v^+ + v^-)/2, & E \in \mathcal{E}_I, \\ v^+, & E \in \mathcal{E}_D, \end{cases} \tag{2.11}$$

for the average and

$$[v] = \begin{cases} v^+ - v^-, & E \in \mathcal{E}_I, \\ v^+, & E \in \mathcal{E}_D, \end{cases} \tag{2.12}$$

for the jump at an edge E, where $u^{\pm}(x) = \lim_{t \to 0, t > 0} u(x \mp tn)$, $x \in E$, and n is the exterior unit normal to E for $E \in \mathcal{E}_D \cup \mathcal{E}_N$ and a fixed, but arbitrary, unit normal to E for $E \in \mathcal{E}_I$; see Figure 2.1.

LEMMA 2.1. *If $f \in L^2$, $g_D \in H^{1/2}(\Gamma_D)$, and $g_N \in L^2(\Gamma_N)$, then the linear functional $l(\cdot)$ is bounded on $\mathcal{V}$ and the exact solution u of* (2.1) *satisfies*

$$a(u, v) = l(v) \tag{2.13}$$

for all $v \in \mathcal{V}$.

Proof. The first statement is obvious. For the second we note that the normal trace $\sigma_n(u)$ of $\sigma(u)$ is well defined in $L^2(E)$ on all edges $E \in \mathcal{E}$ since the stability estimate $\|\sigma(u)\| + \|\nabla \cdot \sigma(u)\| \le c(\|f\| + \|g_D\|_{1/2,\Gamma_D} + \|g_N\|_{\Gamma_N})$ holds. ∎

Here and below we let $\|v\|_{s,\omega}$ and $|v|_{s,\omega}$ denote the standard Sobolev norms and seminorms, respectively, for $v \in H^s(\omega)$ on the set $\omega \subset \Omega$. For brevity we write $\|v\|_s = \|v\|_{s,\Omega}$, $\|v\|_\omega = \|v\|_{0,\omega}$, and $\|v\| = \|v\|_{0,\Omega}$ for the L^2 norm.

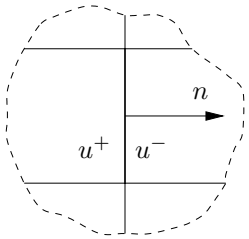


FIG. 2.1. *The plus and minus sides of an edge.*

2.4. The energy norm and some useful inequalities. We equip $\mathcal{V}$ with the mesh dependent energy norm

$$|||v|||^2 = |||v|||_{\mathcal{K}}^2 + \|\langle\sigma_n(v)\rangle\|_{\mathcal{E}}^2 + \|h^{-1}[v]\|_{\mathcal{E}}^2, \tag{2.14}$$

where

$$|||v|||_{\mathcal{K}}^2 = \sum_{K\in\mathcal{K}} (A\nabla v, \nabla v)_K, \tag{2.15}$$

$$\|w\|_{\mathcal{E}}^2 = \sum_{E\in\mathcal{E}_I\cup\mathcal{E}_D} \|h^{1/2}w\|_E^2. \tag{2.16}$$

Next we recall some useful standard inequalities which we will need in our developments. First we have the trace inequality

$$\|v\|_{\partial K}^2 \le c\|v\|_K\left(h_K^{-1}\|v\|_K + \|v\|_{1,K}\right) \quad \text{for } v\in H^1(K), \tag{2.17}$$

where c is a constant independent of h. This inequality follows by mapping to the unit size reference element $\tilde{K}$, employing the trace inequality

$$\|v\|_{\partial\tilde{K}}^2 \le c\|v\|_{\tilde{K}}\|v\|_{1,\tilde{K}} \quad \text{for } v\in H^1(\tilde{K}) \tag{2.18}$$

(see Brenner and Scott [5]) and finally transforming back to K. Furthermore, the following inverse estimate will be useful:

$$\|\langle\sigma_n(v)\rangle\|_{\mathcal{E}} \le C|||v|||_{\mathcal{K}} \quad \text{for } v\in\mathcal{V}, \tag{2.19}$$

with constant C dependent on the degree of polynomials p but not on the meshsize h. This estimate can be shown by scaling; see Thomée [15] for details.

3. A two-scale formulation of the dG method.

3.1. A splitting of $\mathcal{V}$.

THEOREM 3.1. *For $p\ge 2$ there is a decomposition of $\mathcal{V}$ into a direct sum*

$$\mathcal{V} = \mathcal{V}_c + \mathcal{V}_d, \tag{3.1}$$

where

$$\mathcal{V}_d = \{v\in\mathcal{V} : a_{\mathcal{K}}(w,v) - a_{\mathcal{E}}(w,v) = 0 \text{ for all } w\in\mathcal{V}\}, \tag{3.2}$$

$$\mathcal{V}_c = \{v\in\mathcal{V} : a_{\mathcal{E}}(w,v) = 0 \text{ for all } w\in\mathcal{V}_d\}, \tag{3.3}$$

with bilinear forms defined in (2.8) *and* (2.9). *Furthermore, for* $p \geq 2$ *the following norm equivalence holds:*

$$c_1 |||v|||^2 \leq |||v_c|||_{\mathcal{K}}^2 + |||v_d|||_{\mathcal{K}}^2 \leq c_2 |||v|||^2, \tag{3.4}$$

with constants c_1 *and* c_2 *independent of* h *but dependent on* p.

For the proof of Theorem 3.1 we need the following two lemmas.

LEMMA 3.1. *For each edge* $E \in \mathcal{E}_I \cup \mathcal{E}_D$ *there is a function* $\varphi_E \in \mathcal{V}_d$ *such that*

$$[\varphi_E] = 1 \quad \text{on } E, \tag{3.5}$$

$$\int_{E'} [\varphi_E] v = 0 \quad \text{for all } v \in \mathcal{P}_{p-1}(E') \text{ and } E' \in \mathcal{E} \setminus E, \tag{3.6}$$

where $\mathcal{P}_{p-1}(E')$ *denotes the space of polynomials of order* $p-1$ *defined on* E'.

Proof. We consider the case $E \in \mathcal{E}_I$. The case $E \in \mathcal{E}_D$ is similar, and it is also easy to see that the proof does not work out for $E \in \mathcal{E}_N$. We construct φ_E elementwise. Let $K^+, K^- \in \mathcal{K}$ be the triangles which share an interior edge E. Let z denote the coordinate orthogonal to E with positive direction into K^+, let $H^\pm$ be the height of $K^\pm$, and let

$$\varphi_E|_{K^\pm} = \begin{cases} -L_p(2(z/H^\pm) \mp 1)/2, & \text{odd } p, \\ \pm L_p(2(z/H^\pm) \mp 1)/2, & \text{even } p, \end{cases} \tag{3.7}$$

where L_p denotes the Legendre polynomial (see [1]) of order p defined on $[-1, 1]$. We begin by verifying that $\varphi_E \in \mathcal{V}_d$. Note that the condition

$$a_{\mathcal{K}}(w, v) - a_{\mathcal{E}}(w, v) = 0 \tag{3.8}$$

for all $w \in \mathcal{V}$ is equivalent to

$$-(\nabla \cdot \sigma(w), v)_K = (\sigma_n(w), \langle v \rangle)_{\partial K} \tag{3.9}$$

for all $w \in \mathcal{V}_K$ and $K \in \mathcal{K}$. Note that, from the fact that the Legendre polynomial L_p is orthogonal to all polynomials of order $p-1$, it follows that φ_E satisfies

$$-(\nabla \cdot \sigma(w), \varphi_E)_K = 0, \tag{3.10}$$

$$(\sigma_n(w), \langle \varphi_E \rangle)_{\partial K} = 0, \tag{3.11}$$

where in the last equality we also used that $\langle \varphi_E \rangle = 0$ on E. Thus φ_E is in $\mathcal{V}_d$. The properties (3.5) and (3.6) of φ_E are direct consequences of the construction. ∎

LEMMA 3.2. *For* $p \geq 2$ *there is a* $w \in \mathcal{V}_d$ *for each* $v \in \mathcal{V}$ *such that*

$$\|h^{-1} P_0[v]\|_{\mathcal{E}}^2 = \sum_{E \in \mathcal{E}_I \cup \mathcal{E}_D} (\langle \sigma_n(w) \rangle, P_0[v])_E, \tag{3.12}$$

$$|||w|||_{\mathcal{K}} \leq c \|h^{-1} P_0[v]\|_{\mathcal{E}}, \tag{3.13}$$

with constant c *independent of* h *and* p, *and* P_0 *the edgewise* L^2*-projection on constant functions.*

Proof. Let K be a triangle, E one of the edges of K, H the height of K orthogonal to E, and $z \in [0, H]$ the coordinate orthogonal to E. Then the normal derivative of the function $z(z/H - 1)$ is one on E and has average zero on the two other edges.

Based on this observation and the fact that A is positive definite, we conclude that for $p \geq 2$ we can construct a $w' \in \mathcal{V}_d$ for each $v \in \mathcal{V}$ such that

$$\sum_{E \in \mathcal{E}_I \cup \mathcal{E}_D} (\langle \sigma_n(w') \rangle, P_0[v])_E = \|h^{-1} P_0[v]\|_{\mathcal{E}}^2, \tag{3.14}$$

$$|||w'|||_{\mathcal{K}} \leq c \|h^{-1} P_0[v]\|_{\mathcal{E}}. \tag{3.15}$$

Next, for $p \geq 2$, we define $w \in \mathcal{V}_d$ by

$$a_{\mathcal{K}}(w, v) = a_{\mathcal{K}}(w', v) \quad \text{for all } v \in \mathcal{V}_d. \tag{3.16}$$

We note that setting $v = w$ and using the Cauchy–Schwarz inequality give

$$|||w|||_{\mathcal{K}} \leq |||w'|||_{\mathcal{K}}, \tag{3.17}$$

and thus it follows that

$$|||w|||_{\mathcal{K}} \leq c \|h^{-1} P_0[v]\|_{\mathcal{E}}. \tag{3.18}$$

Using the definition of $\mathcal{V}_d$, we get

$$a_{\mathcal{E}}(w, v) = a_{\mathcal{E}}(w', v) \quad \text{for all } v \in \mathcal{V}_d, \tag{3.19}$$

and choosing $v = \varphi_E$ (see Lemma 3.1), we find that

$$P_0 \langle \sigma_n(w) \rangle = P_0 \langle \sigma_n(w') \rangle \quad \text{on } E \tag{3.20}$$

for each edge $E \in \mathcal{E}_I \cup \mathcal{E}_D$. ∎

Remark 3.1. The construction of w' is a consequence of the classical nonconforming quadratic Morley element [10]. The degrees of freedom of the Morley element are the nodal values and the values of the normal derivative at the midpoints of the edges.

LEMMA 3.3. *It holds that*

$$\|h^{-1}(I - P_0)[v]\|_{\mathcal{E}} \leq c |||v|||_{\mathcal{K}} \quad \textit{for all } v \in \mathcal{V}, \tag{3.21}$$

with constant c independent of h and p, and with P_0 the edgewise L^2-projection on constant functions.

Proof. Note that we may subtract the projection of v onto piecewise constants $\pi_0 v$ as follows:

$$\|h^{-1}(I - P_0)[v]\|_{\mathcal{E}}^2 = \|h^{-1}(I - P_0)[v - \pi_0 v]\|_{\mathcal{E}}^2 \tag{3.22}$$

$$\leq c \sum_{K \in \mathcal{K}} h^{-1} \|v - \pi_0 v\|_K \Big(h^{-1} \|v - \pi_0 v\|_K + \|v - \pi_0 v\|_{1,K} \Big) \tag{3.23}$$

$$\leq c |||v|||^2, \tag{3.24}$$

where we finally used the interpolation estimate (4.12) below together with the fact that the H^1 seminorm can be estimated by the energy norm. ∎

Proof of Theorem 3.1. Clearly $\mathcal{V} = \mathcal{V}_c + \mathcal{V}_d$ by the definition. Assume that $v \in \mathcal{V}_c \cap \mathcal{V}_d$. Then we conclude that $a_{\mathcal{K}}(v, v) = 0$, and thus v is a piecewise constant function. It follows that $a_{\mathcal{E}}(w, v) = 0$ for all $w \in \mathcal{V}_d$. Invoking Lemma 3.2, we find that $v = 0$. Therefore, the sum is direct for $p \geq 2$.

Starting with the left inequality in (3.4), we first observe that, using the inverse inequality (2.19) and the triangle inequality, we have

$$\begin{aligned} |||v|||^2 &\le c|||v|||_{\mathcal{K}}^2 + \|h^{-1}[v]\|_{\mathcal{E}}^2 \\ &\le c\Big(|||v_c|||_{\mathcal{K}}^2 + |||v_d|||_{\mathcal{K}}^2\Big) + \|h^{-1}[v]\|_{\mathcal{E}}^2, \end{aligned} \tag{3.25}$$

and thus we need to estimate $\|h^{-1}[v]\|_{\mathcal{E}}^2$. Using the triangle inequality, we have

$$\|h^{-1}[v]\|_{\mathcal{E}} \le \|h^{-1}(I-P_0)[v]\|_{\mathcal{E}} + \|h^{-1}P_0[v]\|_{\mathcal{E}}. \tag{3.26}$$

For the first term on the right-hand side in (3.26) we have, using Lemma 3.3,

$$\|h^{-1}(I-P_0)[v]\|_{\mathcal{E}} \le c|||v|||_{\mathcal{K}} \le c\Big(|||v_d|||_{\mathcal{K}} + |||v_c|||_{\mathcal{K}}\Big). \tag{3.27}$$

Next, for the second, invoking Lemma 3.2 gives

$$\|h^{-1}P_0[v]\|_{\mathcal{E}}^2 = \sum_{E\in\mathcal{E}_I\cup\mathcal{E}_D} (\langle\sigma_n(w)\rangle, P_0[v])_E \tag{3.28}$$

$$= \sum_{E\in\mathcal{E}_I\cup\mathcal{E}_D} (\langle\sigma_n(w)\rangle, [v])_E - \sum_{E\in\mathcal{E}_I\cup\mathcal{E}_D} (\langle\sigma_n(w)\rangle, (I-P_0)[v])_E. \tag{3.29}$$

For the first term on the right-hand side in (3.29) we have the estimate

$$\sum_{E\in\mathcal{E}_I\cup\mathcal{E}_D} (\langle\sigma_n(w)\rangle, [v])_E = \sum_{E\in\mathcal{E}_I\cup\mathcal{E}_D} (\langle\sigma_n(w)\rangle, [v_d])_E \tag{3.30}$$

$$= \sum_{K\in\mathcal{K}} (\sigma(w), \nabla v_d)_K \tag{3.31}$$

$$\le |||w|||_{\mathcal{K}}|||v_d|||_{\mathcal{K}} \tag{3.32}$$

$$\le c\|h^{-1}P_0[v]\|_{\mathcal{E}}|||v_d|||_{\mathcal{K}}, \tag{3.33}$$

where we used the fact that $w\in\mathcal{V}_d$ in (3.30), the definition of $\mathcal{V}_d$ in (3.31), the Cauchy–Schwarz inequality in (3.32), and finally the stability estimate (3.13) in (3.33). For the second term,

$$\sum_{E\in\mathcal{E}_I\cup\mathcal{E}_D} (\langle\sigma_n(w)\rangle, (I-P_0)[v])_E \le \|\langle\sigma_n(w)\rangle\|_{\mathcal{E}}\|(I-P_0)h^{-1}[v]\|_{\mathcal{E}} \tag{3.34}$$

$$\le c|||w|||_{\mathcal{K}}|||v|||_{\mathcal{K}} \tag{3.35}$$

$$\le c\|h^{-1}P_0[v]\|_{\mathcal{E}}|||v|||_{\mathcal{K}}, \tag{3.36}$$

where we used the Cauchy–Schwarz inequality in (3.34), the inverse inequality (2.19) and Lemma 3.3 in (3.35), and finally the stability estimate (3.13) in (3.36).

Starting from (3.29) and using the triangle inequality together with estimates (3.33) and (3.36) and finally dividing with $\|h^{-1}P_0[v]\|_{\mathcal{E}}$ give

$$\|h^{-1}P_0[v]\|_{\mathcal{E}} \le c\Big(|||v_d|||_{\mathcal{K}} + |||v_c|||_{\mathcal{K}}\Big), \tag{3.37}$$

which together with (3.25), (3.26), and (3.27) proves the left inequality in (3.4).

We now turn to the proof of the right inequality in (3.4). Starting from the definition (3.2) of $\mathcal{V}_d$ in Theorem 3.1 and setting $w = v_d$, we get

$$|||v_d|||_{\mathcal{K}}^2 = a_{\mathcal{K}}(v_d, v_d) \tag{3.38}$$

$$= a_{\mathcal{E}}(v_d, v) \tag{3.39}$$

$$\leq \|\langle\sigma_n(v_d)\rangle\|_{\mathcal{E}} \|h^{-1}[v]\|_{\mathcal{E}} \tag{3.40}$$

$$\leq c|||v_d|||_{\mathcal{K}}|||v|||, \tag{3.41}$$

where we used the Cauchy–Schwarz inequality, and, at last, the inverse inequality (2.19) and the obvious fact that $\|h^{-1}[v]\|_{\mathcal{E}} \leq |||v|||$. Finally, dividing by $|||v_d|||_{\mathcal{K}}$ and squaring both sides give

$$|||v_d|||_{\mathcal{K}}^2 \leq c|||v|||^2. \tag{3.42}$$

Next, for v_c we simply have

$$|||v_c|||_{\mathcal{K}}^2 = |||v - v_d|||_{\mathcal{K}}^2 \tag{3.43}$$

$$\leq c\Big(|||v|||_{\mathcal{K}}^2 + |||v_d|||_{\mathcal{K}}^2\Big) \tag{3.44}$$

$$\leq c|||v|||^2, \tag{3.45}$$

which together with (3.42) prove the right inequality in (3.4). At last, tracing constants, we find that both c_1^{-1} and c_2 are of the form $cC^2 + c$, where c denotes constants independent of both h and p, and C is the constant in the inverse inequality (2.19), which depends on p. ∎

3.2. A two-scale formulation of the dG method. Here we shall derive a system of equations corresponding to (2.6) using the splitting given in Theorem 3.1. Writing $u = u_c + u_d$ and $v = v_c + v_d$ and using the identities

$$a(u_c, v_d) = 0, \tag{3.46}$$

$$a(u_d, v_c) = 2a_{\mathcal{K}}(u_d, v_c), \tag{3.47}$$

$$a(u_d, v_d) = a_{\mathcal{K}}(u_d, v_d), \tag{3.48}$$

which are direct consequences of Theorem 3.1, we obtain a triangular system of the following form: find $u = u_c + u_d \in \mathcal{V}_c + \mathcal{V}_d$ such that

$$\begin{aligned} a(u_c, v_c) + 2a_{\mathcal{K}}(u_d, v_c) &= l(v_c), \\ a_{\mathcal{K}}(u_d, v_d) &= l(v_d). \end{aligned} \tag{3.49}$$

We note that, with this particular splitting of $\mathcal{V}$, the discontinuous scales $\mathcal{V}_d$ are in fact not coupled to the continuous scales $\mathcal{V}_c$.

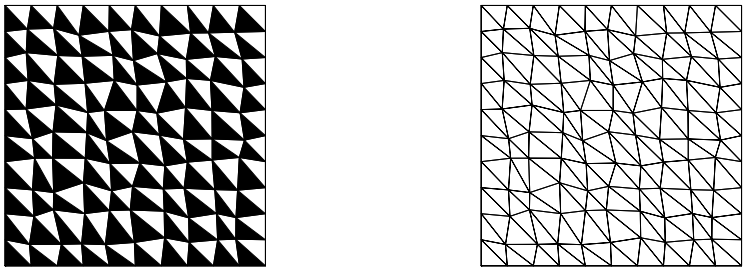
3.3. Checkerboard solutions for $p = 1$. For $p = 1$ the splitting (3.1) in Theorem 3.1 is not direct and the norm equivalence (3.4) does not hold in general. This fact can be seen as follows. Using Green's formula, we have

$$\sum_{K\in\mathcal{K}} (\nabla w, A\nabla v)_K = \sum_{K\in\mathcal{K}} (-\nabla\cdot\nabla w, v)_K + \sum_{E\in\mathcal{E}_I} ([\sigma_n(w)], \langle v\rangle)_E + (\langle\sigma_n(w)\rangle, [v])_E + \sum_{E\in\mathcal{E}_D\cup\mathcal{E}_N} (\sigma_n(w), v)_E.$$

Now if v is a piecewise constant function, then $\nabla v = 0$, and if w is a piecewise linear function, then $-\nabla \cdot A\nabla w = 0$ (recall that A is piecewise constant). Using these facts, we get

$$a_{\mathcal{E}}(w,v) = -\sum_{E\in\mathcal{E}_I} ([\sigma_n(w)], \langle v\rangle)_E - \sum_{E\in\mathcal{E}_N} (\sigma_n(w), v)_E,$$

and thus if $\mathcal{E}_N$ is empty and $\langle v\rangle = 0$ on each edge, then $a_{\mathcal{E}}(w,v) = 0$ for all $w \in \mathcal{V}$. Going back to the splitting $\mathcal{V} = \mathcal{V}_c + \mathcal{V}_d$, in Theorem 3.1 we find that $v \in \mathcal{V}_c \cap \mathcal{V}_d$ and thus the splitting is not direct. Further it is easy to see that $|||v|||_{\mathcal{K}} = 0$, while $|||v|||^2 \neq 0$ and thus c_1 must be zero; i.e., (3.4) does not hold. However, a piecewise constant function v, with $\langle v\rangle = 0$ on each $E \in \mathcal{E}_I$, does exist only on a checkerboard mesh, i.e., a mesh which could be colored as a checkerboard with two colors. In Figure 3.1 we give an example of such a function v on an unstructured checkerboard triangulation of the unit square. In the case when $\mathcal{E}_N$ is not empty but the mesh is a checkerboard mesh, we instead get that $c_1 \to 0$ as $h \to 0$. However, a general unstructured triangulation is usually quite far from being a checkerboard mesh, and in such a situation the norm equivalence will in general hold even for $p = 1$. See the computations of the inf-sup constant presented below.



(a) (b)

FIG. 3.1. (a) *Checkerboard solution with black* $= -1$ *and white* $= 1$ *and* (b) *the corresponding triangulation of the unit square.*

4. Stability analysis and error estimates in the energy norm.

4.1. Stability analysis. Our main result in this section is a proof that the inf-sup constant (see, for instance, [5]) is positive and independent of the meshsize. This stability result is, as is well known, key for proving existence and uniqueness of the discrete solution as well as error estimates in the energy norm.

THEOREM 4.1. *If* $p \geq 2$*, then there is a constant* $m > 0$ *such that*

$$\inf_{u\in\mathcal{V}} \sup_{v\in\mathcal{V}} \frac{a(u,v)}{|||u||| \, |||v|||} \geq m. \tag{4.1}$$

The constant m *is independent of* h *but depends on* p*.*

Proof. Using identities (3.46)–(3.48), we have

$$a(u_c + u_d, v_c + v_d) = a_{\mathcal{K}}(u_c, v_c) + 2a_{\mathcal{K}}(u_d, v_c) + a_{\mathcal{K}}(u_d, v_d). \tag{4.2}$$

Setting

$$v_c + v_d = u_c + \gamma u_d, \tag{4.3}$$

where $\gamma \in \mathbf{R}$ is a parameter, we get

$$a(u_c + u_d, v_c + v_d) = |||u_c|||_{\mathcal{K}}^2 + 2a_{\mathcal{K}}(u_d, u_c) + \gamma|||u_d|||_{\mathcal{K}}^2 \tag{4.4}$$

$$\geq |||u_c|||_{\mathcal{K}}^2 - 2\,|||u_d|||_{\mathcal{K}}|||u_c|||_{\mathcal{K}} + \gamma|||u_d|||_{\mathcal{K}}^2 \tag{4.5}$$

$$\geq (1-\epsilon)|||u_c|||_{\mathcal{K}}^2 + (\gamma - \epsilon^{-1})|||u_d|||_{\mathcal{K}}^2. \tag{4.6}$$

Here we used the Cauchy–Schwarz inequality and the inequality $2ab \leq \epsilon a^2 + \epsilon^{-1}b^2$ for any a, b, and $\epsilon \in \mathbf{R}$ with $\epsilon > 0$. Choosing ϵ such that $1 - \epsilon \geq m'$ and $\gamma \geq 1$ such that $\gamma - \epsilon^{-1} \geq m'$, we get

$$a(u_c + u_d, u_c + \gamma u_d) \geq m'\Big(|||u_c|||_{\mathcal{K}}^2 + |||u_d|||_{\mathcal{K}}^2\Big). \tag{4.7}$$

Next, using the norm equivalence (3.4), we note that, for $\gamma \geq 1$, we have

$$c_1|||u_c + \gamma u_d|||^2 \leq |||u_c|||_{\mathcal{K}}^2 + \gamma^2|||u_d|||_{\mathcal{K}}^2 \tag{4.8}$$

$$\leq \gamma^2\Big(|||u_c|||_{\mathcal{K}}^2 + |||u_d|||_{\mathcal{K}}^2\Big), \tag{4.9}$$

and thus we conclude that

$$|||u_c + u_d|||\;|||u_c + \gamma u_d||| \leq c_1^{-1}\gamma\Big(|||u_c|||_{\mathcal{K}}^2 + |||u_d|||_{\mathcal{K}}^2\Big). \tag{4.10}$$

Combining (4.7) and (4.10), we immediately get the desired inf-sup bound

$$\inf_{u\in\mathcal{V}}\sup_{v\in\mathcal{V}}\frac{a(u,v)}{|||u|||\;|||v|||} \geq \frac{c_1 m'}{\gamma} = m. \qquad \square \tag{4.11}$$

Remark 4.1. In [2] the concept of weak stability $m|||v|||_{\mathcal{K}}^2 \leq a(v,v)$ is discussed. We note that, while weak stability is obvious for the nonsymmetric dG method, it is nontrivial to derive error estimates of $|||u - u_h|||$ since the bilinear form is not bounded with respect to the $|||\cdot|||_{\mathcal{K}}$. Further, $|||\cdot|||_{\mathcal{K}}$ is only a seminorm while $|||\cdot|||$ is a norm.

Example: Computation of the inf-sup constant. We compute the inf-sup constant for the discrete Laplacian defined by (2.6) on the unit square $\Omega = [0,1]^2$ with homogenous Dirichlet conditions on Γ. The triangulations are quasi-uniform unstructured with N elements. For details on such computations we refer to Oden, Babuška, and Baumann [12]. In Table 4.1 we present the inf-sup constant m for a variety of triangulations and $p = 1, \ldots, 4$. We note that the inf-sup constant is independent of the number of elements (or meshsize) and decreases with increasing $p \geq 2$, as expected. Note also that for $p = 1$ the inf-sup constant is indeed strictly positive due to the fact that these computations are done on an unstructured grid in two spatial dimensions, which is typically not close to a checkerboard mesh.

4.2. Error estimates in the energy norm. We first recall that given $u \in H^s(K)$, there is $\pi_K u \in \mathcal{P}_p(K)$ such that the following estimate holds:

$$\|u - \pi_K u\|_{r,K} \leq c p_K^{r-s} h_K^{\mu-r}|u|_{s,K}, \tag{4.12}$$

TABLE 4.1
The inf-sup constant m for different p and meshes with N elements.

N	$p=1$	$p=2$	$p=3$	$p=4$
72	0.054	0.116	0.071	0.047
290	0.022	0.115	0.068	0.044
1300	0.022	0.115	0.067	0.044
2604	0.021	0.116	0.070	–
5366	0.023	0.115	–	–

where $0 \leq r \leq s$, $\mu = \min(p+1, s)$, and c is a constant independent of h and p; see [4]. Further, we let $\pi u \in \mathcal{V}$ be defined by $(\pi v)|_K = \pi_K(v|_K)$. Using (4.12), we get the following lemma.

LEMMA 4.1. *The following interpolation error estimate holds:*

$$|||u - \pi u||| \leq c\left(\sum_{K\in\mathcal{K}} p_K^{-(2s-3)} h_K^{2(\mu-1)} |u|_{s,K}^2\right)^{1/2}. \tag{4.13}$$

Proof. With $\eta = u - \pi u$ we have

$$|||\eta|||^2 = |||\eta|||_{\mathcal{K}}^2 + \|\langle\sigma_n(\eta)\rangle\|_{\mathcal{E}}^2 + \|h^{-1}[\sigma_n(\eta)]\|_{\mathcal{E}}^2.$$

Using the boundedness of A, we get $|||\eta|||_{\mathcal{K}}^2 \leq c\sum_{K\in\mathcal{K}} \|\eta\|_{1,K}^2$. For the second term we invoke the trace inequality (2.17) elementwise to obtain

$$\begin{aligned}\|\langle\sigma_n(\eta)\rangle\|_{\mathcal{E}}^2 &\leq c\sum_{K\in\mathcal{K}} h\|\nabla\eta\|_K\Big(h^{-1}\|\nabla\eta\|_K + \|\nabla\eta\|_{1,K}\Big)\\ &\leq c\sum_{K\in\mathcal{K}} \|\eta\|_{1,K}\Big(\|\eta\|_{1,K} + h\|\eta\|_{2,K}\Big).\end{aligned}$$

For the third term we get in the same way

$$\|h^{-1}[\eta]\|_{\mathcal{E}}^2 \leq c\sum_{K\in\mathcal{K}} h^{-1}\|\eta\|_K\Big(h^{-1}\|\eta\|_K + \|\eta\|_{1,K}\Big).$$

Now (4.13) follows directly from the interpolation error estimate (4.12). ∎

Here we used the multiplicative trace inequality to estimate the edge contributions. We refer to [7] for a discussion of the suboptimality with respect to p resulting from this trace inequality and an alternative estimate.

Using the stability estimate in Theorem 4.1 and the interpolation error estimate in Lemma 4.1, we obtain the following energy norm error estimate using standard arguments. The error estimate is optimal in h but suboptimal in p by a factor $1/2$ modulo the dependence of m on p.

THEOREM 4.2. *The following energy norm error estimate holds:*

$$|||u - u_h|||^2 \leq c(1 + m^{-1})\left(\sum_{K\in\mathcal{K}} p_K^{-(2s-3)} h_K^{2(\mu-1)} |u|_{s,K}^2\right)^{1/2},$$

where c denotes constants independent of h and p. The constant m, defined in Theorem 4.1, is independent of h but depends on p.

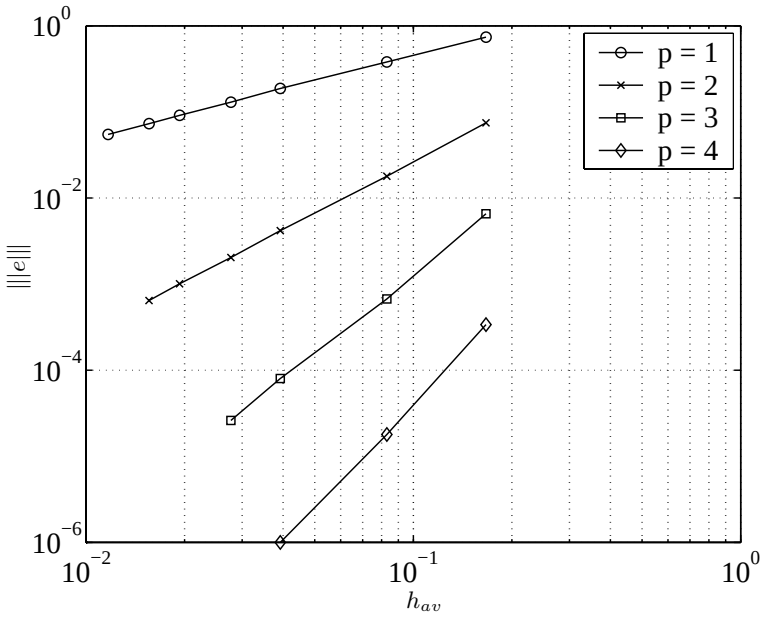


FIG. 4.1. *The energy error as a function of the average meshsize h_{av} for $p = 1, \ldots, 4$.*

Example: The error in the energy norm. We consider the Poisson equation (2.1) on the unit square, $\Omega = [0,1]^2$, with homogeneous Dirichlet boundary conditions, $u = 0$, on the boundary Γ and the right-hand side f chosen so that the exact solution is $u(x,y) = \sin(\pi x)\sin(\pi y)$. The triangulation is unstructured and all triangles are approximately the same size. In Figure 4.1 we plot the error as a function of the average meshsize h_{av} defined by $h_{av} = \sqrt{2N}$, where N is the number of elements. We observe the expected convergence of order $p-1$ for polynomials of order p.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1974.

[2] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] I. BABUŠKA, C. E. BAUMANN, AND J. T. ODEN, *A discontinuous hp finite element method for diffusion problems: 1-d analysis*, Comput. Math. Appl., 37 (1999), pp. 103–122.

[4] I. BABUŠKA AND M. SURI, *The h-p version of the finite element method with quasiuniform meshes*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 199–238.

[5] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1994.

[6] V. GIRAULT AND P.-A. RAVIART, *Finite Element Methods for Navier-Stokes Equations Theory and Algorithms*, Springer-Verlag, New York, 1986.

[7] P. HOUSTON, C. SCHWAB, AND E. SÜLI, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[8] T. J. R. HUGHES, G. ENGEL, L. MAZZEI, AND M. G. LARSON, *A comparision of discontinuous and continuous Galerkin methods based on error estimates, conservation, robustness, and efficiency*, in Discontinuous Galerkin Methods, Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin, 2000, pp. 135–146.

[9] M. G. LARSON AND A. J. NIKLASSON, *Analysis of a Family of Discontinuous Galerkin Methods for Elliptic Problems: The One Dimensional Case*, Preprint 12, Chalmers Finite Element Center, Chalmers University of Technology, Göteborg, Sweden, 2001.

[10] L. S. D. Morley, *The triangular equilibrium element in the solution of plate bending problems*, Aero. Quart., 19 (1968), pp. 149–169.
[11] J. Nitsche, *Über ein Variationzprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.
[12] J. T. Oden, I. Babuška, and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.
[13] B. Rivière, M. F. Wheeler, and V. Girault, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems* I, Comput. Geosci., 3 (2000), pp. 337–360.
[14] B. Rivière, M. F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.
[15] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 265–282

PARALLEL TWO-STEP W-METHODS WITH PEER VARIABLES*

BERNHARD A. SCHMITT† AND RÜDIGER WEINER‡

Abstract. A new class of methods for the solution of stiff initial value problems is introduced that is parallel by design. It has a two-step character and propagates s different "peer" solution variables with essentially identical characteristics from step to step. The main work lies in the solution of s independent linear stage equations which may be solved in parallel. Convergence of order $s-1$ and stability for general stepsize sequences are proved. Conditions for order s and stronger stability criteria are addressed as well. Promising methods up to order 7 are identified by numerical tests with some widely used stiff test problems. Some of these are competitive with existing software even in sequential computations.

Key words. stiff initial value problems, parallel two-step methods, parallel peer W-methods

AMS subject classifications. 65L06, 65Y05

DOI. 10.1137/S0036142902411057

1. Introduction.

Parallel solution methods for large and stiff initial value problems

$$y' = f(t,y), \quad t_0 \le t \le t_e, \quad y(t_0) = y_0 \in \mathbb{R}^n, \tag{1}$$

may be based on many different strategies. In this paper we will consider a new class of time integration schemes with inherent "method parallelism" [8]. This feature is independent of "parallelism across the system" since both kinds of parallelism may be applied simultaneously in many situations. Approaches for method parallelism may start with well-known classical methods and try to parallelize expensive parts of its numerical components, like the solution of the large nonlinear systems in implicit Runge–Kutta methods (e.g., [1]). Similarly, parallel iteration schemes may be applied to classical one-step or multistep methods as in some works of van der Houwen and Sommeijer and van der Houwen and de Swart [11, 12] and Burrage and Suhartanto [3]. For a recent textbook on parallel methods, see [2]. Other approaches are based on new kinds of methods that are parallel by design. Within the wide class of general linear methods (GLMs), the subclass of diagonally implicit multistage integration methods (DIMSIMs) suitable for parallel implementation has been proposed by Butcher [5]. Here, the equations for the internal stages of the GLM may be solved in parallel since the corresponding part of the coefficient matrix is diagonal. The papers [6, 7] are concerned with implementation issues of such methods. Unfortunately, the practical experience with these methods is not overly positive compared to RADAU [10] or parallel software for implicit differential equations (PSIDE) [12].

In order to circumvent some of the theoretical bottlenecks in the structure of classical one-step and multistep methods, the class of linearly implicit *parallel two-step W-methods* (PTSW-methods) has been discussed recently by Podhaisky and the authors [15]. These PTSW-methods are similar to ordinary Rosenbrock–Wanner (ROW)- or

*Received by the editors July 16, 2002; accepted for publication (in revised form) July 11, 2003; published electronically January 28, 2004.
http://www.siam.org/journals/sinum/42-1/41105.html
†Fachbereich Mathematik und Informatik, Universität Marburg, D-35032 Marburg, Germany (schmitt@mathematik.uni-marburg.de).
‡Fachbereich Mathematik und Informatik, Universität Halle, D-06099 Halle, Germany (weiner@mathematik.uni-halle.de).

W-methods (cf. [13], [10, IV.7]) but use s stage increments from the previous time step only. So, all s current stages can be processed in parallel. The PTSW-methods are already quite competitive [16, 19] even in sequential computations. They are particularly attractive in very stiff or singularly perturbed problems since they do not suffer an order reduction due to high stage orders. One of their weaknesses, however, is some critical dependence on stepsize ratios in the stiff case.

Literally speaking, many parallel and classical time integration methods have a set of distinguished or "master" variables and compute additional "slave" variables to improve accuracy or stability properties of the masters. In fact, these methods usually employ only one n-dimensional master approximation for the solution in each time interval with distinguished accuracy and stability properties. In contrast to these schemes, we now consider methods having only peer variables sharing the same accuracy and stability properties (with minor modifications). An immediate advantage of this approach is the existence of a continuous extension for these methods by using an interpolating polynomial. In this paper we restrict the discussion to two-step methods that compute several solution approximations Y_{mi}, $i = 1, \dots, s$, associated with a time interval $[t_m, t_{m+1}]$ from the information contained in the variables $Y_{m-1,i}$ from the previous interval. Generalization to methods using even earlier information is obvious but will not be considered here. Moreover, we will concentrate on linearly implicit methods avoiding the solution of nonlinear systems of equations. In parallel "peer" two-step W-methods (PPSW-methods), the solutions Y_{mi}, $i = 1, \dots, s$, are related to points

$$t_{mi} := t_m + h_m c_i, \quad i = 1, \dots, s, \tag{2}$$

associated with the time interval $[t_m, t_{m+1}]$ but not necessarily contained in it. The PPSW-methods are given by

$$\begin{aligned}(I - \gamma_i h_m T_m) Y_{mi} = &\sum_{j=1}^{s} \big(b_{ij} I + h_m \gamma_{ij} T_m\big) Y_{m-1,j} \\ &+ h_m \sum_{j=1}^{s} a_{ij} f\big(t_{m-1,j}, Y_{m-1,j}\big), \quad i = 1, \dots, s.\end{aligned} \tag{3}$$

The terms $\gamma_i > 0$, $b_{ij}, \gamma_{ij}, a_{ij}$ are the parameters of the method. The matrix T_m should be an approximation of the Jacobian $f_y(t_m, y(t_m))$ for stability reasons only. In fact, the accuracy of these methods is derived for arbitrary T_m in the sense of W-methods. In this context it is no essential restriction to consider autonomous problems, and we will do this for simplicity. The subclass of these methods with $\gamma_i = 0$, $\gamma_{ij} = 0$ may also be attractive for nonstiff problems but will not be discussed here.

Introducing the stage vectors $Y_m = (Y_{mi})_{i=1}^{s} \in \mathbb{R}^{sn}$ and coefficient matrices

$$G := \operatorname{diag}(\gamma_i),\ A = (a_{ij}),\ B = (b_{ij}),\ \Gamma = (\gamma_{ij}),\ \beta := A + \Gamma, \tag{4}$$

a more compact version of the PPSW-method is

$$(I - h_m G \otimes T_m) Y_m = (B \otimes I + h_m \Gamma \otimes T_m) Y_{m-1} + h_m (A \otimes I) f(Y_{m-1}). \tag{5}$$

The matrix β introduced in (4) will play an important role in the analysis. In (5) it is easily seen that for methods using $G = \gamma I$ the matrix $I - h_m G \otimes T_m$ commutes with all other matrices in the scheme. Hence the stability analysis simplifies considerably.

Similar to the notion in implicit Runge–Kutta methods, we will also call methods using

$$\gamma_i \equiv \gamma, \quad G = \gamma I \tag{6}$$

singly implicit. We will concentrate on this kind of method in the present paper and display its specific form for ease of reference:

$$\big(I \otimes (I - \gamma h_m T_m)\big) Y_m = (B \otimes I + h_m \Gamma \otimes T_m) Y_{m-1} + h_m (A \otimes I) f(Y_{m-1}). \tag{7}$$

Such methods may be attractive already in a sequential computing environment since the matrix decomposition of $I - \gamma h_m T_m$ may be used in all stages. A situation of interest on parallel machines is the use of expensive parallel preconditioners for this matrix. Still, multi-implicit methods with a general diagonal matrix G offer additional design options and will be discussed in [18].

The emphasis of our discussion lies on higher-order methods with an order near the number of stages s. Here, a critical source of problems is the fact that the coefficients of the methods depend on the current stepsize ratio σ_m defined in (8). Stepsize increases will be restricted by an upper bound

$$\sigma_m := h_m / h_{m-1} \le \bar{\sigma} \tag{8}$$

for all steps, but no lower bound is assumed since this might be a severe restriction in practice. When discussing nonlocal effects we will occasionally add an additional step-index m to the coefficient matrices A, B, Γ, β (see (4)) to indicate this dependence, and the same notation will be used for other quantities. Convergence results are formulated in terms of the maximal stepsize

$$H := \max_{j \ge 0} h_j. \tag{9}$$

Further abbreviations are $\mathbb{1} = (1, \ldots, 1)^{\mathsf{T}}$, e_i for the ith unit vector. The spectral radius of a matrix is denoted by ρ.

After this introduction the paper continues with basic aspects of stability and accuracy of the schemes. Conditions for stability with general stepsizes and the structure of the stability matrix are derived as well as basic order conditions. The conditions for order $s - 1$ lead to explicit representations of the coefficients of the scheme. We concentrate on these high-order schemes in section 3 and identify one class where the general stability conditions from section 2 can be established easily. We also discuss conditions for improving the order to s. For the singly implicit methods in the form (7) it is unlikely that this can be achieved by improving all local errors by one order. However, we show that there is some superconvergence effect for the global error for certain parameter choices. Unfortunately, this effect can be conveniently exploited for constant stepsizes only. In section 4 we finally discuss implementation issues like error estimation and present numerical results with different test problems and methods with up to eight stages.

2. Basic properties.

2.1. Stability issues. Due to the two-step structure of the scheme and the σ-dependence of its coefficients, the stability analysis encounters many of the difficulties of multistep methods. In addition to the step recursion (5), we consider a second one

with additional perturbations $h_m g_m = h_m(g_{mi})_{i=1}^s \in \mathbb{R}^{sn}$ and solutions $Y_m + X_m$. So, for the error X_m we have the recursion

$$(I - h_m G_m \otimes T_m)X_m = (B_m \otimes I + h_m \Gamma_m \otimes T_m)X_{m-1} \tag{10}$$

$$+ h_m(A_m \otimes I)(f(Y_{m-1} + X_{m-1}) - f(Y_{m-1})) + h_m g_m,$$

where the coefficient matrices are supplemented by the step-index m. The stability of the recursion (10) is covered by the theory for multistep methods; see [9]. For ease of reference we formulate it in the following lemma. The crucial assumption (11) will be verified for our methods in section 3.1.

LEMMA 2.1. *Assume that for some fixed $\bar\sigma > 1$ and stepsize sequences with $h_m \le \bar\sigma h_{m-1}$ a uniform bound $\bar b$ exists for all products*

$$\|B_{m+k} \cdots B_{m+1} B_m\| \le \bar b, \quad m, k \ge 0, \tag{11}$$

where $\sum_{j=m}^{m+k} h_j \le t_e - t_0$. Let the maps $X \mapsto (I - h_m G_m \otimes T_m)^{-1}((G_m\beta_m + \Gamma_m) \otimes T_m)X + (A_m \otimes I)(f(Y_{m-1}+X) - f(Y_{m-1}))$ be uniformly Lipschitz continuous in some neighborhood of zero. Then there exists a constant C such that

$$\|X_m\| \le C(\|X_0\| + \max_{j=1}^{m} \|g_j\|) \quad \textit{for} \quad t_0 \le t_m \le t_e.$$

The standard application of this stability lemma concerns the convergence of the scheme, where the "perturbed" solution $Y_m + X_m = y(t_m)$ is the solution of the initial value problem and the perturbations g_m are the corresponding residuals defined in (13). Here, Lemma 2.1 shows that the scheme (3) converges with order p, i.e., $X_m = O(H^p)$, if the local error is of order p too; i.e., $g_m = O(H^p)$. So we do not need to distinguish between the local and global order of the method except in the case of superconvergence; cf. section 3.2.

In the stiff context it is appropriate to consider the linear autonomous problem

$$y' = Jy$$

as a test equation, where the eigenvalues of J lie in the left complex halfplane. With the choice

$$T_m = J$$

the scheme (3) reduces to the recursion (see (4))

$$Y_m = M_m(h_m J)Y_{m-1}, \quad M_m(z) := (I - zG_m)^{-1}(B_m + z\beta_m), \tag{12}$$

where $M_m(\cdot)$ is the stability matrix of the scalar test equation $y' = \lambda y$ and $z = h_m\lambda$. Unfortunately, M_m depends on the current stepsize ratio σ_m, and only for constant stepsizes the spectral radius of M is an appropriate measure for the longtime behavior of the recursion (12). In the case $\sigma \equiv 1$ we may apply some standard stability notions to the scalar function $\rho(M(z))$.

DEFINITION 2.2. *Let $M(.)$ be the stability matrix of a PPSW-method* (3) *as defined in* (12) *with $\sigma = 1$. Then, the method is called* zero-stable *if $\rho(M(0)) = 1$ and the eigenvalues on the unit circle are simple. It is* A*-stable if $\rho(M(z)) < 1$ in the open left complex halfplane, $z \in \mathbb{C}_- := \{z \in \mathbb{C} : \ \mathrm{Re}\ z < 0\}$, and* L*-stable if, additionally, $\rho(M(\infty)) = 0$.*

Remarks. (a) Any consistent method must reproduce constant solutions for $\lambda = 0$. Since this requirement leads to the identity $B\mathbb{1} = \mathbb{1}$, the spectral radius of $M(0) = B$ cannot be smaller than one.

(b) For singly implicit methods (7) we have $M(\infty) = -\frac{1}{\gamma}\beta$, and L-stability requires $\rho(\beta) = 0$.

(c) It is convenient to introduce the variable $w = z/(1-\gamma z)$ satisfying $(1+\gamma w)(1-\gamma z) = 1$. For $z \in \mathbb{C}_-$ the variable w is contained in the circle centered at $-1/(2\gamma)$ and going through the origin. Now, the stability matrix of singly implicit methods (7) is a linear function of w given by

$$M(z) = (1 + \gamma w)B + w\beta = B + w(\gamma B + \beta).$$

This form may be conveniently employed to check the A-stability of the scheme since M has an eigenvalue $\lambda = e^{i\theta}$ on the unit circle if w is an eigenvalue of the generalized eigenvalue problem [17],

$$(e^{i\theta} I - B)x = w(\gamma B + \beta)x, \quad x \in \mathbb{C}^s.$$

A thorough and general analysis of stability properties of our methods is quite difficult since the stability matrix M_m depends on the actual stepsize ratio σ_m for accuracy reasons. However, this is a common problem for multistep methods where even proving zero-stability is nontrivial for general stepsize sequences. For PPSW-methods this case will be dealt with in section 3.1 by verifying (11). A similar result covering the stiff limit $z = h\lambda \to \infty$ for arbitrary stepsize ratios is presented there too. However, stronger z-uniform stability results for these methods with nonconstant stepsize sequences are not yet available.

2.2. Accuracy conditions. The structure of the coefficient matrices is determined to a large extent by accuracy requirements. The accuracy of the PPSW-methods may be analyzed in a standard way by considering the residuals Δ_{mi} obtained when the exact solution is put into the method. From now on we discuss singly implicit methods (7) only. Using the information $f(t_{m-1,j}, y(t_{m-1,j})) = y'(t_{m-1,j})$, we consider

$$h_m\Delta_{mi} := (I - \gamma h_m T_m)y(t_{mi}) - \sum_{j=1}^{s} \big(b_{ij} I + h_m\gamma_{ij} T_m\big)y(t_{m-1,j}) - h_m \sum_{j=1}^{s} a_{ij} y'(t_{m-1,j}), \quad i = 1, \dots, s. \tag{13}$$

We will use Taylor expansion at t_m but write the error in terms of h_{m-1} since σ_m is bounded from above only, so $h_m = O(h_{m-1})$ but not vice versa. Separating the terms depending on T_m from the others, we see that

$$h_m T_m \left(\gamma y(t_{mi}) + \sum_{j=1}^{s} \gamma_{ij} y(t_{m-1,j}) \right) = O(h_{m-1}^{q+1})$$

holds if condition (see (2))

$$\Gamma(q): \quad \gamma c_i^k + \sum_{j=1}^{s} \gamma_{ij}(c_j - 1)^k \sigma_m^{-k} = 0, \quad k = 0, \dots, q-1, \tag{14}$$

is satisfied for a sufficiently smooth solution. Second, it holds that

$$y(t_{mi}) - \sum_{j=1}^{s} b_{ij} y(t_{m-1,j}) - h_m \sum_{j=1}^{s} a_{ij} y'(t_{m-1,j}) = O(h_{m-1}^q)$$

if condition

$$AB(q):\ c_i^k - \sum_{j=1}^{s} b_{ij} \left(\frac{c_j - 1}{\sigma_m}\right)^k - k \sum_{j=1}^{s} a_{ij} \left(\frac{c_j - 1}{\sigma_m}\right)^{k-1} = 0,\ k = 0, \ldots, q-1, \tag{15}$$

is true, where the tacit convention is used that the second sum is not evaluated for $k = 0$ and cancelled by the factor k holds. In both conditions q denotes the number of conditions.

LEMMA 2.3. *If the conditions* $\Gamma(q)$ *and* $AB(q+1)$ *are satisfied with* $q \geq 1$ *and the solution* y *of* (1) *is sufficiently smooth, then the residuals* (13) *for the scheme* (3) *are of order* q*; i.e.,* $\|\Delta_m\| = O(h_{m-1}^q)$.

With respect to the order, the most critical condition seems to be $\Gamma(q)$ in (14) since it depends on the $s^2 + 1$ coefficients γ, γ_{ij} only. In fact, requiring $\Gamma(s)$ leads to the explicit relation

$$\Gamma_m = -\gamma \Theta_m, \quad \Theta_m := V S_m P V^{-1}, \tag{16}$$

with the Vandermonde matrix V and the Pascal matrix P defined by

$$V = \left(c_i^{j-1}\right)_{i,j=1}^{s}, \quad P = \left(\binom{j-1}{i-1}\right)_{i,j=1}^{s}. \tag{17}$$

We also need the diagonal matrices

$$S_m = \operatorname{diag}(1, \sigma_m, \ldots, \sigma_m^{s-1}),\ D := \operatorname{diag}(1, \ldots, s). \tag{18}$$

The matrix Θ_m will be encountered quite often since it describes the polynomial extrapolation from the subgrid $\{t_{m-1} + h_{m-1} c_i,\ i = 1, \ldots, s\}$ to $\{t_m + h_m c_i,\ i = 1, \ldots, s\}$.

In a similar way, explicit solutions for the order conditions (15) may be obtained by requiring $AB(s)$. Here, we note that the matrix $\left(\left(\frac{c_j - 1}{\sigma_m}\right)^{k-1}\right)_{j,k=1}^{s} = V P^{-1} S_m^{-1}$ multiplies B (see (4)) from the right in (15). The corresponding factor for A is obtained from this one by scaling and shifting its columns to the right. The shift is described by the matrix $F_0 = (\delta_{i-1,j})$, which gives rise to the identities

$$P D F_0^{\mathsf{T}} = D F_0^{\mathsf{T}} P, \quad F_0^{\mathsf{T}} S_m = \sigma_m S_m F_0^{\mathsf{T}}, \tag{19}$$

with the matrices defined in (17), (18). The first relation in (19) corresponds to the identity $\binom{j-2}{i-1}(j-1) = i\binom{j-1}{i}$. Now, condition (15) has the matrix form $0 = V - BVP^{-1}S_m^{-1} - AVP^{-1}S_m^{-1}DF_0^{\mathsf{T}}$. Multiplying with $S_m P V^{-1}$ and using the relations (19), we obtain

$$B_m = \Theta_m - \sigma_m A_m V D F_0^{\mathsf{T}} V^{-1}. \tag{20}$$

Since stability properties depend on the matrix $\beta_m = A_m + \Gamma_m$, by (12) it is convenient to replace the coefficient A_m in the last equation by using β_m and the first order condition (14) as well.

LEMMA 2.4. *If the PPSW-method* (7) *satisfies* $AB(s)$ *and* $\Gamma(s-1)$*, then its coefficient matrices* B_m *and* β_m *are related through*

$$B_m = (I+\gamma E)\Theta_m + \sigma_m\beta_m E, \quad E = -VDF_0^{\mathsf{T}}V^{-1}. \tag{21}$$

The matrices E *and* Θ_m *commute for* $\sigma_m = 1$*, since* $E\Theta_m = \sigma_m\Theta_m E$.

Proof. In matrix form the condition $\Gamma(s-1)$ amounts to $\gamma VS_mP + \Gamma_m V = ue_s^{\mathsf{T}}$ with an arbitrary last column u. Using this equation and replacing $A_m = \beta_m - \Gamma_m$ in (20) give

$$B_m = \Theta_m - \sigma_m(\beta_m V + \gamma VS_mP)DF_0^{\mathsf{T}}V^{-1},$$

since the last column of $\Gamma_m V$ drops out after multiplication with the shift matrix F_0^{T}. Combining the two identities in (19) yields $DF_0^{\mathsf{T}}S_mP = \sigma_m S_mPDF_0^{\mathsf{T}}$. So we see that two factorizations $\sigma_m\Theta_m E = \sigma_m VS_mPV^{-1}VDF_0^{\mathsf{T}}V^{-1} = \sigma_m V(S_mPDF_0^{\mathsf{T}})V^{-1} = VDF_0^{\mathsf{T}}S_mPV^{-1} = E\Theta_m$ are possible. ∎

3. High-order methods.

3.1. Convergence for variable stepsize. In the case of one single parameter value γ as considered here, all coefficient matrices have a simpler structure in the monomial basis $(c^{j-1})_{j=1}^s$ contained in the Vandermonde matrix V. The transformed version of the matrix B of (21) is given by (see (16) and (18))

$$\tilde B_m := V^{-1}B_mV = S_mP - \gamma DF_0^{\mathsf{T}}S_mP - \sigma_m\tilde\beta_m DF_0^{\mathsf{T}},\ \tilde\beta_m := V^{-1}\beta_m V. \tag{22}$$

Note that the first column of $\tilde B_m$ is the first unit vector. According to (12), and considering that $G_m = \gamma I$, the stability matrix at infinity is $M_m(\infty) = -\frac{1}{\gamma}\beta_m$, and optimal stiff damping could be obtained by the choice $\beta_m = 0$. However, (22) shows that this choice violates the condition of zero-stability for $\sigma_m \ge 1$ since the leading matrix $S_mP - \gamma DF_0^{\mathsf{T}}S_mP$ is upper triangular and has the eigenvalues σ_m^{i-1}. Hence $\tilde\beta_m$ is needed for the stabilization of B_m. The following theorem provides a convenient compromise between stability for stiff problems and zero-stability.

THEOREM 3.1. *Let the method* (3) *satisfy the conditions* $AB(s)$ *and* $\Gamma(s-1)$ *and use one single* $\gamma = \gamma_i$*,* $i = 1,\dots,s$*. Then, with the choice*

$$\tilde\beta_m = V^{-1}\beta_m V = F_0S_mD^{-1}, \tag{23}$$

the method is both zero-stable and $\mathrm{L}(\alpha)$*-stable if it is* $\mathrm{A}(\alpha)$*-stable. In fact, both matrices* $B_m - 1e_1^{\mathsf{T}}$ *and* β_m *are nilpotent.*

Proof. The only nonzero elements of $\tilde\beta_m$ are $\tilde\beta_{i+1,i} = \frac{1}{i}\sigma_m^{i-1}$, $i = 1,\dots,s-1$. So, clearly, $\tilde\beta_m$ is nilpotent. This choice leads to $\sigma_m\tilde\beta_m DF_0^{\mathsf{T}} = S_mF_0F_0^{\mathsf{T}} = S_m\,\mathrm{diag}(0,1,\dots,1)$ in (22) and cancels the main diagonal of $\tilde B_m$ beyond the first entry. In fact, we have

$$\tilde B_m := V^{-1}B_mV = e_1e_1^{\mathsf{T}} + S_m(P-I) - \gamma DF_0^{\mathsf{T}}S_mP, \tag{24}$$

and $\tilde B_m - e_1u^{\mathsf{T}}$ is nilpotent for any vector u with $u^{\mathsf{T}}e_1 = 1$. So, $V(\tilde B_m - e_1u^{\mathsf{T}})V^{-1} = B_m - \mathbb{1}e_1^{\mathsf{T}}$ is nilpotent with $u = e_1^{\mathsf{T}}V = (1, c_1, \dots)$. ∎

By Theorem 3.1, the transformed coefficient matrix $\tilde B_m$ is fully specified. Due to the upper triangular structure, for any s, this matrix really is part of the large matrix

$$\begin{pmatrix} 1 & 1-\gamma\sigma & 1-2\gamma\sigma & 1-3\gamma\sigma & \cdots \\ & 0 & 2\sigma(1-\gamma\sigma) & 3\sigma(1-2\gamma\sigma) & \cdots \\ & & 0 & 3\sigma^2(1-\gamma\sigma) & \cdots \\ & & & & \ddots \end{pmatrix}. \tag{25}$$

Although the ratio σ (and even γ) may change between intervals, this representation shows that the nilpotency discussed in Theorem 3.1 is structurally stable. It really holds for general stepsize sequences since the Vandermonde transformation used in (24) is the same everywhere.

LEMMA 3.2. *Let the matrices B_m, $m \geq 1$, be given by* (22) *and the matrices β_m be as defined in Theorem* 3.1. *Then each product of at least $s-1$ matrices β_j, respectively, B_j, has rank one at most. In fact, it holds that*

$$\begin{aligned} \beta_m \beta_{m-1} \cdots \beta_{m-k} &= 0, \quad m, k \geq s-1, \\ B_m B_{m-1} \cdots B_{m-k} &= \mathbb{1}\, v^{\mathsf{T}}, \quad m, k \geq s-2, \end{aligned} \tag{26}$$

with $v^{\mathsf{T}}\mathbb{1} = 1$.

Proof. Without restriction we may consider the matrices $\tilde{\beta}_j, \tilde{B}_j$ defined in (22) since they are all transformed by the same matrix V. The result for the matrices $\tilde{\beta}_j$ is trivial due to their common triangular structure. Equation (24) shows that each $\tilde{B}_j$ has the structure

$$\tilde{B}_j = \begin{pmatrix} 1 & v_j^{\mathsf{T}} \\ 0 & \tilde{B}_{22}^{(j)} \end{pmatrix}, \tag{27}$$

where $\tilde{B}_{22}^{(j)}$ is *strictly* upper triangular. Hence $\tilde{B}_{22}^{(m)} \tilde{B}_{22}^{(m-1)} \cdots \tilde{B}_{22}^{(m-k)} = 0$ for $k \geq s-2$ and in the product $\tilde{B}_m \tilde{B}_{m-1} \cdots \tilde{B}_{m-k}$ only the first row is nontrivial and starts with the element 1: let us denote it by w^{T}. Consequently, from (24) it follows that $V^{-1} B_m \cdots B_{m-k} V = e_1 w^{\mathsf{T}}$, and then $B_m \cdots B_{m-k} = \mathbb{1} w^{\mathsf{T}} V^{-1} =: \mathbb{1} v^{\mathsf{T}}$. Finally, $v^{\mathsf{T}}\mathbb{1} = w^{\mathsf{T}} V^{-1} V e_1 = w^{\mathsf{T}} e_1 = 1$. ∎

Remark. We note that the row vector v^{T} in (26) depends on the stepsize ratios of the $s-1$ rightmost matrices in the product. However, since stepsize changes are bounded above by $\bar{\sigma}$, the vector v in (26) is bounded by some fixed power of $\bar{\sigma}$ (and γ).

This verifies the main assumption of the stability Lemma 2.1, and we may summarize most of the previous results in the following theorem.

THEOREM 3.3. *Let a grid $(t_m)_{m \geq 0}$ be given with bounded stepsize changes $\sigma_m \leq \bar{\sigma}$, and let $0 \leq \gamma \leq \bar{\gamma}$. Let the method* (7) *satisfy $AB(s)$ and $\Gamma(s-1)$, and let β_m be chosen by* (23). *If the initial values are accurate, i.e., $\|Y_{0i} - y(t_{0i})\| = O(h_0^{s-1})$, $i = 1, \ldots, s$, then the PPSW-method converges with order $s-1$, i.e.,*

$$\|Y_{mi} - y(t_{mi})\| = O(H^{s-1}), \quad t_0 \leq t_m \leq t_e, \ i = 1, \ldots, s.$$

Proof. For $X_{mi} = y(t_{mi}) - Y_{mi}$ the recursion (10) holds with $g_m = \Delta_m$ defined in (13) which is of order $O(h_{m-1}^{s-1})$ by assumption and Lemma 2.3. Since the main assumption (11) of the stability Lemma 2.1 has been verified in Lemma 3.2, with the present assumptions the assertion follows. ∎

The representations (21) and (24) show that the extrapolation matrix Θ_m defined in (16) is an important contribution in both coefficients B_m and Γ_m (see (21) and (16)). Applying a standard reformulation of W-methods to the scheme (7), avoiding unnecessary multiplications with T_m, reveals a different interpretation. For high-order methods satisfying $\Gamma(s)$ this version reads

$$\begin{aligned} (I - \gamma h_m I \otimes T_m)\big(Y_m - (\Theta_m \otimes I) Y_{m-1}\big) = \big((B_m - \Theta_m) \otimes I\big) Y_{m-1} \\ + h_m (A_m \otimes I) f(Y_{m-1}), \end{aligned} \tag{28}$$

where $A_m = \gamma\Theta_m + \beta_m$ (see (4) and (16)). It may be interpreted as a corrector equation for the predictor $(\Theta_m \otimes I)Y_{m-1}$ obtained by polynomial extrapolation. The implementation of PPSW-methods will be based on this formulation in section 4.

3.2. Superconvergence of PPSW-methods. So far, we have considered methods of order $s-1$ only. It would be of interest, of course, to achieve even higher orders in PPSW-methods. In light of Lemma 2.3, order s requires that the conditions $\Gamma(s)$ and $AB(s+1)$ hold. While the first condition is already satisfied by $\Gamma_m = -\gamma\Theta_m$ defined in (16), the second one, $AB(s+1)$, leads to s additional nonlinear restrictions. Before formulating this result, we point out that the representation (24) of the coefficient matrix B_m may be written as the rule

$$\sum_{j=1}^{s} b_{ij}\psi(c_j) = \psi(1+\sigma_m c_i) - \psi(\sigma_m c_i) + \psi(0) - \gamma\sigma_m\psi'(1+\sigma_m c_i), \tag{29}$$

$i = 1, \dots, s$, for any polynomial ψ of degree $s-1$.

LEMMA 3.4. *Consider the PPSW-method with $\Gamma(s)$ and the coefficients defined in Theorem 3.1, let $\phi(t) := \prod_{i=1}^{s}(t-c_i)$ be the knot polynomial of the scheme, and let the solution y be sufficiently smooth. Then, for each i, $1 \le i \le s$, the condition*

$$\gamma\sigma_m\phi'(1+\sigma_m c_i) = \phi(1+\sigma_m c_i) - \phi(\sigma_m c_i) + (\sigma_m c_i)^s + \phi(0) \tag{30}$$

implies $\Delta_{mi} = O(h_{m-1}^s)$.

Proof. With $\Gamma(s)$ and (24) the residual (13) has the form

$$h_m\Delta_{mi} = \frac{h_{m-1}^s}{s!}\left((1+\sigma_m c_i)^s - \sum_{j=1}^{s} b_{ij}c_j^s - \sigma_m s\sum_{j=1}^{s} a_{ij}c_j^{s-1}\right) y^{(s)}(t_{m-1}) + O(h_{m-1}^{s+1}), \tag{31}$$

where Taylor expansion at t_{m-1} was used for convenience. For the terms (see (4)) $a_{ij}c_j^{s-1} = (\beta_{ij} - \gamma_{ij})c_j^{s-1}$ the condition $\Gamma(s)$ gives $\sum_j \gamma_{ij}c_j^{s-1} = -\gamma(1+\sigma_m c_i)^{s-1}$. Since the last column of $\tilde{\beta}_m$ in Theorem 3.1 is zero, the corresponding contribution $\sum_{j=1}^{s}\beta_{ij}c_j^{s-1} = 0$ is missing. So, the leading bracket in (31) has the form

$$(1+\sigma_m c_i)^s - \sigma_m s\gamma(1+\sigma_m c_i)^{s-1} - \sum_{j=1}^{s} b_{ij}c_j^s =: u_i. \tag{32}$$

By defining the polynomial $\psi(t) := \phi(t) - t^s$, we observe that $\psi(c_j) = -c_j^s$. Moreover, it has degree $s-1$ so that (29) applies. Consequently, from (32) we obtain

$$\begin{aligned} u_i &= (1+\sigma_m c_i)^s - \sigma_m s\gamma(1+\sigma_m c_i)^{s-1} + \sum_{j=1}^{s} b_{ij}\psi(c_j) \\ &= (1+\sigma_m c_i)^s + \psi(1+\sigma_m c_i) - \sigma_m\gamma\big(s(1+\sigma_m c_i)^{s-1} + \psi'(1+\sigma_m c_i)\big) \\ &\quad - \psi(\sigma_m c_i) + \psi(0) \\ &= \phi(1+\sigma_m c_i) - \sigma_m\gamma\phi'(1+\sigma_m c_i) - \psi(\sigma_m c_i) + \psi(0) \\ &= \phi(1+\sigma_m c_i) + (\sigma_m c_i)^s - \phi(\sigma_m c_i) + \phi(0) - \sigma_m\gamma\phi'(1+\sigma_m c_i). \end{aligned} \tag{33}$$

So, $u_i = 0$ implies $\Delta_{mi} = O(h_{m-1}^s)$ in (31). ∎

Since the recursion (10) combines all previous errors, the gain of one order from $s-1$ to s in the global error requires that condition (30) hold for all stages $i = 1,\dots,s$. However, these s conditions are very severe restrictions for the remaining $s+1$ parameters $\gamma, c_1, \dots, c_s$ and are unlikely to be satisfied for some interesting range of σ-values. So, we will not pursue this line of research further. However, later on, we will discuss an application of using one single superconsistency condition (30) in error control. For multi-implicit methods, i.e., methods in the general form (3), the corresponding conditions (30) can easily be satisfied by using different γ_i; see [18].

For some PPSW-methods there is a different global effect leading to order-s convergence in the error. Unfortunately, this superconvergence may be conveniently exploited for constant stepsizes only.

By a careful choice of parameters the structural information (26) may be used to eliminate the leading term in the global error. In fact, adding the error residuals Δ_m from (13), the exact solution $y_m = \big(y(t_{mi})\big)_{i=1}^{s}$ obeys the modified time step recursion

$$I \otimes (I - \gamma h_m T_m) y_m = (B_m \otimes I + h_m \Gamma_m \otimes T_m) y_{m-1} + h_m (A_m \otimes I) f(y_{m-1}) + h_m \Delta_m,$$

and the error $X_m = (Y_{mi} - y(t_{mi}))_{i=1}^{s}$ satisfies (10), which we now write as

$$X_m = (B_m \otimes I + h_m \mathbf{D_m}) X_{m-1} - h_m \big(I \otimes (I - \gamma h_m T_m)^{-1}\big) \Delta_m. \tag{34}$$

The matrix $\mathbf{D_m} = I \otimes (I - \gamma h_m T_m)^{-1}((\Gamma_m + \gamma B_m) \otimes T_m + (A_m \otimes I)\mathbf{J_m})$ is obtained by using the representation

$$f(Y_{mi}) - f(y_{mi}) = \mathbf{J_{mi}}(Y_{mi} - y_{mi}), \quad \mathbf{J_{mi}} = \int_0^1 f'(y_{mi} + t(Y_{mi} - y_{mi}))\, dt$$

with the block diagonal matrix $\mathbf{J_m} = \operatorname{diag}(\mathbf{J_{mi}})$. The structural result in Lemma 3.2 shows that the following computations may still be performed for nonconstant stepsizes. However, the dependence on the different stepsize ratios becomes overly complicated for higher-order methods and may be of no practical use. So we will restrict this discussion to the case $\sigma_m \equiv \sigma = 1$. Then, all matrices B_m are identical and we drop the index. Furthermore, the vector v in (26) is a fixed vector, namely, the left eigenvector of B to the eigenvalue 1 and will be given explicitly below. As a first step we simplify the recursion (34) by assuming $O(\|\mathbf{D_m} X_{m-1}\|) = O(\|X_{m-1}\|)$. Since our main argument is independent of the dimension n, we write only the scalar case $n = 1$ for simplicity. We obtain

$$\begin{aligned}
X_m &= B X_{m-1} - h\Delta_m + O(h\|X_{m-1}\| + h^2\|\Delta_m\|) \\
&= -h \sum_{j=0}^{m-1} B^j \Delta_{m-j} + B^m X_0 + \sum_{j=1}^{m-1} O(h\|X_{m-j}\| + h^2\ \|\Delta_{m-j+1}\|) \\
&= -h \mathbb{1} \sum_{j=s-1}^{m-1} v^{\mathsf{T}} \Delta_{m-j} \\
&\quad - h \sum_{j=0}^{s-2} B^j \Delta_{m-j} + B^m X_0 + \sum_{j=1}^{m} O(h\|X_{m-j}\| + h^2\|\Delta_{m-j+1}\|).
\end{aligned}$$

Now, if the method has order $s-1$ and the initial error X_0 is of appropriate size, all terms in the last equation are of order h^s except the first one, where (26) was used.

However, by an appropriate choice of the vector v in $B^\infty = \mathbb{1}v^\mathsf{T}$ the leading error term in $v^\mathsf{T}\Delta_{m-j}$, $j \ge s-1$, may be eliminated as well.

THEOREM 3.5. *Let the method* (7) *satisfy* $AB(s)$ *and* $\Gamma(s)$*, and use the matrix* β_m *from Theorem* 3.1*. Then, for constant stepsizes the vector* v^T *in* (26) *depends only on* γ *and the knots* c_i*. In fact,* $v^\mathsf{T}V$ *depends on* γ *only and is given by the first* s *components of the vector*

$$\tilde{v}^\mathsf{T} = (1, 1-\gamma, 3-6\gamma+2\gamma^2, 13-39\gamma+30\gamma^2-6\gamma^3, \ldots).$$

If the solution y *of* (1) *is smooth enough, then*

$$v^\mathsf{T}\Delta_m = \tilde{v}_{s+1}\frac{h^{s-1}}{s!}\,y^{(s)}(t_{m-1}) + O(h^s).$$

Proof. Since the stepsize ratio is constant, $\sigma = 1$, the upper triangular matrix $\tilde{B}$ (see (27)) depends on γ only and is explicitly given by the principal submatrix of order s of (25). Hence its left eigenvector $v^\mathsf{T}V$ to the eigenvalue 1 is a function of γ, and its ith component $\tilde{v}_i$ depends on the first i columns of $\tilde{B}$ only. These components may be computed by computer algebra and the first ones are given in the statement. Introducing the s-vector $\varphi = (\phi_0, \ldots, \phi_{s-1})^\mathsf{T}$ of coefficients of the knot polynomial $\phi(t) = (t-c_1)\cdots(t-c_s) = \sum_{k=0}^{s}\phi_k t^k$ defined above, we may write $0 = (\phi(c_i))_{i=1}^s = V\varphi + (c_i^s)_{i=1}^s$. In order to obtain a vector representation of the local residual (31) we introduce additional terms in (33), again canceling the highest power c_i^s. By considering that now $\sigma = 1$, this yields

$$\begin{aligned} u &= \left((1+c_i)^s - \phi(c_i) - s\gamma(1+c_i)^{s-1} + \sum_{j=1}^{s} b_{ij}\big(\phi(c_j) - c_j^s\big)\right)_{i=1}^s \\ &= \left(\sum_{k=1}^{s}\binom{s}{k-1}c_i^{k-1}\right)_{i=1}^s - V\varphi - s\gamma VPe_s + BV\varphi. \end{aligned}$$

Since the coefficient vector φ is multiplied by the matrix $(B-I)V$, it may be observed here that φ drops out in the product $v^\mathsf{T}u$ since v is the left eigenvector of B, and so $v^\mathsf{T}(B-I) = 0$. We show now that the terms

$$V^{-1}(u - (B-I)V\varphi) = \left(\binom{s}{i-1}\right)_{i=1}^s - s\gamma Pe_s$$

are elements of column $s+1$ in (24) if this explicit representation of $\tilde{B}$ is extended to size $(s+1)\times(s+1)$; see also (17), (18). For all these matrices we denote the extended versions by bold face and recall that the main diagonal of the upper triangular matrix $\tilde{\mathbf{B}} - \mathbf{I}$ is -1 below the first row. We obtain

$$\begin{aligned} \begin{pmatrix} V^{-1}u \\ -1 \end{pmatrix} &= \begin{pmatrix} \left(\binom{s}{i-1}\right)_{i=1}^s - \gamma s Pe_s + (\tilde{B}-I)\varphi \\ -1 \end{pmatrix} \\ &= (\mathbf{P} - \mathbf{I} - \gamma\mathbf{D}\mathbf{F}_\mathbf{0}^\mathsf{T}\mathbf{P})e_{s+1} + \begin{pmatrix} (\tilde{B}-I)\varphi \\ -1 \end{pmatrix}, \end{aligned}$$

where sPe_s contains the nontrivial elements of the last column of $\mathbf{D}\mathbf{F}_\mathbf{0}^\mathsf{T}\mathbf{P}$; see (18), (19). Comparing with (24) for dimension $s+1$ and $\sigma = 1$, i.e.,

$$\mathbf{P} - \mathbf{I} - \gamma\mathbf{D}\mathbf{F}_\mathbf{0}^\mathsf{T}\mathbf{P} = \tilde{\mathbf{B}} - \mathbf{e}_1\mathbf{e}_1^\mathsf{T},$$

yields

$$\begin{pmatrix} V^{-1}u \\ -1 \end{pmatrix} = \begin{pmatrix} (\tilde{B}-I)\varphi \\ 0 \end{pmatrix} + (\tilde{\mathbf{B}}-\mathbf{I})\mathbf{e}_{s+1} = (\tilde{\mathbf{B}}-\mathbf{I})\begin{pmatrix} \varphi \\ 1 \end{pmatrix}. \tag{35}$$

Due to the triangular shape of $\tilde{\mathbf{B}}$, its left eigenvector to the eigenvalue 1 has the form $(v^{\mathsf{T}}V, \tilde{v}_{s+1})$ and from (35) follows

$$v^{\mathsf{T}}u - \tilde{v}_{s+1} = (v^{\mathsf{T}}V, \tilde{v}_{s+1})\begin{pmatrix} V^{-1}u \\ -1 \end{pmatrix} = (v^{\mathsf{T}}V, \tilde{v}_{s+1})(\tilde{\mathbf{B}}-\mathbf{I})\begin{pmatrix} \varphi \\ 1 \end{pmatrix} = 0.$$

Recalling that u (see (32)) contains the error constants in (31), the statement follows. ∎

Remark. The theorem shows that the leading error constant of the scheme is a multiple of the component $\tilde{v}_{s+1}$, which itself is a polynomial in the parameter γ. So there are a few exceptional values for γ where the method has global order s for any set of off-step knots c_i. In Table 1 we present explicit formulas for $\tilde{v}_{s+1}(\gamma)$ and some of the parameters $\hat{\gamma}$ with superconvergence, $\tilde{v}_{s+1}(\hat{\gamma}) = 0$, that were used in specific PPSW-methods. The fourth column contains a numerical estimate of the angle of $A(\alpha)$ stability in degrees, and the last column contains the name of the corresponding method consisting of the number of stages and a letter labeling the zero of $\tilde{v}_{s+1}(\gamma)$ ("a" is the leftmost zero). A special name was given only to those methods used in the numerical tests in section 4.

TABLE 1
Superconvergence conditions and name of methods.

s	$\tilde{v}_{s+1}$	$\hat{\gamma}$	α	Method
2	$3-6\gamma+2\gamma^2$	$\frac{1}{2}(3\pm\sqrt{3})$	90	
3	$13-39\gamma+30\gamma^2-6\gamma^3$	1.32088	90	
4	$75-300\gamma+372\gamma^2-168\gamma^3+24\gamma^4$	0.468147	52.4	4a
		0.912763	89.9	4b
5	$541-2705\gamma+4660\gamma^2-3420\gamma^3+1080\gamma^4$ $-120\gamma^5$	0.722499	79.6	5b
6	$4683-28098\gamma+62130\gamma^2-64200\gamma^3$	0.619411	57.5	6b
	$+32760\gamma^4-7920\gamma^5+720\gamma^6$	1.087080	87.9	6c
7	$47293-331051\gamma+894810\gamma^2-201410\gamma^3$	0.557132	23.9	7 b
	$+864360\gamma^4-335160\gamma^5+65520\gamma^6-5040\gamma^7$	0.885444	81.3	7c
8	$545835-4366680\gamma+13959176\gamma^2$ $-23146032\gamma^3+21724080\gamma^4-11847360\gamma^5$ $+3689280\gamma^6-604800\gamma^7+40320\gamma^8$	0.758671	67.2	8c

The superconvergence property can easily be observed in numerical computations with constant stepsizes. For variable stepsizes a similar effect might still be obtained by clever strategies for a step-dependent choice of the parameter γ. However, for higher-order methods such strategies are probably too complicated to be of practical use. Still, we expect that superconvergence for $\sigma = 1$ has some beneficial effect on the general performance, especially for sharp tolerances where stepsize ratios are clustered around one.

4. Implementation and numerical tests. For practical implementations the form (28) of the PPSW-methods is preferable. Since the coefficients of the method depend on the actual stepsize ratio, they have to be recomputed before (28) can be solved. However, since the expense of these computations is $O(s^3)$ only, it is negligible

compared to the solution of the stage equations for large dimensions n. We also note that in case of a step rejection it is possible to reuse the function evaluations $f(Y_{m-1})$ and the matrix T_m, which is a difference approximation to $f_y(t_{m-1,s}, Y_{m-1,s})$.

Estimates for the local error are required for stepsize selection procedures. As mentioned before, (28) may be interpreted as a corrector equation and its solution $K_m := Y_m - (\Theta_m \otimes I)Y_{m-1}$ is an obvious candidate for an error estimate. Yet, it is of the same order $s-1$ as the error in Y_m that it should estimate. However, this situation can be improved by requiring order s of consistency (30) for one single stage only—for instance, $i = s$. Then, K_{ms} becomes an asymptotically correct order-s estimate for the local error of Y_{ms}. Unfortunately, only for low-order methods ($s \leq 4$) is it possible to find methods satisfying this additional condition and minimal stability requirements, i.e., A(0)-stability for $\sigma = 1$. So, only Method 4a uses this error estimate with the superconvergence value $\gamma = 0.468147$ and knots c_i, yielding (30) for $i = s$ and $\sigma = 1$. In order to obtain a sound error estimate in higher-order methods ($s \geq 5$) too, we compare the latest approximate Y_{ms} with the polynomial predictor of order $s-2$ that ignores $Y_{m-1,1}$ and interpolates the values $Y_{m-1,i}$, $i = 2, \ldots, s$, only.

With the exception of Method 4a, which has a superconsistent stage, the knots are nearly equidistributed in $[-1, 1]$ and $c_s = 1$ was always used. For the sake of completeness we show them in Table 2.

TABLE 2
Knots c_i of specific methods.

Method	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
4a	−1	−0.6779303684	0.5	1				
4b	−1	−0.3	0.7	1				
5b	−1	−0.5	0	0.5	1			
6b, c	−1	−0.5	0	0.4	0.8	1		
7b, c	−1	−0.7	−0.2	0	0.2	0.6	1	
8c	−1	−0.7	−0.3	0	0.3	0.5	0.7	1

Based on the error estimates described above, the new stepsize was computed in a standard manner; cf. [16]. The maximally allowed increase of the stepsize was set to $\sigma_{max} = 1.5$ for all PPSW-methods. Presently the stage equations (28) are solved by LU decomposition. Since this decomposition is the most expensive part and a severe sequential bottleneck for large systems, a remarkable parallel speed-up with PPSW-methods is not likely to be achieved yet. From our experience with PTSW-methods we expect superior parallel performance for large problems only in combination with Krylov solvers for the s independent linear stage systems (see [16, 17]). The first results for Krylov implementations of PPSW methods are reported in the article [20]. We decided to present tests here with sequential computations only, but we compare these with the state-of-the-art linearly implicit one-step code RODAS [10] with the default coefficient set (iwork(2)=1). RODAS was also used to compute the starting values Y_1 for the PPSW-methods. We think that the benchmark with RODAS gives a good insight into basic properties and the potential of these methods and assists in choosing suitable ones. We used Delphi 5 for comfortable programming. Since the reference results have been computed with the original Fortran-code RODAS, we compare the number of steps instead of the computing time. Note that the computational amount of work per step of RODAS is comparable to that of a six-stage PPSW-method in sequential runs. An advantage of our W-type methods over RODAS is that the Jacobian could be kept constant for several steps, but we did not exploit this possibility here.

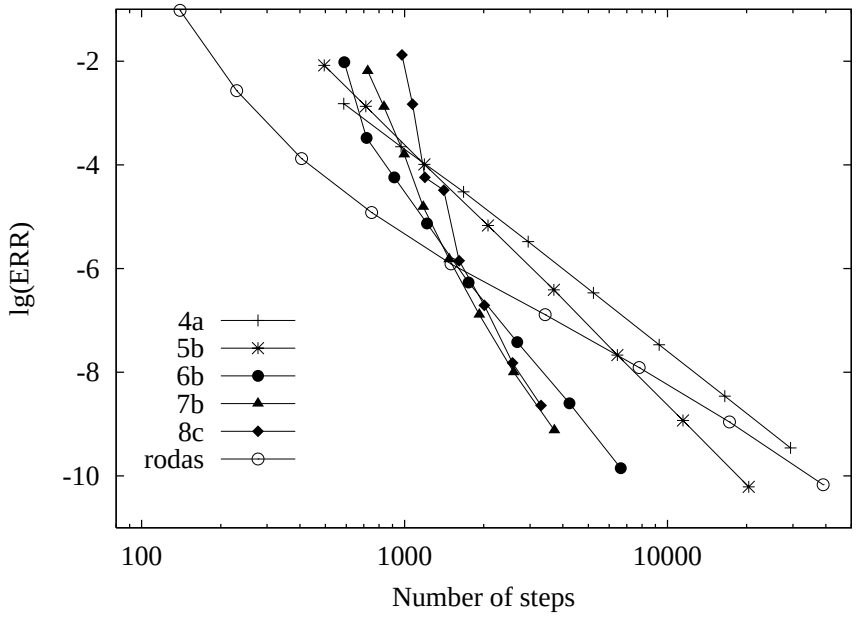


FIG. 1. *Results for OREGO.*

The following stiff test problems from [10] were used in our numerical tests: OREGO with $t_e = 360$, ROBER with $t_e = 10^8$, VDPOL with $t_e = 11$, and PLATE. A problem with a strongly varying Jacobian is KREISS from [14], defined by

$$y' = R(t)\Lambda(\varepsilon)R^{-1}(t), \quad y(0) = \begin{pmatrix} 1 \\ 2.6 \end{pmatrix}, \quad 0 \le t \le 1,$$

$$R(t) = \begin{pmatrix} \cos(-\theta t) & \sin(-\theta t) \\ -\sin(-\theta t) & \cos(-\theta t) \end{pmatrix}, \quad \Lambda(\varepsilon) = \begin{pmatrix} -\frac{1}{\varepsilon} & 0 \\ 0 & -1 \end{pmatrix}.$$

We used $\varepsilon = 10^{-6}$ and $\theta = 1$ here. A reference solution for these problems was computed with RADAU5 [10] and high accuracy. Computations were performed for $atol = rtol = 10^{-2}, \ldots, 10^{-9}$ except for ROBER, where $atol = 10^{-6} rtol$ was set (cf. [10]). In Figures 1–5 we present results for some special methods from Table 1. The logarithm of the final error at the endpoint for Y_{ms} is shown versus the number of steps including the rejected ones. Although the work per step of the methods (7) is one LU decomposition and s function evaluations and back substitutions, we think that this presentation is appropriate for two reasons. First, for larger problems the LU decomposition dominates the computational effort. However, more importantly, we recall that these methods were designed for parallel implementation with s processors.

A general observation for most PPSW-methods is that both the error and the angle α of $L(\alpha)$-stability increase with the parameter γ. So for all examples except PLATE we present the results for the methods with smallest γ from Table 1 since these produced slightly better results. The problem PLATE, however, requires a stability angle $\alpha \ge 71^\circ$ (cf. [10]). Some of the previously mentioned methods had difficulties here, namely, Method 4a for all tolerances and Method 6b for some of

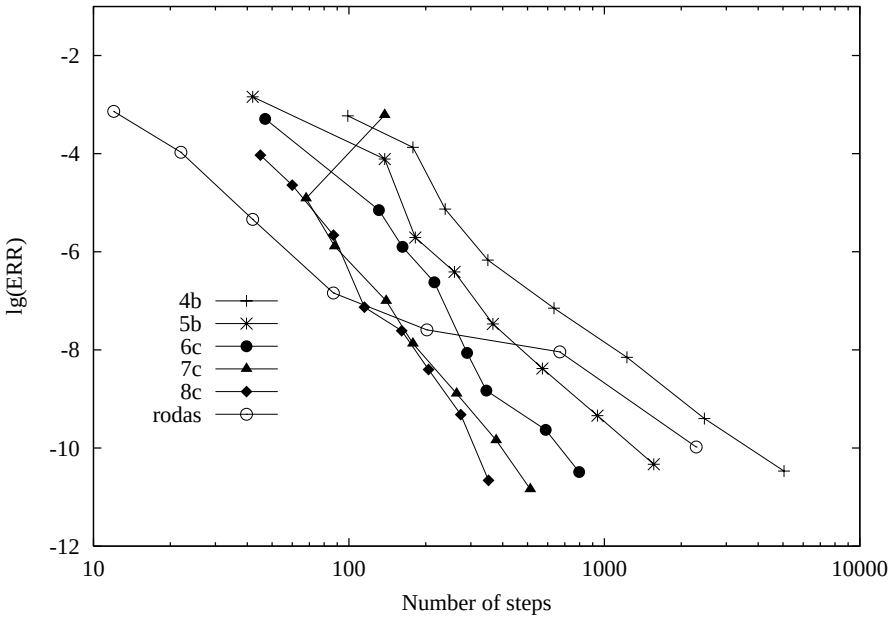


FIG. 2. *Results for PLATE.*

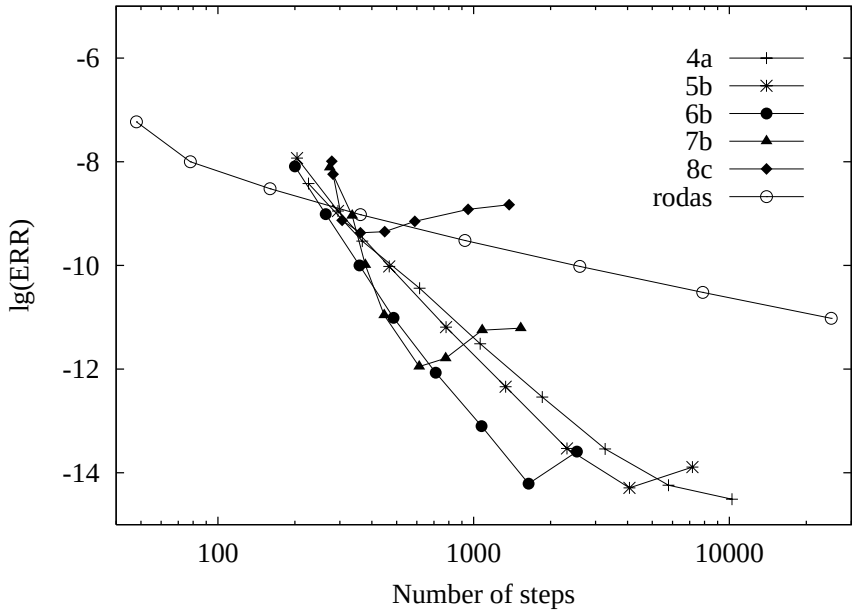


FIG. 3. *Results for ROBER.*

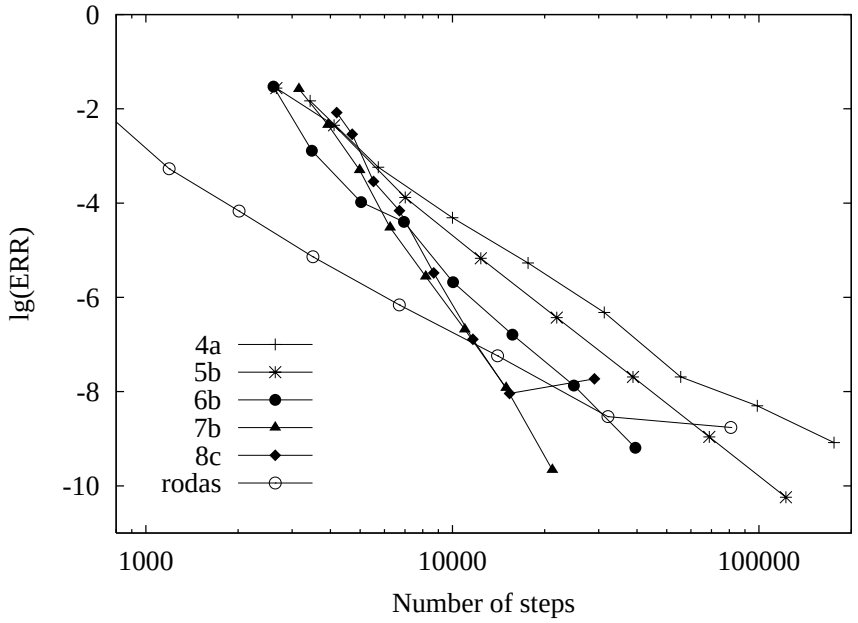


FIG. 4. *Results for VDPOL.*

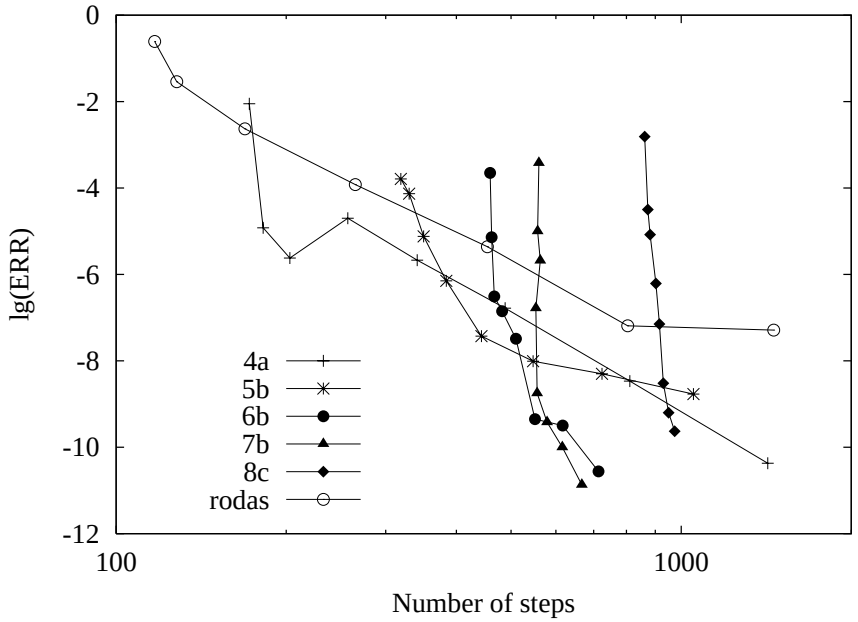


FIG. 5. *Results for the Kreiss problem.*

them. By choosing methods from Table 1 with larger γ and having larger stability angles, suitable PPSW-methods for this problem could be identified, too. Surprisingly, Methods 7b and 8c solved this problem without difficulties for all tolerances.

From these figures it is seen that the presented methods are interesting and have a large potential for stiff equations. The smooth curves show that the stepsize control works reliably but may still be improved. The irregular behavior of the seven- and eight-stage methods in the problem ROBER is an exception that is probably caused by rounding errors due to large values in the coefficients. This problem has been observed with other methods [4] too, and we tried to ameliorate it by distributing the knots c_i over the interval $[-1, 1]$ instead of $[0, 1]$. In comparison to PPSW-methods, RODAS is clearly superior for weak tolerances. This is mainly due to the ability of RODAS to increase the stepsize faster than our methods. Another reason may be that the initial stepsize for the PPSW-methods were too small. However, for medium and stringent tolerances especially, the higher-order PPSW-methods are often superior to RODAS. This is also the case for the six-stage methods whose computational expense is comparable to that of RODAS (in sequential implementations). The choice of the knots c_i is still heuristic, and their influence on accuracy and robustness of the methods needs further research.

5. Conclusions. A new class of linearly implicit methods for stiff initial value problems has been presented with a nearly optimal potential of method parallelism. Its essential new feature is the use of s peer solution variables per time interval. Methods of order $s - 1$ with s stages have been derived possessing good stability properties. The global error improves to order s for special parameter values and constant stepsizes since a certain superconvergence property has been identified. In a sequential implementation using automatic stepsize control and LU-decompositions, several methods with up to $s = 8$ stages have been tested and were found to be competitive with the code RODAS. The full potential of parallelization for PPSW-methods in the solution of large stiff ODEs is expected only in combination with Krylov techniques. Here the use of multi-implicit methods in the general form (3) using different parameters γ_i offers additional options and will be investigated in [18].

REFERENCES

[1] C. Bendtsen, *A parallel stiff ODE solver based on MIRKs*, Adv. Comput. Math., 7 (1997), pp. 27–36.

[2] K. Burrage, *Parallel and Sequential Methods for Ordinary Differential Equations*, Clarendon Press, Oxford, UK, 1995.

[3] K. Burrage and H. Suhartanto, *Parallel iterated methods based on variable step-size multistep Runge-Kutta methods of Radau type for stiff problems*, Adv. Comput. Math., 13 (2000), pp. 257–270.

[4] K. Burrage and T. Tian, *Parallel half-block methods for initial value problems*, Appl. Numer. Math., 32 (2000), pp. 255–271.

[5] J. C. Butcher, *Diagonally-implicit multi-stage integration methods*, Appl. Numer. Math., 11 (1993), pp. 347–363.

[6] J. C. Butcher and Z. Jackiewicz, *Implementation of diagonally implicit multistage integration methods for ordinary differential equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2119–2141.

[7] J. C. Butcher and A. D. Singh, *The choice of parameters in parallel general linear methods for stiff problems*, Appl. Numer. Math., 34 (2000), pp. 59–84.

[8] C. W. Gear and X. Xuhai, *Parallelism across time in ODEs*, Appl. Numer. Math., 11 (1993), pp. 45–68.

[9] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I, 2nd ed., Springer-Verlag, New York, 1993.

[10] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential–Algebraic Problems*, Springer-Verlag, Berlin, Heidelberg, New York 1996.
[11] P. J. van der Houwen and B. P. Sommeijer, *CWI contributions to the development of parallel Runge-Kutta methods*, Appl. Numer. Math., 22 (1996), pp. 327–344.
[12] P. J. van der Houwen and J. J. B. de Swart, *Parallel linear system solvers for Runge-Kutta methods*, Adv. Comput. Math., 7 (1997), pp. 157–181.
[13] P. Kaps and P. Rentrop, *Generalized Runge-Kutta methods of order four with stepsize control for stiff ordinary differential equations*, Numer. Math., 33 (1979), pp. 55–68.
[14] H.-O. Kreiss, *Difference methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 15 (1978), pp. 21–58.
[15] H. Podhaisky, B. A. Schmitt, and R. Weiner, *Two-step W-methods with parallel stages*, Tech. report 22, Universität Halle, Halle, Germany, 1999, http://www.mathematik.uni-halle.de/reports/sources/1999/99-22report.ps.
[16] H. Podhaisky, B. A. Schmitt, and R. Weiner, *Design, analysis and testing of some parallel two–step W–methods for stiff systems*, Appl. Numer. Math., 42 (2002), pp. 381–395.
[17] H. Podhaisky, *Parallele Zweischritt–W–Methoden*, Ph.D. thesis, Martin-Luther-Universität, Halle-Wittenberg, Germany, 2002.
[18] B. A. Schmitt, R. Weiner, and H. Podhaisky, *Multi-implicit Peer Two-Step W-Methods for Parallel Time Integration*, manuscript.
[19] R. Weiner, B. A. Schmitt, and H. Podhaisky, *Two–step W–methods on singular perturbation problems*, in PPAM 2001, Lecture Notes in Comput. Sci. 2328, R. Wyrzykowski et al., eds., Spinger-Verlag, New York, 2002, pp. 778–785.
[20] R. Weiner, B. A. Schmitt, and H. Podhaisky, *Two-step W-methods and their application to MOL-systems*, Tech. Report 02, Universität Halle, Halle, Germany, 2003, http://www.mathematik.uni-halle.de/reports/sources/2003/03-02report.ps.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 283–301

A CHARACTERIZATION OF HYBRIDIZED MIXED METHODS FOR SECOND ORDER ELLIPTIC PROBLEMS*

BERNARDO COCKBURN† AND JAYADEEP GOPALAKRISHNAN‡

This paper is dedicated to Jim Douglas, Jr., on the occasion of his 75th birthday

Abstract. In this paper, we give a new characterization of the approximate solution given by hybridized mixed methods for second order self-adjoint elliptic problems. We apply this characterization to obtain an explicit formula for the entries of the matrix equation for the Lagrange multiplier unknowns resulting from hybridization. We also obtain necessary and sufficient conditions under which the multipliers of the Raviart–Thomas and the Brezzi–Douglas–Marini methods of similar order are identical.

Key words. mixed finite elements, hybrid methods, elliptic problems

AMS subject classifications. Primary, 54C40, 14E20; Secondary, 46E25, 20C20

DOI. 10.1137/S0036142902417893

1. Introduction.

In this paper, we give a new characterization of hybridized mixed methods. This characterization allows us to obtain an explicit formula for the entries of the matrix equation for the so-called Lagrange multipliers. It also allows comparison of hybridized versions of different mixed methods. For example, we give conditions under which the multipliers of the Raviart–Thomas (RT) method and those of the Brezzi–Douglas–Marini (BDM) method of comparable order coincide.

We consider the hybridized version [1] of the standard RT mixed method [13] for the elliptic boundary value problem

$$-\nabla \cdot (a\nabla u) + d\, u = f \qquad \text{in } \Omega \subset \mathbb{R}^2, \tag{1.1}$$

$$u = g \qquad \text{on } \partial\Omega, \tag{1.2}$$

where $a(\boldsymbol{x})$ is a symmetric positive definite matrix-valued function, $d(\boldsymbol{x})$ is a nonnegative function, and Ω is a polygonal domain in $\mathbb{R}^2$. We assume that $a(\boldsymbol{x})$ and $d(\boldsymbol{x})$ are bounded. We consider this a simple setting for transparent presentation of the main ideas. As will be clear later, our techniques can be applied to other hybridized methods and more general second order elliptic problems.

Before describing the results, recall that mixed finite element methods seek approximations $(\boldsymbol{q}_h, u_h)$ to $(-a\,\nabla u, u)$ in appropriate finite element spaces. They give rise to a matrix equation of the form

$$\begin{pmatrix} \mathbb{A} & -\mathbb{B}^t \\ \mathbb{B} & \mathbb{D} \end{pmatrix} \begin{pmatrix} \mathbb{Q} \\ \mathbb{U} \end{pmatrix} = \begin{pmatrix} \mathbb{G} \\ \mathbb{F} \end{pmatrix},$$

where $\mathbb{Q}$ and $\mathbb{U}$ are the vectors of coefficients of $\boldsymbol{q}_h$ and u_h with respect to their corresponding finite element basis, respectively. Since the system is not positive definite,

*Received by the editors November 13, 2002; accepted for publication (in revised form) July 28, 2003; published electronically January 28, 2004.
http://www.siam.org/journals/sinum/42-1/41789.html

†School of Mathematics, University of Minnesota, 206 Church Street S.E., Minneapolis, MN 55455 (cockburn@math.umn.edu). The research of this author was partially supported by the National Science Foundation (grant DMS-0107609) and by the University of Minnesota Supercomputer Institute.

‡Department of Mathematics, University of Florida, Gainesville, FL 32611–8105 (jayg@math.ufl.edu).

solving for $\mathbb{Q}$ and $\mathbb{U}$ is not always easy. Although one can arrive at a positive definite system by elimination of $\mathbb{Q}$ from the equations, this requires inverting $\mathbb{A}$ and maintaining $\mathbb{A}^{-1}$, which is typically a full matrix. Fortunately, by "hybridizing" the mixed method, this difficulty can be overcome. Let us now briefly recall the hybridization procedure.

First, the so-called Lagrange multiplier λ_h is introduced. This gives rise to a matrix equation of the form

$$\begin{pmatrix} A & -B^t & -C^t \\ B & D & 0 \\ C & 0 & 0 \end{pmatrix} \begin{pmatrix} Q \\ U \\ \Lambda \end{pmatrix} = \begin{pmatrix} \mathcal{G} \\ \mathcal{F} \\ 0 \end{pmatrix}, \tag{1.3}$$

where Λ is the vector of degrees of freedom associated to the multiplier λ_h. We will precisely state the underlying finite element spaces later. As is now well known, the new vectors of degrees of freedom Q and U actually define the *same* approximation $(\boldsymbol{q}_h, u_h)$ as the original mixed method. Moreover, both Q and U can now be easily eliminated to obtain an equation for the multiplier only, namely,

$$\mathbb{E}\,\Lambda = \mathbb{H},$$

where $\mathbb{E}$ and $\mathbb{H}$ are given by

$$\begin{aligned} \mathbb{E} &= CA^{-1}\left(A - B^t(BA^{-1}B^t + D)^{-1}B\right)A^{-1}C^t, \\ \mathbb{H} &= \mathbb{H}_g + \mathbb{H}_f, \\ \mathbb{H}_g &= -CA^{-1}\left(A - B^t(BA^{-1}B^t + D)^{-1}B\right)A^{-1}\mathcal{G}, \\ \mathbb{H}_f &= -CA^{-1}B^t(BA^{-1}B^t + D)^{-1}\,\mathcal{F}. \end{aligned} \tag{1.4}$$

That the inverses taken above exist follows from the properties of the underlying finite element spaces. Considering this matrix equation instead of the previous one has several advantages: (i) the matrix $\mathbb{E}$ is symmetric and positive definite, so it can be numerically inverted by using methods like the conjugate gradient method; (ii) the number of degrees of freedom of the multiplier is remarkably smaller than the number of degrees of freedom of the original mixed method; (iii) once Λ has been obtained, both Q and U can be efficiently computed element by element; and (iv) the multiplier λ_h can actually be used to *improve* the approximation to u by means of a local postprocessing, as shown in [1]. This shows that the use of hybridized mixed methods is indeed very advantageous; however, the complicated relation between the matrices $\mathbb{E}$ and $\mathbb{H}$, and the matrices A, B, C, $\mathcal{F}$ and $\mathcal{G}$, can easily dissuade one from basing an implementation on $\mathbb{E}$ and $\mathbb{H}$.

In this paper, we show that the entries of the matrices $\mathbb{E}$ and $\mathbb{H}$ can be expressed as a weighted L^2-inner product of some *discontinuous* auxiliary functions, the weights being nothing but the matrix a^{-1} and the function d. These auxiliary functions are easily constructed in terms of the geometry of the mesh, the matrix a, the function d, and the spaces of the hybridized mixed finite element method. Their definition induces a natural decomposition of the approximate solution $(\boldsymbol{q}_h, u_h)$ of the form

$$(\boldsymbol{q}_h, u_h) = (\boldsymbol{q}_h, u_h)_{\lambda_h} + (\boldsymbol{q}_h, u_h)_g + (\boldsymbol{q}_h, u_h)_f,$$

where $(\boldsymbol{q}_h, u_h)_{\lambda_h}$ is a *lifting* of the Lagrange multiplier λ_h and $(\boldsymbol{q}_h, u_h)_g$ and $(\boldsymbol{q}_h, u_h)_f$ can be computed *locally* only in terms of the data. The introduction of other discrete

lifting operators has proved useful in another context earlier, namely, the analysis of discontinuous Galerkin methods for elliptic problems [2, 7, 8].

We then present two applications of this result. As a first application, we present a technique to assemble the matrix of the Lagrange multiplier equation using simple local element matrices. Next, we compare the matrices $\mathbb{E}$ and $\mathbb{H}$ of the RT method with those of the BDM method of similar order and give necessary and sufficient conditions for the multipliers to be *exactly* the same. This happens, for example, when $d = 0$ and $f = 0$, a case that occurs in many situations of practical interest.

The paper is organized as follows. In section 2, we introduce the hybridized version of the mixed method of Raviart and Thomas and then state, discuss, and prove the characterization result, Theorem 2.1. In section 3, we show how to assemble the matrices for the multipliers, and in section 4, we compare the matrices of the RT method with those of the BDM method of similar order. Finally, in section 5, we end with some concluding remarks.

2. The main result. We begin this section by introducing the classical mixed method of Raviart and Thomas [13]; then, following [1], we hybridize the method. Finally, we state, discuss, and prove the main result, Theorem 2.1.

2.1. The hybridized mixed method. Given a triangulation of Ω, $\mathcal{T}_h$, made of triangles, the mixed method seeks an approximation $(\boldsymbol{q}_h, u_h)$ to the solution $(\boldsymbol{q}, u)$ of the model problem

$$c\,\boldsymbol{q} = -\nabla u \qquad \text{in } \Omega, \tag{2.1}$$

$$\nabla \cdot \boldsymbol{q} + d\,u = f \qquad \text{in } \Omega, \tag{2.2}$$

$$u = g \qquad \text{on } \partial\Omega, \tag{2.3}$$

where $c = a^{-1}$. The approximation $(\boldsymbol{q}_h, u_h)$ is sought in the finite element space $\mathbb{V}_h \times \mathbb{W}_h$ given by

$$\mathbb{V}_h = \{\boldsymbol{v} \in H(\mathrm{div}, \Omega) : \quad \boldsymbol{v}|_K \in P^k(K) \times P^k(K) + \boldsymbol{x}\,P^k(K) \quad \text{for all } K \in \mathcal{T}_h\},$$

$$\mathbb{W}_h = \{w \in L^2(\Omega) : \quad w|_K \in P^k(K) \quad \text{for all } K \in \mathcal{T}_h\},$$

where $P^k(K)$ denotes the space of polynomials on K of degree at most k, $k \geq 0$, and is defined by requiring that, for all $(\boldsymbol{v}, w) \in V_h \times W_h$,

$$\int_\Omega c\,\boldsymbol{q}_h \cdot \boldsymbol{v}\,dx - \int_\Omega u_h \nabla \cdot \boldsymbol{v}\,dx = -\int_{\partial\Omega} g\,\boldsymbol{v} \cdot \boldsymbol{n}\,ds, \tag{2.4}$$

$$\int_\Omega w \nabla \cdot \boldsymbol{q}_h\,dx + \int_\Omega d\,u_h\,w\,dx = \int_\Omega f\,w\,dx. \tag{2.5}$$

It is easy to see that the above weak formulation gives rise to a system of equations of the form

$$\begin{pmatrix} \mathbb{A} & -\mathbb{B}^t \\ \mathbb{B} & \mathbb{D} \end{pmatrix} \begin{pmatrix} \mathbb{Q} \\ \mathbb{U} \end{pmatrix} = \begin{pmatrix} \mathbb{G} \\ \mathbb{F} \end{pmatrix}.$$

We can try to solve this equation by first eliminating $\mathbb{Q}$ from the equations and then solving the resulting equation for $\mathbb{U}$, namely,

$$\left(\mathbb{B}\,\mathbb{A}^{-1}\mathbb{B}^t + \mathbb{D}\right)\,\mathbb{U} = \mathbb{F} + \mathbb{B}\,\mathbb{A}^{-1}\mathbb{G}.$$

Unfortunately, the matrix $\mathbb{A}$ is not easy to invert since the elements of V_h, being functions in $H(\mathrm{div},\Omega)$, have their normal components continuous across element interfaces. If $\boldsymbol{q}_h$ were totally discontinuous, $\mathbb{A}$ would be block-diagonal and hence easily invertible. The idea of the hybridized mixed methods is to relax this continuity constraint to render $\mathbb{A}$ block-diagonal.

Indeed, it was Fraejis de Veubeque [10], back in 1965, who realized that this can be achieved by introducing additional unknowns λ_h, associated to element interfaces, called Lagrange multipliers. Let $\mathcal{E}_{i,h}$ denote the set of edges of the mesh $\mathcal{T}_h$ that are in the interior of the domain Ω. The multipliers are then nothing but approximations to the trace of u on each $e \in \mathcal{E}_{i,h}$. As we show next, their introduction allows elimination of *both* $\boldsymbol{q}_h$ and u_h and reduction of the system to a single matrix equation for the multipliers.

In our particular case, the hybridized mixed method seeks an approximation $(\boldsymbol{q}_h, u_h, \lambda_h)$ to $(\boldsymbol{q}, u, u|_{\mathcal{E}_{i,h}})$ in the finite element space $V_h \times W_h \times M_h$ given by

$$
\begin{aligned}
V_h &= \{\boldsymbol{v} \in L^2(\Omega)\times L^2(\Omega): \quad \boldsymbol{v}|_K \in P^k(K)\times P^k(K) + \boldsymbol{x}P^k(K) \quad \text{for all } \ K \in \mathcal{T}_h\},\\
W_h &= \{w \in L^2(\Omega): \quad w|_K \in P^k(K) \quad \text{for all } \ K \in \mathcal{T}_h\},\\
M_h &= \{\mu \in L^2(\mathcal{E}_{i,h}): \quad \mu|_e \in P^k(e) \quad \text{for all } \ e \in \mathcal{E}_{i,h}\}.
\end{aligned}
$$

It is defined by requiring that, for all $(\boldsymbol{v}, u, \mu) \in V_h \times W_h \times M_h$,

$$
\int_\Omega c\,\boldsymbol{q}_h\cdot\boldsymbol{v}\,dx - \sum_{K\in\mathcal{T}_h}\int_K u_h\nabla\cdot\boldsymbol{v}\,dx + \sum_{e\in\mathcal{E}_{i,h}}\int_e \lambda_h\,[\![\boldsymbol{v}]\!]\,ds = -\int_{\partial\Omega} g\,[\![\boldsymbol{v}]\!]\,ds, \tag{2.6}
$$

$$
\sum_{K\in\mathcal{T}_h}\int_K w\nabla\cdot\boldsymbol{q}_h\,dx + \int_\Omega d\,u_h\,w\,dx = \int_\Omega f\,w\,dx, \tag{2.7}
$$

$$
\sum_{e\in\mathcal{E}_{i,h}}\int_e \mu\,[\![\boldsymbol{q}_h]\!]\,ds = 0, \tag{2.8}
$$

where $[\![\boldsymbol{v}]\!] = \boldsymbol{v}\cdot\boldsymbol{n}$ on $\partial\Omega$ and $[\![\boldsymbol{v}]\!] = \boldsymbol{v}_e^+\cdot\boldsymbol{n}_e^+ + \boldsymbol{v}_e^-\cdot\boldsymbol{n}_e^-$ on $e \in \mathcal{E}_h$. Here, $\boldsymbol{n}$ denotes the outward unit normal to Ω, $\boldsymbol{n}_e^+ = -\boldsymbol{n}_e^-$ is an arbitrary unit vector normal to the $e \in \mathcal{E}_{i,h}$, and $\boldsymbol{v}_e^\pm(\boldsymbol{x}) = \lim_{\epsilon\downarrow 0}\boldsymbol{v}(\boldsymbol{x} - \epsilon\,\boldsymbol{n}_e^\pm)$.

2.2. Two local mappings. Next, we introduce two mappings in terms of which the characterization result will be expressed. They are defined using (2.6) and (2.7).

The first mapping lifts functions on edges of the triangulation $\mathcal{T}_h$ to functions on Ω. Let $\mathcal{E}_h$ be the set of all edges of the triangulation $\mathcal{T}_h$. Notwithstanding a slight abuse of notation, we shall denote the set of all square integrable functions on the union of all edges of $\mathcal{E}_h$ by $L^2(\mathcal{E}_h)$. The lifting associates to each $\mathsf{m} \in L^2(\mathcal{E}_h)$ the pair of functions $(\boldsymbol{q}_h, u_h)_{\mathsf{m}} \equiv (\boldsymbol{q}_{h,\mathsf{m}}, u_{h,\mathsf{m}}) \in V_h \times W_h$ defined by requiring that

$$
\int_\Omega c\,\boldsymbol{q}_{h,\mathsf{m}}\cdot\boldsymbol{v}\,dx - \sum_{K\in\mathcal{T}_h}\int_K u_{h,\mathsf{m}}\nabla\cdot\boldsymbol{v}\,dx = -\sum_{e\in\mathcal{E}_h}\int_e \mathsf{m}\,[\![\boldsymbol{v}]\!]\,ds, \tag{2.9}
$$

$$
\sum_{K\in\mathcal{T}_h}\int_K w\nabla\cdot\boldsymbol{q}_{h,\mathsf{m}}\,dx + \int_\Omega d\,u_{h,\mathsf{m}}\,w\,dx = 0 \tag{2.10}
$$

hold for all $(\boldsymbol{v}, w) \in V_h \times W_h$.

The second mapping associates to the function $f \in L^2(\Omega)$ the element $(\boldsymbol{q}_h, u_h)_f \equiv (\boldsymbol{q}_{h,f}, u_{h,f}) \in V_h \times W_h$ and is defined by requiring that

$$\int_\Omega c\,\boldsymbol{q}_{h,f} \cdot \boldsymbol{v}\,dx - \sum_{K\in\mathcal{T}_h} \int_K u_{h,f} \nabla\cdot \boldsymbol{v}\,dx = 0, \tag{2.11}$$

$$\sum_{K\in\mathcal{T}_h} \int_K w \nabla\cdot \boldsymbol{q}_{h,f}\,dx + \int_\Omega d\,u_{h,f}\,w\,dx = \int_\Omega f\,w\,dx \tag{2.12}$$

hold for all $(\boldsymbol{v}, w) \in V_h \times W_h$.

Note that these mappings can be computed in an element-by-element fashion. Indeed, they are uniquely defined on each element because of the surjectivity of the map $(\nabla\cdot) : V_h \mapsto W_h$ restricted to an element. Moreover, on each element $K \in \mathcal{T}_h$, the lifting $(\boldsymbol{q}_h, u_h)_{\mathsf{m}}$ can be thought of as a result of a one element discretization of the boundary value problem

$$\begin{aligned} c\,\boldsymbol{q}_{\mathsf{m}} &= -\nabla u_{\mathsf{m}} && \text{in } K,\\ \nabla\cdot\boldsymbol{q}_{\mathsf{m}} + d\,u_{\mathsf{m}} &= 0 && \text{in } K,\\ u_{\mathsf{m}} &= \mathsf{m} && \text{on } \partial K, \end{aligned}$$

and that the mapping $(\boldsymbol{q}_h, u_h)_f$ is an approximation to the solution of

$$\begin{aligned} c\,\boldsymbol{q}_f &= -\nabla u_f && \text{in } K,\\ \nabla\cdot\boldsymbol{q}_f + d\,u_f &= f && \text{in } K,\\ u_f &= 0 && \text{on } \partial K. \end{aligned}$$

2.3. Characterization of the approximate solution. Before stating the result, let us introduce the following convention: The extension by zero of the function $\eta \in L^2(\mathcal{F}_h)$, where $\mathcal{F}_h$ is a subset of $\mathcal{E}_h$, to $\mathcal{E}_h$ is also denoted by η. In this way, if $\mathsf{m} = \lambda_h$ on $\mathcal{E}_{i,h}$ and $\mathsf{m} = g$ on $\partial\Omega$, we simply write $\mathsf{m} = \lambda_h + g$; as a consequence we also write

$$(\boldsymbol{q}_h, u_h)_{\mathsf{m}} = (\boldsymbol{q}_h, u_h)_{\lambda_h} + (\boldsymbol{q}_h, u_h)_g.$$

We now have all that is needed to state the main result.

THEOREM 2.1 (characterization of $(\boldsymbol{q}_h, u_h, \lambda_h)$). *Let $(\boldsymbol{q}_h, u_h, \lambda_h)$ be the solution of the hybridized RT method* (2.6), (2.7), *and* (2.8). *Then*

$$(\boldsymbol{q}_h, u_h) = (\boldsymbol{q}_h, u_h)_{\lambda_h} + (\boldsymbol{q}_h, u_h)_g + (\boldsymbol{q}_h, u_h)_f.$$

The Lagrange multiplier $\lambda_h \in M_h$ is the unique solution of

$$a_h(\lambda_h, \mu) = b_h(\mu) \quad \textit{for all } \mu \in M_h, \tag{2.13}$$

where

$$a_h(\lambda_h, \mu) = \int_\Omega c\,\boldsymbol{q}_{h,\lambda_h} \cdot \boldsymbol{q}_{h,\mu}\,dx + \int_\Omega d\,u_{h,\lambda_h}\,u_{h,\mu}\,dx$$

and

$$b_h(\mu) = \int_{\partial\Omega} g\,[\![\boldsymbol{q}_{h,\mu}]\!]\,ds + \int_\Omega f\,u_{h,\mu}\,dx.$$

Remark 2.1. Although the normal components of the functions $\boldsymbol{q}_{h,g}$, $\boldsymbol{q}_{h,f}$ (which can be computed locally only in terms of the data) and $\boldsymbol{q}_{h,\lambda_h}$ are not necessarily continuous across interelement boundaries, the normal components of their sum, namely, $\boldsymbol{q}_h$, are. Marini [12] pointed out this fact for the lowest order RT method when $d = 0$ and a is a piecewise-constant scalar function.

Remark 2.2. Just as for classical finite element methods, the variational formulation (2.13) gives rise to a matrix equation for the degrees of freedom of the multiplier Λ of the form

$$\mathbb{E}\,\Lambda = \mathbb{H}.$$

Thus the entries of the matrices $\mathbb{E}$ and $\mathbb{H}$ are obtained as a weighted L^2-inner product of discontinuous functions, as claimed in the introduction.

We end this section with a proof of Theorem 2.1.

2.4. Proof of Theorem 2.1. We prove this result in three steps. In the first, we observe some identities that result from the equations of the method. In the next step, we show that the continuity condition on the jumps of the fluxes results in a variational equation for the Lagrange multiplier unknowns. In the last step, we collect these two results and conclude.

2.4.1. Step 1.

LEMMA 2.2 (elementary identities). *We have, for any* $\mathsf{m}, \mu \in L^2(\mathcal{E}_h)$*, and* $f \in L^2(\Omega)$*,*

$$\text{(i)} \; -\sum_{e\in\mathcal{E}_h}\int_e \mu\,[\![\boldsymbol{q}_{h,\mathsf{m}}]\!]\,ds = \int_\Omega c\,\boldsymbol{q}_{h,\mathsf{m}}\cdot\boldsymbol{q}_{h,\mu}\,dx + \int_\Omega d\,u_{h,\mathsf{m}}\,u_{h,\mu}\,dx,$$

$$\text{(ii)} \; \int_\Omega c\,\boldsymbol{q}_{h,f}\cdot\boldsymbol{q}_{h,\mathsf{m}}\,dx + \int_\Omega d\,u_{h,f}\,u_{h,\mathsf{m}}\,dx = 0,$$

$$\text{(iii)} \; -\sum_{e\in\mathcal{E}_h}\int_e \mathsf{m}\,[\![\boldsymbol{q}_{h,f}]\!]\,ds = -\int_\Omega f\,u_{h,\mathsf{m}}\,dx.$$

Proof. Let us begin by proving the identity (i). First, take $\boldsymbol{v} = \boldsymbol{q}_{h,\mu}$ in (2.9). Then, replace m by μ in (2.10) and take $w = u_{h,\mathsf{m}}$. The identity (i) follows by simply adding these two equations.

To prove (ii), simply take $w = u_{h,f}$ in (2.10) and $\boldsymbol{v} = \boldsymbol{q}_{h,\mathsf{m}}$ in (2.11) and add the equations.

Finally, to prove (iii), take $\boldsymbol{v} = \boldsymbol{q}_{h,f}$ in (2.9) and $w = u_{h,\mathsf{m}}$ in (2.12) and add the two equations to obtain

$$-\sum_{e\in\mathcal{E}_h}\int_e \mathsf{m}\,[\![\boldsymbol{q}_{h,f}]\!]\,ds = -\int_\Omega f\,u_{h,\mathsf{m}}\,dx + \Theta,$$

where

$$\Theta = \int_\Omega c\,\boldsymbol{q}_{h,f}\cdot\boldsymbol{q}_{h,\mathsf{m}}\,dx + \int_\Omega d\,u_{h,f}\,u_{h,\mathsf{m}}\,dx.$$

Thus (iii) follows from (ii). This completes the proof. ∎

2.4.2. Step 2. In the following lemma, we explore equivalent characterizations of the continuity requirement on fluxes imposed by the method.

LEMMA 2.3 (jump condition). *Let* $(\boldsymbol{q}_h, u_h, \lambda_h)$ *be the solution of the hybridized RT method* (2.6), (2.7), *and* (2.8), *and let* m *be an arbitrary member of* M_h. *Then the following statements are equivalent:*

1. $\sum_{e\in\mathcal{E}_{i,h}} \int_e \mu\, [\![\boldsymbol{q}_{h,\mathsf{m}} + \boldsymbol{q}_{h,g} + \boldsymbol{q}_{h,f}]\!]\, ds = 0$ *for all* $\mu \in M_h$.
2. $(\boldsymbol{q}_h, u_h) = (\boldsymbol{q}_h, u_h)_{\mathsf{m}} + (\boldsymbol{q}_h, u_h)_g + (\boldsymbol{q}_h, u_h)_f$.
3. $\mathsf{m} = \lambda_h$.
4. $a_h(\mathsf{m}, \mu) = b_h(\mu)$ *for all* $\mu \in M_h$.

Proof. (1) $\Longrightarrow$ (2): Set

$$(\widetilde{\boldsymbol{q}}_h, \widetilde{u}_h) = (\boldsymbol{q}_h, u_h)_{\mathsf{m}} + (\boldsymbol{q}_h, u_h)_g + (\boldsymbol{q}_h, u_h)_f.$$

Then, combining the equations defining the local mappings, we get that

$$\int_\Omega c\, \widetilde{\boldsymbol{q}}_h \cdot \boldsymbol{v}\, dx - \sum_{K\in\mathcal{T}_h} \int_K \widetilde{u}_h \nabla\cdot \boldsymbol{v}\, dx + \sum_{e\in\mathcal{E}_{i,h}} \int_e \mathsf{m}\, [\![\boldsymbol{v}]\!]\, ds = -\int_{\partial\Omega} g\, [\![\boldsymbol{v}]\!]\, ds,$$

$$\sum_{K\in\mathcal{T}_h} \int_K w \nabla\cdot \widetilde{\boldsymbol{q}}_h\, dx + \int_\Omega d\, \widetilde{u}_h\, w\, dx = \int_\Omega f\, w\, dx$$

for all $(\boldsymbol{v}, w) \in V_k \times W_h$. Therefore, whenever (1) holds, $\widetilde{\boldsymbol{q}}_h$ and $\widetilde{u}_h$ satisfy all the equations of the hybridized RT method. By uniqueness of solutions of the method, $\widetilde{\boldsymbol{q}}_h = \boldsymbol{q}_h$ and $\widetilde{u}_h = u_h$, so (2) follows.

(2) $\Longrightarrow$ (3): By linear superposition, $\boldsymbol{q}_h = \boldsymbol{q}_{h,\lambda_h} + \boldsymbol{q}_{h,g} + \boldsymbol{q}_{h,f}$. Moreover, $\int_e \mu [\![\boldsymbol{q}_h]\!]\, ds = 0$ for all $e \in \mathcal{E}_{i,h}$. From the implication (1) $\Longrightarrow$ (2), it follows that

$$(\boldsymbol{q}_h, u_h) = (\boldsymbol{q}_h, u_h)_{\lambda_h} + (\boldsymbol{q}_h, u_h)_g + (\boldsymbol{q}_h, u_h)_f.$$

Consequently, (2) implies that

$$(\boldsymbol{q}_h, u_h)_{\mathsf{m}-\lambda_h} = 0,$$

from which it follows that $\mathsf{m} - \lambda_h = 0$.

(3) $\Longrightarrow$ (4): Now we observe that the following identity holds for any $\mathsf{m} \in M_h$:

$$\begin{aligned} -\int_\Omega c\, \boldsymbol{q}_{h,\mathsf{m}} \cdot \boldsymbol{q}_{h,\mu}\, dx - \int_\Omega d\, u_{h,\mathsf{m}}\, u_{h,\mu}\, dx + \int_{\partial\Omega} g\, [\![\boldsymbol{q}_{h,\mu}]\!]\, ds + \int_\Omega f\, u_{h,\mu}\, dx \\ = \sum_{e\in\mathcal{E}_h} \int_e \mu\, [\![\boldsymbol{q}_{h,\mathsf{m}} + \boldsymbol{q}_{h,g} + \boldsymbol{q}_{h,f}]\!]\, ds \quad \text{for all } \mu \in M_h. \end{aligned} \tag{2.14}$$

This equality follows from the identities of Lemma 2.2. Moreover, whenever $\mathsf{m} = \lambda_h$, the last equation of the hybridized mixed method asserts that the right-hand side of (2.14) is zero. From the definition of the forms a_h and b_h in Theorem (2.1), we see that (4) follows.

(4) $\Longrightarrow$ (1): We apply (2.14) again. Whenever (4) holds, the left-hand side of (2.14) is zero. Therefore, (1) follows. This completes the proof. ∎

2.4.3. Step 3. To conclude the proof of Theorem 2.1 observe that the first assertion of the theorem follows from the equivalence of the identities (1) and (2) of Lemma 2.3. The second assertion of the theorem follows again from Lemma 2.3, this time from the equivalence of (3) and (4). This completes the proof of Theorem 2.1.

Remark 2.3. The characterization theorem we just proved states that the solution $(\boldsymbol{q}_h, u_h)$ is the sum of the lifting of $\mathsf{m} = \lambda_h + g$, $(\boldsymbol{q}_h, u_h)_{\mathsf{m}}$ and the contribution from f, $(\boldsymbol{q}_h, u_h)_f$. By the identity (ii) of Lemma 2.2, we see that these two are orthogonal with respect to the bilinear form on $(V_h \times W_h)^2$ defined by

$$\langle\!\langle\, (\boldsymbol{q}_1, u_1), (\boldsymbol{q}_2, u_2)\,\rangle\!\rangle = \int_\Omega c\,\boldsymbol{q}_1 \cdot \boldsymbol{q}_2\,dx + \int_\Omega d\,u_1\,u_2\,dx.$$

3. The matrix entries.

3.1. The local matrices. To compute the matrices $\mathbb{E}$ and $\mathbb{H}$, we can proceed in the traditional finite element way. Let e denote an interior edge of the mesh $\mathcal{T}_h$, and let $\{\, L_{e,\ell}\,\}_{\ell=0}^k$ denote a basis for the set of polynomials of degree at most k on e. For example, we can choose properly scaled Legendre polynomials. Then, we set

$$(\boldsymbol{q}_{e,\ell}, u_{e,\ell}) = (\boldsymbol{q}_h, u_h)_{L_{e,\ell}}.$$

Now, for each element K, we compute the so-called local matrices, whose entries are

$$\begin{aligned}
\mathbb{E}_{K;e,\ell;e',\ell'} &= \int_K c\,\boldsymbol{q}_{e,\ell} \cdot \boldsymbol{q}_{e',\ell'}\,dx + \int_K d\,u_{e,\ell}\,u_{e',\ell'}\,dx,\\
\mathbb{H}_{g\,K;e,\ell} &= \int_{\partial K \cap \partial\Omega} g\,\boldsymbol{q}_{e,\ell} \cdot \boldsymbol{n}\,dx,\\
\mathbb{H}_{f\,K;e,\ell} &= \int_K u_{e,\ell}\,f\,dx.
\end{aligned}$$

Then the global matrices can be easily assembled by noting that

$$\begin{aligned}
\Lambda_{e,\ell}^t \mathbb{E}\,\Lambda_{e',\ell'} &= \sum_{K\in\mathcal{T}_h} \mathbb{E}_{K;e,\ell;e',\ell'},\\
\Lambda_{e,\ell}^t\,\mathbb{H}_g &= \sum_{K\in\mathcal{T}_h} \mathbb{H}_{g\,K;e,\ell},\\
\Lambda_{e,\ell}^t\,\mathbb{H}_f &= \sum_{K\in\mathcal{T}_h} \mathbb{H}_{f\,K;e,\ell}.
\end{aligned}$$

Since the lifting $(\boldsymbol{q}_{e,\ell}, u_{e,\ell})$ is supported only on the triangles sharing the edge e, to compute the local matrices, we have to provide only the numbers $\mathbb{E}_{K;e,\ell;e',\ell'}$, $\mathbb{H}_{g\,K;e,\ell}$, and $\mathbb{H}_{f\,K;e,\ell}$ for any two edges e and e' of K and any two integers ℓ and ℓ' between 0 and k; all the remaining entries are equal to zero. This also implies that the matrix $\mathbb{E}$ is a matrix of $(k+1)\times(k+1)$ blocks which has *at most four* off-diagonal blocks in each block column.

3.2. An example. The above computations can easily be carried out for the hybridized version of the RT method of lowest order. In this example, as the subscript ℓ in $\boldsymbol{q}_{e,\ell}$ and $u_{e,\ell}$ is superfluous, we drop it.

We begin by computing the lifting $\mathsf{m} \mapsto (\boldsymbol{q}_h, u_h)_{\mathsf{m}}$. Let m take the constant value λ_i on the edge e_i of the triangle K, $i = 1, 2, 3$. Then, on K,

$$(\boldsymbol{q}_h, u_h)_{\mathsf{m}} = \sum_{i=1}^{3} (\boldsymbol{q}_{e_i}, u_{e_i})\,\lambda_i,$$

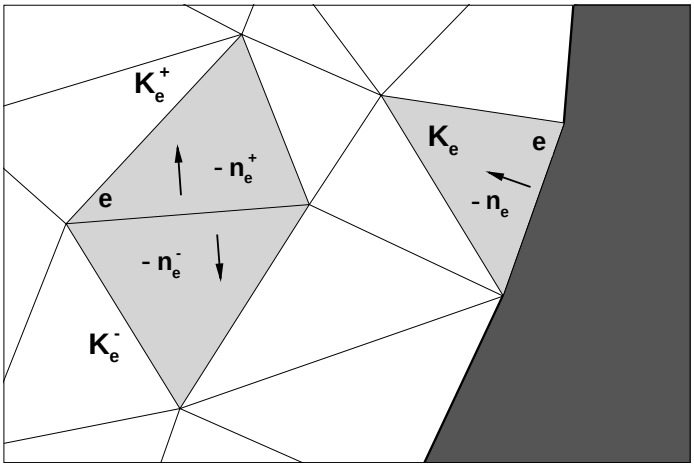


FIG. 1. *A representation of the lifting* $\boldsymbol{q}_{h,\mathsf{m}}$ *when* $\mathsf{m} = \chi_e$ *for interior and boundary edges* e. *In this case, we have taken* $d = 0$ *and* $c = Id$.

where, for $\boldsymbol{x} \in K$ and $i = 1, 2, 3$,

$$\boldsymbol{q}_{e_i}(\boldsymbol{x}) = -\frac{|\, e_i \,|}{|\, K \,|}\,(\bar{c})^{-1}\boldsymbol{n}_i - \frac{1}{2}\left(\frac{\bar{d}\,\rho_i}{1+\bar{d}\,\mathsf{h}^2}\right)(\boldsymbol{x}-\boldsymbol{B}), \qquad u_{e_i}(\boldsymbol{x}) = \frac{\rho_i}{1+\bar{d}\,\mathsf{h}^2},$$

and

$$\bar{c} = \frac{1}{|\, K \,|}\int_K c\,dx, \qquad \bar{d} = \frac{1}{|\, K \,|}\int_K d\,dx, \qquad \rho_i = \frac{|\, e_i \,|\,(\boldsymbol{m}_i - \boldsymbol{B})\cdot\boldsymbol{n}_i}{2\,|\, K \,|},$$
$$\mathsf{h}^2 = \frac{1}{4\,|\, K \,|}\int_K c\,(\boldsymbol{x}-\boldsymbol{B})\cdot(\boldsymbol{x}-\boldsymbol{B})\,dx, \qquad \boldsymbol{B} = (\bar{c})^{-1}\frac{\int_K c\,\boldsymbol{x}\,dx}{|\, K \,|}.$$

Here $\boldsymbol{m}_i$ denotes the midpoint of the edge e_i and $\boldsymbol{n}_i$ its outward unit normal. Note that if c is the identity, $\boldsymbol{B}$ is the barycenter of the triangle; if, moreover, K is an equilateral triangle of diameter h, then $\rho_i = 1/3$ and $\mathsf{h}^2 = h^2/48$. The case of c equaling the 2×2 identity and vanishing d is illustrated in Figure 1.

Now, it is easy to compute the entries of the local matrices:

$$\mathbb{E}_{K;e_i;e_j} = \frac{|\, e_i \,|\,|\, e_j \,|}{|\, K \,|}\,\boldsymbol{n}_j\cdot(\bar{c})^{-1}\boldsymbol{n}_j + \frac{\bar{d}\,|\, K \,|}{1+\bar{d}\,\mathsf{h}^2}\,\rho_i\,\rho_j,$$
$$\mathbb{H}_{g\,K;e_i} = -\sum_{e_j\subset\partial K\cap\partial\Omega}\left(\frac{|\, e_i \,|}{|\, K \,|}(\bar{c})^{-1}\boldsymbol{n}_i\cdot\boldsymbol{n}_j + \frac{|\, K \,|}{|\, e_j \,|}\left(\frac{\bar{d}}{1+\bar{d}\,\mathsf{h}^2}\right)\rho_i\,\rho_j\right)\int_{e_j} g(s)\,ds,$$
$$\mathbb{H}_{f\,K;e_i} = \frac{\rho_i}{1+\bar{d}\,\mathsf{h}^2}\int_K f(x)\,dx.$$

3.3. The reference element. A convenient implementation results if the local matrices for the multiplier can be computed by using quadratures on a reference element alone.

To achieve it, we need to define local mappings on the reference element and map spaces on the reference element $\widehat{K}$ to corresponding ones on any triangle K. Let $\widehat{K}$ be mapped one-to-one onto K by the standard affine mapping

$$\boldsymbol{x} = D_K\,\widehat{\boldsymbol{x}} + b_K,$$

and let us set

$$\widehat{u}(\widehat{\boldsymbol{x}}) \equiv u(\boldsymbol{x}), \quad \widehat{\boldsymbol{q}}(\widehat{\boldsymbol{x}}) \equiv |\det D_K|\, D_K^{-1}\, \boldsymbol{q}(\boldsymbol{x})$$

for scalar-valued functions u and vector-valued functions $\boldsymbol{q}$, respectively. Finally, set

$$\begin{aligned}
\widehat{V} &= P^k(\widehat{K}) \times P^k(\widehat{K}) + \boldsymbol{x}\, P^k(\widehat{K}),\\
\widehat{W} &= P^k(\widehat{K}),\\
\widehat{M} &= \{\widehat{\mathsf{m}} \in L^2(\partial\widehat{K}) : \quad \widehat{\mathsf{m}}|_{\widehat{e}} \in P^k(\widehat{e}) \quad \text{for all } \ \widehat{e} \in \partial\widehat{K}\}.
\end{aligned}$$

Now, suppose we are given a symmetric positive definite 2×2 matrix function $\widehat{C}$ and a scalar nonnegative function $\widehat{D}$ on $\widehat{K}$. For each $\widehat{\mathsf{m}}$ in $\widehat{M}$, we define the element $(\widehat{\boldsymbol{Q}}_{\widehat{\mathsf{m}}}, \widehat{U}_{\widehat{\mathsf{m}}}) \in \widehat{V} \times \widehat{W}$ by requiring that

$$\begin{aligned}
&\int_{\widehat{K}} \widehat{C}\, \widehat{\boldsymbol{Q}}_{\widehat{\mathsf{m}}} \cdot \boldsymbol{V}\, d\widehat{\boldsymbol{x}} - \int_{\widehat{K}} \widehat{U}_{\widehat{\mathsf{m}}} \widehat{\nabla} \cdot \boldsymbol{V}\, d\widehat{\boldsymbol{x}} = -\sum_{\widehat{e}\in\partial\widehat{K}} \int_{\widehat{e}} \widehat{\mathsf{m}}\, [\![\boldsymbol{V}]\!]\, d\widehat{s},\\
&\int_{\widehat{K}} W \widehat{\nabla} \cdot \widehat{\boldsymbol{Q}}_{\widehat{\mathsf{m}}}\, d\widehat{\boldsymbol{x}} + \int_{\widehat{K}} \widehat{D}\, \widehat{U}_{\widehat{\mathsf{m}}}\, W\, d\widehat{\boldsymbol{x}} = 0
\end{aligned}$$

hold for all $(\boldsymbol{V}, W) \in \widehat{V} \times \widehat{W}$. Then, we have the following result.

PROPOSITION 3.1. *Let K be any triangle and e be one of its edges. Set*

$$\widehat{C}(\widehat{\boldsymbol{x}}) = |\det D_K|^{-1}\, D_K^t\, c(\boldsymbol{x})\, D_K \quad \textit{and} \quad \widehat{D}(\widehat{\boldsymbol{x}}) = |\det D_K|\, d(\boldsymbol{x}).$$

Then, the lifting $(\boldsymbol{q}_{e,\ell}, u_{e,\ell})$ on K when mapped to $\widehat{K}$ satisfies

$$\widehat{\boldsymbol{q}}_{e,\ell} = \widehat{\boldsymbol{Q}}_{\widehat{L_{e,\ell}}} \qquad \textit{and} \qquad \widehat{u}_{e,\ell} = \widehat{U}_{\widehat{L_{e,\ell}}}.$$

Moreover,

$$\begin{aligned}
\mathbb{E}_{K;e,\ell;e',\ell'} &= -\int_{\widehat{e}} \widehat{L_{e,\ell}}\, \widehat{\boldsymbol{Q}}_{\widehat{L_{e',\ell'}}} \cdot \widehat{\boldsymbol{n}}\, d\widehat{s},\\
\mathbb{H}_{g\,K;e,\ell} &= \int_{\widehat{e}} \widehat{g}\, \widehat{\boldsymbol{Q}}_{\widehat{L_{e,\ell}}} \cdot \widehat{\boldsymbol{n}}\, d\widehat{s},\\
\mathbb{H}_{f\,K;e,\ell} &= \int_{\widehat{K}} \widehat{U}_{\widehat{L_{e,\ell}}}\, \widehat{f}\, |\det D_K|\, d\widehat{\boldsymbol{x}}.
\end{aligned}$$

This result can be easily proved by a straightforward change of variables and application of Lemma 2.2. Note that if the functions $\widehat{C}$ and $\widehat{D}$ are constant, there is no need to use quadrature rules to find the matrix entries.

4. Comparison with the hybridized BDM method. Now we compare the multipliers given by the hybridized version of the RT method and those given by the hybridized version of the corresponding BDM method.

4.1. Statement of the results. To state our comparison results, we first introduce the hybridized BDM method. The approximate solution given by this method, $(\boldsymbol{q}_h^{\mathrm{BDM}}, u_h^{\mathrm{BDM}}, \lambda_h^{\mathrm{BDM}})$, is sought in the finite element space $V_h^{\mathrm{BDM}} \times W_h^{\mathrm{BDM}} \times M_h$ given by

$$\begin{aligned}
V_h^{\mathrm{BDM}} &= \{\boldsymbol{v} \in L^2(\Omega)\times L^2(\Omega) : \quad \boldsymbol{v}|_K \in P^k(K)\times P^k(K) \quad \text{for all } \ K \in \mathcal{T}_h\},\\
W_h^{\mathrm{BDM}} &= \{w \in L^2(\Omega) : \quad w|_K \in P^{k-1}(K) \quad \text{for all } \ K \in \mathcal{T}_h\},
\end{aligned}$$

where $k \geq 1$, and is defined by requiring that, for all $(\boldsymbol{v}, u, \mu) \in V_h^{\mathrm{BDM}} \times W_h^{\mathrm{BDM}} \times M_h$,

$$\int_\Omega c\, \boldsymbol{q}_h^{\mathrm{BDM}} \cdot \boldsymbol{v}\, dx - \sum_{K \in \mathcal{T}_h} \int_K u_h^{\mathrm{BDM}} \nabla \cdot \boldsymbol{v}\, dx + \sum_{e \in \mathcal{E}_{i,h}} \int_e \lambda_h^{\mathrm{BDM}} [\![\boldsymbol{v}]\!]\, ds = -\int_{\partial\Omega} g\, [\![\boldsymbol{v}]\!]\, ds,$$

$$\sum_{K \in \mathcal{T}_h} \int_K w \nabla \cdot \boldsymbol{q}_h^{\mathrm{BDM}}\, dx + \int_\Omega d\, u_h^{\mathrm{BDM}}\, w\, dx = \int_\Omega f\, w\, dx,$$

$$\sum_{e \in \mathcal{E}_{i,h}} \int_e \mu\, [\![\boldsymbol{q}_h^{\mathrm{BDM}}]\!]\, ds = 0.$$

Note that the approximate solution of the BDM method satisfies exactly the same weak formulation as the approximate solution of the RT method; the only difference is the choice of the finite element spaces. As a consequence, the characterization theorem (Theorem 2.1) holds for the hybridized version of the BDM method. This is the key fact that allows us to compare the hybridized versions of the RT and the BDM methods.

In comparing the RT and BDM methods, for the sake of readability, we shall superscript the notation previously introduced in connection with the RT method by "RT." When superscripted by "BDM," such notation is to be understood as defined exactly as before except that the RT spaces are replaced by the BDM spaces. For example, $(\boldsymbol{q}_h, u_h)_{\mathsf{m}}^{\mathrm{BDM}} \equiv (\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}, u_{h,\mathsf{m}}^{\mathrm{BDM}}) \in V_h^{\mathrm{BDM}} \times W_h^{\mathrm{BDM}}$ is defined by requiring that

$$\int_\Omega c\, \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}} \cdot \boldsymbol{v}\, dx - \sum_{K \in \mathcal{T}_h} \int_K u_{h,\mathsf{m}}^{\mathrm{BDM}} \nabla \cdot \boldsymbol{v}\, dx = -\sum_{e \in \mathcal{E}_h} \int_e \mathsf{m}\, [\![\boldsymbol{v}]\!]\, ds,$$

$$\sum_{K \in \mathcal{T}_h} \int_K w \nabla \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\, dx + \int_\Omega d\, u_{h,\mathsf{m}}^{\mathrm{BDM}}\, w\, dx = 0$$

hold for all $(\boldsymbol{v}, w) \in V_h^{\mathrm{BDM}} \times W_h^{\mathrm{BDM}}$.

To state our first comparison theorem, we need the following additional notation: Denote by $\mathcal{P}_k$ the L^2-orthogonal projection into the space of functions which are piecewise polynomials of degree k on each triangle $K \in \mathcal{T}_h$. Let $\mathcal{R}f = \mathcal{P}_k f - \mathcal{P}_{k-1} f$ for all $f \in L^2(\Omega)$. Define the form

$$b_{h,\mathcal{R}f}^{\mathrm{RT}}(\mu) = \int_\Omega \mathcal{R}f\, u_{h,\mu}^{\mathrm{RT}}\, dx \quad \text{for all } \mu \in M_h$$

and the function $\rho_h \equiv \rho_h(\mathcal{R}f) \in M_h$ by

$$a_h^{\mathrm{RT}}(\rho_h, \mu) = b_{h,\mathcal{R}f}^{\mathrm{RT}}(\mu) \quad \text{for all } \mu \in M_h.$$

Also set

$$\Psi(\mu, g, f) = (\mathbf{0}, \mathcal{R}u_{h,\mu}^{\mathrm{RT}}) + (\mathbf{0}, \mathcal{R}u_{h,g}^{\mathrm{RT}}) + (\mathbf{0}, \mathcal{R}u_{h,\mathcal{P}_{k-1}f}^{\mathrm{RT}}) \quad \text{and}$$
$$\Upsilon(\mathcal{R}f) = (\boldsymbol{q}_h, u_h)_{\rho_h}^{\mathrm{RT}} + (\boldsymbol{q}_h, u_h)_{\mathcal{R}f}^{\mathrm{RT}}.$$

Now we can state the theorem.

THEOREM 4.1 (comparison of the RT and BDM methods for $d = 0$).
Assume that $d(\boldsymbol{x}) = 0$ almost everywhere in Ω.

1. *Suppose $f \in L^2(\Omega)$ is such that $(\mathcal{P}_k - \mathcal{P}_{k-1})f = 0$. Then, the Lagrange multiplier components of the RT and BDM solutions coincide:*

$$\lambda_h^{\mathrm{RT}} = \lambda_h^{\mathrm{BDM}}.$$

2. *If $f \in L^2(\Omega)$ is arbitrary, then the following statements hold:*

$$
\begin{aligned}
&(\alpha) \quad a_h^{\mathrm{RT}}(\mathsf{m},\mu) = a_h^{\mathrm{BDM}}(\mathsf{m},\mu) && \text{for all } \mathsf{m},\mu \in L^2(\mathcal{E}_h).\\
&(\beta) \quad b_h^{\mathrm{RT}}(\mu) = b_h^{\mathrm{BDM}}(\mu) + b_{h,\mathcal{R}f}^{\mathrm{RT}}(\mu) && \text{for all } \mu \in L^2(\mathcal{E}_h).\\
&(\gamma) \quad \lambda_h^{\mathrm{RT}} = \lambda_h^{\mathrm{BDM}} + \rho_h(\mathcal{R}f).\\
&(\delta) \quad (\boldsymbol{q}_h^{\mathrm{RT}}, u_h^{\mathrm{RT}}) = (\boldsymbol{q}_h^{\mathrm{BDM}}, u_h^{\mathrm{BDM}}) + \Psi(\lambda_h^{\mathrm{BDM}}, g, f) + \Upsilon(\mathcal{R}f).
\end{aligned}
$$

Before proving the theorem, let us discuss the result and some of its consequences.

Remark 4.1. Statements (α) and (β) can be easily rewritten in matrix form as follows:

$$
\begin{aligned}
&(\alpha') \quad \mathbb{E}^{\mathrm{RT}} = \mathbb{E}^{\mathrm{BDM}},\\
&(\beta') \quad \mathbb{H}^{\mathrm{RT}} = \mathbb{H}^{\mathrm{BDM}} + \mathbb{H}_{\mathcal{R}f}^{\mathrm{RT}},
\end{aligned}
$$

where the matrix $\mathbb{E}^{\mathrm{N}}$ is the stiffness matrix associated to the bilinear form $a_h^{\mathrm{N}}(\cdot,\cdot)$, and the matrix $\mathbb{H}^{\mathrm{N}}$ is the right-hand side matrix associated to the linear form $b_h^{\mathrm{N}}(\cdot)$ for $\mathrm{N} \in \{\mathrm{RT}, \mathrm{BDM}\}$. The matrix $\mathbb{H}_{\mathcal{R}f}^{\mathrm{RT}}$ is, of course, the right-hand side matrix associated to the linear form $b^{\mathrm{RT}}(\cdot)_{h,\mathcal{R}f}$. This means that, when $d = 0$, the stiffness matrices of the multipliers of both methods *coincide*. However, the right-hand side matrices differ. But they differ by a matrix which vanishes when $\mathcal{R}f = 0$—hence the coincidence of the Lagrange multipliers whenever $\mathcal{R}f = 0$.

Remark 4.2. The coincidence of Lagrange multipliers asserted by the theorem in the case $\mathcal{R}f = 0$ (and $d = 0$) appears to have gone unnoticed hitherto even numerically. This case occurs, e.g., when f is a polynomial of degree $k-1$ on every element of the mesh and in several applications of practical interest, e.g., incompressible flow in porous media (where $f = 0$ usually). The condition $\mathcal{R}f = 0$ is not only sufficient but also necessary for such a coincidence: From the characterization theorem, it is clear that the only part of f that determines λ_h^{RT} is $\mathcal{P}_k f$, while the only part of f that determines λ_h^{BDM} is $\mathcal{P}_{k-1} f$. Therefore, setting f to a polynomial of degree k for which $0 = \mathcal{P}_{k-1} f \neq \mathcal{P}_k f$, we can make $\lambda_h^{\mathrm{RT}} \neq \lambda_h^{\mathrm{BDM}}$.

Remark 4.3. Statement (δ) of the theorem shows how the solution components other than the multipliers are related. Obviously, $\Upsilon(\mathcal{R}f)$ depends linearly on $\mathcal{R}f$. Therefore, when $\mathcal{R}f = 0$, the solution $(\boldsymbol{q}_h^{\mathrm{RT}}, u_h^{\mathrm{RT}})$ differs from $(\boldsymbol{q}_h^{\mathrm{BDM}}, u_h^{\mathrm{BDM}})$ only by $\Psi(\lambda_h^{\mathrm{BDM}}, g, f)$, a function that can be computed locally element by element. In particular, this means that it is possible to implement the less expensive BDM method and locally recover the RT solution u_h^{RT}, which is one order higher in accuracy (under certain regularity assumptions). In this sense, (δ) can be thought of as yielding a postprocessing technique. Of course, one can then further postprocess the RT solution by the technique of [1] and gain one further order in accuracy.

Remark 4.4. It is well known that the Lagrange multiplier of both the RT and the BDM methods approximates the traces of the exact solution u on mesh edges. Specifically, under certain regularity assumptions, [5, Lemma 4.1] asserts that when $\mathcal{T}_h$ is a quasi-uniform mesh with mesh size h, as $h \to 0$,

$$
\|\lambda_h^{\mathrm{BDM}} - \mathcal{P}_{M_h} u\|_{\mathcal{E}_h} = O(h^{k+3/2-\delta_{k1}}), \tag{4.1}
$$

where $\mathcal{P}_{M_h}$ denotes the L^2-orthogonal projection onto M_h, $\|\cdot\|_{\mathcal{E}_h}$ denotes the $L^2(\mathcal{E}_h)$-norm, and δ_{k1} is zero for all k except for $k = 1$, in which case it equals one. The analogous estimate for the RT method [1, Corollary 1.5] is

$$
\|\lambda_h^{\mathrm{RT}} - \mathcal{P}_{M_h} u\|_{\mathcal{E}_h} = O(h^{k+3/2}).
$$

However, obviously whenever $\lambda_h^{\mathrm{RT}} = \lambda_h^{\mathrm{BDM}}$, there can be no difference in the convergence rates. Therefore, when $k = 1$ and $\mathcal{R}f = 0$, by virtue of Theorem 4.1, we conclude that although (4.1) provides for only $O(h^{3/2})$-convergence, in fact the convergence rate is at least $O(h^{5/2})$.

When $d \neq 0$, the liftings $\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}}$ and $\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}$ are no longer divergence-free on each element, in general. In fact, as we show later, there are multipliers m for which $\mathcal{R}u_{h,\mathsf{m}}^{\mathrm{RT}} \neq 0$. This property implies that the statements of Theorem 4.1 do not hold in general. In particular, the following theorem provides a case wherein statement (α) does not hold. Obviously, whenever statement (α) fails to hold, one cannot expect coincidence of RT and BDM Lagrange multipliers.

THEOREM 4.2 (comparison of the RT and BDM methods for $d \geq 0$). *Assume that $c(\boldsymbol{x})$ and $d(\boldsymbol{x})$ are constant on each element of the mesh. Then, whenever $d(\boldsymbol{x})$ is positive on at least one element, statement (α) of Theorem 4.2 does not hold.*

Now, we prove Theorems 4.1 and 4.2.

4.2. Proof of Theorem 4.1. The proof proceeds by establishing a connection between the BDM and RT liftings and applying the characterization theorem. Two properties of the finite element spaces of the RT and BDM methods play a crucial role. The first is simply that the multipliers of both methods share the same space. The second is that the elements of V_h^{RT} whose divergence on each element is a polynomial of degree $k-1$ also belong to the space V_h^{BDM}. Let us begin by proving the latter property.

LEMMA 4.3. *The following containment holds:*

$$\{\boldsymbol{q}_h \in V_h^{\mathrm{RT}} : \nabla \cdot \boldsymbol{q}_h|_K \in P^{k-1}(K) \ \text{ for all } \ K \in \mathcal{T}_h\} \subset V_h^{\mathrm{BDM}}.$$

Proof. If $\boldsymbol{q}_h \in V_h^{\mathrm{RT}}$, then $\boldsymbol{q}_h|_K = \boldsymbol{v}_k + \boldsymbol{x}\widetilde{p}_k$ for some $\boldsymbol{v}_k \in P^k(K) \times P^k(K)$ and some homogeneous polynomial $\widetilde{p}_k(\boldsymbol{x})$ of degree k on K. Taking the divergence, we find that

$$\nabla \cdot (\boldsymbol{q}_h|_K) = \nabla \cdot \boldsymbol{v}_k + (k+2)\widetilde{p}_k.$$

Therefore, $\nabla \cdot (\boldsymbol{q}_h|_K) \in P^{k-1}(K)$ implies that $\widetilde{p}_k = 0$, and consequently $\boldsymbol{q}_h|_K \in P^k(K) \times P^k(K)$. Hence $\boldsymbol{q}_h \in V_h^{\mathrm{BDM}}$. This completes the proof. ∎

The next result uses the above lemma and the definition of the local mappings to establish key relations between the local mappings of the RT and BDM methods.

LEMMA 4.4. *Assume that $d = 0$. Then, for all $\mathsf{m} \in L^2(\mathcal{E}_h)$ and $f \in L^2(\Omega)$,*

$$\begin{aligned} &\text{(i)} \quad \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} = \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}, && u_{h,\mathsf{m}}^{\mathrm{RT}} = u_{h,\mathsf{m}}^{\mathrm{BDM}} + \mathcal{R}u_{h,\mathsf{m}}^{\mathrm{RT}}, \\ &\text{(ii)} \quad \boldsymbol{q}_{h,f}^{\mathrm{RT}} = \boldsymbol{q}_{h,f}^{\mathrm{BDM}} + \boldsymbol{q}_{h,\mathcal{R}f}^{\mathrm{RT}}, && u_{h,f}^{\mathrm{RT}} = u_{h,f}^{\mathrm{BDM}} + \mathcal{R}u_{h,\mathcal{P}_{k-1}f}^{\mathrm{RT}} + u_{h,\mathcal{R}f}^{\mathrm{RT}}. \end{aligned}$$

Proof. Let us begin by proving (i). From (2.9), we have

$$\int_\Omega c\,(\,\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} - \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\,) \cdot \boldsymbol{v}\,dx - \sum_{K\in\mathcal{T}_h} \int_K (\,u_{h,\mathsf{m}}^{\mathrm{RT}} - u_{h,\mathsf{m}}^{\mathrm{BDM}}\,)\,\nabla\cdot\boldsymbol{v}\,dx = 0 \quad \text{for all } \ \boldsymbol{v} \in V_h^{\mathrm{BDM}},$$

and from (2.10), $\nabla \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}}|_K = 0$ and $\nabla \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}|_K = 0$ for all $K \in \mathcal{T}_h$. Since by Lemma 4.3 $\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} \in V_h^{\mathrm{BDM}}$, we can take $\boldsymbol{v} = \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} - \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}$ in the first equation of this proof to get that $\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} = \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}$. It immediately follows that

$$-\sum_{K\in\mathcal{T}_h} \int_K (\,u_{h,\mathsf{m}}^{\mathrm{RT}} - u_{h,\mathsf{m}}^{\mathrm{BDM}}\,)\,\nabla\cdot\boldsymbol{v}\,dx = 0 \qquad \text{for all } \ \boldsymbol{v} \in V_h^{\mathrm{BDM}},$$

which implies that $\mathcal{P}_{k-1}\,u_{h,\mathsf{m}}^{\mathrm{RT}} = u_{h,\mathsf{m}}^{\mathrm{BDM}}$. This proves (i).

Now, let us prove (ii). It suffices to show that for $p = \mathcal{P}_{k-1} f$,

$$\boldsymbol{q}_{h,p}^{\mathrm{RT}} = \boldsymbol{q}_{h,p}^{\mathrm{BDM}} \quad \text{and} \quad u_{h,p}^{\mathrm{RT}} = u_{h,p}^{\mathrm{BDM}} + \mathcal{R} u_{h,p}^{\mathrm{RT}}. \tag{4.2}$$

Indeed, once we have (4.2), by linearity and the obvious equality

$$(\boldsymbol{q}_h, u_h)^{\mathrm{BDM}}_{\mathcal{P}_{k-1} f} = (\boldsymbol{q}_h, u_h)^{\mathrm{BDM}}_{f},$$

we get that

$$\begin{aligned}(\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{f} &= (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{p} + (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{\mathcal{R} f} \\ &= (\boldsymbol{q}_h, u_h)^{\mathrm{BDM}}_{p} + (\boldsymbol{0}, \mathcal{R} u_{h,p}^{\mathrm{RT}}) + (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{\mathcal{R} f},\end{aligned}$$

and (ii) follows. To show (4.2), first observe that since $p|_K = (\mathcal{P}_{k-1} f)|_K \in P^{k-1}(K)$, (2.12) implies that

$$\nabla \cdot \boldsymbol{q}_{h,p}^{\mathrm{RT}}|_K = p \quad \text{and} \quad \nabla \cdot \boldsymbol{q}_{h,p}^{\mathrm{BDM}}|_K = p \quad \text{for all } \ K \in \mathcal{T}_h.$$

Therefore, using Lemma 4.3 again, $\boldsymbol{q}_{h,p}^{\mathrm{RT}} \in V_h^{\mathrm{BDM}}$. Now, (2.11) yields

$$\int_\Omega c \, \big(\boldsymbol{q}_{h,p}^{\mathrm{RT}} - \boldsymbol{q}_{h,p}^{\mathrm{BDM}} \big) \cdot \boldsymbol{v} \, dx - \sum_{K \in \mathcal{T}_h} \int_K \big(u_{h,p}^{\mathrm{RT}} - u_{h,p}^{\mathrm{BDM}} \big) \nabla \cdot \boldsymbol{v} \, dx = 0 \quad \text{for all } \ \boldsymbol{v} \in V_h^{\mathrm{BDM}}.$$

Since we have shown that $\boldsymbol{q}_{h,p}^{\mathrm{RT}} \in V_h^{\mathrm{BDM}}$, we can choose $\boldsymbol{v} = \boldsymbol{q}_{h,p}^{\mathrm{RT}} - \boldsymbol{q}_{h,p}^{\mathrm{BDM}}$ above. Then the first equality of (4.2), namely, $\boldsymbol{q}_{h,p}^{\mathrm{RT}} = \boldsymbol{q}_{h,p}^{\mathrm{BDM}}$, immediately follows. The second equality of (4.2) also follows, because by (i) we have $\mathcal{P}_{k-1} \, u_{h,p}^{\mathrm{RT}} = u_{h,p}^{\mathrm{BDM}}$. This completes the proof. ∎

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. First, observe that the first conclusion of the theorem, namely, $\lambda_h^{\mathrm{RT}} = \lambda_h^{\mathrm{BDM}}$ whenever $\mathcal{R} f = 0$, follows from statement (γ), because $\rho_h(\mathcal{R} f)$ depends linearly on $\mathcal{R} f$. Therefore, we shall prove only statements (α)–(δ).

Applying the characterization theorem with $d = 0$, we find that

$$a_h^{\mathrm{N}}(\mathsf{m}, \mu) = \int_\Omega c \, \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{N}} \cdot \boldsymbol{q}_{h,\mu}^{\mathrm{N}} \quad \text{for } {}_{\mathrm{N}} \in \{{}_{\mathrm{RT}}, {}_{\mathrm{BDM}}\},$$

so the equality of (α) follows from the first equality of Lemma 4.4(i). Similarly, since

$$b_h^{\mathrm{N}}(\mu) = \int_{\partial\Omega} g \, [\![\boldsymbol{q}_{h,\mu}^{\mathrm{N}}]\!] \, ds + \int_\Omega f \, u_{h,\mu}^{\mathrm{N}} \, dx \quad \text{for } {}_{\mathrm{N}} \in \{{}_{\mathrm{RT}}, {}_{\mathrm{BDM}}\},$$

the second equality, namely, (β), also follows from Lemma 4.4(i). Now, statement (γ) obviously follows from (α), (β), and the definitions of the multipliers.

To prove statement (δ), we start again from the characterization theorem, apply statement (γ), and use the identities of Lemma 4.4 in succession:

$$\begin{aligned}(\boldsymbol{q}_h^{\mathrm{RT}}, u_h^{\mathrm{RT}}) &= (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{\lambda_h^{\mathrm{RT}}} + (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{g} + (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{f} \\ &= (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{(\lambda_h^{\mathrm{BDM}} + \rho_h)} + (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{g} + (\boldsymbol{q}_h, u_h)^{\mathrm{RT}}_{f} \\ &= (\boldsymbol{q}_h, u_h)^{\mathrm{BDM}}_{\lambda_h^{\mathrm{BDM}}} + (\boldsymbol{0}, \mathcal{R} u_{h,\lambda_h^{\mathrm{BDM}}}^{\mathrm{RT}} + u_{h,\rho_h}^{\mathrm{RT}}) + (\boldsymbol{q}_{h,g}^{\mathrm{BDM}}, u_{h,g}^{\mathrm{BDM}} + \mathcal{R} u_{h,g}^{\mathrm{RT}}) \\ &\qquad + (\boldsymbol{q}_{h,f}^{\mathrm{BDM}} + \boldsymbol{q}_{h,\mathcal{R} f}^{\mathrm{RT}}, \ u_{h,f}^{\mathrm{BDM}} + \mathcal{R} u_{h,\mathcal{P}_{k-1} f}^{\mathrm{RT}} + u_{h,\mathcal{R} f}^{\mathrm{RT}}) \\ &= (\boldsymbol{q}_h, u_h)^{\mathrm{BDM}}_{\lambda_h^{\mathrm{BDM}}} + (\boldsymbol{q}_h, u_h)^{\mathrm{BDM}}_{g} + (\boldsymbol{q}_h, u_h)^{\mathrm{BDM}}_{f} \\ &\qquad + \Psi(\lambda_h^{\mathrm{BDM}}, g, f) + \Upsilon(\mathcal{R} f).\end{aligned}$$

The first three terms on the right-hand side of the last equality sum to the BDM solution. This completes the proof of Theorem 4.1. ∎

4.3. Proof of Theorem 4.2. To prove this result, we begin by studying the local spaces of the RT method, namely,

$$\begin{aligned} V_K^{\mathrm{RT}} &= P^k(K) \times P^k(K) + \boldsymbol{x}\, P^k(K), \\ W_K^{\mathrm{RT}} &= P^k(K), \\ M_K^{\mathrm{RT}} &= \{\mu \in L^2(\partial K) : \mu|_e \in P^k(e) \text{ for each edge of } K\}. \end{aligned}$$

It turns out that, when $d > 0$, the lifting maps induce a natural orthogonal decomposition of the local space V_K^{RT}, namely,

$$V_K^{\mathrm{RT}} = V_K^0 \oplus V_K^{\perp},$$

where

$$\begin{aligned} V_K^0 &= \{\boldsymbol{v} \in V_K^{\mathrm{RT}} : \boldsymbol{v} \cdot \boldsymbol{n}_e = 0 \text{ on each edge } e \text{ of } K\}, \\ V_K^{\perp} &= \{\boldsymbol{v} \in V_K^{\mathrm{RT}} : (\!(\boldsymbol{v}, \boldsymbol{q})\!) = 0 \ \text{ for all } \boldsymbol{q} \in V_K^0\}, \end{aligned}$$

and

$$(\!(\boldsymbol{v}_1, \boldsymbol{v}_2)\!) = \int_K c\, \boldsymbol{v}_1 \cdot \boldsymbol{v}_2 \, dx + \int_K \frac{1}{d} \nabla \cdot \boldsymbol{v}_1 \, \nabla \cdot \boldsymbol{v}_2 \, dx.$$

Indeed, the following result states that the local space $V_K^{\perp}$ is nothing but the image of the lifting map $\mathsf{m} \mapsto \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}}$.

LEMMA 4.5. *Assume that $d(\boldsymbol{x})$ is a positive constant on an element $K \in \mathcal{T}_h$, and let $\{\mathsf{m}_i\}_{i=1}^{3(k+1)}$ be a basis of M_K^{RT}. Then, $\{\boldsymbol{q}_{h,\mathsf{m}_i}^{\mathrm{RT}}\}_{i=1}^{3(k+1)}$ is a basis of $V_K^{\perp}$.*

Proof. By (2.10), we have for any $\mathsf{m} \in M_K^{\mathrm{RT}}$

$$u_{h,\mathsf{m}}^{\mathrm{RT}} = -\frac{1}{d} \nabla \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}}. \tag{4.3}$$

Substituting this expression for $u_{h,\mathsf{m}}^{\mathrm{RT}}$ into (2.9), we see that (2.9) can be rewritten as follows:

$$(\!(\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}}, \boldsymbol{v})\!) = -\sum_{e \in \partial K} \int_e \mathsf{m}\, [\![\boldsymbol{v}]\!]\, ds.$$

As a consequence

$$\{\boldsymbol{q}_{h,\mathsf{m}_i}^{\mathrm{RT}}\}_{i=1}^{3(k+1)} \subset V_K^{\perp}.$$

Since the dimension of $V_K^{\perp}$ is $3(k+1)$, it remains only to show that the elements of the set $\{\boldsymbol{q}_{h,\mathsf{m}_i}^{\mathrm{RT}}\}_{i=1}^{3(k+1)}$ are linearly independent. So, assume that there are scalars α_i such that

$$\sum_{i=1}^{3(k+1)} \alpha_i\, \boldsymbol{q}_{h,\mathsf{m}_i}^{\mathrm{RT}} = 0.$$

By the linearity of the lifting, this implies that $\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} = 0$, where

$$\mathsf{m} = \sum_{i=1}^{3(k+1)} \alpha_i\, \mathsf{m}_i.$$

However, since by (4.3)

$$u_{h,\mathsf{m}}^{\mathrm{RT}} = -\frac{1}{d}\nabla \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} = 0,$$

we have that the lifting $(\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}}, u_{h,\mathsf{m}}^{\mathrm{RT}})$ is zero. Consequently, $\mathsf{m} = 0$. Since $\{\mathsf{m}_i\}_{i=1}^{3(k+1)}$ is a basis for M_K^{RT}, it follows that $\alpha_i = 0$ for all $i = 1, \ldots, 3(k+1)$. This completes the proof. ∎

Now, we use the above result to show that the lifting $u_{h,\mathsf{m}}^{\mathrm{RT}}$ is not always a polynomial of degree at most $k-1$ on all the elements of the triangulation.

LEMMA 4.6. *Assume that $c(\boldsymbol{x})$ is constant on an element $K \in \mathcal{T}_h$. Also assume that $d(\boldsymbol{x})$ is a positive constant on K. Then there is a function $\mathsf{m} \in M_K^{\mathrm{RT}}$ such that*

$$\mathcal{R} u_{h,\mathsf{m}}^{\mathrm{RT}} \neq 0 \qquad \text{on } K.$$

Proof. Since by (4.3)

$$u_{h,\mathsf{m}}^{\mathrm{RT}} = -\frac{1}{d}\nabla \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}},$$

an application of Lemma 4.3 shows that $\mathcal{R} u_{h,\mathsf{m}}^{\mathrm{RT}} = 0$ if and only if

$$\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} \in P^k(K) \times P^k(K) \quad \text{for all } \mathsf{m} \in M_K^{\mathrm{RT}}. \tag{4.4}$$

We claim that this is not possible for all $\mathsf{m} \in M_K^{\mathrm{RT}}$. Indeed, if this were the case, $V_K^{\perp} \subset P^k(K) \times P^k(K)$. This implies that the orthogonal complement of $P^k(K) \times P^k(K)$ in V_K^{RT} with respect to the inner product $(\!(\cdot,\cdot)\!)$, which we denote by W_K, satisfies

$$W_K \subset V_K^0. \tag{4.5}$$

However, as we shall now see, this implies that $W_K = \{\mathbf{0}\}$, which is a contradiction.

In the orthogonality relation

$$\int_K c\boldsymbol{\phi} \cdot \boldsymbol{v}\, dx + \frac{1}{d}\int_K (\nabla \cdot \boldsymbol{\phi})(\nabla \cdot \boldsymbol{v})\, dx = 0 \quad \text{for all } \boldsymbol{v} \in P^k(K) \times P^k(K),$$

let us choose $\boldsymbol{v} = c^{-T}\nabla\eta$ for some $\eta \in P^{k+1}(K)$ (where c^{-T} denotes the inverse of the transpose of c):

$$\int_K \boldsymbol{\phi} \cdot \nabla\eta\, dx + \frac{1}{d}\int_K (\nabla \cdot \boldsymbol{\phi})(\nabla \cdot c^{-T}\nabla\eta)\, dx = 0. \tag{4.6}$$

By (4.5) and integration by parts

$$\int_K \boldsymbol{\phi} \cdot \nabla\eta\, dx + \int_K \eta \nabla \cdot \boldsymbol{\phi}\, dx = \int_{\partial K} \eta(\boldsymbol{\phi} \cdot \boldsymbol{n})\, ds = 0. \tag{4.7}$$

Subtracting (4.6) from (4.7), we have

$$\int_K \left(\eta - \frac{1}{d}\nabla \cdot c^{-T}\nabla\eta\right)(\nabla \cdot \boldsymbol{\phi})\, dx = 0 \quad \text{for all } \eta \in P^{k+1}(K). \tag{4.8}$$

Now we show that (4.8) implies that $\nabla \cdot \boldsymbol{\phi} = 0$. Choosing $\eta \in P^1(K)$ in (4.8), we conclude that $\nabla \cdot \boldsymbol{\phi}$ is $L^2(K)$-orthogonal to $P^1(K)$. For $k \geq 2$, if $\nabla \cdot \boldsymbol{\phi}$ is $L^2(K)$-orthogonal to $P^{k-1}(K)$, then choosing $\eta \in P^k(K)$ we find that $\nabla \cdot \boldsymbol{\phi}$ is $L^2(K)$-orthogonal to $P^k(K)$ as well, because $\nabla \cdot (c^{-T}\nabla\eta) \in P^{k-2}(K)$. Thus, by induction, $\nabla \cdot \boldsymbol{\phi}$ is zero.

It follows from $\nabla \cdot \boldsymbol{\phi} = 0$ that $\boldsymbol{\phi} = 0$: Indeed, any $\boldsymbol{\phi} \in W$ can be written as $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}\widetilde{p}_k - \mathcal{Q}_k(\boldsymbol{x}\widetilde{p}_k)$, where $\mathcal{Q}_k$ is the orthogonal projection onto $P^k(K) \times P^k(K)$ in the $(\!(\cdot,\cdot)\!)$-inner product, and $\widetilde{p}_k$ is a homogeneous polynomial of degree k. Therefore,

$$0 = \nabla \cdot \boldsymbol{\phi} = (k+2)\widetilde{p}_k - \nabla \cdot \mathcal{Q}_k(\boldsymbol{x}\widetilde{p}_k).$$

Since the latter term is in $P^{k-1}(K)$, we conclude that $\widetilde{p}_k = 0$, so $\boldsymbol{\phi} = \mathbf{0}$. Thus (4.4) does not hold, and the lemma is proved. ∎

The next result establishes an equivalent criterion for statement (α) of Theorem 4.1 in terms of the liftings.

LEMMA 4.7. *Assume that $d(\boldsymbol{x})$ is constant on every element of the mesh. Let $\mathsf{m} \in L^2(\partial K)$. Then*

$$a_h^{\mathrm{RT}}(\mathsf{m},\mathsf{m}) = a_h^{\mathrm{BDM}}(\mathsf{m},\mathsf{m})$$

if and only if

(i) $\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} = \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}$ *on Ω and*

(ii) $u_{\mathsf{m}}^{\mathrm{RT}} = u_{\mathsf{m}}^{\mathrm{BDM}}$ *on all elements $K \in \mathcal{T}_h$, where $d > 0$.*

Proof. Set $\mathbb{J}(\mathsf{m}) = a_h^{\mathrm{RT}}(\mathsf{m},\mathsf{m}) - a_h^{\mathrm{BDM}}(\mathsf{m},\mathsf{m})$. Since, by Theorem 2.1, we have, for $\mathrm{N} \in \{\mathrm{RT}, \mathrm{BDM}\}$,

$$a_h^{\mathrm{N}}(\mathsf{m},\mathsf{m}) = \int_\Omega c\,\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{N}} \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{N}}\,dx + \int_\Omega d\,u_{h,\mathsf{m}}^{\mathrm{N}}\,u_{h,\mathsf{m}}^{\mathrm{N}}\,dx,$$

a straightforward use of the identity $a^2 - b^2 = (a-b)^2 + 2\,b\,(a-b)$ allows us to write $\mathbb{J}(\mathsf{m}) = \Theta(\mathsf{m}) + \mathbb{D}(\mathsf{m})$, where

$$\Theta(\mathsf{m}) = \int_\Omega c\,(\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} - \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}) \cdot (\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} - \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}})\,dx + \int_\Omega d\,(u_{\mathsf{m}}^{\mathrm{RT}} - u_{\mathsf{m}}^{\mathrm{BDM}})^2\,dx$$

and

$$\mathbb{D}(\mathsf{m}) = 2\int_\Omega c\,\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}} \cdot \left(\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} - \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\right)\,dx + 2\int_\Omega d\,u_{h,\mathsf{m}}^{\mathrm{BDM}}\left(u_{h,\mathsf{m}}^{\mathrm{RT}} - u_{h,\mathsf{m}}^{\mathrm{BDM}}\right)\,dx.$$

We will now show that $\mathbb{D}(\mathsf{m}) = 0$. Consider the first term in the definition of $\mathbb{D}(\mathsf{m})$. By (2.9),

$$\begin{aligned}\int_\Omega c\,\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}} \cdot \left(\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} - \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\right)\,dx &= \int_\Omega c\,\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\,dx - \int_\Omega c\,\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}} \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\,dx \\ &= -\sum_{K\in\mathcal{T}_h}\int_K \left(u_{h,\mathsf{m}}^{\mathrm{RT}} - u_{h,\mathsf{m}}^{\mathrm{BDM}}\right)\,\nabla \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\,dx.\end{aligned}$$

Hence, using also (2.10), we have

$$\begin{aligned}\int_\Omega c\,\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}} \cdot \left(\boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{RT}} - \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\right)\,dx &= -\sum_{K\in\mathcal{T}_h}\int_K \left(\mathcal{P}_{k-1}u_{h,\mathsf{m}}^{\mathrm{RT}} - u_{h,\mathsf{m}}^{\mathrm{BDM}}\right)\,\nabla \cdot \boldsymbol{q}_{h,\mathsf{m}}^{\mathrm{BDM}}\,dx \\ &= -\int_\Omega d\,u_{h,\mathsf{m}}^{\mathrm{BDM}}\left(\mathcal{P}_{k-1}u_{h,\mathsf{m}}^{\mathrm{RT}} - u_{h,\mathsf{m}}^{\mathrm{BDM}}\right)\,dx.\end{aligned}$$

Inserting this expression in the definition of $\mathbb{D}$, we get

$$\mathbb{D}(\mathsf{m}) = 2\int_\Omega d\, u_{h,\mathsf{m}}^{\mathrm{BDM}} \left(u_{h,\mathsf{m}}^{\mathrm{RT}} - \mathcal{P}_{k-1} u_{h,\mathsf{m}}^{\mathrm{RT}} \right) dx = 0,$$

because d is constant on each element of the mesh. In other words, $\mathbb{J}(\mathsf{m}) = \Theta(\mathsf{m})$.

This implies that $\mathbb{J}(\mathsf{m}) = 0$ if and only if $\Theta(\mathsf{m}) = 0$. The lemma follows from the definition of $\Theta(\mathsf{m})$ and the fact that $c(\boldsymbol{x})$ is positive definite. ∎

We are now ready to prove Theorem 4.2.

Proof of Theorem 4.2. Since there is at least one element $K \in \mathcal{T}_h$ wherein d is a positive constant, Lemma 4.6 asserts the existence of at least one function $\mathsf{m} \in M_K^{\mathrm{RT}}$ for which $\mathcal{R} u_{h,\mathsf{m}}^{\mathrm{RT}} \neq 0$ on K. This implies that $u_{h,\mathsf{m}}^{\mathrm{RT}} \neq u_{h,\mathsf{m}}^{\mathrm{BDM}}$ on K. Therefore, by Lemma 4.7, for any $\mu \in M_h$ such that $\mu|_{\partial K} = \mathsf{m}$, we have $a_h^{\mathrm{RT}}(\mu,\mu) \neq a_h^{\mathrm{BDM}}(\mu,\mu)$. Consequently, statement (α) of Theorem 4.1 does not hold. This completes the proof. ∎

5. Concluding remarks. The characterization theorem obtained in this paper for the hybridized RT method on triangular meshes also holds for various other methods. For example, it holds for the RT method on simplicial meshes in any space dimension as well as on rectangular and cubic meshes. It also holds for the hybridized versions of the mixed methods of Brezzi, Douglas, and Marini [5, 6] on rectangles, the method of Brezzi et al. [3] on tetrahedra and bricks, and the method of Brezzi et al. [4] on triangles, rectangles, tetrahedra, and bricks.

As a consequence, the matrix entries for the multipliers of all of the above-mentioned methods can be computed as described in section 3. Moreover, a result similar to the comparison theorem (Theorem 4.1) holds for the RT and BDM methods on multidimensional simplices. However, when the elements are rectangles or bricks, the subspace of divergence-free members of the RT and BDM spaces on an element are not identical. Therefore, in general, we cannot expect a result analogous to Theorem 4.1 to hold in this case.

Other applications of the characterization theorem are studied elsewhere. Indeed, in [11] it is shown how to use the characterization theorem to construct a Schwarz preconditioner for the multiplier equation of the RT and BDM hybridized mixed methods of *any* order. Such preconditioners were known only for the lowest order hybridized RT method. Also, in a forthcoming paper, we show how to use the characterization result to obtain error estimates for the multipliers *without* relying on error estimates on the other variables, as is customarily done.

Finally, let us briefly comment on the relation between the hybridized mixed methods and the discontinuous Galerkin methods. It is not difficult to see that the counterpart of the multiplier λ_h is nothing but the so-called numerical trace of the approximation u_h given by the discontinuous Galerkin method. How to exploit this link to achieve a better theoretical understanding of *both* methods remains a challenging open problem; see [9].

Acknowledgment. The authors would like to thank Wolfgang Dahmen, whose stimulating visit to the I.M.A., University of Minnesota, in the Spring of 2001, prompted them to explore ways of characterizing stiffness matrices of mixed methods by using discontinuous test functions. This paper is the unexpected outcome of such exploration.

REFERENCES

[1] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] F. Brezzi, J. Douglas, Jr., R. E. Durán, and M. Fortin, *Mixed finite element methods for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.

[4] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

[5] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[6] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Variable degree mixed methods for second order elliptic problems*, Mat. Apl. Comput., 4 (1985), pp. 19–34.

[7] F. Brezzi, G. Manzini, L. D. Marini, P. Pietra, and A. Russo, *Discontinuous finite elements for diffusion problems*, in Atti Convegno in onore di F. Brioschi (Milan 1997), Istituto Lombardo, Accademia di Scienze e Lettere, Milan, Italy, 1999, pp. 197–217.

[8] F. Brezzi, G. Manzini, L. D. Marini, P. Pietra, and A. Russo, *Discontinuous Galerkin approximations for elliptic problems*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 365–378.

[9] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods. Theory, Computation and Applications, Lect. Notes Comput. Sci. Engrg. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, New York, 2000, pp. 3–50.

[10] B. Fraejis de Veubeke, *Displacement and equilibrium models in the finite element method*, in Stress Analysis, O. C. Zienkiewicz and G. Holister, eds., Wiley, New York, 1965.

[11] J. Gopalakrishnan, *A Schwarz preconditioner for a hybridized mixed method*, Comput. Methods Appl. Math., 3 (2003), pp. 116–134.

[12] L. D. Marini, *An inexpensive method for the evaluation of the solution of the lowest order Raviart–Thomas mixed method*, SIAM J. Numer. Anal., 22 (1985), pp. 493–496.

[13] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, New York, 1977, pp. 292–315.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 302–325

ADJOINT-BASED ITERATIVE METHOD FOR ROBUST CONTROL PROBLEMS IN FLUID MECHANICS*

T. TACHIM MEDJO† AND L. R. TCHEUGOUE TEBOU†

Abstract. In this article we study the convergence of an adjoint-based iterative method recently proposed in [T. R. Bewley, R. Temam, and M. Ziane, *Phys. D*, 138 (2000), pp. 360–392] for the numerical solution of a class of nonlinear robust control problems in fluid mechanics. Under weaker assumptions than those of [T. Tachim Medjo, *Numer. Funct. Anal. Optim.*, 23 (2002), pp. 849–873], we prove the convergence of the algorithm, and we obtain an estimate of the convergence rate. Numerical solutions of a robust control problem related to data assimilation in oceanography are presented to illustrate the method.

Key words. robust control, fluid mechanics, Navier–Stokes, iterative methods

AMS subject classifications. 35, 49, 65, 76

DOI. 10.1137/S0036142902416231

1. Introduction.

Optimal control of fluid flows such as those governed by the Navier–Stokes equations has long been the subject of extensive studies, and applications to engineering and sciences are tremendous. In recent years, the control of fluid flows has also become the subject of intense study in the mathematical community. Due to advances in computer technology, the application of optimal control theory to complex phenomena such as weather forecasting and oceanography is becoming a reality [1, 2, 6, 9, 11, 14, 15, 16, 17, 25, 26, 27, 28]. A classical control problem arising in meteorology and oceanography, and related to *data assimilation*, is the adjustment of initial conditions in order to obtain a flow that agrees with the observations. This has been the subject of extensive studies in the past, and a lot of progress has been made both mathematically and computationally to understand the subject [1, 2, 6, 9, 11, 14, 15, 16, 17, 25, 26, 27, 28].

In general, the application of optimal control theory to complex problems in fluid mechanics has proven to be quite effective when complete state information from high resolution numerical simulations is available [6]. As suggested in [6], in order to extend this infinite-dimensional optimization approach to control externally disturbed flows in which the controls must be determined based on limited noisy flow measurements alone, it is necessary that the controls computed be insensitive to both state disturbances and measurement noise.

As described in [6], robust control theory, which generalizes optimal control theory, can be represented as a differential game between an engineer seeking the "best" control that stabilizes the flow perturbation with limited control effort and, simultaneously, nature seeking the "maximally malevolent" disturbance that destabilizes the flow perturbation with limited disturbance magnitude. In [6], the authors present a general framework for robust control problems in fluid mechanics. Given a fairly general cost functional $\mathcal{J} = \mathcal{J}(\psi, \phi)$, the authors in [6] proved the existence of a saddle point $(\bar{\psi}, \bar{\phi})$, which maximizes $\mathcal{J}$ with respect to the disturbance ψ and minimizes $\mathcal{J}$ with respect to the control ϕ, subject to the Navier–Stokes equations. Chaotic prob-

*Received by the editors October 17, 2002; accepted for publication (in revised form) May 19, 2003; published electronically January 28, 2004.
http://www.siam.org/journals/sinum/42-1/41623.html
†Department of Mathematics, Florida International University, University Park, Miami, FL 33199 (tachimt@fiu.edu, teboul@fiu.edu).

lems, such as weather systems, are highly susceptible to the small disturbances present in all physical systems. The robust control framework presented in [6] should help to reduce the component of the initial state most susceptible to external disturbance and therefore to improve the accuracy of the forecast. Numerical solutions of linear robust control problems of small state dimension can be accurately obtained using traditional methods such as the Riccati solver [5]. However, for linear problems with large state dimension or nonlinear problems in fluid mechanics, methods that require the storage of a large quantity of fields become less appropriate; as suggested in [1], iterative methods seem to be very attractive for these types of problems. Iterative methods based on repeated computations of the adjoint flow have been proven to be very successful in numerical solutions of optimal control problems [1, 15, 17, 18]. The analysis of stability and convergence of a class of such methods is given in [17, 18, 28]. In [6], the authors proposed some iterative algorithms based on repeated computations of the adjoint flow for the numerical solution of a class of robust control problems. In [22], the author proved the convergence of the iterative methods proposed in [6] for the linearized problems. The nonlinear problem was considered in [21], in which the author proved the convergence of the iterative methods proposed in [6] when the cost parameters (γ, l) are large enough and for only one particular choice of the iteration parameter. Although the result of [21] is interesting because the iteration parameters are given explicitly, it is not physically reasonable that the condition (γ, l) be large enough. In fact, let us recall that l large enough corresponds to prohibitive control. In this article, we revisit the method in [21], and we prove the convergence of the adjoint-based iterative method of [6] for γ large enough and $l > 0$.

The article is divided as follows. In the next section, we present the mathematical models. The third section recalls from [6] the robust control problems and proves the existence and uniqueness of their solutions using a contraction mapping argument. This section ends by proving the convergence of the algorithm of [6] under a weaker assumption than those in [21]. The fourth section presents some numerical simulations of a robust control problem related to data assimilation in oceanography in order to illustrate the algorithm.

2. Governing equations and mathematical setting.

2.1. Governing equations. Hereafter, Ω is an open domain in R^2. The flow U and the forcing F satisfy the Navier–Stokes equations in $\Omega \times (0, \infty)$ given by

$$\begin{cases} \dfrac{\partial U}{\partial t} - \nu \Delta U + (U \cdot \nabla) U + \nabla P = F, \\ \text{div } U = 0, \\ U = 0 \text{ on } \partial\Omega, \\ U = U_0 \text{ at } t = 0. \end{cases} \tag{2.1}$$

A stationary or nonstationary solution U with the corresponding forcing F will be referred to as target flow for the control problem. If no target flow is given, U and F are taken as zero. We are interested in the robust regulation of the deviation of the flow from the target (U, F). As in [6], we will consider the control of the linearized equation, which models small perturbations (u, f) to the target flow (U, F) with Dirichlet boundary conditions and known initial conditions such that, in $\Omega \times (0, \infty)$, we have

$$\frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) U + (U \cdot \nabla) u + \nabla p = f, \quad u = 0 \text{ on } \partial\Omega, \ u = g \text{ at } t = 0. \tag{2.2}$$

We will also consider the control of the full nonlinear equation which models large perturbations (u, f) to the target flow (U, F) such that, in $\Omega \times (0, \infty)$, we have

(2.3)
$$\frac{\partial u}{\partial t} - \nu\Delta u + (u \cdot \nabla)U + (U \cdot \nabla)u + (u \cdot \nabla)u + \nabla p = f, \quad u = 0 \text{ on } \partial\Omega, \ u = g \text{ at } t = 0.$$

We will generalize our results to control problems related to data assimilation in oceanography.

2.2. Mathematical setting. Let Ω be a bounded open set of R^2 with boundary $\partial\Omega$, and let $\vec{n}$ be the unit outward normal vector to $\partial\Omega$. We denote by $H^s(\Omega)$, for $s \in R$, the Sobolev spaces constructed on $L^2(\Omega)$, and by $H_0^s(\Omega)$, for $s > 1/2$, the closure of $C_0^\infty(\Omega)$ in $H^s(\Omega)$. As in [6, 30], we set $M = \{u \in (C_0^\infty)^2;\ \text{div } u = 0\}$ and denote by H (resp., V) the closure of M in $(L^2(\Omega))^2$ (resp., $(H^1(\Omega))^2$); we have

$$H = \{u \in (L^2(\Omega))^2;\ \text{div } u = 0 \text{ in } \Omega,\ u \cdot \vec{n} = 0 \text{ on } \partial\Omega\}$$

and

$$V = \{u \in (H_0^1(\Omega))^2;\ \text{div } u = 0 \text{ in } \Omega\},$$

where $\vec{n}$ denotes the outward normal vector to $\partial\Omega$. The scalar product in H is denoted by $(u, v) = \int_\Omega u \cdot v dx$, that on V is denoted by $((u, v)) = \int_\Omega \nabla u \cdot \nabla v dx$, and the associated norms are denoted by $|\cdot|$ and $\|\cdot\|$, respectively. Hereafter, if Z is any other Hilbert space, we will denote by $\langle\cdot,\cdot\rangle_Z$ the scalar product in Z and by $\|\cdot\|_Z$ the associated norm. We also set $Y = L^2(0, T; V)$ and $W = \{u \in L^2(0, T; V), \frac{du}{dt} \in L^2(0, T; V')\}$ endowed with the norm

$$\|u\|_W = \left(\|u\|^2_{L^2(0,T;V)} + \left\| \frac{du}{dt} \right\|^2_{L^2(0,T;V')} \right)^{\frac{1}{2}}.$$

We denote by A the Stokes operator, defined as an isomorphism from V onto the dual V' of V such that, for $u \in V$, Au is defined by

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

where $\langle\cdot,\cdot\rangle_{V',V}$ is the duality bracket between V' and V. The operator A is extended to H as a linear unbounded operator with domain $D(A) = (H^2(\Omega))^2 \cap V$ when $\partial\Omega$ is a C^2 surface. We also denote by $\mathcal{P}$ the Leray–Hopf projector, which is the orthogonal projector of the space $\mathbf{L}^2(\Omega) \equiv (L^2(\Omega))^2$ onto the divergence-free space H. The Stokes operator is related to $\mathcal{P}$ by

$$Au = -\mathcal{P}(\Delta u) \quad \forall u \in D(A).$$

We also define the bilinear mapping B by

$$B(u, v) = \mathcal{P}\left((u \cdot \nabla)v\right) \quad \forall u, v \in V,$$

which is a bilinear mapping from V into V'. To simplify the notation, we define the nonlinear application B from V into V' by

$$B(u) = B(u, u) \quad \forall u \in V. \tag{2.4}$$

Then B is differentiable, and we have

$$B'(u)v = B(u,v) + B(v,u) \quad \forall u, v \in V. \tag{2.5}$$

We will denote by $B'(u)^*$ the adjoint operator of $B'(u)$ for the duality between V and V'. Define a continuous trilinear form b on V such that, with $u, v, w \in (H^1(\Omega))^2$, we have

$$b(u,v,w) = \langle B(u,v), w\rangle_{V',V} = \int (u \cdot \nabla)v \cdot w dx. \tag{2.6}$$

Then the following properties hold true (see [6]):

$$\begin{aligned}
&b(u,v,v) = 0 \quad \forall u \in V, v \in V,\\
&|b(u,v,w)| \le c|u|_{L^2}^{\frac{1}{2}}\|u\|^{\frac{1}{2}}\|v\|^{\frac{1}{2}}|Av|_{L^2}^{\frac{1}{2}}|w|_{L^2} \quad \forall u \in V, v \in D(A), w \in H,\\
&|b(u,v,w)| \le c|u|_{L^2}^{\frac{1}{2}}|Au|_{L^2}^{\frac{1}{2}}\|v\||w|_{L^2} \quad \forall u \in D(A), v \in V, w \in H,\\
&|b(u,v,w)| \le c|u|_{L^2}^{\frac{1}{2}}\|u\|^{\frac{1}{2}}\|v\||w|_{L^2}^{\frac{1}{2}}\|w\|^{\frac{1}{2}} \quad \forall u \in V, v \in V, w \in V,
\end{aligned} \tag{2.7}$$

where $c = c(\Omega)$ is a constant depending only on Ω. The estimates developed in this work involve integration by parts and the following inequalities, which are repeated here for the sake of clarity: the Cauchy–Schwarz inequality $|\langle u, v\rangle| \le |u|_{L^2}|v|_{L^2}$; Hölder's inequality

$$\int fg dx \le \left(\int |f|^p\right)^{\frac{1}{p}} \left(\int |g|^q\right)^{\frac{1}{q}}, \qquad \frac{1}{p} + \frac{1}{q} = 1; \tag{2.8}$$

the Poincaré inequality $|u|_{L^2} \le c\|u\|$; Young's inequality

$$ab \le \frac{\epsilon}{p}a^p + \frac{\epsilon^{-q/p}}{q}b^q, \qquad 1 < p < \infty,\ \frac{1}{p} + \frac{1}{q} = 1; \tag{2.9}$$

and Gronwall's lemma

$$\begin{aligned}
&\frac{dy}{dt} \le gy + h \quad \forall t \ge 0\\
&\quad \Longrightarrow y(t) \le y(0)\exp\left(\int_0^t g(\tau)d\tau\right) + \int_0^t h(s)\exp\left(\int_s^t g(\tau)d\tau\right)ds \quad \forall t \ge 0.
\end{aligned} \tag{2.10}$$

Using the operators A and B, the linearized Navier–Stokes equations (2.2) become

$$\begin{cases}
\dfrac{du}{dt} + \nu Au + B(U,u) + B(u,U) = \mathcal{P}f,\\
u \in V,\\
u = g \text{ at } t = 0,
\end{cases} \tag{2.11}$$

where the regularity required on f, g is

$$f \in L^2(0,T;\mathbf{L}^2(\Omega)) \quad \forall T > 0,\ g \in V; \quad U \in L^\infty(0,T;V) \cap L^2(0,T;D(A)). \tag{2.12}$$

Similarly, application of the Leray projector to the fully nonlinear problem (2.3) gives (with $\tau = 1$)

$$\begin{cases} \dfrac{du}{dt} + \nu Au + B(U,u) + B(u,U) + \tau B(u,u) = \mathcal{P}f, \\ u \in V, \\ u = g \text{ at } t = 0. \end{cases} \tag{2.13}$$

Hereafter, $\tau \in R$ is a nondimensional parameter.

3. Robust control.

3.1. Body force control framework. Following the framework presented in [6], the interior forcing f is decomposed into a disturbance $\psi \in L^2(0,T;\mathbf{L}^2(\Omega))$ and a control $\phi \in L^2(0,T;\mathbf{L}^2(\Omega))$, with $T > 0$. Thus, we write

$$f = B_1\psi + B_2\phi,$$

where B_1 and B_2 are bounded operators on $(L^2(\Omega))^2$. We will also write

$$\mathcal{P}f = \mathcal{B}_1\psi + \mathcal{B}_2\phi,$$

where $\mathcal{B}_1 = \mathcal{P}B_1$ and $\mathcal{B}_2 = \mathcal{P}B_2$ are mappings from $\mathbf{L}^2(\Omega)$ to H. In this section, the cost functional is given by

$$\begin{aligned} \mathcal{J}(\psi,\phi) = {} & \frac{1}{2}\int_0^T |\mathcal{C}_1 u|_{L^2}^2\, dt + \frac{1}{2}\,|\mathcal{C}_2 u(T)|_{L^2}^2 - \frac{1}{2}\int_0^T \left\langle \mathcal{C}_3 \nu \frac{\partial u}{\partial n}, \vec{r} \right\rangle_{\mathbf{L}^2(\partial\Omega)} dt \\ & + \frac{1}{2}\int_0^T \left[l^2\,|\phi|_{L^2}^2 - \gamma^2\,|\psi|_{L^2}^2 \right] dt, \end{aligned} \tag{3.1}$$

where the scalar control parameters γ and l are given, $\vec{r}$ is a known vector field on $\partial\Omega$, $\vec{n}$ is the unit normal vector to $\partial\Omega$, and $\mathcal{C}_3^*\vec{r}\cdot\vec{n} = 0$. The operators $\mathcal{C}_1$ and $\mathcal{C}_2$ are unbounded linear operators on $\mathbf{L}^2(\Omega)$ satisfying

$$|\mathcal{C}_i v|_{\mathbf{L}^2(\Omega)} \le c\|v\| \quad \text{for } i = 1,2, \quad \forall v \in V, \tag{3.2}$$

and $\mathcal{C}_3$ is a bounded linear operator of $\mathbf{L}^2(\partial\Omega)$. The reader is referred to [6] for more details. Some particular interesting cases are

- $\mathcal{C}_1 = d_1 I$ and $\mathcal{C}_2 = \mathcal{C}_3 = 0 \Longrightarrow$ regulation of the turbulent kinetic energy;
- $\mathcal{C}_1 = d_2\nabla\times$ and $\mathcal{C}_2 = \mathcal{C}_3 = 0 \Longrightarrow$ regulation of the square of the vorticity;
- $\mathcal{C}_2 = d_3 I$ and $\mathcal{C}_1 = \mathcal{C}_3 = 0 \Longrightarrow$ terminal control of the turbulent kinetic energy;
- $\mathcal{C}_3 = d_4 I$ and $\mathcal{C}_1 = \mathcal{C}_2 = 0 \Longrightarrow$ minimization of the time-average skin-friction in the direction of the vector $\vec{r}$ integrated over the boundary of the domain.

We consider the following robust control problem, referred to as Problem I.

Problem I. Find $(\bar{\psi},\bar{\phi}) \in L^2(0,T;\mathbf{L}^2(\Omega)) \times L^2(0,T;\mathbf{L}^2(\Omega))$ such that

$$\mathcal{J}(\bar{\psi},\bar{\phi}) = \min_{\phi\in D}\sup_{\psi\in X} \mathcal{J}(\psi,\phi) = \max_{\psi\in X}\inf_{\phi\in D} \mathcal{J}(\psi,\phi),$$

subject to the Navier–Stokes equations (2.13), where $X = D = L^2(0,T;\mathbf{L}^2(\Omega))$. With Problem I we associate the following coupled system (3.3)–(3.5):

$$\begin{cases} \dfrac{du}{dt} + \nu Au + B(U,u) + B(u,U) + \tau B(u,u) = \mathcal{B}_1\psi + \mathcal{B}_2\phi, \\ u \in V, \\ u = g \text{ at } t = 0, \end{cases} \tag{3.3}$$

$$\text{(3.4)}\quad \begin{cases} -\dfrac{d\tilde{u}}{dt} + \nu A^*\tilde{u} + B'(U)^*\tilde{u} + \tau B'(u)^*\tilde{u} = \mathcal{C}_1^*\mathcal{C}_1 u, \\ \tilde{u}(t) \in V_r = \left\{ v \in (H^1(\Omega))^2;\ \text{div } v = 0 \text{ in } \Omega,\ v = \mathcal{C}_3^*\vec{r} \text{ on } \partial\Omega \right\},\ t < T, \\ \tilde{u}(T) = \mathcal{C}_2^*\mathcal{C}_2 u(T) \in H, \end{cases}$$

$$\text{(3.5)}\quad \begin{cases} \gamma^2\psi - \mathcal{B}_1^*\tilde{u} = 0, \\ l^2\phi + \mathcal{B}_2^*\tilde{u} = 0, \end{cases}$$

where A^* is the unbounded operator on $H \cap (H^1(\Omega))^2$ defined by

$$\text{(3.6)}\quad \langle u', A^*\tilde{u}\rangle = \langle Au', \tilde{u}\rangle + \left\langle \frac{\partial u'}{\partial n}, \tilde{u} \right\rangle_{\mathbf{L}^2(\partial\Omega)} \quad \text{for } u' \in D(A),\ \tilde{u} \in H \cap (H^1(\Omega))^2.$$

In [6], the authors proposed an iterative method for the numerical solution of Problem I. This method is based on a repeated computation of the state equation (3.3) and the adjoint equation (3.4). The purpose of this article is to prove the convergence of the algorithm proposed in [6]. To achieve that goal, we rewrite the system (3.3)–(3.5) as a fixed-point problem. Then, using a contraction mapping argument, we prove the convergence of the algorithm, and we obtain its rate of convergence as well.

Linear case. Hereafter, we assume that $\mathcal{C}_3 = 0$. We consider the problem (3.3)–(3.5) in which $\tau = 0$. For $\tau = 0$ and $\mathcal{C}_3 = 0$, we obtain the following system:

$$\text{(3.7)}\quad \begin{cases} \dfrac{du}{dt} + \nu Au + B(U,u) + B(u,U) = \mathcal{B}_1\psi + \mathcal{B}_2\phi, \\ u \in V, \\ u = g \text{ at } t = 0, \end{cases}$$

$$\text{(3.8)}\quad \begin{cases} -\dfrac{d\tilde{u}}{dt} + \nu A^*\tilde{u} + B'(U)^*\tilde{u} = \mathcal{C}_1^*\mathcal{C}_1 u, \\ \tilde{u}(t) \in V,\ t < T, \\ \tilde{u}(T) = \mathcal{C}_2^*\mathcal{C}_2 u(T) \in H, \end{cases}$$

$$\text{(3.9)}\quad \begin{cases} \gamma^2\psi - \mathcal{B}_1^*\tilde{u} = 0, \\ l^2\phi + \mathcal{B}_2^*\tilde{u} = 0. \end{cases}$$

The following result was proven in [6].

PROPOSITION 3.1. *If γ is large enough, then* (3.7)–(3.9) *has a unique solution.*

The purpose of this part is to prove the existence and uniqueness of solutions to (3.7)–(3.9) using a contraction mapping method, and therefore we derive an iterative method (fixed-point iteration method) for the numerical solution of (3.7)–(3.9). Let $\rho > 0$ be a constant, and consider the operator G defined from $\mathbf{Z} \equiv L^2(0,T;\mathbf{L}^2(\Omega)) \times L^2(0,T;\mathbf{L}^2(\Omega))$ into itself by

$$\text{(3.10)}\quad G(\psi,\phi) = (\psi,\phi) - \rho\,(\gamma^2\psi - \mathcal{B}_1^*\tilde{u}, l^2\phi + \mathcal{B}_2^*\tilde{u}),$$

where $\tilde{u}$ is given by (3.7)–(3.8). It is easy to check that (ψ,ϕ) is a fixed point of G if and only if (ψ,ϕ) solves (3.7)–(3.8). The following result holds true.

PROPOSITION 3.2. *Assume that γ is large enough. Then there exists $\rho_0 > 0$ such that G is a contraction from $\mathbf{Z}$ into itself for $\rho < \rho_0$.*

Proof. Let $(\psi_i, \phi_i) \in \mathbf{Z} = L^2(0,T;\mathbf{L}^2(\Omega)) \times L^2(0,T;\mathbf{L}^2(\Omega))$, $i = 1, 2$. Let $(\psi, \phi) = (\psi_1, \phi_1) - (\psi_2, \phi_2)$, $u = u_1 - u_2$, $\tilde{u} = \tilde{u}_1 - \tilde{u}_2$. Then we have

$$\begin{cases} \dfrac{du}{dt} + \nu Au + B(u,U) + B(U,u) = \mathcal{B}_1\psi + \mathcal{B}_2\phi, \\ u \in V, \\ u = 0 \text{ at } t = 0, \end{cases} \tag{3.11}$$

$$\begin{cases} -\dfrac{d\tilde{u}}{dt} + \nu A^*\tilde{u} + B'(U)^*\tilde{u} = \mathcal{C}_1^*\mathcal{C}_1 u, \\ \tilde{u}(t) \in V,\ t < T, \\ \tilde{u}(T) = \mathcal{C}_2^*\mathcal{C}_2 u(T) \in H. \end{cases} \tag{3.12}$$

If we set $G(\psi,\phi) = G(\psi_1,\phi_1) - G(\psi_2,\phi_2)$, then we have

$$\begin{aligned} \|G(\psi,\phi)\|_{\mathbf{Z}}^2 = {} & \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds - 2\rho\gamma^2 \int_0^T |\psi(s)|^2 ds - 2\rho l^2 \int_0^T |\phi(s)|^2 ds \\ & + 2\rho \int_0^T (\psi, \mathcal{B}_1^*\tilde{u}) ds - 2\rho \int_0^T (\phi, \mathcal{B}_2^*\tilde{u}) ds + \rho^2 \int_0^T |\gamma^2\psi - \mathcal{B}_1^*\tilde{u}|^2 ds \\ & + \rho^2 \int_0^T |l^2\phi + \mathcal{B}_2^*\tilde{u}|^2 ds. \end{aligned} \tag{3.13}$$

Thanks to (3.11) and (3.12), we have

$$-2\rho \int_0^t (\phi, \mathcal{B}_2^*\tilde{u}) ds = 2\rho \int_0^t (\psi, \mathcal{B}_1^*\tilde{u}) ds - 2\rho \int_0^t |\mathcal{C}_1 u(s)|^2 ds - 2\rho |\mathcal{C}_2 u(T)|^2. \tag{3.14}$$

Reporting (3.14) in (3.13), we get

$$\begin{aligned} \|G(\psi,\phi)\|_{\mathbf{Z}}^2 = {} & \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds - 2\rho\gamma^2 \int_0^T |\psi(s)|^2 ds - 2\rho l^2 \int_0^T |\phi(s)|^2 ds \\ & + 4\rho \int_0^T (\psi, \mathcal{B}_1^*\tilde{u}) ds - 2\rho \int_0^T |\mathcal{C}_1 u(s)|^2 ds \\ & - 2\rho |\mathcal{C}_2 u(T)|^2 + \rho^2 \int_0^T |\gamma^2\psi - \mathcal{B}_1^*\tilde{u}|^2 ds + \rho^2 \int_0^T |l^2\phi + \mathcal{B}_2^*\tilde{u}|^2 ds. \end{aligned} \tag{3.15}$$

Using Young's inequality and the continuity of the operators $\mathcal{B}_1^*$ and $\mathcal{B}_2^*$ in (3.15), we obtain the estimate

$$\begin{aligned} \|G(\psi,\phi)\|_{\mathbf{Z}}^2 \le {} & (1 - 2\rho\gamma^2 + \tfrac{2\rho}{\varepsilon} + 2\rho^2\gamma^4) \int_0^T |\psi(s)|^2 ds + 2\rho c_1 \varepsilon \int_0^T |\tilde{u}(s)|^2 ds \\ & + (1 - 2\rho l^2 + 2\rho^2 l^4) \int_0^T |\phi(s)|^2 ds + 4\rho^2 c_1 \int_0^T |\tilde{u}(s)|^2 ds - 2\rho \int_0^T |\mathcal{C}_1 u(s)|^2 ds \\ & - 2\rho |\mathcal{C}_2 u(T)|^2, \end{aligned} \tag{3.16}$$

where here and in what follows, c_1 denotes different positive constants depending only on Ω, U, and T.

To complete the proof of this proposition, it remains to conveniently estimate the integrals involving $\tilde{u}$.

From (3.12), by applying the Cauchy–Schwarz, Poincaré, and Young's inequalities, we obtain that

$$\begin{aligned} -\tfrac{1}{2}\tfrac{d}{dt}|\tilde{u}|^2 + \nu\|\tilde{u}\|^2 &\le |b(\tilde{u},U,\tilde{u})| + c|\mathcal{C}_1 u|\|\tilde{u}\| \\ &\le c|\mathcal{C}_1 u|^2 + \tfrac{\nu}{4}\|\tilde{u}\|^2 + c\|U\||\tilde{u}|\|\tilde{u}\|, \end{aligned} \tag{3.17}$$

which gives

$$-\tfrac{1}{2}\tfrac{d}{dt}|\tilde{u}|^2 + \nu\|\tilde{u}\|^2 \le c|\mathcal{C}_1 u|^2 + \tfrac{\nu}{2}\|\tilde{u}\|^2 + c\|U\|^2|\tilde{u}|^2. \tag{3.18}$$

Gronwall's inequality gives

$$\begin{aligned} |\tilde{u}(t)|^2 &\le e^{N_0(t)}|\tilde{u}(T)|^2 + ce^{N_0(t)}\int_t^T |\mathcal{C}_1 u(s)|^2 ds \\ &\le c_1|\mathcal{C}_2 u(T)|^2 + c_1\int_0^T |\mathcal{C}_1 u(s)|^2 ds, \end{aligned} \tag{3.19}$$

where $N_0(t) = c\int_t^T \|U(s)\|^2 ds$ and c is a constant depending only on Ω.

From (3.11), we also have (see [6])

$$\begin{aligned} |u(t)|^2 &\le ce^{M_0(t)}\int_0^t \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds \\ &\le c_1\int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds, \end{aligned} \tag{3.20}$$

where $M_0(t) = c\int_0^t \|U(s)\|^2 ds$. Thanks to (3.20) and the continuity of the operators $\mathcal{C}_1$ and $\mathcal{C}_2$, we derive from (3.19) that

$$|\tilde{u}(t)|^2 \le c_1\int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds. \tag{3.21}$$

Reporting (3.19) and (3.21) in (3.16), we find

$$\begin{aligned} \|G(\psi,\phi)\|_{\mathbf{Z}}^2 &\le (1 - 2\rho\gamma^2 + \tfrac{2c_1\rho}{\varepsilon} + 2\rho^2\gamma^4)\int_0^T |\psi(s)|^2 ds + 2\rho c_1\varepsilon\int_0^T |\mathcal{C}_1 u(s)|^2 ds \\ &+ 2\rho c_1\varepsilon|\mathcal{C}_2 u(T)|^2 + (1 - 2\rho l^2 + 2\rho^2 l^4)\int_0^T |\phi(s)|^2 ds + 4\rho^2 c_1\int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds \\ &- 2\rho\int_0^T |\mathcal{C}_1 u(s)|^2 ds - 2\rho|\mathcal{C}_2 u(T)|^2. \end{aligned} \tag{3.22}$$

Choosing $\varepsilon = c_1^{-1}$ in (3.22) yields

$$\begin{aligned} \|G(\psi,\phi)\|_{\mathbf{Z}}^2 &\le (1 - 2\rho\gamma^2 + 2c_1\rho + 2\rho^2(\gamma^4 + 2c_1))\int_0^T |\psi(s)|^2 ds \\ &\quad + (1 - 2\rho l^2 + 2\rho^2(l^4 + 2c_1))\int_0^T |\phi(s)|^2 ds \\ &\le \max\{1 - 2\rho\gamma^2 + 2c_1\rho + 2\rho^2(\gamma^4 + 2c_1), 1 - 2\rho l^2 + 2\rho^2(l^4 + 2c_1)\} \\ &\quad \times \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds. \end{aligned} \tag{3.23}$$

It is then obvious to check that G is a contraction if γ is large enough and $\rho < \min\{\frac{\gamma^2-c_1}{\gamma^4+2c_1}, \frac{l^2}{l^4+2c_1}\}$; this completes the proof of Proposition 3.2. ∎

Nonlinear case. In this part we study the nonlinear problem, that is (3.3)–(3.5) with $\tau = 1$. For $(\psi, \phi) \in \mathbf{Z} \equiv L^2(0,T;\mathbf{L}^2(\Omega)) \times L^2(0,T;\mathbf{L}^2(\Omega))$, we set

$$G(\psi,\phi) = (\psi,\phi) - \rho\,(\gamma^2\psi - \mathcal{B}_1^*\tilde{u}, l^2\phi + \mathcal{B}_2^*\tilde{u}), \tag{3.24}$$

where $\rho > 0$ is a constant and $\tilde{u}$ is given by (3.3)–(3.4). We will prove that G is a contraction defined on an appropriate subset $\mathcal{X}$ of $\mathbf{Z}$.

Some a priori estimates. Now let $R > 0$ and

$$\mathcal{X}(R) = \left\{(\psi,\phi) \in \mathbf{Z};\ \int_0^T (|\psi(s)|_{L^2}^2 + |\phi(s)|_{L^2}^2)ds \le R^2\right\}.$$

Hereafter we will denote by c_1 a positive constant that depends on Ω, T, and U and whose values may be different in each inequality. Multiplying $(3.3)_1$ by u and applying the Cauchy–Schwarz, Poincaré, and Young's inequalities, we obtain (see [6])

$$\begin{aligned} &|u(t)|_{L^2}^2 \le e^{M_0(t)}|u_0|_{L^2}^2 + ce^{M_0(t)}\int_0^t (|\psi(s)|_{L^2}^2 + |\phi|_{L^2}^2)ds,\\ &\frac{1}{t}\int_0^t \|u(s)\|ds \le \frac{1}{t}e^{M_0(t)}|g|_{L^2}^2 + \frac{1}{t}e^{M_0(t)}\int_0^t (|\psi(s)|_{L^2}^2 + |\phi|_{L^2}^2)ds, \end{aligned} \tag{3.25}$$

where

$$M_0(t) = c\int_0^t \|U\|^2 d\tau \le c_1. \tag{3.26}$$

Now, multiplying $(3.4)_1$ by $\tilde{u}$ and applying the Cauchy–Schwarz, Poincaré, and Young's inequalities, we obtain

$$\begin{aligned} -\tfrac{1}{2}\tfrac{d}{dt}|\tilde{u}|_{L^2}^2 + \nu\|\tilde{u}\|^2 &\le |b(\tilde{u},U+u,\tilde{u})| + |b(U+u,\tilde{u},\tilde{u})| + c|\mathcal{C}_1 u|_{L^2}\|\tilde{u}\|\\ &\le c|\mathcal{C}_1 u|^2 + \tfrac{\nu}{4}\|\tilde{u}\|^2 + c|\tilde{u}|_{L^2}\|\tilde{u}\|\|U+u\|\\ &\le c|\mathcal{C}_1 u|^2 + \tfrac{\nu}{2}\|\tilde{u}\|^2 + c(\|U\|^2 + \|u\|^2)|\tilde{u}|^2, \end{aligned} \tag{3.27}$$

from which we derive

$$\begin{aligned} |\tilde{u}(t)|_{L^2}^2 &\le e^{N_0(t)}|\tilde{u}(T)|_{L^2}^2 + e^{N_0(t)}\int_t^T |u(s)|_{L^2}^2 ds\\ &\le c_1|\mathcal{C}_2 u(T)|_{L^2}^2 + c_1\int_0^T |\mathcal{C}_1 u(s)|_{L^2}^2 ds, \end{aligned} \tag{3.28}$$

where $N_0(t) = c\int_t^T (\|U(s)\|^2 + \|u(s)\|^2)ds \le c_1$.

PROPOSITION 3.3. *Assume that γ is large enough and that R and $|g|_{L^2}$ are small enough. Then there exists $\rho_1 > 0$ such that G maps $\mathcal{X}(R)$ into $\mathcal{X}(R)$ for $\rho \le \rho_1$.*

Proof. Let $(\psi,\phi) \in \mathcal{X}(R)$. We shall prove that $G(\psi,\phi) \in \mathcal{X}(R)$ for γ large enough, R, $|g|_{L^2}$, and ρ small enough. As in the linear case, it is easy to check that

(3.29)
$$\begin{aligned}\|G(\psi,\phi)\|_{\mathbf{Z}}^2 = &\int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds - 2\rho\gamma^2 \int_0^T |\psi(s)|^2 ds - 2\rho l^2 \int_0^T |\phi(s)|^2 ds \\ &+ 2\rho \int_0^T (\psi, \mathcal{B}_1^* \tilde{u}) ds - 2\rho \int_0^T (\phi, \mathcal{B}_2^* \tilde{u}) ds + \rho^2 \int_0^T |\gamma^2 \psi - \mathcal{B}_1^* \tilde{u}|^2 ds \\ &+ \rho^2 \int_0^T |l^2 \phi + \mathcal{B}_2^* \tilde{u}|^2 ds.\end{aligned}$$

It follows from (3.3) and (3.4) that

$$\begin{aligned}-2\rho \int_0^T (\phi, \mathcal{B}_2^* \tilde{u}) ds = &2\rho \int_0^T (\psi, \mathcal{B}_1^* \tilde{u}) ds - 2\rho \int_0^T |\mathcal{C}_1 u(s)|^2 ds - 2\rho |\mathcal{C}_2 u(T)|^2 \\ &+ 2\rho (g, \tilde{u}(0)) + 2\rho \int_0^T b(u,u,\tilde{u}) ds.\end{aligned} \tag{3.30}$$

Reporting (3.30) in (3.29), we get

(3.31)
$$\begin{aligned}\|G(\psi,\phi)\|_{\mathbf{Z}}^2 = &\int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds - 2\rho\gamma^2 \int_0^T |\psi(s)|^2 ds - 2\rho l^2 \int_0^T |\phi(s)|^2 ds \\ &+ 4\rho \int_0^T (\psi, \mathcal{B}_1^* \tilde{u}) ds - 2\rho \int_0^T |\mathcal{C}_1 u(s)|^2 ds + 2\rho \int_0^T b(u,u,\tilde{u}) ds \\ &- 2\rho |\mathcal{C}_2 u(T)|^2 + \rho^2 \int_0^T |\gamma^2 \psi - \mathcal{B}_1^* \tilde{u}|^2 ds + \rho^2 \int_0^T |l^2 \phi + \mathcal{B}_2^* \tilde{u}|^2 ds + 2\rho (g, \tilde{u}(0)).\end{aligned}$$

Thanks to the continuity of the operators $\mathcal{B}_1^*$ and $\mathcal{B}_2^*$, we have

$$\int_0^T |\mathcal{B}_i^* \tilde{u}(s)|^2 ds \le c \int_0^T |\tilde{u}(s)|^2 ds \le c_1 \int_0^T |\mathcal{C}_1 u(s)|^2 ds + c_1 |\mathcal{C}_2 u(T)|^2. \tag{3.32}$$

On the other hand, (2.7), (3.25), and (3.28) show that

$$\int_0^T |b(u,u,\tilde{u})| ds \le c_1 \{|g|^2 + R^2\} \left\{ \int_0^T |\mathcal{C}_1 u(s)|^2 ds + |\mathcal{C}_2 u(T)|^2 \right\}^{\frac{1}{2}}. \tag{3.33}$$

Using Young's inequality as well as (3.32) and (3.33) in (3.31), we find

(3.34)
$$\begin{aligned}\|G(\psi,\phi)\|_{\mathbf{Z}}^2 \le &\left(1 - 2\rho\gamma^2 + \frac{2c_1\rho}{\varepsilon} + 2\rho^2\gamma^4\right) \int_0^T |\psi(s)|^2 ds + 2\rho\varepsilon \int_0^T |\mathcal{C}_1 u(s)|^2 ds \\ &+ 2\rho\varepsilon |\mathcal{C}_2 u(T)|^2 + (1 - 2\rho l^2 + 2\rho^2 l^4) \int_0^T |\phi(s)|^2 ds \\ &+ 4\rho^2 c_1 \left\{ \int_0^T |\mathcal{C}_1 u(s)|^2 ds + |\mathcal{C}_2 u(T)|^2 \right\} \\ &+ \rho\eta \left\{ \int_0^T |\mathcal{C}_1 u(s)|^2 ds + |\mathcal{C}_2 u(T)|^2 \right\} + \frac{\rho}{\eta} c_1 |g|^2 \\ &+ \rho\delta \left\{ \int_0^T |\mathcal{C}_1 u(s)|^2 ds + |\mathcal{C}_2 u(T)|^2 \right\} + \frac{\rho}{\delta} c_1 (|g|^2 + R^2)^2 \\ &- 2\rho \int_0^T |\mathcal{C}_1 u(s)|^2 ds - 2\rho |\mathcal{C}_2 u(T)|^2 \quad \forall \varepsilon, \eta, \delta > 0.\end{aligned}$$

Choosing $\varepsilon = \eta = \delta = \frac{1}{4}$ and

$$\rho \leq \frac{1}{4c_1} \tag{3.35}$$

in (3.34), it follows that

(3.36)

$$\|G(\psi,\phi)\|_{\mathbf{Z}}^2 \leq (1 - 2\rho\gamma^2 + 8c_1\rho + 2\rho^2\gamma^4)\int_0^T |\psi(s)|^2 ds + 4\rho c_1|g|^2 + 4\rho c_1(|g|^2 + R^2)^2 + (1 - 2\rho l^2 + 2\rho^2 l^4)\int_0^T |\phi(s)|^2 ds.$$

We notice that, for the inequality $\|G(\psi,\phi)\|_{\mathbf{Z}} \leq R$ to hold, we shall have

$$1 - 2\rho\gamma^2 + 8c_1\rho + 2\rho^2\gamma^4 + 4\rho c_1\left\{\frac{|g|^2 + (|g|^2 + R^2)^2}{R^2}\right\} \leq 1 \tag{3.37}$$

and

$$1 - 2\rho l^2 2\rho^2 l^4 + 4\rho c_1\left\{\frac{|g|^2 + (|g|^2 + R^2)^2}{R^2}\right\} \leq 1. \tag{3.38}$$

It is obvious that, for γ large enough, (3.37) holds true with

$$\rho \leq \frac{\gamma^2 - 4c_1 - 2c_1\left\{\frac{|g|^2 + (|g|^2 + R^2)^2}{R^2}\right\}}{\gamma^4}. \tag{3.39}$$

As for (3.38), the inequality holds for l large enough; however, choosing l large enough makes the control very expensive, so for practical reasons, we will avoid this choice. The price to pay is to choose R and g small enough. We now make precise what we mean by choosing R and g small. Let $\theta \in (0,1)$, and assume that $|g|^2 \leq R^{2+\theta}$. Then (3.38) holds true, provided

$$-2l^2 + 2\rho l^4 + 4c_1\left\{R^\theta + R^2(R^\theta + 1)^2\right\} \leq 0. \tag{3.40}$$

For (3.40) to hold, it is enough to choose R and ρ so that

$$\begin{aligned} &-l^2 + 4c_1\left\{R^\theta + R^2(R^\theta + 1)^2\right\} \leq 0, \\ &-l^2 + 2\rho l^4 \leq 0. \end{aligned} \tag{3.41}$$

Since we may choose R as small as we want, this completes the proof. ∎

PROPOSITION 3.4. *Assume that γ is large enough and that R and $|g|_{L^2}$ are small enough. Then there exists $\rho_2 > 0$ such that G is a contraction from $\mathcal{X}(R)$ into $\mathcal{X}(R)$ for $\rho < \rho_2$.*

Proof. Let $(u_i, \tilde{u}_i)$ be the solution of (3.3)–(3.4) with (ψ,ϕ) replaced by (ψ_i,ϕ_i), $i = 1,2$. Let $u = u_1 - u_2$, $\psi = \psi_1 - \psi_2$, and $\phi = \phi_1 - \phi_2$. Then $(u, \tilde{u}, \psi, \phi)$ satisfies

$$\begin{cases} \dfrac{du}{dt} + \nu Au + B(u,U) + B(U,u) + B(u_1,u) + B(u,u_2) = \mathcal{B}_1\psi + \mathcal{B}_2\phi, \\ u \in V, \\ u = 0 \text{ at } t = 0, \end{cases} \tag{3.42}$$

$$\begin{cases} -\dfrac{d\tilde{u}}{dt} + \nu A^*\tilde{u} + B'(U+u_1)^*\tilde{u} + B'(u)^*\tilde{u}_2 = \mathcal{C}_1^*\mathcal{C}_1 u, \\ \tilde{u}(t) \in V,\ t < T, \\ \tilde{u}(T) = \mathcal{C}_2^*\mathcal{C}_2 u(T) \in H, \end{cases} \tag{3.43}$$

$$\begin{cases} \gamma^2\psi - \mathcal{B}_1^*\tilde{u} = 0, \\ l^2\phi + \mathcal{B}_2^*\tilde{u} = 0. \end{cases} \tag{3.44}$$

Now set $G(\psi,\phi) = G(\psi_1,\phi_1) - G(\psi_2,\phi_2)$, where G is given as in (3.24). We shall prove that G is a contraction on $\mathcal{X}(R)$. To this end, it suffices to show that $\|G(\psi,\phi)\|_{\mathbf{Z}} \le k\|(\psi,\phi)\|_{\mathbf{Z}}$ for some $0 < k < 1$. For this purpose we need some estimates on u and $\tilde{u}$.

Multiplying $(3.42)_1$ by u, we derive

$$\begin{aligned} &|u(t)|_{L^2}^2 \le e^{M_1(t)} \int_0^t (|\psi(s)|_{L^2}^2 + |\phi|_{L^2}^2)ds, \\ &\frac{1}{t}\int_0^t \|u(s)\|^2 ds \le \frac{1}{t} e^{M_1(t)} \int_0^t (|\psi(s)|_{L^2}^2 + |\phi|_{L^2}^2)ds, \end{aligned} \tag{3.45}$$

where

$$M_1(t) = c\int_0^t (\|U(s)\|^2 + \|u_2(s)\|^2)ds \le c_1. \tag{3.46}$$

Let us notice that

$$\begin{aligned} \langle B'(U+u_1)^*\tilde{u} + B'(u)^*\tilde{u}_2, \tilde{u}\rangle_{V',V} &= b(U+u_1,\tilde{u},\tilde{u}) \\ &\quad + b(\tilde{u},U+u_1,\tilde{u}) + b(u,\tilde{u},\tilde{u}_2) + b(\tilde{u},u,\tilde{u}_2). \end{aligned} \tag{3.47}$$

Moreover,

$$\begin{aligned} |b(\tilde{u},U+u_1,\tilde{u})| &\le c|\tilde{u}|_{L^2}\|\tilde{u}\|(\|U\| + \|u_1\|) \\ &\le \tfrac{\nu}{8}\|\tilde{u}\|^2 + c(\|U\|^2 + \|u_1\|^2)|\tilde{u}|_{L^2}^2, \end{aligned} \tag{3.48}$$

$$\begin{aligned} |b(u,\tilde{u},\tilde{u}_2)| &\le c|u|_{L^2}^{\frac{1}{2}}\|u\|^{\frac{1}{2}}\|\tilde{u}\||\tilde{u}_2|_{L^2}^{\frac{1}{2}}\|\tilde{u}_2\|^{\frac{1}{2}} \\ &\le \tfrac{\nu}{8}\|\tilde{u}\|^2 + c|u|_{L^2}\|u\||\tilde{u}_2|_{L^2}\|\tilde{u}_2\|, \end{aligned} \tag{3.49}$$

$$\begin{aligned} |b(\tilde{u},u,\tilde{u}_2)| &\le c|\tilde{u}|_{L^2}^{\frac{1}{2}}\|\tilde{u}\|^{\frac{1}{2}}\|u\||\tilde{u}_2|_{L^2}^{\frac{1}{2}}\|\tilde{u}_2\|^{\frac{1}{2}} \\ &\le \tfrac{\nu}{8}\|\tilde{u}\|^2 + c|\tilde{u}|_{L^2}^{\frac{2}{3}}\|u\|^{\frac{4}{3}}|\tilde{u}_2|_{L^2}^{\frac{2}{3}}\|\tilde{u}_2\|^{\frac{2}{3}} \\ &\le \tfrac{\nu}{8}\|\tilde{u}\|^2 + c\|u\|^2|\tilde{u}_2|_{L^2} + c|\tilde{u}|_{L^2}^2\|\tilde{u}_2\|^2. \end{aligned} \tag{3.50}$$

Multiplying $(3.43)_1$ by $\tilde{u}$ and using (3.47)–(3.50), we obtain

$$\begin{aligned} -\tfrac{1}{2}\tfrac{d}{dt}|\tilde{u}|_{L^2}^2 + \nu\|\tilde{u}\|^2 &\le c|\mathcal{C}_1 u|_{L^2}^2 + \tfrac{\nu}{2}\|\tilde{u}\|^2 + c\|u\|^2|\tilde{u}_2|_{L^2} + c(\|U\|^2 + \|u_1\|^2)|\tilde{u}|_{L^2}^2 \\ &\quad + c|u|_{L^2}\|u\||\tilde{u}_2|_{L^2}\|\tilde{u}_2\| + c|\tilde{u}|_{L^2}^2\|\tilde{u}_2\|^2, \end{aligned} \tag{3.51}$$

from which we derive (using (3.45))

(3.52)
$$
\begin{aligned}
|\tilde{u}(t)|^2_{L^2} &+ \int_t^T \|\tilde{u}(s)\|^2 ds \leq e^{N_1(t)}|\mathcal{C}_2 u(T)|^2_{L^2} + e^{N_1(t)}\int_t^T |\mathcal{C}_1 u(s)|^2_{L^2} ds \\
&+ e^{N_1(t)}\int_t^T (\|u\|^2 |\tilde{u}_2|_{L^2} + |u(s)|_{L^2}\|u(s)\||\tilde{u}_2|_{L^2}\|\tilde{u}_2(s)\|) ds \\
&\leq e^{N_1(t)}|\mathcal{C}_2 u(T)|^2_{L^2} + e^{N_1(t)}\int_t^T |\mathcal{C}_1 u(s)|^2_{L^2} ds + e^{N_1(t)}\int_t^T \|u\|^2 |\tilde{u}_2|_{L^2} ds \\
&+ e^{N_1(t)} \sup_{s\in[0,T]} |u(s)|_{L^2} \int_t^T |\tilde{u}_2(s)|_{L^2} \left(\int_t^T \|u(s)\|^2 ds\right)^{\frac{1}{2}} \left(\int_t^T \|\tilde{u}_2(s)\|^2 ds\right)^{\frac{1}{2}} \\
&\leq c_1\big((R^2+|g|^2)^{\frac{1}{2}} + R^2 + |g|^2\big)\int_0^T (|\psi(s)|^2_{L^2} + |\phi(s)|^2_{L^2}) ds \\
&+ c_1 \int_0^T |\mathcal{C}_1 u(s)|^2 ds + c_1 |\mathcal{C}_2 u(T)|^2,
\end{aligned}
$$

where $N_1(t) = c\int_t^T (\|U(s)\|^2 + \|u_1(s)\|^2 + \|\tilde{u}_2(s)\|^2) ds \leq c_1$.

Now, proceeding as in the proof of Proposition 3.3, we get

(3.53)
$$
\begin{aligned}
\|G(\psi,\phi)\|^2_{\mathbf{Z}} &= \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds - 2\rho\gamma^2 \int_0^T |\psi(s)|^2 ds - 2\rho l^2 \int_0^T |\phi(s)|^2 ds \\
&+ 4\rho \int_0^T (\psi, \mathcal{B}_1^* \tilde{u}) ds - 2\rho \int_0^T |\mathcal{C}_1 u(s)|^2 ds + 2\rho \int_0^T b(u,u,\tilde{u}) ds + 4\rho \int_0^T b(u,u,\tilde{u}_2) ds \\
&- 2\rho |\mathcal{C}_2 u(T)|^2 + \rho^2 \int_0^T |\gamma^2 \psi - \mathcal{B}_1^* \tilde{u}|^2 ds + \rho^2 \int_0^T |l^2 \phi + \mathcal{B}_2^* \tilde{u}|^2 ds.
\end{aligned}
$$

Thanks to Young's inequality and the continuity of the operators $\mathcal{B}_1^*$ and $\mathcal{B}_2^*$, it follows from (3.53) that

(3.54)
$$
\begin{aligned}
\|G(\psi,\phi)\|^2_{\mathbf{Z}} &\leq \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds - 2\rho\gamma^2 \int_0^T |\psi(s)|^2 ds - 2\rho l^2 \int_0^T |\phi(s)|^2 ds \\
&+ \tfrac{4\rho c_1}{\varepsilon} \int_0^T |\psi|^2 ds + (\varepsilon\rho + 4\rho^2 c_1) \int_0^T |\tilde{u}|^2 ds - 2\rho \int_0^T |\mathcal{C}_1 u(s)|^2 ds + 2\rho \int_0^T b(u,u,\tilde{u}) ds \\
&+ 4\rho \int_0^T b(u,u,\tilde{u}_2) ds - 2\rho |\mathcal{C}_2 u(T)|^2 + 2\rho^2\gamma^4 \int_0^T |\psi|^2 ds \\
&+ 2\rho^2 l^4 \int_0^T |\phi|^2 ds.
\end{aligned}
$$

On the other hand, using (3.45) and (3.52), we get

(3.55)
$$
\begin{aligned}
\int_0^T |b(u,u,\tilde{u})| ds &\leq c_1 \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds \Bigg\{ \int_0^T |\mathcal{C}_1 u(s)|^2 ds + |\mathcal{C}_2 u(T)|^2 \\
&\qquad + \left[(R^2+|g|^2) + (R^2+|g|^2)^{\frac{1}{2}}\right] \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds \Bigg\}^{\frac{1}{2}} \\
&\leq c_1 R \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds \quad \text{for } R \text{ small enough.}
\end{aligned}
$$

Similarly, one can show that

$$\int_0^T |b(u,u,\tilde{u}_2)|ds \le c_1 R \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds. \tag{3.56}$$

Reporting (3.55) and (3.56) in (3.53), using (3.52), and choosing $\varepsilon = \frac{2}{c_1}$, we find

(3.57)

$$\begin{aligned}\|G(\psi,\phi)\|_{\mathbf{Z}}^2 \le & \int_0^T \left(|\psi(s)|^2 + |\phi(s)|^2\right) ds \\ & + (-2\rho\gamma^2 + 2\rho c_1 + \tfrac{2\rho}{c_1}R + 2\rho^2\gamma^4 + 4\rho^2 c_1 + 6\rho R c_1) \int_0^T |\psi(s)|^2 ds \\ & + (-2\rho l^2 + 2\rho^2 l^4 + 4\rho^2 c_1 + \tfrac{2\rho}{c_1}R + 6\rho R c_1) \int_0^T |\phi(s)|^2 ds.\end{aligned}$$

One easily derives from (3.57) that, for G to be a contraction, one shall have

$$1 - 2\rho\gamma^2 + 2\rho c_1 + \frac{2\rho}{c_1}R + 2\rho^2\gamma^4 + 4\rho^2 c_1 + 6\rho R c_1 < 1 \tag{3.58}$$

and

$$1 - 2\rho l^2 + 2\rho^2 l^4 + 4\rho^2 c_1 + \frac{2\rho}{c_1}R + 6\rho R c_1 < 1. \tag{3.59}$$

It is obvious that, for γ large enough, (3.58) holds true with

$$\rho < \frac{\gamma^2 - c_1 - 3c_1 R - \frac{R}{c_1}}{\gamma^4 + 2c_1}. \tag{3.60}$$

As for (3.59), it is easily seen that it holds for R and ρ satisfying

$$\rho < \frac{l^2}{2(l^4 + 2c_1)}, \qquad R < \frac{l^2}{2(3c_1 + \frac{2}{c_1})}, \tag{3.61}$$

which completes the proof of Proposition 3.4. ∎

PROPOSITION 3.5. *Under the hypotheses of Propositions* 3.3 *and* 3.4*, the system* (3.3)–(3.5) *has a unique solution* $(u,\tilde{u},\psi,\phi)$ *for which* $(\psi,\phi) \in \mathcal{X}(R)$.

Proof. It follows from Propositions 3.3 and 3.4 and the fact that if (ψ,ϕ) is a fixed point of G, then $(u,\tilde{u},\psi,\phi)$ is a solution to (3.3)–(3.5), where $(u,\tilde{u})$ is given by (3.3)–(3.4). ∎

3.2. Data assimilation. A classical control problem arising in meteorology and oceanography in relation to data assimilation is the adjustment of initial conditions in order to obtain a flow that agrees with a desired target flow (i.e., the observations). Chaotic problems, such as weather system, are highly susceptible to the small disturbance present in all physical systems. Given a set of measurements of some actual flow $\mathbf{v}$ on $[0,T]$, the problem is to determine a "best" estimate as to the initial state of the model u that leads to the observed system behavior, while simultaneously forcing the model with the worst-case disturbance which perturbs u away from the observed system behavior $\mathbf{v}$. Define $w = u - \mathbf{v}$ as the amount by which the estimated flow

u differs from the observed flow $\mathbf{v}$. The cost function considered for this problem is given by

$$\begin{aligned}\mathcal{J}(\psi,\phi) = \frac{1}{2}\int_0^T |\mathcal{C}_1 w|_{L^2}^2\,dt + \frac{1}{2}\,|\mathcal{C}_2 w(T)|_{L^2}^2 + \frac{1}{2}\int_0^T \left\|\mathcal{C}_3\nu\frac{\partial w}{\partial n}\right\|_{\mathbf{L}^2(\partial\Omega)}^2 dt \\ + \frac{l^2}{2}\,|\phi|_{L^2}^2 - \gamma^2\int_0^T |\psi|_{L^2}^2\,dt,\end{aligned} \tag{3.62}$$

where $\mathcal{C}_3^*\mathcal{C}_3\nu(\partial w/\partial n)\cdot\vec{n} = 0$. The measurements of the actual flow $\mathcal{C}_1\mathbf{v}$, $\mathcal{C}_2\mathbf{v}(T)$, and $\mathcal{C}_3\nu(\partial\mathbf{v}/\partial n)|_{\partial\Omega}$ are assumed to be given. More detail on the functional $\mathcal{J}$ is given in [6]. The results given in this subsection are generalizations of the ones given in the previous subsection; therefore we will omit the proofs. We consider the following robust control problem, related to data assimilation and referred to as Problem II.

Problem II. Find $(\bar{\psi},\bar{\phi}) \in L^2(0,T;\mathbf{L}^2(\Omega)) \times \mathbf{L}^2(\Omega)$ such that

$$\mathcal{J}(\bar{\psi},\bar{\phi}) = \min_{\phi\in D}\sup_{\psi\in X}\mathcal{J}(\psi,\phi) = \max_{\psi\in X}\inf_{\phi\in D}\mathcal{J}(\psi,\phi),$$

subject to the Navier–Stokes equation

$$\begin{cases} \dfrac{du}{dt} + \nu Au + B(U,u) + B(u,U) + \tau B(u,u) = \mathcal{B}_1\psi, \\ u \in V, \\ u = \mathcal{B}_2\phi \text{ at } t = 0, \end{cases} \tag{3.63}$$

where $X = L^2(0,T;\mathbf{L}^2(\Omega))$ and $D = \mathbf{L}^2(\Omega)$. Hereafter, we assume that $\mathcal{C}_3 \equiv 0$. To Problem II we associate the following coupled system:

$$\begin{cases} \dfrac{du}{dt} + \nu Au + B(U,u) + B(u,U) + \tau B(u,u) = \mathcal{B}_1\psi, \\ u \in V, \\ u = \mathcal{B}_2\phi \text{ at } t = 0, \end{cases} \tag{3.64}$$

$$\begin{cases} -\dfrac{d\tilde{u}}{dt} + \nu A\tilde{u} + B'(U)^*\tilde{u} + \tau B'(u)^*\tilde{u} = \mathcal{C}_1^*\mathcal{C}_1(u-\mathbf{v}), \\ \tilde{u}(t) \in V,\ t < T, \\ \tilde{u}(T) = \mathcal{C}_2^*\mathcal{C}_2(u-\mathbf{v})(T) \in H, \end{cases} \tag{3.65}$$

$$\begin{cases} \gamma^2\psi - \mathcal{B}_1^*\tilde{u} = 0, \\ l^2\phi + \mathcal{B}_2^*\tilde{u}(0) = 0, \end{cases} \tag{3.66}$$

where A^* is defined by

$$\langle u', A^*\tilde{u}\rangle = \langle Au', \tilde{u}\rangle \quad \text{for } u' \in D(A) \text{ and } \tilde{u} \in V, \tag{3.67}$$

and $\mathcal{B}_2$ is a mapping from $\mathbf{L}^2(\Omega)$ to V. For the data assimilation (3.64)–(3.66), the operator G is defined from $\mathbf{Z} \equiv L^2(0,T;\mathbf{L}^2(\Omega)) \times \mathbf{L}^2(\Omega)$ into itself by

$$G(\psi,\phi) = (\psi,\phi) - \rho\left(\gamma^2\psi - \mathcal{B}_1^*\tilde{u}, l^2\phi + \mathcal{B}_2^*\tilde{u}(0)\right), \tag{3.68}$$

where ρ is a positive constant and $\tilde{u}$ is given by (3.64)–(3.65). The results obtained for the body force control extend to the data assimilation problem. In particular, if we set

$$\chi(R) = \{(\psi, \phi) \in \mathbf{Z}, \|(\psi, \phi)\|_{\mathbf{Z}} \leq R\},$$

then we have the following result.

PROPOSITION 3.6. *Assume that γ is large enough and that ρ, $\mathbf{v}$, and R are small enough. Then G is a strict contraction from $\mathcal{X}(R)$ into itself.*

Proof. The proof is completely similar to that of Proposition 3.4. ∎

3.3. Iterative method. In this section, we study an iterative method for the numerical solution of the robust control problems presented in this article. This method was proposed in [6] for the numerical solution of Problems I and II. Hereafter, we prove the convergence of the algorithm and we obtain an estimate to the convergence rate. We restrict ourselves to the nonlinear body force control problem (3.3)–(3.5), although the method also applies to the data assimilation problem (3.64)–(3.66). For the sake of clarity, we recall the algorithm proposed in [6].

ALGORITHM.

(1) Initialize $k = 0$ and $(\psi^0, \phi^0) = 0$ on $t \in [0, T]$, where k is the iteration index and (ψ^k, ϕ^k) is the numerical approximation of the disturbance and the control at the kth iteration of the algorithm.

(2) Determine the state u^k on $[0, T]$ from the state equation based on the initial conditions g with the forcing (ψ^k, ϕ^k).

(3) Determine the state $\tilde{u}^k$ on $[0, T]$ from the adjoint equation based on the state u^k.

(4) Determine the local expression of the gradients

$$\frac{\mathrm{D}\mathcal{J}}{\mathrm{D}\psi}(\psi^k, \phi^k) \quad \text{and} \quad \frac{\mathrm{D}\mathcal{J}}{\mathrm{D}\phi}(\psi^k, \phi^k).$$

(5) Determine the updated disturbance ψ^{k+1} with

$$\psi^{k+1} = \psi^k + \alpha^k \frac{\mathrm{D}\mathcal{J}}{\mathrm{D}\psi}(\psi^k, \phi^k),$$

where $0 < M_1 \leq \alpha^k \leq M_2 < 1$, where M_1 and M_2 depend on the second derivative of $\mathcal{J}$.

(6) Determine the updated control ϕ^{k+1} with

$$\phi^{k+1} = \phi^k - \beta^k \frac{\mathrm{D}\mathcal{J}}{\mathrm{D}\phi}(\psi^k, \phi^k),$$

where $0 < M_1 \leq \beta^k \leq M_2 < 1$.

(7) Increment index: $k = k + 1$. Repeat from step (2) until convergence.

Hereafter, we will prove the convergence of the previous algorithm when the iteration parameters α^k and β^k are constant and small enough.

To proceed, set

$$\mathcal{Y} = \begin{cases} L^2(0, T; L^2(\Omega)) & \text{for the linear case,} \\ \mathcal{X}(R) & \text{for the nonlinear case.} \end{cases} \tag{3.69}$$

Propositions 3.2, 3.4, and 3.6 prove that, under certain conditions, G is a strict contraction from $\mathcal{Y}$ into $\mathcal{Y}$. From the contraction mapping principle, G has a unique fixed point $(\psi, \phi) \in \mathcal{Y}$, which is a solution to the robust control problem (3.3)–(3.5). To approximate the solution (ψ, ϕ), we consider the following well known fixed-point iterative method:

$$\begin{aligned} &\text{Choose } (\psi^0, \phi^0) \in \mathcal{Y}, \\ &(\psi^{k+1}, \phi^{k+1}) = G(\psi^k, \phi^k) \quad \text{for } k \geq 0. \end{aligned} \tag{3.70}$$

Then the following results hold true.

PROPOSITION 3.7. *For γ and l large enough, the sequence (ψ^k, ϕ^k) defined by* (3.70) *converges to the unique solution $(\psi, \phi) \in \mathcal{Y}$ of the robust control problem* (3.3)–(3.5) *given by Proposition* 3.5. *Moreover, we have the following estimate of the convergence rate:*

$$\left(\|\psi^k - \psi\|_Y^2 + \|\phi^k - \phi\|_Y^2\right)^{\frac{1}{2}} \leq Q^k (1-Q)^{-1} \left(\|\psi^0 - \psi\|_Y^2 + \|\phi^0 - \phi\|_Y^2\right)^{\frac{1}{2}}, \tag{3.71}$$

where $0 < Q < 1$, with

$$Q^2 = \begin{cases} \max\{1 - 2\rho\gamma^2 + 2c_1\rho + 2\rho^2(\gamma^4 + 2c_1), 1 - 2\rho l^2 + 2\rho^2(l^4 + 2c_1)\} \\ \qquad \textit{for the linear case}, \\ \max\{1 - 2\rho\gamma^2 + 2\rho c_1 + \frac{2\rho}{c_1}R + 2\rho^2\gamma^4 + 4\rho^2 c_1 + 6\rho R c_1, \\ \qquad 1 - 2\rho l^2 + 2\rho^2 l^4 + 4\rho^2 c_1 + \frac{2\rho}{c_1}R + 6\rho R c_1\} \textit{ for the nonlinear case}, \end{cases} \tag{3.72}$$

where γ is large enough and ρ and R are small enough.

Proof. Notice that

$$\|G(\psi_1, \phi_1) - G(\psi_2, \phi_2)\|_{\mathbf{Z}} \leq Q \left(\|\psi_1 - \psi_2\|_X^2 + \|\phi_1 - \phi_2\|_X^2\right)^{\frac{1}{2}}$$

for all $(\psi_1, \phi_1), (\psi_2, \phi_2) \in \mathcal{Y}$. Therefore the proof follows directly from the contraction mapping theory [17]. ∎

Remark 3.1. Let us notice that $(\psi^{k+1}, \phi^{k+1}) = G(\psi^k, \phi^k)$ is computed as

$$\begin{cases} \dfrac{du^k}{dt} + \nu A u^k + B(U, u^k) + B(u^k, U) + B(u^k, u^k) = \mathcal{B}_1 \psi^k + \mathcal{B}_2 \phi^k, \\ u^k \in V, \\ u^k = g \text{ at } t = 0, \end{cases} \tag{3.73}$$

$$\begin{cases} -\dfrac{d\tilde{u}^k}{dt} + \nu A^* \tilde{u}^k + B'(U + u^k)^* \tilde{u}^k = \mathcal{C}_1^* \mathcal{C}_1 u^k, \\ \tilde{u}^k(t) \in V, \ t < T, \\ \tilde{u}^k(T) = \mathcal{C}_2^* \mathcal{C}_2 u^k(T) \in H, \end{cases} \tag{3.74}$$

$$\begin{cases} \psi^{k+1} = \psi^k - \rho(\gamma^2 \psi^k - \mathcal{B}_1^* \tilde{u}^k), \\ \phi^{k+1} = \phi^k - \rho(l^2 \phi^k + \mathcal{B}_2^* \tilde{u}^k). \end{cases} \tag{3.75}$$

Since (see [6])

$$\frac{DJ}{D\psi}(\psi^k, \phi^k) = \mathcal{B}_1^* \tilde{u}^k - \gamma^2 \psi^k \quad \text{and} \quad \frac{DJ}{D\phi}(\psi^k, \phi^k) = \mathcal{B}_2^* \tilde{u}^k + l^2 \phi^k, \tag{3.76}$$

it follows that (3.75) can be rewritten as

$$\left\{\begin{array}{l} \psi^{k+1} = \psi^k + \rho \frac{DJ}{D\psi}(\psi^k, \phi^k), \\ \phi^{k+1} = \phi^k - \rho \frac{DJ}{D\phi}(\psi^k, \phi^k), \end{array}\right. \tag{3.77}$$

which is exactly steps (5) and (6) of the algorithm proposed in [6]. The first to fourth steps are given by (3.73)–(3.75). Therefore, the fixed-point iteration method (3.70) is a particular case of the algorithm proposed in [6] with $\beta^k = \rho$ and $\alpha^k = \rho$. This proves the convergence of the aforementioned iterative method. Moreover, Proposition 3.7 gives an estimate of the convergence rate.

4. Numerical results. In this section, we present some numerical solutions of the data assimilation problem (3.64)–(3.66) obtained using the fixed-point iteration method described in the previous section. A major advantage of the algorithm is that it requires less storage of the fields than do conventional methods such as the Riccati solver. The method is based on repeated computations of the adjoint field (which is just as complicated as the state field). The equations considered in this section are the quasi-geostrophic equations of the ocean given by

$$\left\{\begin{array}{l} \dfrac{\partial \omega}{\partial t} - \nu \Delta \omega + J(\psi, \omega) + \beta \dfrac{\partial \psi}{\partial x} = f, \\ \omega = 0 \text{ on } \partial\Omega, \\ \omega(x, 0) = \omega_0, \end{array}\right. \tag{4.1}$$

where ψ is defined by

$$\left\{\begin{array}{l} \Delta \psi = \omega, \\ \psi = 0 \text{ on } \partial\Omega. \end{array}\right. \tag{4.2}$$

The unknown function ω represents the vorticity of the fluid, and the function ψ is the streamfunction. The constant $\beta > 0$ is the meridional gradient of the Coriolis parameter, and $\nu > 0$ is a constant. The Jacobian operator J is defined by

$$J(u, v) = \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}.$$

The system (4.1)–(4.2) has been extensively used in analytical and numerical study of ocean models [4, 7, 8, 10, 24]. This system is relatively simple as compared to other ocean models such as the primitive equations of the ocean [12, 13] or the multilayer shallow water equations in [29], but we prefer to emphasize in this article the control aspects and do not look for the most involved ocean model. Another advantage of the system (4.1)–(4.2) is that it captures the key features of large scale oceanic circulation and filters out undesired fast (high frequency) oscillations, which are not easy to handle numerically. Using the notation of [20], we can rewrite the quasi-geostrophic equations into the form

$$\left\{\begin{array}{l} \dfrac{d\omega}{dt} + \nu A\omega + B(\omega, \omega) + E\omega = f, \\ \omega \in V, \\ \omega = \omega_0 \text{ at } t = 0. \end{array}\right. \tag{4.3}$$

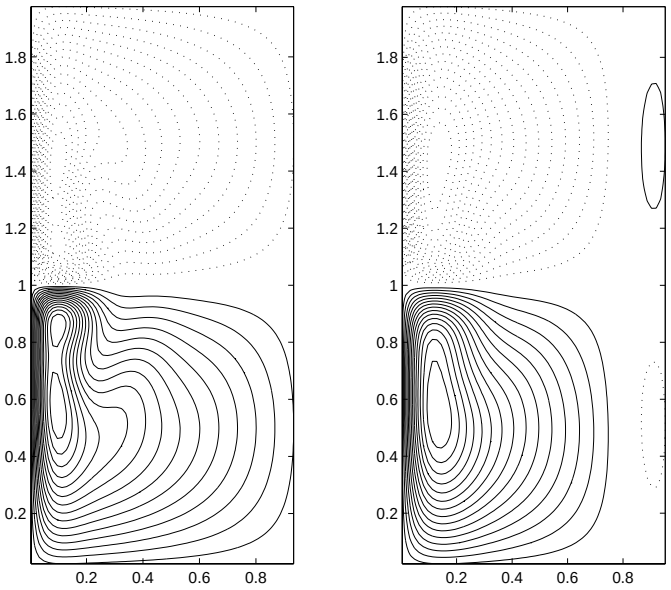


FIG. 1. *Target (streamfunction) flow and state (streamfunction) flow at* $t = T = 0.1$*, for* $\nu = 0.1$*,* $\gamma = 20.0$*,* $l = 20.0$*.*

Define $w = \omega - \mathbf{v}$ as the amount by which the estimated flow ω differs from the observed flow $\mathbf{v}$. The cost function considered in this section is given by (3.62). For simplification, the linear operators $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ are chosen as $\mathcal{C}_1 w = d_1 w$, $\mathcal{C}_2 w = d_2 w$, and $\mathcal{C}_3 = 0$, where d_1 and d_2 are constants. The functional $\mathcal{J}(f_1, f_2)$ measures the errors $\mathcal{C}_1 w$ and $\mathcal{C}_2 w$ on the interior and at the final time, respectively. To find the best estimate ω of the actual flow $\mathbf{v}$, we seek the best initial condition f_2 subject to the worst-case disturbance forcing f_1 such that $\mathcal{J}$ is minimized with respect to f_1 and maximized with respect to f_1, where ω is a solution of the quasi-geostrophic equations (4.3). More detail on the numerical results presented hereafter will appear elsewhere.

Experiment. The focus of this numerical experiment is to apply the algorithm described previously to approximate the solutions of the data assimilation problem (3.64)–(3.66). In our simulations, the (nondimensional) domain is the rectangular basin $[0, 1] \times [0, 2]$. In order to compute the solutions of (3.64)–(3.66), we need to discretize the problem in both time and space. For the time discretization, the numerical scheme presented in [19] is used for the time integration of the state and the adjoint equation. For the space discretization, we use a centered finite-difference scheme of order 2. For the Jacobian operator, we use Arakawa's method [3]. The total (nondimensional) integration time is $T = 0.1$, and the time step $\Delta t = 1.10^{-4}$ in our computations. The total number of grid points is 60×120 (i.e., $\Delta x = \Delta y = \frac{1}{60}$). For the iteration parameters, we set $\rho = 10^{-2}$ in the computations presented in this article. In all the computations presented hereafter, the observed (vorticity) flow $\mathbf{v}$ is obtained by running the quasi-geostrophic model (4.1) with the forcing $f = -\Gamma \sin \pi y$ and the initial conditions $\omega(0) = 0$ until steady state is reached. We then use the results as given data. For the values of the (nondimensional) viscosity ν used in our computations, the observed flow is time-independent, and it is obtained as a steady state of the quasi-geostrophic equation (4.1). The purpose of our simulations is to re-

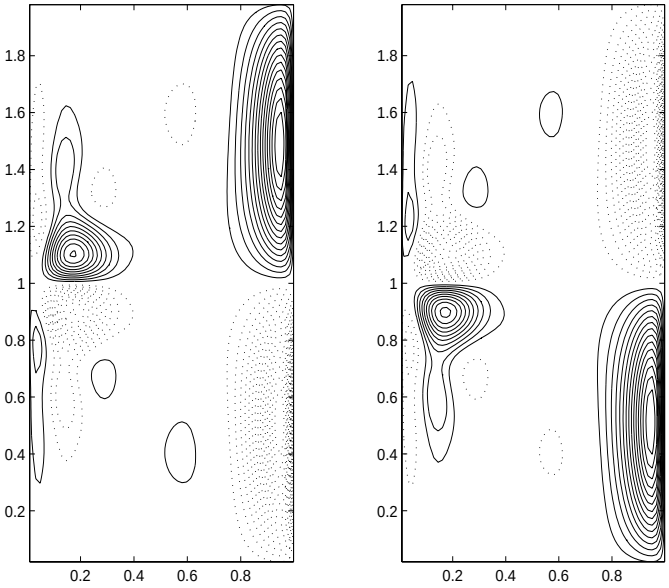


FIG. 2. *Control ϕ and disturbance ψ at $t = T = 0.1$, for $\nu = 0.1$, $\gamma = 20.0$, $l = 20.0$.*

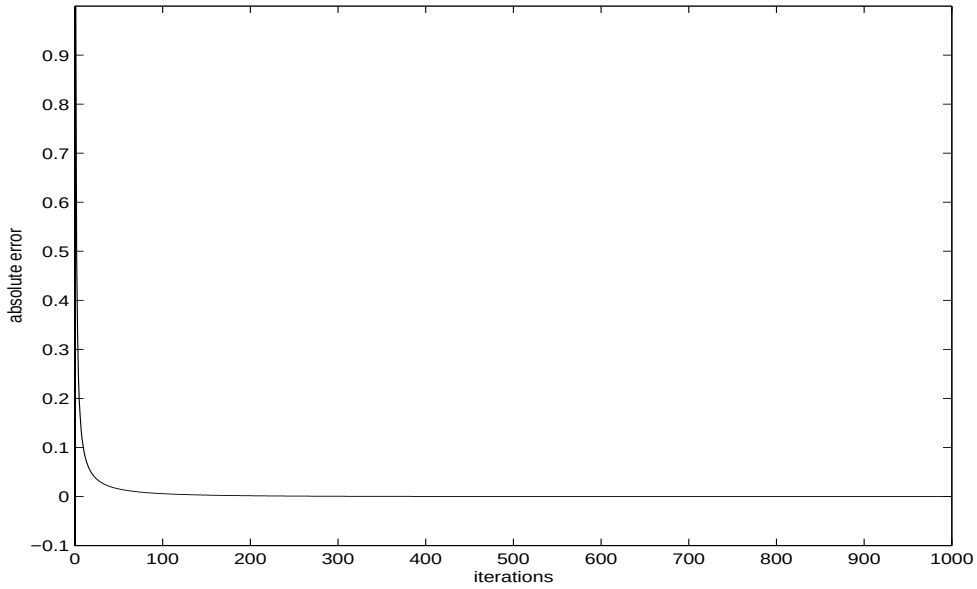


FIG. 3. *Absolute error* (4.4) *for $\nu = 0.05$, $\gamma = 20.0$, $l = 20.0$.*

construct the observed flow using the robust control model (3.64)–(3.66). To achieve that, we apply the fixed-point iteration method described in the previous section until convergence is reached. The criterion for the termination of the algorithm is given by

$$\left(\frac{|\psi_{k+1} - \psi_k|^2 + |\phi_{k+1} - \phi_k|^2}{|\psi_{k+1}|^2 + |\phi_{k+1}|^2} \right)^{\frac{1}{2}} < \epsilon, \tag{4.4}$$

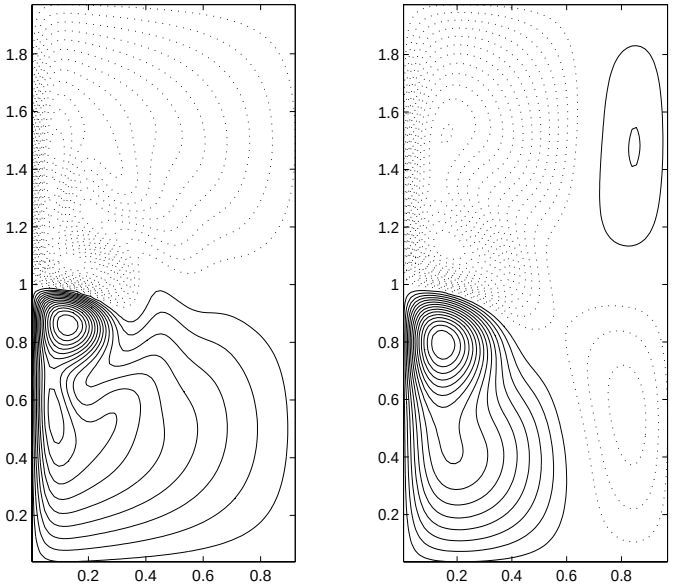


FIG. 4. *Target (streamfunction) flow and state (streamfunction) flow at* $t = T = 0.3$*, for* $\nu = 0.05$*,* $\gamma = 1.0$*,* $l = 10.0$*.*

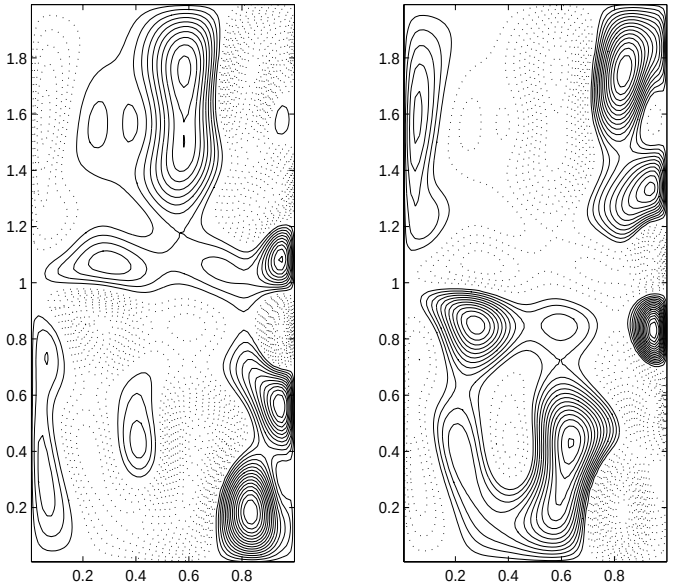


FIG. 5. *Control* ϕ *and disturbance* ψ *at* $t = T = 0.3$*, for* $\nu = 0.05$*,* $\gamma = 1.0$*,* $l = 10.0$*.*

where $\epsilon = h^2$, h is the space step in the finite-difference approximation, and $|\cdot|$ is a finite-difference approximation of the $L^2(0,T;\mathbf{L}^2(\Omega))$-norm. The following figures present the numerical results obtained from the simulations at time $t = T$ for different values of the "cost" parameter (γ, l) and (nondimensional) viscosity number ν. Figure 1 presents the target (streamfunction) flow at $t = T$, the state (streamfunction) flow at $t = T$ for the (nondimensional) viscosity number $\nu = 0.1$, and the cost

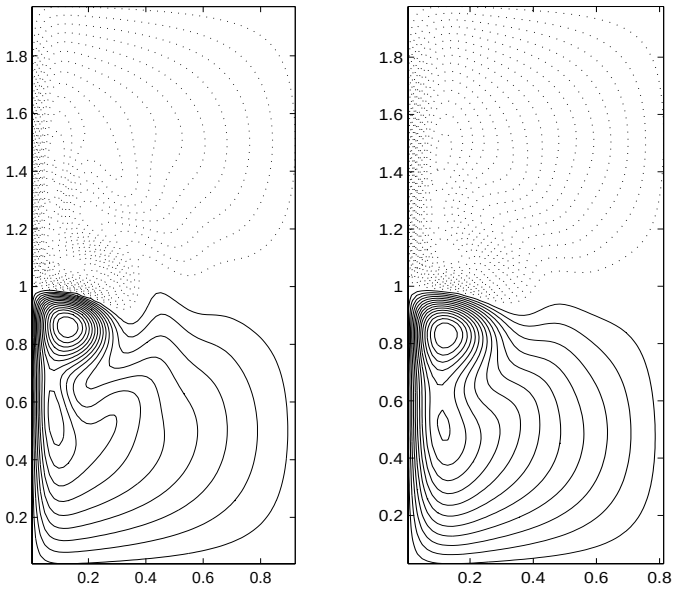


FIG. 6. *Target (streamfunction) flow and state (streamfunction) flow at* $t = T = 0.1$, *for* $\nu = 0.05$, $\gamma = 10.0$, $l = 10.0$.

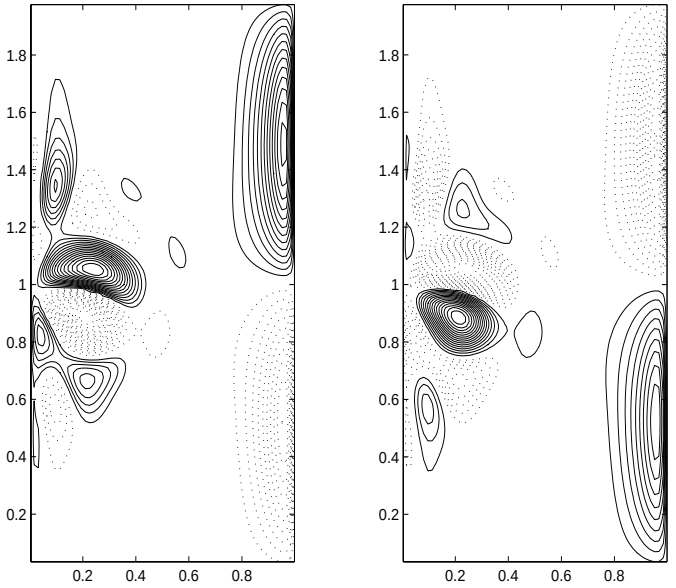


FIG. 7. *Control* ϕ *and disturbance* ψ *at* $t = T = 0.1$, *for* $\nu = 0.05$, $\gamma = 10.0$, $l = 10.0$.

parameter $(\gamma, l) = (20.0, 20.0)$. Figure 2 presents the perturbation ψ at $t = T$ and the control ϕ for the (nondimensional) viscosity number $\nu = 0.1$ and the cost parameter $(\gamma, l) = (20.0, 20.0)$. For these parameters, the state flow accurately approximates the target flow. In fact, as for the target flow, the state flow is characterized by two gyres, one cyclonic in the northern part of the basin and one anticyclonic in the southern part. The two gyres are separated by a meandering jet [23, 24]. As shown in Figure 2,

the contour lines for the control and the disturbances are the same in shape except that they have opposite signs. This suggests that the algorithm has converged to a saddle point of the functional $\mathcal{J}$. In fact, a saddle point must satisfy the condition (3.66) and in our simulations $\mathcal{B}_i = d_i I$, where I is the identity operator and $d_i > 0$ is a constant. This is also confirmed by the convergence rate given in Figure 3. A similar phenomenon appears in Figures 4–5 and Figures 6–7, for which the parameters (ν, γ, l) are $(0.05, 1.0, 10.0)$ and $(0.05, 10.0, 10.0)$, respectively. More numerical simulations will be presented elsewhere.

REFERENCES

[1] F. Abergel and R. Temam, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynam., 1 (1990), pp. 303–325.

[2] V. I. Agoshkov and G. I. Marchuck, *On the solvability and numerical solution of data assimilation problems*, Russ. J. Numer. Anal. Math. Modelling, 8 (1993), pp. 1–16.

[3] A. Arakawa, *Computational design for long-term numerical integrations of equations of atmospheric motion*, J. Comput. Phys., 1 (1966), pp. 119–143.

[4] V. Barcilon, P. Constantin, and E. S. Titi, *Existence of solutions to the Stommel–Charney model of the Gulf Stream*, SIAM J. Math. Anal., 19 (1988), pp. 1355–1364.

[5] T. R. Bewley and S. Liu, *Optimal control and robust control and estimation of linear paths to transition*, J. Fluid Mech., 365 (1998), pp. 305–349.

[6] T. R. Bewley, R. Temam, and M. Ziane, *A general framework for robust control in fluid mechanics*, Phys. D, 138 (2000), pp. 360–392.

[7] A. J. Bourgeois and J. T. Beale, *Validity of the quasigeostrophic model for the large-scale flow in the atmosphere and ocean*, SIAM J. Math. Anal., 25 (1994), pp. 1023–1068.

[8] E. P. Chassignet, *Vorticity dissipation by western boundary currents in the presence of outcropping layers*, J. Phys. Oceanogr., 25 (1995), pp. 242–255.

[9] M. D. Gunzburger and H. Kim, *Existence of an optimal solution of a shape control problem for the stationary Navier–Stokes equations*, SIAM J. Control Optim., 36 (1998), pp. 895–909.

[10] W. R. Holland and P. B. Rhines, *An example of eddy-induced ocean circulation*, J. Phys. Oceanogr., 10 (1980), pp. 1010–1031.

[11] K. Ito and S. S. Ravindran, *Optimal control of thermally convected fluid flows*, SIAM J. Sci. Comput., 19 (1998), pp. 1847–1869.

[12] J. L. Lions, R. Temam, and S. Wang, *Models of the coupled atmosphere and ocean (CAO* I*)*, Comput. Mech. Adv., 1 (1993), pp. 3–54.

[13] J. L. Lions, R. Temam, and S. Wang, *Numerical analysis of the coupled atmosphere and ocean models (CAO* II*)*, Comput. Mech. Adv., 1 (1993), pp. 55–120.

[14] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1970.

[15] G. I. Marchuck and V. P. Shutyaev, *Iterative methods for solving a data assimilation problem*, Russ. J. Numer. Anal. Math. Modelling, 9 (1994), pp. 265–279.

[16] G. I. Marchuck and V. B. Zalesny, *A numerical technique for geophysical data assimilation problems using Pontryagin's principle and splitting-up method*, Russ. J. Numer. Anal. Math. Modelling, 8 (1993), pp. 311–326.

[17] G. I. Marchuk, *Methods of Numerical Mathematics*, Springer-Verlag, New York, 1975.

[18] G. I. Marchuk and V. I. Lebedev, *Numerical Methods in the Theory of Neutron Transport*, Harwood Academic Publishers, New York, 1986.

[19] T. Tachim Medjo, *Numerical simulations of a two-layer quasi-geostrophic equation of the ocean*, SIAM J. Numer. Anal., 37 (2000), pp. 2000–2022.

[20] T. Tachim Medjo, *Numerical solutions of a robust control problem associated with the quasi-geostrophic equation of the ocean*, Nonlinear Anal. Real World Appl., 3 (2002), pp. 317–337.

[21] T. Tachim Medjo, *Fixed-point iteration method for nonlinear robust control problems in fluid mechanics*, Numer. Funct. Anal. Optim., 23 (2002), pp. 849–873.

[22] T. Tachim Medjo, *Iterative methods for a class of robust control problems in fluid mechanics*, SIAM J. Numer. Anal., 39 (2002), pp. 1625–1647.

[23] T. M. Özgökmen and E. P. Chassignet, *The emergence of inertial gyres in a two-layer quasigeostrophic model*, J. Phys. Oceanogr., 28 (1998), pp. 461–484.

[24] T. M. Özgökmen, E. P. Chassignet, and A. M. Paiva, *Impact of wind forcing, bottom*

topography, and inertia on midlatitude jet separation in a quasigeostrophic model, J. Phys. Oceanogr., 27 (1997), pp. 2460–2476.

[25] E. I. Parmuzin and V. P. Shutyaev, *Numerical analysis of iterative methods for solving evolution data assimilation problems*, Russ. J. Numer. Anal. Math. Modelling, 14 (1999), pp. 275–289.

[26] V. P. Shutyaev, *An algorithm for computing functionals for a class of nonlinear problems using the conjugate equation*, Russ. J. Numer. Anal. Math. Modelling, 6 (1991), pp. 169–178.

[27] V. P. Shutyaev, *Some properties of the control operator in the problem of data assimilation and iterative algorithms*, Russ. J. Numer. Anal. Math. Modelling, 10 (1995), pp. 357–371.

[28] V. P. Shutyaev, *Control operators and iterative algorithms in problems of reconstructing source functions and initial data*, Russ. J. Numer. Anal. Math. Modelling, 14 (1999), pp. 137–176.

[29] E. Simonnet, R. Temam, S. Wang, M. Ghil, and K. Ide, *Successive bifurcation in a shallow-water ocean model*, in Proceedings of the 16th International Conference on Numerical Methods in Fluid Dynamics, Arcachan, France, 1998, Lecture Notes in Phys. 515, C. H. Bruneau, ed., Springer-Verlag, Berlin, 1998.

[30] R. Temam, *Navier–Stokes Equations*, AMS-Chelsea Series, AMS, Providence, RI, 2001.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 326–349

ERROR ANALYSIS FOR MAPPED LEGENDRE SPECTRAL AND PSEUDOSPECTRAL METHODS*

JIE SHEN† AND LI-LIAN WANG‡

Abstract. A general framework is introduced to analyze the approximation properties of mapped Legendre polynomials and of interpolations based on mapped Legendre–Gauss–Lobatto points. Optimal error estimates featuring explicit expressions on the mapping parameters for several popular mappings are derived. These results not only play an important role in numerical analysis of mapped Legendre spectral and pseudospectral methods for differential equations but also provide quantitative criteria for the choice of parameters in these mappings.

Key words. spectral approximation, mapped Legendre polynomials, interpolation, orthogonal system

AMS subject classifications. 65N35, 65N15, 65N50

DOI. 10.1137/S0036142903422065

1. Introduction.

In a spectral method, global polynomials are used as trial functions to approximate solutions of partial differential equations (PDEs); if the underlying solutions are smooth throughout the domain, the spectral method will provide very accurate approximations with significantly fewer degrees of freedom when compared with finite difference or finite element methods (cf. [13, 8, 7]). However, if the solutions of PDEs exhibit localized rapid variations such as spikes, sharp interfaces, or internal layers, standard spectral methods usually fail to produce accurate approximations with a reasonable number of degrees of freedom, for the grid is fixed in a standard spectral method and does not take into account the localized solution behaviors. Thus, for problems with localized rapid variations, it is advisable to use a grid adapted to the localized solution behaviors rather than a standard fixed grid.

However, unlike in a finite difference or finite element method, spectral methods cannot gracefully handle an arbitrarily locally refined grid, for the spectral accuracy will usually be lost due to the fact that the locally refined grid cannot, in general, be "smoothly" mapped to the standard spectral grid. Thus the adaptivity for spectral methods is best realized through a "smooth" map which transforms a function having sharp interfaces in the physical domain to a slow varying function on the computational domain. Hence two questions need to be addressed: (i) what is the influence of the mapping on the accuracy of the spectral methods? (ii) how do we adaptively determine a suitable mapping? In this paper, we aim to provide a complete answer to the first question, which is a first step toward a long-term goal of designing a robust adaptive spectral method for solving PDEs.

In general, a coordinate transformation takes the form

$$x = g(y;\lambda), \qquad y \in [-1,1], \quad \lambda \in D_\lambda, \tag{1.1}$$

*Received by the editors January 27, 2003; accepted for publication (in revised form) July 5, 2003; published electronically March 3, 2004.

http://www.siam.org/journals/sinum/42-1/42206.html

†Department of Mathematics, Purdue University, West Lafayette, IN 47907 (shen@math.purdue.edu). The work of this author is partially supported by NFS grants DMS-0074283 and DMS-0311915.

‡Department of Mathematics, Purdue University, West Lafayette, IN 47907 (lwang@math.purdue.edu), and Department of Mathematics, Shanghai Normal University, Shanghai, P.R. China. The work of this author is partially supported by The Shanghai Natural Science Foundation N.00JC14057 and The Shanghai Natural Science Foundation for Youth N. 01QN85.

such that

$$g'(y;\lambda) > 0, \qquad g(\pm 1,\lambda) = \pm 1, \qquad \lambda \in D_\lambda, \tag{1.2}$$

where λ is a parameter vector and D_λ is the feasible domain of λ, and $'$ denotes the derivative with respect to y so (1.1) maps the interval $[-1,1]$ univalently onto itself. Without loss of generality, we assume that the mapping (1.1) is explicitly invertible and denote

$$y = g^{-1}(x;\lambda) := h(x;\lambda), \qquad x, y \in [-1,1], \qquad \lambda \in D_\lambda.$$

Several interesting mappings have been proposed and implemented in practice. In particular, Kosloff and Tal-Ezer [20] introduced the one-parameter mapping

$$x = g(y;\lambda) = \frac{\arcsin(\lambda y)}{\arcsin\lambda}, \qquad 0 < \lambda < 1. \tag{1.3}$$

This mapping stretches the Chebyshev–Gauss–Lobatto grid toward a uniform grid as $\lambda \to 1^-$. Bayliss et al. [3] used a mapped Chebyshev method to treat the boundary layer problem with the mapping

$$x = g(y;\lambda) = (4/\pi)\arctan(\tan(\pi(y-1)/4)/\lambda) + 1, \qquad \lambda > 0. \tag{1.4}$$

The mapping clusters more and more points near $x = -1$ (resp., $x = 1$) as $\lambda \to 0^+$ (resp., as $\lambda \to +\infty$). Bayliss and Turkel [4] introduced a two-parameter mapping

$$x = g(y;\lambda) = \lambda_2 + \tan(a_1(y - a_0))/\lambda_1, \qquad \lambda_1 > 0, \quad -1 \le \lambda_2 < 1, \tag{1.5}$$

where a_0 and a_1 are chosen to satisfy (1.2). Here, as λ_1 increases, more and more points are clustered near $x = \lambda_2$. These mappings have been successfully used to treat some practical problems with localized rapid variations. We note that in [1], the authors used properties special to Chebyshev polynomials to derive some error estimates on projection and interpolation errors of the mapped Chebyshev methods with the mapping (1.3). However, as far as we know, there is neither a systematic framework for analyzing the mapped spectral methods for solving PDEs nor a precise rigorous analysis on how the mapping parameter(s) would affect the accuracy. For example, there have been controversies as to whether λ (with λ close to 1) in [20] would degrade the accuracy [20, 9, 1, 23].

The main purposes of this paper are: (i) to establish a general framework for analyzing the mapped Legendre spectral methods as a first step toward an efficient adaptive spectral method; (ii) to provide precise information on how the mapping parameters affect the accuracy of the mapped spectral method.

For a given mapping, there are essentially two approaches to implement (and analyze) a mapped spectral method. In the first approach, we use $x = g(y;\lambda)$ to transform the original equation (with localized rapid variations in x) to a mapped equation (with smooth behaviors in y), and then apply a standard spectral method (in y) to the mapped equation (see, for instance, [14, 17]). The main advantage of this approach is that standard spectral approximation results can be used for the analysis, but its main disadvantage is that the mapped equation is usually very complicated and its analysis is often cumbersome. In the second approach, we do not transform the equation, but we approximate its solution using a new family of orthogonal functions $\{p_k(h(x;\lambda))\}$, which are obtained by applying the mapping $y = h(x;\lambda)$ to classical

orthogonal polynomials $\{p_k(y)\}$ (see, for instance, [6, 18, 16]) and which are suitable for capturing the localized rapid variations in the solution of the given problem. The analysis of this approach will require approximation results by using the new family of orthogonal functions. The advantage of this approach is that once these approximation results are established, it can be directly (i.e., without using a transform) applied to a large class of problems. We shall take the second approach and establish approximation results for the mapped Legendre polynomials. We emphasize that the two approaches will yield essentially the same approximate solutions (although the two implementations can be quite different). Hence the dependence of the error estimates on the mapping parameters established here for the second approach is essentially valid for the first approach.

The remainder of the paper is organized as follows. In the next section, we introduce the general framework for the mapped Legendre spectral and pseudospectral approximations. In section 3, we apply our general results to the specific mappings (1.3)–(1.5). In section 4, we consider the mapped Legendre approximations for a model problem and present some illustrative numerical results. Some concluding remarks are given in section 5.

2. The general framework. In this section, we introduce a general framework for the error analysis of Legendre spectral methods using mapping (1.1) with (1.2). We assume that for a certain positive integer $r \geq 1$,

$$h(x;\lambda) \in C^r((-1,1)), \qquad \lambda \in D_\lambda. \tag{2.1}$$

2.1. Preliminaries. We first introduce some notation. Let $I = (-1,1)$, and let $\chi(x) > 0$ be a given weight function on I. We define

$$L^2_\chi(I) = \{v \mid v \text{ is measurable on } I \text{ and } \|v\|_\chi < \infty\},$$

equipped with the following inner product and norm:

$$(u,v)_\chi = \int_I u(x)v(x)\chi(x)dx, \qquad \|v\|_\chi = (v,v)_\chi^{\frac{1}{2}}.$$

The weighted Sobolev spaces $H^m_\chi(I)$ and $H^m_{0,\chi}(I)$ are defined as usual. The norm of $H^m_\chi(I)$ is defined as

$$\|v\|_{m,\chi} = \left(\sum_{k=0}^m \|\partial_x^k v\|_\chi^2\right)^{\frac{1}{2}}.$$

In case $\chi(x) \equiv 1$, we shall drop the subscript χ in the notation for the sake of simplicity.

Let $\omega^{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta$ be the Jacobi weight function and $\mathbb{N}$ be the set of all nonnegative integers. For any $m \in \mathbb{N}$, we define the nonuniformly weighted Hilbert space

$$A^m(I) = \{v \mid \partial_x^k v \in L^2_{\omega^{k,k}}(I),\ 0 \leq k \leq m\} \tag{2.2}$$

equipped with the inner product, the seminorm, and the norm as follows:

$$(u,v)_{m,A} = \sum_{k=0}^m (\partial_x^k u, \partial_x^k v)_{\omega^{k,k}}, \quad |v|_{m,A} = \|\partial_x^m v\|_{\omega^{m,m}}, \quad \|v\|_{m,A} = (v,v)_{m,A}^{\frac{1}{2}}.$$

For any real $r > 0$, we define the space $A^r(I)$ and its norm by space interpolation.

We shall use the expression $A \lesssim B$ to mean that there exist a generic positive constant c, independent of any function, N, and the parameters of the mappings, such that $A \leq cB$.

Let $L_l(y)$ be the Legendre polynomial of degree l, which is the eigenfunction of the Sturm–Liouville problem

$$\partial_y((1-y^2)\partial_y v(y)) + \mu v(y) = 0, \qquad y \in I, \tag{2.3}$$

with the corresponding eigenvalues $\mu_l = l(l+1),\ l = 0, 1, 2, \ldots$. We have $L_l(\pm 1) = (\pm 1)^l$ and the following recurrence relations:

$$L_{l+1}(y) = \frac{2l+1}{l+1} y L_l(y) - \frac{l}{l+1} L_{l-1}(y), \qquad l \geq 1, \tag{2.4}$$

$$(2l+1)L_l(y) = \partial_y L_{l+1}(y) - \partial_y L_{l-1}(y), \qquad l \geq 1. \tag{2.5}$$

The set of Legendre polynomials forms an $L^2(I)$-orthogonal system, i.e.,

$$\int_I L_l(y) L_m(y) dy = \gamma_l \delta_{l,m}, \qquad \text{with} \quad \gamma_l = \frac{2}{2l+1}. \tag{2.6}$$

For any $v \in L^2(I)$, we write

$$v(y) = \sum_{l=0}^{\infty} \hat{v}_l L_l(y), \quad \text{with} \quad \hat{v}_l = \frac{1}{\gamma_l}(v, L_l).$$

We have the following equivalence (see [17]):

$$\|\partial_y^r v\|_{\omega^{r,r}} \sim \left(\sum_{l=r}^{\infty} \mu_l^r \hat{v}_l^2 \gamma_l \right)^{\frac{1}{2}} \quad \forall v \in A^r(I). \tag{2.7}$$

We now recall some results on the Legendre spectral approximations. Let $\mathcal{P}_N$ be the set of all polynomials of degree less than or equal to N and $\mathcal{P}_N^0 = \{\phi \in \mathcal{P}_N : \phi(\pm 1) = 0\}$. We define $\widehat{P}_N : L^2(I) \to \mathcal{P}_N$ the $L^2(I)$-orthogonal projector by

$$(\widehat{P}_N v - v, \phi) = 0 \qquad \forall \phi \in \mathcal{P}_N. \tag{2.8}$$

The following result was proved in [11] (see also [2, 15]).

LEMMA 2.1.

$$\|\partial_y^\mu (\widehat{P}_N v - v)\|_{\omega^{\mu,\mu}} \lesssim N^{\mu - r} \|\partial_y^r v\|_{\omega^{r,r}}, \qquad 0 \leq \mu \leq r,\ v \in A^r(I). \tag{2.9}$$

We define the $H^1(I)$-orthogonal projector $\widehat{P}_N^1 : H^1(I) \to \mathcal{P}_N$ by

$$(\widehat{P}_N^1 v - v, \phi)_1 = 0 \qquad \forall \phi \in \mathcal{P}_N \tag{2.10}$$

and the $H_0^1(I)$-orthogonal projector $\widehat{P}_N^{1,0} : H_0^1(I) \to \mathcal{P}_N^0$ by

$$(\partial_y(\widehat{P}_N^{1,0} v - v), \partial_y \phi) = 0 \qquad \forall \phi \in \mathcal{P}_N^0. \tag{2.11}$$

As two special cases of Theorem 3.1 and Theorem 3.4 in Guo and Wang [19], we have the following lemma.

LEMMA 2.2. *If $v \in H^1(I)$ and $\partial_y^r v \in L^2_{\omega^{r-1,r-1}}(I)$, then*

$$\|\widehat{P}^1_N v - v\|_\mu \lesssim N^{\mu-r}\|\partial_y^r v\|_{\omega^{r-1,r-1}}, \quad 0 \le \mu \le 1 \le r. \tag{2.12}$$

If $v \in H^1_0(I)$ and $\partial_y^r v \in L^2_{\omega^{r-1,r-1}}(I)$, then

$$\|\widehat{P}^{1,0}_N v - v\|_\mu \lesssim N^{\mu-r}\|\partial_y^r v\|_{\omega^{r-1,r-1}}, \quad 0 \le \mu \le 1 \le r. \tag{2.13}$$

Next, let $\zeta_{N,j},\ 0 \le j \le N$, be the Legendre–Gauss–Lobatto (LGL) points, which are the zeros of $(1-y^2)\partial_y L_N(y)$. We assume that they are arranged in ascending order. There exists a unique set of Christoffel numbers $\{\omega_{N,j}\}$ such that

$$\int_I \phi(y)dy = \sum_{j=0}^{N} \phi(\zeta_{N,j})\omega_{N,j} \qquad \forall \phi \in \mathcal{P}_{2N-1}. \tag{2.14}$$

In fact, we have

$$\omega_{N,0} = \omega_{N,N} = \frac{2}{N(N+1)},\ \ \omega_{N,j} = \frac{2}{N(N+1)}(L_N(\zeta_{N,j}))^{-2},\ \ 1 \le j \le N-1. \tag{2.15}$$

We define the discrete inner product and discrete norm as

$$(u,v)_N = \sum_{j=0}^{N} u(\zeta_{N,j})v(\zeta_{N,j})\omega_{N,j}, \quad \|v\|_N = (v,v)_N^{\frac{1}{2}}.$$

Note that we have (see, for instance, formula (21.8) of Bernardi and Maday [5])

$$\|\phi\| \le \|\phi\|_N \le \sqrt{2+N^{-1}}\|\phi\| \quad \forall \phi \in \mathcal{P}_N. \tag{2.16}$$

On the other hand, we also have (see Theorem 4.9 of Guo and Wang [15])

$$\|v\|_N \lesssim \|v\| + N^{-1}\|\partial_y v\|_{\omega^{1,1}} \qquad \forall v \in H^1_0(I). \tag{2.17}$$

2.2. Mapped Legendre orthogonal approximations. For a given mapping $y = h(x;\lambda)$, we define the mapped Legendre polynomials by

$$\mathcal{L}_l^{(\lambda)}(x) = L_l(y) = L_l(h(x;\lambda)), \qquad l = 0,1,2,\ldots. \tag{2.18}$$

Due to $h(\pm 1;\lambda) = \pm 1$, we have $\mathcal{L}_l^{(\lambda)}(\pm 1) = (\pm 1)^l$. We denote the weight function

$$\omega_\lambda(x) := h'(x;\lambda) = \frac{1}{g'(y;\lambda)}. \tag{2.19}$$

Thanks to (2.5), we have the recurrence relation

$$(2l+1)\omega_\lambda(x)\mathcal{L}_l^{(\lambda)}(x) = \partial_x \mathcal{L}_{l+1}^{(\lambda)}(x) - \partial_x \mathcal{L}_{l-1}^{(\lambda)}(x), \qquad l \ge 1. \tag{2.20}$$

By virtue of (2.6), the set $\{\mathcal{L}_l^{(\lambda)}\}_{l=0}^{\infty}$ forms a complete orthogonal system in $L^2_{\omega_\lambda}(I)$, and consequently, for any $v \in L^2_{\omega_\lambda}(I)$, we can write

$$v(x) = \sum_{l=0}^{\infty} \hat{v}_l^{(\lambda)} \mathcal{L}_l^{(\lambda)}(x), \qquad \text{with} \qquad \hat{v}_l^{(\lambda)} = \frac{1}{\gamma_l}(v, \mathcal{L}_l^{(\lambda)})_{\omega_\lambda}. \tag{2.21}$$

Moreover, $\mathcal{L}_l^{(\lambda)}$ is the eigenfunction of the Sturm–Liouville problem

$$w_\lambda^{-1}(x)\partial_x(\tilde{\omega}_\lambda(x)\partial_x \mathcal{L}_l^{(\lambda)}(x)) + \mu_l \mathcal{L}_l^{(\lambda)}(x) = 0, \quad x \in I,$$

with $\tilde{\omega}_\lambda(x) = (1-h^2(x;\lambda))\omega_\lambda^{-1}(x)$. This implies that $\{\partial_x \mathcal{L}_l^{(\lambda)}\}_{l=1}^{\infty}$ forms an orthogonal system in $L^2_{\tilde{\omega}_\lambda}(I)$, i.e.,

$$\int_I \partial_x \mathcal{L}_l^{(\lambda)}(x)\partial_x \mathcal{L}_m^{(\lambda)}(x)\tilde{\omega}_\lambda(x)dx = \mu_l \gamma_l \delta_{l,m}. \tag{2.22}$$

We now consider error estimates for approximations using the orthogonal system $\{\mathcal{L}_l^{(\lambda)}\}_{l=0}^{\infty}$. For $\lambda \in D_\lambda$, we set $\mathcal{V}_{N,\lambda} = \text{span}\{\mathcal{L}_0^{(\lambda)}, \mathcal{L}_1^{(\lambda)}, \dots, \mathcal{L}_N^{(\lambda)}\}$. Let $P_{N,\lambda} : L^2_{\omega_\lambda}(I) \to \mathcal{V}_{N,\lambda}$ be the $L^2_{\omega_\lambda}(I)$-orthogonal projector defined by

$$(P_{N,\lambda}v - v, \phi)_{\omega_\lambda} = 0 \qquad \forall \phi \in \mathcal{V}_{N,\lambda}. \tag{2.23}$$

For clarity, the following notation will be used in what follows:

$$V_\lambda(y) = v \circ g(y;\lambda) = v(x), \;\; \Phi_\lambda(y) = \phi \circ g(y;\lambda) = \phi(x), \;\; x, y \in I, \;\; \lambda \in D_\lambda. \tag{2.24}$$

For $\lambda \in D_\lambda$ and $r \in \mathbb{N}$, we define

$$A_\lambda^r(I) = \{v \in L^2_{\omega_\lambda}(I) \mid |v|_{A_\lambda^r} = \|(1-y^2)^{\frac{r}{2}}\partial_y^r V_\lambda\| < \infty, \; y = h(x;\lambda)\} \tag{2.25}$$

and

$$B_\lambda^r(I) = \{v \in L^2_{\omega_\lambda}(I) \mid |v|_{B_\lambda^r} = \|(1-y^2)^{\frac{r-1}{2}}\partial_y^r V_\lambda\| < \infty, \; y = h(x;\lambda)\}. \tag{2.26}$$

The following is a fundamental result for the mapped Legendre spectral approximations.

THEOREM 2.1. *For any $v \in A_\lambda^r(I)$, $\lambda \in D_\lambda$, and $r \geq 1$,*

$$\|\partial_x(P_{N,\lambda}v - v)\|_{\tilde{\omega}_\lambda} + N\|P_{N,\lambda}v - v\|_{\omega_\lambda} \lesssim N^{1-r}|v|_{A_\lambda^r}. \tag{2.27}$$

Moreover, for $r > 1$,

$$|P_{N,\lambda}v(\pm 1) - v(\pm 1)| \lesssim N^{1-r}|v|_{A_\lambda^r}. \tag{2.28}$$

Proof. For $v \in L^2_{\omega_\lambda}(I)$, we have $V_\lambda \in L^2(I)$, so we can write

$$v(x) = \sum_{l=0}^{\infty} \hat{v}_l^{(\lambda)} \mathcal{L}_l^{(\lambda)}(x) = \sum_{j=0}^{\infty} \widehat{V}_l^{(\lambda)} L_l(y) = V_\lambda(y) \tag{2.29}$$

with

$$\hat{v}_l^{(\lambda)} = \frac{1}{\gamma_l}(v, \mathcal{L}_l^{(\lambda)})_{\omega_\lambda} = \frac{1}{\gamma_l}(V_\lambda, L_l) = \widehat{V}_l^{(\lambda)}. \tag{2.30}$$

Let $\widehat{P}_N$ and $P_{N,\lambda}$ be the projectors defined in (2.8) and (2.23). We derive from (2.30) and Lemma 2.1 with $\mu = 0$ that

$$\begin{aligned}\|P_{N,\lambda}v - v\|^2_{\omega_\lambda} &= \sum_{l=N+1}^{\infty} (\hat{v}_l^{(\lambda)})^2\gamma_l = \sum_{l=N+1}^{\infty} (\widehat{V}_l^{(\lambda)})^2\gamma_l = \|\widehat{P}_N V_\lambda - V_\lambda\|^2 \\ &\lesssim N^{-2r}\|(1-y^2)^{\frac{r}{2}}\partial_y^r V_\lambda\|^2 = N^{-2r}|v|^2_{A^r_\lambda}.\end{aligned} \tag{2.31}$$

On the other hand, since $\{\partial_y L_l\}$ is $L^2_{\omega^{1,1}}(I)$-orthogonal, we derive from (2.22), (2.30), and Lemma 2.1 with $\mu = 1$ that

$$\begin{aligned}\|\partial_x(P_{N,\lambda}v - v)\|^2_{\tilde{\omega}_\lambda} &= \sum_{l=N+1}^{\infty} \mu_l\gamma_l(\hat{v}_l^{(\lambda)})^2 = \sum_{l=N+1}^{\infty} \mu_l\gamma_l(\widehat{V}_l^{(\lambda)})^2 \\ &= \|\partial_y(\widehat{P}_N V_\lambda - V_\lambda)\|^2_{\omega^{1,1}} \lesssim N^{2(1-r)}|v|^2_{A^r_\lambda}.\end{aligned}$$

Next, since $|\mathcal{L}_l^{(\lambda)}(\pm 1)| = 1$, we derive from (2.7), (2.29), and (2.30) that

$$\begin{aligned}|P_{N,\lambda}v(\pm 1) - v(\pm 1)| &\le \sum_{l=N+1}^{\infty} |\widehat{V}_l^{(\lambda)}| \le C_{N,r}\left(\sum_{l=N+1}^{\infty} \mu_l^r(\widehat{V}_l^{(\lambda)})^2\gamma_l\right)^{\frac{1}{2}} \\ &\lesssim C_{N,r}\|\partial_y^r V_\lambda\|_{\omega^{r,r}} \lesssim C_{N,r}|v|_{A^r_\lambda},\end{aligned}$$

where for $r > 1$,

$$C_{N,r} = \left(\sum_{l=N+1}^{\infty} \mu_l^{-r}\gamma_l^{-1}\right)^{\frac{1}{2}} \lesssim \left(\sum_{l=N+1}^{\infty} l^{1-2r}\right)^{\frac{1}{2}} \lesssim \left(\int_N^{\infty} x^{1-2r}dx\right)^{\frac{1}{2}} \lesssim N^{1-r}.$$

The proof is complete. ∎

When analyzing mapped Legendre spectral methods for numerical solutions of PDEs, we often need to consider the $H^1_{\omega_\lambda}(I)$-orthogonal projection $P^1_{N,\lambda} : H^1_{\omega_\lambda}(I) \to \mathcal{V}_{N,\lambda}$ defined by

$$(P^1_{N,\lambda}v - v, \phi)_{1,\omega_\lambda} = 0 \qquad \forall \phi \in \mathcal{V}_{N,\lambda}.$$

THEOREM 2.2. *For any $v \in H^1_{\omega_\lambda}(I) \cap B^r_\lambda(I)$, $\lambda \in D_\lambda$, and $r \ge 1$,*

$$\|P^1_{N,\lambda}v - v\|_{1,\omega_\lambda} \lesssim (d_{\lambda,1}+1)N^{1-r}|v|_{B^r_\lambda}, \tag{2.32}$$

where $d_{\lambda,1} = \max_{x\in\bar{I}} |\omega_\lambda(x)|$.

Proof. By (2.19) and (2.24),

$$\begin{aligned}\|\phi - v\|^2_{1,\omega_\lambda} &= \int_I (\partial_y(\Phi_\lambda(y) - V_\lambda(y)))^2\Big(\frac{dy}{dx}\Big)^2 dy + \int_I (\Phi_\lambda(y) - V_\lambda(y))^2 dy \\ &\le (d_{\lambda,1}+1)^2\|\Phi_\lambda - V_\lambda\|^2_1.\end{aligned} \tag{2.33}$$

Next, we take $\phi(x) = \Phi_\lambda(y) = \widehat{P}^1_N V_\lambda(y)$ in (2.33), where $\widehat{P}^1_N$ is defined in (2.10), and we obtain from the projection theorem and Lemma 2.2 that

$$\begin{aligned}\|P^1_{N,\lambda}v - v\|_{1,\omega_\lambda} &= \inf_{\phi\in\mathcal{V}_{N,\lambda}} \|\phi - v\|_{1,\omega_\lambda} \le (d_{\lambda,1}+1)\|\widehat{P}^1_N V_\lambda - V_\lambda\|_1 \\ &\lesssim (d_{\lambda,1}+1)N^{1-r}\|\partial_y^r V_\lambda\|_{\omega^{r-1,r-1}} \lesssim (d_{\lambda,1}+1)N^{1-r}|v|_{B^r_\lambda}.\end{aligned}$$ ∎

Next, we consider the bilinear form

$$a_{\omega_\lambda}^{(\nu)}(u,v)=\nu_1(\partial_x u,\partial_x(v\omega_\lambda))+\nu_2(u,v)_{\omega_\lambda}$$

(where $\nu=(\nu_1,\nu_2)$ and $\nu_i>0, i=1,2$) associated to the mapped Legendre spectral approximation of the model elliptic equation

$$-\mu_1 v_{xx}+\mu_2 v=f,\quad v(\pm 1)=0. \tag{2.34}$$

Due to the nonuniform weight function $\omega_\lambda(x)$, the bilinear form $a_{\omega_\lambda}^{(\nu)}(u,v)$ is not symmetric. We first study its continuity and coercivity.

LEMMA 2.3. *For any $u,v\in H^1_{\omega_\lambda}(I)$,*

$$a_{\omega_\lambda}^{(\nu)}(u,v)\le \nu_1(d_{\lambda,2}+1)|u|_{1,\omega_\lambda}\|v\|_{1,\omega_\lambda}+\nu_2\|u\|_{\omega_\lambda}\|v\|_{\omega_\lambda}, \tag{2.35}$$

where

$$d_{\lambda,2}=\max_{x\in\bar I}|\omega_\lambda^{-1}(x)\partial_x\omega_\lambda(x)|. \tag{2.36}$$

For any $v\in H^1_{0,\omega_\lambda}(I)$,

$$a_{\omega_\lambda}^{(\nu)}(v,v)\ge \nu_1|v|^2_{1,\omega_\lambda}+\Big(\nu_2-\frac{\nu_1}{2}d_{\lambda,3}\Big)\|v\|^2_{\omega_\lambda}, \tag{2.37}$$

where

$$d_{\lambda,3}=\max_{x\in\bar I}\{\omega_\lambda^{-1}(x)\partial_x^2\omega_\lambda(x)\}. \tag{2.38}$$

Proof. By the Cauchy inequality,

$$\begin{aligned}a_{\omega_\lambda}^{(\nu)}(u,v)&\le \nu_1|(\partial_x u,\partial_x v)_{\omega_\lambda}+(\partial_x u,v\partial_x\omega_\lambda)|+\nu_2|(u,v)_{\omega_\lambda}|\\&\le \nu_1(|u|_{1,\omega_\lambda}|v|_{1,\omega_\lambda}+\max_{x\in\bar I}|\omega_\lambda^{-1}(x)\partial_x\omega_\lambda(x)||u|_{1,\omega_\lambda}\|v\|_{\omega_\lambda})+\nu_2\|u\|_{\omega_\lambda}\|v\|_{\omega_\lambda}\\&\le \nu_1(d_{\lambda,2}+1)|u|_{1,\omega_\lambda}\|v\|_{1,\omega_\lambda}+\nu_2\|u\|_{\omega_\lambda}\|v\|_{\omega_\lambda}.\end{aligned}$$

On the other hand,

$$\begin{aligned}a_{\omega_\lambda}^{(\nu)}(v,v)&=\nu_1|v|^2_{1,\omega_\lambda}+\nu_2\|v\|^2_{\omega_\lambda}+\frac{\nu_1}{2}\int_I\partial_x(v^2(x))\partial_x\omega_\lambda(x)dx\\&=\nu_1|v|^2_{1,\omega_\lambda}+\nu_2\|v\|^2_{\omega_\lambda}-\frac{\nu_1}{2}\int_I v^2(x)\partial_x^2\omega_\lambda(x)dx \qquad (2.39)\\&\ge \nu_1|v|^2_{1,\omega_\lambda}+\Big(\nu_2-\frac{\nu_1}{2}d_{\lambda,3}\Big)\|v\|^2_{\omega_\lambda}.\end{aligned}$$

∎

This lemma indicates that if $\nu_2>\frac{\nu_1}{2}d_{\lambda,3}$, then $|||v|||_{1,\omega_\lambda}:=\sqrt{a_{\omega_\lambda}^{(\nu)}(v,v)}$ is a norm for the space $H^1_{0,\omega_\lambda}(I)$.

Next, we set $\mathcal{V}^0_{N,\lambda}=H^1_{0,\omega_\lambda}(I)\cap\mathcal{V}_{N,\lambda}$ and define the orthogonal projector $P^{1,0}_{N,\lambda}: H^1_{0,\omega_\lambda}(I)\to\mathcal{V}^0_{N,\lambda}$ by

$$a_{\omega_\lambda}^{(\nu)}(P^{1,0}_{N,\lambda}v-v,\phi)=0\qquad\forall\phi\in\mathcal{V}^0_{N,\lambda}. \tag{2.40}$$

THEOREM 2.3. *If* $\nu_2 > \frac{\nu_1}{2} d_{\lambda,3}$, *then for any* $v \in H^1_{0,\omega_\lambda}(I) \cap B^r_\lambda(I)$, $\lambda \in D_\lambda$, *and* $r \geq 1$,

$$|||P^{1,0}_{N,\lambda} v - v|||_{1,\omega_\lambda} \lesssim (\nu_1^{\frac{1}{2}}(d_{\lambda,1}+1)(d_{\lambda,2}+1)^{\frac{1}{2}} + \nu_2^{\frac{1}{2}} N^{-1}) N^{1-r} |v|_{B^r_\lambda}, \tag{2.41}$$

where $d_{\lambda,i}, i = 1, 2, 3$, *are the same as before.*

Proof. By the projection theorem, (2.33), and (2.35),

$$\begin{aligned} |||P^{1,0}_{N,\lambda} v - v|||^2_{1,\omega_\lambda} &= \inf_{\phi \in \mathcal{V}^0_{N,\lambda}} |||\phi - v|||^2_{1,\omega_\lambda} \\ &\leq \nu_1 (d_{\lambda,2}+1) |\phi - v|_{1,\omega_\lambda} \|\phi - v\|_{1,\omega_\lambda} + \nu_2 \|\phi - v\|^2_{\omega_\lambda} \\ &\leq \nu_1 (d_{\lambda,1}+1)^2 (d_{\lambda,2}+1) \|\Phi_\lambda - V_\lambda\|^2_1 + \nu_2 \|\Phi_\lambda - V_\lambda\|^2. \end{aligned} \tag{2.42}$$

Let $\widehat{P}^{1,0}_N$ be the $H^1_0(I)$-orthogonal projection as in Lemma 2.2. Hence, by taking $\phi(x) = \Phi_\lambda(y) = \widehat{P}^{1,0}_N V_\lambda(y)$ in (2.42), where $\widehat{P}^{1,0}_N$ is defined in (2.11), we can obtain the desired result thanks to Lemma 2.2. ∎

2.3. Mapped Legendre pseudospectral approximations. In this subsection, we consider the interpolation operator based on the mapped Legendre–Gauss–Lobatto (MLGL) points.

Let $\{\zeta_{N,j}\}_{j=0}^N$ and $\{\omega_{N,j}\}_{j=0}^N$ be the LGL points and weights. The MLGL points and weights are defined by

$$\xi^{(\lambda)}_{N,j} = g(\zeta_{N,j}; \lambda), \quad \omega^{(\lambda)}_{N,j} = \omega_{N,j}, \quad 0 \leq j \leq N, \ \ \lambda \in D_\lambda. \tag{2.43}$$

It is clear that $\xi^{(\lambda)}_{N,0} = -1$ and $\xi^{(\lambda)}_{N,N} = 1$, and thanks to (2.14),

$$\int_I \phi(x) \omega_\lambda(x) dx = \int_I \phi(g(y;\lambda)) dy = \sum_{j=0}^N \phi(\xi^{(\lambda)}_{N,j}) \omega^{(\lambda)}_{N,j} \qquad \forall \phi \in \mathcal{V}_{2N-1,\lambda}. \tag{2.44}$$

Let the discrete inner product and discrete norm be defined as

$$(u,v)_{\omega_\lambda,N} = \sum_{j=0}^N u(\xi^{(\lambda)}_{N,j}) v(\xi^{(\lambda)}_{N,j}) \omega^{(\lambda)}_{N,j}, \qquad \|v\|_{\omega_\lambda,N} = (v,v)^{\frac{1}{2}}_{\omega_\lambda,N}.$$

We have from (2.16) that

$$\|\phi\|_{\omega_\lambda} \leq \|\phi\|_{\omega_\lambda,N} \leq \Big(2 + \frac{1}{N}\Big)^{\frac{1}{2}} \|\phi\|_{\omega_\lambda} \qquad \forall \phi \in \mathcal{V}_{N,\lambda}. \tag{2.45}$$

Let $\mathcal{I}_{N,\lambda} v$ be the interpolation of $v(x)$ in $\mathcal{V}_{N,\lambda}$ at points $\xi^{(\lambda)}_{N,j}$. We first establish a result on the stability of $\mathcal{I}_{N,\lambda}$.

LEMMA 2.4. *For any* $v \in L^2_{\omega_\lambda}(I) \cap H^1_{\tilde{\omega}_\lambda}(I)$ *and* $\lambda \in D_\lambda$,

$$\|\mathcal{I}_{N,\lambda} v\|_{\omega_\lambda} \lesssim \|v\|_{\omega_\lambda} + N^{-1}(|v(1)| + |v(-1)| + |v|_{1,\tilde{\omega}_\lambda}). \tag{2.46}$$

Proof. By (2.15), (2.43), and (2.45),

$$\begin{aligned} \|\mathcal{I}_{N,\lambda} v\|^2_{\omega_\lambda} \leq \|\mathcal{I}_{N,\lambda} v\|^2_{\omega_\lambda,N} &= \frac{2}{N(N+1)}(v^2(-1) + v^2(1)) + \sum_{j=1}^{N-1} v^2(\xi^{(\lambda)}_{N,j}) \omega^{(\lambda)}_{N,j} \\ &= \frac{2}{N(N+1)}(v^2(-1) + v^2(1)) + \sum_{j=1}^{N-1} V^2_\lambda(\zeta_{N,j}) \omega_{N,j}. \end{aligned} \tag{2.47}$$

Thanks to (2.17), we have from (2.24) that

$$\sum_{j=1}^{N-1} V_\lambda^2(\zeta_{N,j})\omega_{N,j} \le \|V_\lambda\|_N^2 \lesssim \|V_\lambda\|^2 + N^{-2}\|\partial_y V_\lambda\|_{\omega^{1,1}}^2 \lesssim \|v\|_{\omega_\lambda}^2 + N^{-2}|v|_{1,\tilde{\omega}_\lambda}^2. \tag{2.48}$$

This completes the proof. ∎

Remark 2.1. For the treatment of nonlinear problems, we often need to estimate the terms such as $\mathcal{I}_{N,\lambda}\phi$ with $\phi \in \mathcal{V}_{M,\lambda}^0, M > N$. By the formula (5.9) in [5], we have the following inverse inequality:

$$|\phi|_{1,\tilde{\omega}_\lambda} = |\Phi_\lambda|_{1,\omega^{1,1}} \le \sqrt{2}N\|\Phi_\lambda\| = \sqrt{2}N\|\phi\|_{\omega_\lambda} \quad \forall \phi \in \mathcal{V}_{N,\lambda}. \tag{2.49}$$

So we obtain from (2.46) that for any $\phi \in \mathcal{V}_{M,\lambda}^0$ and $\psi \in \mathcal{V}_{L,\lambda}^0$,

$$\|\mathcal{I}_{N,\lambda}\phi\|_{\omega_\lambda} \lesssim \|\phi\|_{\omega_\lambda} + N^{-1}|\phi|_{1,\tilde{\omega}_\lambda} \lesssim \Big(1 + \frac{M}{N}\Big)\|\phi\|_{\omega_\lambda}. \tag{2.50}$$

On the other hand, by (2.45),

$$\begin{aligned} |(\phi,\psi)_{\omega_\lambda,N}| &= |(\mathcal{I}_{N,\lambda}\phi, \mathcal{I}_{N,\lambda}\psi)_{\omega_\lambda,N}| \le \|\mathcal{I}_{N,\lambda}\phi\|_{\omega_\lambda,N}\|\mathcal{I}_{N,\lambda}\psi\|_{\omega_\lambda,N} \\ &\le \Big(2 + \frac{1}{N}\Big)\|\mathcal{I}_{N,\lambda}\phi\|_{\omega_\lambda}\|\mathcal{I}_{N,\lambda}\psi\|_{\omega_\lambda} \\ &\lesssim \Big(1 + \frac{M}{N}\Big)\Big(1 + \frac{L}{N}\Big)\|\phi\|_{\omega_\lambda}\|\psi\|_{\omega_\lambda}. \end{aligned} \tag{2.51}$$

The following is the main result on the MLGL interpolation.

THEOREM 2.4. *For any* $v \in A_\lambda^r(I)$, $\lambda \in D_\lambda$, *and* $r > 1$,

$$\|\partial_x(\mathcal{I}_{N,\lambda}v - v)\|_{\tilde{\omega}_\lambda} + N\|\mathcal{I}_{N,\lambda}v - v\|_{\omega_\lambda} \lesssim N^{1-r}|v|_{A_\lambda^r}. \tag{2.52}$$

Proof. By (2.27), (2.28), and (2.46),

$$\begin{aligned} \|\mathcal{I}_{N,\lambda}v - P_{N,\lambda}v\|_{\omega_\lambda} &= \|\mathcal{I}_{N,\lambda}(P_{N,\lambda}v - v)\|_{\omega_\lambda} \\ &\lesssim \|P_{N,\lambda}v - v\|_{\omega_\lambda} + N^{-1}(|P_{N,\lambda}v(1) - v(1)| \\ &\quad + |P_{N,\lambda}v(-1) - v(-1)| + \|\partial_x(P_{N,\lambda}v - v)\|_{\tilde{\omega}_\lambda}) \\ &\lesssim N^{-r}|v|_{A_\lambda^r}. \end{aligned} \tag{2.53}$$

Due to (2.49), we obtain from (2.53) that

$$|\mathcal{I}_{N,\lambda}v - P_{N,\lambda}v|_{1,\tilde{\omega}_\lambda} \lesssim N\|\mathcal{I}_{N,\lambda}v - P_{N,\lambda}v\|_{\omega_\lambda} \lesssim N^{1-r}|v|_{A_\lambda^r}.$$

We then derive from (2.27) that

$$\begin{aligned} |\mathcal{I}_{N,\lambda}v - v|_{1,\tilde{\omega}_\lambda} + N\|\mathcal{I}_{N,\lambda}v - v\|_{\omega_\lambda} &\le |\mathcal{I}_{N,\lambda}v - P_{N,\lambda}v|_{1,\tilde{\omega}_\lambda} + N\|\mathcal{I}_{N,\lambda}v - P_{N,\lambda}v\|_{\omega_\lambda} \\ &\quad + |P_{N,\lambda}v - v|_{1,\tilde{\omega}_\lambda} + N\|P_{N,\lambda}v - v\|_{\omega_\lambda} \lesssim N^{1-r}|v|_{A_\lambda^r}. \end{aligned}$$

∎

Remark 2.2. As a direct consequence, we can estimate the difference between the continuous and discrete inner products. In fact, we deduce from (2.27), (2.44), (2.45), and (2.52) that for any $v \in A_\lambda^r(I)$ and $\phi \in \mathcal{V}_{N,\lambda}$ with $r > 1$ and $\lambda \in D_\lambda$,

$$\begin{aligned} &|(v,\phi)_{\omega_\lambda} - (v,\phi)_{\omega_\lambda,N}| \\ &\quad \le |(v,\phi)_{\omega_\lambda} - (P_{N-1,\lambda}v, v)_{\omega_\lambda}| + |(P_{N-1,\lambda}v,\phi)_{\omega_\lambda,N} - (\mathcal{I}_{N,\lambda}v,\phi)_{\omega_\lambda,N}| \\ &\quad \le \|P_{N-1,\lambda}v - v\|_{\omega_\lambda}\|\phi\|_{\omega_\lambda} + \|P_{N-1,\lambda}v - \mathcal{I}_{N,\lambda}v\|_{\omega_\lambda,N}\|\phi\|_{\omega_\lambda,N} \\ &\quad \lesssim N^{-r}|v|_{A_\lambda^r}\|\phi\|_{\omega_\lambda}. \end{aligned} \tag{2.54}$$

3. The upper bounds of $|v|_{A^r_\lambda}$ and $|v|_{B^r_\lambda}$. In this section, we provide upper bounds of $|v|_{A^r_\lambda}$ and $|v|_{B^r_\lambda}$ in terms of derivatives of $v(x)$ for the mappings (1.3)–(1.5). We also derive explicit bounds on the positive constants $d_{\lambda,i}$ $(i = 1, 2, 3)$ defined in the previous section. These bounds provide, in particular, explicit information on how the mapping parameters affect the accuracy of the mapped Legendre approximation.

3.1. The mapping (1.3). In this case, we have $D_\lambda = (0, 1)$. The inverse of the mapping (1.3) is

$$y = h(x;\lambda) = \frac{\sin(ax)}{\lambda}, \qquad a = \arcsin\lambda, \quad \lambda \in (0,1). \tag{3.1}$$

Moreover,

$$\omega_\lambda(x) = \frac{dy}{dx} = \frac{a}{\lambda}\cos(ax) = \Big(\frac{dx}{dy}\Big)^{-1} = \frac{a}{\lambda}\sqrt{1-\lambda^2y^2}. \tag{3.2}$$

Since $a \to \lambda$ as $\lambda \to 0$, $\frac{\lambda}{a}$ is uniformly bounded for $\lambda \in (0, 1)$. For clarity, let $Q_l(y;\lambda)$ be a polynomial of degree l with respect to y. Then for any integer $k \geq 1$,

$$\begin{aligned}
\frac{d^kx}{dy^k} &= \frac{\lambda}{a}\sum_{j=0}^{k-1}\binom{k-1}{j}\Big((1-\lambda y)^{-\frac12}\Big)^{(j)}\Big((1+\lambda y)^{-\frac12}\Big)^{(k-j-1)}\\
&= \sum_{j=0}^{k-1}E_j^k(\lambda)(1-\lambda y)^{-\frac12-j}(1+\lambda y)^{\frac12+j-k}\\
&= (1-\lambda^2y^2)^{\frac12-k}\sum_{j=0}^{k-1}E_j^k(\lambda)(1-\lambda y)^{k-1-j}(1+\lambda y)^j\\
&= (1-\lambda^2y^2)^{\frac12-k}Q_{k-1}(y;\lambda),
\end{aligned} \tag{3.3}$$

where $E_j^k(\lambda)$ is a constant in terms of j, k, and λ. By direct calculations,

$$\begin{aligned}
\partial_yV_\lambda(y) &= \partial_xv(x)\frac{dx}{dy} = (1-\lambda^2y^2)^{-\frac12}Q_0(y;\lambda)\partial_xv(x),\\
\partial_y^2V_\lambda(y) &= \partial_x^2v(x)\Big(\frac{dx}{dy}\Big)^2 + \partial_xv(x)\frac{d^2x}{dy^2}\\
&= (1-\lambda^2y^2)^{-1}Q_0(y;\lambda)\partial_x^2v(x) + (1-\lambda^2y^2)^{-\frac32}Q_1(y;\lambda)\partial_xv(x),\\
\partial_y^3V_\lambda(y) &= \partial_x^3v(x)\Big(\frac{dx}{dy}\Big)^3 + 3\partial_x^2v(x)\frac{dx}{dy}\frac{d^2x}{dy^2} + \partial_xv(x)\frac{d^3x}{dy^3}\\
&= (1-\lambda^2y^2)^{-\frac32}Q_0(y;\lambda)\partial_x^3v(x) + (1-\lambda^2y^2)^{-2}Q_1(y;\lambda)\partial_x^2v(x)\\
&\quad + (1-\lambda^2y^2)^{-\frac52}Q_2(y;\lambda)\partial_xv(x).
\end{aligned} \tag{3.4}$$

Thus an induction argument leads to

$$\partial_y^kV_\lambda(y) = \sum_{j=1}^{k}(1-\lambda^2y^2)^{\frac{j}{2}-k}Q_{k-j}(y;\lambda)\partial_x^jv(x), \qquad k \geq 1, \tag{3.5}$$

where $Q_l(y;\lambda)$ $(0 \leq l \leq k)$ are uniformly bounded for all $y \in \bar I$ and $\lambda \in (0, 1)$. Then, by the definition of $|v|_{A^r_\lambda}$, we derive from (3.5) that for $r \geq 1$,

$$|v|^2_{A^r_\lambda} = \|\partial_y^rV_\lambda\|^2_{\omega^{r,r}} \lesssim \sum_{j=1}^{r}\int_I(1-\lambda^2y^2)^{j-2r}(1-y^2)^r(\partial_x^jv(x))^2\omega_\lambda(x)dx, \tag{3.6}$$

where $y = g(x;\lambda) \in \bar{I}$. Since $1-y^2 \le 1-\lambda^2 y^2$ and $A^0_\lambda(I) = L^2_{\omega_\lambda}(I)$, we conclude that

$$|v|_{A^r_\lambda} \lesssim (1-\lambda^2)^{\frac{1}{2}-\frac{r}{2}} \|v\|_{r,\omega_\lambda}, \quad r \ge 0, \ \ \lambda \in (0,1). \tag{3.7}$$

Similarly, we derive from by the definition of $B^r_\lambda(I)$ that

$$|v|_{B^r_\lambda} \lesssim (1-\lambda^2)^{-\frac{r}{2}} \|v\|_{r,\omega_\lambda}, \quad r \ge 1, \ \ \lambda \in (0,1). \tag{3.8}$$

Next, we derive from (3.1) and (3.2) that the values of $d_{\lambda,i}, 1 \le i \le 3$, in (2.32), (2.35), and (2.38) are

$$d_{\lambda,1} = \frac{a}{\lambda}, \qquad d_{\lambda,2} = \frac{a\lambda}{\sqrt{1-\lambda^2}}, \qquad d_{\lambda,3} = -a^2. \tag{3.9}$$

In summary, we have proved the following results.

COROLLARY 3.1. *For any $v \in H^r_{\omega_\lambda}(I)$, $\lambda \in (0,1)$, and $r \ge 1$,*

$$\|\partial_x(P_{N,\lambda}v - v)\|_{\tilde{\omega}_\lambda} + N\|P_{N,\lambda}v - v\|_{\omega_\lambda} \lesssim (1-\lambda^2)^{\frac{1}{2}-\frac{r}{2}} N^{1-r} \|v\|_{r,\omega_\lambda}, \tag{3.10}$$

$$\|P^1_{N,\lambda}v - v\|_{1,\omega_\lambda} \lesssim (1-\lambda^2)^{-\frac{r}{2}} N^{1-r} \|v\|_{r,\omega_\lambda}, \tag{3.11}$$

and for $r > 1$,

$$\|\partial_x(\mathcal{I}_{N,\lambda}v - v)\|_{\tilde{\omega}_\lambda} + N\|\mathcal{I}_{N,\lambda}v - v\|_{\omega_\lambda} \lesssim (1-\lambda^2)^{\frac{1}{2}-\frac{r}{2}} N^{1-r} \|v\|_{r,\omega_\lambda}. \tag{3.12}$$

If $\nu_1, \nu_2 > 0$ and $v \in H^1_{0,\omega_\lambda}(I) \cap H^r_{\omega_\lambda}(I)$ with $r \ge 1$, then

$$|||P^{1,0}_{N,\lambda}v - v|||_{1,\omega_\lambda} \lesssim (\nu_1^{\frac{1}{2}}(1-\lambda^2)^{-\frac{1}{4}} + \nu_2^{\frac{1}{2}} N^{-1})(1-\lambda^2)^{-\frac{r}{2}} N^{1-r} \|v\|_{r,\omega_\lambda}. \tag{3.13}$$

Remark 3.1. For $\lambda = 0$, (1.3) becomes the identity map. So we obtain the same results as in the standard Legendre case, i.e.,

$$|v|_{A^r_\lambda} \lesssim \|(1-x^2)^{\frac{r}{2}} \partial^r_x v\|, \qquad |v|_{B^r_\lambda} \lesssim \|(1-x^2)^{\frac{r-1}{2}} \partial^r_x v\|, \qquad \lambda \to 0.$$

Remark 3.2. For $\lambda = 1$, (1.3) becomes $y = \sin(\pi x/2)$. This mapping has singularities at $x = \pm 1$, and therefore, (3.7)–(3.8) are no longer valid. However, we find from (3.1), (3.2), and (3.6) that

$$\begin{aligned} |v|_{A^r_1} &\lesssim \left(\sum_{j=1}^r \int_I (1-y^2)^{j-r} (\partial^j_x v(x))^2 \omega_\lambda(x) dx \right)^{\frac{1}{2}} \\ &\lesssim \left(\sum_{j=1}^r \int_I (1-x^2)^{j-r+\frac{1}{2}} (\partial^j_x v(x))^2 dx \right)^{\frac{1}{2}}. \end{aligned}$$

This implies that $|v|_{A^r_1}$ is bounded if $v \in H^r(I)$ and for some $\sigma < 1$,

$$(1-x^2)^{j-r+\frac{1}{2}} (\partial^j_x v(x))^2 \lesssim \frac{1}{(1-x^2)^\sigma} \quad \text{as} \quad |x| \to 1, \quad 1 \le j \le r. \tag{3.14}$$

In particular, one can verify that (3.14) is satisfied if

$$\partial^j_x v(\pm 1) = 0, \qquad 1 \le j \le r-2, \ v \in H^r(I). \tag{3.15}$$

Indeed, by the Hardy's inequality (see [12]), we have that for $\alpha < 1$,

$$\int_I u^2(x)(1-x^2)^{\alpha-1}dx \lesssim \int_I u^2(x)(1-x^2)^{\alpha-2}dx \lesssim \int_I (\partial_x u(x))^2(1-x^2)^{\alpha}dx,$$

provided that $u(\pm 1) = 0$ and the right-hand side of the inequality is finite. Using this equality and the condition (3.15), we have for $1 \le j \le r-2$,

$$\begin{aligned}\int_I (\partial_x^j v(x))^2\ (1-x^2)^{j-r+\frac{1}{2}}dx &\lesssim \int_I (\partial_x^{j+1} v(x))^2(1-x^2)^{j-r+\frac{3}{2}}dx\\ &\lesssim \cdots \lesssim \int_I (\partial_x^{r-1} v(x))^2(1-x^2)^{-\frac{1}{2}}dx.\end{aligned}$$

Then, by the inequality (13.5) in [5], we have that for $\alpha > 1$,

$$\int_I u^2(x)(1-x^2)^{\alpha-2}dx \lesssim \int_I ((\partial_x u(x))^2 + u^2(x))(1-x^2)^{\alpha}dx,$$

which implies that

$$\int_I (\partial_x^{r-1} v(x))^2(1-x^2)^{-\frac{1}{2}}dx \lesssim \int_I ((\partial_x^r v(x))^2 + (\partial_x^{r-1} v(x))^2)(1-x^2)^{\frac{3}{2}}dx.$$

A combination of the above estimates show that under the condition (3.15), we have

$$|v|_{A_1^r} \lesssim (\|(1-x^2)^{\frac{3}{4}}\partial_x^{r-1} v\| + \|(1-x^2)^{\frac{1}{4}}\partial_x^r v\|). \tag{3.16}$$

Similar results can also be derived for $|v|_{B_1^r}$. The estimates indicate, in particular, that, for the mapping (1.3) with $\lambda = 1$, the convergence rate of the mapped Legendre method is of order r if (3.15) is satisfied. In particular, only a second-order convergence rate can be expected if the function does not vanish at the end-points.

Remark 3.3. If the parameter λ was chosen as (cf. [1] and [10])

$$\lambda = \lambda(N,\epsilon) = \operatorname{sech}\Big(\frac{|\ln\epsilon|}{N}\Big) = \frac{2}{\epsilon^{1/N} + \epsilon^{-1/N}} \sim 1 - \frac{1}{2}(\ln^2\epsilon)N^{-2} \quad \text{for } N \gg 1, \tag{3.17}$$

where ϵ is the desired accuracy, then we find from (3.7) and (3.8) that for any $v \in H_{\omega_\lambda}^r(I)$,

$$|v|_{A_\lambda^r} \sim (|\ln\epsilon|)^{1-r}N^{r-1}, \quad |v|_{B_\lambda^r} \sim (|\ln\epsilon|)^{-r}N^r,$$

which, along with Corollary 3.1, implies that

$$\|P_{N,\lambda}v - v\|_{\omega_\lambda} \sim |\ln\epsilon|^{1-r}N^{-1}, \quad \|P_{N,\lambda}^1 v - v\|_{1,\omega_\lambda} \sim |\ln\epsilon|^{-r}N.$$

Thus a lower order ($< r$) of accuracy is expected by choosing (3.17), except when ϵ and N are such that $\epsilon \lesssim \exp(-\gamma N)$, $\gamma > 0$.

3.2. The mapping (1.4). In this case, $D_\lambda = \{\lambda \mid \lambda > 0\}$, and (1.4) is

$$y = h(x;\lambda) = (4/\pi)\arctan(\lambda\tan(\pi(x-1)/4)) + 1, \qquad \lambda > 0, \tag{3.18}$$

$$\omega_\lambda(x) = \frac{dy}{dx} = \frac{\lambda}{1 + (\lambda^2-1)\sin^2(\pi(x-1)/4)}. \tag{3.19}$$

In particular, $h(x;\lambda)=x$ and $\omega_\lambda(x)=1$ for $\lambda=1$.

Let us denote

$$C_\lambda=\begin{cases}\lambda^{-1}, & 0<\lambda\le 1,\\ \lambda, & \lambda>1,\end{cases}\quad C_1=\begin{cases}1, & 0<\lambda\le 1,\\ \lambda^{-2}, & \lambda>1,\end{cases}\quad C_2=\begin{cases}\lambda^2, & 0<\lambda\le 1,\\ 1, & \lambda>1.\end{cases}$$

We have $C_1,\ C_2\le 1$ for $\lambda>0$, and by (3.19),

$$\begin{aligned}\frac{dx}{dy}&=(\cos^2(\pi(x-1)/4)+\lambda^2\sin^2(\pi(x-1)/4))/\lambda\\ &=C_\lambda(C_1\cos^2(\pi(x-1)/4)+C_2\sin^2(\pi(x-1)/4)).\end{aligned}\tag{3.20}$$

We set

$$\mathcal{T}_l=\operatorname{span}\{\cos(k\pi(x-1)/4),\ \sin(k\pi(x-1)/4),\ 1\le k\le l\},\quad l\in\mathbb{N}.$$

For $j\ge 1$, we denote by $T_{k,j}(x)$ some functions in $\mathcal{T}_l$ with coefficients in terms of C_1 and C_2. Then, by (3.18) and (3.20),

$$\begin{aligned}\partial_y V_\lambda(y)&=\partial_x v(x)\frac{dx}{dy}=C_\lambda T_{2,1}(x)\partial_x v(x),\\ \partial_y^2 V_\lambda(y)&=\frac{d(\partial_y V_\lambda(y))}{dx}\frac{dx}{dy}=C_\lambda^2(T_{4,2}(x)\partial_x^2 v(x)+T_{4,1}(x)\partial_x v(x)).\end{aligned}$$

Hence, by an induction argument, we find that for $k\ge 1$,

$$\partial_y^k V_\lambda(y)=C_\lambda^k\sum_{j=1}^{k}T_{k,j}(x)\partial_x^j v(x),\tag{3.21}$$

where $T_{k,j}(x)$ $(1\le j\le k)$ are uniformly bounded for all $x\in\bar I$ and $\lambda>0$. Let

$$\omega_\lambda^{(r)}(x):=\omega_\lambda(x)(1-y^2)^r=\omega_\lambda(x)(1-h^2(x;\lambda))^r\ (\lesssim\omega_\lambda(x),\quad x\in I,\ \lambda>0).$$

By (3.18) and (3.20),

$$\lim_{x\to 1}\frac{1-h(x;\lambda)}{1-x}=\lambda,\qquad \lim_{x\to -1}\frac{1+h(x;\lambda)}{1+x}=\lambda^{-1}.$$

By virtue of (3.21) and the definitions of $|v|_{A_\lambda^r}$ and $|v|_{B_\lambda^r}$, we derive that

$$|v|_{A_\lambda^r}\lesssim C_\lambda^r\|v\|_{r,\omega_\lambda^{(r)}},\qquad |v|_{B_\lambda^r}\le C_\lambda^r\|v\|_{r,\omega_\lambda^{(r-1)}},\quad \lambda>0.\tag{3.22}$$

Next, we deduce from (3.18)–(3.20) that the values of the constants $d_{\lambda,i},\ 1\le i\le 3$, are

$$d_{\lambda,1}=C_\lambda,\quad d_{\lambda,2}=\frac{\pi/4|\lambda^2-1|\,|\sin(\pi(x-1)/2)|}{\cos^2(\pi(x-1)/4)+\lambda^2\sin^2(\pi(x-1)/4)}\le\frac{\pi|\lambda^2-1|}{4\lambda},$$

$$d_{\lambda,3}=\begin{cases}\max\{\frac{\pi^2}{8}(1-\lambda^2),S_\lambda(z_0)\} & \text{if}\quad 0<\lambda\le e_0,\\ \frac{\pi^2}{8}(1-\lambda^2) & \text{if}\quad e_0<\lambda\le 1,\\ \frac{\pi^2(\lambda^2-1)}{8\lambda^2} & \text{if}\quad 1<\lambda\le e_1,\\ \max\{\frac{\pi^2(\lambda^2-1)}{8\lambda^2},S_\lambda(z_0)\} & \text{if}\quad \lambda>e_1,\end{cases}\tag{3.23}$$

where $e_0 = \sqrt{\frac{\sqrt{97}-5}{6}}, e_1 = \sqrt{\frac{5}{3}}, z_0 = \frac{5\lambda^2-3}{3(1-\lambda^4)}$, and

$$S_\lambda(z) = \frac{\pi^2 b(-2bz^2 + (3b-2)z + 1)}{8(1-bz)^2}, \quad \text{with} \quad b = 1-\lambda^2.$$

The estimate on $d_{\lambda,2}$ is derived using a simple inequality $a^2 + b^2 \geq 2ab$. The estimate on $d_{\lambda,3}$ is nontrivial and its derivation is given in Appendix A.

In summary, we obtained the following approximation results for mapping (1.4).

COROLLARY 3.2. *For any $v \in H^r_{\omega_\lambda^{(r)}}(I)$, $\lambda \in (0,1)$, and $r \geq 1$,*

$$\|\partial_x(P_{N,\lambda}v - v)\|_{\tilde{\omega}_\lambda} + N\|P_{N,\lambda}v - v\|_{\omega_\lambda} \lesssim C_\lambda^r N^{1-r}\|v\|_{r,\omega_\lambda^{(r)}}, \tag{3.24}$$

and for $r > 1$,

$$\|\partial_x(\mathcal{I}_{N,\lambda}v - v)\|_{\tilde{\omega}_\lambda} + N\|\mathcal{I}_{N,\lambda}v - v\|_{\omega_\lambda} \lesssim C_\lambda^r N^{1-r}\|v\|_{r,\omega_\lambda^{(r)}}, \tag{3.25}$$

while for any $v \in H^r_{\omega_\lambda^{(r-1)}}(I)$ and $r \geq 1$,

$$\|P^1_{N,\lambda}v - v\|_{1,\omega_\lambda} \lesssim C_\lambda^{r+1} N^{1-r}\|v\|_{r,\omega_\lambda^{(r-1)}}. \tag{3.26}$$

If, in addition, $\nu_2 > \frac{\nu_1}{2} d_{\lambda,3}$ and $v \in H^1_{0,\omega_\lambda}(I)$, then

$$|||P^{1,0}_{N,\lambda}v - v|||_{1,\omega_\lambda} \lesssim (\nu_1^{\frac{1}{2}}(d_{\lambda,1}+1)(d_{\lambda,2}+1)^{\frac{1}{2}} + \nu_2^{\frac{1}{2}}N^{-1})N^{1-r}\|v\|_{r,\omega_\lambda^{(r-1)}}, \tag{3.27}$$

where $d_{\lambda,i}, i = 1,2,3$ are given in (3.23).

3.3. The mapping (1.5). In this case, $\lambda = (\lambda_1, \lambda_2)$ and $D_\lambda = \{(\lambda_1, \lambda_2) \mid \lambda_1 > 0, \ -1 \leq \lambda_2 < 1\}$. The mapping (1.5) is explicitly invertible:

$$y = h(x;\lambda) = a_0 + \arctan(\lambda_1(x - \lambda_2))/a_1. \tag{3.28}$$

The values of a_0 and a_1 are

$$a_0 = a_0(\lambda) = \frac{\kappa_1 - \kappa_2}{\kappa_1 + \kappa_2}, \qquad a_1 = a_1(\lambda) = \frac{\kappa_1 + \kappa_2}{2}, \tag{3.29}$$

where

$$\kappa_1 = \arctan(\lambda_1(1+\lambda_2)), \qquad \kappa_2 = \arctan(\lambda_1(1-\lambda_2)). \tag{3.30}$$

With the above choice, we find that

$$-1 \leq a_0 < 1, \qquad 0 < a_1 < \frac{\pi}{2}. \tag{3.31}$$

The weight functions are

$$\omega_\lambda(x) = \frac{dy}{dx} = \frac{\lambda_1}{a_1(1+\lambda_1^2(x-\lambda_2)^2)} = \Big(\frac{dx}{dy}\Big)^{-1} = \frac{\lambda_1}{a_1}\cos^2(a_1(y-a_0)), \tag{3.32}$$

$$\tilde{\omega}_\lambda(x) = a_1^{-1}\lambda_1^{-1}(\kappa_1 - q(x;\lambda))(\kappa_2 + q(x;\lambda))(1+\lambda_1^2(x-\lambda_2)^2), \tag{3.33}$$

where $q(x;\lambda) = \arctan(\lambda_1(x-\lambda_2))$.

For simplicity, we rewrite (3.32) as

(3.34)
$$\frac{dx}{dy} = \frac{a_1(\lambda_1+1)^2}{\lambda_1}\left(\frac{1}{(\lambda_1+1)^2} + \frac{\lambda_1^2}{(\lambda_1+1)^2}(x-\lambda_2)^2\right) := C_\lambda(D_1 + D_2(x-\lambda_2)^2).$$

Since $a_1 \to \lambda_1$ as $\lambda_1 \to 0$, we have that for all $\lambda \in D_\lambda$,

$$C_\lambda \lesssim \lambda_1 + 1, \qquad 0 < C_\lambda^{-1}, D_1, D_2 \le 1. \tag{3.35}$$

Let us denote by $Q_l(x-\lambda_2)$ a polynomial of degree l with respect to $x-\lambda_2$ with coefficients in terms of D_1, D_2, and C_λ^{-1}. Then, by (3.28) and (3.34),

$$\partial_y V_\lambda(y) = \partial_x v(x)\frac{dx}{dy} = C_\lambda Q_2(x-\lambda_2)\partial_x v(x),$$
$$\partial_y^2 V_\lambda(y) = \frac{d(\partial_y V_\lambda(y))}{dx}\frac{dx}{dy} = C_\lambda^2(Q_4(x-\lambda_2)\partial_x^2 v(x) + Q_3(x-\lambda_2)\partial_x v(x)).$$

Hence, by an induction argument, we find that for $k \ge 1$,

$$\partial_y^k V_\lambda(y) = C_\lambda^k \sum_{j=1}^{k} Q_{k+j}(x-\lambda_2)\partial_x^j v(x), \tag{3.36}$$

where $Q_{k+j}(x-\lambda_2)$ $(1 \le j \le k)$ are uniformly bounded for all $x \in \bar{I}$ and $\lambda \in D_\lambda$. Let us denote

(3.37)
$$S_\lambda(x;r) := (1-y^2)^r = \frac{1}{a_1^{2r}}\Big(\kappa_2 - \arctan(\lambda_1(x-\lambda_2))\Big)^r\Big(\kappa_1 + \arctan(\lambda_1(x-\lambda_2))\Big)^r.$$

We have $S_\lambda(x;r) \le 1$, and

$$\lim_{x\to 1}\frac{\kappa_2 - \arctan(\lambda_1(x-\lambda_2))}{1-x} = \frac{\lambda_1}{1+\lambda_1^2(1-\lambda_2)^2},$$

$$\lim_{x\to -1}\frac{\kappa_1 + \arctan(\lambda_1(x-\lambda_2))}{1+x} = \frac{\lambda_1}{1+\lambda_1^2(1+\lambda_2)^2}.$$

Consequently,

(3.38)
$$\lim_{|x|\to 1} S_\lambda(x;r) = G_\lambda^r(1-x^2)^r \quad \text{with} \quad G_\lambda = \frac{\lambda_1^2}{a_1^{2r}(1+\lambda_1^2(1-\lambda_2)^2)(1+\lambda_1^2(1+\lambda_2)^2)}.$$

Next, let

$$\varpi_\lambda^{(r)}(x) := \omega_\lambda(x)S_\lambda(x;r) \lesssim \omega_\lambda(x), \quad x \in I,\ \lambda \in D_\lambda.$$

By the definition of $|v|_{A_\lambda^r}$ and (3.35),

$$|v|_{A_\lambda^r} \lesssim (\lambda_1+1)^r \|v\|_{r,\varpi_\lambda^{(r)}}, \qquad r \ge 0, \quad \lambda \in D_\lambda. \tag{3.39}$$

Similarly, we deduce that

$$|v|_{B^r_\lambda} \lesssim (\lambda_1+1)^r \|v\|_{r,\varpi^{(r-1)}_\lambda}, \qquad r\geq 1, \quad \lambda\in D_\lambda. \tag{3.40}$$

Remark 3.4. We find from (3.28)–(3.30) and (3.32) that

$$a_0\to\lambda_2, \quad a_1\to 1, \quad h(x;\lambda)\to x, \quad \omega_\lambda(x)\to 1, \quad \text{as} \quad \lambda_1\to 0.$$

So we have the same estimate as for the standard Legendre case:

$$|v|_{A^r_\lambda} \lesssim \|(1-x)^{\frac{r}{2}}\partial_x^r v\|, \qquad |v|_{B^r_\lambda} \lesssim \|(1-x)^{\frac{r-1}{2}}\partial_x^r v\|.$$

Remark 3.5. We observe from the derivation of (3.36) that if the function v possesses certain special properties as specified below, more precise estimates can be derived. For instance, if the rapid variational region of $v(x)$ is contained in $O_\varepsilon(\lambda_2) := (\lambda_2-\varepsilon,\lambda_2+\varepsilon)$ for some $\varepsilon>0$, we can assume

$$\sup_{x\in I_\epsilon} |\partial_x^j v(x)| \leq \delta(\varepsilon) C_\lambda^{-r}, \qquad 0\leq j\leq r,$$

where $I_\varepsilon = \bar I \setminus O_\varepsilon(\lambda_2)$, and $\delta(\varepsilon)$ is a small positive number corresponding to ε. Then

$$|v|_{A^r_\lambda} \lesssim \Big(\delta(\varepsilon) + a_1^r\lambda_1^{-r}(1+\lambda_1^2\varepsilon^2)^r \|v\|_{H^r_{\omega^{(r)}_\lambda}(O_\varepsilon(\lambda_2))}\Big).$$

In particular, if

$$\mathrm{supp}\{\partial_x^j v(x)\} \subseteq \overline{O_\varepsilon(\lambda_2)} \subseteq [-1,1], \qquad 0\leq j\leq r,$$

then we have

$$|v|_{A^r_\lambda} \lesssim \Big(a_1^r\lambda_1^{-r}(1+\lambda_1^2\varepsilon^2)^r \|v\|_{H^r_{\omega^{(r)}_\lambda}(O_\varepsilon(\lambda_2))}\Big).$$

The above analysis is also valid for $|v|_{B^r_\lambda}$.

Next, we compute the values of $d_{\lambda,i}, i=1,2,3$. Using (3.28) and (3.32) yields

$$d_{\lambda,1} = a_1\lambda_1^{-1}, \quad d_{\lambda,2} = \max_{x\in\bar I} \frac{2\lambda_1^2|x-\lambda_2|}{1+\lambda_1^2(x-\lambda_2)^2} \leq \lambda_1, \quad d_{\lambda,3} \leq \frac{3}{2}\lambda_1^2. \tag{3.41}$$

The derivation of $d_{\lambda,3}$ is a little complicated, so we defer it to Appendix B.

A combination of Theorems 2.1–2.4 and the above estimates leads to the following approximation results.

COROLLARY 3.3. *For any $v\in H^r_{\varpi^{(r)}_\lambda}(I)$, $\lambda\in D_\lambda$, and $r\geq 1$,*

$$\|\partial_x(P_{N,\lambda}v-v)\|_{\tilde\omega_\lambda} + N\|P_{N,\lambda}v-v\|_{\omega_\lambda} \lesssim (\lambda_1+1)^r N^{1-r}\|v\|_{r,\varpi^{(r)}_\lambda}, \tag{3.42}$$

and for $r>1$,

$$\|\partial_x(\mathcal{I}_{N,\lambda}v-v)\|_{\tilde\omega_\lambda} + N\|\mathcal{I}_{N,\lambda}v-v\|_{\omega_\lambda} \lesssim (\lambda_1+1)^r N^{1-r}\|v\|_{r,\varpi^{(r)}_\lambda}, \tag{3.43}$$

while for any $v\in H^r_{\varpi^{(r-1)}_\lambda}(I)$ and $r\geq 1$,

$$\|P^1_{N,\lambda}v-v\|_{1,\omega_\lambda} \lesssim (\lambda_1+1)^r N^{1-r}\|v\|_{r,\varpi^{(r-1)}_\lambda}. \tag{3.44}$$

If, in addition, $\nu_2 > \frac{\nu_1}{2}d_{\lambda,3}$ and $v\in H^1_{0,\omega_\lambda}(I)$, then

$$|||P^{1,0}_{N,\lambda}v-v|||_{1,\omega_\lambda} \lesssim (\nu_1^{\frac12}(d_{\lambda,1}+1)(d_{\lambda,2}+1)^{\frac12} + \nu_2^{\frac12}N^{-1})N^{1-r}\|v\|_{r,\varpi^{(r-1)}_\lambda}, \tag{3.45}$$

where $d_{\lambda,i}, i=1,2,3$, are given in (3.41).

3.4. Some other mappings. We consider here several useful mappings which do not quite fit into our framework.

Let us consider first the mapping

$$C(s) = (1-b)\frac{\sinh(\theta s)}{\sinh(\theta)} + b\frac{\tanh(\theta(s+1/2)) - \tanh(\theta/2)}{2\tanh(\theta/2)}, \quad s \in [-1,0], \tag{3.46}$$

which was first introduced by Song and Haidvogel [26] as part of the so-called s-coordinates in their ocean circulation model. The two parameters $0 \le \theta \le 20$ and $0 \le b \le 1$ are used to fit the surface and bottom topography.

In order to apply our general framework, we set $s = \frac{y-1}{2}$, $x = 2C(s)+1$, $\lambda_1 = \theta$, and $\lambda_2 = b$ in (3.46) to get

$$x = g(y;\lambda) = 2(1-\lambda_2)\frac{\sinh(\lambda_1(y-1)/2)}{\sinh(\lambda_1)} + \lambda_2\frac{\tanh(\lambda_1 y/2) - \tanh(\lambda_1/2)}{\tanh(\lambda_1/2)} + 1, \tag{3.47}$$
$$y \in [-1,1], \quad \lambda_1 \ge 0, \quad 0 \le \lambda_2 \le 1.$$

Clearly, it maps the interval $[-1,1]$ univalently onto itself with $g(\pm 1;\lambda) = \pm 1$, and it is an identity mapping when $\lambda_1 = 0$. However, this mapping is not explicitly invertible. For simplicity, we consider only the special case $\lambda_2 = 0$ and denote $\lambda := \lambda_1$. In this case, the weight function is

$$\omega_\lambda(x) = \Big(\frac{dx}{dy}\Big)^{-1} = \frac{\sinh\lambda}{\lambda}\mathrm{sech}(\lambda(y-1)/2) > 0, \quad x, y \in I,\ \lambda > 0. \tag{3.48}$$

One can verify readily that

$$\frac{\tanh\lambda}{\lambda} \le \omega_\lambda(x) \le \frac{\sinh\lambda}{\lambda}, \quad x \in I,\ \lambda > 0.$$

To estimate the corresponding upper bounds of $|v|_{A^r_\lambda}$ and $|v|_{B^r_\lambda}$ (cf. section 2), we can follow the same procedure as for the mapping (1.3). Since

$$\frac{d^k x}{dy^k} = \frac{2}{\sinh\lambda}\Big(\frac{\lambda}{2}\Big)^k \begin{cases} \sinh(\lambda(y-1)/2) & \text{if } k \text{ is even,} \\ \cosh(\lambda(y-1)/2) & \text{if } k \text{ is odd,} \end{cases}$$

we find

$$\Big|\frac{d^k x}{dy^k}\Big| \lesssim \lambda^k \coth\lambda, \quad x \in \bar{I},\ \lambda > 0,\ k \ge 1.$$

As in the derivations of (3.4)–(3.6), we obtain that for any $v \in H^r_{\omega_\lambda}(I)$,

$$|v|_{A^r_\lambda},\ |v|_{B^r_\lambda} \lesssim \lambda^r (\coth\lambda)^r ||v||_{r,\omega_\lambda}. \tag{3.49}$$

In view of the facts

$$\coth\lambda \sim \lambda^{-1} \text{ if } \lambda \ll 1, \quad \coth\lambda \sim 1 \text{ if } \lambda \gg 1,$$

we conclude that for small λ, this mapping is close to the identity mapping, while for large λ, an extra factor λ^r appears in the error estimates.

The mapping techniques have been successfully used in spectral methods to resolve boundary layers. For instance, the following mapping is used in [21] and [22]:

(3.50)
$$x = g(y;m) = -1 + \sigma_m \int_{-1}^{y} (1-t^2)^m dt, \ \text{ with } \sigma_m = 2\Big/\int_{-1}^{1}(1-y^2)^m dy, \ \ m \in \mathbb{N}.$$

Clearly, we have $g(\pm 1;m) = \pm 1$, and

$$\omega_m(x) = \sigma_m^{-1}(1-y^2)^{-m}, \quad x, y \in I. \tag{3.51}$$

As m increases, more and more Gauss-type collocation points are clustered near the end-points ± 1 so it is suitable for resolving very thin boundary layers. However, the mapping is singular at the end-points, which implies in particular that $d_{\lambda,1} = \infty$. The same is true for the iterated mappings introduced by Tang and Trummer [27],

$$x_0 = y, \quad x_m = \sin\Big(\frac{\pi}{2}x_{m-1}\Big), \quad m \geq 1, \tag{3.52}$$

which are very effective mappings for problems with thin boundary layers. Hence we cannot directly apply our general framework to these mappings. Although it is possible to derive some special estimates as we did in Remark 3.2, the computations would be very tedious. However, we shall consider in a forthcoming paper the mapped Jacobi method in which we will be able to handle mappings with singularities at the end-points.

4. The mapped Legendre methods for a model equation. To illustrate how the results we developed in previous sections can be applied to analyze the mapped Legendre spectral and pseudospectral methods for PDEs, we consider the following model equation:

$$\begin{cases} -\varepsilon\partial_x^2 u(x) + u(x) = f(x), & x \in I, \\ u(\pm 1) = 0. \end{cases} \tag{4.1}$$

Let $\nu_1 = \varepsilon$ and $\nu_2 = 1$ and $\omega_\lambda(x)$, $a^{(\nu)}_{\omega_\lambda}(\cdot,\cdot)$ be the same as in section 2. A weighted variational formulation for (4.1) is to find $u \in H^1_{0,\omega_\lambda}(I)$ such that

$$a^{(\nu)}_{\omega_\lambda}(u,v) = (f,v)_{\omega_\lambda} \qquad \forall v \in H^1_{0,\omega_\lambda}(I). \tag{4.2}$$

It is clear from Lemma 2.3 that, if $\varepsilon d_{\lambda,3} < 2$ and $f \in L^2_{\omega_\lambda}(I)$, (4.2) admits a unique solution.

4.1. Error estimates. The mapped Legendre spectral approximation for (4.2) is to find $u_N \in \mathcal{V}^0_{N,\lambda}$ such that

$$a^{(\nu)}_{\omega_\lambda}(u_N, v_N) = (f, v_N)_{\omega_\lambda} \qquad \forall v_N \in \mathcal{V}^0_{N,\lambda}. \tag{4.3}$$

Let $P^{1,0}_{N,\lambda}$ be the projector as in Theorem 2.3. Then, by (2.40), (4.2), and (4.3),

$$a^{(\nu)}_{\omega_\lambda}(u - u_N, v_N) = a^{(\nu)}_{\omega_\lambda}(P^{1,0}_{N,\lambda}u - u_N, v_N) = 0 \qquad \forall v_N \in \mathcal{V}^0_{N,\lambda}.$$

As a consequence of Theorem 2.3, we have the following theorem.

THEOREM 4.1. *Let u and u_N be, respectively, the solutions of* (4.2) *and* (4.3). *If* $\varepsilon d_{\lambda,3} < 2$, $u \in H^1_{0,\omega_\lambda}(I) \cap B^r_\lambda(I)$, $\lambda \in D_\lambda$, *and* $r \geq 1$, *then*

$$|||u - u_N|||_{1,\omega_\lambda} \lesssim (\varepsilon^{\frac{1}{2}}(d_{\lambda,1}+1)(d_{\lambda,2}+1)^{\frac{1}{2}} + N^{-1})N^{1-r}|u|_{B^r_\lambda}, \tag{4.4}$$

where $d_{\lambda,i}, i = 1, 2, 3$, *are the same as in Theorem* 2.3. *For the mappings* (1.3)–(1.5), *the upper bound of* $|u|_{B^r_\lambda}$ *and the values of* $d_{\lambda,i}, i = 1, 2, 3$, *are given in section* 3.

Unlike in the standard Legendre–Galerkin method, where the linear system can be made sparse by choosing suitable basis functions [24], the linear system associated to (4.3) is in general full (unless a very special mapping is used), and furthermore, it is very costly to evaluate the entries of the linear system. Hence it is often convenient to use the mapped Legendre collocation method: find $u_N \in \mathcal{V}^0_{N,\lambda}$ such that

$$-\varepsilon\partial_x^2 u_N(\xi^{(\lambda)}_{N,j}) + u_N(\xi^{(\lambda)}_{N,j}) = f(\xi^{(\lambda)}_{N,j}), \qquad 1 \leq j \leq N-1, \tag{4.5}$$

where $\{\xi^{(\lambda)}_{N,j}\}_{j=0}^N$ are the mapped LGL points defined in (2.43). Taking the discrete inner product of (4.5) with any $v_N \in \mathcal{V}^0_{N,\lambda}$, thanks to (2.44), we find that (4.5) is equivalent to the following: find $u_N \in \mathcal{V}^0_{N,\lambda}$ such that

$$\varepsilon(\partial_x u_N, \partial_x(\omega_\lambda v_N)) + (u_N, v_N)_{\omega_\lambda,N} = (f, v_N)_{\omega_\lambda,N} \qquad \forall v_N \in \mathcal{V}^0_{N,\lambda}. \tag{4.6}$$

We note that the linear system associated with the above formulation is full and ill conditioned. However, as demonstrated in [24, 25], it can be efficiently solved by using a preconditioned conjugate gradient–type iterative method with the standard Legendre–Galerkin method for (4.1) as a preconditioner. Note that with the collocation approach, there is no additional cost involved if the original PDE (4.1) has variable coefficients.

THEOREM 4.2. *Let u and u_N be, respectively, the solutions of* (4.2) *and* (4.6). *If* $\varepsilon d_{\lambda,3} < 2$, $u \in H^1_{0,\omega_\lambda}(I) \cap B^r_\lambda(I)$, *and* $f \in A^s_\lambda(I), \lambda \in D_\lambda$ *with* $r \geq 1$ *and* $s > 1$, *then*

$$|||u - u_N|||_{1,\omega_\lambda} \lesssim (\varepsilon^{\frac{1}{2}}(d_{\lambda,1}+1)(d_{\lambda,2}+1)^{\frac{1}{2}} + N^{-1})N^{1-r}|u|_{B^r_\lambda} + N^{-s}|f|_{A^s_\lambda}, \tag{4.7}$$

where $d_{\lambda,i}, i = 1, 2, 3$, *are the same as in Theorem* 2.3. *For the mappings* (1.3)–(1.5), *the upper bounds of* $|u|_{B^r_\lambda}, |f|_{A^s_\lambda}$ *and the values of* $d_{\lambda,i}, i = 1, 2, 3$, *are given in section* 3.

Proof. Let $U_N = P^{1,0}_{N,\lambda}u$ and $\hat{e}_N = U_N - u_N$. Then by (2.40), (4.2), and (4.6),

$$\begin{aligned} \varepsilon(\partial_x \hat{e}_N, \partial_x(\omega_\lambda v_N)) + (\hat{e}_N, v_N)_{\omega_\lambda,N} &= (U_N, v_N)_{\omega_\lambda,N} - (U_N, v_N)_{\omega_\lambda} \\ &\quad + (f, v_N)_{\omega_\lambda} - (f, v_N)_{\omega_\lambda,N} \qquad \forall v_N \in \mathcal{V}^0_{N,\lambda}. \end{aligned} \tag{4.8}$$

Taking $v_N = \hat{e}_N$ in (4.8), we have from (2.45) that

$$\begin{aligned} a^{(\nu)}_\lambda(\hat{e}_N, \hat{e}_N) &\leq \varepsilon(\partial_x \hat{e}_N, \partial_x(\omega_\lambda \hat{e}_N)) + \|\hat{e}_N\|^2_{\omega_\lambda,N} \\ &\leq |(U_N, \hat{e}_N)_{\omega_\lambda,N} - (U_N, \hat{e}_N)_{\omega_\lambda}| + |(f, \hat{e}_N)_{\omega_\lambda} - (f, \hat{e}_N)_{\omega_\lambda,N}|. \end{aligned} \tag{4.9}$$

By Remark 2.3,

$$|(f, \hat{e}_N)_{\omega_\lambda} - (f, \hat{e}_N)_{\omega_\lambda,N}| \lesssim N^{-s}|f|_{A^s_\lambda}\|\hat{e}_N\|_{\omega_\lambda}.$$

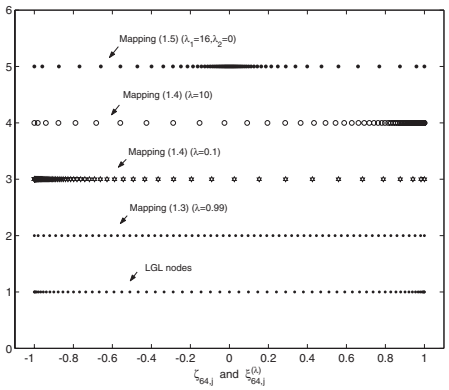


FIG. 1. *LGL vs. MLGL points.*

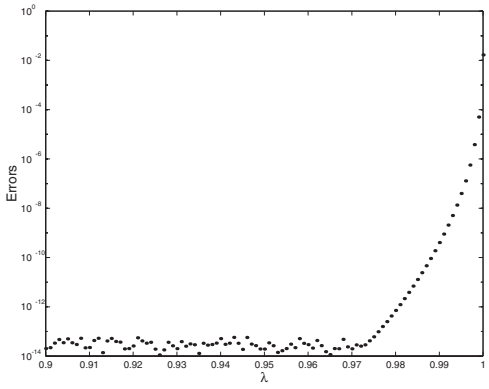


FIG. 2. *Errors for mapping* (1.3).

Moreover, by (2.44), (2.45), and Theorem 2.3,

$$\begin{aligned}|(U_N,\hat{e}_N)_{\omega_\lambda,N}-(U_N,\hat{e}_N)_{\omega_\lambda}| &\le |(U_N-P^{1,0}_{N-1,\lambda}u,\hat{e}_N)_{\omega_\lambda,N}|+|(U_N-P^{1,0}_{N-1,\lambda}u,\hat{e}_N)_{\omega_\lambda}|\\ &\lesssim \|U_N-P^{1,0}_{N-1,\lambda}u\|_{\omega_\lambda}\|e_N\|_{\omega_\lambda}\lesssim (|||U_N-u|||_{1,\omega_\lambda}+|||P^{1,0}_{N-1,\lambda}u-u|||_{1,\omega_\lambda})\|\hat{e}_N\|_{\omega_\lambda}\\ &\lesssim (\varepsilon^{\frac{1}{2}}(d_{\lambda,1}+1)(d_{\lambda,2}+1)^{\frac{1}{2}}+N^{-1})N^{1-r}|u|_{B^r_\lambda}\|\hat{e}_N\|_{\omega_\lambda}.\end{aligned}$$

The desired results follow from the above estimates. □

4.2. Numerical results. We now present some numerical results with emphasis on how the accuracy depends on the choice of the parameters in the mappings (1.3)–(1.5).

We first illustrate the effects of the parameters on the distributions of the MLGL points. In Figure 1, we plot the LGL points vs. the MLGL points ($N = 64$) with several typical parameters. It is clear that the mapping (1.3) stretches the grid evenly as $\lambda \to 1$; the mapping (1.4) clusters the points to $x = -1$ (resp., $x = 1$) for $\lambda < 1$ (resp., $\lambda > 1$); and the mapping (1.5) clusters the points to $x = \lambda_2$ for $\lambda_1 > 1$.

Example 1. We consider (4.1) with the exact solution

$$u(x) = \frac{e^{(1+x)/\sqrt{\varepsilon}} - e^{-(1+x)/\sqrt{\varepsilon}}}{e^{2/\sqrt{\varepsilon}} - e^{-2/\sqrt{\varepsilon}}} - \frac{1+x}{2}.$$

This solution exhibits a boundary layer of width $O(\sqrt{\varepsilon})$ at $x = -1$.

We first take $\varepsilon = 0.1$ so the solution is smoothly varying throughout the domain. We use (4.5) with the mapping (1.3) and $N = 100$ to approximate (4.1). In Figure 2, we plot the maximum absolute errors between u and u_N at the MLGL points with $\lambda \in [0.9, 1]$. We see that the error increases very quickly as $\lambda \to 1$, which is in agreement with the theoretic analysis in Corollary 3.1 and Theorem 4.2. Hence it is not advisable to use mapping (1.3) with λ close to 1.

Next we take $\varepsilon = 10^{-8}$ so the solution has a thin boundary layer at $x = -1$, and we use (4.5) with the mapping (1.4) and $N = 100$ to approximate (4.1). We plot in Figure 3 the errors with $\lambda \in [20, 200]$. The results indicate that the errors grow as λ increases, as predicted by Corollary 3.2 and Theorem 4.2.

Example 2. We take $u(x) = \tanh(ax)$ with $a = 150$. This solution has a large derivative at $x = 0$; see Figure 4. We use (4.5) with mapping (1.5) and $N = 100$. In

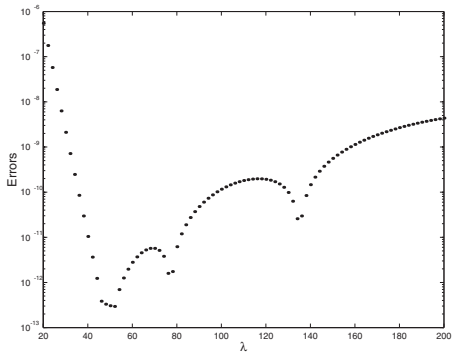


FIG. 3. *Errors for mapping* (1.4).

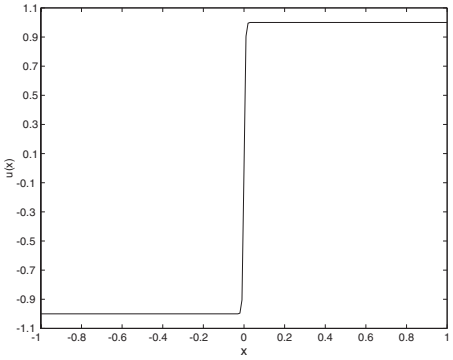
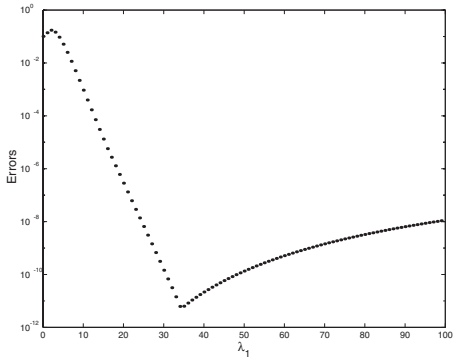


FIG. 4. *Solution in Example* 2.



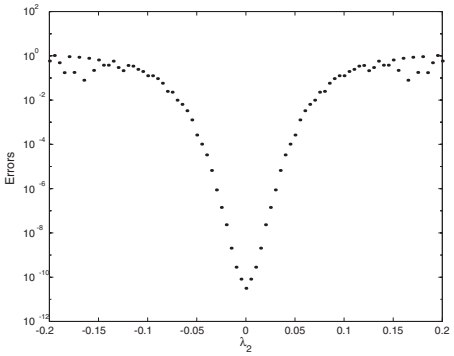


FIG. 5. *Errors for mapping* (1.5) *with* $\lambda_2 = 0$ *and various* λ_1 *(left), and with* $\lambda_1 = 51$ *and various* λ_2 *(right).*

Figure 5, we plot the maximum absolute errors at the MLGL points with $\lambda_2 = 0$ and various λ_1 (left panel) and with $\lambda_1 = 51$ and different λ_2 (right panel). Note that the accuracy is very sensitive to the choice of the parameter λ_2, which should be at the location of large variation, but less sensitive to the values of λ_1, which represents the intensity of the mapping at $x = \lambda_2$. Again, the numerical results are in agreement with Corollary 3.3 and Theorem 4.2.

5. Concluding remarks. We presented a general framework for analyzing the approximation properties of mapped Legendre polynomials and of interpolations based on MLGL points and derived optimal error estimates for general mappings. More precisely, we introduced a new family of orthogonal functions which are obtained by applying the mapping to Legendre polynomials, and we analyzed various projection and interpolation operators based on these mapped Legendre functions.

As an application of our general results, we considered the popular mappings (1.3)–(1.5) introduced in [20, 3, 4] and derived error estimates featuring explicit expressions on the mapping parameters. We used a model equation to show that these results not only play an important role in numerical analysis of mapped Legendre spectral and pseudospectral methods for differential equations but also provide quantitative criteria for the choice of parameters in these mappings.

This paper is a first step toward a long-term goal of designing a robust adaptive spectral method for solving PDEs.

Appendix A. The derivation of $d_{\lambda,3}$ in (3.23). For $\lambda = 1$, we have $d_{\lambda,3} \equiv 0$. We next consider $\lambda \neq 1$. For simplicity, let $z = \sin^2(\pi(x-1)/4)$ and $b = 1 - \lambda^2$. By (3.19) and a direct calculation, we find that

$$S_\lambda(z) := \omega_\lambda^{-1}(x)\partial_x^2\omega_\lambda(x) = \frac{\pi^2 b(-2bz^2 + (3b-2)z + 1)}{8(1-bz)^2}$$

and

$$S'_\lambda(z) = \frac{\pi^2 b(3b(b-2)z + 5b - 2)}{8(1-bz)^3}.$$

Let us denote

$$z_0 = -\frac{5b-2}{3b(b-2)} = \frac{5\lambda^2 - 3}{3(1-\lambda^4)}, \quad e_0 = \sqrt{\frac{\sqrt{97}-5}{6}}, \quad e_1 = \sqrt{\frac{5}{3}}.$$

We find that

$$\text{if } \ 0 < \lambda \le e_0 \quad \text{or} \quad \lambda \ge e_1, \ \text{ then } |z_0| \le 1.$$

Hence

$$\begin{aligned} d_{\lambda,3} &= \max_{z\in[0,1]} S_\lambda(z) = \max\{S_\lambda(0), S_\lambda(z_0), S_\lambda(1)\} \\ &= \begin{cases} \max\{\frac{\pi^2}{8}(1-\lambda^2), S_\lambda(z_0)\} & \text{if} \quad 0 < \lambda \le e_0, \\ \frac{\pi^2}{8}(1-\lambda^2) & \text{if} \quad e_0 < \lambda \le 1, \\ \frac{\pi^2(\lambda^2-1)}{8\lambda^2} & \text{if} \quad 1 < \lambda \le e_1, \\ \max\{\frac{\pi^2(\lambda^2-1)}{8\lambda^2}, S_\lambda(z_0)\} & \text{if} \quad \lambda > e_1. \end{cases} \end{aligned}$$

Appendix B. The derivation of $d_{\lambda,3}$ in (3.41). By (3.28), (3.32), and a direct calculation, we find

$$W_\lambda(x) := \omega_\lambda^{-1}(x)\partial_x^2\omega_\lambda(x) = \frac{2\lambda_1^2(3\lambda_1^2(x-\lambda_2)^2 - 1)}{(1+\lambda_1^2(x-\lambda_2)^2)^2}$$

and

$$\partial_x W_\lambda(x) = \frac{4\lambda_1^2(x-\lambda_2)(5 - 3\lambda_1^2(x-\lambda_2)^2)}{(1+\lambda_1^2(x-\lambda_2)^2)^3}.$$

Clearly, $W_\lambda(\lambda_2) = -2\lambda_1^2$, and if x is such that $\lambda_1^2(x-\lambda_2)^2 = \frac{5}{3}$, we have $W_\lambda(x) = \frac{9}{8}\lambda_1^2$. Hence

$$W_\lambda(\lambda_2) \le W_\lambda(\pm 1) = \frac{2\lambda_1^2(3\lambda_1^2(\pm 1 - \lambda_2)^2 - 1)}{(1+\lambda_1^2(\pm 1-\lambda_2)^2)^2} \le 6\lambda_1^2\Big(\frac{\lambda_1|\pm 1 - \lambda_2|}{1+\lambda_1^2(\pm 1-\lambda_2)^2}\Big)^2 \le \frac{3}{2}\lambda_1^2.$$

Therefore,

$$-2\lambda_1^2 \le W_\lambda(x) \le \frac{3}{2}\lambda_1^2, \qquad x \in \bar{I}.$$

This completes the proof. ∎

REFERENCES

[1] M. R. Abril-Raymundo and B. Garcí-Archilla, *Approximation properties of a mapped Chebyshev method*, Appl. Numer. Math., 32 (2000), pp. 119–136.

[2] I. Babuška and B. Guo, *Optimal estimates for lower and upper bounds of approximation errors in the p-version of the finite element method in two dimensions*, Numer. Math., 85 (2000), pp. 219–255.

[3] A. Bayliss, D. Gottlieb, B. J. Matkowsky, and M. Minkoff, *An adaptive pseudospectral method for reaction diffusion problems*, J. Comput. Phys., 81 (1989), pp. 421–443.

[4] A. Bayliss and E. Turkel, *Mappings and accuracy for Chebyshev pseudo-spectral approximations*, J. Comput. Phys., 101 (1992), pp. 349–359.

[5] C. Bernardi and Y. Maday, *Spectral method*, in Handbook of Numerical Analysis Vol. 5, Handb. Numer. Anal. 5, P. G. Ciarlet and L. L. Lions, eds., North–Holland, Amsterdam, 1997, pp. 209–485.

[6] J. P. Boyd, *Orthogonal rational functions on a semi-infinite interval*, J. Comput. Phys., 70 (1987), pp. 63–88.

[7] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover Publications, Mineola, NY, 2001.

[8] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1987.

[9] P. G. Dinesen, J. S. Hesthaven, and J. P. Lynov, *A pseudospectral collocation time-domain method for diffractive optics*, in Proceedings of the Fourth International Conference on Spectral and High Order Methods (Herzliya, 1998), Appl. Numer. Math., 33 (2000), pp. 199–206.

[10] W. S. Don and D. Gottlieb, *Spectral simulation of supersonic reactive flows*, SIAM J. Numer. Anal., 35 (1998), pp. 2370–2384.

[11] D. Funaro, *Polynomial Approximations of Differential Equations*, Springer-Verlag, New York, 1992.

[12] J. E. Littlewood, G. H. Hardy, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1952.

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

[14] C. E. Grosch and S. A. Orszag, *Numerical solution of problems in unbounded regions: Coordinates transforms*, J. Comput. Phys., 25 (1977), pp. 273–296.

[15] B.-Y. Guo and L. L. Wang, *Jacobi interpolation approximations and their applications to singular differential equations*, Adv. Comput. Math., 14 (2001), pp. 227–276.

[16] B.-Y. Guo, J. Shen, and Z.-Q. Wang, *Chebyshev rational spectral and pseudospectral methods on a semi-infinite interval*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 65–84.

[17] B.-Y. Guo, *Jacobi approximations in certain Hilbert spaces and their applications to singular differential equations*, J. Math. Anal. Appl., 243 (2000), pp. 373–408.

[18] B.-Y. Guo, J. Shen, and Z.-Q. Wang, *A rational approximation and its applications to differential equations on the half line*, J. Sci. Comput., 15 (2000), pp. 117–147.

[19] B.-Y. Guo and L.-L. Wang, *Jacobi approximations in certain Besov spaces*, J. Approx. Theory, submitted.

[20] D. Kosloff and H. Tal-Ezer, *A modified Chebyshev pseudospectral method with an $o(n^{-1})$ time step restriction*, J. Comput. Phys., 104 (1993), pp. 457–469.

[21] W. B. Liu and J. Shen, *A new efficient spectral-Galerkin method for singular perturbation problems*, J. Sci. Comput., 11 (1996), pp. 411–437.

[22] W. B. Liu and T. Tang, *Error analysis for a Galerkin spectral method with coordinate transformation for solving singularly perturbed problems*, Appl. Numer. Math., 38 (2001), pp. 315–345.

[23] J. L. Mead and R. A. Renaut, *Accuracy, resolution, and stability properties of a modified Chebyshev method*, SIAM J. Sci. Comput., 24 (2002), pp. 143–160.

[24] J. Shen, *Efficient spectral-Galerkin method* I. *Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[25] J. Shen, *Efficient Chebyshev-Legendre Galerkin methods for elliptic problems*, in Proceedings of ICOSAHOM'95, A. V. Ilin and R. Scott, eds., Houston J. Math., Houston, 1996, pp. 233–240.

[26] Y. H. Song and D. Haidvogel, *A semi-implicit ocean circulation model using a generalized topography-following coordinate system*, J. Comput. Phys., 115 (1994), pp. 228–244.

[27] T. Tang and M. R. Trummer, *Boundary layer resolving pseudospectral methods for singular pertubation problems*, SIAM J. Sci. Comput., 17 (1996), pp. 430–438.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 350–382

CONVERGENCE OF A SEMI-LAGRANGIAN SCHEME FOR THE ONE-DIMENSIONAL VLASOV–POISSON SYSTEM*

NICOLAS BESSE†

Abstract. A semi-Lagrangian scheme is proposed for solving the periodic one-dimensional Vlasov–Poisson system in phase space on unstructured meshes. The distribution function $f(t, x, v)$ and the electric field $E(t, x)$ are shown to converge to the exact solution values in the L^∞ norm. The rate of convergence is in $O(h^{4/3})$.

Key words. Vlasov–Poisson system, semi-Lagrangian methods, convergence analysis

AMS subject classifications. 65M12, 82D10

DOI. 10.1137/S0036142902410775

1. Introduction.

The numerical resolution of the Vlasov equation is usually performed by Lagrangian methods like particles-in-cell methods (PIC), which consist of approximating the plasma by a finite number of macroparticles. The trajectories of these particles are computed from the characteristic curves given by the Vlasov equation, whereas self-consistent fields are computed by gathering the charge and current densities of the particles on a mesh of the physical space (see Birdsall and Langdon [10] for more details). Although this method allows us to obtain satisfying results with a small number of particles, it is well known that the numerical noise inherent to the particle method becomes too large to allow a precise description of the tail of the distribution function, which plays an important role in charged particle beams. To remedy this problem, Eulerian methods have been proposed which consist of discretizing the Vlasov equation on a mesh of phase space. For example, finite volume schemes, which are known to be robust and computationally cheap, have been implemented by Boris and Book [11], Cheng and Knorr [13], and more recently Mineau [32], Fijalkow [19], and Filbet, Sonnendrücker, and Bertrand [21]. Nevertheless, finite volume schemes are low order, too dissipative, and restricted by a CFL condition.

Other kinds of Eulerian method are the semi-Lagrangian methods which, in some particular cases, can be regarded as local versions of characteristic Galerkin methods [3, 4], which have been used in convection-diffusion problems [17, 35, 25]. Semi-Lagrangian methods were introduced at the beginning of the 1980s for the time-advection of various atmospheric and fluid dynamics models [43, 42, 37], which can be formulated as abstract Liouville systems (ALS). Semi-Lagrangian advection attempts to combine the advantages of both Eulerian and Lagrangian advection schemes while ameliorating their drawbacks. Eulerian advection schemes have good resolution properties, but CFL condition number, which is a necessary condition for achieving numerical stability, often leads to overly restrictive time steps. On the other hand, Lagrangian advection schemes allow one to use larger time steps, but, at later times, Lagrangian distortion (an initial regularly spaced set of particles will generally become highly irregularly spaced over long times) implies that important features of the flow

*Received by the editors July 5, 2002; accepted for publication (in revised form) June 11, 2003; published electronically March 3, 2004.

http://www.siam.org/journals/sinum/42-1/41077.html

†Commissariat à l'Energie Atomique, Direction des Applications Militaires–Ile de France, BP No 12, F-91680 Bruyères-le-Châtel, France (nicolas.besse@cea.fr). Current address: Institut de Recherche Mathématique Avancée, Université Louis Pasteur - CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France (besse@math.u-strasbg.fr).

may not be well described. A semi-Lagrangian method uses a regular Cartesian mesh and different sets of particles. At each time step the set of particles is chosen such that they arrive exactly at the points of the mesh at the end of the time step, and is advected by the characteristic curves of the ALS. More precisely, the method consists of directly computing the distribution function of the ALS on a fixed Cartesian grid of phase space, by integrating (or following) the characteristic curves backward (from the end of the characteristic, which is a point of the fixed mesh, to the beginning of characteristic, during a time step) at each time step and interpolating the value at the base of the characteristics. In recent applications of semi-Lagrangian methods to lower-dimensional relativistic Vlasov–Maxwell (RVM) calculations [1, 2, 40], cubic splines are used for the interpolation scheme, linear interpolation being too dissipative. Semi-Lagrangian methods have been efficiently implemented using parallel computers [41] and give considerable promise for displaying the detailed structure of distribution functions in weak density regions.

The author extends semi-Lagrangian schemes on unstructured meshes with a different kind of high order local interpolation operator and with the possibility of having a positive and conservative method by introducing a linear combination of low order solutions and high order solutions tempered by a limiter coefficient (cf. [9]). Here we present the convergence of the method for the simplest interpolation operator, that is, the Lagrange first order interpolation operator. The scheme preserves positivity because the basis functions associated with the Lagrange first order interpolation operator are always positive. Additionally, the scheme is not limited by a CFL condition. More complicated interpolation on a triangle, which involves knowledge of the gradient of the distribution function, has been implemented successfully (cf. [9]), but it seems to be a challenge to show the convergence of these methods because we advect not only the distribution function f but also its gradients. A first result on the convergence analysis of semi-Lagrangian methods with propagation of gradients is stated in [8].

Let us note that a first work on convergence of one-dimensional particle methods is [33], where Neunzert and Wick consider nonuniform initial loadings of particles asymptotically distributed with respect to initial data. Cottet and Raviart [16] present a mathematical analysis of the particle method for solving the one-dimensional Vlasov–Poisson system, where uniform initial loadings of particles are considered. A number of additional authors have studied the convergence of particle methods for the multidimensional Vlasov–Poisson system [22, 45, 46, 49]. They have also proved convergence results on random and deterministic particle methods for the Vlasov–Poisson–Fokker–Planck kinetic equations [26, 27]. Finally, Glassey and Schaeffer have done the convergence analysis of a particle method for the RVM system [24]. Schaeffer [39] has also proved the convergence of a finite difference scheme for the one-dimensional Vlasov–Poisson–Fokker–Planck system, and Filbet [20] has shown the convergence of a finite volume scheme for the one-dimensional Vlasov–Poisson system.

Although a number of papers present satisfactory numerical results using semi-Lagrangian methods [43, 13, 40, 1, 2, 18, 9], few rigorous mathematical results on convergence analysis of semi-Lagrangian methods have been stated. Although interesting a priori estimates have been pointed out (cf. [4, 5, 18]), a lot of work still remains to give complete and rigorous results in more general situations. The more difficult step in the convergence analysis of semi-Lagrangian methods is obtaining a stability result for the interpolation operators. If stability results in the L^∞ norm seem inaccessible for high order interpolation operators because of the Runge phenomena (artificial os-

cillations, whose amplitude increases with the degree of the polynomial in the case of Lagrange interpolation, appear at the edges of finite elements), a more appropriate mathematical framework is L^2 stability. If Fourier analysis tools as Fourier series are useful for proving L^2 stability in the case of grids, convenient mathematical tools are lacking for unstructured meshes such as triangulation and have to be developed in the future. Nevertheless new results on the convergence analysis of classes of high order schemes can be found in [7, 8, 6].

This paper is organized as follows. In the first part we present the continuous problem. In the second part we expose the discrete problem and the numerical scheme to solve it. Then we study the convergence of our numerical scheme. In the last section we give refined convergence results.

2. The continuous problem. We consider a noncollisional plasma of charged particles (electrons and ions) in one dimension. We take into account the electrostatic forces and neglect the magnetic effects. Due to the great inertia of the ions compared to the electrons, we assume that the ions form a neutralizing uniform background.

Denoting by $f(t,x,v) \geq 0$ the distribution function of electrons in phase space (with mass normalized to one, the charge to plus one), and by $E(t,x)$ the self-consistent electric field, the adimensional Vlasov–Poisson system reads

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} + E(t,x)\frac{\partial f}{\partial v} = 0, \tag{2.1}$$

$$\frac{dE}{dx}(t,x) = \rho(t,x) = \int_{-\infty}^{+\infty} f(t,x,v)dv - 1. \tag{2.2}$$

We consider a periodic plasma of period L. Hence in (2.1) and (2.2) we have $x \in [0,L]$, $v \in \mathbb{R}$, $t \geq 0$, and the functions f and E satisfy the periodic boundary conditions

$$f(t,0,v) = f(t,L,v), \quad v \in \mathbb{R}, \quad \mathrm{t} \geq 0, \tag{2.3}$$

and

$$E(t,0) = E(t,L) \Longleftrightarrow \frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f(t,x,v)dvdx = 1, \quad t \geq 0, \tag{2.4}$$

which means that the plasma is globally neutral. In order to have a well-posed problem, we add to (2.1)–(2.4) a zero-mean electrostatic condition,

$$\int_0^L E(t,x)dx = 0, \quad t \geq 0, \tag{2.5}$$

and an initial condition,

$$f(0,x,v) = f_0(x,v), \quad x \in [0,L],\ v \in \mathbb{R}. \tag{2.6}$$

If we introduce the electrostatic potential $\phi = \phi(t,x)$ such that

$$E(t,x) = -\frac{\partial \phi}{\partial x}(t,x),$$

and if we denote by $G = G(x,y)$ the Green function associated with our problem—that is to say, for $y \in]0,L[$, $G(.,y)$ is the solution of

$$-\frac{\partial^2 G}{\partial x^2}(x,y) = \delta(x-y), \quad x \in [0,L], \quad G(0,y) = G(L,y),$$

where δ is the Dirac distribution—then $G(x,y)$ and $K(x,y) = -\partial_x G(x,y)$ are given by

$$G(x,y) = \begin{cases} x\left(1-\dfrac{y}{L}\right), & 0 \le x \le y, \\ y\left(1-\dfrac{x}{L}\right), & y \le x \le L, \end{cases} \qquad K(x,y) = \begin{cases} \left(\dfrac{y}{L}-1\right), & 0 \le x < y, \\ \dfrac{y}{L}, & y < x \le L. \end{cases}$$

Therefore ϕ is given by

$$\phi(t,x) = \int_0^L G(x,y)\left(\int_{-\infty}^{+\infty} f(t,y,v)dv - 1\right)dy,$$

and E can be rewritten as

$$E(t,x) = \int_0^L K(x,y)\left(\int_{-\infty}^{+\infty} f(t,y,v)dv - 1\right)dy. \tag{2.7}$$

In addition, assuming that the electric field E is smooth enough, we can solve (2.1), (2.3), and (2.6) in the classical sense as follows. For the existence, uniqueness, and regularity of the solutions of the following differential system we refer the reader to [12] and [36].

We consider the first order differential system

$$\begin{aligned} &\frac{dX}{dt}(t;s,x,v) = V(t;s,x,v), \\ &\frac{dV}{dt}(t;s,x,v) = E(t,X(t;s,x,v)) \end{aligned} \tag{2.8}$$

and denote by $t \to (X(t;s,x,v), V(t;s,x,v))$ the characteristic curves, which are the solution of (2.8) with the initial conditions

$$X(s;s,x,v) = x, \qquad V(s;s,x,v) = v. \tag{2.9}$$

Then the solution of problem (2.1), (2.6) is given by

$$f(t,x,v) = f_0(X(0;t,x,v), V(0;t,x,v)), \quad x,v \in \mathbb{R},\ \ t \ge 0. \tag{2.10}$$

We note that the periodicity in x of $f_0(x,v)$ and $E(t,x)$ implies the periodicity in x of $f(t,x,v)$. Moreover, as

$$\left|\frac{\partial(X,V)}{\partial(x,v)}\right| = 1,$$

we get

$$\frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f(t,x,v)dvdx = \frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f_0(x,v)dvdx = 1.$$

Therefore, according to the previous considerations, an equivalent form of the Vlasov–Poisson periodic problem is to find a pair (f,E), smooth enough, periodic with respect to x, with period L, and solving (2.7), (2.8), (2.9), and (2.10).

2.1. Definitions and notation. We now introduce basic notation. If $\mathbb{N}$ denotes the set of nonnegative integers, a multi-index α is an n-tuple of nonnegative integers $\alpha := (\alpha_1, \ldots, \alpha_n)$, $\alpha_i \in \mathbb{N}$, $i = 1, \ldots, n$. We have the following definitions:

$$|\alpha| = \alpha_1 + \cdots + \alpha_n,$$

$$D^\alpha = \partial_{x_1}^{\alpha_1} \ldots \partial_{x_n}^{\alpha_n}.$$

Let Ω be a domain in $\mathbb{R}^n$. For any nonnegative integer m let $\mathscr{C}^m(\Omega)$ be the vector space consisting of all functions ϕ that, together with all their partial derivatives $D^\alpha\phi$ of orders $|\alpha| \leq m$, are continuous on Ω.

We define the vector space $\mathscr{C}_b^m(\Omega)$ of all functions $\phi \in \mathscr{C}^m(\Omega)$ for which $D^\alpha\phi$ is bounded and uniformly continuous on Ω for $0 \leq |\alpha| \leq m$. $\mathscr{C}_b^m(\Omega)$ is a Banach space with the norm given by

$$||\phi||_{\mathscr{C}_b^m(\Omega)} = \max_{0 \leq |\alpha| \leq m} \sup_{z \in \Omega} |D^\alpha\phi(z)|.$$

We define $\mathscr{C}_c^m(\Omega)$ as the subspace of $\mathscr{C}_b^m(\Omega)$ consisting of those functions ϕ for which, for $0 \leq |\alpha| \leq m$, $D^\alpha\phi$ has compact support in Ω.

If $0 < \lambda \leq 1$, we define $\mathscr{C}^{m,\lambda}(\Omega)$ to be the subspace of $\mathscr{C}_b^m(\Omega)$ consisting of those functions ϕ for which, for $0 \leq |\alpha| \leq m$, $D^\alpha\phi$ satisfies in Ω a Hölder condition of exponent λ; that is, there exists a constant K such that

$$|D^\alpha\phi(x) - D^\alpha\phi(y)| \leq K|x - y|^\lambda, \quad x, y \in \Omega.$$

$\mathscr{C}^{m,\lambda}(\Omega)$ is a Banach space with norm given by

$$||\phi||_{\mathscr{C}^{m,\lambda}(\Omega)} = ||\phi||_{\mathscr{C}_b^m(\Omega)} + \max_{0 \leq |\alpha| \leq m} \sup_{\substack{x,\, y \,\in\, \Omega \\ x \neq y}} \frac{|D^\alpha\phi(x) - D^\alpha\phi(y)|}{|x - y|^\lambda}.$$

For all $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ we let

$$\mathrm{Lip}(\phi) = \sup_{\substack{x,\, y \,\in\, \Omega \\ x \neq y}} \frac{|\phi(x) - \phi(y)|}{|x - y|}.$$

Furthermore,

$$\mathrm{Lip}(\Omega) = \{\phi : \mathbb{R}^n \longrightarrow \mathbb{R} \mid \ \mathrm{Lip}(\phi) < \infty\}$$

is a Banach space with the norm given by

$$||\phi||_{\mathrm{Lip}(\Omega)} = ||\phi||_{\mathscr{C}^{0,1}(\Omega)}.$$

We define $\mathscr{C}_{b,per_{x_i}}^m(\Omega_{x_i} \times \Omega_{n-1})$ as the subspace of $\mathscr{C}_b^m(\Omega)$ consisting of those functions ϕ which are periodic with respect to the variable x_i and bounded with respect to other variables. We also define $\mathscr{C}_{c,per_{x_i}}^m(\Omega_{x_i} \times \Omega_{n-1})$ as the subspace of $\mathscr{C}_c^m(\Omega)$ consisting of those functions ϕ which are periodic with respect to the variable x_i and compactly supported with respect to the other variables.

We denote by $L^p(\Omega)$, $1 \le p \le \infty$, the space of all equivalence classes of real-valued Lebesgue-measurable functions. $L^p(\Omega)$ is a Banach space with the norm given by

$$||\phi||_{L^p(\Omega)} = \left\{ \int_\Omega |\phi|^p d\Omega \right\}^{1/p}, \quad 1 \le p < \infty,$$
$$||\phi||_{L^\infty(\Omega)} = \text{ess } \sup_{z\in\Omega} |\phi(z)|.$$

We define $W^{m,p}(\Omega)$ to be the Sobolev space consisting of all functions ϕ which, together with all their partial derivatives $D^\alpha\phi$ taken in the sense of distribution of orders $|\alpha| \le m$, belong to the $L^p(\Omega)$ space. If we define the seminorm as

$$|\phi|_{W^{k,p}(\Omega)} = \left\{ \sum_{|\alpha|=k} |D^\alpha\phi|^p_{L^p(\Omega)} \right\}^{1/p}, \quad 1 \le p < \infty,$$
$$|\phi|_{W^{k,\infty}(\Omega)} = \max_{|\alpha|=m} \text{ ess } \sup_{z\in\Omega} |D^\alpha\phi(z)|,$$

then we provide $W^{m,p}(\Omega)$ with the norm

$$||\phi||_{W^{m,p}(\Omega)} = \left\{ \sum_{k=0}^{m} |\phi|^p_{W^{k,p}(\Omega)} \right\}^{1/p}, \quad 1 \le p < \infty,$$
$$||\phi||_{W^{m,\infty}(\Omega)} = \max_{0\le k\le m} |\phi|_{W^{k,\infty}(\Omega)}.$$

Let X be a Banach space with norm $||\cdot||_X$. We denote by $\mathscr{C}^m(0,T;X)$, $0 < T < +\infty$, the space of m-times continuously differentiable functions from $(0,T)$ into X, and by $L^p(0,T;X)$ the space of all strongly measurable functions $\phi : t \longrightarrow \phi(t)$ from $(0,T)$ into X. The following norms are defined:

$$||\phi||_{\mathscr{C}(0,T;X)} = \sup_{t\in[0,T]} ||\phi(t)||_X,$$

$$||\phi||_{\mathscr{C}^m(0,T;X)} = \sum_{k=0}^{m} \left\| \frac{d^k\phi}{dt^k} \right\|_{\mathscr{C}(0,T;X)},$$

$$||\phi||_{L^p(0,T;X)} = \left\{ \int_0^T ||\phi(t)||^p_X dt \right\}^{1/p}, \quad 1 \le p < \infty,$$

$$||\phi||_{L^\infty(0,T;X)} = \text{ess } \sup_{0<t<T} ||\phi(t)||_X.$$

Finally, we introduce the space $\ell^\infty(0,T;X)$ defined by

$$\ell^\infty(0,T;X) := \left\{ f : \{t^0,\dots,t^M\} \to X \mid \ ||f||_{\ell^\infty(0,T;X)} = \max_{1\le n\le M} ||f(t^n)||_X < \infty \right\},$$

where X denotes a functional space (in our context X should be L^p, $p \in [1,\infty]$), and the space $L^{1,\infty}$ defined by

$$L^{1,\infty} = \left\{ f \in L^1 \cap L^\infty \mid \ \|f\|_{L^{1,\infty}} < \infty \right\},$$

where

$$\|f\|_{L^{1,\infty}} = \|f\|_{L^1} + \|f\|_{L^\infty}.$$

2.2. Existence, uniqueness, and regularity of the solution of the continuous problem. In this section we recall a theorem of existence of a classical solution for the Vlasov–Poisson system. The following theorem gives the existence, uniqueness, and regularity of the classical solutions, global in time, of the Vlasov–Poisson periodic system in one dimension.

THEOREM 2.1. *Assuming $f_0 \in \mathscr{C}^1_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)$, positive, periodic with respect to the variable x with period L, and $Q(0) \leq R$ with $R > 0$ and $Q(t)$ defined as*

$$Q(t) = 1 + \sup\{|v| : \exists x \in [0, L], \quad \tau \in [0, t] \mid f(\tau, x, v) \neq 0\}$$

and

$$\frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f_0(x, v) dv dx = 1,$$

then the periodic Vlasov–Poisson system has a unique classical solution (f, E), periodic in x, with period L, for all time t in $[0, T]$, such that

$$f \in \mathscr{C}^1_b\left(0, T; \mathscr{C}^1_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)\right),$$
$$E \in \mathscr{C}^1_b\left(0, T; \mathscr{C}^1_{b,per_x}(\mathbb{R})\right),$$

and there exists a constant $C = C(R, f_0)$ dependent on R and f_0 such that

$$Q(T) \leq CT.$$

Moreover, if we assume $f_0 \in \mathscr{C}^m_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)$, then $(f, E) \in \mathscr{C}^m_b(0, T; \mathscr{C}^m_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)) \times \mathscr{C}^m_b(0, T; \mathscr{C}^m_{b,per_x}(\mathbb{R}))$ for all finite time T.

Proof. We do not write out the proof because it is a straightforward adaptation of the proof done by Schaeffer in [38]. We refer the reader to the articles [34, 28, 29, 23, 15, 30, 31]. ∎

2.3. Regularity assumptions for the continuous problem. For our purpose, we first suppose that $f_0(x, v)$ satisfies the following regularity assumptions:

$$f_0 \in \mathscr{C}^2_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v).$$

Then, as is proven in Glassey [23], if f_0 is smooth and compactly supported, the solution of the Vlasov–Poisson system remains smooth and compactly supported for all time. Theorem 2.1 gives the existence and uniqueness of the solution (f, E) such that

$$f \in \mathscr{C}^2_b\left(0, T; \mathscr{C}^2_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)\right), \tag{2.11}$$

$$E \in \mathscr{C}^2_b\left(0, T; \mathscr{C}^2_{b,per_x}(\mathbb{R})\right). \tag{2.12}$$

Further, we prove that we still have convergence under weaker regularity assumptions.

3. The discrete problem.

3.1. Space of approximation and the interpolation operator. Let $Q = [0, L] \times \mathbb{R}$, $\Omega = [0, L] \times [-R, R]$ with $R > 0$, and $\mathcal{T}_h$ be a triangulation of Q.

Before going further we impose some regularity assumptions on the triangulation $\mathcal{T}_h$ as follows:

(H1) The triangulation $\mathcal{T}_h$ is regular; that is to say, there exists a constant σ such that

$$\frac{h_T}{\rho_T} \leq \sigma \quad \forall T \in \mathcal{T}_h,$$

and the quantity $h = \max_{\{T \in \mathcal{T}_h\}} h_T$ approaches zero, where h_T and ρ_T denote, respectively, the exterior and the interior diameter of a finite element T.

(H2) All the finite elements (T, P_T, Σ_T), $T \in \mathcal{T}_h$, are affine equivalent to a single reference finite element $(\widehat{T}, \widehat{P}, \widehat{\Sigma})$ (see [14]).

Let P_m be the space Lagrange polynomial of degree less than or equal to m, and let X_h be the space defined by

$$X_h = \{g \in W^{1,\infty} \cap W^{1,p}(Q),\ g_{|_T} \in P_m\ \forall T \in \mathcal{T}_h\}.$$

Let π_h be a continuous linear interpolation operator from $W^{m+1,\infty} \cap W^{m+1,p}(Q)$, $1 \leq p < \infty$, onto X_h. The interpolation error estimations in Sobolev spaces (see [14]) give, with $k \in \{0, 1\}$ and $q \in \{p, \infty\}$,

$$||f - \pi_h f||_{W^{k,q}(Q)} \leq C h^{m+1-k} |f|_{W^{m+1,q}} \quad \forall f \in W^{m+1,\infty} \cap W^{m+1,q}(Q). \tag{3.1}$$

The space X_h is characterized by its basis functions, denoted by $\{\psi_k\}$.

3.2. Transport operators. Now we introduce some transport operators. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the operators defined as

$$\mathcal{T}_1 g(t, x, v) = g\left(t, x - v\frac{\Delta t}{2}, v\right),$$

$$\mathcal{T}_2 g(t, x, v) = g(t, x, v - \Delta t \widetilde{E}(t, x)),$$

where $\widetilde{E}(t, x)$ is the solution of the following problem:

$$\begin{cases} \dfrac{d\widetilde{E}}{dx}(t, x) = \displaystyle\int_v \mathcal{T}_1 g(t, x, v) dv - 1, \\ \displaystyle\int_0^L \widetilde{E}(t, x) dx = 0. \end{cases} \tag{3.2}$$

Let $\widetilde{\mathcal{T}}_1$ be the transport operator defined as

$$\widetilde{\mathcal{T}}_1 g(t, x, v) = \pi_h g\left(t, x - v\frac{\Delta t}{2}, v\right),$$

where

$$\pi_h g(t, x, v) = \sum_k g(t, x_k, v_k) \psi_k(x, v),$$

and let $\widetilde{\mathcal{T}}_2$ be defined as

$$\widetilde{\mathcal{T}}_2 g(t, x, v) = \pi_h g(t, x, v - \Delta t \widetilde{E}(t, x)).$$

Finally we introduce

$$\widetilde{\mathcal{T}}_2^{\star} g(t,x,v) = \pi_h g(t,x,v - \Delta t E_h(t,x)),$$

where $E_h(t,x)$ is the solution of the following problem:

$$\begin{cases} \dfrac{dE_h}{dx}(t,x) = \displaystyle\int_v \widetilde{\mathcal{T}}_1 g(t,x,v)dv - 1, \\ \displaystyle\int_0^L E_h(t,x)dx = 0. \end{cases} \tag{3.3}$$

Notice that (2.7) implies that $\widetilde{E}(t,x)$ and $E_h(t,x)$ are respectively given by

$$\widetilde{E}(t,x) = \int_0^L K(x,y) \left(\int_{-\infty}^{+\infty} \mathcal{T}_1 g(t,y,v)dv - 1 \right) dy$$

and

$$E_h(t,x) = \int_0^L K(x,y) \left(\int_{-\infty}^{+\infty} \widetilde{\mathcal{T}}_1 g(t,y,v)dv - 1 \right) dy.$$

4. The numerical scheme. We suppose that we know $f_h(t^n)$ defined on $\mathcal{T}_h$. Therefore the numerical scheme which allows us to go from time t^n to t^{n+1} and compute $f_h(t^{n+1})$ can be described in four steps:

- (A1) We evaluate the distribution at time t^n at the foot of the field-free characteristics starting at (x,v) at time $t^{n+1/2}$ using a Lagrange interpolation operator. This action is described by the transport operator $\widetilde{\mathcal{T}}_1$.
- (A2) The output from (A1) is integrated with respect to velocity to provide an approximation for the density at time $t^{n+1/2}$, which is then substituted into the Poisson equation (3.3) to compute the approximation of the electric field at time $t^{n+1/2}$.
- (A3) The result obtained from (A1) is evaluated at the foot of the velocity characteristic starting at (x,v) at time t^{n+1} with the acceleration field found in (A2) using a Lagrange interpolation operator. This action is described by the transport operator $\widetilde{\mathcal{T}}_2^{\star}$
- (A4) Between time $t^{n+1/2}$ and t^{n+1}, we apply step (A1) to the output from (A3). This action is described by the transport operator $\widetilde{\mathcal{T}}_1$. Then we obtain $f_h(t^{n+1})$, which is the new initial data for the algorithm (A1)–(A4).

Using transport operators defined above in section 3.2, the numerical scheme can be written as

$$f_h(t^{n+1},x,v) = \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^{\star} \circ \widetilde{\mathcal{T}}_1 f_h(t^n,x,v),$$

where $f_h(0,x,v) = \pi_h f_0(x,v)$ is a discretization of f_0 for the initial data,

$$f_h(t^n, x+L, v) = f_h(t^n,x,v) \quad \forall |v| \leq Q(T)$$

is the boundary condition in x, and

$$f_h(t^n,x,v) = 0 \quad \forall |v| > Q(T),\ \forall x \in [0,L]$$

is the boundary condition in v.

5. Convergence analysis.

5.1. Main theorem. We next give the convergence theorem.

THEOREM 5.1. *Assuming $f_0 \in \mathscr{C}^2_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)$, positive, periodic with respect to the variable x with period L, then the numerical solution of the Vlasov–Poisson system (f_h, E_h), computed by the numerical scheme exposed in section 4, converges toward the solution (f, E) of the periodic Vlasov–Poisson system, and there exists a constant $C = C(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))})$ independent of Δt, h such that*

$$||f - f_h||_{\ell^\infty(0,T;L^\infty(Q))} \leq C\left(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\left(\Delta t^2 + h^2 + \frac{h^2}{\Delta t}\right)$$

and

$$||E - E_h||_{\ell^\infty(0,T;L^\infty([0,L]))} \leq C\left(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\left(\Delta t^2 + h^2 + \frac{h^2}{\Delta t}\right).$$

Remark 5.2. In Theorem 5.1 we have a lot of choices for the time step. We note that the convergence rate is slightly better than first order: If we make the choice $\Delta t = h^{2/3}$, then the error estimate involves $h^{4/3}$ rather than h to the first power. Therefore we see that the main reason for using semi-Lagrangian schemes in lieu of particle schemes comes from the nice flexibility of the error estimates stated in Theorem 5.1, because they allow us to choose larger time steps and get convergence rates higher than one.

5.2. Idea of the proof. We want to evaluate the global error at time t^{n+1}:

$$e^{n+1} = ||f(t^{n+1}, x, v) - f_h(t^{n+1}, x, v)||_{L^\infty(Q)}.$$

Therefore we decompose $f(t^{n+1}, x, v) - f_h(t^{n+1}, x, v)$ as

$$\begin{aligned} f(t^{n+1}, x, v) - f_h(t^{n+1}, x, v) = {} & f(t^{n+1}, x, v) - \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n, x, v) \\ & + \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n, x, v) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n, x, v) \\ & + \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n, x, v) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 f(t^n, x, v) \\ & + \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 (f(t^n, x, v) - f_h(t^n, x, v)). \end{aligned}$$

In order to estimate e^{n+1} we will estimate the four terms on the right-hand side of this equation. These estimations are described in the following section.

5.3. A priori estimates. We begin with the following lemma, which gives an estimate of the time discretization error.

LEMMA 5.3. *Assume that $f \in \mathscr{C}^2_b(0,T;\mathscr{C}^2_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$; then there exists a constant C such that*

$$\left\| f(t^{n+1}) - \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) \right\|_{L^\infty(Q)} \leq C\left(\|f\|_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\Delta t^3.$$

Proof. As f is constant along the characteristic curves, we have

$$\begin{aligned} f(t^{n+1}, x, v) &= f(t^{n+1}, X(t^{n+1}; t^{n+1}, x, v), V(t^{n+1}; t^{n+1}, x, v)) \\ &= f(t^n, X(t^n; t^{n+1}, x, v), V(t^n; t^{n+1}, x, v)) \\ &= f(t^n, X(t^n), V(t^n)), \end{aligned}$$

where $X(t^n) = X(t^n; t^{n+1}, x, v)$ and $V(t^n) = V(t^n; t^{n+1}, x, v)$. On the other hand, we have

$$\begin{aligned}
\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) &= \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n, x, v) \\
&= \mathcal{T}_1 \circ \mathcal{T}_2 f\Big(t^n, x - v\tfrac{\Delta t}{2}, v\Big) \\
&= \mathcal{T}_1 f\left(t^n, x - v\tfrac{\Delta t}{2} + \tfrac{\Delta t^2}{2}\widetilde{E}(t^{n+1/2}, x), v - \Delta t \widetilde{E}(t^{n+1/2}, x)\right) \\
&= f\Big(t^n, x - v\Delta t + \tfrac{\Delta t^2}{2}\widetilde{E}\Big(t^{n+1/2}, x - v\tfrac{\Delta t}{2}\Big), v - \Delta t \widetilde{E}\Big(t^{n+1/2}, x - v\tfrac{\Delta t}{2}\Big)\Big) \\
&= f(t^n, \widetilde{X}(t^n; t^{n+1}, x, v), \widetilde{V}(t^n; t^{n+1}, x, v)) \\
&= f(t^n, \widetilde{X}(t^n), \widetilde{V}(t^n)),
\end{aligned}$$

where

$$\widetilde{X}(t^n) = x - v\Delta t + \frac{\Delta t^2}{2}\widetilde{E}\left(t^{n+1/2}, x - v\frac{\Delta t}{2}\right)$$

and

$$\widetilde{V}(t^n) = v - \Delta t \widetilde{E}\left(t^{n+1/2}, x - v\frac{\Delta t}{2}\right).$$

In order to justify the following Taylor expansion, we remember that assumption (2.12) gives $E \in \mathscr{C}^2_b(0, T; \mathscr{C}^2_{b,per_x}(\mathbb{R}))$. We notice that $\widetilde{E}$ has the same regularity in space as E, as the source terms in Poisson equations (3.2) and (2.2) also have the same regularity.

Hence a Taylor expansion gives

$$\begin{aligned}
X(t^{n+1/2}) - \big(x - v\tfrac{\Delta t}{2}\big) &= X(t^{n+1/2}) - \big(X(t^{n+1}) - V(t^{n+1})\tfrac{\Delta t}{2}\big) \\
&= X(t^{n+1/2}) - \big(X(t^{n+1}) - \tfrac{\Delta t}{2}\dot{X}(t^{n+1})\big) \\
&= O(\Delta t^2).
\end{aligned} \tag{5.1}$$

As $f \in \mathscr{C}^2_b(0, T; \mathscr{C}^2_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$, we have

$$\begin{aligned}
\frac{f(t^{n+1/2}, x, v) - f\big(t^n, x - v\frac{\Delta t}{2}, v\big)}{\frac{\Delta t}{2}} &= \partial_t f(t^{n+1/2}, x, v) + v\partial_x f(t^{n+1/2}, x, v) + O(\Delta t) \\
&= -E(t^{n+1/2}, x)\partial_v f(t^{n+1/2}, x, v) + O(\Delta t).
\end{aligned} \tag{5.2}$$

Then, using (2.7) and (5.2), we get

$$\begin{aligned}
&E(t^{n+1/2}, x) - \widetilde{E}(t^{n+1/2}, x) \\
&\quad = \int_0^L K(x, y)\left(\int_{-\infty}^{+\infty}\Big[f(t^{n+1/2}, y, v) - f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big)\Big]dv\right)dy \\
&\quad \leq C\left(\|f\|_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\Delta t^2.
\end{aligned} \tag{5.3}$$

Using (5.1) and (5.3), we obtain

$$\begin{aligned}
V(t^n) - \widetilde{V}(t^n) &= V(t^n) - \Big(V(t^{n+1}) - \Delta t \widetilde{E}\Big(t^{n+1/2}, X(t^{n+1}) - V(t^{n+1})\tfrac{\Delta t}{2}\Big)\Big) \\
&= V(t^n) - \Big(V(t^{n+1}) - \Delta t \widetilde{E}\left(t^{n+1/2}, X(t^{n+1/2}) + O(\Delta t^2)\right)\Big) \\
&= V(t^n) - \left(V(t^{n+1}) - \Delta t E\left(t^{n+1/2}, X(t^{n+1/2}) + O(\Delta t^2)\right)\right) \\
&\quad + \Delta t \Big(\widetilde{E}\ \left(t^{n+1/2}, X(t^{n+1/2}) + O(\Delta t^2)\right) \\
&\qquad - E\left(t^{n+1/2}, X(t^{n+1/2}) + O(\Delta t^2)\right)\Big) \\
&= V(t^n) - \left(V(t^{n+1}) - \Delta t E(t^{n+1/2}, X(t^{n+1/2}))\right) + O(\Delta t^3) \\
&= V(t^n) - V(t^{n+1}) + \Delta t \dot{V}(t^{n+1/2}) + O(\Delta t^3) \\
&\le C\left(\|f\|_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\Delta t^3
\end{aligned}$$

and

$$\begin{aligned}
X(t^n) - \widetilde{X}(t^n) &= X(t^n) - \Big(X(t^{n+1}) - \Delta t V(t^{n+1}) \\
&\qquad + \tfrac{\Delta t^2}{2}\widetilde{E}\Big(t^{n+1/2}, X(t^{n+1}) - V(t^{n+1})\tfrac{\Delta t}{2}\Big)\Big) \\
&= X(t^n) - \Big(X(t^{n+1}) - \Delta t V(t^{n+1}) \\
&\qquad + \tfrac{\Delta t^2}{2}\widetilde{E}\big(t^{n+1/2}, X(t^{n+1/2}) + O(\Delta t^2)\big)\Big) \\
&= X(t^n) - \Big(X(t^{n+1}) - \Delta t V(t^{n+1}) \\
&\qquad + \tfrac{\Delta t^2}{2}E\big(t^{n+1/2}, X(t^{n+1/2}) + O(\Delta t^2)\big)\Big) \\
&\quad - \tfrac{\Delta t^2}{2}\Big(\widetilde{E}(t^{n+1/2}, X(t^{n+1/2}) + O(\Delta t^2)) \\
&\qquad - E(t^{n+1/2}, X(t^{n+1/2}) + O(\Delta t^2))\Big) \\
&= X(t^n) - \Big(X(t^{n+1}) - \Delta t V(t^{n+1}) \\
&\qquad + \tfrac{\Delta t^2}{2}E(t^{n+1/2}, X(t^{n+1/2}))\Big) + O(\Delta t^4) \\
&= X(t^n) - \Big(X(t^{n+1}) - \Delta t \dot{X}(t^{n+1}) + \tfrac{\Delta t^2}{2}\ddot{X}(t^{n+1/2})\Big) + O(\Delta t^4) \\
&= X(t^n) - \Big(X(t^{n+1}) - \Delta t \dot{X}(t^{n+1}) + \tfrac{\Delta t^2}{2}\ddot{X}(t^{n+1})\Big) + O(\Delta t^3) \\
&\le C\left(\|f\|_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\Delta t^3.
\end{aligned}$$

Finally, we deduce that

$$\begin{aligned}
\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) &= f(t^n, X(t^n) + O(\Delta t^3), V(t^n) + O(\Delta t^3)) \\
&= f(t^n, X(t^n), V(t^n)) + \nabla f(t^n, X(t^n), V(t^n)) \cdot O(\Delta t^3) \\
&= f(t^{n+1}, X(t^{n+1}), V(t^{n+1})) + \nabla f(t^n, X(t^n), V(t^n)) \cdot O(\Delta t^3) \\
&= f(t^{n+1}, x, v) + \nabla f(t^n, X(t^n), V(t^n)) \cdot O(\Delta t^3)
\end{aligned}$$

and

$$||f(t^{n+1}) - \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n)||_{L^p(Q)} \leq C \left(\|f\|_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))} \right) ||\nabla f||_{L^\infty([0,T]\times Q)} \Delta t^3. \qquad \square$$

We continue with the following result.

PROPOSITION 5.4. *Assume that* $f \in L^\infty(0,T;\mathscr{C}^{m+1}_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$, $m \geq 0$, *and* π_h *is a continuous linear interpolation operator from* $W^{m+1,\infty}(Q)$ *onto* X_h; *then there exists a constant* C *such that for* $i = 1, 2$, $1 \leq p \leq \infty$,

$$||\mathcal{T}_i f||_{L^\infty(0,T;W^{m+1,p}(Q))} \leq C||f||_{L^\infty(0,T;W^{m+1,p}(Q))}, \tag{5.4}$$

$$||\widetilde{\mathcal{T}}_i f||_{L^\infty(0,T;L^p(Q))} \leq C||f||_{L^\infty(0,T;W^{m+1,p}(Q))}, \tag{5.5}$$

and

$$\|(\mathcal{T}_i - \widetilde{\mathcal{T}}_i)f\|_{L^\infty(0,T;L^p(Q))} \leq Ch^{m+1}||f||_{L^\infty(0,T;W^{m+1,p}(Q))}. \tag{5.6}$$

Proof. It is obvious that

$$\left\| f\left(t, x - v\tfrac{\Delta t}{2}, v\right) \right\|_{L^\infty(0,T;L^p(Q))} = \|f\|_{L^\infty(0,T;L^p(Q))} \tag{5.7}$$

and

$$\|f(t,x,v - \widetilde{E}(t,x)\Delta t)\|_{L^\infty(0,T;L^p(Q))} = \|f\|_{L^\infty(0,T;L^p(Q))}. \tag{5.8}$$

On one side the gradient of $f(t, x - v\Delta t/2, v)$ gives

$$\left\| \partial_x \left(f\left(t, x - v\tfrac{\Delta t}{2}, v\right) \right) \right\|_{L^\infty(0,T;L^p(Q))} = \|\partial_x f\|_{L^\infty(0,T;L^p(Q))}$$

and

$$\left\| \partial_v \left(f(t, x - v\tfrac{\Delta t}{2}, v) \right) \right\|_{L^\infty(0,T;L^p(Q))} \leq \tfrac{\Delta t}{2} \|\partial_x f\|_{L^\infty(0,T;L^p(Q))} + \|\partial_v f\|_{L^\infty(0,T;L^p(Q))}.$$

Hence

$$\left\| f\left(t, x - v\tfrac{\Delta t}{2}, v\right) \right\|_{L^\infty(0,T;W^{1,p}(Q))} \leq C \|f\|_{L^\infty(0,T;W^{1,p}(Q))}.$$

In the same way we get

$$\left\| f\left(t, x - v\tfrac{\Delta t}{2}, v\right) \right\|_{L^\infty(0,T;W^{m+1,p}(Q))} \leq C \|f\|_{L^\infty(0,T;W^{m+1,p}(Q))}.$$

On the other side the gradient of $f(t,x,v - \widetilde{E}(t,x)\Delta t)$ gives

$$\begin{aligned} &\|\partial_x(f(t,x,v - \widetilde{E}(t,x)\Delta t))\|_{L^\infty(0,T;L^p(Q))} \\ &\quad \leq \|\partial_x f\|_{L^\infty(0,T;L^p(Q))} + \Delta t||\partial_x \widetilde{E}||_{L^\infty([0,T]\times[0,L])} \|\partial_v f\|_{L^\infty(0,T;L^p(Q))} \end{aligned}$$

and

$$\|\partial_v(f(t,x,v - \widetilde{E}(t,x)\Delta t))\|_{L^\infty(0,T;L^p(Q))} \leq \|\partial_v f\|_{L^\infty(0,T;L^p(Q))}.$$

Hence

$$\begin{aligned}\|f(t,x,v-\widetilde{E}(t,x)\Delta t)\|_{L^\infty(0,T;W^{1,p}(Q))} &\le (1+C\Delta t)\,\|f\|_{L^\infty(0,T;W^{1,p}(Q))}\\ &\le C\,\|f\|_{L^\infty(0,T;W^{1,p}(Q))}.\end{aligned}$$

In the same way, as $\widetilde{E}\in L^\infty(0,T;\mathscr{C}^{m+1}_{b,per_x}(\mathbb{R}))$, we get

$$\|f(t,x,v-\widetilde{E}(t,x)\Delta t)\|_{L^\infty(0,T;W^{m+1,p}(Q))} \le C\,\|f\|_{L^\infty(0,T;W^{m+1,p}(Q))},$$

which completes the proof of (5.4).

π_h is an interpolation operator which is characterized by the basis functions $\{\psi_k\}$. Then $\pi_h f$ can be written as follows:

$$\pi_h f(t,x,v)=\sum_k f(t,x_k,v_k)\psi_k(x,v)=\sum_k f_k(t)\psi_k(x,v).$$

As any $\psi_k\in L^\infty(Q)$ and has compact support, there exists a constant M such that

$$\begin{aligned}\left\|\sum_k|\psi_k(x,v)|\right\|_{L^\infty(Q)} &\le \sup_{T\in\mathcal{T}_h}\left\|\sum_k|\psi_k(x,v)|\right\|_{L^\infty(T)}\\ &\le \mathrm{card}(\Sigma_T)\sup_{T\in\mathcal{T}_h}\max_{(x,v)\in T}|\psi_k(x,v)|\\ &\le M,\end{aligned}$$

where Σ_T is the set of degrees of freedom on the triangle T.

- L^∞ case:

$$\|\pi_h f\|_{L^\infty(Q)}\le\|f\|_{L^\infty(Q)}\sum_k|\psi_k(x,v)|\le M\,\|f\|_{L^\infty(Q)}.$$

- L^1 case:

$$\int_Q|\pi_h f(t)|dvdx\le\sum_k|f_k(t)|\int_Q|\psi_k|dxdv\le M\sum_k|f_k(t)|\mathrm{meas}\,(\mathcal{S}_k),$$

where $\mathcal{S}_k$ is the support of ψ_k. Let $\mathcal{A}_k$ be the geometrical area associated with the node $N_k=(x_k,v_k)$, obtained by joining the barycenter of the triangles that have the vertex N_k in common to the middle of the edges of the triangles; then there exists a constant $K>0$ independent of h such that $(1/K)\mathrm{meas}(\mathcal{S}_k)\le\mathrm{meas}(\mathcal{A}_k)<\mathrm{meas}(\mathcal{S}_k)$. Then we obtain

$$\|\pi_h f(t)\|_{L^1(Q)}\le CMK\sum_k|f_k(t)|\mathrm{meas}\,(\mathcal{A}_k)\le C\,\|f(t)\|_{L^1(Q)}$$

and

$$\|\pi_h f\|_{L^\infty([0,T],L^1(Q))}\le C\,\|f\|_{L^\infty([0,T],L^1(Q))}.$$

- L^p case:

$$\int_Q|\pi_h f(t)|^p dvdx\le\int_Q\left(\sum_k|f_k(t)|\,|\psi_k|\right)^p dvdx.$$

Thanks to the Hölder inequality, we get

$$\int_Q |\pi_h f(t)|^p \leq \int_Q \left(\sum_k |f_k(t)|^p |\psi_k|\right) \left(\sum_k |\psi_k|\right)^{p/p^*} dvdx,$$

with $p^* = p/(p-1)$. Then we get

$$\begin{aligned}
\|\pi_h f(t)\|^p_{L^p(Q)} &\leq M^{p/p^*} \sum_k |f_k(t)|^p \int_Q |\psi_k| dxdv \\
&\leq K M^{p/p^*+1} \sum_k |f_k(t)|^p \text{meas}\,(\mathcal{A}_k) \\
&\leq C \|f\|^p_{L^p(Q)}
\end{aligned}$$

and finally

$$\|\pi_h f\|_{L^\infty(0,T;L^p(Q))} \leq C \|f\|_{L^\infty(0,T;L^p(Q))}.$$

Hence, as $f \in L^\infty(0,T;\mathscr{C}^{m+1}_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$, then

$$\begin{aligned}
\left\|\pi_h f\Big(t, x - v\tfrac{\Delta t}{2}, v\Big)\right\|_{L^\infty(0,T;L^p(Q))} &\leq C \left\|f\Big(t, x - v\tfrac{\Delta t}{2}, v\Big)\right\|_{L^\infty(0,T;L^p(Q))} \\
&\leq C||f||_{L^\infty(0,T;L^p(Q))} \\
&\leq C||f||_{L^\infty(0,T;W^{m+1,p}(Q))}
\end{aligned}$$

and

$$\begin{aligned}
\|\pi_h f(t,x,v - \widetilde{E}(t,x)\Delta t)\|_{L^\infty(0,T;L^p(Q))} &\leq C \|f(t,x,v - \widetilde{E}(t,x)\Delta t)\|_{L^\infty(0,T;L^p(Q))} \\
&\leq C||f||_{L^\infty(0,T;L^p(Q))} \\
&\leq C||f||_{L^\infty(0,T;W^{m+1,p}(Q))},
\end{aligned}$$

which completes the proof of (5.5). Finally, thanks to inequality (3.1), we obtain

$$\begin{aligned}
&\left\|f\Big(t, x - v\tfrac{\Delta t}{2}, v\Big) - \pi_h f\Big(t, x - v\tfrac{\Delta t}{2}, v\Big)\right\|_{L^\infty(0,T;L^p(Q))} \\
&\qquad \leq C h^{m+1} \left\|f\Big(t, x - v\tfrac{\Delta t}{2}, v\Big)\right\|_{L^\infty(0,T;W^{m+1,p}(Q))} \\
&\qquad \leq C h^{m+1} ||f||_{L^\infty(0,T;W^{m+1,p}(Q))}
\end{aligned}$$

and

$$\begin{aligned}
&\|f(t,x,v - \widetilde{E}(t,x)\Delta t) - \pi_h f(t,x,v - \widetilde{E}(t,x)\Delta t)\|_{L^\infty(0,T;L^p(Q))} \\
&\qquad \leq C h^{m+1} ||f(t,x,v - \widetilde{E}(t,x)\Delta t)||_{L^\infty(0,T;W^{m+1,p}(Q))} \\
&\qquad \leq C h^{m+1} ||f||_{L^\infty(0,T;W^{m+1,p}(Q))},
\end{aligned}$$

which completes the proof of the proposition. □

The next lemma gives an estimate of the space discretization error.

LEMMA 5.5. *Assume that* $f \in L^\infty(0,T;\mathscr{C}^{m+1}_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$ *and that* π_h *is a continuous linear interpolation operator from* $W^{m+1,\infty}(Q)$ *onto* X_h*; then there exists a constant* C *such that*

$$\|\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^\infty(Q)} \leq Ch^{m+1}||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))}.$$

Proof. We begin with the following decomposition:

$$\begin{aligned}\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n) &= (\mathcal{T}_1 - \widetilde{\mathcal{T}}_1) \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) \\ &\quad + \widetilde{\mathcal{T}}_1 \circ (\mathcal{T}_2 - \widetilde{\mathcal{T}}_2) \circ \mathcal{T}_1 f(t^n) \\ &\quad + \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ (\mathcal{T}_1 - \widetilde{\mathcal{T}}_1) f(t^n).\end{aligned} \tag{5.9}$$

Using (5.4), (5.5), and (5.6), the decomposition (5.9) gives for the first term

$$\begin{aligned}||(\mathcal{T}_1 - \widetilde{\mathcal{T}}_1) \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n)||_{L^\infty(\Omega)} &\leq Ch^{m+1}|\mathcal{T}_2 \circ \mathcal{T}_1 f(t^n)|_{W^{m+1,\infty}(Q)} \\ &\leq Ch^{m+1}|\mathcal{T}_1 f(t^n)|_{W^{m+1,\infty}(Q)} \\ &\leq Ch^{m+1}|f(t^n)|_{W^{m+1,\infty}(Q)} \\ &\leq Ch^{m+1}||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))},\end{aligned}$$

for the second term of (5.9)

$$\begin{aligned}||\widetilde{\mathcal{T}}_1 \circ (\mathcal{T}_2 - \widetilde{\mathcal{T}}_2) \circ \mathcal{T}_1 f(t^n)||_{L^\infty(Q)} &\leq C||(\mathcal{T}_2 - \widetilde{\mathcal{T}}_2) \circ \mathcal{T}_1 f(t^n)||_{L^\infty(Q)} \\ &\leq Ch^{m+1}|\mathcal{T}_1 f(t^n)|_{W^{m+1,\infty}(Q)} \\ &\leq Ch^{m+1}||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))},\end{aligned}$$

and for the third term of (5.9)

$$\begin{aligned}||\widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ (\mathcal{T}_1 - \widetilde{\mathcal{T}}_1) f(t^n)||_{L^\infty(Q)} &\leq C||(\mathcal{T}_1 - \widetilde{\mathcal{T}}_1) f(t^n)||_{L^\infty(Q)} \\ &\leq Ch^{m+1}||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))},\end{aligned}$$

which proves the lemma. ∎

We continue with the proof of another lemma that gives an estimate of a coupling error between the resolution of the Vlasov and the Poisson equations.

LEMMA 5.6. *Assume that* $f \in L^\infty(0,T;\mathscr{C}^{m+1}_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$ *and that* π_h *is a continuous linear interpolation operator from* $W^{m+1,\infty}(Q)$ *onto* X_h*; then there exists a constant* C *such that*

$$\|\widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^\infty(Q)} \leq C\Delta t \left(e^n + h^{m+1}\right) ||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))},$$

where

$$e^n = \|f(t^n) - f_h(t^n)\|_{L^\infty(Q)}.$$

Proof. On the one hand, we have

$$(\widetilde{\mathcal{T}}_2 - \widetilde{\mathcal{T}}_2^\star)g(t^n) = \pi_h\left(g(t^n, x, v - \Delta t\widetilde{E}^{n+1/2}(x)) - g(t^n, x, v - \Delta t E_h^{n+1/2}(x))\right).$$

On the other hand, we have

$$\begin{aligned}|g(t^n, x, v- &\ \Delta t\widetilde{E}^{n+1/2}(x)) - g(t^n, x, v - \Delta t E_h^{n+1/2}(x))| \\ &\leq \Delta t|\widetilde{E}^{n+1/2}(x) - E_h^{n+1/2}(x)|\,\|\nabla g(t^n)\|_{L^\infty(Q)},\end{aligned}$$

where $\widetilde{E}^{n+1/2}(x)$ and $E_h^{n+1/2}(x)$ can be written as follows:

$$\widetilde{E}^{n+1/2}(x) = \int_0^L K(x,y)\left(\int_{\mathbb{R}} \mathcal{T}_1 f(t^n,y,v)dv - 1\right)dy,$$
$$E_h^{n+1/2}(x) = \int_0^L K(x,y)\left(\int_{\mathbb{R}} \widetilde{\mathcal{T}}_1 f_h(t^n,y,v)dv - 1\right)dy.$$

Then we can write

$$\begin{aligned}
E_h^{n+1/2}(x) - \widetilde{E}^{n+1/2}(x) &= \int_0^L K(x,y)\left(\int_{\mathbb{R}} \left[\widetilde{\mathcal{T}}_1 f_h(t^n,y,v) - \mathcal{T}_1 f(t^n,y,v)\right]dv\right)dy,\\
&= \int_0^L K(x,y)\left(\int_{|v|\le Q(T)} \pi_h\Big[f_h\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) - f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big)\Big]dv\right)dy\\
&\quad + \int_0^L \int_{|v|\le Q(T)} K(x,y)\Big(\pi_h f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) - f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big)\Big)dvdy,
\end{aligned}$$

so that we get

$$\begin{aligned}
&\|E_h^{n+1/2} - \widetilde{E}^{n+1/2}\|_{L^\infty([0,L])}\\
&\qquad \le LQ(T)||K||_{L^\infty}\left\|\pi_h\Big[f_h\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) - f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big)\Big]\right\|_{L^\infty(Q)}\\
&\qquad\quad + LQ(T)||K||_{L^\infty}\left\|\pi_h f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) - f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big)\right\|_{L^\infty(Q)},
\end{aligned} \tag{5.10}$$

and using (5.4), (5.5), and (5.6),

$$\begin{aligned}
\|E_h^{n+1/2} - \widetilde{E}^{n+1/2}\|_{L^\infty([0,L])} &\le LQ(T)||K||_{L^\infty}||\pi_h||_{L^\infty}\,\|f_h(t^n) - f(t^n)\|_{L^\infty(Q)}\\
&\quad + CLQ(T)||K||_{L^\infty}h^{m+1}||f(t^n)||_{W^{m+1,\infty}(Q)}.
\end{aligned} \tag{5.11}$$

Finally, we obtain

$$\|E_h^{n+1/2} - \widetilde{E}^{n+1/2}\|_{L^\infty([0,L])} \le C\left(e^n + h^{m+1}\right)$$

and, as a consequence,

$$\|(\widetilde{\mathcal{T}}_2 - \widetilde{\mathcal{T}}_2^\star)g(t^n)\|_{L^\infty(Q)} \le C\Delta t\left(e^n + h^{m+1}\right)\|\nabla g(t^n)\|_{L^\infty(Q)}. \tag{5.12}$$

Then, using (5.4) and (5.12),

$$\begin{aligned}
\|\widetilde{\mathcal{T}}_1 \circ (\widetilde{\mathcal{T}}_2 - \widetilde{\mathcal{T}}_2^\star) \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^\infty(Q)} &\le C\|(\widetilde{\mathcal{T}}_2 - \widetilde{\mathcal{T}}_2) \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^\infty(Q)}\\
&\le C\Delta t\left(e^n + h^{m+1}\right)\|\nabla(\widetilde{\mathcal{T}}_1 f(t^n))\|_{L^\infty(Q)}.
\end{aligned}$$

Now we estimate the term $\|\nabla(\widetilde{\mathcal{T}}_1 f(t^n))\|_{L^\infty(Q)}$. We can do this in the following way. Using (3.1), we get

$$\begin{aligned}
\|\nabla(\widetilde{\mathcal{T}}_1 f(t^n))\|_{L^\infty(Q)} &\le \left\|\nabla\Big(\pi_h f\Big(t^n, x - v\tfrac{\Delta t}{2}, v\Big)\Big)\right\|_{L^\infty(Q)} \\
&\le \left\|\nabla\Big[(\pi_h f - f)\Big(t^n, x - v\tfrac{\Delta t}{2}, v\Big)\Big]\right\|_{L^\infty(Q)} \\
&\qquad + \left\|\nabla\Big(f\Big(t^n, x - v\tfrac{\Delta t}{2}, v\Big)\Big)\right\|_{L^\infty(Q)} \\
&\le Ch^m ||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))} + ||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))} \\
&\le C||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))}.
\end{aligned}$$

In fact, this estimation is due to the continuity of π_h from $W^{m+1,\infty}(Q)$ onto X_h. Then we finally obtain

$$\|\widetilde{\mathcal{T}}_1 \circ (\widetilde{\mathcal{T}}_2 - \widetilde{\mathcal{T}}_2^\star) \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^\infty(Q)} \le C\Delta t \left(e^n + h^{m+1}\right) ||f||_{L^\infty(0,T;W^{m+1,\infty}(Q))},$$

which completes the proof. ∎

We now state the last lemma, which gives information about the stability of the numerical scheme.

LEMMA 5.7. *Let π_h be the interpolation operator from $W^{2,\infty}(Q)$ onto X_h with $P_m = P_1$; then we have*

$$\|\widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 (f(t^n) - f_h(t^n))\|_{L^\infty(Q)} \le e^n. \tag{5.13}$$

Proof. As π_h is a linear interpolation operator, the basis functions satisfy

$$0 \le \psi_k \le 1$$

and

$$\sum_k \psi_k = 1,$$

and therefore we have

$$\|\pi_h\|_{L^\infty} = \sup_{\substack{f \in L^\infty \\ f \neq 0}} \frac{\|\pi_h f\|_{L^\infty(Q)}}{\|f\|_{L^\infty(Q)}} \le 1.$$

Indeed we have

$$\begin{aligned}
|\pi_h g| &= \left|\sum_k g(x_k, v_k)\psi_k(x, v)\right| \\
&\le \sum_k |g_k|\psi_k(x, v) \\
&\le \|g\|_{L^\infty} \sum_k \psi_k = \|g\|_{L^\infty}.
\end{aligned}$$

As a consequence we obviously obtain

$$\begin{aligned}
\|\widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 (f(t^n) - f_h(t^n))\|_{L^\infty(Q)} &\le \|\widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 (f(t^n) - f_h(t^n))\|_{L^\infty(Q)} \\
&\le \|\widetilde{\mathcal{T}}_1 (f(t^n) - f_h(t^n))\|_{L^\infty(Q)} \\
&\le \|f(t^n) - f_h(t^n)\|_{L^\infty(Q)},
\end{aligned}$$

which completes the proof. ∎

Now we can return to the proof of the main theorem.

Proof of the main theorem. We want to evaluate the global error at time t^{n+1}:

$$e^{n+1} = ||f(t^{n+1},x,v) - f_h(t^{n+1},x,v)||_{L^\infty(Q)}.$$

We decompose $f(t^{n+1},x,v) - f_h(t^{n+1},x,v)$ as

$$\begin{aligned} f(t^{n+1},x,v) - f_h(t^{n+1},x,v) = {} & f(t^{n+1},x,v) - \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n,x,v) \\ & + \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n,x,v) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n,x,v) \\ & + \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n,x,v) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 f(t^n,x,v) \\ & + \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 (f(t^n,x,v) - f_h(t^n,x,v)). \end{aligned} \tag{5.14}$$

Finally if we put together Lemmas 5.3, 5.5, 5.6, 5.7, we obtain the following estimation:

$$e^{n+1} \leq (1 + C\Delta t)e^n + C\left(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\left(\Delta t^3 + h^2 + h^2\Delta t\right).$$

A discrete Gronwall inequality enables us to get

$$e^{n+1} \leq \exp(CT)e^0 + C\left(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\left(\Delta t^2 + \frac{h^2}{\Delta t} + h^2\right).$$

As e^0 is only a fixed interpolation error, we obtain

$$e^n \leq C\left(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\left(\Delta t^2 + \frac{h^2}{\Delta t} + h^2\right).$$

In order to prove the convergence of the electric field, we estimate

$$||E(t^{n+1/2}) - E_h^{n+1/2}||_{L^\infty([0,L])}.$$

To estimate this term we proceed as in the proof of Lemmas 5.6 and 5.3. Then we obtain

$$||\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}||_{L^\infty([0,L])} \leq C\left(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\left(\Delta t^2 + h^2 + \frac{h^2}{\Delta t}\right)$$

and

$$||E(t^{n+1/2}) - \widetilde{E}(t^{n+1/2})||_{L^\infty([0,L])} \leq C\left(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\Delta t^2$$

so that

$$||E(t^{n+1/2}) - E_h^{n+1/2}||_{L^\infty([0,L])} \leq C\left(||f||_{\mathscr{C}^2(0,T;W^{2,\infty}(Q))}\right)\left(\Delta t^2 + h^2 + \frac{h^2}{\Delta t}\right). \qquad \square$$

5.4. Other results. We can prove the convergence of our numerical scheme under weaker regularity assumptions. Following the proof of existence and uniqueness of the solutions of the Cauchy problem for the Vlasov–Maxwell system in one dimension made by Cooper and Klimas [15], if we take f_0 such that

$$f_0 \in \mathscr{C}_{b,per_x} \cap W_c^{1,\infty}(\mathbb{R}_x \times \mathbb{R}_v),$$

the Vlasov–Poisson periodic system given by (2.7), (2.8), (2.9), and (2.10) has a unique solution (f, E) such that

$$f \in \mathscr{C}_b\left(0,T;\mathscr{C}_{b,per_x} \cap W_c^{1,\infty}(\mathbb{R}_x \times \mathbb{R}_v)\right),$$

$$\partial_t f \in L^\infty\left(0,T;L_c^\infty(\mathbb{R}_x \times \mathbb{R}_v)\right),$$

where the derivative is taken in the sense of distribution, and

$$E \in \mathscr{C}^1\left(0,T;\mathscr{C}^1_{b,per_x}(\mathbb{R}_x)\right).$$

Now we state the theorem.

THEOREM 5.8. *Assume that* $f_0 \in \mathscr{C}_{b,per_x} \cap W_c^{1,\infty}(\mathbb{R}_x \times \mathbb{R}_v)$. *Let* $\alpha > 0$, $h \sim \Delta t^{1/\varepsilon}$, *with* $0 < \varepsilon < 1$; *then* (f_h, E_h), *the numerical solution of the periodic Vlasov–Poisson system, converges towards* (f, E), *and there exists a constant* $C = C(\|f\|_{\mathscr{C}_b(0,T;W^{1,\infty}(Q))}, \|\partial_t f\|_{L^\infty(0,T;L^\infty(Q))})$ *independent of* Δt *and* h *such that*

$$||f - f_h||_{\ell^\infty(0,T;L^\infty(Q))} \leq C\left(\Delta t + h + h^{1-\varepsilon}\right)$$

and

$$||E - E_h||_{\ell^\infty(0,T;L^\infty([0,L]))} \leq C\left(\Delta t + h + h^{1-\varepsilon}\right).$$

Proof. In order to prove Theorem 5.8 we have to examine how Lemmas 5.3, 5.5, 5.6, 5.7 and Proposition 5.4 can be adapted to the new regularity assumptions.

We begin with Lemma 5.3. Now we cannot apply Taylor expansion, since the solution is not regular enough. Thus we have to rewrite all the estimates. First we have

$$\begin{aligned} X(t^{n+1/2}) - (x - v\Delta t/2) &= X(t^{n+1/2}) - \left(X(t^{n+1}) - V(t^{n+1})\tfrac{\Delta t}{2}\right) \\ &= \int_{t^{n+1}}^{t^{n+1/2}} \left(V(t) - V(t^{n+1})\right) dt \\ &= \int_{t^{n+1}}^{t^{n+1/2}} \int_{t^{n+1}}^{t} E\left(\tau, X(\tau)\right) d\tau dt \\ &\leq C\Delta t^2 \left\|E\right\|_{L^\infty(0,T;L^\infty([0,L]))} \\ &\leq C\Delta t^2. \end{aligned} \tag{5.15}$$

Next we note that we have the following decomposition:

$$f(t^{n+1/2}, y, v) - f\left(t^n, y - v\tfrac{\Delta t}{2}, v\right) = \int_{t^n}^{t^{n+1/2}} \partial_t f(t, y, v) dt + \int_{y - v\Delta t/2}^{y} \partial_x f(t^n, x, v) dx.$$

As $f \in \mathscr{C}_b(0,T;{W_c}^{1,\infty}(Q))$ and $\partial_t f \in L^\infty\left(0,T;L_c^\infty(Q)\right)$, integrating the previous decomposition, we obtain

$$\begin{aligned} &\int_0^L \int_{\mathbb{R}_v} \left| f(t^{n+1/2}, y, v) - f\left(t^n, y - v\tfrac{\Delta t}{2}, v\right)\right| dy dv \\ &\quad \leq \int_0^L \int_{\mathbb{R}_v} \int_{t^n}^{t^{n+1/2}} \left|\partial_t f\left(t, y, v\right)\right| dt dv dy + \int_0^L \int_{\mathbb{R}_v} \int_{y - v\Delta t/2}^{y} \left|\partial_x f\left(t^n, x, v\right)\right| dx dv dy, \end{aligned}$$

and then

$$\int_0^L \int_{\mathbb{R}_v} \left| f(t^{n+1/2}, y, v) - f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) \right| dydv$$
$$\leq CLQ^2(T)\Delta t \left(\|\partial_t f\|_{L^\infty(0,T,L^\infty(Q))} + \|\partial_x f\|_{L^\infty(0,T;L^\infty(Q))} \right)$$
$$\leq C\Delta t, \tag{5.16}$$

so that, using (2.7),

$$|E(t^{n+1/2}, x) - \widetilde{E}(t^{n+1/2}, x)| \leq C\Delta t. \tag{5.17}$$

Then we have

$$V(t^n) - \widetilde{V}(t^n) = \int_{t^{n+1}}^{t^n} E\left(t, X(t)\right) dt + \Delta t \widetilde{E}\Big(t^{n+1/2}, x - v\tfrac{\Delta t}{2}\Big)$$
$$= \int_{t^{n+1}}^{t^n} \left(E\left(t, X(t)\right) - E(t^{n+1/2}, X(t))\right) dt$$
$$+ \int_{t^{n+1}}^{t^n} \left(E(t^{n+1/2}, X(t)) - E(t^{n+1/2}, X(t^{n+1/2}))\right) dt$$
$$+ \Delta t \Big\{ E\Big(t^{n+1/2}, X(t^{n+1}) - V(t^{n+1})\tfrac{\Delta t}{2}\Big) - E(t^{n+1/2}, X(t^{n+1/2})) \Big\}$$
$$+ \Delta t \Big\{ \widetilde{E}\Big(t^{n+1/2}, X(t^{n+1}) - V(t^{n+1})\tfrac{\Delta t}{2}\Big)$$
$$- E\Big(t^{n+1/2}, X(t^{n+1}) - V(t^{n+1})\tfrac{\Delta t}{2}\Big) \Big\}.$$

As $E \in \mathscr{C}^1(0,T;\mathscr{C}^1_{b,per_x}(\mathbb{R}_x))$, we obtain

$$\sup \left\{ \left| V(t^n; t^{n+1}, x, v) - \widetilde{V}(t^n; t^{n+1}, x, v) \right| \mid \forall (x,v) \in [0,L] \times \mathbb{R} \right\}$$
$$\leq C\Delta t^2 \mathrm{Lip}\left(E(.,x)\right) + CQ(T)\Delta t^2 \mathrm{Lip}\left(E(t,.)\right) + C\Delta t^3 \mathrm{Lip}\left(E(t,.)\right) + C\Delta t^2$$
$$\leq C\Delta t^2. \tag{5.18}$$

We go on with the estimate of $X(t^n) - \widetilde{X}(t^n)$. We have

$$X(t^n) - \widetilde{X}(t^n) = \int_{t^{n+1}}^{t^n} \int_{t^{n+1}}^{t} E\left(\tau, X(\tau)\right) d\tau dt - \tfrac{\Delta t^2}{2} \widetilde{E}\Big(t^{n+1/2}, X(t^{n+1}) - V(t^{n+1})\tfrac{\Delta t}{2}\Big)$$

so that

$$\sup \left\{ \left| X(t^n) - \widetilde{X}(t^n) \right| \mid \forall (x,v) \in [0,L] \times \mathbb{R} \right\} \leq C\Delta t^2 \left(\|E\|_{L^\infty} + \|\widetilde{E}\|_{L^\infty} \right) \leq C\Delta t^2. \tag{5.19}$$

Now we use the estimates (5.18) and (5.19) in order to bound the quantity

$$\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - f\left(t^{n+1}, x, v\right) = f(t^n, \widetilde{X}(t^n), \widetilde{V}(t^n)) - f\left(t^n, X(t^n), V(t^n)\right)$$

in the L^∞ norm. As we have the continuous embedding $W^{1,\infty} \hookrightarrow \mathscr{C}^{0,1}$, then

$$\begin{aligned} \left\|\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - f\left(t^{n+1}\right)\right\|_{L^\infty(Q)} &\leq C\mathrm{Lip}\left(f(t^n, ., .)\right)\Delta t^2, \\ &\leq C \sup_{t\in[0,T]} \mathrm{Lip}\left(f(t, ., .)\right)\Delta t^2, \\ &\leq C\Delta t^2. \end{aligned} \tag{5.20}$$

Following the proof of Proposition 5.4, if we take $f \in \mathscr{C}_b(0, T; W^{1,\infty}_c(Q))$, $E \in \mathscr{C}^1(0, T; \mathscr{C}^1_{b,per_x}(\mathbb{R}))$, and if we take the derivative in the sense of distribution, then, using (3.1), we still have (with $m \in \{0, 1\}$)

$$||\mathcal{T}_i f||_{L^\infty(0,T;W^{m,\infty}(Q))} \leq C||f||_{L^\infty(0,T;W^{m,\infty}(Q))}, \tag{5.21}$$

$$||\widetilde{\mathcal{T}}_i f||_{L^\infty(0,T;L^\infty(Q))} \leq C||f||_{L^\infty(0,T;W^{m,\infty}(Q))}, \tag{5.22}$$

and

$$\|(\mathcal{T}_i - \widetilde{\mathcal{T}}_i)f\|_{L^\infty(0,T;L^\infty(Q))} \leq Ch||f||_{L^\infty(0,T;W^{1,\infty}(Q))}. \tag{5.23}$$

As a consequence, Lemma 5.5 supplies the estimate

$$\|\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^\infty(Q)} \leq Ch.$$

The estimate of Lemma 5.6 has to be replaced by

$$\|\widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^\infty(Q)} \leq C\Delta t\left(e^n + h\right).$$

In order to justify this inequality we just have to show that $\mathrm{Lip}(\widetilde{\mathcal{T}}_1 f(t^n))$ is bounded. Indeed we have

$$\mathrm{Lip}(\widetilde{\mathcal{T}}_1 f(t^n)) = \mathrm{Lip}\Big(\pi_h f\Big(t^n, x - v\tfrac{\Delta t}{2}, v\Big)\Big) \leq \|\pi_h\|_{L^\infty}\mathrm{Lip}\left(f(t^n, ., .)\right) < +\infty.$$

Finally, we get all the desired a priori estimates by seeing that the stability result (5.13) still holds. Then the proof of the theorem is the same as that for Theorem 5.1, and we get

$$||f - f_h||_{\ell^\infty(0,T;L^\infty(Q))} \leq C\left(\Delta t + h + \frac{h}{\Delta t}\right)$$

and

$$||E - E_h||_{\ell^\infty(0,T;L^\infty([0,L]))} \leq C\left(\Delta t + h + \frac{h}{\Delta t}\right).$$

Now if we take $\Delta t \sim h^\varepsilon$ with $0 < \varepsilon < 1$, we get the desired result. In fact the best ε to choose is $1/2$ so that convergence holds with order $1/2$. ∎

Remark 5.9. Under the regularity assumptions $f_0 \in \mathscr{C}^{m+1}_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)$, if there exists an interpolation operator π_h that satisfies both a consistency condition such as

$$\|f - \pi_h f\|_{L^\infty(0,T;L^p(Q))} \leq Ch^{m+1}||f||_{L^\infty(0,T;W^{m+1,p}(Q))} \tag{5.24}$$

and a stability condition such as

$$\|\pi_h f\|_{L^\infty(0,T;L^p(Q))} \leq (1+Ch)||f||_{L^\infty(0,T;L^p(Q))}, \tag{5.25}$$

then our method can easily be applied to prove the convergence of high order schemes in the L^p norm and to find error estimates such as

$$||f - f_h||_{\ell^\infty(0,T;L^p(Q))} \leq C\left(||f||_{\mathscr{C}^2(0,T;W^{m+1,p}(Q))}\right)\left(\Delta t^2 + h^{m+1} + \frac{h^{m+1}}{\Delta t}\right)$$

and

$$||E - E_h||_{\ell^\infty(0,T;L^\infty([0,L]))} \leq C\left(||f||_{\mathscr{C}^2(0,T;W^{m+1,p}(Q))}\right)\left(\Delta t^2 + h^{m+1} + \frac{h^{m+1}}{\Delta t}\right).$$

Unfortunately Lagrange interpolations of high order do not satisfy the stability condition (5.25). Besides, it seems difficult but not impossible to build interpolation operators π_h which satisfy both conditions (5.24) and (5.25).

If we use a Lagrange interpolation operator of high order, the discrete solution $f_h(t^n)$ belongs to $W^{1,p}(Q)$. The numerical scheme consists of a succession of transport and projection on the finite element space generated by the Lagrange finite element of high order. The transport operation leaves the norm of the solution unchanged. Then the scheme is stable if the interpolation operator π_h is stable, i.e., $\|\pi_h\|_{L^p} \leq 1 + \varepsilon(h)$ with $\lim_{h\to 0}\varepsilon(h) = 0$. Unfortunately Lagrange interpolation does not have nice properties of stability. Let $\tau_{h,\xi}$ be a translation operator such that $\tau_{z,\xi}f_h(t^n,x,v) = f_h(t^n,x-z,v-\xi) = g_h(t^n,x,v)$. Therefore $g_h(t^n) \in W^{1,p}(Q)$, and we have

$$\begin{aligned}
\|\pi_h \circ \tau_{z,\xi} f_h(t^n)\|_{L^p(Q)} &= \|\pi_h g_h(t^n)\|_{L^p(Q)} \\
&\leq \|g_h(t^n)\|_{L^p(Q)} + \|\pi_h g_h(t^n) - g_h(t^n)\|_{L^p(Q)} \\
&\leq \|g_h(t^n)\|_{L^p(Q)} + Ch\,|g_h(t^n)|_{W^{1,p}(Q)} \\
&\leq \|g_h(t^n)\|_{L^p(Q)} + C,
\end{aligned}$$

since $|g_h(t^n)|_{W^{1,p}(Q)} \sim O(h^{-1})$ and with C independent of h and such that $C > 1$.

We can also prove the convergence of our numerical scheme with noncompactly supported initial data. If we take f_0 such that

$$f_0 \in \mathscr{C}_{b,per_x} \cap W^{1,\infty} \cap W^{1,1}(\mathbb{R}_x \times \mathbb{R}_v),$$

$$0 < f_0 \leq (1+|v|)^{-\lambda}, \quad v\nabla f_0 \in L^\infty_x\left(L^1_v\right),$$

and if we suppose that there exists a constant $R > 0$ such that

$$\mathcal{L}(f_0,R)(\xi) = \sup\left\{ \frac{|f_0(x,v) - f_0(y,w)|}{\|(x,v)-(y,w)\|_2} \;\middle|\; x,y \in [0,L],\ v,w \in \mathbb{R},\right.$$
$$\left.(x,v) \neq (y,w),\ |v-\xi| \leq R,\ |w-\xi| \leq R\right\}(1+|\xi|) \in L^\infty \cap L^1(\mathbb{R}_\xi), \tag{5.26}$$

where $(x,v) \in [0,L] \times \mathbb{R}$ and $\|(x,v)\|_2 = \sqrt{x^2+v^2}$, the periodic Vlasov–Poisson system given by (2.7), (2.8), (2.9), and (2.10) has a unique solution (f,E) such that

$$0 < f(t,x,v) \le (1+|v|)^{-\lambda}, \tag{5.27}$$

$$f \in \mathscr{C}_b\left(0,T;\mathscr{C}_{b,per_x} \cap W^{1,\infty} \cap W^{1,1}(\mathbb{R}_x \times \mathbb{R}_v)\right),$$

$$v\nabla f,\ \partial_t f \in L^\infty\left(0,T;L^\infty_x\left(L^1_v\right)\right),$$

where the derivative is taken in the sense of distribution and

$$E \in \mathscr{C}^1\left(0,T;\mathscr{C}^1_{b,per_x}(\mathbb{R}_x)\right).$$

In addition, there exists a constant $C(T) > 0$ such that $\forall t \in [0,T]$,

$$\begin{aligned}
&\mathcal{L}(f(t), R + C(T))(\xi) \\
&\quad = \sup\left\{ \frac{|f(t,x,v) - f(t,y,w)|}{\|(x,v)-(y,w)\|_2} \;\middle|\; x,y \in [0,L],\ v,w \in \times\mathbb{R},\right. \\
&\qquad\qquad \left.(x,v) \neq (y,w),\ |v-\xi| \le R + C(T),\ |w-\xi| \le R + C(T)\right\}(1+|\xi|) \\
&\quad \in L^\infty \cap L^1(\mathbb{R}_\xi).
\end{aligned}$$

Now we state the theorem.

THEOREM 5.10. *Assume that $f_0 \in \mathscr{C}_{b,per_x} \cap W^{1,\infty} \cap W^{1,1}(\mathbb{R}_x \times \mathbb{R}_v)$, $0 \le f_0 \le (1+|v|)^{-\lambda}$, $\forall \lambda > 1$ and that f_0 satisfies (5.26). Let α be such that $0 < \alpha < \lambda$, and suppose that the bound of velocity support R evolves as $h^{-1/\alpha}$. Then (f_h, E_h), the numerical solution of the periodic Vlasov–Poisson system, converges towards (f,E), and there exists a positive function μ such that $\lim_{h\to 0}\mu(h) = 0$, and a constant $C = C(\|f\|_{L^\infty(0,T;W^{1,\infty}(Q))}, \|f\|_{L^\infty(0,T;W^{1,1}(Q))}, \|\partial_t f\|_{L^\infty(0,T;L^\infty_x(L^1_v))}, \|v\nabla f\|_{L^\infty(0,T;L^\infty_x(L^1_v))})$ independent of Δt, h such that*

$$||f - f_h||_{\ell^\infty(0,T;L^{1,\infty}(Q))} \le C\left(\Delta t + h + (h+\mu(h))^{1-1/\sigma} + h^{\lambda/\alpha}\right)$$

and

$$||E - E_h||_{\ell^\infty(0,T;L^\infty([0,L]))} \le C\left(\Delta t + h + (h+\mu(h))^{1-1/\sigma} + h^{\lambda/\alpha}\right),$$

where $\Delta t \sim (h+\mu(h))^{1/\sigma}$, with $\sigma > 1$.

Before giving the proof of Theorem 5.10, we need to establish the L^1 stability of π_h by proving the following two lemmas.

LEMMA 5.11. *Assume that $0 \le f_0(x,v) \le \zeta(x,v) \sim (1+|v|)^{-\lambda}$, for $\lambda > 1$. Then there exists a constant C, depending only on T, L, and f_0, such that*

$$0 \le f_h(t,x,v) \le C\zeta_h(x,v), \quad t \in [0,T], \quad (x,v) \in Q, \tag{5.28}$$

where

$$\zeta_h(x,v) = \sum_k \frac{1}{(1+|v_k|)^\lambda}\psi_k(x,v).$$

There also exists a constant $C > 0$ such that

$$||f_h(t)||_{L^1(Q)} \le C, \quad t \in [0,T]. \tag{5.29}$$

Proof. We begin with the transport in x. Let us notice that there exists a constant R independent of h such that for every triangle T_m of the triangulation $\mathcal{T}_h$ there exists a ball $B(a_m, Rh)$ of center a_m and radius Rh which contains T_m. Let N_k be a vertex of triangle T_m. If we consider transport in x, the origin of the characteristic, $x_k^\star = x_k - v_k\Delta t/2$, which ends at N_k, belongs to a triangle $T_m^\star$. Let $\mu(h)$ be a positive function such that $\lim_{h\to 0}\mu(h) = 0$. If N_o, N_p, and N_q are the vertices of the triangle $T_m^\star$, we have

$$|v_k - v_o| \leq 2Rh \leq 2R(h + \mu(h)) \leq 2R\varepsilon(h),$$

$$|v_k - v_p| \leq 2Rh \leq 2R(h + \mu(h)) \leq 2R\varepsilon(h),$$

and

$$|v_k - v_q| \leq 2Rh \leq 2R(h + \mu(h)) \leq 2R\varepsilon(h),$$

where $\varepsilon(h) = h + \mu(h)$. On the other hand, we note that

$$\frac{\zeta_h(x_j, v_j)}{\zeta_h(x_k, v_k)} = \frac{(1 + |v_k|)^\lambda}{(1 + |v_j|)^\lambda} \leq 1 + C_1(\lambda, R)\varepsilon(h), \quad j = \{o, p, q\}. \tag{5.30}$$

Now, if we consider the transport in v, the origin of the characteristic $v_k^* = v_k - E_h(t^{n+1/2})\Delta t$ which ends at N_k belongs to a triangle T_m^*. If N_i, N_s, and N_l are the vertices of a triangle T_m^*, as $\|E_h\|_{\ell^\infty(0,T;L^\infty([0,L]))}$ is bounded we have

$$|v_k - v_i| \leq C\Delta t, \quad |v_k - v_s| \leq C\Delta t, \quad \text{and } |v_k - v_l| \leq C\Delta t,$$

and then we have

$$\frac{\zeta_h(x_j, v_j)}{\zeta_h(x_k, v_k)} = \frac{(1 + |v_k|)^\lambda}{(1 + |v_j|)^\lambda} \leq 1 + C_2(\lambda, R)\Delta t, \quad j = \{i, s, l\}. \tag{5.31}$$

If we set $b_1 = 1 + C_1(\lambda, R)\varepsilon(h)$, $b_2 = 1 + C_2(\lambda, R)\Delta t$, and $b = b_1 b_2 b_1$, then we have

$$f_h(0, x_k, v_k) \leq \frac{1}{(1 + |v_k|)^\lambda} \leq \frac{b^0}{(1 + |v_k|)^\lambda}$$

and consequently

$$f_h(0, x, v) \leq b^0 \zeta_h(x, v).$$

If we assume that

$$f_h(t^n, x_k, v_k) \leq b^n \zeta_h(x_k, v_k)$$

and consequently

$$f_h(t^n, x, v) \leq b^n \zeta_h(x, v),$$

the numerical scheme gives for the first half advection with the respect to the variable x

$$f_h(t^{n+1/2}, x_k, v_k) = \sum_l f_h(t^n, x_l, v_l)\psi_l\left(x_k - v_k\frac{\Delta t}{2}, v_k\right).$$

Let $T_m^\star$ be the triangle which contains the origin of the characteristic coming from the node N_k. Let N_o, N_p, and N_q be the three vertices of $T_m^\star$. Then we can write

$$\frac{f_h(t^{n+1/2}, x_k, v_k)}{\zeta_h(x_k, v_k)} = \lambda_o\frac{f_{h,o}^n}{\zeta_h(x_k, v_k)} + \lambda_p\frac{f_{h,p}^n}{\zeta_h(x_k, v_k)} + \lambda_q\frac{f_{h,q}^n}{\zeta_h(x_k, v_k)},$$

where

$$\lambda_l = \psi_l\left(x_k - v_k\frac{\Delta t}{2}, v_k\right)$$

and

$$f_{h,l}^n = f_h(t^n, x_l, v_l).$$

Using the property (5.30) of ζ_h and the property

$$\lambda_o + \lambda_p + \lambda_q = 1,$$

we obtain

$$\frac{f_h(t^{n+1/2}, x_k, v_k)}{\zeta_h(x_k, v_k)} \le b^n\lambda_o\frac{\zeta_h(x_o, v_o)}{\zeta_h(x_k, v_k)} + b^n\lambda_p\frac{\zeta_h(x_p, v_p)}{\zeta_h(x_k, v_k)} + b^n\lambda_q\frac{\zeta_h(x_q, v_q)}{\zeta_h(x_k, v_k)} \le b_1 b^n.$$

In the same way for the two other advections we finally obtain

$$\frac{f_h(t^{n+1}, x_k, v_k)}{\zeta_h(x_k, v_k)} \le b^{(n+1)} \quad \forall\ N_k \in \mathcal{T}_h.$$

For a finite time T and $\forall n \in \{0, \dots, T/\Delta t\}$, if we consider $\varepsilon(h) \le \Delta t$, we have $b \le 1 + C(C_1, C_2)\Delta t$, $b^{(n+1)} \le \exp(C(C_1, C_2)T)$, and as in the continuous case there exists a majorizing function of the discrete distribution

$$f_h(t, x, v) \le C\zeta_h(x, v) \quad \forall t \in [0, T], \quad \forall (x, v) \in Q.$$

In order to prove (5.29), we note that

$$\begin{aligned}\int_{\mathbb{R}\zeta_h(x,v)dxdv} &= \sum_k \frac{1}{(1 + |v_k|)^\lambda}\int_{\mathbb{R}} \psi_k(x, v)dxdv \\ &= \sum_k \frac{\text{meas}(\mathcal{A}_k)}{(1 + |v_k|)^\lambda} \\ &\le C\int_{\mathbb{R}} \frac{1}{(1 + |v|)^\lambda} < +\infty. \qquad \square\end{aligned}$$

Let $\mathcal{A}_k$ be the area associated with the node N_k and $\psi_k \in P_1$; then we have

$$\text{meas}(\mathcal{A}_k) = \int_{\mathbb{R}} \psi_k(x, v)dxdv = \frac{|\text{supp}\psi_k|}{3}.$$

We introduce χ_k, the characteristic function defined as follows:

$$\chi_k(x,v) = \begin{cases} 1 \text{ if } (x,v) \in \mathcal{A}_k, \\ 0 \text{ otherwise.} \end{cases}$$

Then we introduce the function $g_h^n(x,v)$ defined by

$$g_h^n(x,v) = \sum_k g_h^n(x_k,v_k)\chi_k(x,v),$$

with

$$g_h^n(x_k,v_k) = f_h(t^n,x_k,v_k).$$

We note that

$$||f_h(t^n)||_{L^1(Q)} = ||g_h^n||_{L^1(Q)}.$$

Moreover, as for the proof of the Lemma 5.11, we can prove that

$$0 \leq g_h^n(x,v) \leq C\gamma_h(x,v) \ \ \forall n \in [0,N], \ N = \left[\frac{T}{\Delta t}\right], \ \ (x,v) \in Q,$$

where

$$\gamma_h(x,v) = \sum_k \frac{1}{(1+|v_k|)^\lambda}\chi_k(x,v).$$

We notice that there exists another constant C independent of h such that

$$0 < \gamma_h \leq C(1+|v|)^{-\lambda}.$$

Now we state the lemma which shows the L^1 stability of the interpolation operator π_h.

LEMMA 5.12. *Let $g \in \mathscr{C}_b \cap L^1(Q)$ and $0 < g \leq C(1+|v|)^{-\lambda}$; then there exists a positive function μ, where $\lim_{h\to 0}\mu(h) = 0$, such that*

$$||\pi_h g||_{L^1(Q)} \leq ||g||_{L^1(Q)} + \mu(h).$$

Proof. We have

$$\begin{aligned} ||\pi_h g||_{L^1(Q)} &= \sum_k g_k \int_{\mathbb{R}}\int_0^L \psi_k(x,v)dxdv = \sum_k g_k \text{meas}(\mathcal{A}_k) \\ &= \sum_k g_k \int_Q \chi_k(x,v)dxdv = \int_Q \sum_k g_k\chi_k(x,v)dxdv \\ &= ||g_h||_{L^1(Q)}. \end{aligned}$$

As

$$g_h(x,v) \leq C(1+|v|)^{-\lambda}$$

and

$$\lim_{h\to 0} g_h = g \quad \text{a.e.},$$

the dominated convergence theorem asserts that

$$\lim_{h\to 0} \int_Q |g_h - g|\, dxdv = 0,$$

and as a consequence there exists a positive function μ with $\lim_{h\to 0} \mu(h) = 0$ such that

$$\int_Q |g_h - g|\, dxdv \le \mu(h).$$

Then we deduce that

$$\left| \|\pi_h g\|_{L^1(Q)} - \|g\|_{L^1(Q)} \right| = \left| \|g_h\|_{L^1(Q)} - \|g\|_{L^1(Q)} \right| \le \int_Q |g_h - g|\, dxdv \le \mu(h).$$

Finally we deduce that

$$||\pi_h g||_{L^1(Q)} \le ||g||_{L^1(Q)} + \mu(h). \qquad \square$$

Now we can return to the proof of Theorem 5.10.

Proof of Theorem 5.10. In order to prove the theorem we have to see how the a priori estimates (5.16), (5.17), (5.18), (5.19), and (5.20) are modified and obtain the same kind of a priori estimates in the L^1 norm.

As $v\nabla f,\ \partial_t f \in L^\infty(0,T; L^\infty_x(L^1_v))$ the estimate (5.16) becomes

$$\begin{aligned}\int_0^L \int_{\mathbb{R}_v} |f(t^{n+1/2}, y, v) - f\left(t^n, y - v\Delta t/2, v\right)| dydv \\ &\le CL\Delta t \left(\|\partial_t f\|_{L^\infty(0,T,L^\infty_x(L^1_v))} + \|v\partial_x f\|_{L^\infty(0,T;L^\infty_x(L^1_v))} \right) \\ &\le C\Delta t,\end{aligned}$$

so that we still have

$$|E(t^{n+1/2}, x) - \widetilde{E}(t^{n+1/2}, x)| \le C\Delta t.$$

The estimate (5.19) still holds, but the estimate (5.18) changes into

$$|V(t^n) - \widetilde{V}(t^n)| \le C(1 + |v|)\Delta t^2.$$

Then the estimate (5.20) becomes

$$\begin{aligned}\left\| \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - f\left(t^{n+1}\right) \right\|_{L^\infty(Q)} &\le \sup_{v\in\mathbb{R}} \left\{ \mathcal{L}(f(t^n), C(T))(v) \right\} \Delta t^2 \\ &\le C\Delta t^2,\end{aligned}$$

and in the L^1 norm we have

$$
\begin{aligned}
&\left\|\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - f\left(t^{n+1}\right)\right\|_{L^1(Q)} \\
&\leq \int_0^L \int_{\mathbb{R}} \Big| f\left(t^n, \widetilde{X}(t^n; t^{n+1}, x, v), \widetilde{V}(t^n; t^{n+1}, x, v)\right) \\
&\qquad\qquad - f\left(t^n, X(t^n; t^{n+1}, x, v), V(t^n; t^{n+1}, x, v)\right)\Big| dvdx \\
&\leq \int_0^L \int \sup\Big\{ \|(\chi,\xi)-(y,w)\|_2^{-1} \cdot |f(t^n,\chi,\xi) - f(t^n,y,w)| \ \Big| \\
&\qquad\qquad (\chi,\xi),(y,w) \in [0,L]\times\mathbb{R},\ (\chi,\xi)\neq(y,w), |\xi - v|, |w-v| \leq C(T)\Big\} \\
&\qquad\qquad \times \left\|\left(\widetilde{X}(t^n; t^{n+1}, x, v), \widetilde{V}(t^n; t^{n+1}, x, v)\right)\right. \\
&\qquad\qquad\quad \left. - \left(X(t^n; t^{n+1}, x, v), V(t^n; t^{n+1}, x, v)\right)\right\|_2 dvdx \\
&\leq \Delta t^2 \int_0^L \int_{\mathbb{R}} \mathcal{L}(f(t^n), C(T))(v) dvdx \\
&\leq C\Delta t^2,
\end{aligned}
$$

where

$$
\begin{aligned}
\sup\left\{\left|V(t^n; t^{n+1}, x, v) - v\right| \mid x \in [0,L], v \in \mathbb{R}\right\} &\leq \int_{t^n}^{t^{n+1}} \|E(\tau,.)\|_{L^\infty} dt \\
&\leq T\,\|E\|_{L^\infty(0,T;L^\infty)} \leq C(T) < +\infty.
\end{aligned}
$$

Then we conclude that the estimate of Lemma 5.3 has to be replaced by

$$
\left\|\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - f\left(t^{n+1}\right)\right\|_{L^{1,\infty}(Q)} \leq C\Delta t^2.
$$

Following the proof of Proposition 5.4, if we take $f \in \mathscr{C}_b(0,T; W^{1,\infty} \cap W^{1,1}(Q))$ and $E \in \mathscr{C}^1(0,T; \mathscr{C}^1_{b,per_x}(\mathbb{R}))$, then using (3.1) and taking the derivative in the sense of distribution, we still have (with $m \in \{0,1\}$, $p \in \{1,\infty\}$)

$$
||\mathcal{T}_i f||_{L^\infty(0,T;W^{m,p}(Q))} \leq C||f||_{L^\infty(0,T;W^{m,p}(Q))},
$$

$$
||\widetilde{\mathcal{T}}_i f||_{L^\infty(0,T;L^p(Q))} \leq C||f||_{L^\infty(0,T;W^{m,p}(Q))},
$$

and

$$
\|(\mathcal{T}_i - \widetilde{\mathcal{T}}_i) f\|_{L^\infty(0,T;L^p(Q))} \leq Ch||f||_{L^\infty(0,T;W^{1,p}(Q))}.
$$

As a consequence, Lemma 5.5 supplies the estimate

$$
\|\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^{1,\infty}(Q)} \leq Ch.
$$

The estimate of Lemma 5.6 has to be replaced by

(5.32)
$$\|\widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2 \circ \widetilde{\mathcal{T}}_1 f(t^n) - \widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 f(t^n)\|_{L^{1,\infty}(Q)} \leq C\Delta t \left(e^n + \frac{1}{(1+R)^\lambda} + h \right),$$

where

$$e^n = \|f(t^n) - f_h(t^n)\|_{L^{1,\infty}(Q)}.$$

The proof of Lemma 5.6 holds, except for the estimate of $E_h^{n+1/2}(x) - \widetilde{E}^{n+1/2}(x)$ that we slightly modify as follows. We rewrite

$$\begin{aligned}
&E_h^{n+1/2}(x) - \widetilde{E}^{n+1/2}(x) \\
&\quad = \int_0^L K(x,y) \left(\int_{\mathbb{R}} \left[\widetilde{\mathcal{T}}_1 f_h(t^n, y, v) - \mathcal{T}_1 f(t^n, y, v) \right] dv \right) dy \\
&\quad = \int_0^L K(x,y) \left(\int_{|v|\leq R} \pi_h \Big[f_h\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) - f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) \Big] dv \right) dy \\
&\qquad + \int_0^L \int_{|v|>R} K(x,y) f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) dv dy \\
&\qquad + \int_0^L \int_{|v|\leq R} K(x,y) \Big(\pi_h f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) - f\Big(t^n, y - v\tfrac{\Delta t}{2}, v\Big) \Big) dv dy,
\end{aligned}$$

so that we get

$$\begin{aligned}
\|E_h^{n+1/2} - \widetilde{E}^{n+1/2}\|_{L^\infty([0,L])} \leq\ & ||K||_{L^\infty} ||\pi_h||_{L^\infty} \|f_h(t^n) - f(t^n)\|_{L^{1,\infty}(Q)} \\
& + ||K||_{L^\infty} ||f(t^n)||_{L^1(Q\setminus\Omega)} \\
& + C||K||_{L^\infty} h ||f(t^n)||_{W^{1,1}(Q)}.
\end{aligned} \tag{5.33}$$

Thanks to assumption (5.27), for the second term of (5.33) we obtain

$$\left\| E_h^{n+1/2} - \widetilde{E}^{n+1/2} \right\|_{L^\infty([0,L])} \leq C \left(e^n + \frac{1}{(1+R)^\lambda} + h \right).$$

In order to finish justifying the inequality (5.32), we now just have to show that $\mathcal{L}(\widetilde{\mathcal{T}}_1 f(t^n, C(T)))(\xi)$ belongs to $L^\infty \cap L^1$. Indeed we have

$$\begin{aligned}
\mathcal{L}(\widetilde{\mathcal{T}}_1 f(t^n, C(T)))(\xi) &= \mathcal{L}\Big(\pi_h f\Big(t^n, x - v\tfrac{\Delta t}{2}, v\Big), C(T) \Big)(\xi) \\
&\leq \|\pi_h\|_{L^\infty} \mathcal{L}\left(f(t^n), C(T) \right)(\xi) \in L^\infty \cap L^1.
\end{aligned}$$

Finally, thanks to Lemma 5.12, we get the $L^{1,\infty}$ stability of the interpolation operator π_h; that is to say, there exists a constant C such that

$$||\pi_h f||_{L^{1,\infty}} \leq ||f||_{L^{1,\infty}} + \mu(h) \quad \forall f \in \mathscr{C}_b\left(0, T; \mathscr{C}_{b,per_x} \cap L^1(\mathbb{R}_x \times \mathbb{R}_v)\right).$$

Then it is obvious that the estimate of Lemma 5.7 becomes

$$\|\widetilde{\mathcal{T}}_1 \circ \widetilde{\mathcal{T}}_2^\star \circ \widetilde{\mathcal{T}}_1 (f(t^n) - f_h(t^n))\|_{L^{1,\infty}(Q)} \leq e^n + 3\mu(h).$$

As in the proof of the main theorem, a discrete Gronwall inequality enables us to get

$$e^{n+1} \leq \exp(CT)e^0 + C\left(\Delta t + h + \frac{h+\mu(h)}{\Delta t} + \frac{1}{(1+R)^\lambda}\right).$$

If we suppose that $R = \frac{1}{h^{1/\alpha}}$, $\alpha > 0$, and since e^0 is only a fixed interpolation error, we obtain

$$e^{n+1} \leq C\left(\Delta t + h + \frac{h+\mu(h)}{\Delta t} + h^{\lambda/\alpha}\right).$$

Then the end of the proof is the same as the proof of the main Theorem 5.1, and we get

$$||f - f_h||_{\ell^\infty(0,T;L^{1,\infty}(Q))} \leq C\left(\Delta t + h + \frac{h+\mu(h)}{\Delta t} + h^{\lambda/\alpha}\right)$$

and

$$||E - E_h||_{\ell^\infty(0,T;L^\infty([0,L]))} \leq C\left(\Delta t + h + \frac{h+\mu(h)}{\Delta t} + h^{\lambda/\alpha}\right).$$

If we choose $\Delta t \sim (h+\mu(h))^{1/\sigma}$, $\sigma > 1$, we get the estimates of Theorem 5.1. □

REFERENCES

[1] M. L. Begue, A. Ghizzo, and P. Bertrand, *Two-dimensional Vlasov simulation of Raman scattering and plasma beatwave acceleration on parallel computers*, J. Comput. Phys., 151 (1999), pp. 458–478.

[2] M. L. Begue, A. Ghizzo, P. Bertrand, E. Sonnendrücker, and O. Coulaud, *Two-dimensional semi-Lagrangian Vlasov simulations of laser-plasma interaction in the relativistic regime*, J. Comput. Phys., 62 (1999), pp. 367–388.

[3] R. Bermejo, *On the equivalence of semi-Lagrangian schemes and particle-in-cell finite element methods*, Monthly Weather Review, 118 (1990), pp. 979–987.

[4] R. Bermejo, *Analysis of an algorithm for the Galerkin-characteristic method*, Numer. Math., 60 (1991), pp. 163–194.

[5] R. Bermejo, *Analysis of a class of quasi-monotone and conservative semi-Lagrangian advection schemes*, Numer. Math., 87 (2001), pp. 597–623.

[6] N. Besse, *Etude mathématique et numérique de l'équation de Vlasov non linéaire sur des maillages non structurés de l'espace des phases*, Ph.D. thesis, Institut de Recherche Mathématique Avancée, Université Louis Pasteur, Strasbourg, France, 2003.

[7] N. Besse, *Convergence of Classes of High Order Semi-Lagrangian Schemes for the Vlasov–Poisson System*, manuscript, Institut de Recherche Mathématique Avancée, Université Louis Pasteur, Strasbourg, France.

[8] N. Besse, *Convergence of a High Order Semi-Lagrangian Scheme with Propagation of Gradients for the Vlasov-Poisson System*, manuscript, Institut de Recherche Mathématique Avancée, Université Louis Pasteur, Strasbourg, France.

[9] N. Besse and E. Sonnendrücker, *Semi-Lagrangian schemes for the Vlasov equation on an unstructured mesh of phase space*, J. Comput. Phys., 191 (2003), pp. 341–376.

[10] C. K. Birdsall and A. B. Langdon, *Plasma Physics via Computer Simulation*, McGraw–Hill, New York, 1985.

[11] J. P. Boris and D. L. Book, *Solution of continuity equations by the method of flux-corrected transport*, J. Comput. Phys., 20 (1976), pp. 397–431.

[12] F. Bouchut, F. Golse, and M. Pulvirenti, *Kinetic Equations and Asymptotic Theory*, P. G. Ciarlet and P.-L. Lions, eds., Ser. Appl. Math., Gauthier-Villars, Paris, 2000.

[13] C. Z. Cheng and G. Knorr, *The integration of the Vlasov equation in configuration space*, J. Comput. Phys., 22 (1976), pp. 330–351.

[14] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Finite element methods (Part 1), Vol. II, P. G. Ciarlet and J. L. Lions, eds., North–Holland, New York, 1991, pp. 17–351.

[15] J. Cooper and A. Klimas, *Boundary value problems for the Vlasov–Maxwell equation in one dimension*, J. Math. Anal. Appl., 75 (1980), pp. 306–329.
[16] G.-H. Cottet and P.-A. Raviart, *Particle methods for the one-dimensional Vlasov–Poisson equations*, SIAM J. Numer. Anal., 21 (1984), pp. 52–76.
[17] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.
[18] M. Falcone and R. Ferretti, *Convergence analysis for a class of high-order semi-Lagrangian advection schemes*, SIAM J. Numer. Anal., 35 (1998), pp. 909–940.
[19] E. Fijalkow, *A numerical solution to the Vlasov equation*, Comput. Phys. Commun., 116 (1999), pp. 319–328.
[20] F. Filbet, *Convergence of a finite volume scheme for the Vlasov–Poisson system*, SIAM J. Numer. Anal., 39 (2001), pp. 1146–1169.
[21] F. Filbet, E. Sonnendrücker, and P. Bertrand, *Conservative numerical schemes for the Vlasov equation*, J. Comput. Phys., 172 (2001), pp. 166–187.
[22] K. Ganguly and H. D. Victory, Jr., *On the convergence of particle methods for multidimensional Vlasov–Poisson systems*, SIAM J. Numer. Anal., 26 (1989), pp. 249–288.
[23] R. T. Glassey, *The Cauchy Problem in Kinetic Theory*, SIAM, Philadelphia, 1996.
[24] R. Glassey and J. Schaeffer, *Convergence of a particle method for the relativistic Vlasov–Maxwell system*, SIAM J. Numer. Anal., 28 (1991), pp. 1–25.
[25] Y. Hasbani, E. Livne, and M. Bercovier, *Finite elements and characteristics applied to advection-diffusion equations*, Comput. Fluids, 11 (1983), pp. 71–83.
[26] K. J. Havlak and H. D. Victory, Jr., *The numerical analysis of random particle methods applied to Vlasov–Poisson–Fokker–Planck kinetic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 291–317.
[27] K. J. Havlak and H. D. Victory, Jr., *On deterministic particle methods for solving Vlasov–Poisson–Fokker–Planck systems*, SIAM J. Numer. Anal., 35 (1998), pp. 1473–1519.
[28] E. Horst, *On classical solutions of the initial value problem for the unmodified non-linear Vlasov equation*, Math. Methods Appl. Sci., 3 (1981), pp. 229–248.
[29] E. Horst, *On the asymptotic growth of the solutions of the Vlasov–Poisson system*, Math. Methods Appl. Sci., 16 (1993), pp. 75–85.
[30] S. V. Iordanskii, *The Cauchy problem for the kinetic equation of plasma*, Amer. Math. Soc. Transl., 35 (1964), pp. 351–363.
[31] P. L. Lions and B. Perthame, *Propagation of moments and regularity for the* 3*-dimensional Vlasov–Poisson system*, Invent. Math., 105 (1991), pp. 415–430.
[32] P. Mineau, *Simulation en Physique des Plasmas*, Ph.D. thesis, Laboratoire Mathématiques et Applications, Physique Mathématique d'Orléans (MAPMO), Université d'Orléans, Orléans, France, 1997.
[33] H. Neunzert and J. Wick, *Die theorie de asymptotischen verteiling und die numerische losung von integrodifferentialgleichungen*, Numer. Math., 21 (1973), pp. 243–243.
[34] K. Pfaffelmoser, *Global classical solutions of the Vlasov–Poisson system in three dimensions for general initial data*, J. Differential Equations, 95 (1992), pp. 281–303.
[35] O. Pironneau, *On the transport-diffusion algorithm and its applications to the Navier–Stokes equations*, Numer. Math., 38 (1982), pp. 309–332.
[36] P.-A. Raviart, *An analysis of particle methods*, in Numerical Methods in Fluid Dynamics, Lecture Notes in Math. 1127, Springer, New York, 1985, pp. 243–324.
[37] A. Robert, *A stable numerical integration scheme for the primitive meteorological equations*, Atmos. Ocean., 19 (1981), pp. 35–46.
[38] J. Schaeffer, *Global existence of smooth solutions to the Vlasov–Poisson system in three dimensions*, Comm. Partial Differential Equations, 16 (1991), pp. 1313–1335.
[39] J. Schaeffer, *Convergence of a difference scheme for the Vlasov–Poisson–Fokker–Planck system in one dimension*, SIAM J. Numer. Anal., 35 (1998), pp. 1149–1175.
[40] E. Sonnendrücker, J. Roche, P. Bertrand, and A. Ghizzo, *The semi-Lagrangian method for the numerical resolution of Vlasov equations*, J. Comput. Phys., 149 (1996), pp. 841–872.
[41] O. Coulaud, E. Sonnendrücker, E. Dillon, P. Bertrand, and A. Ghizzo, *Parallelisation of semi-Lagrangian Vlasov codes*, J. Comput. Phys., 61 (1999), pp. 435–448.
[42] P. K. Smolarkiewicz and J. A. Pudykiewicz, *A class of semi-Lagrangian approximation for fluids*, J. Atmospheric Sci., 49 (1992), pp. 2082–2096.
[43] A. Staniforth and J. Cote, *Semi-Lagrangian integration schemes for atmospheric models—A review*, Monthly Weather Review, 119 (1991), pp. 2206–2223.
[44] S. Ukai and T. Okabe, *On classical solutions in the large in time of two-dimensional Vlasov's equation*, Osaka J. Math., 15 (1978), pp. 245–261.

[45] H. D. Victory, Jr., and E. J. Allen, *The convergence theory of particle-in-cell methods for multidimensional Vlasov–Poisson systems*, SIAM J. Numer. Anal., 28 (1991), pp. 1207–1241.

[46] H. D. Victory, Jr., G. Tucker, and K. Ganguly, *The convergence analysis of the fully discretized particle methods for solving Vlasov–Poisson systems*, SIAM J. Numer. Anal., 28 (1991), pp. 955–989.

[47] S. Wollman, *Global in time solutions of the two dimensional Vlasov–Poisson system*, Comm. Pure Appl. Math., 33 (1988), pp. 173–197.

[48] S. Wollman, *Local existence and uniqueness theory of the Vlasov–Maxwell system*, J. Math. Anal. Appl., 127 (1987), pp. 103–121.

[49] S. Wollman, *On the approximation of the Vlasov–Poisson system by particle methods*, SIAM J. Numer. Anal., 37 (2000), pp. 1369–1398.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 383–408

EXCLUSION REGIONS FOR SYSTEMS OF EQUATIONS*

HERMANN SCHICHL† AND ARNOLD NEUMAIER†

Abstract. Branch and bound methods for finding all zeros of a nonlinear system of equations in a box frequently have the difficulty that subboxes containing no solution cannot be easily eliminated if there is a nearby zero outside the box. This has the effect that near each zero many small boxes are created by repeated splitting, whose processing may dominate the total work spent on the global search.

This paper discusses the reasons for the occurrence of this so-called cluster effect and how to reduce the cluster effect by defining exclusion regions around each zero found that are guaranteed to contain no other zero and hence can safely be discarded.

Such exclusion regions are traditionally constructed using uniqueness tests based on the Krawczyk operator or the Kantorovich theorem. These results are reviewed; moreover, refinements are proved that significantly enlarge the size of the exclusion region. Existence and uniqueness tests are also given.

Key words. zeros, system of equations, validated enclosure, existence test, uniqueness test, inclusion region, exclusion region, branch and bound, cluster effect, Krawczyk operator, Kantorovich theorem, backboxing, affine invariant

AMS subject classifications. Primary, 65H20; Secondary, 65G30

DOI. 10.1137/S0036142902418898

1. Introduction.

Branch and bound methods for finding all zeros of a nonlinear system of equations in a box [10, 23] frequently have the difficulty that subboxes containing no solution cannot be easily eliminated if there is a nearby zero outside the box. This has the effect that near each zero many small boxes are created by repeated splitting, whose processing may dominate the total work spent on the global search.

This paper discusses in section 3 the reasons for the occurrence of this so-called cluster effect and how to reduce the cluster effect by defining exclusion regions around each zero found that are guaranteed to contain no other zero and hence can safely be discarded. Such exclusion boxes (possibly first used by Jansson [4]) are the basis for the backboxing strategy by van Iwaarden [24] (see also Kearfott [8, 9]) that eliminates the cluster effect near well-conditioned zeros.

Exclusion regions are traditionally constructed using uniqueness tests based on the Krawczyk operator (see, e.g., Neumaier [16, Chapter 5]) or the Kantorovich theorem (see, e.g., Ortega and Rheinboldt [19, Theorem 12.6.1]); both provide existence and uniqueness regions for zeros of systems of equations. Shen and Neumaier [22] proved that the Krawczyk operator with slopes always provides an existence region which is at least as large as that computed by Kantorovich's theorem. Deuflhard and Heindl [2] proved an affine invariant version of the Kantorovich theorem.

In section 2, these results are reviewed, together with recent works on improved preconditioning by Hansen [3] and on Taylor models by Berz and Hoefkens [1] that are related to our present work. In sections 4–7, we discuss componentwise and affine invariant existence, uniqueness, and nonexistence regions given a zero or any other

*Received by the editors December 3, 2002; accepted for publication (in revised form) May 13, 2003; published electronically March 3, 2004.
http://www.siam.org/journals/sinum/42-1/41889.html
†Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria (Hermann.Schichl@esi.ac.at, Arnold.Neumaier@univie.ac.at, http://www.mat.univie.ac.at/∼neum/).

point of the search region. They arise from a more detailed analysis of the properties of the Krawczyk operator with slopes used in [22].

Numerical examples given in section 8 show that the refinements introduced in this paper significantly enlarge the sizes of the exclusion regions.

In the following, the notation is as in the book [17]. In particular, inequalities are interpreted componentwise, I denotes the identity matrix, intervals and boxes (= interval vectors) are in boldface, and $\operatorname{rad}\mathbf{x} = \frac{1}{2}(\overline{x} - \underline{x})$ denotes the radius of a box $\mathbf{x} = [\underline{x}, \overline{x}] \in \mathbb{IR}^n$. The interior of a set $S \subseteq \mathbb{R}^n$ is denoted by $\operatorname{int}(S)$ and the interval hull by $\Box S$.

We consider the nonlinear system of equations

$$F(x) = 0, \tag{1}$$

where $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is twice continuously differentiable in a convex domain D. (For some results, weaker conditions suffice; it will be clear from the arguments used that continuity and the existence of the quantities in the hypothesis of the theorems are sufficient.)

Since F is twice continuously differentiable, we can always (e.g., using the mean value theorem) write

$$F(x) - F(z) = F[z, x](x - z) \tag{2}$$

for any two points x and z with a suitable matrix $F[z, x] \in \mathbb{R}^{n\times n}$, continuously differentiable in x and z; any such $F[z, x]$ is called a *slope matrix* for F. While (in dimension $n > 1$) $F[z, x]$ is not uniquely determined, we always have (by continuity)

$$F[z, z] = F'(z). \tag{3}$$

Thus $F[z, x]$ is a slope version of the Jacobian. There are recursive procedures to calculate a slope $F[z, x]$, given x and z; see Krawczyk and Neumaier [14], Rump [20], and Kolev [13]; a Matlab implementation is in Intlab [21].

Since the slope matrix $F[z, x]$ is continuously differentiable, we can write similarly

$$F[z, x] = F[z, z'] + \sum (x_k - z'_k) F_k[z, z', x] \tag{4}$$

with *second order slope matrices* $F_k[z, z', x]$, continuous in z, z', x. Here, as throughout this paper, the summation extends over $k = 1, \ldots, n$. Second order slope matrices can also be computed recursively; see Kolev [13]. Moreover, if F is quadratic, the slope is linear in x and z, and the coefficients of x determine constant second order slope matrices without any work.

If $z = z'$ the formula above somewhat simplifies, because of (3), to

$$F[z, x] = F'(z) + \sum (x_k - z_k) F_k[z, z, x]. \tag{5}$$

Throughout the paper we shall make the following assumption, without mentioning it explicitly.

Assumption A. The point z and the convex subset X lie in the domain of definition of F. The center, $z \in X$, and the second order slope (5) are fixed. Moreover, for a fixed preconditioning matrix $C \in \mathbb{R}^{m\times n}$, the componentwise bounds

$$\begin{aligned} \overline{b} &\geq |CF(z)| \geq \underline{b}, \\ B_0 &\geq |CF'(z) - I|, \\ B'_0 &\geq |CF'(z)|, \\ B_k(x) &\geq |CF_k[z, z, x]| \quad (k = 1, \ldots, n) \end{aligned} \tag{6}$$

are valid for all $x \in X$.

Example 1.1. We consider the system of equations

$$\begin{aligned} x_1^2 + x_2^2 &= 25, \\ x_1 x_2 &= 12. \end{aligned} \tag{7}$$

The system has the form (1) with

$$F(x) = \begin{pmatrix} x_1^2 + x_2^2 - 25 \\ x_1 x_2 - 12 \end{pmatrix}. \tag{8}$$

With respect to the center $z = \binom{3}{4}$, we have

$$F(x) - F(z) = \begin{pmatrix} x_1^2 - 3^2 + x_2^2 - 4^2 \\ x_1 x_2 - 3 \cdot 4 \end{pmatrix} = \begin{pmatrix} (x_1 + 3)(x_1 - 3) + (x_2 + 4)(x_2 - 4) \\ x_2(x_1 - 3) + 3(x_2 - 4) \end{pmatrix}$$

so that we can take

$$F[z, x] = \begin{pmatrix} x_1 + 3 & x_2 + 4 \\ x_2 & 3 \end{pmatrix}$$

as a slope. (Note that other choices would be possible.) The interval slope $F[z, \mathbf{x}]$ in the box $\mathbf{x} = [2, 4] \times [3, 5]$ is then

$$F[z, x] = \begin{pmatrix} [5, 7] & [7, 9] \\ [3, 5] & 3 \end{pmatrix}.$$

The slope can be put in form (5) with

$$F'(z) = \begin{pmatrix} 6 & 8 \\ 4 & 3 \end{pmatrix}, \quad F_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad F_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and we obtain

$$B_1 = \frac{1}{14} \begin{pmatrix} 3 & 0 \\ 4 & 0 \end{pmatrix}, \qquad B_2 = \frac{1}{14} \begin{pmatrix} 8 & 3 \\ 6 & 4 \end{pmatrix}.$$

Since we calculated without rounding errors and z happens to be a zero of F, both B_0 and $\bar{b}$ vanish.

2. Known results. The oldest semilocal existence theorem for zeros of systems of equations is due to Kantorovich [7], who obtained as a by-product of a convergence guarantee for Newton's method (which is not of interest in our context) the following result.

THEOREM 2.1 (Kantorovich). *Let z be a vector such that $F'(z)$ is invertible, and let α and β be constants with*

$$\|F'(z)^{-1}\|_\infty \le \alpha, \qquad \|F'(z)^{-1} F(z)\|_\infty \le \beta. \tag{9}$$

Suppose further that $z \in \mathbf{x}$ and that there exists a constant $\gamma > 0$ such that for all $x \in \mathbf{x}$,

$$\max_i \sum_{j,k} \left| \frac{\partial^2 F_i(x)}{\partial x_j \partial x_k} \right| \le \gamma. \tag{10}$$

If $2\alpha\beta\gamma < 1$, then $\Delta := \sqrt{1 - 2\alpha\beta\gamma}$ is real and we have the following:

1. *There is no zero $x \in \mathbf{x}$ with*

$$\underline{r} < \|x - z\|_\infty < \overline{r},$$

where

$$\underline{r} = \frac{2\beta}{1+\Delta}, \qquad \overline{r} = \frac{1+\Delta}{\alpha\gamma}.$$

2. *At most one zero x is contained in $\mathbf{x}$ with*

$$\|x - z\|_\infty < \frac{2}{\alpha\gamma}.$$

3. *If*

$$\max_{x\in\mathbf{x}} \|x - z\|_\infty < \overline{r},$$

then there is a unique zero $x \in \mathbf{x}$, and this zero satisfies

$$\|x - z\|_\infty \le \underline{r}.$$

The affine invariant version of the Kantorovich theorem given in Deuflhard and Heindl [2] essentially amounts to applying the theorem to $F'(z)^{-1}F(x)$ in place of $F(x)$. In practice, rounding errors in computing $F'(z)^{-1}$ are made, which requires the use of a preconditioning matrix $C \approx F'(z)^{-1}$ and $CF(x)$ in place of $F(x)$ to get the benefits of affine invariance in floating point computations.

Kahan [5] used the Krawczyk operator, which needs only first order slopes, to make existence statements. Together with later improvements using slopes, his result is contained in the following statement.

THEOREM 2.2 (Kahan). *Let $z \in \mathbf{z} \subseteq \mathbf{x}$. If there is a matrix $C \in \mathbb{R}^{n\times n}$ such that the Krawczyk operator*

$$K(\mathbf{z}, \mathbf{x}) := z - CF(z) - (CF[\mathbf{z}, \mathbf{x}] - I)(\mathbf{x} - z) \tag{11}$$

satisfies $K(\mathbf{z}, \mathbf{x}) \subseteq \mathbf{x}$, then $\mathbf{x}$ contains a zero of (1). *Moreover, if $K(\mathbf{x}, \mathbf{x}) \subseteq \operatorname{int}(\mathbf{x})$, then $\mathbf{x}$ contains a unique zero of* (1).

Shen and Neumaier [22] proved that the Krawczyk operator with slopes always provides existence regions which are at least as large as those computed by Kantorovich's theorem, and, since the Krawczyk operator is affine invariant, this also covers the affine invariant Kantorovich theorem.

Recent work by Hansen [3] shows that there is scope for gain in Krawczyk's method by improved preconditioning; but he gives only heuristic recipes for how to proceed. For quadratic problems, where the slope is linear in x, his recipe suggests evaluating $CF[z, x]$ term by term before substituting intervals. Indeed, by subdistributivity, we always have

$$CA_0 + \sum CA_k(\mathbf{x}_k - \mathbf{z}_k) \subseteq C\Big(A_0 + \sum A_k(\mathbf{x}_k - \mathbf{z}_k)\Big)$$

so that, for quadratic functions, Hansen's recipe is never worse than the traditional recipe. We adapt it as follows to general functions, using second order slopes; in the general case, the preconditioned slope takes the form

$$CF[z, x] = CF[z, z'] + \sum (x_k - z'_k) CF_k[z, z', x] \tag{12}$$

or, with $z = z'$, as we use it most of the time,

$$CF[z,x] = CF'(z) + \sum (x_k - z_k) CF_k[z,z,x]. \tag{13}$$

In the following, the consequences of this formulation, combined with ideas from Shen and Neumaier [22], are investigated in detail.

Recent work on Taylor models by Berz and Hoefkens [1] (see also Neumaier [18]) uses expansions to even higher than second order, although at a significantly higher cost. This may be of interest for systems suffering a lot from cancellation, where using low order methods may incur much overestimation, leading to tiny inclusion regions. Another recent paper on exclusion boxes is Kalovics [6].

3. The cluster effect. As explained by Kearfott and Du [11], many branch and bound methods used for global optimization suffer from the so-called *cluster effect*. As is apparent from the discussion below, this effect is also present for branch and bound methods using constraint propagation methods to find and verify *all* solutions of nonlinear systems of equations. (See, e.g., Van Hentenryck, Michel, and Deville [23] for constraint propagation methods.)

The cluster effect consists of excessive splitting of boxes close to a solution and failure to remove many boxes not containing the solution. As a consequence, these methods slow down considerably once they reach regions close to the solutions. The mathematical reason for the cluster effect and how to avoid it will be investigated in this section.

Let us assume that for arbitrary boxes $\mathbf{x}$ of maximal width ε the computed expression $F(\mathbf{x})$ overestimates the range of F over $\mathbf{x}$ by $O(\varepsilon^k)$:

$$F(\mathbf{x}) \in (1 + C\varepsilon^k) \square \{F(x) \mid x \in \mathbf{x}\} \tag{14}$$

for $k \leq 2$ and ε sufficiently small. The exponent k depends on the method used for the computation of $F(\mathbf{x})$.

Let x^* be a regular solution of (1) (so that $F'(x^*)$ is nonsingular), and assume (14). Then any box of diameter ε that contains a point x with

$$\|F'(x^*)(x - x^*)\|_\infty \leq \Delta = C\varepsilon^k \tag{15}$$

might contain a solution. Therefore, independent of the pruning scheme used in a branch and bound method, no box of diameter ε can be eliminated. The inequality (15) describes a parallelepiped of volume

$$V = \frac{\Delta^n}{\det F'(x^*)}.$$

Thus, any covering of this region by boxes of diameter ε contains at least V/ε^n boxes.

The number of boxes of diameter ε which cannot be eliminated is therefore proportional to at least

$$\frac{C^n}{\det F'(x^*)} \quad \text{if } k = 1,$$

$$\frac{(C\varepsilon)^n}{\det F'(x^*)} \quad \text{if } k = 2.$$

For $k = 1$ this number grows exponentially with the dimension, with a growth rate determined by the relative overestimation C and a proportionality factor related to the condition of the Jacobian.

In contrast, for $k = 2$ the number is guaranteed to be small for sufficiently small ε. The size of ε, the diameter of the boxes most efficient for covering the solution, is essentially determined by the nth root of the determinant, which, for a well-scaled problem, reflects the condition of the zero. However, for ill-conditioned zeros (with a tiny determinant in naturally scaled coordinates), one already needs quite narrow boxes before the cluster effect subsides.

So, to avoid the cluster effect, we need at least the quadratic approximation property $k = 2$. Hence, Jacobian information is essential, as well as techniques to discover the shape of the uncertainty region.

A comparison of the typical techniques used for box elimination shows that constraint propagation techniques lead to overestimation of order $k = 1$; hence they suffer from the cluster effect. Centered forms using first order information (Jacobians) as in Krawczyk's method provide estimates with $k = 2$ and are therefore sufficient to avoid the cluster effect, except near ill-conditioned or singular zeros. Second order information as used, e.g., in the theorem of Kantorovich still provides only $k = 2$ in estimate (15); the cluster effect is avoided under the same conditions.

For singular (and hence for sufficiently ill-conditioned) zeros, the argument does not apply, and no technique is known to remove the cluster effect in this case. A heuristic that limits the work in this case by retaining a single but *larger* box around an ill-conditioned approximate zero is described in Algorithm 7 (Step 4(c)) of Kearfott [10].

4. Componentwise exclusion regions close to a zero. Suppose that x^* is a solution of the nonlinear system of equations (1). We want to find an *exclusion region* around x^* with the property that in the interior of this region x^* is the only solution of (1). Such an exclusion region need not be further explored in a branch and bound method for finding all solutions of (1); hence we get the name.

In this section we take an approximate zero z of F, and we choose C to be an approximation of $F'(z)^{-1}$. Suitable candidates for z can easily be found within a branch and bound algorithm by trying Newton steps from the midpoint of each box, iterating while x^ℓ remains in a somewhat enlarged box and either $\|x^{\ell+1} - x^\ell\|$ or $\|F(x^\ell)\|$ decreases by a factor of, say, 1.5 below the best previous value in the iteration. This works locally well even at nearly singular zeros and gives a convenient stop in case no nearby solution exists.

PROPOSITION 4.1. *For every solution $x \in X$ of* (1)*, the deviation*

$$s := |x - z|$$

satisfies

$$0 \le s \le \Big(B_0 + \sum s_k B_k(x)\Big)s + \bar{b}. \tag{16}$$

Proof. By (2) we have $F[z, x](x - z) = F(x) - F(z) = -F(z)$, because x is a zero. Hence, using (5), we compute

$$\begin{aligned} -(x - z) &= -(x - z) + C(F[z, x](x - z) + F(z) + F'(z)(x - z) - F'(z)(x - z)) \\ &= C(F[z, x] - F'(z))(x - z) + (CF'(z) - I)(x - z) + CF(z) \\ &= \Big(CF'(z) - I + \sum (x_k - z_k)CF_k[z, z, x]\Big)(x - z) + CF(z). \end{aligned}$$

Now we take absolute values, use (6), and get

$$\begin{aligned} s = |x - z| &\le \Big(|CF'(z) - I| + \sum |x_k - z_k|\,|CF_k[z, z, x]|\Big)\,|x - z| + |CF(z)| \\ &\le \Big(B_0 + \sum s_k B_k(x)\Big)\,s + \bar{b}. \qquad \square \end{aligned}$$

Using this result we can give a first criterion for existence regions.

THEOREM 4.2. *Let $0 < u \in \mathbb{R}^n$ be such that*

$$\Big(B_0 + \sum u_k \overline{B}_k\Big)u + \bar{b} \le u \tag{17}$$

with $B_k(x) \le \overline{B}_k$ for all $x \in M_u$, where

$$M_u := \{x \mid |x - z| \le u\} \subseteq X. \tag{18}$$

Then (1) *has a solution $x \in M_u$.*

Proof. For arbitrary x in the domain of definition of F we define

$$K(x) := x - CF(x).$$

Now take any $x \in M_u$. We get

$$\begin{aligned} K(x) = x - CF(x) &= z - CF(z) - (CF[z, x] - I)(x - z) \\ &= z - CF(z) - \Big(C\Big(F'(z) + \sum F_k[z, z, x](x_k - z_k)\Big) - I\Big)\,(x - z); \end{aligned}$$

hence

$$K(x) = z - CF(z) - \Big(CF'(z) - I + \sum CF_k[z, z, x](x_k - z_k)\Big)\,(x - z). \tag{19}$$

Taking absolute values we find

$$\begin{aligned} |K(x) - z| &= \Big|-CF(z) - \Big(CF'(z) - I + \sum CF_k[z, z, x](x_k - z_k)\Big)\,(x - z)\Big| \\ &\le |CF(z)| + \Big(|CF'(z) - I| + \sum |CF_k[z, z, x]|\,|x_k - z_k|\Big)\,|x - z| \qquad (20) \\ &\le \bar{b} + \Big(B_0 + \sum u_k \overline{B}_k\Big)\,u. \end{aligned}$$

Now assume (17). Then (20) gives

$$|K(x) - z| \le u,$$

which implies by Theorem 2.2 that there exists a solution of (1) which lies in M_u. $\square$

Note that (17) implies $B_0 u \le u$; thus the spectral radius $\rho(B_0) \le 1$. In the applications, we can make both B_0 and $\bar{b}$ very small by choosing z as an approximate zero and C as an approximate inverse of $F'(z)$.

Now the only thing that remains is the construction of a suitable vector u for Theorem 4.2.

THEOREM 4.3. *Let $S \subseteq X$ be any set containing z, and take*

$$\overline{B}_k \ge B_k(x) \quad \textit{for all } x \in S. \tag{21}$$

For $0 < v \in \mathbb{R}^n$, set

$$w := (I - B_0)v, \qquad a := \sum v_k \overline{B}_k v. \tag{22}$$

We suppose that

$$D_j = w_j^2 - 4a_j\overline{b}_j > 0 \tag{23}$$

for all $j = 1, \ldots, n$, and we define

$$\lambda_j^e := \frac{w_j + \sqrt{D_j}}{2a_j}, \qquad \lambda_j^i := \frac{\overline{b}_j}{a_j \lambda_j^e}, \tag{24}$$

$$\lambda^e := \min_{j=1,\ldots,n} \lambda_j^e, \qquad \lambda^i := \max_{j=1,\ldots,n} \lambda_j^i. \tag{25}$$

If $\lambda^e > \lambda^i$, then there is at least one zero x^* of (1) in the (inclusion) region

$$R^i := [z - \lambda^i v, z + \lambda^i v] \cap S. \tag{26}$$

The zeros in this region are the only zeros of F in the interior *of the (exclusion) region*

$$R^e := [z - \lambda^e v, z + \lambda^e v] \cap S. \tag{27}$$

Proof. Let $0 < v \in \mathbb{R}^n$ be arbitrary, and set $u = \lambda v$. We check for which λ the vector u satisfies property (17) of Theorem 4.2. The requirement

$$\begin{aligned} \lambda v &\geq \Big(B_0 + \sum u_k \overline{B}_k\Big)u + \overline{b} = \Big(B_0 + \sum \lambda v_k \overline{B}_k\Big)\lambda v + \overline{b} \\ &= \overline{b} + \lambda B_0 v + \lambda^2 \sum v_k \overline{B}_k v = \overline{b} + \lambda(v - w) + \lambda^2 a \end{aligned}$$

leads to the sufficient condition $\lambda^2 a - \lambda w + \overline{b} \leq 0$. The jth component of this inequality requires that λ lies between the solutions of the quadratic equation $\lambda^2 a_j - \lambda w_j + \overline{b}_j = 0$, which are λ_j^i and λ_j^e. Hence, for every $\lambda \in [\lambda^i, \lambda^e]$ (this interval is nonempty by assumption), the vector u satisfies (17).

Now assume that x is a solution of (1) in $\operatorname{int}(R^e) \setminus R^i$. Let λ be minimal with $|x - z| \leq \lambda v$. By construction, $\lambda^i < \lambda < \lambda^e$. By the properties of the Krawczyk operator, we know that $x = K(z, x)$; hence

$$\begin{aligned} |x - z| &\leq |CF(z)| + \Big(|CF'(z) - I| + \sum |CF_k[z, z, x]|\,|x_k - z_k|\Big)|x - z| \\ &\leq \overline{b} + \lambda B_0 v + \lambda^2 \sum v_k \overline{B}_k v < \lambda v, \end{aligned} \tag{28}$$

since $\lambda > \lambda^i$. But this contradicts the minimality of λ. So there are indeed no solutions of (1) in $\operatorname{int}(R^e) \setminus R^i$. ∎

This is a componentwise analogue of the Kantorovich theorem. We show in Example 8.1 that it is best possible in some cases.

We observe that the inclusion region from Theorem 4.3 can usually be further improved by noting that $x^* = K(z, x^*)$ and (19) imply

$$\begin{aligned} x^* &\in K(z, \mathbf{x}^i) \\ &= z - CF(z) - \Big(CF'(z) - I + \sum CF_k[z, z, \mathbf{x}^i](\mathbf{x}_k^i - z_k)\Big)\,(\mathbf{x}^i - z) \\ &\subset \operatorname{int}(\mathbf{x}^i). \end{aligned}$$

An important special case is when $F(x)$ is quadratic in x. For such a function $F[z,x]$ is linear in x, and therefore all $F_k[z,z,x]$ are constant in x. This, in turn, means that $B_k(x) = B_k$ is constant as well. So we can set $\overline{B}_k = B_k$, and the estimate (21) becomes valid everywhere.

COROLLARY 4.4. *Let F be a quadratic function. For arbitrary $0 < v \in \mathbb{R}^n$ define*

$$w := (I - B_0)v, \qquad a := \sum v_k B_k v. \tag{29}$$

We suppose that

$$D_j = w_j^2 - 4a_j\overline{b}_j > 0 \tag{30}$$

for all $j = 1, \ldots, n$, and we set

$$\lambda_j^e := \frac{w_j + \sqrt{D_j}}{2a_j}, \qquad \lambda_j^i := \frac{\overline{b}_j}{a_j \lambda_j^e}, \tag{31}$$

$$\lambda^e := \min_{j=1,\ldots,n} \lambda_j^e, \qquad \lambda^i := \max_{j=1,\ldots,n} \lambda_j^i. \tag{32}$$

If $\lambda^e > \lambda^i$, then there is at least one zero x^* of (1) in the (inclusion) box

$$\mathbf{x}^i := [z - \lambda^i v, z + \lambda^i v]. \tag{33}$$

The zeros in this region are the only zeros of F in the interior *of the (exclusion) box*

$$\mathbf{x}^e := [z - \lambda^e v, z + \lambda^e v]. \tag{34}$$

The examples later will show that the choice of v greatly influences the quality of the inclusion and exclusion regions. The main difficulty for choosing v is the positivity requirement for every D_j. In principle, a vector v, if it exists, could be found by local optimization. A method worth trying could be to choose v as a local optimizer of the problem

$$\begin{aligned} &\max n \log \lambda^e + \sum_{j=1}^{n} \log v_j \\ &\text{s.t. } D_j \geq \eta \quad (j = 1, \ldots, n), \end{aligned}$$

where η is the smallest positive machine number. This maximizes locally the volume of the excluded box. However, since λ^e is nonsmooth, solving this needs a nonsmooth optimizer (such as SolvOpt [15]).

The $\overline{B}_k$ can be constructed using interval arithmetic for a given reference box $\mathbf{x}$ around z. Alternatively, they could be calculated once in a bigger reference box $\mathbf{x}_{\text{ref}}$ and later reused on all subboxes of $\mathbf{x}_{\text{ref}}$. Saving the $\overline{B}_k$ (which needs the storage of n^3 numbers per zero) provides a simple exclusion test for other boxes. This takes $O(n^3)$ operations, while recomputing the $\overline{B}_k$ costs $O(n^4)$ operations.

5. Exclusion polytopes. Instead of boxes, we can use more general polytopes to describe exclusion and inclusion regions. With the notation as in the introduction, we assume the upper bounds

$$\overline{B}_k \geq |B_k(x)| \qquad \text{for all } x \in X. \tag{35}$$

THEOREM 5.1. *For $0 \le v \le w \in \mathbb{R}^n$, define*

$$P(w) = (\overline{B}_1{}^T w, \ldots, \overline{B}_n{}^T w) \in \mathbb{R}^{n\times n}, \tag{36}$$

$$\Pi^i = \{x \in \mathbb{R}^n \mid (w-v)^T|x-z| \le \bar{b}^T w\}. \tag{37}$$

Then any zero $x \in X$ of (1) contained in the polytope

$$\Pi^e = \{x \in \mathbb{R}^n \mid P(w)|x-z| + B_0^T w \le v\} \tag{38}$$

lies already in Π^i.

Proof. Suppose $x \in \Pi^e$ satisfies $F(x) = 0$. By Proposition 4.1, $s = |x-z|$ satisfies

$$\begin{aligned} s^T w &\le s^T\Big(B_0^T w + \sum s_k \overline{B}_k{}^T w\Big) + \bar{b}^T w \\ &= s^T(B_0^T w + P(w)s) + \bar{b}^T w \\ &\le s^T v + \bar{b}^T w. \end{aligned}$$

Hence $s^T(w-v) \le \bar{b}^T w$, giving

$$(w-v)^T|x-z| \le \bar{b}^T w; \tag{39}$$

hence $x \in \Pi^i$. ∎

COROLLARY 5.2. *Let $\mathbf{x} \subseteq X$ be a box and $z \in \mathbf{x}$ be an approximate zero. If there is a vector $0 \le w \in \mathbb{R}^n$ with*

$$v := P(w)u + B_0^T w \le w, \tag{40}$$

where $u := |\mathbf{x} - z|$, then all solutions $x \in \mathbf{x}$ of (1) satisfy (39), and, in particular,

$$|x-z|_i \le \bar{b}^T w\,(w_i - v_i)^{-1} \quad \text{for all } i \text{ with } w_i > v_i. \tag{41}$$

Proof. Let $x \in \mathbf{x}$ be a solution of (1). Then $x \in \Pi^e$ by (40), and, due to Theorem 5.1, $\mathbf{x} \in \Pi^i$. Therefore (39) holds. In particular, $(w-v)_i|x-z|_i \le \bar{b}^T w$. This implies the result. ∎

In contrast to (32), the test (40) needs only $O(n^2)$ operations (once $P(w)$ is computed) and the storage of n^2+n numbers per zero. Since $P(w)$ can be calculated columnwise, it is not even necessary to keep all $\overline{B}_k$ in store.

Since B_0 and $\bar{b}$ usually are very tiny (they contain only roundoff errors), this is a powerful box reduction technique if we can find a suitable vector w.

The result is most useful, of course, if $w > v$, but in some cases this is not possible. In these cases boxes are at least reduced in some components.

A suitable choice for w may be an approximation $w > 0$ to a Perron eigenvector [16, section 3.2] of the nonnegative matrix

$$M = \sum_k u_k \overline{B}_k{}^T,$$

where $u > 0$ is proportional to the width of the box of interest. Then

$$\lambda w = Mw = \sum u_k \overline{B}_k{}^T w = P(w)u.$$

If

$$\frac{\max(B_0^T w)_i}{w_i} < \alpha < 1, \qquad \mu := (1-\alpha)\lambda^{-1},$$

we can conclude from Corollary 5.2 (with μu in place of u) that the box $[z-\mu u, z+\mu u]$ can be reduced to $[z-\hat{u}, z+\hat{u}]$, where (with $c/0 = \infty$)

$$\hat{u}_i := \min\left(\mu u_i, \frac{\overline{b}^T w}{\max(0, \alpha w_i - (B_0^T w))_i}\right).$$

6. Uniqueness regions. Regions in which there is a unique zero can be found most efficiently as follows. First, one verifies as in the previous sections an exclusion box $\mathbf{x}^e$ which contains no zero except in a much smaller inclusion box $\mathbf{x}^i$. The inclusion box can usually be refined further by some iterations with Krawczyk's method, which generally converges quickly if the initial inclusion box is already verified. Thus we may assume that $\mathbf{x}^i$ is really tiny, with width determined by rounding errors only.

Clearly, $\mathrm{int}(\mathbf{x}^e)$ contains a unique zero iff $\mathbf{x}^i$ contains at most one zero. Thus it suffices to have a condition under which a tiny box contains at most one zero. This can be done even in fairly ill-conditioned cases by the following test.

THEOREM 6.1. *Take an approximate solution $z \in X$ of (1), and let B be a matrix such that*

$$|CF[z,\mathbf{x}] - I| + \sum |\mathbf{x}_k - z_k|\,|CF_k[\mathbf{x}, z, \mathbf{x}]| \le B. \tag{42}$$

If $\|B\| < 1$ for some monotone norm, then $\mathbf{x}$ contains at most one solution x of (1).

Proof. Assume that x and x' are two solutions. Then we have

$$0 = F(x') - F(x) = F[x,x'](x'-x) = \Big(F[x,z] + \sum (x'_k - z_k)F_k[x,z,x']\Big)(x'-x). \tag{43}$$

Using an approximate inverse C of $F'(z)$ we further get

$$x - x' = \Big((CF[z,x] - I) + \sum (x'_k - z_k)CF_k[x,z,x']\Big)(x'-x). \tag{44}$$

Applying absolute values, and using (42), we find

$$|x'-x| \le \Big(|CF[z,x]-I| + \sum |CF_k[x,z,x']|\,|x'_k - z_k|\Big)|x'-x| \le B|x'-x|. \tag{45}$$

This, in turn, implies $\|x'-x\| \le \|B\|\,\|x'-x\|$. If $\|B\| < 1$ we immediately conclude $\|x'-x\| \le 0$; hence $x = x'$. ∎

Since B is nonnegative, $\|B\| < 1$ holds for some norm iff the spectral radius of B is less than one (see, e.g., Neumaier [16, Corollary 3.2.3]); a necessary condition for this is that $\max B_{kk} < 1$, and a sufficient condition is that $|B|u < u$ for some vector $u > 0$.

So one first checks whether $\max B_{kk} < 1$. If this holds, one checks whether $\|B\|_\infty < 1$; if this fails, one computes an approximate solution u of $(I-B)u = e$, where e is the all-one vector, and checks whether $u > 0$ and $|B|u < u$. If this fails, the spectral radius of B is very close to 1 or larger. (Essentially, this amounts to testing $I - B$ for being an H-matrix; cf. [16, Proposition 3.2.3].)

We can find a matrix B satisfying (42) by computing $\hat{B}_k \geq |CF_k[\mathbf{x}, z, \mathbf{x}]|$, for example by interval evaluation, using (5), and observing

$$\begin{aligned}|CF[z,\mathbf{x}] - I| &\leq |CF'(z) - I| + \sum |\mathbf{x}_k - z_k|\,|CF_k[z,z,\mathbf{x}]| \\ &\leq |CF'(z) - I| + \sum |\mathbf{x}_k - z_k|\,|CF_k[\mathbf{x},z,\mathbf{x}]|.\end{aligned}$$

Then, using (6), we get

$$|CF[z,\mathbf{x}] - I| + \sum |\mathbf{x}_k - z_k|\,|CF_k[\mathbf{x},z,\mathbf{x}]| \leq B_0 + 2\sum |\mathbf{x}_k - z_k|\,\hat{B}_k =: B, \tag{46}$$

where B can be computed using rounding towards $+\infty$.

If F is quadratic, the results simplify again. In this case all $F_k[x', z, x] =: F_k$ are constant, and we can replace $\hat{B}_k$ by $B_k := |CF_k|$. Hence (46) becomes

$$B = B_0 + 2\sum |\mathbf{x}_k - z_k| B_k.$$

7. Componentwise exclusion regions around arbitrary points. In a branch–and–bound-based method for finding all solutions to (1), we not only need to exclude regions close to zeros but also boxes far away from all solutions. This is usually done by interval analysis on the range of F, by constraint propagation methods (see, e.g., Van Hentenryck, Michel, and Deville [23]), or by Krawczyk's method or preconditioned Gauss–Seidel iteration (see, e.g., [16]). An affine invariant, componentwise version of the latter is presented in this section.

Let z be an arbitrary point in the region of definition of F. Throughout this section, $C \in \mathbb{R}^{m\times n}$ denotes an arbitrary rectangular matrix. M_u is as in (18).

THEOREM 7.1. *Let $0 < u \in \mathbb{R}^n$, and take $\overline{B}_k \geq B_k(x)$ for all $x \in M_u$. If there is an index $i \in \{1,\dots,n\}$ such that the inequality*

$$\underline{b}_i - (B_0'u)_i - \sum u_k(\overline{B}_k u)_i > 0 \tag{47}$$

is valid, then (1) *has no solution $x \in M_u$.*

Proof. We set $\mathbf{x} = [z-u, z+u]$. For a zero $x \in M_u$ of F, we calculate, using (5), similar to the proof of Theorem 4.2,

(48)
$$\begin{aligned}0 = |K(x) - x| &= \Big| - CF(z) - \Big(CF'(z) - \sum CF_k[z,z,x](x_k - z_k)\Big)(x-z)\Big| \\ &\geq |CF(z)| - \Big|(CF'(z) - I)(x-z) + \sum (x_k - z_k) CF_k[z,z,x](x-z)\Big|.\end{aligned}$$

Now we use (6) and (47) to compute

$$\begin{aligned}|CF(z)|_i \geq \underline{b}_i &> (B_0'u)_i + \sum (u_k \overline{B}_k u)_i \\ &\geq \Big(|CF'(z)|u\Big)_i + \sum \Big(u_k |CF_k[z,z,x]\,|u\Big)_i \\ &\geq \Big|CF'(z)(x-z)\Big|_i + \sum \Big|(x_k - z_k)CF_k[z,z,x](x-z)\Big|_i \\ &\geq \Big|(CF'(z) - I)(x-z) + \sum (x_k - z_k) CF_k[z,z,x](x-z)\Big|_i .\end{aligned}$$

This calculation and (47) imply

$$\begin{aligned}|CF(z)|_i - \Big|CF'(z)(x-z) + \sum (x_k - z_k) CF_k[z,z,x](x-z)\Big|_i \\ \geq \underline{b}_i - (B_0'u)_i - \sum (u_k \overline{B}_k u)_i > 0,\end{aligned}$$

contradicting (48). ∎

Again, we need a method to find good vectors u satisfying (47). The following theorem provides that.

THEOREM 7.2. *Let $S \subseteq X$ be a set containing z, and take $\overline{B}_k \geq B_k(x)$ for all $x \in S$. If for any $0 < v \in \mathbb{R}^n$ we define*

$$\begin{aligned} w^\times &:= B_0' v, \\ a^\times &:= \sum v_k \overline{B}_k v, \\ D_i^\times &:= {w_i^\times}^2 + 4\underline{b}_i a_i^\times, \\ \lambda_i^\times &:= \frac{\underline{b}_i}{w_i^\times + \sqrt{D_i^\times}}, \\ \lambda^\times &:= \max_{i=1,\ldots,n} \lambda_i^\times, \end{aligned} \tag{49}$$

then F has no zero in the interior of the exclusion region

$$R^\times := [z - \lambda^\times v, z + \lambda^\times v] \cap S. \tag{50}$$

Proof. We set $u = \lambda v$ and check the result (47) of Theorem 7.1:

$$0 < \underline{b}_i - (B_0' u)_i - \sum (u_k \overline{B}_k u)_i = \underline{b}_i - \lambda (B_0' v)_i - \lambda^2 \sum (v_k \overline{B}_k v)_i.$$

This quadratic inequality has to be satisfied for some $i \in \{1, \ldots, n\}$. The ith inequality is true for all $\lambda \in [0, \lambda_i^\times[$, so we can take the maximum of all these numbers and still have the inequality satisfied for at least one i. Bearing in mind that the estimates are only true in the set S, the result follows from Theorem 7.1. ∎

As in the last section, a vector v could be calculated by local optimization, e.g., as a local optimizer of the problem

$$\max n \log \lambda^\times + \sum_{j=1}^{n} \log v_j.$$

This maximizes locally the volume of the excluded box. Solving this also needs a nonsmooth optimizer since $\lambda^\times$ is nonsmooth like λ^e. However, in contrast to the v needed in Theorem 4.3, there is no positivity requirement which has to be satisfied. In principle, every choice of v leads to some exclusion region.

Finding a good choice for C is a subtle problem and could be attacked by methods similar to Kearfott, Hu, and Novoa [12]. Example 8.3 below shows that a pseudoinverse of $F'(z)$ usually yields reasonable results. However, improving the choice of C sometimes widens the exclusion box by a considerable amount.

Again, for quadratic F the result can be made global, due to the fact that the $F_k[z, z, x]$ are independent of x.

COROLLARY 7.3. *Let F be quadratic, and $0 < v \in \mathbb{R}^n$. Choose $\overline{B}_k \geq |CF_k|$, $w_i^\times$, $a_i^\times$, $D_i^\times$, $\lambda_i^\times$, and $\lambda^\times$ as in Theorem 7.2. Then F has no zero in the interior of the exclusion box*

$$\mathbf{x}^\times := [z - \lambda^\times v, z + \lambda^\times]. \tag{51}$$

Proof. This is a direct consequence of Theorem 7.2 and the fact that all $F_k[z, z, x]$ are constant in x. ∎

Results analogous to Theorems 4.3, 5.1, 6.1, and 7.2 can be obtained for exclusion regions in global optimization problems by applying the above techniques to the first order optimality conditions. Since nothing new happens mathematically, we refrain from giving details.

8. Examples. We illustrate the theory with a few examples.

Example 8.1. We continue Example 1.1, doing all calculations symbolically, hence free of rounding errors, assuming a known zero. (This idealizes the practically relevant case where a good approximation of a zero is available from a standard zero-finder.)

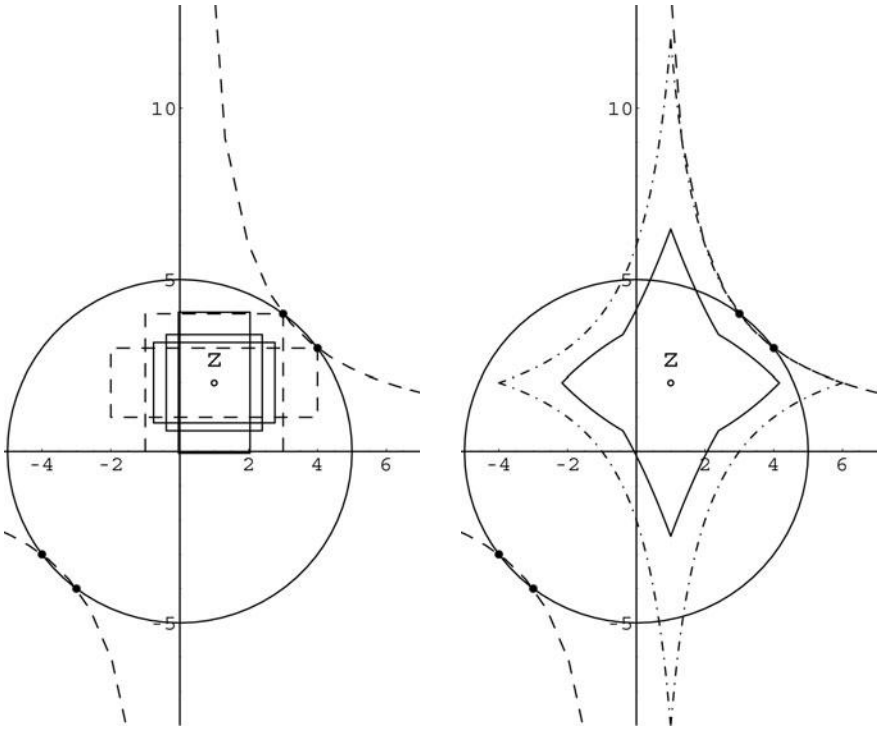


FIG. 1. *Maximal exclusion boxes around* $\binom{1}{2}$ *and the total excluded region for Example* 8.1.

We consider the system of equations (7), which has the four solutions $\pm\binom{3}{4}$ and $\pm\binom{4}{3}$; cf. Figure 1. The system has the form (1) with F given by (8). If we take the solution $x^* = \binom{3}{4}$ as center z, we can use the slope calculations from the introduction. From (29) we get

$$w_j = v_j, \qquad D_j = v_j^2 \quad (j = 1, 2),$$

$$a_1 = \tfrac{1}{14}(3v_1^2 + 8v_1v_2 + 3v_2^2), \qquad a_2 = \tfrac{1}{14}(4v_1^2 + 6v_1v_2 + 4v_2^2),$$

and, for the particular choice $v = \binom{1}{1}$, we get from (31)

$$\lambda^i = 0, \qquad \lambda^e = 1. \tag{52}$$

Thus, Corollary 4.4 implies that the interior of the box

$$[x^* - v, x^* + v] = \begin{pmatrix} [2,4] \\ [3,5] \end{pmatrix}$$

contains no solution apart from $\binom{3}{4}$. This is best possible, since there is another solution $\binom{4}{3}$ at a vertex of this box. The choice $v = \binom{1}{2}$, $\omega(v) = \frac{8}{7}$, gives another exclusion box, neither contained in nor containing the other box.

If we consider the point $z = \binom{1}{2}$, we find

$$F(z) = \begin{pmatrix} -20 \\ -10 \end{pmatrix}, \quad F'(z) = \begin{pmatrix} 2 & 4 \\ 2 & 1 \end{pmatrix}, \quad C = \frac{1}{6}\begin{pmatrix} -1 & 4 \\ 2 & -2 \end{pmatrix},$$

$$\underline{b} = \frac{10}{3}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad B_0 = 0, \quad B_1 = \frac{1}{6}\begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}, \quad B_2 = \frac{1}{6}\begin{pmatrix} 4 & 1 \\ 2 & 2 \end{pmatrix},$$

$$w^\times = v, \qquad a^\times = \frac{1}{6}\begin{pmatrix} v_1^2 + 4v_1v_2 + v_2^2 \\ 2v_1^2 + 2v_1v_2 + 2v_2^2 \end{pmatrix},$$

$$D_1^\times = \frac{1}{9}(29v_1^2 + 80v_1v_2 + 20v_2^2), \qquad D_2^\times = \frac{1}{9}(40v_1^2 + 40v_1v_2 + 49v_2^2).$$

Since everything is affine invariant and $v > 0$, we can set $v = (1, v_2)$, and we compute

$$\lambda^\times = \begin{cases} \frac{20}{3v_2 + \sqrt{40 + 40v_2 + 49v_2^2}} & \text{if } v_2 \le 1, \\ \frac{30}{3 + \sqrt{29 + 80v_2 + 20v_2^2}} & \text{if } v_2 > 1. \end{cases}$$

Depending on the choice of v_2, the volume of the exclusion box varies. There are three locally best choices $v_2 \approx 1.97228$, $v_2 \approx 0.661045$, and $v_2 = 1$, the first providing the globally maximal exclusion box.

For any two different choices of v_2 the resulting boxes are never contained in one another. Selected maximal boxes are depicted in Figure 1 (left) in solid lines; the total region which can be excluded by Corollary 7.3 is shown in solid lines in the right part of the figure.

The optimal preconditioner for exclusion boxes, however, does not need to be an approximate inverse to $F'(z)$. In this case, it turns out that $C = (0\ 1)$ is optimal for every choice of v. Two clearly optimal boxes and the total excluded region for every possible choice of v with $C = (0\ 1)$ can be found in Figure 1 in dashed lines.

Example 8.2. The system of equations (1) with

$$F(x) = \begin{pmatrix} x_1^2 + x_1x_2 + 2x_2^2 - x_1 - x_2 - 2 \\ 2x_1^2 + x_1x_2 + 3x_2^2 - x_1 - x_2 - 4 \end{pmatrix} \tag{53}$$

has the solutions $\binom{1}{1}$, $\binom{1}{-1}$, $\binom{-1}{1}$; cf. Figure 2. It is easily checked that

$$F[z, x] = \begin{pmatrix} x_1 + x_2 + z_1 - 1 & 2x_2 + z_1 + 2z_2 - 1 \\ 2x_1 + x_2 + 2z_1 - 1 & 3x_2 + z_1 + 3z_2 - 1 \end{pmatrix}$$

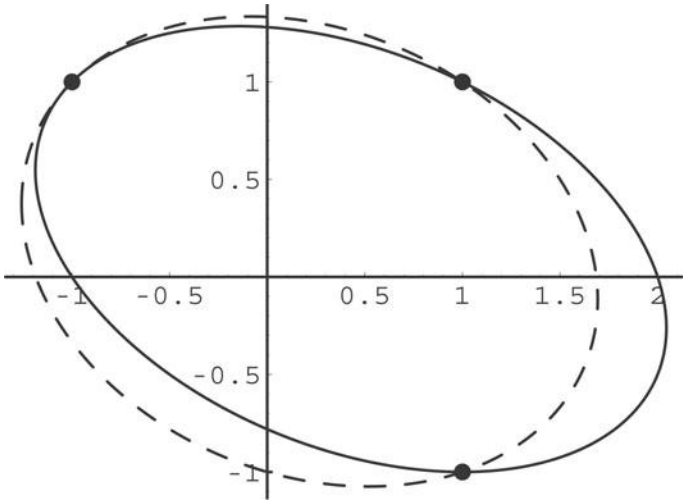


FIG. 2. *Two quadratic equations in two variables; Example* 8.2.

satisfies (2). Thus (5) holds with

$$F'(z) = \begin{pmatrix} 2z_1 + z_2 - 1 & z_1 + 4z_2 - 1 \\ 4z_1 + z_2 - 1 & z_1 + 6z_2 - 1 \end{pmatrix}, \qquad F_1 = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}, \qquad F_2 = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}.$$

We consider boxes centered at the solution $z = x^* = \binom{1}{1}$. For

$$\mathbf{x} = [x^* - \varepsilon u, x^* + \varepsilon u] = \begin{pmatrix} [1-\varepsilon, 1+\varepsilon] \\ [1-\varepsilon, 1+\varepsilon] \end{pmatrix},$$

we find

$$F'[x^*, \mathbf{x}] = \begin{pmatrix} [2-2\varepsilon, 2+2\varepsilon] & [4-2\varepsilon, 4+2\varepsilon] \\ [4-3\varepsilon, 4+3\varepsilon] & [6-3\varepsilon, 6+3\varepsilon] \end{pmatrix},$$

$$F'(\mathbf{x}) = \begin{pmatrix} [2-3\varepsilon, 2+3\varepsilon] & [4-5\varepsilon, 4+5\varepsilon] \\ [4-5\varepsilon, 4+5\varepsilon] & [6-7\varepsilon, 6+7\varepsilon] \end{pmatrix}.$$

The midpoint of $F'(\mathbf{x})$ is here $F'(z)$, and the optimal preconditioner is

$$C := F'(x^*)^{-1} = \begin{pmatrix} -1.5 & 1 \\ 1 & -0.5 \end{pmatrix};$$

from this, we obtain

$$B_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0.5 \end{pmatrix}.$$

The standard uniqueness test checks for a given box $\mathbf{x}$ whether the matrix $F'(\mathbf{x})$ is strongly regular (Neumaier [16]). But given the zero x^* (or, in finite precision

calculations, a tiny enclosure for it), it suffices to show strong regularity of $F[x^*, \mathbf{x}]$. We find

$$|I - CF'(\mathbf{x})| = \frac{\varepsilon}{2}\begin{pmatrix} 19 & 29 \\ 11 & 17 \end{pmatrix},$$

with spectral radius $\varepsilon(9 + 4\sqrt{5}) \approx 17.944\varepsilon$. Thus $F'(\mathbf{x})$ is strongly regular for $\varepsilon < 1/17.944 = 0.0557$. The exclusion box constructed from slopes is better, since

$$|I - CF[x^*, \mathbf{x}]| = \varepsilon\begin{pmatrix} 6 & 6 \\ 3.5 & 3.5 \end{pmatrix}$$

has spectral radius 9.5ε. Thus $F[x^*, \mathbf{x}]$ is strongly regular for $\varepsilon < 1/9.5$, and we get an exclusion box of radius $1/9.5$.

The Kantorovich theorem, Theorem 2.1, yields the following results:

$$\begin{aligned} &F'' = \left(\begin{pmatrix} 2 & 1 \\ 4 & 1 \end{pmatrix} \quad \begin{pmatrix} 4 & 1 \\ 1 & 6 \end{pmatrix}\right), \\ &\alpha = 2.5, \quad \beta = 0, \quad \gamma = 12, \quad \Delta = 1, \\ &\underline{r} = 0, \qquad \overline{r} = \frac{2}{2.5 \cdot 12} = \frac{1}{15}; \end{aligned}$$

hence it provides an even smaller (i.e., inferior) exclusion box of radius $\frac{1}{15}$.

If we apply Kahan's theorem, Theorem 2.2, with $F'(\mathbf{x})$, we have to check that $K(\mathbf{x}, \mathbf{x}) \subseteq \operatorname{int}(\mathbf{x})$. Now

$$K(\mathbf{x}, \mathbf{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{\varepsilon}{2}\begin{pmatrix} 19 & 29 \\ 11 & 17 \end{pmatrix}\begin{pmatrix} [-\varepsilon, \varepsilon] \\ [-\varepsilon, \varepsilon] \end{pmatrix}$$

is in $\operatorname{int}(\mathbf{x})$ if

$$\begin{pmatrix} [1 - 24\varepsilon^2, 1 + 24\varepsilon^2] \\ [1 - 14\varepsilon^2, 1 + 14\varepsilon^2] \end{pmatrix} \subseteq \begin{pmatrix} [1 - \varepsilon, 1 + \varepsilon] \\ [1 - \varepsilon, 1 + \varepsilon] \end{pmatrix},$$

which holds for $\varepsilon < 1/24$. This result can be improved if we use slopes instead of interval derivatives. Indeed,

$$K(z, \mathbf{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \varepsilon\begin{pmatrix} 6 & 6 \\ 3.5 & 3.5 \end{pmatrix}\begin{pmatrix} [-\varepsilon, \varepsilon] \\ [-\varepsilon, \varepsilon] \end{pmatrix}$$

is in $\operatorname{int}(\mathbf{x})$ if

$$\begin{pmatrix} [1 - 12\varepsilon^2, 1 + 12\varepsilon^2] \\ [1 - 7\varepsilon^2, 1 + 7\varepsilon^2] \end{pmatrix} \subseteq \begin{pmatrix} [1 - \varepsilon, 1 + \varepsilon] \\ [1 - \varepsilon, 1 + \varepsilon] \end{pmatrix},$$

i.e., for $\varepsilon < 1/12$.

Now we consider the new results. From (31) we get

$$\lambda^e = \frac{2}{v_1 + v_2}. \tag{54}$$

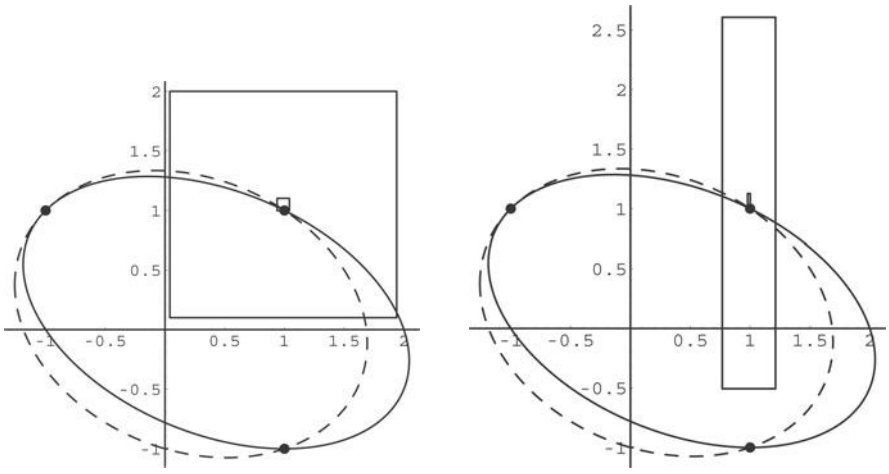


FIG. 3. $\mathbf{x}^e$ *and* $\mathbf{x}^i$ *calculated for Example* 8.2 *with three significant digits for* $v = (1, 1)$ *and* $v = (1, 7)$ *at* $z = (0.99, 1.05)$.

In exact arithmetic, we find $\lambda^e = 1$ so that Corollary 4.4 implies that the interior of the box

$$[x^* - v, x^* + v] = \begin{pmatrix} [0, 2] \\ [0, 2] \end{pmatrix} \tag{55}$$

contains no solution apart from z. In this example, the box is not as large as desirable, since in fact the larger box

$$[x^* - 2v, x^* + 2v] = \begin{pmatrix} [-1, 3] \\ [-1, 3] \end{pmatrix}$$

contains no other solution. However, *the box* (55) *is still one order of magnitude larger than that obtained from the standard uniqueness tests or the Kantorovich theorem.*

If we use inexact arithmetic (we used Mathematica with three significant digits, using this artificially low precision to make the inclusion regions visible in the pictures) and only approximative zeros, the results do not change too much, which can be seen in the pictures of Figure 3.

Corollary 7.3 also gives very promising results. The size of the exclusion boxes again depends on the center z and the vector v. The results for various choices can be found in Figure 4.

To utilize Corollary 5.2 at the exact zero $z = \binom{1}{1}$ we first choose for $u = \binom{1}{1}$ the Perron eigenvector $w_p = \binom{1}{0}$. Its eigenvalue is $\lambda = 1$, and, since $B_0 = 0$ and $\overline{b} = 0$, we conclude that Corollary 5.2 reduces the *first* component of every box $\mathbf{x}$ in the parallelogram P,

$$|x_1 - 1| + |x_2 - 1| < 2, \tag{56}$$

to the thin value $[1, 1]$. That the second component is not reduced is caused by the degeneracy of u. If we choose instead a positive approximation $w = \binom{1}{\varepsilon}$ to w_p and consider any box $\mathbf{x} \subseteq P$, there is $\alpha < 1$ with

$$|x_1 - 1| + |x_2 - 1| < 2\alpha < 2,$$

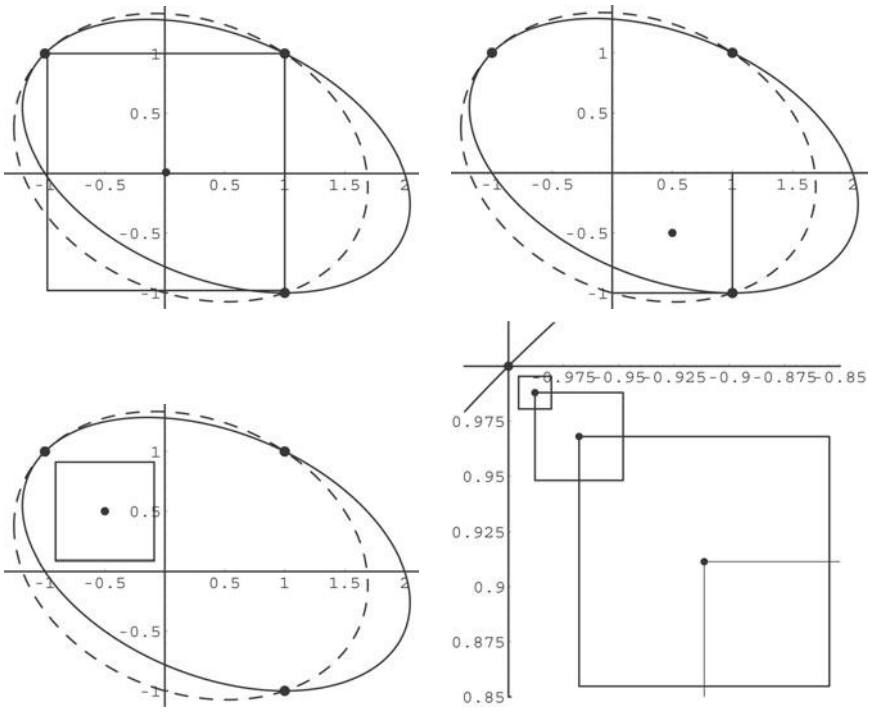


FIG. 4. $\mathbf{x}^\times$ *for Example* 8.2 *and various choices of* z *and* $v = (1, 1)$.

because $\mathbf{x}$ is compact. For $\varepsilon \le 1/\alpha - 1$, we therefore get

$$v = \frac{1}{2}\begin{pmatrix} |x_1 - 1| + (1+\varepsilon)|x_2 - 1| \\ \varepsilon|x_2 - 1| \end{pmatrix} \le \frac{1}{2}\begin{pmatrix} (1+\varepsilon)(|x_1 - 1| + |x_2 - 1|) \\ \varepsilon|x_2 - 1| \end{pmatrix} < w.$$

Then Corollary 5.2 implies that $|x_i - 1| \le 0$ for $i = 1, 2$.

The parallelogram P is best possible in the sense that it contains the other two solutions on its boundary. (But, for general systems, the corresponding maximal exclusion set need not reach another zero and has no simple geometric shape.)

For a nonquadratic polynomial function, all calculations become more complex, and the exclusion sets found are usually far from optimal, though still much better than those from the traditional methods. The $F_k[z, z, x]$ are no longer independent of x, so Theorems 4.3 and 7.2 have to be applied. This involves the computation of a suitable upper bound $\overline{B}_k$ of $F_k[z, z, x]$ by interval arithmetic.

Example 8.3. Figure 5 displays the following system of equations $F(x) = 0$ in two variables, with two polynomial equations of degree 2 and 8:

(57)

$$\begin{aligned} F_1(x) &= x_1^2 + 2x_1x_2 - 2x_2^2 - 2x_1 - 2x_2 + 3, \\ F_2(x) &= x_1^4x_2^4 + x_1^3x_2^4 + x_1^4x_2^3 + 15x_1^2x_2^4 - 8x_1^3x_2^3 + 10x_1^4x_2^2 + 3x_1x_2^4 + 5x_1^2x_2^3 \\ &\quad + 7x_1^3x_2^2 + x_1^4x_2 - 39x_2^4 + 32x_1x_2^3 - 57x_1^2x_2^2 + 21x_1^3x_2 - 17x_1^4 - 27x_2^3 - 17x_1x_2^2 \\ &\quad - 8x_1^2x_2 - 18x_1^3 - 478x_2^2 + 149x_1x_2 - 320x_1^2 - 158x_2 - 158x_1 + 1062. \end{aligned}$$

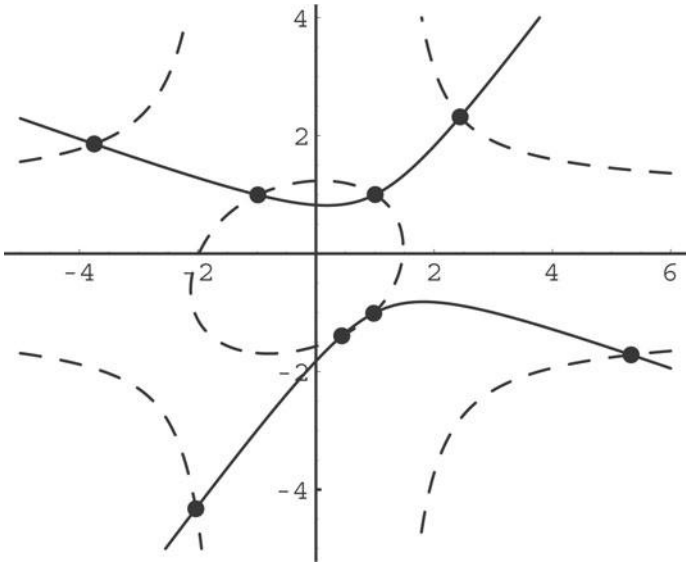


FIG. 5. *Two polynomial equations in two variables; Example* 8.3.

The system (57) has 8 solutions, at approximately

$$\begin{pmatrix}1.0023149901708083\\1.0011595047756938\end{pmatrix},\quad \begin{pmatrix}0.4378266929701329\\-1.3933047617799774\end{pmatrix},\quad \begin{pmatrix}0.9772028387127761\\-1.0115934531170049\end{pmatrix},$$

$$\begin{pmatrix}-0.9818234823156266\\0.9954714636375825\end{pmatrix},\quad \begin{pmatrix}-3.7502535429488344\\1.8585101451403585\end{pmatrix},\quad \begin{pmatrix}2.4390986061035260\\2.3174396617957018\end{pmatrix},$$

$$\begin{pmatrix}5.3305903297000243\\-1.7161362016394848\end{pmatrix},\quad \begin{pmatrix}-2.0307311621763933\\-4.3241016906293375\end{pmatrix}.$$

We consider the approximate solution $z = \binom{0.99}{1.01}$. For the set S we choose the box $[z-u, z+u]$ with $u = \binom{1}{1}$. In this case we have

$$F(z) \approx \begin{pmatrix}-0.0603\\-1.170\end{pmatrix},\qquad F'(z) \approx \begin{pmatrix}2 & -4.06\\-717.55 & -1147.7\end{pmatrix},$$

$$F_1[z,z,x] = \begin{pmatrix}1 & 0\\f_1 & 0\end{pmatrix},\qquad F_2[z,z,x] = \begin{pmatrix}2 & -2\\f_2 & f_3\end{pmatrix},$$

where

$$\begin{aligned}f_1 \approx &-405.63 - 51.66x_1 - 17x_1^2 + 36.52x_2 + 23x_1x_2 + x_1^2x_2\\ &- 13.737x_2^2 + 26.8x_1x_2^2 + 10x_1^2x_2^2 - 7.9x_2^3 - 6.02x_1x_2^3 + x_1^2x_2^3\\ &+ 19.92x_2^4 + 2.98x_1x_2^4 + x_1^2x_2^4,\\ f_2 \approx &\quad 191.04 - 7.6687x_2 + 62.176x_2^2 + 39.521x_2^3,\\ f_3 \approx &-588.05 - 36.404x_2 - 19.398x_2^2.\end{aligned}$$

We further compute

$$C = \begin{pmatrix} 0.22035 & -0.00077947 \\ -0.13776 & -0.00038397 \end{pmatrix},$$

$$B_0 = 10^{-5}\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \overline{B}_1 = \begin{pmatrix} 1.0636 & 0 \\ 0.5027 & 0 \end{pmatrix}, \quad \overline{B}_2 = \begin{pmatrix} 0.3038 & 0.1358 \\ 0.5686 & 0.5596 \end{pmatrix}, \quad \overline{b} = \begin{pmatrix} 0.0124 \\ 0.0088 \end{pmatrix}.$$

If we use Theorem 4.3 for $v = \binom{1}{1}$, we get

$$w = \begin{pmatrix} 0.99999 \\ 0.99998 \end{pmatrix}, \quad a = \begin{pmatrix} 1.5032 \\ 1.6309 \end{pmatrix}, \quad D = \begin{pmatrix} 0.925421 \\ 0.942575 \end{pmatrix},$$

$$\lambda^i = 0.0126403, \qquad \lambda^e = 0.604222,$$

so we may conclude that there is exactly one zero in the box

$$\mathbf{x}^i = \begin{pmatrix} [0.97736, 1.00264] \\ [0.99736, 1.02264] \end{pmatrix},$$

and this zero is the only zero in the interior of the exclusion box

$$\mathbf{x}^e = \begin{pmatrix} [0.385778, 1.59422] \\ [0.405778, 1.61422] \end{pmatrix}.$$

In Figure 6 the two boxes are displayed.

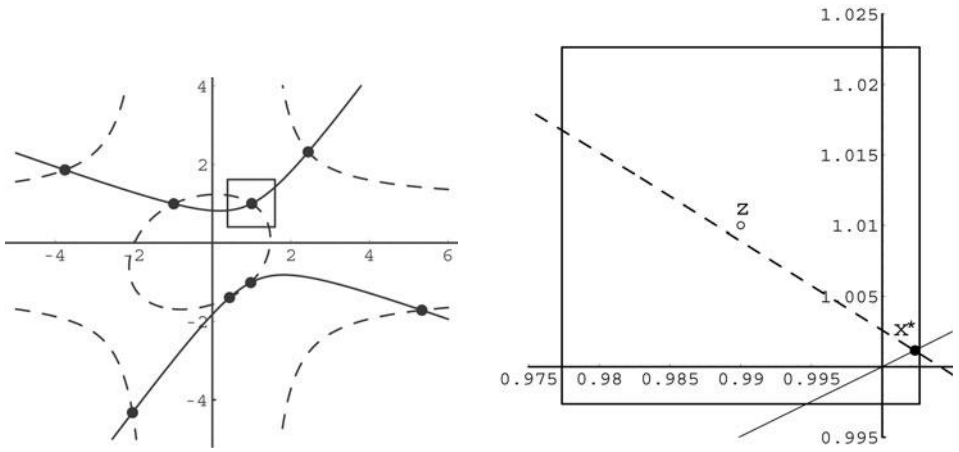


FIG. 6. *Exclusion and inclusion boxes for Example* 8.3 *at* $z = (0.99, 1.01)$.

Next we consider the point $z = \binom{1.5}{-1.5}$ to test Theorem 7.2. We compute

$$F(z) \approx \begin{pmatrix} -3.75 \\ -1477.23 \end{pmatrix}, \qquad F_1[z, z, x] \approx \begin{pmatrix} 1 & 0 \\ g_1 & 0 \end{pmatrix},$$

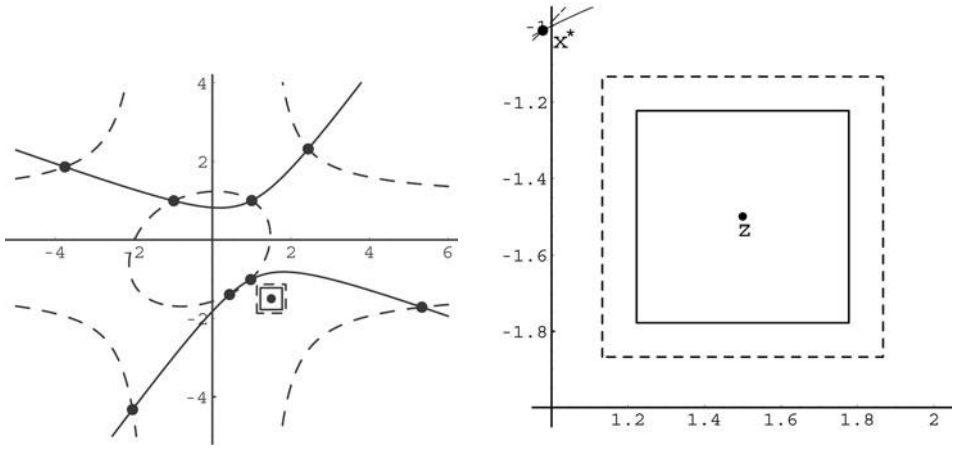


FIG. 7. *Exclusion boxes for Example* 8.3 *at* $z = (1.5, -1.5)$.

$$F'(z) \approx \begin{pmatrix} -2 & 7 \\ -1578.73 & 1761.77 \end{pmatrix}, \qquad F_2[z,z,x] = \begin{pmatrix} 2 & -2 \\ g_2 & g_3 \end{pmatrix},$$

with

$$\begin{aligned} g_1 \approx & -488.75 - 69x_1 - 17x_1^2 + 61.75x_2 + 24x_1x_2 + x_1^2x_2 \\ & + 31.5x_2^2 + 37x_1x_2^2 + 10x_1^2x_2^2 - 12.25x_2^3 - 5x_1x_2^3 + x_1^2x_2^3 \\ & + 24.75x_2^4 + 4x_1x_2^4 + x_1^2x_2^4, \\ g_2 \approx & \quad 73.1563 + 138.063x_2 - 95.875x_2^2 + 68.25x_2^3, \\ g_3 \approx & -536.547 - 12.75x_2 + 7.6875x_2^2. \end{aligned}$$

Performing the necessary computations, we find for $\mathbf{x} = [z-u, z+u]$ with $u = \frac{1}{2}\binom{1}{1}$

$$F'(z)^{-1} \approx \begin{pmatrix} 0.234 & -0.00093 \\ 0.21 & -0.000266 \end{pmatrix}, \qquad \underline{b} = \begin{pmatrix} 0.496 \\ 0.3939 \end{pmatrix},$$

$$\overline{B}_1 = \begin{pmatrix} 1.2895 & 0 \\ 0.5113 & 0 \end{pmatrix}, \quad B_0' = \begin{pmatrix} 1 & 10^{-5} \\ 10^{-5} & 1.00001 \end{pmatrix}, \quad \overline{B}_2 = \begin{pmatrix} 1.5212 & 0.0215 \\ 0.7204 & 0.2919 \end{pmatrix}.$$

Now we use Theorem 7.2 for $v = \binom{1}{1}$ and $C = F'(z)^{-1}$ and get

$$w^\times = \begin{pmatrix} 1.00001 \\ 1.00002 \end{pmatrix}, \quad a^\times = \begin{pmatrix} 2.8322 \\ 1.5236 \end{pmatrix}, \quad D^\times = \begin{pmatrix} 6.6191 \\ 3.4006 \end{pmatrix}, \quad \lambda^\times = 0.277656;$$

so we conclude that there are no zeros of F in the interior of the exclusion box

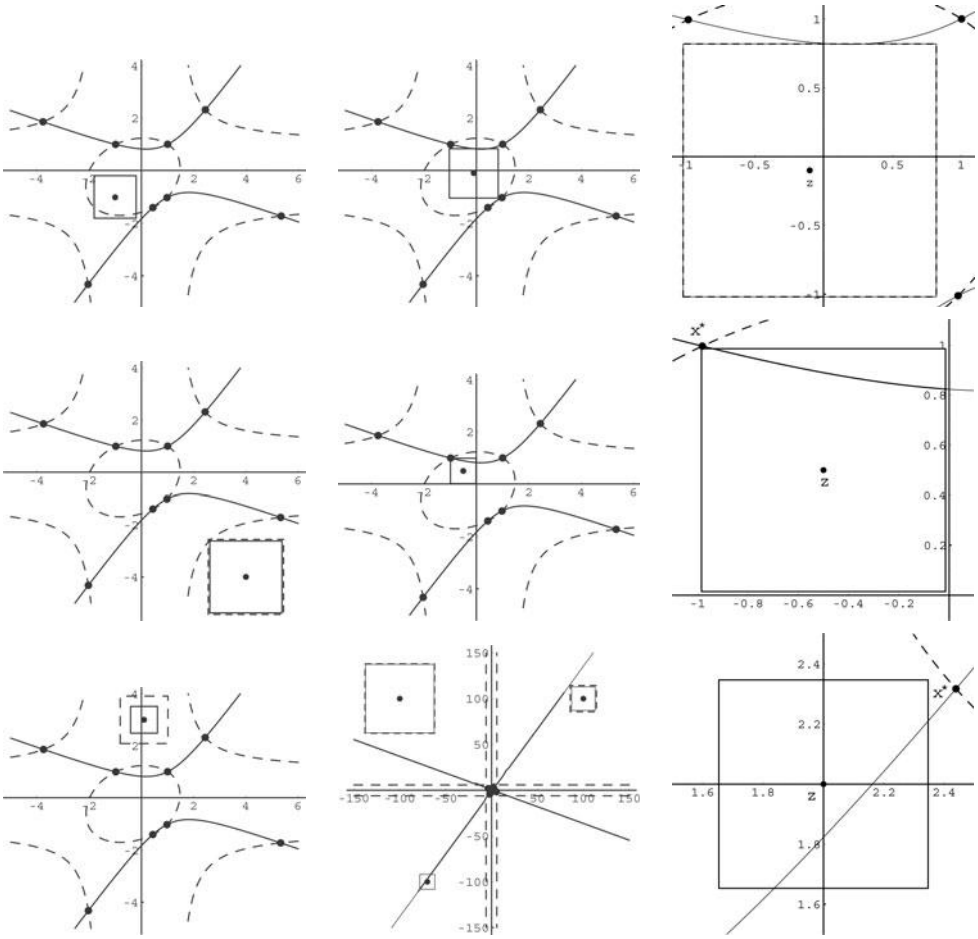


FIG. 8. *Exclusion boxes for Example* 8.3 *in various regions of* $\mathbb{R}^2$.

$$\mathbf{x}^{\times} = \begin{pmatrix} [1.22234, 1.77766] \\ [-1.77766, -1.22234] \end{pmatrix}.$$

However, the choice $C = F'(z)^{-1}$ is not best possible in this situation. If we take

$$C = \begin{pmatrix} 1 & 0.002937 \end{pmatrix},$$

we compute $\lambda^{\times} = 0.367223$ and find the considerably larger exclusion box

$$\mathbf{x}^{\times} = \begin{pmatrix} [1.13278, 1.86722] \\ [-1.86722, -1.13278] \end{pmatrix}.$$

Figure 7 shows both boxes, the bigger one in dashed lines.

Finally, Figure 8 shows various exclusion boxes for nonzeros, and Figure 9 contains exclusion boxes and some inclusion boxes for all of the zeros of F.

While the previous examples were low dimensional, our final example shows that the improvements over traditional results may even be more pronounced for higher dimensional problems with poorly conditioned zeros.

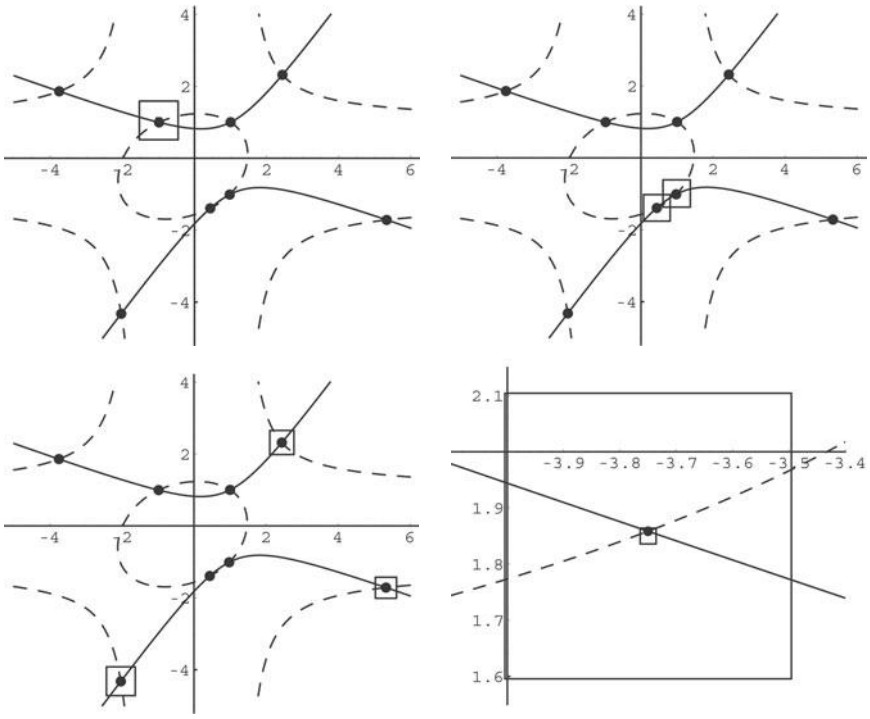


FIG. 9. *Exclusion boxes for all zeros of F in Example* 8.3.

Example 8.4. We consider the set of equations

$$\sum_{k=1}^{n} x_k^i = H(n,-i) \quad \text{for } i = 1,\dots,n,$$

where the *harmonic numbers* $H(n,m)$ are defined as

$$H(n,m) := \sum_{k=1}^{n} k^{-m}.$$

Clearly, $x_k^* = k$ is a solution, and the complete set of solutions is given by all permutations of this vector.

We compare the results provided by Theorem 4.3 with the exclusion box obtained by strong regularity of the slope $F[z,\mathbf{x}]$ (which in the previous examples was the best among the traditional choices). The vector v needed in Theorem 4.3 was chosen as the all-one vector e. All numerical calculations were performed in double precision arithmetic.

The results are collected in Table 1; R denotes the radius of the exclusion box computed by Theorem 4.3, r the radius of the exclusion box implied by strong regularity of $F[z,\mathbf{x}]$, and κ the condition number of $F'(x^*)$. All numbers are approximate.

From the logarithmic plot in Figure 10, we see that the radii of the exclusion boxes decrease in both cases exponentially with n. However, the quotient of the two radii increases exponentially with n. This shows that our new method suffers much less from the double deterioration due to the increase of both dimension and the Jacobian condition number at the zero.

TABLE 1

n	R	r	R/r	κ
2	1	1	1.000	10.91
3	0.41316	0.127017	3.253	153.155
4	0.197355	0.0206925	9.538	3021.56
5	0.082	0.00359092	22.835	76819.8
6	0.034	0.00063524	53.523	$2.38489 \cdot 10^{6}$
7	0.013	0.00011303	115.007	$8.7331 \cdot 10^{7}$
8	0.005	0.000020137	248.296	$3.68207 \cdot 10^{9}$
9	0.00185847	$3.58494 \cdot 10^{-6}$	518.408	$1.75585 \cdot 10^{11}$
10	0.00068	$6.3732199 \cdot 10^{-7}$	1066.960	$9.34062 \cdot 10^{12}$
11	0.00025	$1.1311565 \cdot 10^{-7}$	2210.130	$5.48274 \cdot 10^{14}$
12	0.000092	$2.00428 \cdot 10^{-8}$	4590.190	$3.52073 \cdot 10^{16}$
13	0.000034	$3.5455649 \cdot 10^{-9}$	9589.450	$2.46174 \cdot 10^{18}$
14	0.0000125	$6.26252 \cdot 10^{-10}$	19960.000	$5.6081 \cdot 10^{19}$
15	$4.5043 \cdot 10^{-6}$	$1.1045 \cdot 10^{-10}$	40781.400	$2.64518 \cdot 10^{20}$
16	$1.6527 \cdot 10^{-6}$	$1.94493 \cdot 10^{-11}$	84975.400	$9.40669 \cdot 10^{21}$

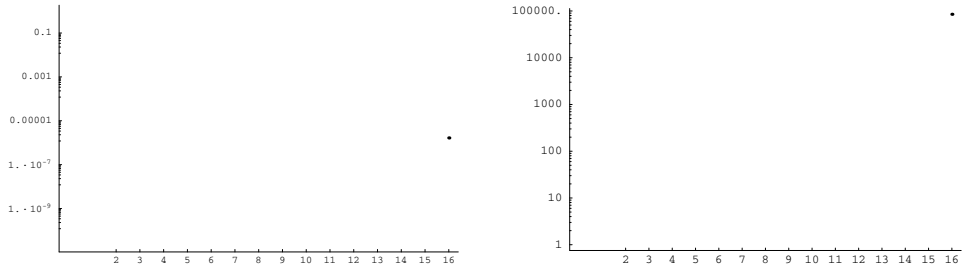


FIG. 10. *Radii of the exclusion boxes and quotient of the radii for Example* 8.4.

REFERENCES

[1] M. BERZ AND J. HOEFKENS, *Verified high-order inversion of functional dependencies and interval Newton methods*, Reliab. Comput., 7 (2001), pp. 379–398.

[2] P. DEUFLHARD AND G. HEINDL, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal., 16 (1979), pp. 1–10.

[3] E. HANSEN, *Preconditioning linearized equations*, Computing, 58 (1997), pp. 187–196.

[4] C. JANSSON, *On self-validating methods for optimization problems*, in Topics in Validated Computations, J. Herzberger, ed., Elsevier, Amsterdam, 1994, pp. 381–438.

[5] W. M. KAHAN, *A More Complete Interval Arithmetic*, Lecture notes for an engineering summer course in numerical analysis, University of Michigan, Ann Arbor, MI, 1968.

[6] F. KALOVICS, *Creating and handling box valued functions used in numerical methods*, J. Comput. Appl. Math., 147 (2002), pp. 333–348.

[7] L. B. KANTOROVICH, *Functional analysis and applied mathematics*, Uspekhi Mat. Nauk, 3 (1948), pp. 89–185 (in Russian). Translated by C. D. Benster, Nat. Bur. Stand. Rep. 1509, Washington, DC, 1952.

[8] R. B. KEARFOTT, *Empirical evaluation of innovations in interval branch and bound algorithms for nonlinear systems*, SIAM J. Sci. Comput., 18 (1997), pp. 574–594.

[9] R. B. KEARFOTT, *A review of techniques in the verified solution of constrained global optimization problems*, in Applications of Interval Computations, R. B. Kearfott and V. Kreinovich, eds., Kluwer, Dordrecht, The Netherlands, 1996, pp. 23–60.

[10] R. B. KEARFOTT, *Rigorous Global Search: Continuous Problems*, Kluwer, Dordrecht, The Netherlands, 1996.

[11] R. B. Kearfott and K. Du, *The cluster problem in multivariate global optimization*, J. Global Optim., 5 (1994), pp. 253–265.
[12] R. B. Kearfott, C. Hu, and M. Novoa, III, *A review of preconditioners for the interval Gauss-Seidel method*, Interval Comput., 1 (1991), pp. 59–85.
[13] L. V. Kolev, *Use of interval slopes for the irrational part of factorable functions*, Reliab. Comput., 3 (1997), pp. 83–93.
[14] R. Krawczyk and A. Neumaier, *Interval slopes for rational functions and associated centered forms*, SIAM J. Numer. Anal., 22 (1985), pp. 604–616.
[15] A. Kuntsevich and F. Kappel, *SolvOpt*, http://www.kfunigraz.ac.at/imawww/kuntsevich/solvopt/(1997).
[16] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.
[17] A. Neumaier, *Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, UK, 2001.
[18] A. Neumaier, *Taylor forms—use and limits*, Reliab. Comput., 9 (2002), pp. 43–79.
[19] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Classics Appl. Math. 30, SIAM, Philadelphia, 2000.
[20] S. M. Rump, *Expansion and estimation of the range of nonlinear functions*, Math. Comp., 65 (1996), pp. 1503–1512.
[21] S. M. Rump, *INTLAB—INTerval LABoratory*, in Developments in Reliable Computing, T. Csendes, ed., Kluwer, Dordrecht, The Netherlands, 1999, pp. 77–104.
[22] Z. Shen and A. Neumaier, *The Krawczyk operator and Kantorovich's theorem*, J. Math. Anal. Appl., 149 (1990), pp. 437–443.
[23] P. Van Hentenryck, L. Michel, and Y. Deville, *Numerica. A Modeling Language for Global Optimization*, MIT Press, Cambridge, MA, 1997.
[24] R. J. Van Iwaarden, *An Improved Unconstrained Global Optimization Algorithm*, Ph.D. thesis, University of Colorado, Denver, CO, http://www-math.cudenver.edu/graduate/thesis/rvan.ps.gz (1996).

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 409–433

PERFECTLY MATCHED LAYERS FOR THE CONVECTED HELMHOLTZ EQUATION*

E. BÉCACHE†, A.-S. BONNET-BEN DHIA‡, AND G. LEGENDRE§

Abstract. In this paper, we propose and analyze perfectly matched absorbing layers for a problem of time-harmonic acoustic waves propagating in a duct in the presence of a uniform flow. The absorbing layers are designed for the pressure field, satisfying the convected scalar Helmholtz equation. A difficulty, compared to the Helmholtz equation, comes from the presence of so-called inverse upstream modes which become unstable, instead of evanescent, with the classical Bérenger's perfectly matched layers (PMLs). We investigate here a PML model, recently introduced for time-dependent problems, which makes all outgoing waves evanescent. We then analyze the error due to the truncation of the domain and prove that the convergence is exponential with respect to the size of the layers for both the classical and the new PML models. Numerical validations are finally presented.

Key words. acoustic waves, convected Helmholtz equation, duct modes, absorbing layers, perfectly matched layer, instabilities

AMS subject classifications. 35J05, 65N12, 76Q05

DOI. 10.1137/S0036142903420984

1. Introduction.

Perfectly matched layers (PMLs) were introduced by Bérenger [3] in order to design efficient numerical absorbing boundary conditions (more precisely, absorbing layers) for the computation of time-dependent solutions of Maxwell's equations in unbounded domains. They have since been used for numerous applications, mostly in the time domain [4, 28, 5, 23] but also for time-harmonic wave-like equations [27, 15].

In particular, PMLs have been used for the solution in the time domain of the linearized Euler equations [19, 13, 16, 26], which model acoustic propagation in the presence of a flow. In this case, it has been observed that PMLs can lead to instabilities, due to the presence of waves whose phase and group velocities have opposite signs [26] (see [2] for a general analysis of this phenomenon). Some techniques have been developed to overcome this difficulty, making the layers stable but, unfortunately, no longer perfectly matched [16, 1]. More recently, ideas for designing stable PMLs for this problem have emerged from several teams independently. These new approaches, which seem to be very closely related, have been developed for time-dependent applications in [20, 11, 14] and for time-harmonic applications in the present paper. These different works all deal with the case of a parallel flow, which is orthogonal to the layers.

*Received by the editors January 9, 2003; accepted for publication (in revised form) August 15, 2003; published electronically March 3, 2004.
http://www.siam.org/journals/sinum/42-1/42098.html

†Laboratoire POems, UMR 2706 CNRS/ENSTA/INRIA, Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay cedex, France (eliane.becache@inria.fr).

‡Laboratoire POems, UMR 2706 CNRS/ENSTA/INRIA, École Nationale Supérieure des Techniques Avancées, 32 boulevard Victor, 75739 Paris cedex 15, France (bonnet@ensta.fr).

§Office National d'Études et de Recherches Aérospatiales, BP 72, 29 avenue de la Division Leclerc, 92322 Châtillon cedex, France. Current address: Laboratoire POems, UMR 2706 CNRS/ENSTA/INRIA, École Nationale Supérieure des Techniques Avancées, 32 boulevard Victor, 75739 Paris cedex 15, France (legendre@ensta.fr).

We are concerned with the propagation of acoustic waves in a duct in the presence of a uniform flow. For such a mean flow, thetime-harmonic linearized Euler equations reduce into a scalar convected Helmholtz equation for the pressure. In this particular case, one could, of course, use a Dirichlet-to-Neumann (DtN) operator to obtain an equivalent problem in a bounded domain. However, the PMLs, being local, are easier to implement, and we intend to extend this method to vectorial cases, involving more general flows in a forthcoming paper.

When applying the *classical* (i.e., Bérenger's) PMLs to the convected Helmholtz equation in a duct, a simple modal analysis shows that the presence of the so-called inverse upstream modes produces an exponential blow-up of the solution in the space variable. This is easy to see, remembering the interpretation of the PMLs as a complex change of variable [8, 24, 10, 9]. This change of variable corresponds to a similarity applied on the axial wave numbers of the modes. For the classical Helmholtz equation, this similarity makes all outgoing modes become evanescent. But in the presence of a flow, the transformation sends the inverse upstream modes into the "bad" part of the complex plane, leading to the instabilities observed in the time domain.

The idea proposed here, which is similar to those developed independently in [20, 14] for time domain applications, consists of applying a translation before the similarity to the axial wave numbers. This removes the unstable modes. We will call the PMLs thus obtained *new* PMLs.

The object of this paper is the analysis of the convergence of both PML models as the thickness of the sponge layer tends to infinity. Similar convergence analyses have already been carried out for the Helmholtz equation, via boundary integral equation techniques in [21] or using the pole condition in [18]. Surprisingly, we prove that, for the convected Helmholtz equation, the two models always converge. In other words, contrary to time domain applications, the presence of unstable modes does not affect the efficiency of the classical PMLs.

Finally, let us emphasize that in most papers concerning PMLs for time-harmonic applications, coefficients are designed in order to satisfy requirements established for time domain applications. We show that this choice is too restrictive: for instance, the particular dependence of these coefficients regarding the frequency has no more justification for the present case.

The outline of this paper is as follows: the equations of the scattering problem are presented in section 2. A formulation in a bounded domain is given, involving DtN conditions on the fictitious boundaries, which are known explicitly through modal expansions. Finally, the well posedness is proven using Fredholm theory.

Classical and new PML techniques, with constant coefficients, are described in section 3. A modal analysis indicates that these layers are "perfectly matched." Besides, they are absorbing, except the classical PMLs in the presence of the so-called inverse upstream modes.

Section 4 is devoted to the analysis of the error due to the truncation of the layers. An equivalent formulation of the problem with PMLs is written in the physical domain, the thickness of the layers appearing in the expression of the DtN maps. In this way, we prove that both PML models converge to the physical solution, as the length of the layers tends to infinity. More precisely, the error in the physical domain does not depend on the PML model under consideration and decreases exponentially fast for both models. Note, however, that classical PMLs lead to an exponentially large solution in the layers, whereas the solution computed with new PMLs is evanescent in the layers.

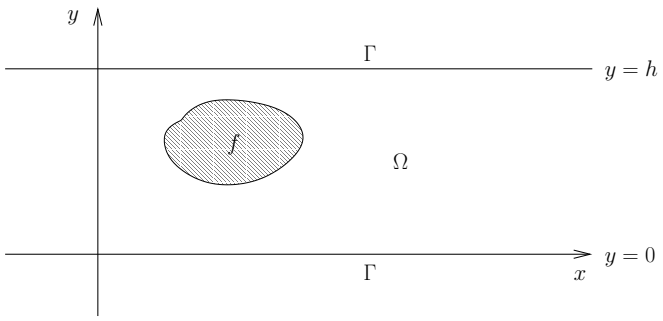


FIG. 1. *The infinite duct.*

Extension to the case of layers with spatially varying coefficients is discussed in section 5, and numerical illustrations are given in the last section.

2. The physical and the mathematical models.

2.1. The problem in the infinite duct. We consider an infinite rigid duct carrying a mean fluid flow; see Figure 1. The problem is two-dimensional, set in the xy-plane, where the x- (resp., y-) axis is parallel (resp., normal) to the walls of the duct. Mathematically, the duct is defined by the unbounded domain $\Omega = \mathbb{R} \times [0, h]$, where h denotes the distance between the rigid walls.

To describe the propagation of acoustic waves in the duct, we assume the following approximations to be valid:

- The fluid is homogeneous, nonviscous, and nonheat conductive.
- The thermodynamic processes are adiabatic.
- The mean velocity v_0 is subsonic and uniform.
- The perturbations are small, and equations are linear in the acoustic quantities.
- A harmonic time dependence $\exp(-i\omega t)$, $\omega > 0$ being the pulsation, is assumed (although this factor is suppressed throughout).

The acoustic pressure field $p(x, y)$ then satisfies the convected Helmholtz equation in the infinite duct:

$$(1 - M^2)\,\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + 2ikM\,\frac{\partial p}{\partial x} + k^2\,p = f \quad \text{in } \Omega, \tag{2.1}$$

where $f \in L^2(\Omega)$ is a compactly supported function and $M = v_0/c_0$ and $k = \omega/c_0$ are, respectively, the Mach number ($-1 < M < 1$) and the wave number, c_0 being the sound velocity in the fluid. In addition to (2.1), the pressure satisfies the Neumann homogeneous boundary condition on the two rigid walls of the duct:

$$\frac{\partial p}{\partial y} = 0 \quad \text{on } \Gamma = \partial\Omega. \tag{2.2}$$

To obtain a well-posed problem, a "radiation condition," which selects the "outgoing" waves, needs to be defined at infinity. This condition is nonlocal and is given in terms of the DtN operator. This requires the introduction of the so-called modes of the duct, which are the solutions of (2.1)–(2.2) in the absence of a source ($f = 0$) and with separated variables. These are given by

$$p_n^{\pm}(x, y) = e^{i\beta_n^{\pm} x}\,\varphi_n(y),$$

where

$$\varphi_0(y) = \sqrt{\frac{1}{h}}, \quad \varphi_n(y) = \sqrt{\frac{2}{h}}\,\cos\Big(\frac{n\pi y}{h}\Big), \quad n \in \mathbb{N}^*, \tag{2.3}$$

and where the axial wave numbers $\beta_n^{\pm}$ are the solutions of

$$-(1-M^2)\,\beta^2 - 2kM\,\beta + k^2 = \frac{n^2\pi^2}{h^2}, \quad n \in \mathbb{N}.$$

Let us introduce

$$K_0 = \frac{kh}{\pi\sqrt{1-M^2}}, \tag{2.4}$$

and let $N_0 = [K_0]$ denote the integer part of K_0. If $n \le N_0$, $\beta_n^{\pm}$ is real and equal to

$$\beta_n^{\pm} = \frac{-kM \pm \sqrt{k^2 - \frac{n^2\pi^2}{h^2}(1-M^2)}}{1-M^2}. \tag{2.5}$$

In this case, $p_n^{\pm}$ is called a propagative mode. The number of propagative modes is an increasing function of the Mach number M, which is assumed to be positive. Simple calculations show that the group velocity $\frac{\partial\omega}{\partial\beta}$ is positive for the p_n^+ modes and negative for the p_n^- modes. A well-known effect of the presence of flow is the existence, when

$$\sqrt{1-M^2}\,\frac{n\pi}{h} < k < \frac{n\pi}{h},$$

of modes p_n^+ which have a negative phase velocity $\frac{\omega}{\beta}$ and a positive group velocity. These are called inverse upstream modes.

The axial wave number $\beta_n^{\pm}$ is complex if $n > N_0$:

$$\beta_n^{\pm} = \frac{-kM \pm i\sqrt{\frac{n^2\pi^2}{h^2}(1-M^2) - k^2}}{1-M^2}. \tag{2.6}$$

In this case, $p_n^{\pm}$ is exponentially decreasing when $x \to \pm\infty$ and is called an evanescent mode.

2.2. Reduction to a bounded domain. We now want to select the outgoing solution (2.1)–(2.2), which corresponds to a superposition of p_n^+ (resp., p_n^-) modes when $x \to +\infty$ (resp., $x \to -\infty$), i.e., either to the propagative modes with a positive (resp., negative) group velocity or to the evanescent modes.

To derive the appropriate DtN boundary condition, we introduce the bounded domain Ω_b, located in between two boundaries $\Sigma_{\pm}$, respectively, located at $x = x_-$ and $x = x_+$ (see Figure 2), such that the support of the source f is included in Ω_b:

$$\Omega_b = \{(x,y) \in \Omega,\ x_- \le x \le x_+\}.$$

We set $\Omega_{\pm}$ the complementary domains

$$\Omega_- = \{(x,y) \in \Omega,\ x < x_-\} \quad \text{and} \quad \Omega_+ = \{(x,y) \in \Omega,\ x > x_+\}.$$

The solution p of (2.1) then satisfies the homogeneous equation

$$(1-M^2)\,\frac{\partial^2 p}{\partial x^2} + 2ikM\,\frac{\partial p}{\partial x} + \frac{\partial^2 p}{\partial y^2} + k^2\,p = 0 \quad \text{in } \Omega_{\pm} \tag{2.7}$$

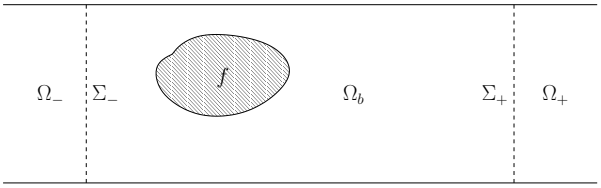


FIG. 2. *The bounded domain.*

and therefore can be decomposed on the modes. Consequently, in Ω_-, i.e., for $x < x_-$, we have

$$p(x,y) = \sum_{n=0}^{+\infty} \left(p(x_-,.),\varphi_n\right)_{L^2(\Sigma_-)}\ \varphi_n e^{i\beta_n^-(x-x_-)},$$

and in Ω_+, i.e., for $x > x_+$,

$$p(x,y) = \sum_{n=0}^{+\infty} \left(p(x_+,.),\varphi_n\right)_{L^2(\Sigma_+)}\ \varphi_n e^{i\beta_n^+(x-x_+)},$$

where $(\cdot,\cdot)_{L^2(\Sigma_+)}$ (resp., $(\cdot,\cdot)_{L^2(\Sigma_-)}$) denotes the $L^2(\Sigma_+)$ (resp., $L^2(\Sigma_-)$) inner product for scalar functions:

$$(u,v)_{L^2(\Sigma_\pm)} \equiv \int_{\Sigma_\pm} u(y)\overline{v}(y)\,\mathrm{d}y.$$

The DtN operators $T_\pm$ can then be defined as

$$\begin{aligned} T_\pm :\ H^{1/2}(\Sigma_\pm) &\to H^{-1/2}(\Sigma_\pm), \\ \phi &\mapsto \mp\sum_{n=0}^{+\infty} i\beta_n^\pm\ (\phi,\varphi_n)_{L^2(\Sigma_\pm)}\ \varphi_n(y), \end{aligned} \tag{2.8}$$

and we have the following boundary conditions on $\Sigma_\pm$ for the solution of (2.1):

$$\frac{\partial p}{\partial \boldsymbol{n}} = -T_\pm p \quad \text{on } \Sigma_\pm, \tag{2.9}$$

where the vector $\boldsymbol{n}$ denotes the unit outward normal to $\Sigma_\pm$.

Having established exact boundary conditions satisfied by p, we can now define a problem in the bounded domain Ω_b: find $p \in H^1(\Omega_b)$ such that

$$\begin{cases} (1-M^2)\,\dfrac{\partial^2 p}{\partial x^2} + 2ikM\,\dfrac{\partial p}{\partial x} + \dfrac{\partial^2 p}{\partial y^2} + k^2\,p = f & \text{in } \Omega_b, \\ \dfrac{\partial p}{\partial y} = 0 & \text{on } \Gamma \cap \partial\Omega_b, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = -T_\pm p & \text{on } \Sigma_\pm. \end{cases} \tag{2.10}$$

The fact that f is compactly supported in Ω_b shows clearly that problems (2.10) and (2.1)–(2.2) are equivalent in the sense of the following proposition.

PROPOSITION 2.1. *If p is a solution of system* (2.1)–(2.2)*, then $p_{|\Omega_b}$ is a solution of* (2.10)*. Conversely, if $\widetilde{p}$ is a solution of* (2.10)*, then $\widetilde{p}$ can be extended in a unique way to a solution of* (2.1)–(2.2)*.*

2.3. Well posedness. Formulation (2.10) has two main advantages. First, from a theoretical point of view, it provides a result of existence and uniqueness of the solution. Second, it can be used to obtain numerical solutions, since it is posed in a bounded domain.

An equivalent weak form of system (2.10) can then be written as follows: find $p \in H^1(\Omega_b)$ such that

$$a_{\Omega_b}(p,q) = -\int_{\Omega_b} f\overline{q}\,\mathrm{d}x\,\mathrm{d}y \quad \forall q \in H^1(\Omega_b), \tag{2.11}$$

where the sesquilinear form $a_{\Omega_b}(\cdot\,,\cdot)$ is defined by

$$a_{\Omega_b}(p,q) = b(p,q) + c(p,q), \tag{2.12}$$

with

$$b(p,q) = \int_{\Omega_b} \left((1-M^2)\,\frac{\partial p}{\partial x}\frac{\partial \overline{q}}{\partial x} + \frac{\partial p}{\partial y}\frac{\partial \overline{q}}{\partial y} + p\overline{q} \right) \mathrm{d}x\,\mathrm{d}y + \langle T_+ p, q\rangle_{\Sigma_+} + \langle T_- p, q\rangle_{\Sigma_-}\,, \tag{2.13}$$

where the brackets $\langle\cdot,\cdot\rangle_{\Sigma_+}$ (resp., $\langle\cdot,\cdot\rangle_{\Sigma_-}$) denote the natural duality pairing between $H^{-1/2}(\Sigma_+)$ and $H^{1/2}(\Sigma_+)$ (resp., $H^{-1/2}(\Sigma_-)$ and $H^{1/2}(\Sigma_-)$), and

$$c(p,q) = \int_{\Omega_b} \left(-2ikM\,\frac{\partial p}{\partial x}\overline{q} - (1+k^2)\,p\overline{q} \right) \mathrm{d}x\,\mathrm{d}y. \tag{2.14}$$

It was shown in [6] that this problem is of Fredholm type. By the Fredholm alternative, problem (2.11) is well posed if and only if the homogeneous problem has no solution except the trivial one, $p = 0$.

THEOREM 2.2. *The problem is well posed if and only if*

$$k \neq \sqrt{1-M^2}\,\frac{n\pi}{h} \quad \forall n \in \mathbb{N}. \tag{2.15}$$

Proof. Suppose that p is a solution of (2.1)–(2.2) with $f = 0$. Then there are complex constants A_n^+ and A_n^- such that

$$p(x,y) = \sum_{n=0}^{+\infty} \left(A_n^+\, e^{i\beta_n^+ x} + A_n^-\, e^{i\beta_n^- x} \right) \varphi_n(y),$$

with definitions (2.3) through (2.6). Boundary condition (2.9) then gives

$$A_n^-(\beta_n^- - \beta_n^+) = A_n^+(\beta_n^- - \beta_n^+) = 0$$

so that p vanishes identically if $\beta_n^+ \neq \beta_n^-$ or, likewise, if $k^2 \neq (1-M^2)\,\frac{n^2\pi^2}{h^2}$.

Suppose conversely that $k = \sqrt{1-M^2}\,\frac{n\pi}{h}$; then

$$\beta_n^+ = \beta_n^- = -\frac{kM}{1-M^2},$$

and $\varphi_n(y)\,e^{i\beta_n^+ x}$ is a nontrivial solution of the homogeneous problem. ∎

In what follows, we assume the problem is well posed, which means that (2.15) is satisfied.

3. The PML model. The PML model introduced by Bérenger for the time-dependent Maxwell equations can be constructed using a complex change of variable in the frequency domain, as shown in [8, 10, 9]. We use this same approach in the present paper. This is closely related to the technique known as dilation analyticity for the study of resonances [17].

In this section, we briefly recall some properties of the classical PML formulation for the Helmholtz equation. Note that in the context of propagation in a waveguide the interpretation of the method relies on the modal approach instead of the usual plane wave approach. This modal analysis allows us to point out the origin of the instabilities in the presence of flow and leads naturally to the introduction of a new model of PMLs as a remedy.

3.1. Modal analysis of Bérenger's model in a waveguide. The purpose of the method is to provide a fictitious, absorbing medium such that its interface with the "physical" bounded domain does not reflect any outgoing mode. Transposing Bérenger's formulation in the frequency domain from its original setting in the time domain consists of making the following substitution:

$$\frac{\partial}{\partial x} \longrightarrow \alpha \, \frac{\partial}{\partial x}, \tag{3.1}$$

where α is a complex function taken to be

$$\alpha(x) = \frac{-i\omega}{-i\omega + \sigma(x)} \tag{3.2}$$

with $\sigma(x)$ a real, positive function such as $\sigma(x) = 0$ (and therefore $\alpha(x) = 1$) in Ω_b, the derivative with respect to y being left unchanged.

In the case of the Helmholtz equation, we obtain

$$\alpha(x) \, \frac{\partial}{\partial x} \left(\alpha(x) \, \frac{\partial p}{\partial x} \right) + \frac{\partial^2 p}{\partial y^2} + k^2 \, p = f \quad \text{in } \Omega. \tag{3.3}$$

Note that the writing of this equation in the weak sense implies the following jump conditions at the interfaces between Ω_b and the PMLs:

$$[p(x,y)] = 0 \quad \text{and} \quad \left[\alpha(x) \, \frac{\partial p}{\partial x}(x,y) \right] = 0. \tag{3.4}$$

For the modal analysis in the waveguide, we now assume that $\alpha(x)$ is a constant in $\Omega \setminus \Omega_b$, which we still denote by α for the sake of simplicity. In other words,

$$\alpha(x) = \begin{cases} 1 & \text{if } x_- \leq x \leq x_+, \\ \alpha & \text{otherwise.} \end{cases} \tag{3.5}$$

For any α, the interface between the PML and the physical domain is perfectly transparent, and we will see that if α is well chosen, the transmitted waves decrease exponentially in the layer.

Classically, the modes in a waveguide are given by

$$p_n^{\pm}(x,y) = e^{\pm i \beta_n x} \varphi_n(y), \quad n \in \mathbb{N}, \tag{3.6}$$

where functions φ_n, for all $n \in \mathbb{N}$, are defined by (2.3) and axial wave numbers β_n are solutions of the dispersion equation

$$\beta_n{}^2 = k^2 - \frac{n^2 \pi^2}{h^2}, \quad n \in \mathbb{N},$$

such that $\beta_n > 0$ for the propagative modes and $\mathrm{Im}(\beta_n) > 0$ for the evanescent modes. Referring to subsection 2.1, note that

$$\beta_n = \beta_n^+ = -\beta_n^-,$$

with the Mach number M taken equal to zero in the definitions (2.5) and (2.6) of $\beta_n^\pm$. In the same manner, one can define the modes in the PML as

$$p_{n,\alpha}^\pm(x,y) = e^{\pm i\beta_{n,\alpha}x}\varphi_n(y), \quad n \in \mathbb{N}, \tag{3.7}$$

with

$$\beta_{n,\alpha} = \frac{\beta_n}{\alpha}.$$

If α satisfies the hypotheses

$$\mathrm{Re}(\alpha) > 0, \quad \mathrm{Im}(\alpha) < 0, \tag{3.8}$$

then $p_{n,\alpha}^\pm$ is exponentially decreasing as $x \to \pm\infty$ for any n corresponding either to a propagative or to an evanescent mode. It is now straightforward to show that an incident mode p_n^+ generates an evanescent transmitted mode $p_{n,\alpha}^+$ in Ω_+ and no reflection at the interface Σ_+. Let us stress that assumption (3.8) is the only requirement on α to obtain a PML. Surprisingly, the fairly restrictive choice (3.2) seems to be used in most time-harmonic applications.

3.2. The new PML formulation for the convected Helmholtz equation. A natural idea for designing a PML for the convected Helmholtz equation, already used in the literature for applications in the time domain, is to apply the technique described in the previous subsection. It has been observed by several authors that this approach leads to instabilities in the time domain [19, 13, 16, 26]. The presence of instabilities have been explained in [2], thanks to an analysis via group velocities. In the context of a duct, this phenomenon can be easily understood using the modal approach.

As in the no-flow case, the axial wave numbers $\beta_{n,\alpha}^\pm$ of the modes $p_{n,\alpha}^\pm$ in the PML are given by

$$\beta_{n,\alpha}^\pm = \frac{\beta_n^\pm}{\alpha}, \quad n \in \mathbb{N}.$$

This can be illustrated by representing the $\beta_n^\pm$ and $\beta_{n,\alpha}^\pm$ in the complex plane. We clearly notice in Figures 3 and 4 that the transformation

$$\begin{array}{rccc} S_\alpha & : & \mathbb{C} & \to & \mathbb{C}, \\ & & z & \mapsto & \dfrac{z}{\alpha}, \end{array}$$

due to the change of variable used in the PML, is a similarity of ratio $\frac{1}{|\alpha|}$ and angle $\arg\left(\frac{1}{\alpha}\right) = -\arg(\alpha)$ around the origin in the complex plane. The main difference between the equation considered here and the Helmholtz equation is the possible existence of inverse upstream modes. Indeed, if p_n^+ is an inverse upstream mode (as defined in subsection 2.1), the corresponding β_n^+ is negative so that $\mathrm{Im}(\frac{\beta_n^+}{\alpha})$ becomes negative for any α satisfying assumption (3.8). This is illustrated in Figure 5, the third propagative downstream mode of the case presented being an inverse upstream

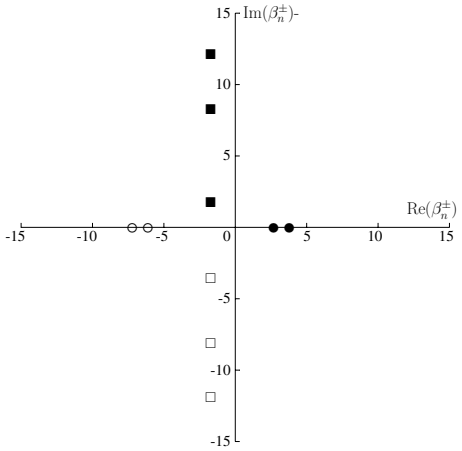


FIG. 3. *First axial wave numbers of the modes for the convected Helmholtz equation ($k = 5$, $M = 0.3$, and $h = 1$). Circles and squares are respectively associated with propagative and evanescent modes, while filled and empty symbols, respectively, refer to downstream and upstream modes.*

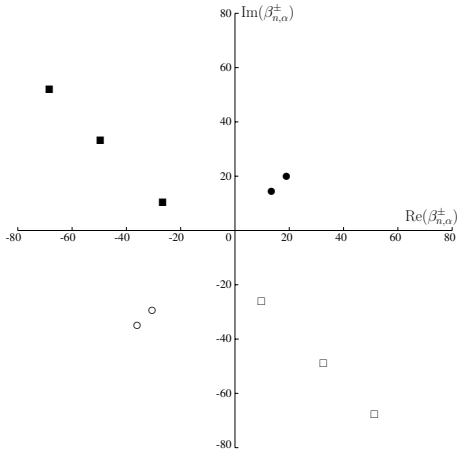


FIG. 4. *Effect of the similarity S_α ($\alpha = 0.1(1-i)$) on the first axial wave numbers of the modes for the convected Helmholtz equation ($k = 5$, $M = 0.3$, and $h = 1$).*

mode. This leads us to the conclusion that the PML model does not produce any unstable (i.e., exponentially growing in the layer) modes if all the axial wave numbers $\beta_{n,\alpha}$ for the propagative downstream (resp., upstream) modes are strictly located in the upper (resp., lower) half of the complex plane.

Guided by the previous geometrical interpretation, we apply a translation in the complex plane prior to the similarity which moves all the $\beta_{n,\alpha}^{+}$'s corresponding to the inverse upstream modes in the right half-plane and keeps the $\beta_{n,\alpha}^{-}$'s associated with propagative modes in the left one. Such a transformation is equivalent to the following substitution in (2.1):

$$\frac{\partial}{\partial x} \longrightarrow \alpha \frac{\partial}{\partial x} + i\lambda$$

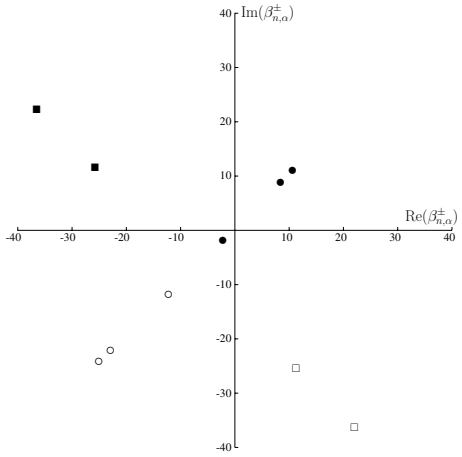


FIG. 5. *Effect of the similarity S_α ($\alpha = 0.2(1-i)$) on the first axial wave numbers of the modes for the convected Helmholtz equation in presence of an inverse upstream mode ($k = 6$, $M = 0.4$, and $h = 1$).*

with $\lambda \in \mathbb{R}$. The resulting axial wave numbers are now given by

$$\beta^{\pm}_{n,\alpha,\lambda} = \frac{\beta^{\pm}_n - \lambda}{\alpha}, \quad n \in \mathbb{N}. \tag{3.9}$$

Although λ could be chosen from among several values, the most appropriate choice is the following:

$$\lambda^* = -\frac{kM}{1-M^2}. \tag{3.10}$$

This value corresponds to the real part of the wave number of each evanescent mode, and, for any α satisfying assumption (3.8), the $\beta^{\pm}_{n,\alpha,\lambda^*}$'s are well located. Other choices for λ would require further restrictions on α in order to ensure that the β_n's associated with evanescent modes also stay in the "good side" of the complex plane (see Figure 6).

We denote in the following by $\lambda(x)$ the function defined by

$$\lambda(x) = \begin{cases} 0 & \text{if } x_- \le x \le x_+, \\ \lambda & \text{otherwise.} \end{cases} \tag{3.11}$$

Finally, the equation in the new PML medium can be written as

$$\left(1-M^2\right)\left(\alpha(x)\,\frac{\partial}{\partial x} + i\lambda(x)\right)^2 p \,+\, \frac{\partial^2 p}{\partial y^2} \,+\, 2ikM\left(\alpha(x)\,\frac{\partial}{\partial x} + i\lambda(x)\right) p \,+\, k^2 p = f \quad \text{in } \Omega, \tag{3.12}$$

where the function $\alpha(x)$ (resp., $\lambda(x)$) is defined in (3.5) (resp., in (3.11)) with $\lambda \in \mathbb{R}$ and $\alpha \in \mathbb{C}$ satisfying assumption (3.8). Writing this equation in a weak sense implies jump conditions at the interfaces between Ω_b and the layers:

$$[p(x,y)] = 0 \quad \text{and} \quad \left[\alpha(x)\,\frac{\partial p}{\partial x}(x,y) + i\lambda(x)\,p(x,y)\right] = 0. \tag{3.13}$$

Remark. This new change of variable can also be used to derive stable PMLs in the time domain, as is done in [20, 14].

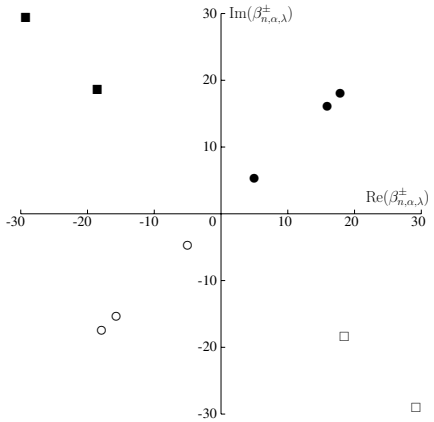


FIG. 6. *Effect of the new transformation (translation prior to similarity S_α) on the first axial wave numbers of the modes for the convected Helmholtz equation in the presence of an inverse upstream mode ($k = 6$, $M = 0.4$, and $h = 1$), $\alpha = 0.2(1 - i)$, and $\lambda = -\frac{kM}{1-M^2}$.*

4. PML truncation. Error estimates.

4.1. Truncation of the absorbing layer and well posedness. Until now, we have considered an absorbing layer of infinite length. In practice, one has to bound the computational domain and layers are of finite length L in this section.

We denote by Ω^L the truncated domain and by $\Sigma^L_\pm$ the external boundaries, presented in Figure 7. For simplicity, we choose to use homogeneous Dirichlet boundary conditions on these boundaries, but the analysis done in the following would still be valid for the natural boundary conditions $\alpha \frac{\partial p}{\partial x} + i\lambda\, p = 0$. Let p^L denote the solution in the truncated domain, satisfying

$$\begin{cases} \left(1 - M^2\right)\left(\alpha(x)\,\dfrac{\partial}{\partial x} + i\lambda(x)\right)^2 p + \dfrac{\partial^2 p}{\partial y^2} \\ \quad + 2ikM\left(\alpha(x)\,\dfrac{\partial}{\partial x} + i\lambda(x)\right) p + k^2\, p = f & \text{in } \Omega^L, \\ \dfrac{\partial p^L}{\partial y} = 0 & \text{on } \Gamma \cap \partial\Omega^L, \\ p^L = 0 & \text{on } \Sigma^L_\pm. \end{cases} \tag{4.1}$$

Denoting $V_L = \left\{q \in H^1(\Omega^L) \mid q = 0 \text{ on } \Sigma^L_\pm\right\}$, a variational formulation of (4.1) can be written as follows: find $p^L \in V_L$ such that

$$a_{\Omega^L}(p^L, q) = -\int_{\Omega^L} \frac{1}{\alpha}\, f\overline{q}\,\mathrm{d}x\,\mathrm{d}y \;\forall q \in V_L, \tag{4.2}$$

where the sesquilinear form $a_{\Omega^L}(\cdot\,,\cdot)$ is defined by

$$a_{\Omega^L}(p, q) = b_L(p, q) + c_L(p, q),$$

with

$$b_L(p, q) = \int_{\Omega^L} \left((1 - M^2)\alpha\,\frac{\partial p}{\partial x}\frac{\partial \overline{q}}{\partial x} + \frac{1}{\alpha}\,\frac{\partial p}{\partial y}\frac{\partial \overline{q}}{\partial y} + p\overline{q}\right) \mathrm{d}x\,\mathrm{d}y$$

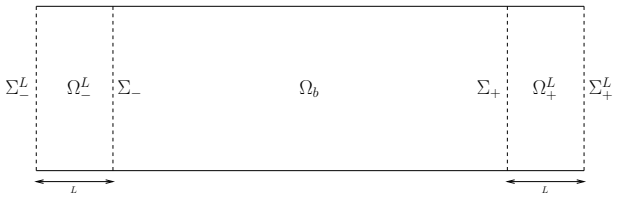


FIG. 7. *The truncated domain* Ω^L.

and

$$\begin{aligned} c_L(p,q) &= \int_{\Omega^L} i\left(\left((M^2-1)\lambda - 2kM\right)\frac{\partial p}{\partial x}\overline{q} + (1-M^2)\lambda\, p\frac{\partial \overline{q}}{\partial x}\right)\mathrm{d}x\,\mathrm{d}y \\ &\quad + \int_{\Omega^L}\left((1-M^2)\frac{\lambda^2}{\alpha} + 2kM\frac{\lambda}{\alpha} - \frac{k^2}{\alpha} - 1\right)p\overline{q}\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

THEOREM 4.1. *If* α *satisfies* (3.8)*, then problem* (4.2) *is of Fredholm type.*

Proof. The bounded operator C_L on $H^1(\Omega^L)$, defined by the Riesz representation theorem as

$$(C_L p, q)_{H^1(\Omega^L)} = c_L(p,q)\ \forall (p,q) \in H^1(\Omega^L)^2,$$

is clearly compact (from the compactness of the embedding of $H^1(\Omega^L)$ into $L^2(\Omega^L)$). On the other hand, the sesquilinear form $b_L(\cdot\,,\cdot)$ is coercive on V_L. To check this, it suffices to take the real part of $b_L(q,q)$:

$$\begin{aligned} \mathrm{Re}\,(b_L(q,q)) &= \int_{\Omega^L}\left(\mathrm{Re}(\alpha)(1-M^2)\left|\frac{\partial q}{\partial x}\right|^2 + \mathrm{Re}\left(\frac{1}{\alpha}\right)\left|\frac{\partial q}{\partial y}\right|^2 + |q|^2\right)\mathrm{d}x\,\mathrm{d}y \\ &\geq C\,\|q\|^2_{H^1(\Omega^L)}, \end{aligned}$$

where, because of assumption (3.8), C is a strictly positive constant depending on the complex constant α and the Mach number M:

$$C = \inf\left((1-M^2)\mathrm{Re}(\alpha), \mathrm{Re}\left(\frac{1}{\alpha}\right), 1\right). \qquad \square$$

4.2. Reduction to a problem posed in Ω_b. Remember that our original problem (2.1)–(2.2) has been proved in section 2.2 to be equivalent to the problem (2.10) posed in Ω_b. Having in mind the comparison between the solution p^L of problem (4.1), posed in the truncated domain, and the solution p of the original problem, we first reformulate (4.1) as a problem posed only in Ω_b. Consider the following problem: find $p_b^L \in H^1(\Omega_b)$ such that

$$\text{(4.3)} \qquad \begin{cases} (1-M^2)\,\dfrac{\partial^2 p_b^L}{\partial x^2} + 2ikM\,\dfrac{\partial p_b^L}{\partial x} + \dfrac{\partial^2 p_b^L}{\partial y^2} + k^2\,p_b^L = f & \text{in } \Omega_b, \\ \dfrac{\partial p_b^L}{\partial y} = 0 & \text{on } \Gamma \cap \partial\Omega_b, \\ \dfrac{\partial p_b^L}{\partial \boldsymbol{n}} = -T_\pm^L p_b^L & \text{on } \Sigma_\pm, \end{cases}$$

where $T_\pm^L$ are operators defined as follows:

$$\begin{aligned} T_\pm^L :\ H^{1/2}(\Sigma_\pm) &\to H^{-1/2}(\Sigma_\pm), \\ \phi &\mapsto \mp \sum_{n=0}^{+\infty} i\nu_n^\pm(L)\ (\phi,\varphi_n)_{L^2(\Sigma_\pm)}\ \varphi_n(y), \end{aligned} \tag{4.4}$$

with

$$\nu_n^\pm(L) = \beta_n^\pm + \frac{\beta_n^\mp - \beta_n^\pm}{1 - e^{i(\beta_n^- - \beta_n^+)L/\alpha}}. \tag{4.5}$$

Note that values $\nu_n^\pm(L)$ are well defined, because of assumption (2.15).

PROPOSITION 4.2. *If p^L is a solution of* (4.1)*, then $p^L{}_{|\Omega_b}$ is a solution of* (4.3)*. Conversely, if p_b^L is a solution of* (4.3)*, then it can be extended in a unique way to a solution of* (4.1)*.*

Proof. The key idea for reformulating the problem as a problem posed in Ω_b is to write an exact boundary condition satisfied by the solution on the boundaries $\Sigma_\pm$. We define the complementary domains $\Omega_\pm^L$ by

$$\Omega_-^L = \left\{(x,y)\in\Omega^L,\ x_- - L < x < x_-\right\} \quad \text{and} \quad \Omega_+^L = \left\{(x,y)\in\Omega^L,\ x_+ < x < x_+ + L\right\}.$$

Since $p_\pm^L = p^L{}_{|\Omega_\pm^L}$ satisfies a homogeneous equation in these domains, it can be given as a modal expansion. Consider, for instance, the solution in the right domain Ω_+^L. Using the Dirichlet boundary condition on the external layer boundary Σ_+^L, the solution can be written as

$$p_+^L(x,y) = \sum_{n=0}^{+\infty} \left(p_+^L(x_+,.),\varphi_n\right)_{L^2(\Sigma_+)} \left(A_n^+\, e^{i\gamma_n^+(x-x_+)} + A_n^-\, e^{i\gamma_n^-(x-x_+)}\right) \varphi_n(y),$$

where we have denoted $\gamma_n^\pm = \beta_{n,\alpha,\lambda}^\pm$, for the sake of clarity, and

$$A_n^\pm = \mp \frac{e^{i\gamma_n^\mp L}}{e^{i\gamma_n^+ L} - e^{i\gamma_n^- L}}.$$

We check easily that these quantities are always defined. Actually, the denominator would vanish if there existed an integer n for which $(\gamma_n^+ - \gamma_n^-)L \in 2\pi\mathbb{Z}$, which means that $(\beta_n^+ - \beta_n^-)L/\alpha \in 2\pi\mathbb{Z}$. If $k^2 \neq (1-M^2)\,\frac{n^2\pi^2}{h^2}$, the quantity $\beta_n^+ - \beta_n^-$ is never zero. Furthermore, with α satisfying assumption (3.8), $(\beta_n^+ - \beta_n^-)L/\alpha$ always has a nonzero imaginary part and thus cannot belong to $2\pi\mathbb{Z}$. We then write an exact boundary condition satisfied by p_+^L on Σ_+:

$$\left(\frac{\partial p_+^L}{\partial x}\right)_{|\Sigma_+} = \sum_{n=0}^{+\infty} \left(p_+^L(x_+,.),\varphi_n\right)_{L^2(\Sigma_+)} \left(A_n^+\, i\gamma_n^+ + A_n^-\, i\gamma_n^-\right) \varphi_n(y).$$

Using the jump conditions (3.13) and relation (3.9), this yields an exact boundary condition satisfied in the interior by p_b^L:

$$\begin{aligned} \left(\frac{\partial p_b^L}{\partial x}\right)_{|\Sigma_+} &= \alpha \left(\frac{\partial p_+^L}{\partial x}\right)_{|\Sigma_+} + i\lambda\, p_{+|\Sigma_+}^L \\ &= i \sum_{n=0}^{+\infty} \left(p_b^L(x_+,.),\varphi_n\right)_{L^2(\Sigma_+)} \left(A_n^+\beta_n^+ + A_n^-\beta_n^-\right) \varphi_n(y). \end{aligned}$$

Setting $\nu_n^+(L) = A_n^+\beta_n^+ + A_n^-\beta_n^-$, this can also be written as

$$\left(\frac{\partial p_b^L}{\partial x}\right)_{|\Sigma_+} = -T_+^L(p_b^L)_{|\Sigma_+},$$

where T_+^L denotes the operator defined in (4.4). ∎

Remark. It clearly appears in the expression (4.5) that operators $T_\pm^L$, and thus problem (4.3), do not depend on λ. In fact, the computed solution p^L depends on λ only in the layers.

4.3. Convergence and error estimates. We have shown that the original system (2.1)–(2.2) and system (4.1) with absorbing layers of finite length are both equivalent to problems posed only in Ω_b (respectively, (2.10) and (4.3)). We are now able to compare the solutions of these two problems, which are solutions of the following variational formulations:

For the original problem (2.10): find $p_b \in H^1(\Omega_b)$ such that

$$a_{\Omega_b}(p_b, q) = -\int_{\Omega_b} f\overline{q}\,\mathrm{d}x\,\mathrm{d}y \;\forall q \in H^1(\Omega_b), \tag{4.6}$$

where the sesquilinear form $a_{\Omega_b}(\cdot\,,\cdot)$ is given by (2.12) and can be written as

$$a_{\Omega_b}(p, q) = (Ap, q)_{H^1(\Omega_b)} + \langle T_+ p, q\rangle_{\Sigma_+} + \langle T_- p, q\rangle_{\Sigma_-}\,, \tag{4.7}$$

with A the bounded linear operator on $H^1(\Omega_b)$ defined by

$$(Ap, q)_{H^1(\Omega_b)} = \int_{\Omega_b}\left(\left(1-M^2\right)\frac{\partial p}{\partial x}\frac{\partial \overline{q}}{\partial x} + \frac{\partial p}{\partial y}\frac{\partial \overline{q}}{\partial y} - 2ikM\,\frac{\partial p}{\partial x}\overline{q} - k^2\,p\overline{q}\right)\mathrm{d}x\,\mathrm{d}y. \tag{4.8}$$

For the problem with absorbing layers of finite length (4.3): find $p_b^L \in H^1(\Omega_b)$ such that

$$a_{\Omega_b}^L(p_b^L, q) = -\int_{\Omega_b} f\overline{q}\,\mathrm{d}x\,\mathrm{d}y \;\forall q \in H^1(\Omega_b), \tag{4.9}$$

where the sesquilinear form $a_{\Omega_b}^L(\cdot\,,\cdot)$ can be written as

$$a_{\Omega_b}^L(p, q) = (Ap, q)_{H^1(\Omega_b)} + \left\langle T_+^L p, q\right\rangle_{\Sigma_+} + \left\langle T_-^L p, q\right\rangle_{\Sigma_-}\,, \tag{4.10}$$

the operator A being defined in (4.8).

To prove convergence and get error estimates, we follow an idea developed in [25], which has also been used in [12].

LEMMA 4.3. *Suppose that assumptions* (2.15) *and* (3.8) *hold. Then there exist strictly positive constants* $\mathcal{C} = \mathcal{C}(k, M)$ *and* $\eta = \eta(\theta, k, h, M)$ *(where* θ *denotes the argument of* α*) such that, for all* $(p, q) \in \left(H^1(\Omega_b)\right)^2$*, we have*

$$\left|a_{\Omega_b}(p, q) - a_{\Omega_b}^L(p, q)\right| \le \mathcal{C}e^{-\eta L/|\alpha|}\,\|p\|_{H^1(\Omega_b)}\,\|q\|_{H^1(\Omega_b)}\,. \tag{4.11}$$

More precisely, the constant η *is determined by*

$$\eta = \frac{2k}{1-M^2}\,\min\left(-\sin(\theta)\sqrt{1-\frac{{N_0}^2}{{K_0}^2}},\ \cos(\theta)\sqrt{\frac{(N_0+1)^2}{{K_0}^2}-1}\right), \tag{4.12}$$

where K_0 *is defined in* (2.4).

Proof. From expressions (4.7) and (4.10), we have

$$a_{\Omega_b}(p,q) - a^L_{\Omega_b}(p,q) = \left(\langle T_+ p, q\rangle_{\Sigma_+} - \langle T^L_+ p, q\rangle_{\Sigma_+}\right) + \left(\langle T_- p, q\rangle_{\Sigma_-} - \langle T^L_- p, q\rangle_{\Sigma_-}\right).$$

Let us focus on the first term in the right-hand side, the estimation of the second one being analogous. From the definitions (2.8) and (4.4) of operators T_+ and T^L_+, we have, for any $\phi \in H^{1/2}(\Sigma_+)$,

$$\left(T_+ - T^L_+\right)\phi = -\sum_{n=0}^{+\infty} i\left(\beta^+_n - \nu^+_n(L)\right)\,\phi_n\,\varphi_n(y) \quad \text{with } \phi_n = (\phi,\varphi_n)_{L^2(\Sigma_+)}.$$

Therefore, for any $(\phi,\psi) \in (H^{1/2}(\Sigma_+))^2$,

$$\left\langle (T_+ - T^L_+)\phi, \psi\right\rangle_{\Sigma_+} = -\sum_{n=0}^{+\infty} i(\beta^+_n - \nu^+_n(L))\,\phi_n\,\overline{\psi}_n,$$

with $\phi_n = (\phi,\varphi_n)_{L^2(\Sigma_+)}$ and $\psi_n = (\psi,\varphi_n)_{L^2(\Sigma_+)}$. This implies the following estimate:

$$\left|\left\langle \left(T_+ - T^L_+\right)\phi, \psi\right\rangle_{\Sigma_+}\right| \le \sum_{n=0}^{+\infty} \left|\beta^+_n - \nu^+_n(L)\right|\,\left|\phi_n\overline{\psi}_n\right|. \tag{4.13}$$

From (4.5), we have

$$\left|\beta^+_n - \nu^+_n(L)\right| = \frac{|\beta^+_n - \beta^-_n|}{\left|1 - e^{i(\beta^-_n - \beta^+_n)L/\alpha}\right|}.$$

Noticing that, for any $z \in \mathbb{C}$,

$$\left|1 - e^{iz}\right| \ge \left|e^{-\mathrm{Im}(z)} - 1\right|$$

so that, if $\mathrm{Im}(z) < 0$ and $|\mathrm{Im}(z)|$ is large enough, we conclude that this quantity is larger than

$$\left|1 - e^{iz}\right| \ge \left|e^{-\mathrm{Im}(z)} - 1\right| \ge \frac{1}{2}\,e^{-\mathrm{Im}(z)}.$$

We can easily check, using assumption (3.8), that we have $\mathrm{Im}\left((\beta^-_n - \beta^+_n)\,L/\alpha\right) < 0$ for all $n \in \mathbb{N}$ so that, for L large enough, the previous estimate gives us

$$\left|\beta^+_n - \nu^+_n(L)\right| \le 2\left|\beta^+_n - \beta^-_n\right|\,e^{\mathrm{Im}\left((\beta^-_n - \beta^+_n)L/\alpha\right)}. \tag{4.14}$$

Let us now distinguish the two cases.

The propagative modes $n \le N_0$. From (2.5), we have

$$\beta^+_n - \beta^-_n = \frac{2k}{1-M^2}\sqrt{1 - \frac{n^2}{{K_0}^2}} = \delta_n > 0.$$

Noting that

$$\delta_{N_0} \le \delta_n \le \frac{2k}{1-M^2},$$

we derive from estimate (4.14)

$$\left|\beta^+_n - \nu^+_n(L)\right| \le 2\delta_n\,e^{-\delta_{N_0} L\,\mathrm{Im}(1/\alpha)} \le \frac{4k}{1-M^2}\,e^{-\delta_{N_0} L\,\mathrm{Im}(1/\alpha)}. \tag{4.15}$$

The evanescent modes $n \geq N_0 + 1$. From (2.6), we have

$$\beta_n^+ - \beta_n^- = \frac{2ik}{1-M^2}\sqrt{\frac{n^2}{{K_0}^2} - 1} = i\delta_n, \quad \delta_n > 0.$$

This time, δ_n is increasing and $\sqrt{\frac{n^2}{{K_0}^2} - 1} \leq \frac{n}{K_0}$. Estimate (4.14) thus yields

$$\left|\beta_n^+ - \nu_n^+(L)\right| \leq \frac{4k}{1-M^2}\frac{n}{K_0}\, e^{-\delta_{N_0+1} L \,\mathrm{Re}(1/\alpha)}. \tag{4.16}$$

By substituting these estimates into (4.13), we see that

$$\begin{aligned}\left|\left\langle \left(T_+ - T_+^L\right)\phi, \psi\right\rangle_{\Sigma_+}\right| \;\leq\; & \frac{4k}{1-M^2}\left(\sum_{n=0}^{N_0} e^{-\delta_{N_0} L \,\mathrm{Im}(1/\alpha)}\, \left|\phi_n\, \overline{\psi}_n\right| \right. \\ & \left. + \sum_{n=N_0+1}^{+\infty} \frac{n}{K_0}\, e^{-\delta_{N_0+1} L \,\mathrm{Re}(1/\alpha)}\, \left|\phi_n \overline{\psi}_n\right|\right).\end{aligned}$$

Setting $\eta = |\alpha| \min(\delta_{N_0} \,\mathrm{Im}(1/\alpha), \delta_{N_0+1}\,\mathrm{Re}(1/\alpha))$, we then have

$$\begin{aligned}\left|\left\langle \left(T_+ - T_+^L\right)\phi, \psi\right\rangle_{\Sigma_+}\right| \;&\leq\; \frac{4k}{1-M^2}\, e^{-\eta L/|\alpha|} \sum_{n=0}^{+\infty}\left(1 + \frac{n^2}{{K_0}^2}\right)^{1/2} \left|\phi_n\, \overline{\psi}_n\right| \\ &\leq\; \mathcal{C}\, e^{-\eta L/|\alpha|}\, |\phi|_{H^{1/2}(\Sigma_+)}\, |\psi|_{H^{1/2}(\Sigma_+)}.\end{aligned}$$

The trace theorem now yields, for any $(p,q) \in \left(H^1(\Omega_b)\right)^2$,

$$\left|\left\langle (T_+ - T_+^L)p, q\right\rangle_{\Sigma_+}\right| \leq \mathcal{C}\, e^{-\eta L/|\alpha|}\, \|p\|_{H^1(\Omega_b)}\, \|q\|_{H^1(\Omega_b)}.$$

One can obviously obtain the same estimate on Σ_- and thus conclude the proof of claim (4.11). ∎

Setting $V = H^1(\Omega_b)$, we introduce linear operators $\mathcal{A}$ and $\mathcal{A}^L$ in $\mathcal{L}(V, V')$, respectively, associated with the sesquilinear forms $a_{\Omega_b}(\cdot\,,\cdot)$ and $a^L_{\Omega_b}(\cdot\,,\cdot)$: for all $(p,q) \in V^2$,

$$\langle \mathcal{A}p, q\rangle_{V',V} = a_{\Omega_b}(p,q) \quad \text{and} \quad \left\langle \mathcal{A}^L p, q\right\rangle_{V',V} = a^L_{\Omega_b}(p,q).$$

Obviously estimate (4.11) implies

$$\left\|\mathcal{A} - \mathcal{A}^L\right\|_{\mathcal{L}(V,V')} \leq \mathcal{C}\, e^{-\eta L/|\alpha|}. \tag{4.17}$$

Problems (4.6) and (4.9) can both be written in terms of these operators:

$$\mathcal{A}p_b = -f, \tag{4.18}$$

$$\mathcal{A}^L p_b^L = -f. \tag{4.19}$$

It follows from taking the difference between (4.18) and (4.19) that the error $p_b - p_b^L$ satisfies the following equation:

$$\mathcal{A}^L\left(p_b - p_b^L\right) = \left(\mathcal{A}^L - \mathcal{A}\right) p_b. \tag{4.20}$$

Using estimate (4.17), we are now able to show the following result.

THEOREM 4.4. *Suppose that assumptions* (3.8) *and* (2.15) *hold. There exists* $L_1 > 0$ *such that for all* $L \geq L_1$, $\mathcal{A}^L$ *is an isomorphism on* $H^1(\Omega_b)$ *and the solution* p_b^L *of problem* (4.3) *converges to the solution* p_b *of problem* (2.10). *Furthermore, there exists a constant* $\mathcal{C}$ *depending on* M *and* k *such that*

$$\left\|p_b - p_b^L\right\|_V \leq \mathcal{C}\, e^{-\eta L/|\alpha|} \left\|p_b\right\|_V , \tag{4.21}$$

with η *being defined in* (4.12).

Proof. For $g \in V'$, we consider the following problem: find $u \in V$ such that

$$\mathcal{A}^L u = g. \tag{4.22}$$

We can rewrite the operator $\mathcal{A}^L$ as $\mathcal{A}^L = \mathcal{A} + \left(\mathcal{A}^L - \mathcal{A}\right)$ and, using that $\mathcal{A}$ is an isomorphism on V,

$$\mathcal{A}^L = \mathcal{A}\left(I + \mathcal{A}^{-1}\left(\mathcal{A}^L - \mathcal{A}\right)\right).$$

Problem (4.22) thus becomes

$$\left(I + \mathcal{A}^{-1}\left(\mathcal{A}^L - \mathcal{A}\right)\right) u = \mathcal{A}^{-1} g.$$

Applying the Banach fixed point theorem, this problem admits a unique solution if

$$\left\|\mathcal{A}^{-1}\left(\mathcal{A}^L - \mathcal{A}\right)\right\|_{\mathcal{L}(V,V')} < 1,$$

which is satisfied as soon as

$$\left\|\mathcal{A}^L - \mathcal{A}\right\|_{\mathcal{L}(V,V')} < \left\|\mathcal{A}^{-1}\right\|^{-1}_{\mathcal{L}(V,V')}.$$

This can be achieved for L large enough, that is, $L \geq L_1$, since $\left\|\mathcal{A}^L - \mathcal{A}\right\|_{\mathcal{L}(V,V')}$ tends to zero as L tends to infinity, because of (4.17). Moreover, we have

$$\left\|\left(I + \mathcal{A}^{-1}\left(\mathcal{A}^L - \mathcal{A}\right)\right)^{-1}\right\|_{\mathcal{L}(V,V')} < \frac{1}{1 - \left\|\mathcal{A}^{-1}\left(\mathcal{A}^L - \mathcal{A}\right)\right\|_{\mathcal{L}(V,V')}},$$

which implies the following estimate:

$$\left\|u\right\|_V < \frac{\left\|\mathcal{A}^{-1} g\right\|_V}{1 - \left\|\mathcal{A}^{-1}(\mathcal{A}^L - \mathcal{A})\right\|_{\mathcal{L}(V,V')}}.$$

Applying this result to the error, the solution of problem (4.20) yields

$$\begin{aligned}\left\|p_b - p_b^L\right\|_V &< \frac{\left\|\mathcal{A}^{-1}\left(\mathcal{A}^L - \mathcal{A}\right) p_b\right\|_V}{1 - \left\|\mathcal{A}^{-1}\left(\mathcal{A}^L - \mathcal{A}\right)\right\|_{\mathcal{L}(V,V')}} \\ &\leq \frac{\left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(V,V')} \left\|\mathcal{A}^L - \mathcal{A}\right\|_{\mathcal{L}(V,V')} \left\|p_b\right\|_V}{1 - \left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(V,V')} \left\|\mathcal{A}^L - \mathcal{A}\right\|_{\mathcal{L}(V,V')}}.\end{aligned}$$

When $\left\|\mathcal{A}^L - \mathcal{A}\right\|_{\mathcal{L}(V,V')}$ is small enough, the quantity in the right-hand side can be bounded by

$$\left\|p_b - p_b^L\right\|_V \leq 2 \left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(V,V')} \left\|\mathcal{A}^L - \mathcal{A}\right\|_{\mathcal{L}(V,V')} \left\|p_b\right\|_V \leq 2\mathcal{C}\, e^{-\eta} \left\|p_b\right\|_V ,$$

the last inequality coming from (4.17). ∎

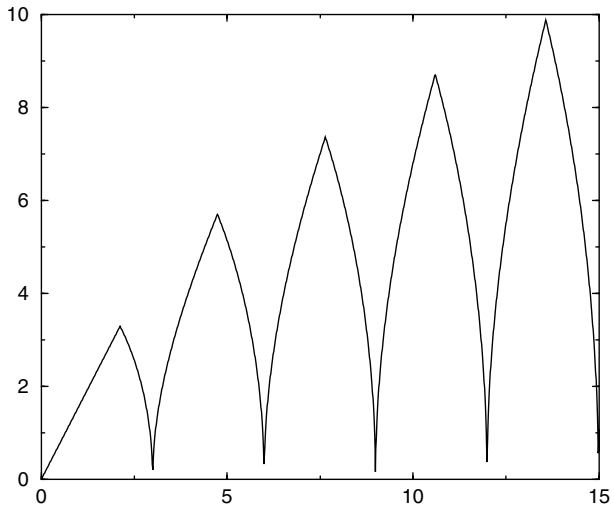


FIG. 8. *Coefficient η plotted as a function of wave number k for $M = 0.3$, $h = 1$, and $\theta = -\frac{\pi}{4}$.*

Remarks. 1. We emphasize that error estimate (4.21) of Theorem 4.4 does not depend on the parameter λ. As a consequence, exponential convergence is obtained for both the classical and the new PML models. This is the main difference between the behavior of the new PMLs and that of the classical PMLs in the time domain, as in that case the layers lead to instabilities in the presence of inverse upstream modes [19, 26, 1, 2].

2. Note that estimate (4.21) also proves that convergence holds when the length of the layers L is fixed and $|\alpha|$ tends to 0. This is useful in practice for numerical computations. Indeed, L has to be small in order to reduce the number of degrees of freedom. Moreover, it is more convenient to change the value of parameter α than the length of the layers, which requires a new mesh of the computational domain.

3. The value of η is strongly related to the position of wave number k with respect to the cut-off frequencies. More precisely, for a given value of the argument θ of coefficient α, the accuracy deteriorates when k is close to a cut-off wave number (see Figure 8).

5. Varying coefficients. In practical computations, it is very common to use a spatially varying coefficient $\alpha(x)$ in the layers. Actually, it has been proven for finite difference schemes that discontinuities in α through the boundaries $\Sigma_\pm$ generate spurious reflections after discretization [10]. In this section, we show that the analysis done previously for constant coefficients α and λ can be easily extended to varying coefficients. Let us point out, however, that the numerical results presented in the next section are obtained with constant coefficients and that no significant effects due to the discontinuities have been observed.

Let α and λ be two functions of the coordinate x, defined from $\mathbb{R}$ to $\mathbb{C}$, such that

$$\alpha(x) = 1 \quad \text{and} \quad \lambda(x) = 0 \quad \text{for } x \in [x_-, x_+].$$

We assume, moreover, that $\alpha(x)$ satisfies (3.8) for all $x > x_+$ or $x < x_-$. Let us consider once more problem (4.1). Since the proof of Theorem 4.1 does not use the fact that α and λ are constant in the layers, the theorem still holds, and the problem is of Fredholm type.

To establish a convergence result with respect to the size of the layers, we follow the steps of sections 4.2 and 4.3. The main point is that the modal solutions in the right-hand side layer, for instance, are now

$$p_n^{\pm}(x,y) = \psi_n^{\pm}(x)\varphi_n(y), \quad n \in \mathbb{N},$$

where φ_n is given by (2.3) and $\psi_n^{\pm}$ is defined by

$$\begin{cases} \left(\alpha(x)\dfrac{d}{dx} + i\lambda(x)\right)\psi_n^{\pm} = i\beta_n^{\pm}\psi_n^{\pm}, \\ \psi_n^{\pm}(x_+) = 1. \end{cases}$$

It then follows from straightforward calculations that problem (4.1) is equivalent to problem (4.3) set in domain Ω_b, with the following new definition for the coefficient $\nu_n^+(L)$:

$$\nu_n^+(L) = \beta_n^+ + \frac{\beta_n^- - \beta_n^+}{1 - \dfrac{\psi_n^-(x_+ + L)}{\psi_n^+(x_+ + L)}}. \tag{5.1}$$

and a similar definition for $\nu_n^-(l)$. Note, moreover, that

$$\frac{\psi_n^-(x_+ + L)}{\psi_n^+(x_+ + L)} = e^{i(\beta_n^- - \beta_n^+)} \int_{x_+}^{x_+ + L} \frac{1}{\alpha(x)}\,\mathrm{d}x$$

so that again the coefficients $\nu_n^{\pm}(L)$ do not depend on λ.

The final result then reads as follows.

THEOREM 5.1. *Problem* (4.1) *is well posed, and the solution* p_b^L *of problem* (4.3)*, with* $\nu_n^{\pm}(L)$ *defined above, exists and converges to the solution* p_b *of problem* (2.10) *as* $L \to +\infty$*. Furthermore, there exist three constants* $\mathcal{C} = \mathcal{C}(k, M)$ *and* $\tau_{\pm} = \tau_{\pm}(k, M, \alpha, L)$ *such that*

$$\left\|p_b - p_b^L\right\|_V \leq \mathcal{C}\,\left(e^{-\tau_+} + e^{-\tau_-}\right)\|p_b\|_V\,,$$

with

$$\tau_{\pm} = \frac{2k}{1 - M^2}\,\min\left(\mathrm{Im}(I_{\pm})\sqrt{1 - \frac{{N_0}^2}{{K_0}^2}},\ \mathrm{Re}(I_{\pm})\sqrt{\frac{(N_0+1)^2}{{K_0}^2} - 1}\right),$$

where $I\pm = \pm\int_{x_\pm}^{x_\pm \pm L} \frac{1}{\alpha(x)}\,\mathrm{d}x$*,* K_0 *is defined in* (2.4)*, and* N_0 *denotes the integer part of* K_0*.*

6. Numerical results. In order to illustrate the conclusions previously drawn concerning the PML models, numerical examples are presented. The following configuration is considered: the computational domain is the same as the one presented in Figure 7, extending from $x = -0.2$ to $x = 2.2$ and $y = 0$ to $y = 1$. The layers occupy the region from $x = -0.2$ to $x = 0$ in the downstream direction and from $x = 2$ to $x = 2.2$ in the upstream direction, the thickness L of the layers then being fixed and equal to 10% of the length of domain Ω_b. A compactly supported source f is given by

$$f = \begin{cases} 1 \text{ if } (x-1)^2 + (y-0.7)^2 \leq 0.04, \\ 0 \text{ elsewhere.} \end{cases}$$

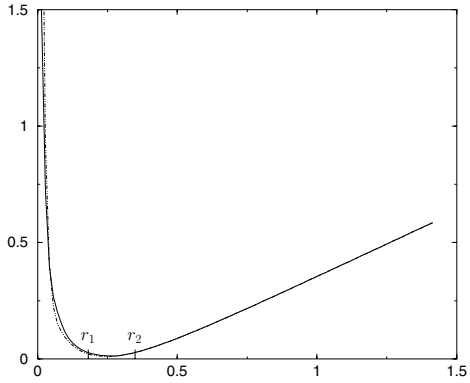


FIG. 9. *Relative error* $\frac{\|p-p_{\rm ref}\|_{H^1(\Omega_b)}}{\|p_{\rm ref}\|_{H^1(\Omega_b)}}$ *as a function of* $|\alpha|$, $k = 10$, *and* $M = 0.3$. *The solid line is the result for the new PML model, while the dotted line refers to the classical PML model.*

The numerical solution of problem (4.2) posed in the domain bounded with PMLs is compared to the computed solution (which is called the *reference solution*) of problem (2.11) posed in the domain bounded with DtN operators. Both approximations are done with a finite element method. The DtN map, usually expressed through an infinite series expansion, is here approximated by truncating the series.

All the simulations have been conducted with the same unstructured mesh, whose mesh size is linked to the problem via a resolution of approximately 20 nodal points per wavelength when using second-order triangular (P_2) Lagrange finite elements. For the computation of the reference solution, the number of terms in the truncated DtN map is 8, which is sufficient for accuracy in each of the cases tested. The coefficient α is chosen to be a complex constant in the layers, whose argument is taken to be equal to $-\frac{\pi}{4}$, and coefficient λ of the new PML model takes the value λ^* defined in (3.10). Homogeneous Dirichlet boundary conditions are imposed on the outer boundaries of the layers. The computations are done with the finite element library MÉLINA [22].

6.1. The no inverse upstream mode case. In this first simulation, we choose $k = 10$ and $M = 0.3$. For such values of the wave number and the Mach number, four modes are propagative, and there is no inverse upstream mode. The solution in the layers is then exponentially decaying for both the classical (i.e., $\lambda = 0$) and new PML models.

In Figure 9, the relative error to the reference solution in the $H^1(\Omega_b)$ norm is plotted as a function of the modulus of α for the two models. We observe no noticeable discrepancy between the classical and new PML models, which behave similarly in this case. Both curves present a minimum plateau, and we can roughly distinguish three zones, as indicated in Figure 9:

• $|\alpha| \in [r_1, r_2]$. A very good agreement between the DtN and the PML solutions, for the classical and new PML models, is obtained for a large range of values of $|\alpha|$ corresponding to the plateau seen in Figure 9. The real part of the corresponding solution is shown in Figure 10. We also observe in this figure the effect of the flow on the propagation of sound, as the wavelength of the solution is longer downstream from the source than upstream.

• $|\alpha| > r_2$. For larger values of $|\alpha|$, the layer is insufficiently absorbing, and a reflection occurs at the end of the layers, as shown in Figure 11.

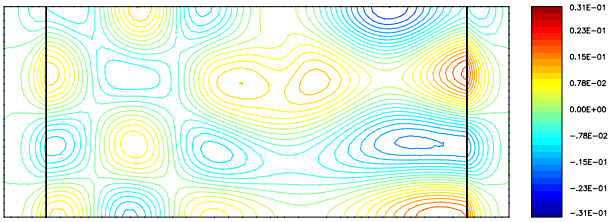


FIG. 10. *Real part of the pressure field;* $k = 10$ *and* $M = 0.3$, $\alpha = .19(1 - i)$, *and* $\lambda = -\frac{kM}{1-M^2}$.

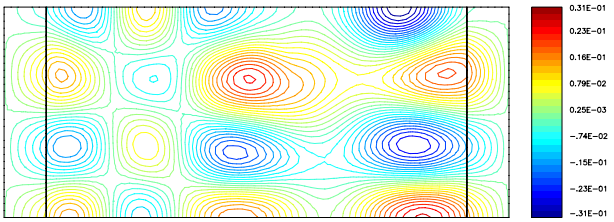


FIG. 11. *Real part of the pressure field;* $k = 10$ *and* $M = 0.3$, $\alpha = 1 - i$, *and* $\lambda = -\frac{kM}{1-M^2}$.

• $|\alpha| < r_1$. For small values of $|\alpha|$, the absorption in the layers is high, but the mesh resolution becomes too coarse to correctly represent modes in the PML medium, thus producing spurious numerical errors, as seen in Figure 12.

We want to confirm the convergence estimate of Theorem 4.4, which implies that

$$-\ln\left(\frac{\left\|p_b - p_b^L\right\|_V}{\|p_b\|_V}\right) \geq \frac{\eta L}{|\alpha|} - \ln(\mathcal{C}).$$

To this end, the opposite of the logarithm of the relative error in $H^1(\Omega_b)$ norm is plotted as a function of the inverse of $|\alpha|$ for both PML models. The exponential convergence of the method, which can be deduced from the slope of curves in Figure 13, agrees satisfactorily with the estimation given by the theory for both PML models, as the two curves coincide for this case.

6.2. The inverse upstream mode case. For the choice of $k = 9$ and $M = 0.4$, the last of the four propagative downstream modes (i.e., $n = 3$) has a negative phase velocity and is therefore called an inverse upstream mode. The solution in the downstream layer is then exponentially decreasing or increasing with the distance, depending on the applied PML model. Results for this case are shown in Figures 14 and 15, where the relative error in the $H^1(\Omega_b)$ norm and the opposite of its logarithm, respectively, are shown.

As one can observe in the zoom in Figure 14, the curves of the relative error have again a minimum plateau for both PML models. This time, the size of the plateau is smaller for the classical PML and the error for this model has a rather erratic behavior for small values of $|\alpha|$. The convergence of the method is nonetheless achieved for both models, with the predicted exponential rate (see Figure 15). However, the new PML model seems better suited to practical computations, as one can choose an appropriate and optimal value of α for convergence more conveniently.

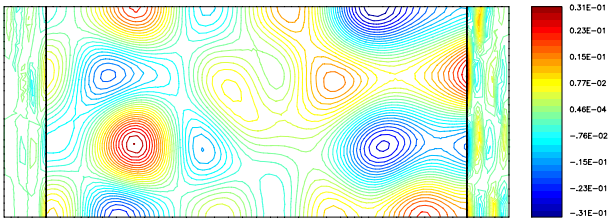


FIG. 12. *Real part of the pressure field;* $k = 10$ *and* $M = 0.3$, $\alpha = .02(1 - i)$, *and* $\lambda = -\frac{kM}{1-M^2}$.

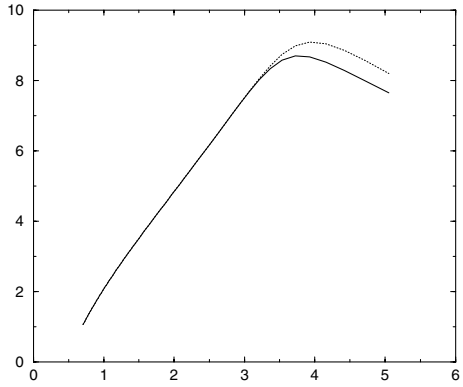


FIG. 13. $-\ln(\frac{\|p-p_{\text{ref}}\|_{H^1(\Omega_b)}}{\|p_{\text{ref}}\|_{H^1(\Omega_b)}})$ *as a function of* $\frac{1}{|\alpha|}$, $k = 10$, *and* $M = 0.3$. *The solid and dotted lines, respectively, refer to results for the new and classical PML models.*

To conclude, Figures 16 and 17, respectively, show the solutions for the new and classical models, the value of $|\alpha|$ for this case corresponding to the minimum of the curves in Figure 14. Note that whatever the behavior of the solution in the layers, the solution in the "physical" domain remains almost the same.

6.3. Some practical remarks on the use of PMLs for time-harmonic problems. We would like to point out that the numerical analysis which has been carried out in this section was based on the knowledge of a reference solution. In practice, it would be useful to have a posteriori criteria which indicate whether the numerical solution is satisfactory or not. In transient applications, the quality of the PML model is ensured as soon as the reflections produced at the interface between the physical and the absorbing layer can be neglected. In particular, if the excitation is a pulse localized in time, the exact solution should vanish after a large time, which gives a criterion for evaluating the efficiency of the absorbing layer. The situation is completely different in time-harmonic applications. For instance, we have the following:

- The notion of reflection is more difficult to exploit: as illustrated in the previous numerical results, it is not clear how to distinguish a "reflected" wave from an "incident" wave.
- The experiment of a pulse localized in time has no counterpart in time-harmonic applications.
- A good choice of the absorbing layer parameters allows one to select the outgoing solution of the problem. A bad choice ($\text{Im}(\alpha) > 0$) would select

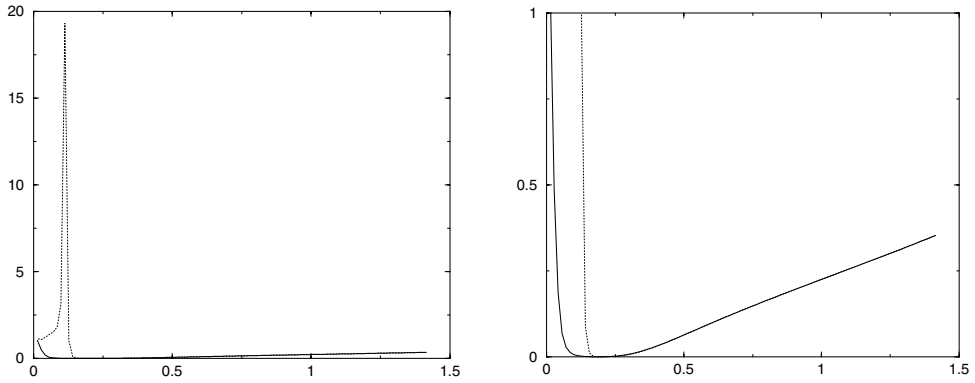


FIG. 14. *Left: relative error* $\frac{\|p-p_{\rm ref}\|_{H^1(\Omega_b)}}{\|p_{\rm ref}\|_{H^1(\Omega_b)}}$ *as a function of* $|\alpha|$, $k = 9$, *and* $M = 0.4$*; right: zoom on the zone of interest. The solid line is the result for the new PML model, while the dotted line refers to the classical PML model.*

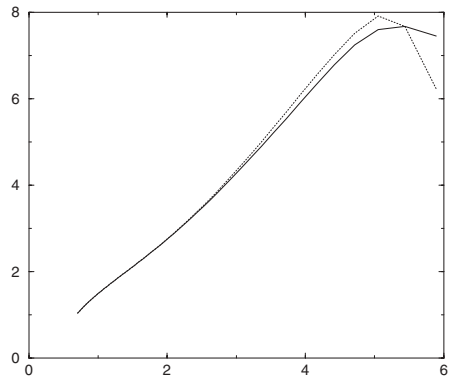


FIG. 15. $-\ln(\frac{\|p-p_{\rm ref}\|_{H^1(\Omega_b)}}{\|p_{\rm ref}\|_{H^1(\Omega_b)}})$ *as a function of* $\frac{1}{|\alpha|}$, $k = 9$, *and* $M = 0.4$. *The solid and dotted lines, respectively, refer to results for the new and classical PML models.*

the ingoing solution, which is difficult to detect when one does not know the exact solution.

- However, one can note that when $|\alpha|$ is too small, spurious numerical errors are observed. They can be removed by refining the mesh in the layer.

7. Conclusion. In this paper, we have studied PMLs for the convected Helmholtz equation. In the presence of inverse upstream modes, the solution can have arbitrarily large values in the classical PMLs, thus causing the instabilities observed in time domain applications. We have investigated a new PML model which always leads to an exponentially decreasing solution in the layer, even in the presence of inverse upstream modes. The error analysis surprisingly showed the convergence for both the classical and new models. Nevertheless, numerical results seem to indicate that the error is best controlled with the new model when inverse upstream modes are present. In order to understand the different numerical behaviors of the two models, there remains to analyze the convergence of the solution of the discretized PML models with respect to both the finite element mesh size and the layer parameters α and L.

This is a preliminary step in dealing with more complex time-harmonic problems.

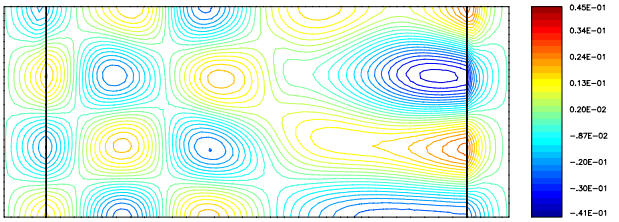


FIG. 16. *Real part of the pressure field;* $k = 9$ *and* $M = 0.4$, $\alpha = .14(1 - i)$, *and* $\lambda = -\frac{kM}{1-M^2}$.

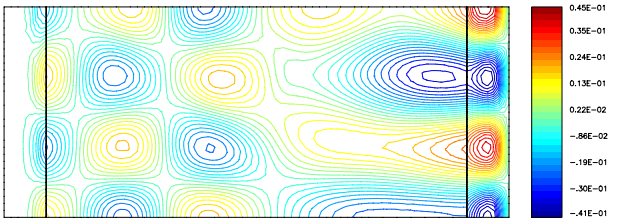


FIG. 17. *Real part of the pressure field;* $k = 9$ *and* $M = 0.4$, $\alpha = .14(1 - i)$, *and* $\lambda = 0$.

In particular, it would be interesting to extend the present method to nonuniform flows. This gives rise to several difficulties. First, even for a parallel flow, the problem can no longer be reduced to a simple scalar equation and has to be modeled with a vectorial model, for instance linearized Euler equations or Galbrun's equation [7]. Furthermore, a modal analysis cannot be done so easily, since the orthogonality of the modes is lost and their completeness is an open question. Finally, for some flows, there exist physical outgoing unstable modes which have to be adequately treated by the absorbing model.

REFERENCES

[1] S. ABARBANEL, D. GOTTLIEB, AND J. S. HESTHAVEN, *Well-posed perfectly matched layers for advective acoustics*, J. Comput. Phys., 154 (1999), pp. 266–283.
[2] E. BÉCACHE, S. FAUQUEUX, AND P. JOLY, *Stability of perfectly matched layers, group velocities and anisotropic waves*, J. Comput. Phys., 188 (2003), pp. 399–433.
[3] J.-P. BÉRENGER, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.
[4] J.-P. BÉRENGER, *Three-dimensional perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 127 (1996), pp. 363–379.
[5] J.-P. BÉRENGER, *Improved PML for the FDTD solution of wave-structure interaction problems*, IEEE Trans. Antennas and Propagation, 45 (1997), pp. 466–473.
[6] A.-S. BONNET-BEN DHIA, L. DAHI, E. LUNÉVILLE, AND V. PAGNEUX, *Acoustic diffraction by a plate in a uniform flow*, Math. Models Methods Appl. Sci., 12 (2002), pp. 625–647.
[7] A.-S. BONNET-BEN DHIA, G. LEGENDRE, AND E. LUNÉVILLE, *Analyse mathématique de l'équation de Galbrun en écoulement uniforme*, C. R. Acad. Sci. Paris Sér. IIb Méc., 329 (2001), pp. 601–606.
[8] W. C. CHEW AND W. H. WEEDON, *A 3D perfectly matched medium from modified Maxwell's equations with stretched coordinates*, IEEE Microwave Opt. Technol. Lett., 7 (1994), pp. 599–604.
[9] F. COLLINO AND P. MONK, *The perfectly matched layer in curvilinear coordinates*, SIAM J. Sci. Comput., 19 (1998), pp. 2061–2090.
[10] F. COLLINO AND P. B. MONK, *Optimizing the perfectly matched layer*, Comput. Methods Appl. Mech. Engrg., 164 (1998), pp. 157–171.
[11] J. DIAZ AND P. JOLY, *Stabilized Perfectly Matched Layers for Advective Wave Equations*,

manuscript.

[12] D. Gómez Pedreira and P. Joly, *A method for computing guided waves in integrated optics. Part* II: *Numerical approximation and error analysis*, SIAM J. Numer. Anal., 39 (2002), pp. 1684–1711.

[13] J. W. Goodrich and T. Hagstrom, *A Comparison of Two Accurate Boundary Treatments for Computational Aeroacoustics*, AIAA paper 97-1585, 1997.

[14] T. Hagstrom and I. Nazarov, *Absorbing layers and radiation boundary conditions for jet flow simulations*, in Proceedings of the 8th AIAA/CEAS Aeroacoustics Conference, Breckenridge, CO, 2002, AIAA paper 2002-2606.

[15] I. Harari, M. Slavutin, and E. Turkel, *Analytical and numerical studies of a finite element PML for the Helmholtz equation*, J. Comput. Acoust., 8 (2000), pp. 121–137.

[16] J. S. Hesthaven, *On the analysis and construction of perfectly matched layers for the linearized Euler equations*, J. Comput. Phys., 142 (1998), pp. 129–147.

[17] P. D. Hislop and I. M. Sigal, *Introduction to Spectral Theory with Applications to Schrödinger Operators*, Appl. Math. Sci. 113, Springer-Verlag, New York, 1996.

[18] T. Hohage, F. Schmidt, and L. Zschiedrich, *Solving Time-Harmonic Scattering Problems Based on the Pole Condition: Convergence of the PML Method*, Tech. report 01-23, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Berlin, Germany, 2001.

[19] F. Q. Hu, *On absorbing boundary conditions for linearized Euler equations by a perfectly matched layer*, J. Comput. Phys., 129 (1996), pp. 201–219.

[20] F. Q. Hu, *A stable, perfectly matched layer for linearized Euler equations in unsplit physical variables*, J. Comput. Phys., 173 (2001), pp. 455–480.

[21] M. Lassas and E. Somersalo, *On the existence and convergence of the solution of PML equations*, Computing, 60 (1998), pp. 229–241.

[22] D. Martin, *On line documentation of* mélina, http://perso.univ-rennes1.fr/daniel.martin/melina/www/homepage.html.

[23] P. G. Petropoulos, *Reflectionless sponge layers as absorbing boundary conditions for the numerical solution of Maxwell equations in rectangular, cylindrical, and spherical coordinates.*, SIAM J. Appl. Math., 60 (2000), pp. 1037–1058.

[24] C. M. Rappaport, *Perfectly matched absorbing conditions based on anisotropic lossy mapping of space*, IEEE Microwave Guided Wave Lett., 5 (1995), pp. 90–92.

[25] J. Razafiarivelo, *Optimisation de formes en électromagnétisme*, Ph.D. thesis, Université de Paris VI, Paris, France, 1996.

[26] C. K. W. Tam, L. Auriault, and F. Cambuli, *Perfectly matched layer as an absorbing boundary condition for the linearized Euler equations in open and ducted domains*, J. Comput. Phys., 144 (1998), pp. 213–234.

[27] E. Turkel and A. Yefet, *Absorbing PML boundary layers for wave-like equations*, Appl. Numer. Math., 27 (1998), pp. 533–557.

[28] L. Zhao and A. C. Cangellaris, *GT-PML: Generalized theory of perfectly matched layers and its application to the reflectionless truncation of finite-difference time-domain grids*, IEEE Trans. Microwave Theory Tech., 44 (1996), pp. 2555–2563.

SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 434–459

MIXED DISCONTINUOUS GALERKIN APPROXIMATION OF THE MAXWELL OPERATOR*

PAUL HOUSTON†, ILARIA PERUGIA‡, AND DOMINIK SCHÖTZAU§

Abstract. We introduce and analyze a discontinuous Galerkin discretization of the Maxwell operator in mixed form. Here, all the unknowns of the underlying system of partial differential equations are approximated by discontinuous finite element spaces of the same order. For piecewise constant coefficients, the method is shown to be stable and optimally convergent with respect to the mesh size. Numerical experiments highlighting the performance of the proposed method for problems with both smooth and singular analytical solutions are presented.

Key words. discontinuous Galerkin methods, mixed methods, Maxwell operator

AMS subject classification. 65N30

DOI. 10.1137/S003614290241790X

1. Introduction.

The origins of discontinuous Galerkin (DG) methods can be traced back to the 1970s, where they were proposed for the numerical solution of the neutron transport equation, as well as for the weak enforcement of continuity in Galerkin methods for elliptic and parabolic problems; see [11] for a historical review. In the meantime, these methods have undergone quite a remarkable development and are used in a wide range of applications; see the recent survey articles [10, 12, 13] and the references cited therein. The main advantages of DG methods lie in their robustness, conservation properties, and great flexibility in the mesh design. Indeed, being based on completely discontinuous finite element spaces, these methods can easily handle elements of various types and shapes, nonmatching grids, and local spaces of different polynomial orders; thus, they are ideal for *hp*-adaptivity.

In recent years, DG methods have begun to find their way into computational electromagnetics. Here, we mention the work in [17], where the full Maxwell system is discretized using unstructured spectral elements in space together with a suitable low-storage Runge–Kutta time stepping scheme; similar spectral DG methods were proposed in [22]. The study of DG methods applied to the time-harmonic Maxwell equations in electric field-based formulation was initiated in [24]; here, a local discontinuous Galerkin (LDG) method was proposed for the low-frequency problem, covering the cases of heterogeneous media and topologically nontrivial domains. The numerical experiments in [19] have confirmed the *hp*-convergence rates proved in [24] for smooth solutions and indicate that DG methods can be effective in a wide range of low-frequency applications where the bilinear forms are coercive.

On the other hand, one of the main difficulties in the numerical solution of Maxwell's equations consists of dealing with divergence-free constraints that need

*Received by the editors November 14, 2002; accepted for publication (in revised form) August 15, 2003; published electronically March 3, 2004.
http://www.siam.org/journals/sinum/42-1/41790.html

†Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK (Paul.Houston@mcs.le.ac.uk). The research of this author was supported by the EPSRC under grants GR/N24230 and GR/R76615.

‡Dipartimento di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy (perugia@dimat.unipv.it).

§Mathematics Department, University of British Columbia, 1984 Mathematics Road, Vancouver, BC V6T 1Z2, Canada (schoetzau@math.ubc.ca). The research of this author was partially supported by the Swiss National Science Foundation under Project 21-068126.02.

to be imposed on the fields, especially in cases where the analytical solutions exhibit strong singularities. Several approaches have been proposed in the literature: we mention here the (weighted) regularization methods studied in [1, 14], the singular field approach of [6], and the Lagrange multiplier techniques used in [9, 15, 26], for example. The methods studied in [24, 19] are DG versions of the regularization approach of [1] and, for singular solutions, were shown to suffer from similar drawbacks as their conforming counterparts. A mixed discontinuous Galerkin approach was recently adopted in [25], where a stabilized interior penalty discretization was proposed for the high-frequency time-harmonic Maxwell equations. For smooth material coefficients, optimal convergence of the method was proved by employing a duality approach, provided that appropriate stabilization terms were included in the method.

In this paper, we introduce and analyze a new mixed DG formulation for the Maxwell operator (consisting of the curl-curl operator subject to a divergence-free constraint). Although this formulation is based on the same mixed approach as the one proposed in [25], here the amount of numerical stabilization is drastically reduced. In particular, we abandon all the volume stabilization terms from [25] and achieve well-posedness of the formulation through a suitable definition of the numerical fluxes. We present a numerical analysis of this method for piecewise constant material coefficients and obtain a priori error bounds in the associated energy norm that are optimal in the mesh size if both the field and the Lagrange multiplier related to the divergence constraint are approximated with piecewise polynomials of the same degree. Here, we consider both the case where the underlying analytical solution is smooth and where only minimal regularity assumptions are assumed. The method proposed in this paper is tested on a set of numerical examples that confirm the convergence rates predicted in the theoretical analysis for both smooth and singular solutions on regular and irregular meshes. The method is also tested within an adaptive procedure on affine quadrilateral meshes where hanging nodes are introduced during the course of the refinement. The numerical results indicate that singularities present in the analytical solution are correctly captured by the proposed scheme.

The stability analysis of the mixed DG formulation is carried out along the following lines. First, we rewrite the mixed system in an augmented form by introducing auxiliary variables, giving rise to a standard mixed saddle point problem with nonconsistent forms. Then we establish coercivity of the curl-curl operator on a suitable kernel. Finally, we prove the inf-sup stability condition for the form related to the divergence constraint. The proof of this result makes use of ideas developed in [8] for the analysis of stabilized mixed methods and relies on a decomposition of the discontinuous Galerkin finite element space for the Lagrange multiplier into the direct sum of its largest conforming (stable) subspace and a corresponding complement. The control over functions in the complement is then ensured by a crucial norm equivalence property that we establish by using an approximation result from [21, section 2.1].

The outline of the paper is as follows. In section 2 we introduce our mixed DG method for the Maxwell operator. Our main theoretical results are the a priori error bounds presented in section 3. Their proofs are carried out in the following sections, where we introduce an auxiliary mixed formulation (section 4), establish the continuity and stability properties of the forms involved (section 5), and, finally, derive the actual error estimates (section 6). The numerical performance of the method is tested in section 7. Concluding remarks are presented in section 8.

2. Model problem and discretization. In this section, we introduce a mixed DG discretization of the curl-curl operator subject to a divergence-free constraint.

2.1. Notation. We start by introducing the notation and function spaces that will be used throughout this paper. Given a bounded domain D in $\mathbb{R}^2$ or $\mathbb{R}^3$, we denote by $H^s(D)$ the standard Sobolev space of functions with the integer or fractional regularity exponent $s \geq 0$ and by $\|\cdot\|_{s,D}$ its norm. We also write $\|\cdot\|_{s,D}$ to denote the norms in the spaces $H^s(D)^d$, $d = 2, 3$. We set $L^2(D) = H^0(D)$. Furthermore, $L^\infty(D)$ is the space of bounded functions on D. Given $D \subset \mathbb{R}^3$ and a positive weight function $w \in L^\infty(D)$, $H(\mathrm{curl}_w; D)$ and $H(\mathrm{div}_w; D)$ are the spaces of vector fields $\boldsymbol{u} \in L^2(D)^3$ with $\nabla \times (w\boldsymbol{u}) \in L^2(D)^3$ and $\nabla \cdot (w\boldsymbol{u}) \in L^2(D)$, respectively, endowed with their corresponding graph norms. $H(\mathrm{curl}^0_w; D)$ and $H(\mathrm{div}^0_w; D)$ are the subspaces of $H(\mathrm{curl}_w; D)$ and $H(\mathrm{div}_w; D)$, respectively, of functions with zero (weighted) curl and divergence, respectively. For $w \equiv 1$, we omit the subscript and write $H(\mathrm{curl}; D)$ and $H(\mathrm{div}; D)$, respectively. We denote by $H^1_0(D)$, $H_0(\mathrm{curl}; D)$, and $H_0(\mathrm{div}; D)$ the subspaces of $H^1(D)$, $H(\mathrm{curl}; D)$, and $H(\mathrm{div}; D)$, respectively, of functions with zero trace, tangential trace, and normal trace, respectively.

2.2. Model problem. Let Ω be a bounded Lipschitz polyhedron in $\mathbb{R}^3$, with $\boldsymbol{n}$ denoting the outward normal unit vector to its boundary $\Gamma = \partial\Omega$. We assume that the domain Ω is simply connected and that Γ is connected. We consider the following mixed model problem: find the vector field $\boldsymbol{u}$ and the scalar field p such that

$$\begin{aligned} \nabla \times (\mu^{-1} \nabla \times \boldsymbol{u}) - \varepsilon \nabla p &= \boldsymbol{j} \quad \text{in } \Omega, \\ \nabla \cdot (\varepsilon \boldsymbol{u}) &= 0 \quad \text{in } \Omega, \\ \boldsymbol{n} \times \boldsymbol{u} &= \boldsymbol{g} \quad \text{on } \Gamma, \\ p &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{2.1}$$

Here, the right-hand side $\boldsymbol{j} \in L^2(\Omega)^3$ is an external source field, and the Dirichlet datum $\boldsymbol{g}$ is a prescribed tangential trace which we assume belongs to $L^2(\Gamma)^3$. The coefficients $\mu = \mu(\boldsymbol{x})$ and $\varepsilon = \varepsilon(\boldsymbol{x})$ are real functions in $L^\infty(\Omega)$ that satisfy

$$0 < \mu_* \leq \mu(\boldsymbol{x}) \leq \mu^* < \infty, \qquad 0 < \varepsilon_* \leq \varepsilon(\boldsymbol{x}) \leq \varepsilon^* < \infty, \quad \text{a.e. } \boldsymbol{x} \in \overline{\Omega}. \tag{2.2}$$

For simplicity, we assume that μ and ε are piecewise constant with respect to a partition of the domain Ω into Lipschitz polyhedra.

REMARK 2.1. *Problem* (2.1) *describes the principal operator of the time-harmonic Maxwell equations in a heterogeneous insulating medium (i.e., with electric conductivity* $\sigma = 0$*). The coefficients* μ *and* ε *are the magnetic permeability and the electric permittivity of the medium, respectively. The divergence constraint is incorporated by means of the Lagrange multiplier* p*; see, e.g.,* [15, 25, 26] *and the references cited therein. Problem* (2.1) *is also a formulation of the magnetostatic problem in terms of the vector potential* $\boldsymbol{u}$ *and with Coulomb's gauge* $\nabla \cdot \boldsymbol{u} = 0$ *(*$\varepsilon \equiv 1$ *in this case).*

REMARK 2.2. *In the discontinuous Galerkin context, the Dirichlet boundary condition* $\boldsymbol{n} \times \boldsymbol{u} = \boldsymbol{g}$ *on* Γ *is enforced weakly by so-called interior penalty stabilization terms. In order to make these terms well defined for each boundary face of a grid on* Ω*, we use the regularity assumption* $\boldsymbol{g} \in L^2(\Gamma)^3$*, which is slightly stronger than the natural assumption for* $\boldsymbol{g}$*. For less regular boundary data, the interior penalty terms need to be defined as suitable duality pairings over the whole boundary* Γ*.*

Setting $\boldsymbol{V} = \{\boldsymbol{v} \in H(\mathrm{curl}; \Omega) : (\boldsymbol{n} \times \boldsymbol{v})|_\Gamma \in L^2(\Gamma)^3\}$ and $Q = H^1_0(\Omega)$, the variational form of (2.1) is as follows: find $\boldsymbol{u} \in \boldsymbol{V}$, with $\boldsymbol{n} \times \boldsymbol{u} = \boldsymbol{g}$ on Γ, and $p \in Q$ such that

$$a(\boldsymbol{u},\boldsymbol{v}) + b(\boldsymbol{v},p) = \int_\Omega \boldsymbol{j}\cdot\boldsymbol{v}\,d\boldsymbol{x}, \tag{2.3}$$

$$b(\boldsymbol{u},q) = 0 \tag{2.4}$$

for all $(\boldsymbol{v},q) \in H_0(\mathrm{curl};\Omega)\times Q$, where the forms a and b are given, respectively, by

$$a(\boldsymbol{u},\boldsymbol{v}) = \int_\Omega \mu^{-1}\nabla\times\boldsymbol{u}\cdot\nabla\times\boldsymbol{v}\,d\boldsymbol{x}, \qquad b(\boldsymbol{v},p) = -\int_\Omega \varepsilon\boldsymbol{v}\cdot\nabla p\,d\boldsymbol{x}.$$

Well-posedness of the formulation (2.3)–(2.4) follows from the standard theory of mixed problems [7], since a is bilinear, continuous, and coercive on the kernel of b, and b is linear and continuous, and satisfies the inf-sup condition; see, e.g., [26] for details.

2.3. Meshes, finite element spaces, and traces. Throughout, we consider shape regular and affine meshes $\mathcal{T}_h$ that partition the domain Ω into tetrahedra and/or parallelepipeds, with possible hanging nodes; we always assume that the meshes are aligned with any discontinuities in the coefficients μ and ε. We denote by h_K the diameter of the element $K\in\mathcal{T}_h$ and set $h=\max_K h_K$. An interior face of $\mathcal{T}_h$ is defined as the (nonempty) two-dimensional interior of $\partial K^+\cap\partial K^-$, where K^+ and K^- are two adjacent elements of $\mathcal{T}_h$, not necessarily matching. A boundary face of $\mathcal{T}_h$ is defined as the (nonempty) two-dimensional interior of $\partial K\cap\Gamma$, where K is a boundary element of $\mathcal{T}_h$. We denote by $\mathcal{F}_h^{\mathcal{I}}$ the union of all interior faces of $\mathcal{T}_h$, $\mathcal{F}_h^{\mathcal{D}}$ the union of all boundary faces of $\mathcal{T}_h$, and set $\mathcal{F}_h=\mathcal{F}_h^{\mathcal{I}}\cup\mathcal{F}_h^{\mathcal{D}}$.

Given a nonnegative integer ℓ and an element $K\in\mathcal{T}_h$, we define $\mathcal{S}^\ell(K)$ as the space $\mathcal{P}^\ell(K)$ of polynomials of degree at most ℓ in K, if K is a tetrahedron, or the space $\mathcal{Q}^\ell(K)$ of polynomials of degree at most ℓ in each variable in K, if K is a parallelepiped. Similarly, for a face $f\subset\mathcal{F}_h$, we write $\mathcal{S}^\ell(f)$ for the space $\mathcal{P}^\ell(f)$ of polynomials of degree at most ℓ in f, if f is a triangle, and the space $\mathcal{Q}^\ell(f)$ of polynomials of degree at most ℓ in each variable in f, if f is a parallelogram. Then the generic finite element space of discontinuous piecewise polynomials is given by

$$S^\ell(\mathcal{T}_h) = \{u\in L^2(\Omega):\, u|_K\in\mathcal{S}^\ell(K)\ \ \forall K\in\mathcal{T}_h\}.$$

For piecewise smooth vector- and scalar-valued functions $\boldsymbol{v}$ and q, respectively, we introduce the following trace operators. Let $f\subset\mathcal{F}_h^{\mathcal{I}}$ be an interior face shared by two neighboring elements K^+ and K^-; we write $\boldsymbol{n}^\pm$ to denote the outward normal unit vectors to the boundaries $\partial K^\pm$, respectively. Denoting by $\boldsymbol{v}^\pm$ and $q^\pm$ the traces of $\boldsymbol{v}$ and q on $\partial K^\pm$ taken from within $K^\pm$, respectively, we define the jumps across f by $[\![\boldsymbol{v}]\!]_T=\boldsymbol{n}^+\times\boldsymbol{v}^+ + \boldsymbol{n}^-\times\boldsymbol{v}^-$, $[\![\boldsymbol{v}]\!]_N=\boldsymbol{v}^+\cdot\boldsymbol{n}^+ + \boldsymbol{v}^-\cdot\boldsymbol{n}^-$, and $[\![q]\!]_N=q^+\boldsymbol{n}^+ + q^-\boldsymbol{n}^-$ and the averages by $\{\!\{\boldsymbol{v}\}\!\}=(\boldsymbol{v}^+ + \boldsymbol{v}^-)/2$ and $\{\!\{q\}\!\}=(q^+ + q^-)/2$. On a boundary face $f\subset\mathcal{F}_h^{\mathcal{D}}$, we set $[\![\boldsymbol{v}]\!]_T=\boldsymbol{n}\times\boldsymbol{v}$, $[\![q]\!]_N=q\,\boldsymbol{n}$, $\{\!\{\boldsymbol{v}\}\!\}=\boldsymbol{v}$, and $\{\!\{q\}\!\}=q$.

2.4. DG discretization. We wish to approximate problem (2.1) by discrete functions $\boldsymbol{u}_h$ and p_h in the finite element spaces $\boldsymbol{V}_h=S^\ell(\mathcal{T}_h)^3$ and $Q_h=S^\ell(\mathcal{T}_h)$, respectively, for a given partition $\mathcal{T}_h$ of Ω and an approximation order $\ell\ge 1$.

To this end, we consider the following DG method: find $(\boldsymbol{u}_h,p_h)\in\boldsymbol{V}_h\times Q_h$ such that

$$a_h(\boldsymbol{u}_h,\boldsymbol{v}) + b_h(\boldsymbol{v},p_h) = f_h(\boldsymbol{v}), \tag{2.5}$$

$$b_h(\boldsymbol{u}_h,q) - c_h(p_h,q) = 0 \tag{2.6}$$

for all $(\boldsymbol{v},q) \in \boldsymbol{V}_h \times Q_h$, where the discrete forms a_h, b_h, and c_h and the linear functional f_h are given, respectively, by

$$
\begin{aligned}
a_h(\boldsymbol{u},\boldsymbol{v}) &= \int_\Omega \mu^{-1}\nabla_h \times \boldsymbol{u} \cdot \nabla_h \times \boldsymbol{v}\, d\boldsymbol{x} - \int_{\mathcal{F}_h} [\![\boldsymbol{u}]\!]_T \cdot \{\!\{\mu^{-1}\nabla_h \times \boldsymbol{v}\}\!\}\, ds \\
&\quad - \int_{\mathcal{F}_h} [\![\boldsymbol{v}]\!]_T \cdot \{\!\{\mu^{-1}\nabla_h \times \boldsymbol{u}\}\!\}\, ds + \int_{\mathcal{F}_h} \mathtt{a}\, [\![\boldsymbol{u}]\!]_T \cdot [\![\boldsymbol{v}]\!]_T\, ds + \int_{\mathcal{F}_h^{\mathcal{I}}} \mathtt{b}\, [\![\varepsilon\boldsymbol{u}]\!]_N [\![\varepsilon\boldsymbol{v}]\!]_N\, ds, \\
b_h(\boldsymbol{v},p) &= -\int_\Omega \varepsilon\boldsymbol{v} \cdot \nabla_h p\, d\boldsymbol{x} + \int_{\mathcal{F}_h} \{\!\{\varepsilon\boldsymbol{v}\}\!\} \cdot [\![p]\!]_N\, ds, \qquad c_h(p,q) = \int_{\mathcal{F}_h} \mathtt{c}[\![p]\!]_N \cdot [\![q]\!]_N\, ds, \\
f_h(\boldsymbol{v}) &= \int_\Omega \boldsymbol{j} \cdot \boldsymbol{v}\, d\boldsymbol{x} - \int_{\mathcal{F}_h^{\mathcal{D}}} \boldsymbol{g} \cdot \mu^{-1}\nabla_h \times \boldsymbol{v}\, ds + \int_{\mathcal{F}_h^{\mathcal{D}}} \mathtt{a}\, \boldsymbol{g} \cdot (\boldsymbol{n} \times \boldsymbol{v})\, ds.
\end{aligned}
$$

Here, ∇_h denotes the elementwise ∇ operator. The form a_h corresponds to the interior penalty discretization of the curl-curl operator [19, 25], with the addition of a normal jump term; the form b_h discretizes the divergence operator in a DG fashion; and the form c_h is a stabilization form that penalizes the jumps of p_h. The parameters $\mathtt{a}$, $\mathtt{b}$, and $\mathtt{c}$ are positive stabilization parameters that will be chosen later on, depending on the mesh size and the coefficients μ and ε. Note that a similar discretization has been investigated in [25] for a time-harmonic high-frequency model of Maxwell's equations; we point out that the additional stabilization forms that have been added there become obsolete with the analysis presented in this paper.

As in [24, Remark 3.2] or [25, Proposition 4], it can be readily seen that the analytical solution $(\boldsymbol{u},p) \in \boldsymbol{V} \times Q$ satisfies (2.5)–(2.6) for all $(\boldsymbol{v},q) \in \boldsymbol{V}_h \times Q_h$.

REMARK 2.3. *All the interface contributions arising in the forms in* (2.5)–(2.6) *can easily be obtained by rewriting the problem* (2.1) *as a first-order system and introducing so-called* numerical fluxes *in the sense of* [5]. *Thus, all the stabilization terms in* (2.5)–(2.6) *are* local, consistent, *and* conservative. *To see this, we rewrite* (2.1) *as*

$$
\boldsymbol{s} - \mu^{-1}\nabla \times \boldsymbol{u} = 0, \qquad \nabla \times \boldsymbol{s} - \varepsilon\nabla p = \boldsymbol{j}, \qquad \nabla \cdot (\varepsilon\boldsymbol{u}) = 0 \qquad \text{in } \Omega,
$$

subject to the boundary conditions $\boldsymbol{n} \times \boldsymbol{u} = \boldsymbol{g}$ *and* $p = 0$ *on* Γ. *Then we consider the following discretization: find* $(\boldsymbol{s}_h, \boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times \boldsymbol{V}_h \times Q_h$ *such that*

$$
\begin{aligned}
&\int_K \boldsymbol{s}_h \cdot \boldsymbol{t}\, d\boldsymbol{x} - \int_K \mu^{-1}\nabla \times \boldsymbol{t} \cdot \boldsymbol{u}_h\, d\boldsymbol{x} + \int_{\partial K} \mu^{-1}\boldsymbol{t} \cdot \widehat{\boldsymbol{u}}_h \times \boldsymbol{n}\, ds = 0, \\
&\int_K \boldsymbol{s}_h \cdot \nabla \times \boldsymbol{v}\, d\boldsymbol{x} - \int_{\partial K} \boldsymbol{v} \cdot \widehat{\boldsymbol{s}}_h \times \boldsymbol{n}\, ds + \int_K p_h\, \nabla \cdot (\varepsilon\boldsymbol{v})\, d\boldsymbol{x} \\
&\qquad\qquad - \int_{\partial K} \widehat{p}_h\, \varepsilon\boldsymbol{v} \cdot \boldsymbol{n}\, ds = \int_K \boldsymbol{j} \cdot \boldsymbol{v}\, d\boldsymbol{x}, \\
&\int_K \varepsilon\boldsymbol{u}_h \cdot \nabla q\, d\boldsymbol{x} - \int_{\partial K} q\, \widehat{\varepsilon\boldsymbol{u}}_h \cdot \boldsymbol{n}\, ds = 0
\end{aligned}
\tag{2.7}
$$

for all $(\boldsymbol{t},\boldsymbol{v},q) \in \boldsymbol{V}_h \times \boldsymbol{V}_h \times Q_h$ *and for all elements* K *in the partition* $\mathcal{T}_h$. *In* (2.7), *the traces of* $\boldsymbol{u}_h$, $\boldsymbol{s}_h$, p_h, *and* $\varepsilon\boldsymbol{u}_h$ *on* ∂K *are approximated by the numerical fluxes*

$$
\begin{aligned}
\widehat{\boldsymbol{u}}_h &= \{\!\{\boldsymbol{u}_h\}\!\}, & \widehat{\boldsymbol{s}}_h &= \{\!\{\mu^{-1}\nabla_h \times \boldsymbol{u}_h\}\!\} - \mathtt{a}[\![\boldsymbol{u}_h]\!]_T, \\
\widehat{p}_h &= \{\!\{p_h\}\!\} - \mathtt{b}[\![\varepsilon\boldsymbol{u}_h]\!]_N, & \widehat{\varepsilon\boldsymbol{u}}_h &= \{\!\{\varepsilon\boldsymbol{u}_h\}\!\} - \mathtt{c}[\![p_h]\!]_N,
\end{aligned}
$$

respectively (these definitions are for interior faces; they must be suitably adapted for

boundary faces). By integration by parts, the first equation of (2.7) *reads as*

$$\int_K \boldsymbol{s}_h \cdot \boldsymbol{t}\, d\boldsymbol{x} = \int_K \mu^{-1} \nabla \times \boldsymbol{u}_h \cdot \boldsymbol{t}\, d\boldsymbol{x} + \int_{\partial K} \mu^{-1} \boldsymbol{t} \cdot (\boldsymbol{u}_h - \widehat{\boldsymbol{u}}_h) \times \boldsymbol{n}\, ds. \tag{2.8}$$

As the numerical flux $\widehat{\boldsymbol{u}}_h$ *is independent of* $\boldsymbol{s}_h$*, the auxiliary variable* $\boldsymbol{s}_h$ *can be locally expressed in terms of* $\boldsymbol{u}_h$ *by inverting the local mass matrix* $\int_K \boldsymbol{s}_h \cdot \boldsymbol{t}\, d\boldsymbol{x}$ *in* (2.8)*. By substituting the resulting expression for* $\boldsymbol{s}_h$ *into the second equation of* (2.7)*, one obtains an elemental formulation for the unknowns* $\boldsymbol{u}_h$ *and* p_h *only. Finally, summing over all elements* $K \in \mathcal{T}_h$ *gives the formulation* (2.5)–(2.6)*. We refer the reader to* [5] *and* [24] *for further details on the formalization of this elimination process.*

3. Main results. In this section, we state our main results for the mixed DG method in (2.5)–(2.6); the proofs of these a priori error bounds will be given in sections 4, 5, and 6.

3.1. Stabilization parameters and DG-norms. We start by defining the stabilization parameters $\mathtt{a}$, $\mathtt{b}$, and $\mathtt{c}$ appearing in (2.5)–(2.6) and introduce the norms employed in the proceeding error analysis. To this end, we first define the function $\mathtt{h}$ in $L^\infty(\mathcal{F}_h)$, representing the local mesh size, as $\mathtt{h}(\boldsymbol{x}) = \min\{h_K, h_{K'}\}$, if $\boldsymbol{x}$ is in the interior of $\partial K \cap \partial K'$ for two neighboring elements in the mesh $\mathcal{T}_h$, and $\mathtt{h}(\boldsymbol{x}) = h_K$ if $\boldsymbol{x}$ is in the interior of $\partial K \cap \Gamma$. Similarly, we define the functions $\mathtt{m}$ and $\mathtt{e}$ in $L^\infty(\mathcal{F}_h)$ by $\mathtt{m}(\boldsymbol{x}) = \min\{\mu_K, \mu_{K'}\}$ and $\mathtt{e}(\boldsymbol{x}) = \max\{\varepsilon_K, \varepsilon_{K'}\}$, if $\boldsymbol{x}$ is in the interior of $\partial K \cap \partial K'$, and $\mathtt{m}(\boldsymbol{x}) = \mu_K$ and $\mathtt{e}(\boldsymbol{x}) = \varepsilon_K$, if $\boldsymbol{x}$ is in the interior of $\partial K \cap \Gamma$, with μ_K and ε_K denoting the restrictions of μ and ε to the element K, respectively. With this notation, we choose the stabilization parameters as follows:

$$\mathtt{a} = \alpha\, \mathtt{m}^{-1}\mathtt{h}^{-1}, \qquad \mathtt{b} = \beta\, \mathtt{e}^{-1}\mathtt{h}, \qquad \mathtt{c} = \gamma\, \mathtt{e}\mathtt{h}^{-1}, \tag{3.1}$$

where α, β, and γ are positive parameters, independent of the mesh size and the coefficients μ and ε.

Further, we set $\boldsymbol{V}(h) = (\boldsymbol{V} \cap H(\mathrm{div}_\varepsilon; \Omega)) + \boldsymbol{V}_h$ and $Q(h) = Q + Q_h$ and define

$$\begin{aligned}
|\boldsymbol{v}|^2_{\boldsymbol{V}(h)} &= \|\mu^{-\frac{1}{2}} \nabla_h \times \boldsymbol{v}\|^2_{0,\Omega} + \|\mathtt{m}^{-\frac{1}{2}} \mathtt{h}^{-\frac{1}{2}} [\![\boldsymbol{v}]\!]_T\|^2_{0,\mathcal{F}_h} + \|\mathtt{e}^{-\frac{1}{2}} \mathtt{h}^{\frac{1}{2}} [\![\varepsilon \boldsymbol{v}]\!]_N\|^2_{0,\mathcal{F}_h^{\mathcal{I}}}, \\
\|\boldsymbol{v}\|^2_{\boldsymbol{V}(h)} &= \|\varepsilon^{\frac{1}{2}} \boldsymbol{v}\|^2_{0,\Omega} + |\boldsymbol{v}|^2_{\boldsymbol{V}(h)}, \\
\|q\|^2_{Q(h)} &= \|\varepsilon^{\frac{1}{2}} \nabla_h q\|^2_{0,\Omega} + \|\mathtt{e}^{\frac{1}{2}} \mathtt{h}^{-\frac{1}{2}} [\![q]\!]_N\|^2_{0,\mathcal{F}_h}.
\end{aligned}$$

We also introduce the space $H^s(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|_K \in H^s(K),\ K \in \mathcal{T}_h\}$, endowed with the norm $\|v\|^2_{s,\mathcal{T}_h} = \sum_{K \in \mathcal{T}_h} \|v\|^2_{s,K}$. On the boundary, we define $H^s(\mathcal{F}_h^{\mathcal{D}}) = \{v \in L^2(\mathcal{F}_h^{\mathcal{D}}) : v|_f \in H^s(f),\ f \subset \mathcal{F}_h^{\mathcal{D}}\}$, equipped with the norm $\|v\|^2_{s,\mathcal{F}_h^{\mathcal{D}}} = \sum_{f \subset \mathcal{F}_h^{\mathcal{D}}} \|v\|^2_{s,f}$.

3.2. Existence and uniqueness. Next, we show that the discrete problem in (2.5)–(2.6) is uniquely solvable, provided that α is sufficiently large. To this end, we first recall the following well-known coercivity result, valid in view of the choice of $\mathtt{a}$ and the assumptions on the meshes; see [5, 19, 25] for details.

LEMMA 3.1. *There exists a parameter* $\alpha_{\min} > 0$*, independent of the mesh size and the coefficients* μ *and* ε*, such that for* $\alpha \geq \alpha_{\min}$ *and* $\beta > 0$ *we have*

$$a_h(\boldsymbol{v}, \boldsymbol{v}) \geq C |\boldsymbol{v}|^2_{\boldsymbol{V}(h)} \quad \forall \boldsymbol{v} \in \boldsymbol{V}_h,$$

with a constant $C > 0$ *independent of the mesh size and the coefficients* μ *and* ε*.*

The condition $\alpha \geq \alpha_{\min} > 0$ is a restriction that is typically encountered with symmetric interior penalty methods and may be omitted by using other DG discretizations of the curl-curl operator, such as the nonsymmetric interior penalty or the LDG method; see, e.g., [5, 24] for details.

PROPOSITION 3.2. *For $\alpha \geq \alpha_{\min}$, $\beta > 0$, and $\gamma > 0$, the mixed DG method* (2.5)–(2.6) *possesses a unique solution.*

Proof. It is enough to show that $\boldsymbol{j} = \mathbf{0}$ and $\boldsymbol{g} = \mathbf{0}$ imply $\boldsymbol{u}_h = \mathbf{0}$ and $p_h = 0$. To this end, take $\boldsymbol{v} = \boldsymbol{u}_h$ in (2.5) and $q = p_h$ in (2.6); then subtract (2.6) from (2.5). With the coercivity of a_h in Lemma 3.1, it follows that $\nabla_h \times \boldsymbol{u}_h = \mathbf{0}$, $[\![\boldsymbol{u}_h]\!]_T = \mathbf{0}$ on $\mathcal{F}_h$, $[\![\varepsilon \boldsymbol{u}_h]\!]_N = 0$ on $\mathcal{F}_h^{\mathcal{I}}$, and $[\![p_h]\!]_N = \mathbf{0}$ on $\mathcal{F}_h$; i.e., $\boldsymbol{u}_h \in H(\mathrm{curl}^0;\Omega) \cap H_0(\mathrm{curl};\Omega)$, $\boldsymbol{u}_h \in H(\mathrm{div}_\varepsilon;\Omega)$, and $p_h \in H_0^1(\Omega)$. Integrating by parts, (2.6) becomes $\int_\Omega q\, \nabla \cdot (\varepsilon \boldsymbol{u}_h)\, d\boldsymbol{x} = 0$, for any $q \in Q_h$, and then, since ε is piecewise constant, $\nabla \cdot (\varepsilon \boldsymbol{u}_h) = 0$. Therefore, $\boldsymbol{u}_h$ also belongs to $H(\mathrm{div}_\varepsilon^0;\Omega)$, which, owing to our assumptions on Ω, implies that $\boldsymbol{u}_h = \mathbf{0}$; cf. [16, section 4]. Equation (2.5) becomes $\int_\Omega \varepsilon \boldsymbol{v} \cdot \nabla p_h\, d\boldsymbol{x} = 0$, for any $\boldsymbol{v} \in V_h$, and then $\nabla p_h = \mathbf{0}$. Since $p_h = 0$ on Γ, we conclude that $p_h = 0$. ∎

For the rest of this article, we shall assume that the hypotheses on the stabilization parameters in the statement of Proposition 3.2 hold.

3.3. A priori error estimates. First, we establish optimal error estimates for smooth solutions on possibly nonconforming meshes subject to the following restrictions: (i) any interior face $f \subset \mathcal{F}_h^{\mathcal{I}}$ has to be an *entire* elemental face of at least one of the two adjacent elements sharing f; (ii) the number of interior faces contained in an elemental face is uniformly bounded with respect to the mesh size h. This implies bounded variation of the local mesh size; i.e., whenever K and K' share a common face and $h_K \geq h_{K'}$, we have $h_K \leq C h_{K'} \leq C h_K$, for a constant $C > 0$, independent of the mesh size. The reason for this restriction is related to the validity of the norm equivalence result of Theorem 5.3.

THEOREM 3.3. *Let $(\boldsymbol{u}, p)$ be the analytical solution of* (2.1) *satisfying $\boldsymbol{u} \in H^{s+1}(\mathcal{T}_h)^3$ and $p \in H^{s+1}(\mathcal{T}_h)$ for a regularity exponent $s > \frac{1}{2}$. Let $(\boldsymbol{u}_h, p_h)$ be the mixed DG approximation obtained by* (2.5)–(2.6) *on possibly nonconforming meshes that satisfy restrictions* (i) *and* (ii) *above. Then we have the a priori error bound*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)} + \|p - p_h\|_{Q(h)} \leq C\, h^{\min\{s,\ell\}} \Big[\|\boldsymbol{u}\|_{s+1,\mathcal{T}_h} + \|p\|_{s+1,\mathcal{T}_h} \Big],$$

with $C > 0$ depending on the bounds (2.2) *on the coefficients μ and ε, the shape-regularity and bounded variation properties of the mesh, the stabilization parameters α, β, and γ, and the polynomial degree ℓ but independent of the mesh size h.*

While the bound in Theorem 3.3 guarantees optimal convergence in the mesh size h with respect to the polynomial degree used in the approximation, the smoothness assumptions on the analytical solution are not minimal. In fact, for $\varepsilon = \mu = 1$ and a homogeneous Dirichlet datum $\boldsymbol{g} \equiv \mathbf{0}$, from the regularity results in [3], it follows that one has only $\boldsymbol{u} \in H^s(\Omega)^3$ and $\nabla \times \boldsymbol{u} \in H^s(\Omega)^3$ for a regularity exponent $s = s(\Omega) > \frac{1}{2}$ (the same actually holds for smooth coefficients ε and μ; see [25, section 2.2]). As far as the assumption $p \in H^{s+1}(\Omega)$ is concerned, although it does not seem to hold for general source terms $\boldsymbol{j}$ in $L^2(\Omega)^3$, it is trivially satisfied in the physically most relevant case of divergence-free source terms $\boldsymbol{j}$, where $p \equiv 0$.

For these reasons, we state a second result under weaker smoothness assumptions for the component $\boldsymbol{u}$ of the analytical solution. In order to do this, we restrict ourselves to the case of conforming meshes (i.e., meshes with no hanging nodes); this restriction is necessary, since the proof requires the use of $H(\mathrm{curl};\Omega)$-conforming projections.

Note that, since the boundary datum $\boldsymbol{g}$ is the tangential trace on Γ of a function in $H(\mathrm{curl};\Omega)$, its restriction to each $f \subset \mathcal{F}_h^{\mathcal{D}}$ is a two-dimensional vector field which lies on the same plane as f. Hence, we can understand $\boldsymbol{g}$ as a function in $L^2(\Gamma)^2$.

THEOREM 3.4. *Let $(\boldsymbol{u},p)$ be the analytical solution of (2.1) satisfying $\varepsilon\boldsymbol{u} \in H^s(\mathcal{T}_h)^3$, $\mu^{-1}\nabla\times\boldsymbol{u} \in H^s(\mathcal{T}_h)^3$, $p \in H^{s+1}(\mathcal{T}_h)$, and $\boldsymbol{g} \in H^{s+\frac{1}{2}}(\mathcal{F}_h^{\mathcal{D}})^2$ for a regularity exponent $s > \frac{1}{2}$. Let $(\boldsymbol{u}_h,p_h)$ be the mixed DG approximation obtained by (2.5)–(2.6) on conforming meshes. Then we have the a priori error bound*

$$\begin{aligned}&\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{V}(h)} + \|p-p_h\|_{Q(h)}\\ &\qquad \le C\,h^{\min\{s,\ell\}}\Big[\|\varepsilon\boldsymbol{u}\|_{s,\mathcal{T}_h} + \|\mu^{-1}\nabla\times\boldsymbol{u}\|_{s,\mathcal{T}_h} + \|p\|_{s+1,\mathcal{T}_h} + \|\boldsymbol{g}\|_{s+\frac{1}{2},\mathcal{F}_h^{\mathcal{D}}}\Big],\end{aligned}$$

with $C > 0$ depending on the bounds (2.2) on the coefficients μ and ε, the shape-regularity of the mesh, the stabilization parameters α, β, and γ, and the polynomial degree ℓ but independent of the mesh size h.

REMARK 3.5. *The numerical results reported in section 7 show that, on a two-dimensional L-shaped domain, the above convergence rates are obtained also on nonconforming affine meshes for the strongest corner singularities.*

The mixed method in (2.5)–(2.6) enforces the divergence constraint in a weak sense only; nevertheless, the convergence rate of the error in the (elementwise) divergence might be of interest. Our last result addresses this issue and proves a rate that is of one order lower than the error measured in the DG-norm. This result is numerically observed to be sharp on conforming finite element meshes; cf. section 7.

THEOREM 3.6. *Let $(\boldsymbol{u},p)$ be the analytical solution of (2.1) and $(\boldsymbol{u}_h,p_h)$ be the mixed DG approximation obtained by (2.5)–(2.6) on a possibly nonconforming mesh satisfying restrictions* (i) *and* (ii) *above. Then we have*

$$\sum_{K\in\mathcal{T}_h} h_K^2\|\nabla\cdot(\varepsilon(\boldsymbol{u}-\boldsymbol{u}_h))\|_{0,K}^2 \le C\left[\|\mathtt{h}^{\frac{1}{2}}[\![\varepsilon\boldsymbol{u}-\varepsilon\boldsymbol{u}_h]\!]_N\|_{\mathcal{F}_h^{\mathcal{I}}}^2 + \|\mathtt{e}^{\frac{1}{2}}\mathtt{h}^{-\frac{1}{2}}[\![q-q_h]\!]_N\|_{0,\mathcal{F}_h}^2\right],$$

with $C > 0$ depending on the shape-regularity and bounded variation properties of the mesh, the stabilization parameter γ, and the polynomial degree ℓ but independent of the mesh size h.

REMARK 3.7. *Under the assumptions of both Theorem 3.3 and Theorem 3.4, the estimate in Theorem 3.6 implies that $[\sum_{K\in\mathcal{T}_h} h_K^2\|\nabla\cdot(\varepsilon(\boldsymbol{u}-\boldsymbol{u}_h))\|_{0,K}^2]^{\frac{1}{2}} \le C\,h^{\min\{s,\ell\}}$, with $C > 0$ independent of the mesh size.*

The proofs of Theorems 3.3, 3.4, and 3.6 are carried out in the next sections and concluded in sections 6.2 and 6.3.

4. Auxiliary mixed formulation. In order to facilitate the error analysis, we rewrite the discrete formulation (2.5)–(2.6) in a different (and perturbed) form by introducing the jumps of p_h as auxiliary unknowns and by employing lifting operators as in [5, 24]. In this way, the resulting bilinear forms have suitable continuity and coercivity properties so that the method can be analyzed using the classical theory of mixed finite methods. We begin by introducing the lifting operators $\mathcal{L}$ and $\mathcal{M}$. For $\boldsymbol{v}$ belonging to $\boldsymbol{V}(h)$ and $q \in Q(h)$, we define $\mathcal{L}(\boldsymbol{v}) \in \boldsymbol{V}_h$ and $\mathcal{M}(q) \in Q_h$ by

$$\int_\Omega \mathcal{L}(\boldsymbol{v})\cdot\boldsymbol{w}\,d\boldsymbol{x} = \int_{\mathcal{F}_h} [\![\boldsymbol{v}]\!]_T\cdot\{\!\{\boldsymbol{w}\}\!\}\,ds, \qquad \int_\Omega \mathcal{M}(q)\cdot\boldsymbol{w}\,d\boldsymbol{x} = \int_{\mathcal{F}_h}\{\!\{\boldsymbol{w}\}\!\}\cdot[\![q]\!]_N\,ds$$

for all $\boldsymbol{w} \in \boldsymbol{V}_h$. We then define the perturbed forms

$$\widetilde{a}_h(\boldsymbol{u},\boldsymbol{v}) = \int_\Omega \mu^{-1}\nabla_h \times \boldsymbol{u} \cdot \nabla_h \times \boldsymbol{v}\, d\boldsymbol{x} - \int_\Omega \mathcal{L}(\boldsymbol{u}) \cdot (\mu^{-1}\nabla_h \times \boldsymbol{v})\, d\boldsymbol{x} - \int_\Omega \mathcal{L}(\boldsymbol{v}) \cdot (\mu^{-1}\nabla_h \times \boldsymbol{u})\, d\boldsymbol{x} + \int_{\mathcal{F}_h} \mathtt{a}\, [\![\boldsymbol{u}]\!]_T \cdot [\![\boldsymbol{v}]\!]_T\, ds + \int_{\mathcal{F}_h^{\mathcal{I}}} \mathtt{b}\, [\![\varepsilon \boldsymbol{u}]\!]_N [\![\varepsilon \boldsymbol{v}]\!]_N\, ds,$$

$$\widetilde{b}_h(\boldsymbol{v},p) = -\int_\Omega \varepsilon \boldsymbol{v} \cdot \big[\nabla_h p - \mathcal{M}(p)\big]\, d\boldsymbol{x}.$$

Note that $a_h = \widetilde{a}_h$ in $\boldsymbol{V}_h \times \boldsymbol{V}_h$ and $b_h = \widetilde{b}_h$ in $\boldsymbol{V}_h \times Q_h$, although this is no longer true in $\boldsymbol{V}(h) \times \boldsymbol{V}(h)$ and $\boldsymbol{V}(h) \times Q(h)$, respectively.

Next, we define the discrete space

$$M_h = \{\lambda \in L^2(\mathcal{F}_h)^3 : \lambda|_f \in \mathcal{S}^\ell(f)^3 \quad \forall f \subset \mathcal{F}_h\},$$

endowed with the norm $\|\eta\|_{M_h} = \|\mathtt{e}^{\frac{1}{2}}\mathtt{h}^{-\frac{1}{2}}\eta\|_{0,\mathcal{F}_h}$, and consider the following auxiliary mixed formulation: find $(\boldsymbol{u}_h, \lambda_h, p_h) \in \boldsymbol{V}_h \times M_h \times Q_h$ such that

$$A_h(\boldsymbol{u}_h,\lambda_h;\boldsymbol{v},\eta) + B_h(\boldsymbol{v},\eta;p_h) = f_h(\boldsymbol{v}) \qquad \forall (\boldsymbol{v},\eta) \in \boldsymbol{V}_h \times M_h, \tag{4.1}$$

$$B_h(\boldsymbol{u}_h,\lambda_h;q) = 0 \qquad \forall q \in Q_h, \tag{4.2}$$

with forms A_h and B_h given, respectively, by

$$A_h(\boldsymbol{u},\lambda;\boldsymbol{v},\eta) = \widetilde{a}_h(\boldsymbol{u},\boldsymbol{v}) + \int_{\mathcal{F}_h} \mathtt{c}\lambda \cdot \eta\, ds, \quad B_h(\boldsymbol{v},\eta;p) = \widetilde{b}_h(\boldsymbol{v},p) - \int_{\mathcal{F}_h} \mathtt{c}[\![p]\!]_N \cdot \eta\, ds.$$

PROPOSITION 4.1. *Problem* (4.1)–(4.2) *admits a unique solution* $(\boldsymbol{u}_h, \lambda_h, p_h) \in \boldsymbol{V}_h \times M_h \times Q_h$, *with* $(\boldsymbol{u}_h, p_h)$ *the solution to* (2.5)–(2.6), *and* $\lambda_h = [\![p_h]\!]_N$.

Proof. By choosing test functions $(\boldsymbol{0}, \eta)$ in (4.1), we have

$$\int_{\mathcal{F}_h} \mathtt{c}\lambda_h \cdot \eta\, ds = \int_{\mathcal{F}_h} \mathtt{c}[\![p_h]\!]_N \cdot \eta\, ds \qquad \forall \eta \in M_h.$$

Since $\mathtt{c}$ is constant on each $f \in \mathcal{F}_h$, we have that $\lambda_h = [\![p_h]\!]_N$. Then (4.1) and (4.2) coincide with (2.5) and (2.6), respectively. Therefore, if $(\boldsymbol{u}_h, \lambda_h, p_h) \in \boldsymbol{V}_h \times M_h \times Q_h$ is a solution to (4.1)–(4.2), then $\lambda_h = [\![p_h]\!]_N$ and $(\boldsymbol{u}_h, p_h)$ is (the unique) solution to (2.5)–(2.6), which proves uniqueness of the solution. Existence follows from uniqueness. ∎

Finally, we introduce the space $\boldsymbol{W}(h) = \boldsymbol{V}(h) \times M_h$ and set

$$|(\boldsymbol{v},\eta)|^2_{\boldsymbol{W}(h)} = |\boldsymbol{v}|^2_{\boldsymbol{V}(h)} + \|\eta\|^2_{M_h}, \qquad \|(\boldsymbol{v},\eta)\|^2_{\boldsymbol{W}(h)} = \|\boldsymbol{v}\|^2_{\boldsymbol{V}(h)} + \|\eta\|^2_{M_h}.$$

The proofs of Theorems 3.3, 3.4, and 3.6 are now carried out by analyzing the auxiliary mixed formulation (4.1)–(4.2). In section 5 we prove the continuity of A_h and B_h, the ellipticity of A_h on the kernel of B_h, as well as the inf-sup condition for B_h. In the proof of the inf-sup condition, we employ a norm equivalence property. Then the error estimates of Theorems 3.3, 3.4, and 3.6 are obtained in section 6.

5. Continuity and stability. In this section, we prove continuity properties of the forms A_h and B_h, the ellipticity of A_h on the kernel of B_h, as well as the inf-sup condition for B_h.

5.1. Continuity properties. The following continuity properties hold.

PROPOSITION 5.1. *There exist constants $a_1 > 0$ and $a_2 > 0$, independent of the mesh size and the coefficients μ and ε, such that*

$$|A_h(\boldsymbol{u},\lambda;\boldsymbol{v},\eta)| \le a_1\,\|(\boldsymbol{u},\lambda)\|_{\boldsymbol{W}(h)}\|(\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)}, \qquad (\boldsymbol{u},\lambda),(\boldsymbol{v},\eta)\in\boldsymbol{W}(h),$$
$$|B_h(\boldsymbol{v},\eta;q)| \le a_2\,\|(\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)}\|q\|_{Q(h)}, \qquad (\boldsymbol{v},\eta)\in\boldsymbol{W}(h),\ q\in Q(h).$$

The linear functional $f_h : \boldsymbol{V}_h \to \mathbb{R}$ on the right-hand side of (4.1) *satisfies*

$$|f_h(\boldsymbol{v})| \le C\big[\varepsilon_*^{-\frac12}\|\boldsymbol{j}\|_{0,\Omega} + \|\mathtt{m}^{-\frac12}\mathtt{h}^{-\frac12}\boldsymbol{g}\|_{0,\Gamma}\big]\|\boldsymbol{v}\|_{\boldsymbol{V}(h)}, \qquad \boldsymbol{v}\in\boldsymbol{V}_h,$$

with a constant $C > 0$, independent of the mesh size and the coefficients μ and ε.

Proof. Proceeding as in [24, Proposition 4.2] or [25, Proposition 12], we have the following stability estimates for $\mathcal{L}$ and $\mathcal{M}$:

$$\|\mu^{-\frac12}\mathcal{L}(\boldsymbol{v})\|_{0,\Omega} \le C\|\mathtt{m}^{-\frac12}\mathtt{h}^{-\frac12}[\![\boldsymbol{v}]\!]_T\|_{0,\mathcal{F}_h}, \qquad \|\varepsilon^{\frac12}\mathcal{M}(q)\|_{0,\Omega} \le C\|\mathtt{e}^{\frac12}\mathtt{h}^{-\frac12}[\![q]\!]_N\|_{0,\mathcal{F}_h},$$

for any $\boldsymbol{v}\in\boldsymbol{V}(h)$, $q\in Q(h)$, with a constant $C > 0$ that is independent of the mesh size and the coefficients μ and ε. With these stability estimates, the continuity of A_h and B_h follows from the Cauchy–Schwarz inequality and the choice of the stabilization parameters in (3.1). The continuity of f_h is obtained by using similar arguments; see [24, Corollary 4.15] for details. ∎

5.2. Ellipticity on the kernel. Define the discrete kernel

$$\mathrm{Ker}(B_h) = \{(\boldsymbol{u},\lambda)\in\boldsymbol{W}_h : B_h(\boldsymbol{u},\lambda;p) = 0 \ \ \forall p\in Q_h\}.$$

PROPOSITION 5.2. *For $\alpha \ge \alpha_{\min}$, $\beta > 0$, $\gamma > 0$, there is a constant $b > 0$, independent of the mesh size, such that*

$$A_h(\boldsymbol{u},\lambda;\boldsymbol{u},\lambda) \ge b\,\|(\boldsymbol{u},\lambda)\|^2_{\boldsymbol{W}(h)} \quad \forall(\boldsymbol{u},\lambda)\in\mathrm{Ker}(B_h).$$

Proof. Throughout the proof, we denote by C any constant independent of the mesh size and the coefficients μ and ε and by C_m any constant that depends on the bounds (2.2) on the coefficients μ and ε but is independent of the mesh size.

From Lemma 3.1, we immediately have

$$A_h(\boldsymbol{u},\lambda;\boldsymbol{u},\lambda) \ge C\,|(\boldsymbol{u},\lambda)|^2_{\boldsymbol{W}(h)}, \qquad (\boldsymbol{u},\lambda)\in\boldsymbol{W}_h. \tag{5.1}$$

Now fix $(\boldsymbol{u},\lambda)\in\mathrm{Ker}(B_h)$, and let $(\boldsymbol{z},\psi)$ be the solution of the auxiliary problem

$$\nabla\times(\mu^{-1}\nabla\times\boldsymbol{z}) - \varepsilon\nabla\psi = \varepsilon\boldsymbol{u}, \qquad \nabla\cdot(\varepsilon\boldsymbol{z}) = 0 \qquad \text{in } \Omega,$$

subject to the boundary conditions $\boldsymbol{n}\times\boldsymbol{z} = \boldsymbol{0}$ and $\psi = 0$ on Γ. Thereby,

$$\begin{aligned}\|\mu^{-\frac12}\nabla\times\boldsymbol{z}\|_{0,\Omega} + \|\varepsilon^{\frac12}\boldsymbol{z}\|_{0,\Omega} + \|\nabla\times(\mu^{-1}\nabla\times\boldsymbol{z})\|_{0,\Omega}\\ + \|\varepsilon^{\frac12}\nabla\psi\|_{0,\Omega} + \|\varepsilon^{\frac12}\psi\|_{0,\Omega} \le C_m\|\varepsilon^{\frac12}\boldsymbol{u}\|_{0,\Omega}.\end{aligned} \tag{5.2}$$

Set $\boldsymbol{w} = \mu^{-1}\nabla\times\boldsymbol{z}$; clearly, $\boldsymbol{w}\in H(\mathrm{curl};\Omega)$. Therefore, from [16, Corollary 7.2], there exists $\boldsymbol{w}_0\in H^1(\Omega)^3$ such that

$$\nabla\times\boldsymbol{w}_0 = \nabla\times\boldsymbol{w}, \qquad \|\boldsymbol{w}_0\|_{1,\Omega} \le C\|\boldsymbol{w}\|_{H(\mathrm{curl};\Omega)} \le C_m\|\varepsilon^{\frac12}\boldsymbol{u}\|_{0,\Omega}. \tag{5.3}$$

Multiplying the first equation of the auxiliary problem by $\boldsymbol{u}$ and integrating by parts over each element, we get

$$\|\varepsilon^{\frac{1}{2}}\boldsymbol{u}\|_{0,\Omega}^2 = \int_\Omega \boldsymbol{w}_0 \cdot \nabla_h \times \boldsymbol{u}\, d\boldsymbol{x} - \int_{\mathcal{F}_h} \boldsymbol{w}_0 \cdot [\![\boldsymbol{u}]\!]_T\, ds + \int_\Omega \psi \nabla_h \cdot (\varepsilon \boldsymbol{u})\, d\boldsymbol{x} - \int_{\mathcal{F}_h^{\mathcal{I}}} \{\!\{\psi\}\!\} [\![\varepsilon \boldsymbol{u}]\!]_N\, ds.$$

Since $(\boldsymbol{u}, \lambda) \in \mathrm{Ker}(B_h)$, we have $B_h(\boldsymbol{u}, \lambda; \psi_h) = 0$ for any $\psi_h \in Q_h$ and obtain, from integration by parts and the fact that $[\![\psi]\!]_N = 0$ on $\mathcal{F}_h$,

$$\|\varepsilon^{\frac{1}{2}}\boldsymbol{u}\|_{0,\Omega}^2 = \int_\Omega \boldsymbol{w}_0 \cdot \nabla_h \times \boldsymbol{u}\, d\boldsymbol{x} - \int_{\mathcal{F}_h} \boldsymbol{w}_0 \cdot [\![\boldsymbol{u}]\!]_T\, ds + \int_\Omega (\psi - \psi_h) \nabla_h \cdot (\varepsilon \boldsymbol{u})\, d\boldsymbol{x} - \int_{\mathcal{F}_h^{\mathcal{I}}} \{\!\{\psi - \psi_h\}\!\} [\![\varepsilon \boldsymbol{u}]\!]_N\, ds - \int_{\mathcal{F}_h} \mathtt{eh}^{-1} \lambda \cdot [\![\psi - \psi_h]\!]_N\, ds.$$

Using (2.2) and (5.3), we have

$$\left| \int_\Omega \boldsymbol{w}_0 \cdot \nabla_h \times \boldsymbol{u}\, d\boldsymbol{x} \right| \le C_m \|\boldsymbol{w}_0\|_{1,\Omega} \|\mu^{-\frac{1}{2}} \nabla_h \times \boldsymbol{u}\|_{0,\Omega} \le C_m \|\varepsilon^{\frac{1}{2}}\boldsymbol{u}\|_{0,\Omega} |\boldsymbol{u}|_{\boldsymbol{V}(h)}.$$

Furthermore, using trace inequalities, (2.2) and (5.3), we obtain

$$\begin{aligned}\left| \int_{\mathcal{F}_h} \boldsymbol{w}_0 \cdot [\![\boldsymbol{u}]\!]_T\, ds \right| &\le C \left(\sum_{K \in \mathcal{T}_h} h_K \mu_K \|\boldsymbol{w}_0\|_{0,\partial K}^2 \right)^{\frac{1}{2}} \|\mathtt{m}^{-\frac{1}{2}} \mathtt{h}^{-\frac{1}{2}} [\![\boldsymbol{u}]\!]_T\|_{0,\mathcal{F}_h} \\ &\le C_m \|\boldsymbol{w}_0\|_{1,\Omega} \|\mathtt{m}^{-\frac{1}{2}} \mathtt{h}^{-\frac{1}{2}} [\![\boldsymbol{u}]\!]_T\|_{0,\mathcal{F}_h} \le C_m \|\varepsilon^{\frac{1}{2}}\boldsymbol{u}\|_{0,\Omega} |\boldsymbol{u}|_{\boldsymbol{V}(h)}.\end{aligned}$$

For the other terms, we choose ψ_h as the L^2-projection of ψ on Q_h. Since ε is piecewise constant, we have $\int_\Omega (\psi - \psi_h) \nabla_h \cdot (\varepsilon \boldsymbol{u})\, d\boldsymbol{x} = 0$. Then, using the Cauchy–Schwarz inequality, the definition of $\mathtt{e}$, and standard approximation properties of the L^2-projection, we obtain

$$\begin{aligned}\left| \int_{\mathcal{F}_h^{\mathcal{I}}} \{\!\{\psi - \psi_h\}\!\} [\![\varepsilon \boldsymbol{u}]\!]_N\, ds \right| &\le C \left(\sum_{K \in \mathcal{T}_h} \varepsilon_K h_K^{-1} \|\psi - \psi_h\|_{0,\partial K}^2 \right)^{\frac{1}{2}} |\boldsymbol{u}|_{\boldsymbol{V}(h)} \\ &\le C \left(\sum_{K \in \mathcal{T}_h} \varepsilon_K \|\nabla \psi\|_{0,K}^2 + \varepsilon_K \|\psi\|_{0,K}^2 \right)^{\frac{1}{2}} |\boldsymbol{u}|_{\boldsymbol{V}(h)} \le C_m \|\varepsilon^{\frac{1}{2}}\boldsymbol{u}\|_{0,\Omega} |\boldsymbol{u}|_{\boldsymbol{V}(h)}.\end{aligned}$$

Similarly,

$$\begin{aligned}\left| \int_{\mathcal{F}_h} \mathtt{eh}^{-1} \lambda \cdot [\![\psi - \psi_h]\!]_N\, ds \right| &\le C \left(\sum_{K \in \mathcal{T}_h} \varepsilon_K h_K^{-1} \|\psi - \psi_h\|_{0,\partial K}^2 \right)^{\frac{1}{2}} \left(\int_{\mathcal{F}_h} \mathtt{eh}^{-1} \lambda^2\, ds \right)^{\frac{1}{2}} \\ &\le C_m \|\varepsilon^{\frac{1}{2}}\boldsymbol{u}\|_{0,\Omega} \|\lambda\|_{M_h}.\end{aligned}$$

The above computations show that $\|\varepsilon^{\frac{1}{2}}\boldsymbol{u}\|_{0,\Omega} \le C_m\, |(\boldsymbol{u}, \lambda)|_{\boldsymbol{W}(h)}$. Combining this with (5.1) and the definition of $\|(\boldsymbol{u}, \lambda)\|_{\boldsymbol{W}(h)}$ completes the proof. ∎

5.3. Inf-sup condition. In this section, we prove the inf-sup condition for the form B_h. Our proof is inspired by recent ideas from [8] used in the analysis of stabilized mixed methods. We will make use of the following crucial norm equivalence result. To this end, let Q_h^c be the subspace $Q_h \cap H_0^1(\Omega)$ of Q_h, and let $Q_h^\perp$ be the orthogonal complement in Q_h of Q_h^c, with respect to the norm $\|\cdot\|_{Q(h)}$. We observe that $\|q\|_{Q_h^\perp} = \|\mathtt{e}^{\frac{1}{2}}\mathtt{h}^{-\frac{1}{2}}[\![q]\!]_N\|_{0,\mathcal{F}_h}$ is a norm on $Q_h^\perp$. Indeed, if $q \in Q_h^\perp$ and $\|q\|_{Q_h^\perp} = 0$, then $q \in Q_h^\perp \cap Q_h^c = \{0\}$. The norms $\|\cdot\|_{Q_h^\perp}$ and $\|\cdot\|_{Q(h)}$ are equivalent in $Q_h^\perp$.

THEOREM 5.3. *There are positive constants C_1 and C_2, independent of the mesh size and the coefficients μ and ε, such that $C_1\,\|q\|_{Q(h)} \le \|q\|_{Q_h^\perp} \le C_2\,\|q\|_{Q(h)}$ for any $q \in Q_h^\perp$.*

Proof. Step 1. The following approximation result holds: for any $q \in Q_h$,

$$\inf_{q^c \in Q_h^c} \|\varepsilon^{\frac{1}{2}}\nabla_h(q - q^c)\|_{0,\Omega} \le C\|\mathtt{e}^{\frac{1}{2}}\mathtt{h}^{-\frac{1}{2}}[\![q]\!]_N\|_{0,\mathcal{F}_h}, \tag{5.4}$$

with a constant $C > 0$ independent of the mesh size and the coefficients μ and ε. This result has been proved in [21, Theorems 2.2 and 2.3] for simplicial meshes. The proof there can be easily generalized to the meshes considered in this paper and readily gives the independence of the constant on the coefficients μ and ε; we refer the reader to [18, Appendix A] for these technical details.

Step 2. The inequality on the right-hand side of the norm equivalence is trivially satisfied with $C_2 = 1$. To show the bound on the left-hand side, let $P_h : Q_h \to Q_h^\perp$ denote the $Q(h)$-orthogonal projection. For $q \in Q_h$, we then have $\|P_h q\|_{Q(h)} = \inf_{\overline{q} \in Q_h^c} \|q - \overline{q}\|_{Q(h)} \le C\|P_h q\|_{Q_h^\perp}$. Here, we have used properties of orthogonal projections, the approximation result (5.4), the fact that $[\![q]\!]_N = [\![P_h q]\!]_N$, and the definition of $\|\cdot\|_{Q_h^\perp}$. Since P_h is surjective, the equivalence follows. ∎

Our main result of this section is the following inf-sup condition.

PROPOSITION 5.4. *We have*

$$\inf_{0 \ne q \in Q_h} \sup_{\mathbf{0} \ne (\boldsymbol{v},\nu) \in \boldsymbol{W}_h} \frac{B_h(\boldsymbol{v}, \nu; q)}{\|q\|_{Q(h)}\|(\boldsymbol{v},\nu)\|_{\boldsymbol{W}(h)}} \ge \kappa > 0,$$

for a constant κ, independent of the mesh size and the coefficients μ and ε.

Proof. Fix $0 \ne q \in Q_h$ arbitrary and consider its $Q(h)$-orthogonal decomposition as $q = q_0 \oplus q_1$, with $q_0 \in Q_h^c$ and $q_1 \in Q_h^\perp$. By choosing $\boldsymbol{v}_0 = -\nabla q_0 \in \boldsymbol{V}_h \cap H(\mathrm{curl}^0;\Omega) \cap H_0(\mathrm{curl};\Omega)$, we have

$$B_h(\boldsymbol{v}_0, 0; q_0) = \|\varepsilon^{\frac{1}{2}}\nabla q_0\|_{0,\Omega}^2 = \|q_0\|_{Q(h)}^2. \tag{5.5}$$

Furthermore, by the definition of $\mathtt{e}$,

$$\begin{aligned} \|(\boldsymbol{v}_0, 0)\|_{\boldsymbol{W}(h)}^2 &= \|\mathtt{e}^{-\frac{1}{2}}\mathtt{h}^{\frac{1}{2}}[\![\varepsilon \boldsymbol{v}_0]\!]_N\|_{0,\mathcal{F}_h^{\mathcal{I}}}^2 + \|\varepsilon^{\frac{1}{2}}\boldsymbol{v}_0\|_{0,\Omega}^2 \\ &\le C \sum_{K \in \mathcal{T}_h} \varepsilon_K h_K \|\nabla q_0\|_{0,\partial K}^2 + \|\varepsilon^{\frac{1}{2}}\nabla q_0\|_{0,\Omega}^2 \le C\|\varepsilon^{\frac{1}{2}}\nabla q_0\|_{0,\Omega}^2 = C\|q_0\|_{Q(h)}^2. \end{aligned} \tag{5.6}$$

Here, we have used the discrete trace inequality $\|\nabla\varphi\|_{0,\partial K} \le C h_K^{-\frac{1}{2}}\|\nabla\varphi\|_{0,K}$, valid for polynomials $\varphi \in \mathcal{S}^\ell(K)$, with a constant $C > 0$ that is independent of the local mesh size h_K. Next, setting $\nu_1 = -[\![q_1]\!]_N$ gives

$$B_h(\mathbf{0}, \nu_1; q_1) = \gamma \int_{\mathcal{F}_h} \mathtt{e}\mathtt{h}^{-1}[\![q_1]\!]_N^2\, ds \ge \gamma C_1^2 \|q_1\|_{Q(h)}^2, \quad \|(\mathbf{0}, \nu_1)\|_{\boldsymbol{W}(h)} \le C_2\|q_1\|_{Q(h)}, \tag{5.7}$$

where C_1 and C_2 are the constants in the norm equivalence of Theorem 5.3.

Then we set $(\boldsymbol{v},\nu) = (\boldsymbol{v}_0,0)+\delta(\boldsymbol{0},\nu_1)$, with a parameter $\delta>0$ still at our disposal. Since $B_h(\boldsymbol{0},\nu_1;q_0)=0$, we obtain from (5.5) and (5.7)

$$\begin{aligned} B_h(\boldsymbol{v},\nu;q) &= B_h(\boldsymbol{v}_0,0;q_0) + B_h(\boldsymbol{v}_0,0;q_1) + \delta B_h(\boldsymbol{0},\nu_1;q_1) \\ &\geq \|q_0\|^2_{Q(h)} + \delta\gamma C_1^2 \|q_1\|^2_{Q(h)} - |B_h(\boldsymbol{v}_0,0;q_1)|. \end{aligned}$$

Combining the continuity of B_h (see Proposition 5.1) with a weighted Cauchy–Schwarz inequality and (5.6), we obtain, for any $\zeta>0$,

$$|B_h(\boldsymbol{v}_0,0;q_1)| \leq C\zeta \|\boldsymbol{v}_0\|^2_{\boldsymbol{V}(h)} + \frac{C}{\zeta}\|q_1\|^2_{Q(h)} \leq C\zeta\|q_0\|^2_{Q(h)} + \frac{C}{\zeta}\|q_1\|^2_{Q(h)}.$$

Hence, by suitably choosing δ and ζ, we have

$$B_h(\boldsymbol{v},\nu;q) \geq \kappa_1\big[\|q_0\|^2_{Q(h)} + \|q_1\|^2_{Q(h)}\big] = \kappa_1\|q\|^2_{Q(h)}, \tag{5.8}$$

where we used the orthogonality of the decomposition of q. From (5.6) and (5.7),

$$\|(\boldsymbol{v},\nu)\|_{\boldsymbol{W}(h)} \leq \kappa_2\|q\|_{Q(h)}. \tag{5.9}$$

The constants κ_1 and κ_2 in (5.8) and (5.9) are independent of the mesh size and the coefficients μ and ε. The proposition follows from (5.8) and (5.9), with $\kappa=\kappa_1/\kappa_2$. ∎

6. Error estimates. In this section, we prove the error estimates stated in Theorems 3.3, 3.4, and 3.6.

6.1. Abstract error estimates. We start by deriving abstract error bounds. To this end, for the analytical solution $(\boldsymbol{u},p)$ to (2.3)–(2.4), we define the residuals

$$R^1_h(\boldsymbol{u},p;\boldsymbol{v},\nu) = A_h(\boldsymbol{u},0;\boldsymbol{v},\nu) + B_h(\boldsymbol{v},\nu;p) - f_h(\boldsymbol{v}) \quad \text{and} \quad R^2_h(\boldsymbol{u};q) = B_h(\boldsymbol{u},0;q)$$

for all $(\boldsymbol{v},\nu)\in\boldsymbol{W}_h$ and $q\in Q_h$ and set

$$\mathcal{R}^1_h(\boldsymbol{u},p) = \sup_{\boldsymbol{0}\neq(\boldsymbol{v},\nu)\in\boldsymbol{W}_h} \frac{|R^1_h(\boldsymbol{u},p;\boldsymbol{v},\nu)|}{\|(\boldsymbol{v},\nu)\|_{\boldsymbol{W}(h)}}, \qquad \mathcal{R}^2_h(\boldsymbol{u}) = \sup_{0\neq q\in Q_h} \frac{|R^2_h(\boldsymbol{u};q)|}{\|q\|_{Q(h)}}.$$

In the following theorem we present abstract error estimates for our DG method. These error bounds are obtained by extending the standard conforming mixed finite element theory [7] to the setting considered here and taking into account the residual terms arising from the nonconsistency of the perturbed formulation.

THEOREM 6.1. *There exist positive constants C such that*

$$\begin{aligned} \|(\boldsymbol{u}-\boldsymbol{u}_h,\lambda_h)\|_{\boldsymbol{W}(h)} \leq C\max\{1,b^{-1}\}\Big[&\inf_{\boldsymbol{v}\in\boldsymbol{V}_h}\|\boldsymbol{u}-\boldsymbol{v}\|_{\boldsymbol{V}(h)} \\ &+ \inf_{q\in Q_h}\|p-q\|_{Q(h)} + \mathcal{R}^1_h(\boldsymbol{u},p) + \mathcal{R}^2_h(\boldsymbol{u})\Big], \\ \|p-p_h\|_{Q(h)} \leq C\Big[&\inf_{q\in Q_h}\|p-q\|_{Q(h)} + \|(\boldsymbol{u}-\boldsymbol{u}_h,\lambda_h)\|_{\boldsymbol{W}(h)} + \mathcal{R}^1_h(\boldsymbol{u},p)\Big]. \end{aligned}$$

Here, b, the ellipticity constant from Proposition 5.2*, depends on the bounds in* (2.2) *but is independent of the mesh size. The constants C depend on the continuity constants a_1 and a_2 in Proposition* 5.1 *and the inf-sup constant κ from Proposition* 5.4 *but are independent of the mesh size and the coefficients μ and ε.*

Proof. By the triangle inequality and the definition of $\|(\cdot,\cdot)\|_{\boldsymbol{W}(h)}$, we have

$$\|(\boldsymbol{u}-\boldsymbol{u}_h,\lambda_h)\|_{\boldsymbol{W}(h)} \le \|(\boldsymbol{u}-\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)} + \|(\boldsymbol{v}-\boldsymbol{u}_h,\eta-\lambda_h)\|_{\boldsymbol{W}(h)} \tag{6.1}$$

for any $(\boldsymbol{v},\eta)\in\boldsymbol{W}_h$. First, we take $(\boldsymbol{v},\eta)\in\mathrm{Ker}(\mathrm{B_h})$. Since $(\boldsymbol{v}-\boldsymbol{u}_h,\eta-\lambda_h)\in\mathrm{Ker}(\mathrm{B_h})$, employing the ellipticity property of Proposition 5.2 and the definition of R_h^1, we have

$$\begin{aligned} b\|(\boldsymbol{v}-\boldsymbol{u}_h,\eta-\lambda_h)\|^2_{\boldsymbol{W}(h)} &\le A_h(\boldsymbol{v}-\boldsymbol{u}_h,\eta-\lambda_h;\boldsymbol{v}-\boldsymbol{u}_h,\eta-\lambda_h)\\ &= A_h(\boldsymbol{v}-\boldsymbol{u},\eta;\boldsymbol{v}-\boldsymbol{u}_h,\eta-\lambda_h)\\ &\quad - B_h(\boldsymbol{v}-\boldsymbol{u}_h,\eta-\lambda_h;p-q) + R_h^1(\boldsymbol{u},p;\boldsymbol{v}-\boldsymbol{u}_h,\eta-\lambda_h) \end{aligned}$$

for any $q\in Q_h$. From the continuity properties of Proposition 5.1, the definition of the norm $\|(\cdot,\cdot)\|_{\boldsymbol{W}(h)}$, and (6.1), we have

$$\begin{aligned} \|(\boldsymbol{u}-\boldsymbol{u}_h,\lambda_h)\|_{\boldsymbol{W}(h)} &\le \Big(1+\frac{a_1}{b}\Big)\inf_{(\boldsymbol{v},\eta)\in\mathrm{Ker}(\mathrm{B_h})}\|(\boldsymbol{u}-\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)}\\ &\quad + \frac{a_2}{b}\inf_{q\in Q_h}\|p-q\|_{Q(h)} + \frac{1}{b}\mathcal{R}_h^1(\boldsymbol{u},p). \end{aligned} \tag{6.2}$$

Next, we prove that

$$\inf_{(\boldsymbol{v},\eta)\in\mathrm{Ker}(\mathrm{B_h})}\|(\boldsymbol{u}-\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)} \le \Big(1+\frac{a_2}{\kappa}\Big)\inf_{(\boldsymbol{v},\eta)\in\boldsymbol{W}_h}\|(\boldsymbol{u}-\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)} + \frac{1}{\kappa}\mathcal{R}_h^2(\boldsymbol{u}). \tag{6.3}$$

To this end, let $(\boldsymbol{v},\eta)$ be any element of $\boldsymbol{W}_h$, and consider the following problem: find $(\boldsymbol{w},\nu)\in\boldsymbol{W}_h$ such that

$$B_h(\boldsymbol{w},\nu;q) = B_h(\boldsymbol{u}-\boldsymbol{v},-\eta;q) - R_h^2(\boldsymbol{u},q) \qquad \forall q\in Q_h. \tag{6.4}$$

Problem (6.4) admits solutions in $\boldsymbol{W}_h$ that are unique up to elements in the kernel of B_h. The discrete inf-sup condition of Proposition 5.4 guarantees the existence of a solution $(\boldsymbol{w},\nu)$ satisfying

$$\begin{aligned} \|(\boldsymbol{w},\nu)\|_{\boldsymbol{W}(h)} &\le \frac{1}{\kappa}\Big[\sup_{q\in Q_h}\frac{|B_h(\boldsymbol{u}-\boldsymbol{v},-\eta;q)|}{\|q\|_{Q(h)}} + \sup_{q\in Q_h}\frac{|R_h^2(\boldsymbol{u},q)|}{\|q\|_{Q(h)}}\Big]\\ &\le \frac{a_2}{\kappa}\|(\boldsymbol{u}-\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)} + \frac{1}{\kappa}\mathcal{R}_h^2(\boldsymbol{u}), \end{aligned} \tag{6.5}$$

where we have used the continuity of B_h, the definition of the norm $\|(\cdot,\cdot)\|_{\boldsymbol{W}(h)}$, and the definition of $\mathcal{R}_h^2$. From (6.4), $B_h(\boldsymbol{w}+\boldsymbol{v},\nu+\eta;q)=0$, for any $q\in Q_h$, so that $(\boldsymbol{w}+\boldsymbol{v},\nu+\eta)\in\mathrm{Ker}(B_h)$. Therefore, since

$$\|(\boldsymbol{u}-(\boldsymbol{v}+\boldsymbol{w}),\eta+\nu)\|_{\boldsymbol{W}(h)} \le \|(\boldsymbol{u}-\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)} + \|(\boldsymbol{w},\nu)\|_{\boldsymbol{W}(h)},$$

for any $(\boldsymbol{v},\eta)\in\boldsymbol{W}_h$, taking into account (6.5), we obtain (6.3). This, together with (6.2), yields

$$\begin{aligned} \|(\boldsymbol{u}-\boldsymbol{u}_h,\lambda_h)\|_{\boldsymbol{W}(h)} \le C\max\{1,b^{-1}\}\Big[&\inf_{(\boldsymbol{v},\eta)\in\boldsymbol{W}_h}\|(\boldsymbol{u}-\boldsymbol{v},\eta)\|_{\boldsymbol{W}(h)}\\ &+ \inf_{q\in Q_h}\|p-q\|_{Q(h)} + \mathcal{R}_h^1(\boldsymbol{u},p) + \mathcal{R}_h^2(\boldsymbol{u})\Big], \end{aligned}$$

where the constant C depends on a_1, a_2, and κ. Choosing $\eta = 0$ gives the error bound for $(\boldsymbol{u} - \boldsymbol{u}_h, \lambda_h)$.

We now turn to the bound for $p - p_h$. Again by the triangle inequality, we have

$$\|p - p_h\|_{Q(h)} \leq \|p - q\|_{Q(h)} + \|q - p_h\|_{Q(h)} \tag{6.6}$$

for any $q_h \in Q_h$. Since

$$A_h(\boldsymbol{u} - \boldsymbol{u}_h, -\lambda_h; \boldsymbol{v}, \eta) + B_h(\boldsymbol{v}, \eta; p - q) + B_h(\boldsymbol{v}, \eta; q - p_h) = R_h^1(\boldsymbol{u}, p; \boldsymbol{v}, \eta),$$

for any $(\boldsymbol{v}, \eta) \in \boldsymbol{W}_h$, the discrete inf-sup condition of Proposition 5.4 gives

$$\begin{aligned}
\|q - p_h\|_{Q(h)} &\leq \frac{1}{\kappa} \sup_{(\boldsymbol{0},0) \neq (\boldsymbol{v},\eta) \in \boldsymbol{W}_h} \frac{B_h(\boldsymbol{v}, \eta; q - p_h)}{\|(\boldsymbol{v}, \eta)\|_{\boldsymbol{W}(h)}} \\
&= \frac{1}{\kappa} \sup_{(\boldsymbol{0},0) \neq (\boldsymbol{v},\eta) \in \boldsymbol{W}_h} \frac{-A_h(\boldsymbol{u} - \boldsymbol{u}_h, -\lambda_h; \boldsymbol{v}, \eta) - B_h(\boldsymbol{v}, \eta; p - q) + R_h^1(\boldsymbol{u}, p; \boldsymbol{v}, \eta)}{\|(\boldsymbol{v}, \eta)\|_{\boldsymbol{W}(h)}} \\
&\leq \frac{a_1}{\kappa} \|(\boldsymbol{u} - \boldsymbol{u}_h, \lambda_h)\|_{\boldsymbol{W}(h)} + \frac{a_2}{\kappa} \|p - q\|_{Q(h)} + \frac{1}{\kappa} \mathcal{R}_h^1(\boldsymbol{u}, p).
\end{aligned}$$

This, together with (6.6), gives the bound for $p - p_h$. ∎

6.2. Proof of Theorems 3.3 and 3.4. We are now ready to prove our main results by making explicit the abstract error estimates of Theorem 6.1. First, we derive bounds on the residuals; we note that the residuals are optimally convergent on possibly nonconforming meshes under minimal smoothness assumptions as in Theorem 3.4, thereby covering both the cases of Theorems 3.3 and 3.4.

PROPOSITION 6.2. *Let $\mathcal{T}_h$ be a possibly nonconforming mesh satisfying restrictions* (i) *and* (ii) *in section* 3.3. *Assume the analytical solution $(\boldsymbol{u}, p)$ of* (2.1) *satisfies $\varepsilon \boldsymbol{u} \in H^s(\mathcal{T}_h)^3$ and $\mu^{-1} \nabla \times \boldsymbol{u} \in H^s(\mathcal{T}_h)^3$ for $s > \frac{1}{2}$. Then we have*

$$\mathcal{R}_h^1(\boldsymbol{u}, p) + \mathcal{R}_h^2(\boldsymbol{u}) \leq C h^{\min\{s, \ell+1\}} \big[\|\varepsilon \boldsymbol{u}\|_{s, \mathcal{T}_h} + \|\mu^{-1} \nabla \times \boldsymbol{u}\|_{s, \mathcal{T}_h} \big],$$

with a constant $C > 0$, independent of the mesh size.

Proof. First, for $\boldsymbol{v} \in \boldsymbol{V}_h$, $\eta \in M_h$, we have

$$|R_h^1(\boldsymbol{u}, p; \boldsymbol{v}, \eta)| = \left| \int_{\mathcal{F}_h} \{\!\{ \mu^{-1} \nabla \times \boldsymbol{u} - \boldsymbol{\Pi}_{\boldsymbol{V}_h} (\mu^{-1} \nabla \times \boldsymbol{u}) \}\!\} \cdot [\![\boldsymbol{v}]\!]_T \, ds \right|,$$

with $\boldsymbol{\Pi}_{\boldsymbol{V}_h}$ denoting the L^2-projection onto $\boldsymbol{V}_h$. This can be easily proved by employing integration by parts, the properties of the L^2-projection, and taking into account the first equation in (2.1). From the Cauchy–Schwarz inequality and standard approximation properties, we obtain

$$\begin{aligned}
|R_h^1(\boldsymbol{u}, \lambda, p; \boldsymbol{v}, \eta)| &\leq C \left(\sum_{K \in \mathcal{T}_h} h_K \|\mu^{-1} \nabla \times \boldsymbol{u} - \boldsymbol{\Pi}_{\boldsymbol{V}_h} (\mu^{-1} \nabla \times \boldsymbol{u})\|_{0, \partial K}^2 \right)^{\frac{1}{2}} \|(\boldsymbol{v}, \eta)\|_{\boldsymbol{W}(h)} \\
&\leq C h^{\min\{s, \ell+1\}} \|\mu^{-1} \nabla \times \boldsymbol{u}\|_{s, \mathcal{T}_h} \|(\boldsymbol{v}, \eta)\|_{\boldsymbol{W}(h)}.
\end{aligned}$$

Furthermore, for $q \in Q_h$,

$$|R_h^2(\boldsymbol{u}; q)| = \left| \int_{\mathcal{F}_h} \{\!\{ \varepsilon \boldsymbol{u} - \boldsymbol{\Pi}_{\boldsymbol{V}_h} (\varepsilon \boldsymbol{u}) \}\!\} \cdot [\![q]\!]_N \, ds \right|.$$

Hence,

$$|R_h^2(\boldsymbol{u};q)| \le C\left(\sum_{K\in\mathcal{T}_h} h_K \|\varepsilon\boldsymbol{u} - \boldsymbol{\Pi}_{\boldsymbol{V}_h}(\varepsilon\boldsymbol{u})\|_{0,\partial K}^2\right)^{\frac{1}{2}} \|q\|_{Q(h)} \le C h^{\min\{s,\ell+1\}} \|\varepsilon\boldsymbol{u}\|_{s,\mathcal{T}_h} \|q\|_{Q(h)}.$$

This completes the proof (the constant C in these estimates actually depends on the bound (2.2) on the coefficients μ and ε). ∎

Next, we prove the result of Theorems 3.3 and 3.4.

Proof of Theorem 3.3. Let $(\boldsymbol{u}_h, p_h, \lambda_h)$ be the solution of the auxiliary mixed system in (4.1)–(4.2). We apply Theorem 6.1 and bound all the terms in the abstract error bounds there. To this end, we set $\boldsymbol{v} = \boldsymbol{\Pi}_{\boldsymbol{V}_h}\boldsymbol{u}$ and $q = \Pi_{Q_h} p$, with $\boldsymbol{\Pi}_{\boldsymbol{V}_h}$ and Π_{Q_h} denoting the L^2-projections onto $\boldsymbol{V}_h$ and Q_h, respectively. Standard approximation properties, together with the bounded variation of the mesh size, give $\|\boldsymbol{u} - \boldsymbol{\Pi}_{\boldsymbol{V}_h}\boldsymbol{u}\|_{\boldsymbol{V}(h)} \le C\, h^{\min\{s,\ell\}} \|\boldsymbol{u}\|_{s+1,\mathcal{T}_h}$. The additional smoothness assumption on $\boldsymbol{u}$ made in this case is required for the estimate of the term containing the tangential jumps. Similarly, we have $\|p - \Pi_{Q_h} p\|_{Q(h)} \le C\, h^{\min\{s,\ell\}} \|p\|_{s+1,\mathcal{T}_h}$. The constants C in the previous estimates depend on the bounds (2.2) on the coefficients μ and ε. Inserting these estimates, together with the residual estimates of Proposition 6.2, in Theorem 6.1 gives the result. ∎

Proof of Theorem 3.4. Let $(\boldsymbol{u}_h, p_h, \lambda_h)$ be the solution of the auxiliary mixed system in (4.1)–(4.2). As in the proof of Theorem 3.3, we apply Theorem 6.1. On conforming meshes, we can choose $\boldsymbol{v} = \boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u} \in H(\mathrm{curl};\Omega)$ as the standard conforming Nédélec interpolant of $\boldsymbol{u}$ of the second type; see [23]. Thereby, we have $[\![\boldsymbol{v}]\!]_T = \boldsymbol{0}$ on $\mathcal{F}_h^{\mathcal{I}}$, and from the approximation results proved in [2] for tetrahedra, but also valid for parallelepipeds, we get

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u}\|_{0,K} + \|\nabla\times(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u})\|_{0,K} \le C\, h_K^{\min\{s,\ell\}} \big[\|\boldsymbol{u}\|_{s,K} + \|\nabla\times\boldsymbol{u}\|_{s,K}\big].$$

Hence,

$$\begin{aligned}\|\varepsilon^{\frac{1}{2}}(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u})\|_{0,\Omega} + \|\mu^{-\frac{1}{2}}\nabla\times(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u})\|_{0,\Omega} \\ \le\ C\, h^{\min\{s,\ell\}} \big[\|\varepsilon\boldsymbol{u}\|_{s,\mathcal{T}_h} + \|\mu^{-1}\nabla\times\boldsymbol{u}\|_{s,\mathcal{T}_h}\big],\end{aligned} \tag{6.7}$$

with C also depending on the bounds (2.2) on the coefficients μ and ε. On a boundary face $f \subset \mathcal{F}_h^{\mathcal{D}}$, the tangential field $\boldsymbol{n}\times\boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u}$ coincides with $\boldsymbol{\Pi}_{\mathrm{div}}(\boldsymbol{n}\times\boldsymbol{u})$, where $\boldsymbol{\Pi}_{\mathrm{div}}$ is the two-dimensional $H(\mathrm{div})$-conforming Nédélec interpolation operator of the second type. In particular, $\boldsymbol{\Pi}_{\mathrm{div}}$ reproduces polynomial tangential fields of degree ℓ on f; see [23] for details. From a standard scaling argument, we obtain

$$\|\boldsymbol{n}\times(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u})\|_{0,f} \le C\, \mathtt{h}_{|f}^{\min\{s,\ell\}+\frac{1}{2}} \|\boldsymbol{g}\|_{s+\frac{1}{2},f}.$$

Therefore, summing over all faces yields

$$\|\mathtt{m}^{-\frac{1}{2}}\mathtt{h}^{-\frac{1}{2}}(\boldsymbol{n}\times\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{curl}}(\boldsymbol{n}\times\boldsymbol{u}))\|_{0,\mathcal{F}_h^{\mathcal{D}}} \le C\, h^{\min\{s,\ell\}} \|\boldsymbol{g}\|_{s+\frac{1}{2},\mathcal{F}_h^{\mathcal{D}}}. \tag{6.8}$$

Moreover, since $\boldsymbol{u} \in H(\mathrm{div}_\varepsilon^0;\Omega)$, we have for an interior face f, shared by K_f and K_f',

$$\|\mathtt{e}^{-\frac{1}{2}}\mathtt{h}^{\frac{1}{2}}[\![\varepsilon(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u})]\!]_N\|_{0,f} \le C\, h^{\min\{s,\ell\}} \big[\|\boldsymbol{u}\|_{s,K_f\cup K_f'} + \|\nabla\times\boldsymbol{u}\|_{s,K_f\cup K_f'}\big].$$

This bound follows from [25, Lemma 23]. Hence,

$$\|\mathtt{e}^{-\frac{1}{2}}\mathtt{h}^{\frac{1}{2}}[\![\varepsilon(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{curl}}\boldsymbol{u})]\!]_N\|_{0,\mathcal{F}_h^{\mathcal{I}}} \le C\, \big[\|\varepsilon\boldsymbol{u}\|_{s,\mathcal{T}_h} + \|\mu^{-1}\nabla\times\boldsymbol{u}\|_{s,\mathcal{T}_h}\big]. \tag{6.9}$$

Moreover, choosing $q = \Pi_{H^1} p$ in Theorem 6.1, where Π_{H^1} is a standard H^1-projector, we have $[\![q]\!]_N = 0$ on $\mathcal{F}_h$ and

$$\|p - \Pi_{H^1} p\|_{Q(h)} \leq C\, h^{\min\{s,\ell\}} \|p\|_{s+1,\mathcal{T}_h}. \tag{6.10}$$

Combining (6.7)–(6.10) with the residual estimates in Proposition 6.2 yields

$$\begin{aligned}\inf_{\boldsymbol{v}\in\boldsymbol{V}_h} \|\boldsymbol{u}-\boldsymbol{v}\|_{\boldsymbol{V}(h)} + \inf_{q\in Q_h} \|p-q\|_{Q(h)} + \mathcal{R}_h^1(\boldsymbol{u},p) + \mathcal{R}_h^2(\boldsymbol{u}) \\ \leq C\, h^{\min\{s,\ell\}} \left[\|\varepsilon\boldsymbol{u}\|_{s,\mathcal{T}_h} + \|\mu^{-1}\nabla\times\boldsymbol{u}\|_{s,\mathcal{T}_h} + \|p\|_{s+1,\mathcal{T}_h} + \|\boldsymbol{g}\|_{s+\frac{1}{2},\mathcal{F}_h^{\mathcal{D}}}\right],\end{aligned}$$

with C also depending on the bounds (2.2) on the coefficients μ and ε. Combining these estimates with the residual estimates of Proposition 6.2 gives the result. ∎

6.3. Proof of Theorem 3.6. To prove Theorem 3.6, we first recall that

$$B_h(\boldsymbol{u}_h, \lambda_h; q) = 0 \qquad \forall q \in Q_h.$$

Note that $\lambda_h = [\![p]\!]_N$ (cf. Proposition 4.1); thereby, for an element $K \in \mathcal{T}_h$, we have

$$-\int_K \varepsilon\boldsymbol{u}_h \cdot \nabla q \, d\boldsymbol{x} + \int_{\partial K} \{\!\{\varepsilon\boldsymbol{u}_h\}\!\} \cdot (q\boldsymbol{n}_K)\, ds - \gamma \frac{\mathtt{e}}{\mathtt{h}} \int_{\partial K} [\![p_h]\!] \cdot (q\boldsymbol{n}_K)\, ds = 0 \tag{6.11}$$

for all $q \in \mathcal{S}^\ell(K)$. Employing integration by parts and the identity (6.11), we get, after some elementary manipulations, the following:

$$\begin{aligned} h_K \int_K \nabla\cdot(\varepsilon\boldsymbol{u}_h) q \, d\boldsymbol{x} &= -h_K \int_K \varepsilon\boldsymbol{u}_h \cdot \nabla q \, d\boldsymbol{x} + h_K \int_{\partial K} \varepsilon\boldsymbol{u}_h \cdot (q\boldsymbol{n}_K)\, ds \\ &= \frac{1}{2} h_K \int_{\partial K^0} [\![\varepsilon\boldsymbol{u}_h]\!]_N q \, ds + h_K \gamma \frac{\mathtt{e}}{\mathtt{h}} \int_{\partial K} [\![p_h]\!] \cdot (q\boldsymbol{n}_K)\, ds, \end{aligned}$$

for $q \in \mathcal{S}^\ell(K)$, where $\partial K^0 = \partial K \cap \mathcal{F}_h^{\mathcal{I}}$. The Cauchy–Schwarz inequality gives

$$\left| h_K \int_K \nabla\cdot(\varepsilon\boldsymbol{u}_h) q \, d\boldsymbol{x} \right| \leq C h_K^{\frac{1}{2}} \|q\|_{0,\partial K} \left(\int_{\partial K^0} \mathtt{h} [\![\varepsilon\boldsymbol{u}_h]\!]_N^2 \, ds + \int_{\partial K} \frac{\mathtt{e}}{\mathtt{h}} [\![p_h]\!]_N^2 \, ds \right)^{\frac{1}{2}},$$

with a constant $C > 0$, independent of the mesh size. Here, we used shape-regularity and bounded variation properties of the mesh. Employing the discrete trace inequality $\|q\|_{0,\partial K} \leq C h_K^{-\frac{1}{2}} \|q\|_{0,K}$ for all $q \in \mathcal{S}^\ell(K)$ and the characterization

$$h_K \|\nabla\cdot(\varepsilon\boldsymbol{u}_h)\|_{0,K} = \sup_{q\in\mathcal{S}^\ell(K)} \frac{h_K \int_K \nabla\cdot(\varepsilon\boldsymbol{u}_h) q \, d\boldsymbol{x}}{\|q\|_{0,K}},$$

we obtain

$$h_K^2 \|\nabla\cdot(\varepsilon\boldsymbol{u}_h)\|_{0,K}^2 \leq C \left(\int_{\partial K^0} \mathtt{h} [\![\varepsilon\boldsymbol{u}_h]\!]_N^2 \, ds + \int_{\partial K} \frac{\mathtt{e}}{\mathtt{h}} [\![p_h]\!]_N^2 \, ds \right).$$

Summing over all elements and taking into account that the analytical solution satisfies $[\![\varepsilon\boldsymbol{u}]\!]_N = 0$ on $\mathcal{F}_h^{\mathcal{I}}$ and $[\![p]\!]_N = \boldsymbol{0}$ on $\mathcal{F}_h$ completes the proof. ∎

7. Numerical experiments. In this section, we present a series of numerical experiments to illustrate the a priori error estimates derived for the mixed DG method introduced in section 2. Here, we restrict ourselves to two-dimensional model problems with constant coefficients $\mu \equiv \varepsilon \equiv 1$. In this case, by identifying two-dimensional vector fields $\boldsymbol{u}(x,y) = (u_1(x,y), u_2(x,y))$ in $\mathbb{R}^2$ with their three-dimensional extensions $\boldsymbol{u}(x,y,z) = (u_1(x,y,0), u_2(x,y,0), 0)$ in $\mathbb{R}^3$, we deduce that

$$\nabla \times (\nabla \times \boldsymbol{u}) = \left(\frac{\partial}{\partial y} \left(\frac{\partial u_2}{\partial x} - \frac{\partial u_1}{\partial y} \right), -\frac{\partial}{\partial x} \left(\frac{\partial u_2}{\partial x} - \frac{\partial u_1}{\partial y} \right) \right).$$

On the boundary, we have $\boldsymbol{n} \times \boldsymbol{u} = \boldsymbol{u} \cdot \boldsymbol{t}$, where $\boldsymbol{t}$ is the counterclockwise oriented tangential unit vector; i.e., if $\boldsymbol{n} = (n_1, n_2)$, then $(t_1, t_2) = (-n_2, n_1)$. Hence, the Dirichlet boundary datum given in (2.1) is a scalar function g. Similarly, the tangential jumps are scalar quantities defined as $[\![\boldsymbol{u}]\!]_T = \boldsymbol{u}^+ \cdot \boldsymbol{t}^+ + \boldsymbol{u}^- \cdot \boldsymbol{t}^-$.

We shall restrict our attention to meshes consisting of quadrilateral elements only. In this case the finite element space $S^\ell(\mathcal{T}_h)$ is constructed by mapping the reference element $\hat{K} = (-1,1)^2$ onto each element K in the computational mesh $\mathcal{T}_h$ via the standard bilinear mapping $F_K : \hat{K} \mapsto K$. Thereby, discrete functions, restricted to a given element K, are defined as $u \circ F_K = \hat{u}$, where $\hat{u} \in \mathcal{Q}^\ell(\hat{K})$. We point out that meshes obtained with nonaffine mappings F_K are not rigorously covered by our analysis, and the underlying stability and approximation properties need to be further investigated; see [4] for related work on the approximation properties of bilinearly mapped quadrilateral elements. Finally, we note that throughout this section we select the constants appearing in the stabilization parameters defined in (3.1) as follows: $\alpha = 10\,\ell^2$, $\beta = 1$, and $\gamma = 1$. We remark that the dependence of α on the polynomial degree ℓ has been formally chosen in order to guarantee the coercivity property in Lemma 3.1 of the underlying DG form a_h independently of ℓ; cf. [19], for example.

7.1. Example 1. Here, we let Ω be the L-shaped domain $(-1,1)^2 \backslash [0,1) \times (-1,0]$; further, we choose $\boldsymbol{j}$ and g so that the analytical solution to the two-dimensional analogue of (2.1) with $\mu \equiv \varepsilon \equiv 1$ is given by

$$\begin{pmatrix} u_1 \\ u_2 \\ p \end{pmatrix} = \begin{pmatrix} -\exp(x)(y\cos(y) + \sin(y)) \\ \exp(x) y \sin(y) \\ \sin(\pi(x-1)/2)\sin(\pi(y-1)/2) \end{pmatrix}; \tag{7.1}$$

this is a variant of the model problem considered in [19]. We investigate the asymptotic behavior of the errors of the mixed DG method (2.5)–(2.6) on a sequence of successively finer square and quadrilateral meshes for different values of the polynomial degree ℓ. In each case we consider two types of quadrilateral meshes which are constructed from a uniform square mesh by (i) randomly perturbing each of the interior nodes by up to 10% of the local mesh size (cf. Figure 1(a)) and (ii) randomly *splitting* each of the interior nodes by a displacement of up to 10% of the local mesh size (cf. Figure 1(b)). The latter meshes are constructed so that all the nodes in the interior of Ω are irregular (i.e., hanging); cf. [20].

In Figure 2 we first present a comparison of the DG-norm $\|\cdot\|_{\boldsymbol{V}(h)}$ of the error in the approximation to $\boldsymbol{u}$ with the mesh function h for $1 \le \ell \le 4$. For consistency, $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$ is plotted against h_u for each mesh type, where h_u denotes the mesh size of the uniform square mesh; this ensures that a fair comparison between the error per degree of freedom for each mesh type can be made. Here, we observe that $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$ converges to zero, for each fixed ℓ, at the rate $\mathcal{O}(h^\ell)$ as the mesh is

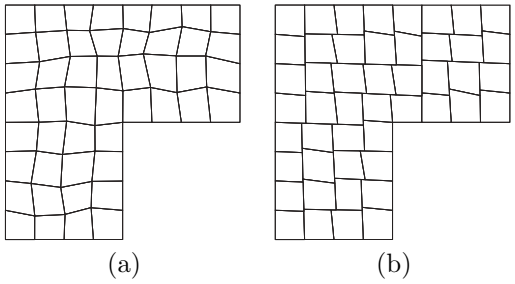


FIG. 1. *Example* 1. (a) *quadrilateral mesh* (i)*;* (b) *quadrilateral mesh* (ii).

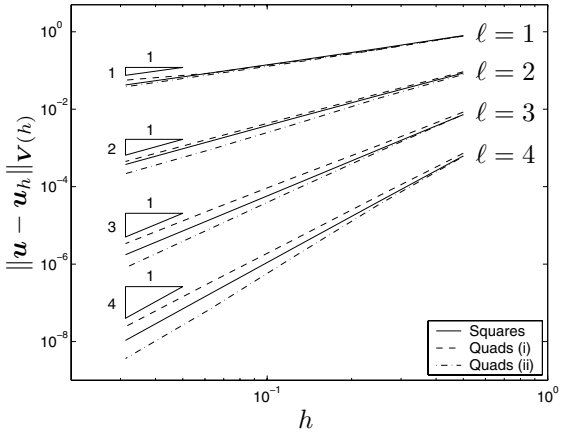


FIG. 2. *Example* 1. *Convergence of* $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$ *with h-refinement.*

refined, thereby confirming Theorem 3.3. In particular, we observe that while the error on the square mesh is smaller than on the randomly generated quadrilateral mesh (i), as we would expect, the error is consistently smaller when the irregular quadrilateral mesh is employed. As in [20], we attribute this improvement in $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$ to the increase in interelement communication on the meshes (ii); when no hanging nodes are present in the mesh, elements may communicate only with their four immediate neighbors. On the other hand, on irregular meshes elements may now communicate with all of their neighbors which share a common node; cf. [20].

Second, in Figure 3 we plot the DG-norm $\|\cdot\|_{Q(h)}$ of the error in approximating p by p_h as the mesh size tends to zero. As for the approximation to $\boldsymbol{u}$, we again observe that $\|p - p_h\|_{Q(h)}$ converges to zero, for each fixed ℓ, at the rate $\mathcal{O}(h^\ell)$ as the mesh is refined; cf. Theorem 3.3. However, in contrast to the approximation to $\boldsymbol{u}$, both the conforming and nonconforming quadrilateral meshes lead to a slight degradation in the size of the error in the approximation to p for each mesh and each polynomial degree employed, though in almost all cases the error in the numerical solution computed on the meshes (ii) was observed to be slightly smaller than the corresponding quantity computed on the meshes (i). Thereby, the increase in interelement communication arising when the meshes (ii) are employed no longer leads to the improvement in the size of the approximation error observed above for $\boldsymbol{u}$ as well as in [20].

The increase in the quality of the numerical approximation $\boldsymbol{u}_h$ to $\boldsymbol{u}$ when the nonconforming meshes (ii) are employed becomes even more apparent when the error

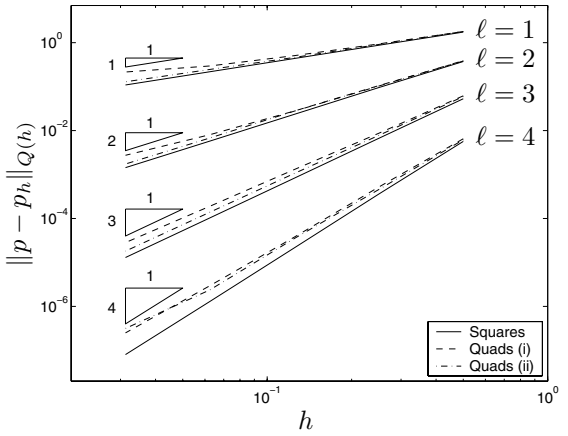


FIG. 3. *Example* 1. *Convergence of* $\|p - p_h\|_{Q(h)}$ *with h-refinement.*

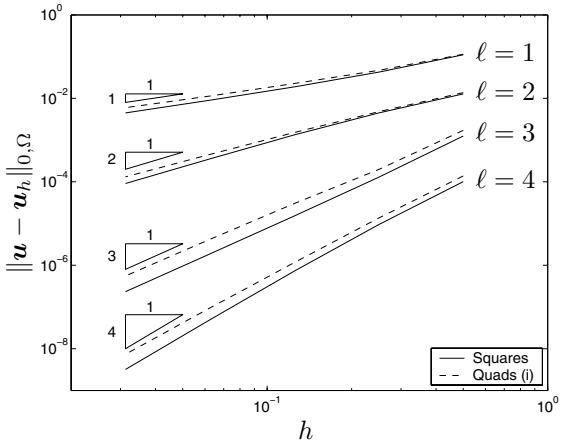


FIG. 4. *Example* 1. *Convergence of* $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}$ *with h-refinement.*

$\boldsymbol{u} - \boldsymbol{u}_h$ is measured in terms of the $L^2(\Omega)$-norm. To this end, in Figure 4, we first plot $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}$ against h for $1 \leq \ell \leq 4$ using the uniform square and randomly generated quadrilateral meshes (i). As predicted by Theorem 3.3, we observe that $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}$ converges to zero, for each fixed ℓ, at the rate $\mathcal{O}(h^\ell)$ as the mesh is refined. While this rate of convergence is one order less than we would expect when using discontinuous piecewise polynomials of degree at most ℓ in each coordinate direction, the numerical results clearly verify the sharpness of the a priori error analysis. However, in contrast, when the nonconforming quadrilateral meshes (ii) are employed, the order of convergence increases by a full power of h; thereby, in this case $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}$ now converges to zero, for each fixed ℓ, at the rate $\mathcal{O}(h^{\ell+1})$ as h tends to zero; cf. Figure 5. Analogous behavior is also observed when the $L^2(\Omega)$-norm of the error in the approximation to the divergence of $\boldsymbol{u}$ is computed. Indeed, from Figure 6, we observe that $\|h\,\nabla_h \cdot \boldsymbol{e}_h\|_{0,\Omega}$ converges to zero at the rate $\mathcal{O}(h^\ell)$ as h tends to zero when the uniform and randomly generated quadrilateral meshes (i) are employed, thereby confirming Theorem 3.6 and Remark 3.7. On the other hand, when the nonconforming quadrilateral meshes (ii) are employed, this rate of convergence increases to $\mathcal{O}(h^{\ell+1})$

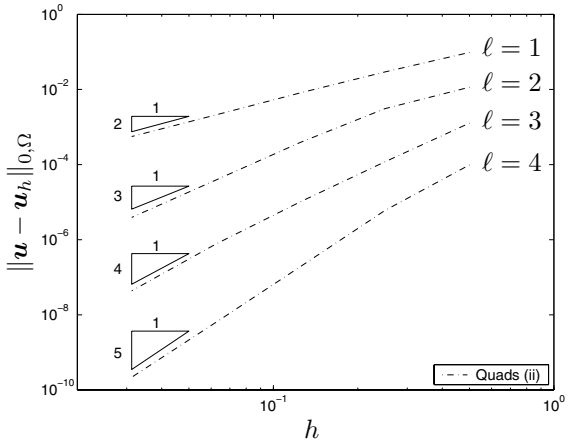


FIG. 5. *Example* 1. *Convergence of* $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}$ *with h-refinement.*

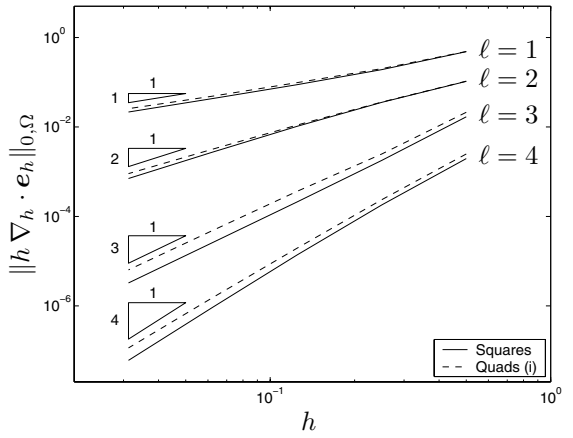


FIG. 6. *Example* 1. *Convergence of* $\|h\,\nabla_h \cdot \boldsymbol{e}_h\|_{0,\Omega}$ *with h-refinement.*

as h tends to zero; cf. Figure 7.

As a final remark, we note that on all the meshes employed, the $L^2(\Omega)$-norm of the error in the approximation to p converges to zero, for each fixed ℓ, at the (optimal) rate $\mathcal{O}(h^{\ell+1})$ as the mesh is refined. As for the DG-norm of $p - p_h$, both the conforming and nonconforming quadrilateral meshes lead to a slight degradation in the size of $\|p - p_h\|_{0,\Omega}$ for each mesh and each polynomial degree employed, though the error in the numerical solution computed on the meshes (ii) was observed to be slightly smaller than the corresponding quantity computed on the meshes (i); for brevity, these results have been omitted.

7.2. Example 2. In this second example, we investigate the performance of the mixed DG method (2.5)–(2.6) for a problem with a corner singularity in $\boldsymbol{u}$. To this end, we again let Ω be the same L-shaped domain as in the first example; here, we set $\boldsymbol{j} = \boldsymbol{0}$, and g is chosen so that the analytical solution $\boldsymbol{u}$ to the two-dimensional analogue of (2.1) with $\mu \equiv \varepsilon \equiv 1$ is given, in terms of the polar coordinates (r, ϑ), by $\boldsymbol{u}(x, y) = \nabla S(r, \vartheta)$, where $S(r, \vartheta) = r^{2/3} \sin(2\vartheta/3)$; thereby, $p \equiv 0$. The analytical solution $\boldsymbol{u}$ contains a singularity at the corner located at the origin of Ω; here, we

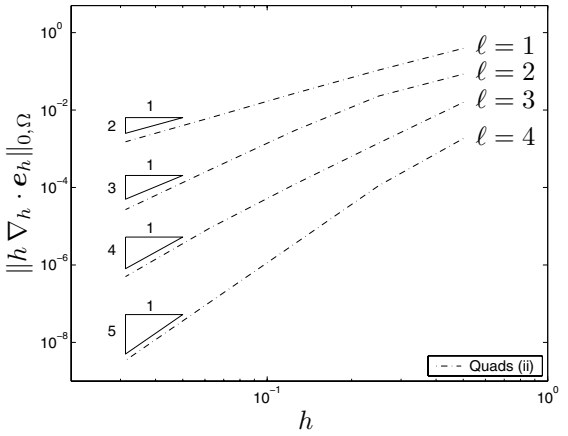


FIG. 7. *Example* 1. *Convergence of* $\|h\,\nabla_h \cdot \boldsymbol{e}_h\|_{0,\Omega}$ *with h-refinement.*

TABLE 1
Example 2. *Convergence of* $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$ *on uniform square meshes with h-refinement.*

	$\ell = 1$		$\ell = 2$		$\ell = 3$	
Elements	$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$	k	$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$	k	$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$	k
12	5.987e-1	-	5.350e-1	-	4.853e-1	-
48	4.300e-1	0.48	3.427e-1	0.64	3.076e-1	0.66
192	2.815e-1	0.61	2.144e-1	0.68	1.929e-1	0.67
768	1.816e-1	0.63	1.344e-1	0.67	1.211e-1	0.67
3072	1.170e-1	0.63	8.452e-2	0.67	7.622e-2	0.67

TABLE 2
Example 2. *Convergence of* $\|p - p_h\|_{Q(h)}$ *on uniform square meshes with h-refinement.*

	$\ell = 1$		$\ell = 2$		$\ell = 3$	
Elements	$\|p - p_h\|_{Q(h)}$	k	$\|p - p_h\|_{Q(h)}$	k	$\|p - p_h\|_{Q(h)}$	k
12	8.742e-1	-	1.341	-	1.598	-
48	8.147e-1	0.10	1.039	0.37	1.178	0.44
192	6.235e-1	0.39	7.159e-1	0.54	7.948e-1	0.57
768	4.253e-1	0.55	4.678e-1	0.61	5.154e-1	0.63
3072	2.763e-1	0.62	2.990e-1	0.65	3.285e-1	0.65

have only $\boldsymbol{u} \in H^{2/3-\varepsilon}(\Omega)^2$, $\varepsilon > 0$.

In this example, let us first confine ourselves to uniform square meshes; we shall return to the more general meshes considered in the previous example later. To this end, in Tables 1 and 2 we present a comparison of the DG-norms of the error in the approximation to both $\boldsymbol{u}$ and p, respectively, with the mesh function h on a sequence of uniform square meshes for $1 \leq \ell \leq 3$. In each case we show the number of elements in the computational mesh, the corresponding DG-norm of the error, and the computed rate of convergence k. Here, we observe that (asymptotically) both $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$ and $\|p - p_h\|_{Q(h)}$ converge to zero at the optimal rate $\mathcal{O}(h^{\min(2/3-\varepsilon,\ell)})$ as h tends to zero, predicted by Theorem 3.4.

Next, in Table 3 we present a comparison of the $L^2(\Omega)$-norm of the error in the numerical approximation to $\boldsymbol{u}$ with h. On the basis of Theorem 3.4, we expect that $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}$ should tend to zero at the rate $\mathcal{O}(h^{\min(2/3-\varepsilon,\ell)})$ as h tends to zero. However, from Table 3, we observe that for $\ell = 1, 2, 3$, the rate of convergence of the

TABLE 3
Example 2. *Convergence of* $\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}$ *on uniform square meshes with* h*-refinement.*

	$\ell=1$		$\ell=2$		$\ell=3$	
Elements	$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}$	k	$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}$	k	$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}$	k
12	2.788e-1	-	2.420e-1	-	2.004e-1	-
48	1.817e-1	0.62	1.252e-1	0.95	9.703e-2	1.05
192	1.057e-1	0.78	6.160e-2	1.02	4.513e-2	1.10
768	6.078e-2	0.80	3.151e-2	0.97	2.187e-2	1.05
3072	3.620e-2	0.75	1.740e-2	0.86	1.154e-2	0.92

TABLE 4
Example 2. *Convergence of* $\|h\nabla_h\cdot\boldsymbol{e}_h\|_{0,\Omega}$ *on uniform square meshes with* h*-refinement.*

	$\ell=1$		$\ell=2$		$\ell=3$	
Elements	$\|h\nabla_h\cdot\boldsymbol{e}_h\|_{0,\Omega}$	k	$\|h\nabla_h\cdot\boldsymbol{e}_h\|_{0,\Omega}$	k	$\|h\nabla_h\cdot\boldsymbol{e}_h\|_{0,\Omega}$	k
12	1.348e-1	-	2.086e-1	-	3.022e-1	-
48	6.438e-2	1.07	1.752e-1	0.25	2.380e-1	0.34
192	3.668e-2	0.81	1.296e-1	0.43	1.678e-1	0.50
768	2.274e-2	0.69	8.844e-2	0.55	1.115e-1	0.59
3072	1.324e-2	0.78	5.800e-2	0.61	7.206e-2	0.63

TABLE 5
Example 2. *Convergence of* $\|p-p_h\|_{0,\Omega}$ *on uniform square meshes with* h*-refinement.*

	$\ell=1$		$\ell=2$		$\ell=3$	
Elements	$\|p-p_h\|_{0,\Omega}$	k	$\|p-p_h\|_{0,\Omega}$	k	$\|p-p_h\|_{0,\Omega}$	k
12	1.906e-1	-	1.627e-1	-	1.361e-1	-
48	1.135e-1	0.75	7.926e-2	1.04	6.327e-2	1.11
192	5.602e-2	1.02	3.452e-2	1.20	2.696e-2	1.23
768	2.487e-2	1.17	1.426e-2	1.28	1.103e-2	1.29
3072	1.049e-2	1.24	5.759e-3	1.31	4.435e-3	1.31

$L^2(\Omega)$-norm of the error in the approximation to $\boldsymbol{u}$ is slightly higher than predicted, although, asymptotically, we expect these convergence rates to slowly tend to the optimal one. Additionally, in Table 4 we show the convergence of $\|h\nabla_h\cdot\boldsymbol{e}_h\|_{0,\Omega}$ with respect to h; here, we again observe that, asymptotically, the rate of convergence tends to the one predicted in Theorem 3.6; cf. Remark 3.7. Finally, in Table 5 we show $\|p-p_h\|_{0,\Omega}$ for $\ell=1,2,3$, based on employing uniform square meshes. In comparison with Table 3, Table 5 indicates that the rate of convergence of $\|p-p_h\|_{0,\Omega}$ is almost twice the optimal rate of $\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}$; indeed, asymptotically, we observe that k is tending towards $4/3$ as the mesh is uniformly refined.

We remark that analogous convergence rates to those reported in Tables 1–5 are also observed when the numerical approximation is computed on the conforming quadrilateral meshes (i); for brevity, these results are omitted. However, convergence of the mixed DG method was *not* observed on the quadrilateral meshes (ii). We recall that the convergence proof presented in the case of weak smoothness assumptions for the component $\boldsymbol{u}$ of the analytical solution (cf. Theorem 3.4) precludes the presence of hanging nodes in the mesh $\mathcal{T}_h$ and, as in the case of smooth analytical solutions, assumes that the elements in $\mathcal{T}_h$ are affine. As noted above, optimal rates of convergence are still observed computationally when the quadrilateral meshes (i), which are conforming, but nonaffine, are employed. In order to test the method in the case when $\mathcal{T}_h$ contains hanging nodes, but the elements are affine, we consider the performance of the mixed DG method (2.5)–(2.6) on a sequence of adaptively refined square

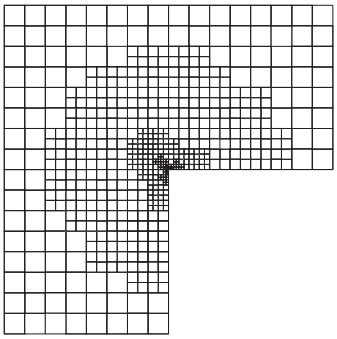


FIG. 8. *Example* 2. *Computational mesh after* 9 *adaptive refinement steps, with* 721 *nodes and* 618 *elements and* $\ell = 1$.

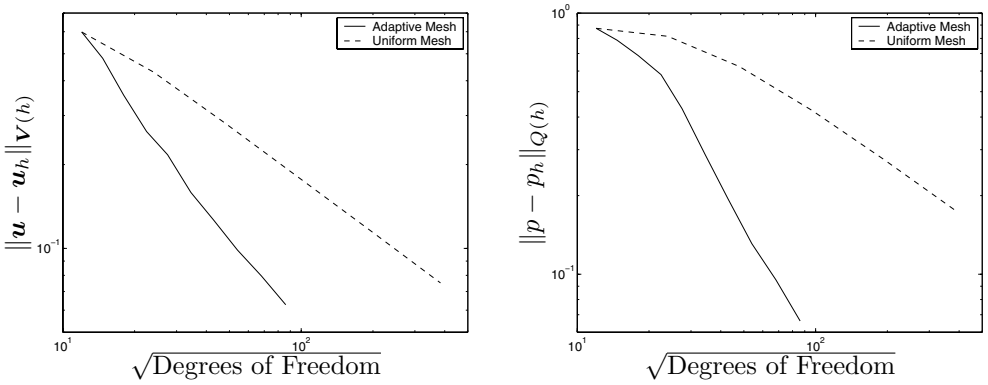


FIG. 9. *Example* 2. *Comparison of adaptive and uniform* h*-refinement.*

meshes. Here, the adaptive meshes are constructed by employing the fixed fraction strategy (with refinement and derefinement fractions set to 25% and 0%, respectively) with a simple error indicator η_K based on the gradient of the numerical approximation. More precisely, given K in $\mathcal{T}_h$, we set $\eta_K = (\|h_K \nabla \boldsymbol{u}_h\|^2_{0,K} + \|h_K \nabla p_h\|^2_{0,K})^{1/2}$. More sophisticated error indicators may be appropriate for this problem; here, we are simply interested in generating a sequence of adaptive meshes (containing hanging nodes) in which to test the hypotheses of our a priori error analysis. A typical mesh generated with this adaptive algorithm is shown in Figure 8.

In Figure 9, we present a comparison of $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{V}(h)}$ and $\|p - p_h\|_{Q(h)}$ with the (square root of the) number of degrees of freedom in $\boldsymbol{V}_h \times Q_h$ for $\ell = 1$ on the sequence of adaptively refined meshes generated as above, as well as on a sequence of uniform square meshes; cf. Tables 1 and 2 above, respectively. Here, we clearly observe that the error in the mixed DG discretization of (2.1) converges to zero as the finite element space $\boldsymbol{V}_h \times Q_h$ is enriched, even when the mesh contains hanging nodes; indeed, we see that the adaptively refined meshes lead to a general improvement in the error when compared to the uniform square meshes. In summary, in the case of weak Sobolev regularity assumptions on $\boldsymbol{u}$, the mixed DG method (2.5)–(2.6) is observed (numerically) to be optimally convergent on both conforming and nonconforming affine square meshes as well as on conforming nonaffine meshes; however, convergence is not observed when the mesh contains hanging nodes *and* the elements are not affine.

8. Conclusions. In this paper, we have presented a new mixed discontinuous Galerkin method for the discretization of the time-harmonic Maxwell operator. This method is based on equal-order finite element spaces, where all the unknowns are approximated with piecewise discontinuous polynomials of the same degree. When compared to the numerical scheme proposed in [25], the amount of numerical stabilization here is drastically reduced. Our error analysis and numerical results show that the method is optimally convergent in the energy norms for smooth as well as for singular solutions. In the latter case, the theoretical analysis is restricted to regular meshes without hanging nodes. However, numerical experiments for a problem with a strongly singular solution have demonstrated that the application of the method within an adaptive procedure on affine quadrilateral meshes, where hanging nodes are introduced during the course of the refinement, still leads to a convergent numerical approximation as the finite element space is enriched. A more delicate issue seems to be the one related to nonaffine meshes. Indeed, our tests seem to indicate that the assumption of affineness on the meshes cannot be eliminated when meshes containing hanging nodes are employed. Future work will be devoted to the study of variants of the proposed method which deliver optimal rates of convergence when the error in the approximation to the vector field $\boldsymbol{u}$ is measured in the $L^2(\Omega)$-norm.

Acknowledgments. The numerical experiments in this paper were performed using the University of Leicester Mathematical Modelling Centre's supercomputer, which was purchased through the EPSRC Strategic Equipment Initiative.

REFERENCES

[1] A. Alonso and A. Valli, *A domain decomposition approach for heterogeneous time-harmonic Maxwell equations*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 97–112.

[2] A. Alonso and A. Valli, *An optimal domain decomposition preconditioner for low-frequency time-harmonic Maxwell equations*, Math. Comp., 68 (1999), pp. 607–631.

[3] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional non-smooth domains*, Math. Models Appl. Sci., 21 (1998), pp. 823–864.

[4] D. Arnold, D. Boffi, and R. Falk, *Quadrilateral H(div) Elements*, Tech. report 1283-02, IAN-CNR Pavia, Pavia, Italy, 2002.

[5] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[6] A.-S. Bonnet-Ben Dhia, C. Hazard, and S. Lohrengel, *A singular field method for the solution of Maxwell's equations in polyhedral domains*, SIAM J. Appl. Math., 59 (1999), pp. 2028–2044.

[7] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer–Verlag, New York, 1991.

[8] F. Brezzi and M. Fortin, *A minimal stabilisation procedure for mixed finite element methods*, Numer. Math., 89 (2001), pp. 457–491.

[9] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37 (2000), pp. 1542–1570.

[10] B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems*, in High-Order Methods for Computational Physics, Lect. Notes Comput. Sci. Eng. 9, T. Barth and H. Deconinck, eds., Springer–Verlag, Berlin, 1999, pp. 69–224.

[11] B. Cockburn, G. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., Springer–Verlag, Berlin, 2000, pp. 3–50.

[12] B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin Methods. Theory, Computation and Applications*, Lect. Notes Comput. Sci. Eng. 11, Springer–Verlag, Berlin, 2000.

[13] B. Cockburn and C.-W. Shu, *Runge–Kutta discontinuous Galerkin methods for convection–dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.
[14] M. Costabel and M. Dauge, *Weighted regularization of Maxwell equations in polyhedral domains*, Numer. Math., 93 (2002), pp. 239–277.
[15] L. Demkowicz and L. Vardapetyan, *Modeling of electromagnetic absorption/scattering problems using hp–adaptive finite elements*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 103–124.
[16] P. Fernandes and G. Gilardi, *Magnetostatic and electrostatic problems in inhomogeneous anisotropic media with irregular boundary and mixed boundary conditions*, Math. Models Methods Appl. Sci., 7 (1997), pp. 957–991.
[17] J. Hesthaven and T. Warburton, *High-order nodal methods on unstructured grids, Part* I. *Time-domain solution of Maxwell's equations*, J. Comput. Phys., 181 (2002), pp. 1–34.
[18] P. Houston, I. Perugia, and D. Schötzau, *Mixed Discontinuous Galerkin Approximation of the Maxwell Operator*, Tech. report 02-16, University of Basel, Department of Mathematics, Basel, Switzerland, 2002.
[19] P. Houston, I. Perugia, and D. Schötzau, *hp-DGFEM for Maxwell's equations*, in Numerical Mathematics and Advanced Applications: ENUMATH 2001, F. Brezzi, A. Buffa, S. Corsaro, and A. Murli, eds., Springer–Verlag, Berlin, 2003, pp. 785–794.
[20] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.
[21] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.
[22] D. Kopriva, S. Woodruff, and M. Hussaini, *Discontinuous spectral element approximation of Maxwell's equations*, in Discontinuous Galerkin Methods. Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., Springer–Verlag, Berlin, 2000, pp. 355–361.
[23] J. Nédélec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 50 (1986), pp. 57–81.
[24] I. Perugia and D. Schötzau, *The hp-local discontinuous Galerkin method for low-frequency time-harmonic Maxwell equations*, Math. Comp., 72 (2003), pp. 1179–1214.
[25] I. Perugia, D. Schötzau, and P. Monk, *Stabilized interior penalty methods for the time-harmonic Maxwell equations*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4675–4697.
[26] L. Vardapetyan and L. Demkowicz, *hp-adaptive finite elements in electromagnetics*, Comput. Methods Appl. Mech. Engrg., 169 (1999), pp. 331–344.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 461–483

FIRST-ORDER SYSTEM LEAST SQUARES AND ELECTRICAL IMPEDANCE TOMOGRAPHY*

H. R. MACMILLAN†, T. A. MANTEUFFEL‡, AND S. F. MCCORMICK†

Abstract. Electrical impedance tomography (EIT) belongs to a family of imaging methods that employ boundary measurements to distinguish interior spatial variation of an electromagnetic parameter. The associated inverse problem is notoriously ill-posed, due to diffusive effects in the quasi-static regime, when electrical impedance reduces to its real part, resistivity. The standard approach to EIT is output least squares (OLS). For a set of applied normal boundary currents, one minimizes the defect between the measured and computed boundary voltages associated, respectively, with the exact impedance and its approximation. In minimizing a boundary functional, OLS implicitly imposes the governing Poisson equation as an optimization constraint. To reconstruct resistivity or, equivalently, conductivity, we introduce a new first-order system least-squares (FOSLS) formulation that incorporates the elliptic PDE as an interior functional in a global unconstrained minimization scheme. We then establish equivalence of our functional to OLS and to an existing interior least-squares functional due to Kohn and Vogelius [*Comm. Pure Appl. Math*, 37 (1984), pp. 289–298]. That the latter may be viewed as a special dual approach (FOSLL*) is an interesting attribute of this equivalence. Because the FOSLS functional implicitly reflects the inherent loss of resolution away from the boundary, relative to the set of applied boundary tests, the need for artificial regularization may be avoided.

Key words. first-order system least squares, electrical impedance tomography

AMS subject classifications. 15A29, 65M32, 65M60

DOI. 10.1137/S0036142902412245

1. Introduction.

Electrical impedance tomography (EIT) belongs to a family of imaging methods based on distinguishing interior spatial variation of an electromagnetic parameter. The basic principle is to approximate this parameter in a manner consistent with given pairings of boundary current and boundary voltage data. In mathematical parlance, given boundary data represents discrete knowledge of the Dirichlet-to-Neumann map associated with the unknown parameter's effect on diffusion. The broad scope of EIT's application, ranging from biomedical imaging [8, 10, 4] to geophysical prospecting [34, 28, 37, 12] to industrial process tomography [31, 14, 15], implies the importance of different parameter regimes and boundary data protocols. Thus, it is reasonable to expect efficient and effective algorithms to be highly problem specific, according to prior information and measurement techniques. In this paper, we discuss the governing equations, the standard least-squares approach, and established theoretical insights, but we do not conduct a comprehensive study of a myriad of practical concerns. Our focus is, instead, on proving equivalence of three least-squares formulations: output least squares [8], a formulation due to Kohn and Vogelius [23], and a new first-order system least-squares (FOSLS) formulation. That the Kohn and Vogelius functional may be viewed as a first-order system LL* (FOSLL* [6]) formulation, combined with the equivalence results, suggests a unifying framework for future

*Received by the editors July 24, 2002; accepted for publication (in revised form) September 8, 2003; published electronically March 11, 2004.
http://www.siam.org/journals/sinum/42-2/41224.html
†School of Computational Science and Information Technology, Florida State University, 467 DSL, Tallahassee, FL 32306-4120 (macmilla@csit.fsu.edu). This author's research was supported by NSF VIGRE grant DMS-9810751.
‡Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CA 80309-0526 (tmanteuf@colorado.edu, stevem@colorado.edu).

studies of EIT. It is our conviction that the ultimate goal of such studies is to quantify the sense in which the notoriously ill-posed problem can be posed well. This is in line with not only classical Backus and Gilbert theory on geophysical inverse problems [1] but also with recent analysis of probabilistic approaches [26]. The equivalence results of this paper, by establishing common ground, set the stage for investigations of the uncertainty inherent in posing least-squares formulations of EIT.

1.1. Mathematical setting. The development of a mathematical model for the EIT inverse problem begins with Maxwell's equations in a source-free, closed, and simply connected domain, which we denote Ω with boundary Γ. A series of assumptions (see [8, 30] and the references therein), including negligible magnetic permeability and quasi-static periodic time variation, leads to a Poisson equation for complex-valued voltage p:

$$\nabla \cdot \gamma \nabla p = 0, \tag{1.1}$$

where $\gamma = \sigma + i\omega\epsilon$ is the admittance. EIT refers to the reconstruction of this parameter, since electrical impedance is defined as $1/\gamma$. Here we make the electrostatic assumption, $\omega = 0$. Thus, the quantities in (1.1) are real-valued, and the aim is reconstruction of the conductivity, σ, based on its associated Dirichlet and Neumann boundary data. This is equivalent to reconstructing resistivity, $1/\sigma$. The electrostatic assumption is an important first principle since the lack of resolution with quasi-static fields, and thus EIT, is attributed to the conductivity term. Extension to electromagnetic fields varying slowly in time remains a future work.

Remark 1. We refer to the process of reconstructing conductivity given Dirichlet and Neumann data as electrical conductivity tomography (ECT) and remark that in the literature it is also known as the DC resistivity problem, applied potential tomography, electric current computed tomography, and electrical resistance tomography.

In what follows, we adopt the usual Sobolev spaces, $\mathrm{L}^2(\Omega)$ and $\mathrm{H}^1(\Omega)$, including the dual boundary spaces, $\mathrm{H}^{1/2}(\Gamma)$ and $\mathrm{H}^{-1/2}(\Gamma)$ [11]. We assume that the boundary, Γ, is Lipschitz and note that the normal current flux is the scaled Neumann data, $\mathbf{n} \cdot \sigma \nabla p$, where $\mathbf{n}$ denotes the outward unit normal. Also, the boundary voltage distribution is the Dirichlet data, $p_{|\Gamma}$. It is practical to impose the normal current and measure the resulting voltage data, and we choose this perspective in the presentation of our theory. Hence, considering the Neumann problem

$$\begin{aligned} \nabla \cdot \sigma \nabla p &= 0 \qquad \text{in } \Omega, \\ \mathbf{n} \cdot \sigma \nabla p &= h \qquad \text{on } \Gamma, \end{aligned} \tag{1.2}$$

where the normal current $h \in \mathrm{H}^{-1/2}(\Gamma)$ satisfies $\int_\Gamma h \, ds = 0$, one may view coefficient σ as determining a Neumann-to-Dirichlet map, i.e., as determining the trace of voltage $p_{|\Gamma} \in \mathrm{H}^{1/2}(\Gamma)$ on the boundary (up to a constant). An alternative and likely more realistic mathematical model uses Robin boundary conditions in place of the Dirichlet conditions as the measured data [8, 40]. However, since the Neumann data is in each case given, such measurements are equivalent to discrete knowledge of the Neumann-to-Dirichlet map. At this stage we assume the most general Sobolev setting and defer addressing practical issues regarding boundary data to a future work [25].

Several definitions associated with system (1.2) now follow to facilitate formally stating the inverse problem of ECT.

DEFINITION 1.1. *Denote* $\mathrm{L}^\infty_+(\Omega) := \{\sigma \in \mathrm{L}^\infty(\Omega) : \sigma > 0\}$. *Also, let*

$$\mathrm{H}^{1/2}_0(\Gamma) := \left\{ g \in \mathrm{H}^{1/2}(\Gamma) : \int_\Gamma g \, d\tau = 0 \right\},$$
$$\mathrm{H}^{-1/2}_0(\Gamma) := \left\{ h \in \mathrm{H}^{-1/2}(\Gamma) : \int_\Gamma h \, d\tau = 0 \right\}.$$

A given conductivity, $\sigma \in \mathrm{L}^\infty_+(\Omega)$, *induces a bounded linear map,*

$$\mathcal{R}_\sigma : \mathrm{H}^{-1/2}_0(\Gamma) \longrightarrow \mathrm{H}^{1/2}_0(\Gamma),$$

from the imposed normal current to the resulting boundary voltage. This map is defined by

$$\mathcal{R}_\sigma(h) := p_{|\Gamma},$$

where p *solves* (1.2), *and we call* $\mathcal{R}_\sigma$ *the Neumann-to-Dirichlet map (NtD map) associated with* σ.

Remark 2. Function space $\mathrm{L}^\infty_+(\Omega)$ is used for admissible conductivity since it is the least restrictive for which the NtD map associated with (1.2) can be defined. However, uniquely defining the NtD map requires further restriction [17]. Regarding the boundary spaces, a compatibility condition on (1.2) implies that we must have $\int_\Gamma h \, d\tau = 0$. Similarly, p is determined by (1.2) only up to a constant, so we are permitted to impose the condition $\int_\Gamma g \, d\tau = 0$.

DEFINITION 1.2. *We say that* $\{p, \mathcal{R}_\sigma(h), h\} \in \mathrm{H}^1(\Omega) \times \mathrm{H}^{1/2}_0(\Gamma) \times \mathrm{H}^{-1/2}_0(\Gamma)$ *is a Dirichlet–Neumann triple (DN triple) for* σ *if* (1.2) *holds. Also, we shall refer to* $\{\mathcal{R}_\sigma(h), h\}$ *as a DN pair.*

Remark 3. The NtD map, $\mathcal{R}_\sigma$, is self-adjoint in the L^2-inner-product. This is apparent since for DN triples $\{p_1, \mathcal{R}_\sigma(h_1), h_1\}$ and $\{p_2, \mathcal{R}_\sigma(h_2), h_2\}$, integration by parts leads to

$$\langle h_1, \mathcal{R}_\sigma(h_2) \rangle_{0,\Gamma} = \int_\Omega \sigma \nabla p_1 \cdot \nabla p_2 \, dx = \langle \mathcal{R}_\sigma(h_1), h_2 \rangle_{0,\Gamma}.$$

Finally, we have the notation to define the inverse problem of ECT.

DEFINITION 1.3. *The inverse problem of ECT is to approximate the exact conductivity,* σ^*, *according to a given graph of measured DN data* $\{g_i, h_i\}_{i=1}^L$, *where* $g_i \approx \mathcal{R}_{\sigma^*}(h_i)$.

Stated otherwise, the inverse problem of ECT is to approximately invert the nonlinear mapping $\sigma \longrightarrow \mathcal{R}_\sigma$. The approximate nature of this inversion is due to the preimposed discreteness of having finite data, the inexactness of this data, and the need to choose a discrete approximation space for conductivity. Next, we give background on these issues as they affect our ability to solve the inverse problem.

1.2. Theoretical background. Classical discussion of existence and uniqueness of the inverse problem is hindered by several concerns. First, assuming exact boundary data is the same as assuming the existence of a conductivity consistent with the boundary data. However, given the intrinsic error in the measurements of the data, it is likely that there is no such consistent conductivity. As for uniqueness, progressively refined results have been established, with a focus on achieving a uniqueness result for piecewise continuous conductivity [21, 22, 35, 27, 17]. The techniques used in theoretical uniqueness results, however, rely on complete and precise

boundary information and cannot accommodate the reality of only a finite number of approximate DN pairs. As such, they do not provide insight into how inexact and incomplete boundary data unleash error in σ that cannot be confidently eliminated. Toward such an understanding, it is instructive to consider Calderon's initial analysis [7]. Let $\delta \in \mathrm{L}^1(\Omega)$ be a perturbation of a background conductivity, σ. For DN triples $\{p_i, \mathcal{R}_\sigma(h_i), h_i\}$ and $\{p_j, \mathcal{R}_\sigma(h_j), h_j\}$, he observes that

$$\langle h_i, \mathcal{R}_{\sigma+\delta}(h_j) - \mathcal{R}_\sigma(h_j)\rangle_{0,\Gamma} = \int_\Omega \delta\, \nabla p_i \cdot \nabla p_j \, dx \; + \; O(\delta^2). \tag{1.3}$$

This classical expansion displays how perturbations in the interior can be practically undetectable at the boundary. For example, observe that, with σ constant, the distribution of voltage, p, is increasingly smooth away from the boundary. Thus, the product of the gradients of solutions to the Laplace equation places decreasing weight on perturbations of conductivity according to distance into the interior. Notably, this weighting appears within the FOSLS formulation when updating the approximate conductivity.

In a now classical work, Backus and Gilbert [1] introduce the notion of "inference" for studying a wide class of geophysical inverse problems, not just ECT. Parker [29] has applied this theory to ECT, considering the linearized one-dimensional problem of determining the conductive layers of stratified earth. Independently, Seagar, Yeo, and Bates [32, 33] assess the "visibility" of circular disks of constant conductivity as perturbations of a background constant conductivity in a two-dimensional circular domain by using conformal mapping techniques. Finally, to determine whether the effect of a certain error can be detected on the boundary, Isaacson [16] introduces "distinguishability." As is observed in each analysis of resolution, the best test to apply in order to detect a specific error, $\sigma - \sigma^*$, is that which attains the operator norm of the defect in the NtD maps, i.e., that which attains the sup in

$$\|\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*}\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{1/2}(\Gamma))} := \sup_{h\in\mathrm{H}_0^{-1/2}(\Gamma)} \frac{\|(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})h\|_{1/2,\Gamma}}{\|h\|_{-1/2,\Gamma}}. \tag{1.4}$$

Thus, given complete and exact boundary data, the natural functional to minimize is the operator norm in (1.4). Of course, with finite data, the sup can be taken only over a finite-dimensional subspace, $span\{h_i\}_{i=1}^L \subset \mathrm{H}_0^{-1/2}(\Gamma)$, where L is the number of available DN data. Despite the preimposed discreteness of this data and its inevitable error, we hope to still recover meaningful components of σ^*.

1.3. Output least squares. In analogy to the Frobenius norm of matrix algebra, a natural alternative to minimizing the sup over the span of the boundary data is to minimize the output least-squares (OLS) functional,

$$\mathcal{F}(\sigma; \{g_i, h_i\}_{i=1}^L) := \sum_{i=1}^{L} \|\mathcal{R}_\sigma(h_i) - g_i\|_{1/2,\Gamma}^2. \tag{1.5}$$

This formulation is an example of unconstrained least-squares optimization and is the standard approach throughout the wide range of EIT applications [3, 8, 12, 40]. It should be noted that the L^2-norm, rather than $\mathrm{H}^{1/2}$-norm, is typically enforced. However, in this work we preserve the most general Sobolev space setting. This top-down approach should prove more beneficial, as future work in computational electromagnetics leads to a better understanding of the fields impinging a given domain, thereby leading to a more accurate representation of the boundary data.

To minimize $\mathcal{F}$ in (1.5), the first variation of the NtD map with respect to σ enables a standard Newton quadraticization. The first-order term in (1.3) is used to represent this variation, determining the effect on the boundary of changing a given degree of freedom for σ. Once the linear system, involving the "sensitivity matrix," is solved to obtain a new approximation, the L forward problems are solved, and the process is repeated (see [8] and the references therein). The physics at the source of EIT's difficulty is manifested by the extremely high condition number of the sensitivity matrix when a uniform grid for σ is chosen. This is consistent with what the physics suggests: an increasingly coarse representation of conductivity should be used to extend the grid into the interior. Typically, some form of regularization is instead performed to ensure that the Hessian is sufficiently positive definite.

Regularization is, in essence, the process of posing an inverse problem well [36], and it has taken several forms in application to EIT. A regularization operator, $\mathcal{B}$, can be incorporated into the OLS functional as an additional least-squares term so that (1.5) is replaced by

$$\mathcal{F}(\sigma; \{g_i, h_i\}_{i=1}^{L}) := \left(\sum_{i=1}^{L} \|\mathcal{R}_\sigma(h_i) - g_i\|_{0,\Gamma}^2 \right) + \alpha \|\mathcal{B}\sigma\|_{0,\Omega}^2, \tag{1.6}$$

where α is a regularization parameter that is usually chosen empirically to balance resolution and computability in some sense. The most common regularization is a penalty on the H^1-seminorm of σ [8], wherein $\mathcal{B} = \nabla$. When piecewise constant basis functions are used in the discrete representation of σ, one can impose a penalty on the jumps between each element [9]. Of course, the resulting reconstruction is artificially smoothed, especially near the boundary where one ought to be able to recover high frequency components. However, the technique has been actively researched for many years, and ways of reducing the artificial smoothness have been developed, including the recent multilevel approach of Borcea [3]. The theory we develop suggests that the FOSLS formulation inherently imposes a smoothing of the conductivity estimate that is relative to both the distance from the boundary and prior information about σ. Indeed, the FOSLS formulation may likewise benefit from an empirical process of determining ideal weightings on each term in the functional, analogous to determining α in (1.6). These present potential directions for future research.

An alternative regularization scheme, referred to as a statistical or Bayesian approach, is implemented by Vauhkonen et al. [39] and Kaipio et al. [18, 19] to penalize the orthogonal complement of an operator which represents prior information. The penalty thus serves to steer the approximation toward an expected solution. A related method, that of basis constraints [38], builds the subspace of admissible conductivity, which may in practice be of very low dimension, by using basis functions selected according to prior information. Though we encourage the use of basis constraints to incorporate what is already known into the solution process, the prevailing conviction is that such information should not serve to "regularize" the inverse problem. Rather, one must accept and account for the inherent lack of resolution [1, 26, 13]. We feel that FOSLS offers a natural foundation for projecting this conviction.

2. FOSLS framework. We now prove equivalence of the OLS formulation to two interior least-squares functionals. To represent Dirichlet data, we write $\mathcal{R}_{\sigma^*}(h)$ instead of g to stress the present assumption on zero error in the data, as well as the dependence on σ^*. Finally, we develop the framework explicitly for two dimensions and anticipate that the setting generalizes to three dimensions, although we

draw attention to those points in the development for which the three-dimensional generalization is not obvious.

2.1. Kohn and Vogelius. The first interior least-squares formulation we address was introduced by Kohn and Vogelius [23] and Kohn and McKenney [20] and is closely related to the original interior functional proposed for ECT by Wexler, Fry, and Neumann [41]. We now review their formulation in two dimensions before proving its equivalence to OLS.

2.1.1. Formulation. To begin, we recall the interior partial differential equation (PDE) and boundary conditions associated with the exact, albeit unknown, conductivity, σ^*:

$$\begin{aligned} \nabla \cdot \sigma^* \nabla p &= 0 && \text{in } \Omega, \\ \mathbf{n} \cdot \sigma^* \nabla p &= h && \text{on } \Gamma, \\ p &= \mathcal{R}_{\sigma^*}(h) && \text{on } \Gamma. \end{aligned}$$

Letting $\nabla^\perp = (-\partial_y, \partial_x)^t$, and recalling that Ω is simply connected, then the divergence-free condition on $\sigma^* \nabla p$ asserted in this PDE implies the existence of $s \in \mathrm{H}^1(\Omega)$, depending on σ^*, h, and Ω, such that

$$\sigma^* \nabla p = \nabla^\perp s.$$

On Γ, we have

$$\mathbf{n} \cdot \sigma^* \nabla p = \mathbf{n} \cdot \nabla^\perp s = \boldsymbol{\tau} \cdot \nabla s = h,$$

where $\boldsymbol{\tau}$ denotes the clockwise unit tangent vector. Note that s satisfies

$$\begin{aligned} \nabla \cdot \frac{1}{\sigma^*} \nabla s = \nabla \times \frac{1}{\sigma^*} \nabla^\perp s &= 0 && \text{in } \Omega, \\ s &= \int_\Gamma h \, d\gamma && \text{on } \Gamma. \end{aligned}$$

To assess the validity of an approximate conductivity, the Kohn and Vogelius (KV) functional is based on defining two separate sets of Dirichlet problems: given $\{\mathcal{R}_{\sigma^*}(h_i), h_i\}_{i=1}^L$, find $p_i \in \mathrm{H}^1(\Omega)$ such that

$$\begin{aligned} \nabla \cdot \sigma \nabla p_i &= 0 && \text{in } \Omega, \\ p_i &= \mathcal{R}_{\sigma^*}(h_i) && \text{on } \Gamma, \end{aligned} \tag{2.1}$$

and find $s_i \in \mathrm{H}^1(\Omega)$ such that

$$\begin{aligned} \nabla \times \frac{1}{\sigma} \nabla^\perp s_i &= 0 && \text{in } \Omega, \\ s_i &= \int_\Gamma h_i \, d\gamma && \text{on } \Gamma. \end{aligned} \tag{2.2}$$

Now, if $\sigma = \sigma^*$, then

$$\| \sigma \nabla p_i - \nabla^\perp s_i \|_{0,\Omega} = 0 \qquad \forall \, i \in \{1, 2, \ldots, L\}. \tag{2.3}$$

With a complete set of boundary data associated with σ^*, the converse statement, which confirms uniqueness, is also true: if (2.1)–(2.3) are satisfied for every DN pair $\{\mathcal{R}_{\sigma^*}(h), h\}$, then $\sigma = \sigma^*$.

The statement in (2.3) suggests minimizing the functional

$$\mathcal{F}_{KV}(\sigma;\{\mathcal{R}_{\sigma^*}(h_i),h_i\}_{i=1}^L) := \sum_{i=1}^{L}\left(\min_{p_i,s_i\,\in\,\mathrm{H}^1(\Omega)} \|\sigma^{1/2}\nabla p_i - \sigma^{-1/2}\nabla^{\perp}s_i\|_{0,\Omega}^2\right), \tag{2.4}$$

where, for each $i = 1, 2, \ldots, L$, we define

$$\begin{aligned} p_{i|_\Gamma} &:= \mathcal{R}_{\sigma^*}(h_i),\\ s_{i|_\Gamma} &:= \int_\Gamma h_i\, d\gamma\,. \end{aligned} \tag{2.5}$$

The rescaling of conductivity used in (2.4) is chosen so that, for a fixed approximation, σ, the minimizations in (2.4) can be conducted separately. This fact is shown explicitly in Lemma 2.2 below.

The scaling of conductivity used in (2.4) also benefits the updating process for conductivity. For a fixed set of approximations, $\{p_i, s_i\}_{i=1}^L$, it is shown in [20] that the best conductivity is determined pointwise to be

$$\sigma = \left(\sum_{i=1}^{L} |\ \nabla^{\perp}s_i\ |^2\right)^{1/2}\left(\sum_{i=1}^{L} |\ \nabla p_i\ |^2\right)^{-1/2}.$$

This sets up a natural minimization process of alternating between the subspace for σ and those for $\{p_i, s_i\}_{i=1}^L$. Indeed, this alternating process is the design of the proof of the equivalence result. In numerical studies of the KV functional in [20], minimization in σ is incorporated explicitly, and, instead of the natural alternating scheme, a Newton quadraticization is used.

2.1.2. Equivalence to OLS. Here we establish equivalence of the KV functional to that of OLS. Because the line of proof follows for each DN pair, it is convenient to denote

$$\mathrm{F}(\sigma,p,s) := \|\sigma^{1/2}\nabla p - \sigma^{-1/2}\nabla^{\perp}s\|_{0,\Omega}.$$

Also, we observe that, for any DN pair $\{\mathcal{R}_{\sigma^*}(h),h\}$, the exact conductivity generates solutions p^* and s^* to (2.1) and (2.2), respectively, such that $\mathrm{F}(\sigma^*,p^*,s^*) = 0$. For any arbitrary $p\,, s \in \mathrm{H}^1(\Omega)$ satisfying (2.5), we can interpret $\mathrm{F}(\sigma,p,s)$ as an expression of error in each variable. Thus, letting $\mathrm{H}_0^1(\Omega) := \{q \in \mathrm{H}^1(\Omega) : q = 0 \text{ on } \Gamma\}$, it is useful to define

$$\begin{aligned} \delta p &:= p - p^* \in \mathrm{H}_0^1(\Omega),\\ \delta s &:= s - s^* \in \mathrm{H}_0^1(\Omega). \end{aligned}$$

Note that since δp and δs are zero on the boundary, we preserve the boundary values of p^* and s^*, as required by the functional, and we can rewrite (2.4) subject to (2.5) as

$$\mathcal{F}_{KV}(\sigma;\{\mathcal{R}_{\sigma^*}(h_i),h_i\}_{i=1}^L) = \sum_{i=1}^{L}\left(\min_{\delta p_i,\delta s_i\,\in\,\mathrm{H}_0^1(\Omega)} \mathrm{F}(\sigma,p_i^*+\delta p_i,s_i^*+\delta s_i)\right). \tag{2.6}$$

The equivalence relation we seek, Corollary 2.3 below, follows from a generalized ellipticity property for functional term $\mathrm{F}(\sigma,p^*+\delta p,s^*+\delta s)$. As already mentioned, this

property of the functional is established in Lemma 2.2. Essentially, in the following theorem we establish that $\mathrm{F}(\sigma, p^*+\delta p, s^*+\delta s)$ is equivalent to an expression of specific quantities $\mathcal{R}_\sigma(h) - \mathcal{R}_{\sigma^*}(h)$, $\delta p - \delta p_D$, and $\delta s - \delta s_D$, where

$$\begin{aligned} \delta p_D &:= \arg \min_{\delta p \,\in\, \mathrm{H}^1_0(\Omega)} \|\sigma^{1/2}\nabla(p^* + \delta p)\|_{0,\Omega}, \\ \delta s_D &:= \arg \min_{\delta s \,\in\, \mathrm{H}^1_0(\Omega)} \|\sigma^{-1/2}\nabla^\perp(s^* + \delta s)\|_{0,\Omega} \end{aligned} \tag{2.7}$$

come from the minimizations in (2.6). Hence by design, proceeding to conduct the minimizations in (2.6) leads to lower and upper bounds for $\mathcal{F}_{KV}(\sigma; \{\mathcal{R}_{\sigma^*}(h_i), h_i\}_{i=1}^L)$ in terms of the defect in the NtD map.

THEOREM 2.1. *Suppose that $\sigma \in \mathrm{L}^\infty_+(\Omega)$ satisfies $\sigma = \sigma^*$ on Γ. Suppose also that $s^* \in \mathrm{H}^1(\Omega)$ is such that $\mathrm{F}(\sigma^*, p^*, s^*) = 0$ for DN triple $\{p^*, \mathcal{R}_{\sigma^*}(h), h\}$ associated with σ^*, where $h \in \mathrm{H}^{-1/2}_0(\Gamma)$. Finally, let δp_D and δs_D be as defined in* (2.7). *Then there exist positive constants c_0 and c_1 such that*

$$\begin{aligned} c_0\|(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})h\|^2_{1/2,\Gamma} + \|\sigma^{1/2}\nabla(\delta p - \delta p_D)\|^2_{0,\Omega} + \|\sigma^{-1/2}\nabla^\perp(\delta s - \delta s_D)\|^2_{0,\Omega} \\ \le F(\sigma, p^* + \delta p, s^* + \delta s) \\ \le c_1\|(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})h\|^2_{1/2,\Gamma} + \|\sigma^{1/2}\nabla(\delta p - \delta p_D)\|^2_{0,\Omega} + \|\sigma^{-1/2}\nabla^\perp(\delta s - \delta s_D)\|^2_{0,\Omega}, \end{aligned} \tag{2.8}$$

where c_0 depends on Ω and $\min_\Omega |\sigma|$, $c_1 = \|\mathcal{R}_\sigma^{-1}\|_{\mathcal{L}(\mathrm{H}^{1/2}_0(\Gamma), \mathrm{H}^{-1/2}_0(\Gamma))}$, and each multiply only the first terms.

Proof. The optimality of δp_D and δs_D, characterized in (2.7), leads to the orthogonality conditions

$$\begin{aligned} \langle \sigma\nabla(p^* + \delta p_D), \nabla\delta p\rangle_{0,\Omega} = 0 \qquad \forall\, \delta p \in \mathrm{H}^1_0(\Omega), \\ \langle \sigma^{-1}\nabla^\perp(s^* + \delta s_D), \nabla^\perp\delta s\rangle_{0,\Omega} = 0 \qquad \forall\, \delta s \in \mathrm{H}^1_0(\Omega). \end{aligned} \tag{2.9}$$

Note that the Neumann data for δp_D and δs_D are generally not zero and that, in preserving the Dirichlet data, the strong form of the first equation in (2.9) means that $\{p^* + \delta p_D, \mathcal{R}_\sigma(h), h\}$ is a DN triple. Now we expand the terms of the functional, use (2.9), and note that all but one cross term vanishes upon integrating by parts:

$$\begin{aligned} F(\sigma, p^* + \delta p, s^* + \delta s) \\ &= \|\sigma^{1/2}\nabla(p^* + \delta p_D + \delta p - \delta p_D) - \sigma^{-1/2}\nabla^\perp(s^* + \delta s_D + \delta s - \delta s_D)\|^2_{0,\Omega} \\ &= \|\sigma^{1/2}(\nabla p^* + \delta p_D)\|^2_{0,\Omega} + \|\sigma^{1/2}\nabla(\delta p - \delta p_D)\|^2_{0,\Omega} - 2\langle \nabla p^*, \nabla^\perp s^*\rangle_{0,\Omega} \\ &\quad + \|\sigma^{-1/2}\nabla^\perp(s^* + \delta s_D)\|^2_{0,\Omega} + \|\sigma^{-1/2}\nabla^\perp(\delta s - \delta s_D)\|^2_{0,\Omega}. \end{aligned} \tag{2.10}$$

By assumption, $\sigma^*\nabla p^* = \nabla^\perp s^*$, so we may split the lone cross term in (2.10) to write

$$\begin{aligned} F(\sigma, p^* + \delta p, s^* + \delta s) = \|\sigma^{1/2}\nabla(p^* + \delta p_D)\|^2_{0,\Omega} - \|\sigma^{*1/2}\nabla p^*\|^2_{0,\Omega} \\ + \|\sigma^{-1/2}\nabla^\perp(s^* + \delta s_D)\|^2_{0,\Omega} - \|\sigma^{*-1/2}\nabla^\perp s^*\|^2_{0,\Omega} \\ + \|\sigma^{1/2}\nabla(\delta p - \delta p_D)\|^2_{0,\Omega} + \|\sigma^{-1/2}\nabla^\perp(\delta s - \delta s_D)\|^2_{0,\Omega}. \end{aligned} \tag{2.11}$$

With the last two terms as desired, consider the other four. The second equation in (2.9) implies that $\sigma^{-1}\nabla^\perp(s^*+\delta s_D)$ is equal to a grad. Thus, there exists a $\delta p_N \in \mathrm{H}^1(\Omega)$ such that

$$\sigma\nabla(p^* + \delta p_N) = \nabla^\perp(s^* + \delta s_D).$$

Together with the Dirichlet data for $s = s^* + \delta s_D$ in (2.5), we have

$$\mathbf{n} \cdot \sigma \nabla (p^* + \delta p_N) = \mathbf{n} \cdot \nabla^\perp (s^* + \delta s_D) = h. \tag{2.12}$$

In other words, we may think of $\{p^* + \delta p_N, \mathcal{R}_\sigma(h), h\}$ as a DN triple for σ. This allows us to write (2.11) as

$$\begin{aligned} F(\sigma, p^* + \delta p, s^* + \delta s) = {} & \|\sigma^{1/2} \nabla (p^* + \delta p_D)\|_{0,\Omega}^2 - \|\sigma^{*1/2} \nabla p^*\|_{0,\Omega}^2 \\ & + \|\sigma^{1/2} \nabla (p^* + \delta p_N)\|_{0,\Omega}^2 - \|\sigma^{*1/2} \nabla p^*\|_{0,\Omega}^2 \\ & + \|\sigma^{1/2} \nabla (\delta p - \delta p_D)\|_{0,\Omega}^2 + \|\sigma^{-1/2} \nabla^\perp (\delta s - \delta s_D)\|_{0,\Omega}^2. \end{aligned} \tag{2.13}$$

Next, integration by parts yields the following relations:

$$\begin{aligned} \langle \mathcal{R}_{\sigma^*}(h), h \rangle_{0,\Gamma} &= \|\sigma^{*1/2} \nabla p^*\|_{0,\Omega}^2, \\ \langle \mathcal{R}_{\sigma}(h), h \rangle_{0,\Gamma} &= \|\sigma^{1/2} \nabla (p^* + \delta p_N)\|_{0,\Omega}^2, \\ \left\langle \mathcal{R}_{\sigma^*}^{-1}(\mathcal{R}_{\sigma^*}(h)), \mathcal{R}_{\sigma^*}(h) \right\rangle_{0,\Gamma} &= \|\sigma^{*1/2} \nabla p^*\|_{0,\Omega}^2, \\ \left\langle \mathcal{R}_{\sigma}^{-1}(\mathcal{R}_{\sigma^*}(h)), \mathcal{R}_{\sigma^*}(h) \right\rangle_{0,\Gamma} &= \|\sigma^{1/2} \nabla (p^* + \delta p_D)\|_{0,\Omega}^2. \end{aligned} \tag{2.14}$$

Thus, (2.13) becomes

$$\begin{aligned} F(\sigma, p^* + \delta p, s^* + \delta s) = {} & \left\langle (\mathcal{R}_\sigma^{-1} - \mathcal{R}_{\sigma^*}^{-1}) \mathcal{R}_{\sigma^*}(h), \mathcal{R}_{\sigma^*}(h) \right\rangle_{0,\Gamma} + \langle (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h), h \rangle_{0,\Gamma} \\ & + \|\sigma^{1/2} \nabla (\delta p - \delta p_D)\|_{0,\Omega}^2 + \|\sigma^{-1/2} \nabla^\perp (\delta s - \delta s_D)\|_{0,\Omega}^2. \end{aligned} \tag{2.15}$$

Since $\mathcal{R}_\sigma$ and $\mathcal{R}_{\sigma^*}$ are self-adjoint, we may combine the first two terms:

$$\begin{aligned} & \left\langle (\mathcal{R}_\sigma^{-1} - \mathcal{R}_{\sigma^*}^{-1}) \mathcal{R}_{\sigma^*}(h), \mathcal{R}_{\sigma^*}(h) \right\rangle_{0,\Gamma} + \langle (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h), h \rangle_{0,\Gamma} \\ & \quad = \left\langle (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*} + \mathcal{R}_{\sigma^*} \mathcal{R}_\sigma^{-1} \mathcal{R}_{\sigma^*} - \mathcal{R}_{\sigma^*})(h), h \right\rangle_{0,\Gamma} \\ & \quad = \left\langle (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*}) \mathcal{R}_\sigma^{-1} (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h), h \right\rangle_{0,\Gamma} \\ & \quad = \left\langle \mathcal{R}_\sigma^{-1} (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h), (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h) \right\rangle_{0,\Gamma}. \end{aligned} \tag{2.16}$$

Also, since

$$\begin{aligned} & \langle \mathcal{R}_\sigma^{-1} (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h), (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*}) h \rangle_{0,\Gamma} \\ & \qquad \leq \|\mathcal{R}_\sigma^{-1}\|_{\mathcal{L}(\mathrm{H}_0^{1/2}(\Gamma), \mathrm{H}_0^{-1/2}(\Gamma))} \|(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h)\|_{1/2,\Gamma}^2, \end{aligned}$$

we may combine (2.15) and (2.16) to arrive at the desired upper bound:

$$\begin{aligned} \mathrm{F}(\sigma, p^* + \delta p, s^* + \delta s) \leq {} & \|\mathcal{R}_\sigma^{-1}\|_{\mathcal{L}(\mathrm{H}_0^{1/2}(\Gamma), \mathrm{H}_0^{-1/2}(\Gamma))} \|\mathcal{R}_\sigma(h) - \mathcal{R}_{\sigma^*}(h)\|_{1/2,\Gamma}^2 \\ & + \|\sigma^{1/2} \nabla (\delta p - \delta p_D)\|_{0,\Omega}^2 + \|\sigma^{-1/2} \nabla^\perp (\delta s - \delta s_D)\|_{0,\Omega}^2. \end{aligned}$$

To achieve the lower bound, we recall (2.7) and observe that

$$(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h) = \delta p_{N|\Gamma} \in \mathrm{H}_0^{-1/2}(\Gamma)$$

by definition of δp_N. Hence, we may now consider the $\tilde{p} \in \mathrm{H}^1(\Omega)$ that solves

$$\begin{aligned} \nabla \cdot \sigma \nabla \tilde{p} &= 0 && \text{in } \Omega, \\ \tilde{p} &= \delta p_N && \text{on } \Gamma. \end{aligned}$$

Appealing to (2.14) and the trace theorem [11], we have

$$\begin{aligned}\langle \mathcal{R}_\sigma^{-1}(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h), (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h)\rangle_{0,\Gamma} &= \|\sigma^{1/2}\nabla\tilde{p}\|_{0,\Omega}^2 \\ &\geq c\,\|\nabla\tilde{p}\|_{0,\Omega}^2 \\ &\geq c_0\,\|\delta p_N\|_{1/2,\Gamma}^2 \\ &= c_0\|(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h)\|_{1/2,\Gamma}^2, \end{aligned} \tag{2.17}$$

where c_0 depends on the smoothness of the boundary and $\min_\Omega |\sigma|$. Finally, incorporating (2.17) with (2.16) into (2.6) establishes the lower bound:

$$\begin{aligned}\mathrm{F}(\sigma, p^* + \delta p, s^* + \delta s) \geq c_0\|(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h)\|_{1/2,\Gamma}^2 \\ +\|\sigma^{1/2}\nabla(\delta p - \delta p_D)\|_{0,\Omega}^2 + \|\sigma^{-1/2}\nabla^\perp(\delta s - \delta s_D)\|_{0,\Omega}^2. \end{aligned}$$ ∎

To make use of the above equivalence, we establish the following lemma using the constructs of the theorem to reassert the ability to minimize the functional in (2.4) for s_i and p_i separately.

LEMMA 2.2. *Suppose that $\sigma = \sigma^*$ on Γ. Also, for each $h \in \mathrm{H}_0^{-1/2}(\Gamma)$, let $\{p^*, \mathcal{R}_{\sigma^*}(h), h\}$ be a DN triple for σ^* and define*

$$\begin{aligned}\delta p_D &:= \arg \min_{\delta p \,\in\, \mathrm{H}_0^1(\Omega)} \|\sigma^{1/2}\nabla(p^* + \delta p)\|_{0,\Omega}, \\ \delta s_D &:= \arg \min_{\delta s \,\in\, \mathrm{H}_0^1(\Omega)} \|\sigma^{-1/2}\nabla^\perp(s^* + \delta s)\|_{0,\Omega}. \end{aligned} \tag{2.18}$$

Then

$$\mathrm{F}(\sigma, p^* + \delta p_D, s^* + \delta s_D) = \min_{\delta p, \delta s \,\in\, \mathrm{H}_0^1(\Omega)} \mathrm{F}(\sigma, p^* + \delta p, s^* + \delta s). \tag{2.19}$$

Proof. From (2.10) and the boundary data for p^* and s^*, we have

$$\begin{aligned} F(\sigma, p^* + \delta p, s^* + \delta s) \\ &= \|\sigma^{1/2}(\nabla p^* + \delta p_D)\|_{0,\Omega}^2 + \|\sigma^{1/2}\nabla(\delta p - \delta p_D)\|_{0,\Omega}^2 - 2\langle \nabla p^*, \nabla^\perp s^*\rangle_{0,\Omega} \\ &\quad + \|\sigma^{-1/2}\nabla^\perp(s^* + \delta s_D)\|_{0,\Omega}^2 + \|\sigma^{-1/2}\nabla^\perp(\delta s - \delta s_D)\|_{0,\Omega}^2 \\ &= \|\sigma^{1/2}(\nabla p^* + \delta p_D)\|_{0,\Omega}^2 + \|\sigma^{1/2}\nabla(\delta p - \delta p_D)\|_{0,\Omega}^2 - 2\langle \mathcal{R}_{\sigma^*}(h), h\rangle_{0,\Gamma} \\ &\quad + \|\sigma^{-1/2}\nabla^\perp(s^* + \delta s_D)\|_{0,\Omega}^2 + \|\sigma^{-1/2}\nabla^\perp(\delta s - \delta s_D)\|_{0,\Omega}^2. \end{aligned}$$

The conclusion is apparent since the boundary term is a known fixed quantity. ∎

A direct result of this lemma and Theorem 2.1 is the following corollary.

COROLLARY 2.3. *For any $h \in \mathrm{H}_0^{-1/2}(\Gamma)$, there exist positive constants c_0 and c_1 such that*

$$\begin{aligned} c_0\|\mathcal{R}_\sigma(h) - \mathcal{R}_{\sigma^*}(h)\|_{1/2,\Gamma}^2 \\ &\leq \min_{p,s \,\in\, \mathrm{H}^1(\Omega)} \|\sigma^{1/2}\nabla p - \sigma^{-1/2}\nabla^\perp s\|_{0,\Omega}^2 \\ &\leq c_1\|\mathcal{R}_\sigma(h) - \mathcal{R}_{\sigma^*}(h)\|_{1/2,\Gamma}^2, \end{aligned} \tag{2.20}$$

where c_0 depends on Ω and $\min_\Omega |\sigma|$, $c_1 = \|\mathcal{R}_\sigma^{-1}\|_{\mathcal{L}(\mathrm{H}_0^{1/2}(\Gamma), \mathrm{H}_0^{-1/2}(\Gamma))}$, and where p and s are subject to the boundary data in (2.5).

It is also clear from Lemma 2.2, as a generalization of the above corollary, that the KV formulation is equivalent to the OLS formulation. We state this formally in the following corollary.

COROLLARY 2.4. *With $\mathcal{F}_{KV}$ defined in* (2.4) *subject to* (2.5)*, there exist positive constants c_0 and c_1 such that*

$$\begin{aligned} c_0 \sum_{i=1}^{L} \|\mathcal{R}_\sigma(h_i) - \mathcal{R}_{\sigma^*}(h_i)\|^2_{1/2,\Gamma} & \\ \leq \mathcal{F}_{KV}(\sigma; \{\mathcal{R}_{\sigma^*}(h_i), h_i\}_{i=1}^{L}) & \\ \leq c_1 \sum_{i=1}^{L} \|\mathcal{R}_\sigma(h_i) - \mathcal{R}_{\sigma^*}(h_i)\|^2_{1/2,\Gamma}, & \end{aligned} \tag{2.21}$$

where $c_1 = \|\mathcal{R}_\sigma^{-1}\|_{\mathcal{L}(\mathrm{H}_0^{1/2}(\Gamma),\mathrm{H}_0^{-1/2}(\Gamma))}$ and c_0 depends on Ω and $\min_\Omega |\sigma|$.

Remark 4. As a final remark, consider (2.20) under the assumption of complete knowledge of the NtD map. Taking the sup over all $h \in \mathrm{H}_0^{-1/2}(\Gamma)$ of each quantity and scaling appropriately then leads to the idealized equivalence result:

$$\begin{aligned} c_0 \|\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*}\|^2_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{1/2}(\Gamma))} & \\ \leq \sup_{h \in \mathrm{H}_0^{-1/2}(\Gamma)} \frac{\left(\min_{p,s \,\in\, \mathrm{H}^1(\Omega)} \|\sigma^{1/2}\nabla p - \sigma^{-1/2}\nabla^\perp s\|^2_{0,\Omega}\right)}{\|h\|_{-1/2,\Gamma}} & \\ \leq c_1 \|\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*}\|^2_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{1/2}(\Gamma))}, & \end{aligned}$$

where, again, p and s are subject to the boundary data in (2.5) for a given h. This suggests a sup version, rather than least-squares version, of the KV functional that is equivalent to the operator norm of the defect in the NtD maps.

2.2. FOSLS. We now present a FOSLS formulation for ECT. The first-order system we use may be viewed as a rescaled version of the Maxwell equations. We will show that the FOSLS formulation not only provides the ability to solve forward problems efficiently but also yields an insightful process for updating the approximate conductivity. Again, the result is a natural alternating minimization scheme. We present the FOSLS formulation and provide an equivalence result akin to that of the previous section.

2.2.1. Formulation. To derive the FOSLS formulation for ECT, define $q := \frac{1}{2}\log(\sigma)$ so that $e^{2q} := \sigma$. This is natural thing to do for deeper reasons than the positivity of conductivity, since the reciprocal of conductivity, resistivity, is also of interest. It is therefore seemingly arbitrary whether one considers an error in conductivity, $\sigma - \sigma^*$, or the associated error in resistivity, $1/\sigma - 1/\sigma^*$. However, by defining error in terms of q, such an inconsistency is avoided since

$$|q - q^*| = \frac{1}{2}\left|\log\left(\frac{\sigma}{\sigma^*}\right)\right| = \frac{1}{2}\left|\log\left(\frac{1/\sigma}{1/\sigma^*}\right)\right|. \tag{2.22}$$

Proceeding with the formulation, we now let $\mathbf{u} := e^q \nabla p$. Then, with $\{p, \mathcal{R}_\sigma(h), h\}$ a DN triple for σ, $\mathbf{u}$ solves the first-order system

$$\begin{aligned} e^{-q}\nabla \cdot e^q \mathbf{u} &= 0 && \text{in } \Omega, \\ e^q \nabla \times e^{-q}\mathbf{u} &= 0 && \text{in } \Omega, \\ \mathbf{n} \cdot e^q \mathbf{u} &= h && \text{on } \Gamma. \end{aligned} \tag{2.23}$$

We may think of (2.23) as determining the tangential current

$$\mathbf{n} \times e^{-q}\mathbf{u} = (\mathbf{n} \times \nabla)\mathcal{R}_\sigma(h) \in (\mathrm{H}_0^{-1/2}(\Gamma))^{n-1}.$$

When $n = 2$, we write $(\boldsymbol{\tau}\cdot)$ in lieu of $(\mathbf{n}\times)$. Practically speaking, deriving tangential current data from voltage data is problematic at best, since simply differentiating noisy data is an unstable process. Though there may be future innovations that enable direct assessment of tangential current, we defer discussion of this important matter to future research.

We now make our first definition regarding the FOSLS formulation.

DEFINITION 2.5. *First-order system* (2.23) *defines the normal-to-tangential current map (NtT map),* $\mathcal{P}_q : \mathrm{H}_0^{-1/2}(\Gamma) \longrightarrow (\mathrm{H}_0^{-1/2}(\Gamma))^{n-1}$, *where*

$$\mathcal{P}_q(h) = \mathbf{n} \times e^{-q}\mathbf{u} = (\mathbf{n} \times \nabla)\mathcal{R}_\sigma(h).$$

We also define the following variations on $\mathrm{H}(\mathrm{div}, \Omega)$ and $\mathrm{H}(curl, \Omega)$.

DEFINITION 2.6. *For* $q \in \mathrm{L}^\infty(\Omega)$, *let* $\mathrm{W}_q = \mathrm{H}(\mathrm{div}\, e^q, \Omega) \cap \mathrm{H}(curl\, e^{-q}, \Omega)$, *where*

$$\mathrm{H}(\mathrm{div}\, e^q, \Omega) = \{\mathbf{u} \in (\mathrm{L}^2(\Omega))^n : \nabla \cdot e^q\mathbf{u} \in \mathrm{L}^2(\Omega)\},$$
$$\mathrm{H}(curl\, e^{-q}, \Omega) = \left\{\mathbf{u} \in (\mathrm{L}^2(\Omega))^n : \nabla \times e^{-q}\mathbf{u} \in (\mathrm{L}^2(\Omega))^{2n-3}\right\},$$

for $n = 2, 3$.

Finally, it is convenient to introduce an analogy to the DN-triple notation.

DEFINITION 2.7. *In the FOSLS setting, if* $\mathbf{u}$ *solves* (2.23) *with boundary data* h, *then we refer to* $\{\mathbf{u}, h, \mathcal{P}_{q^*}(h)\}$ *as a normal-tangential triple (NT triple) for* q *and to* $\{h, \mathcal{P}_{q^*}(h)\}$ *as an NT pair.*

Given the boundary data, $\{h_i, \mathcal{P}_{q^*}(h_i)\}_{i=1}^L$, associated with an unknown q^*, we approximate q^* by minimizing the FOSLS functional

$$\mathcal{F}_{OSLS}(q; \{h_i, \mathcal{P}_{q^*}(h_i)\}_{i=1}^L) = \sum_{i=1}^{L} \left(\min_{\mathbf{u}_i \in \mathrm{W}_q} \mathrm{F}(q, \mathbf{u}_i; h_i, \mathcal{P}_{q^*}(h_i)) \right), \tag{2.24}$$

where

$$\begin{aligned} \mathrm{F}(q, \mathbf{u}_i; h_i, \mathcal{P}_{q^*}(h_i)) &:= \|e^{-q}\nabla \cdot e^q\mathbf{u}_i\|_{0,\Omega}^2 + \|e^q\nabla \times e^{-q}\mathbf{u}_i\|_{0,\Omega}^2 \\ &\quad + \|\mathbf{n} \cdot e^q\mathbf{u}_i - h_i\|_{-1/2,\Gamma}^2 + \|\mathbf{n} \times e^{-q}\mathbf{u}_i - \mathcal{P}_{q^*}(h_i)\|_{-1/2,\Gamma}^2. \end{aligned} \tag{2.25}$$

To minimize $\mathcal{F}_{OSLS}$, first note that, for a fixed approximation, q, each $\mathbf{u}_i$ can be numerically approximated with optimal efficiency [5]. Furthermore, if we assume $q \in \mathrm{H}^1(\Omega)$ and $\mathbf{u} \in \mathrm{W}_0 = \mathrm{H}(\mathrm{div}, \Omega) \cap \mathrm{H}(curl, \Omega)$, then for fixed $\mathbf{u}_i \in (\mathrm{H}^1(\Omega))^n$, with $i = 1, \ldots, L$, we also have an elliptic problem to improve approximation $q \in \mathrm{H}^1(\Omega)$. Taking the first variation in q of $\mathcal{F}_{OSLS}$ for a fixed set, $\{\mathbf{u}_i \in (\mathrm{H}^1(\Omega))^2\}_{i=1}^L$, reveals that the best $q \in \mathrm{H}^1(\Omega)$ satisfies the weak form

$$\langle A\nabla q, \nabla r\rangle_{0,\Omega} = f(r) \qquad \forall\, r \in \mathrm{H}^1(\Omega), \tag{2.26}$$

where

$$A = 2\sum_{i=1}^{L} |\mathbf{u}_i|^2 \tag{2.27}$$

and $f \in \mathrm{H}^{-1}(\Omega)$, for $n = 2$ for example, is given by

$$\langle f, r\rangle_{0,\Omega} = \sum_{i=1}^{L} \left(\langle \nabla \cdot \mathbf{u}_i, \nabla r \cdot \mathbf{u}_i \rangle_{0,\Omega} + \langle \nabla \times \mathbf{u}_i, \nabla^{\perp} r \cdot \mathbf{u}_i \rangle_{0,\Omega} \right),$$

using $(\boldsymbol{\tau}\cdot)$ in lieu of $(\mathbf{n}\times)$ and $(\nabla^{\perp}\cdot)$ in lieu of $(\nabla\times)$.

These properties suggest that we can employ an alternating subspace minimization scheme. Furthermore, it is remarkable that the minimization associated with the weak form in (2.26) incorporates inner products of interior current as a weighting on the components of q, just as it appears in Calderdon's linearization. Thus, the FOSLS functional appears to reduce error in q according to naturally weighted H^1-seminorms. Note that prior information about σ^*, manifested as an improved approximation σ, is reflected in this weighting. Because requiring $q \in \mathrm{H}^1(\Omega)$ is quite restrictive, the OLS or KV approach may be better suited for many applications. However, given the above insights, the role of FOSLS as a framework for EIT is compelling.

2.2.2. Equivalence to OLS. As with the KV functional, we demonstrate equivalence for $n = 2$. To show that the FOSLS functional in (2.24) is equivalent to OLS, we first establish several lemmas and some useful notation. In the first lemma, we relate $\mathcal{P}_q$ to $\mathcal{R}_\sigma$ in two dimensions. Generalization of this lemma to the three-dimensional problem is not obvious, and not addressed in this study. In three dimesions, complications arise since $\mathcal{P}_q$ is then a 2-vector, whereas $\mathcal{R}_\sigma(h)$ remains a scalar quantity. However, the increased dimensionality of the data in three dimensions can be expected to improve the quality of the reconstruction.

LEMMA 2.8. *Let $\sigma = \sigma^*$ on Γ and $n = 2$. Then*

$$\|\mathcal{P}_q - \mathcal{P}_{q^*}\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{-1/2}(\Gamma))} = \|\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*}\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{1/2}(\Gamma))}. \tag{2.28}$$

Proof. We wish to show that

$$\sup_{h\in\mathrm{H}_0^{-1/2}(\Gamma)} \frac{\|(\mathcal{P}_q - \mathcal{P}_{q^*})(h)\|_{-1/2,\Gamma}}{\|h\|_{-1/2,\Gamma}} = \sup_{h\in\mathrm{H}_0^{-1/2}(\Gamma)} \frac{\|(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h)\|_{1/2,\Gamma}}{\|h\|_{-1/2,\Gamma}}.$$

Since $(\mathcal{P}_q - \mathcal{P}_{q^*})(h) = (\boldsymbol{\tau}\cdot\nabla)(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h)$, it suffices to show that

$$\|(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h)\|_{1/2,\Gamma} = \|(\boldsymbol{\tau}\cdot\nabla)(\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})(h)\|_{-1/2,\Gamma}.$$

Noting that there is a $g \in \mathrm{H}^{1/2}(\Gamma)$ such that $g = (\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*})h$, then this reduces to

$$\|g\|_{1/2,\Gamma} = \left\| \frac{dg}{d\tau} \right\|_{-1/2,\Gamma}. \tag{2.29}$$

We now show that this holds for any $g \in \mathrm{H}_0^{1/2}(\Gamma)$ by appealing to the representation of the $\mathrm{H}^{1/2}$- and $\mathrm{H}^{-1/2}$-norms using the Fourier transform. Defining $\hat{g}(\omega)$ as the transform of g, we have [24]

$$\|g\|_{1/2,\Gamma} = \|\omega^{1/2}\hat{g}\|_{0,\Gamma},$$
$$\|h\|_{-1/2,\Gamma} = \|\omega^{-1/2}\hat{h}\|_{0,\Gamma}.$$

Hence, (2.29) follows from

$$\|g\|_{1/2,\Gamma} = \|\omega^{1/2}\hat{g}\|_{0,\Gamma} = \|\omega^{-1/2}\omega\hat{g}\|_{0,\Gamma} = \|\omega^{-1/2}\hat{g'}\|_{0,\Gamma} = \left\| \frac{dg}{d\tau} \right\|_{-1/2,\Gamma}. \qquad \square$$

Remark 5. An alternative proof may be preferable in which (2.29) is derived by interpolating from an analogous statement of integer index.

It is convenient to isolate the interior and boundary functional terms of the FOSLS functional in (2.24) and write

$$\begin{aligned}
\mathrm{F}(q,\mathbf{u}_i;h,\mathcal{P}_{q^*}(h_i)) &:= \mathrm{F_I}(q,\mathbf{u}_i) + \mathrm{F_B}(q,\mathbf{u}_i;h_i,\mathcal{P}_{q^*}(h_i)),\\
\mathrm{F_I}(q,\mathbf{u}_i) &:= \|e^{-q}\nabla\cdot e^{q}\mathbf{u}_i\|^2_{0,\Omega} + \|e^{q}\nabla\times e^{-q}\mathbf{u}_i\|^2_{0,\Omega},\\
\mathrm{F_B}(q,\mathbf{u}_i;h_i,\mathcal{P}_{q^*}(h_i)) &:= \|\mathbf{n}\cdot e^{q}\mathbf{u}_i - h_i\|^2_{-1/2,\Gamma} + \|\boldsymbol{\tau}\cdot e^{-q}\mathbf{u}_i - \mathcal{P}_{q^*}(h_i)\|^2_{-1/2,\Gamma}.
\end{aligned} \tag{2.30}$$

The following lemmas set the stage for stating and proving the equivalence results. The first establishes a useful equivalence relation between F and the norm defined on $\mathrm{W}_q(\Omega)$ by

$$\|\mathbf{w}\|^2_{\mathrm{W}_q} := \|\mathbf{w}\|^2_{0,\Omega} + \|e^{-q}\nabla\cdot e^{q}\mathbf{w}\|^2_{0,\Omega} + \|e^{q}\nabla\times e^{-q}\mathbf{w}\|^2_{0,\Omega}.$$

LEMMA 2.9. *Let* $\mathbf{w}\in\mathrm{W}_q(\Omega)$. *Then there exists a constant, c, depending on q and* Ω, *such that*

$$\frac{1}{c}\|\mathbf{w}\|^2_{\mathrm{W}_q} \le \mathrm{F}(q,\mathbf{w};0,0) \le c\|\mathbf{w}\|^2_{\mathrm{W}_q}. \tag{2.31}$$

Proof. The upper bound in (2.31) follows directly from the trace theorem [11]. For the lower bound, it suffices to show there is a c such that

$$\begin{aligned}
&\|\mathbf{w}\|^2_{0,\Omega}\\
&\quad\le c\left(\|e^{-q}\nabla\cdot e^{q}\mathbf{w}\|^2_{0,\Omega} + \|e^{q}\nabla\times e^{-q}\mathbf{w}\|^2_{0,\Omega} + \|\mathbf{n}\cdot e^{q}\mathbf{w}\|^2_{-1/2,\Gamma} + \|\boldsymbol{\tau}\cdot e^{-q}\mathbf{w}\|^2_{-1/2,\Gamma}\right).
\end{aligned}$$

To this end, we use a scaled Helmholtz decomposition [11]: $\forall\ \mathbf{v}\in(\mathrm{L}^2(\Omega))^2$, there exist $\phi\in\mathrm{H}^1_0(\Omega)$ and $\psi\in\mathrm{H}^1(\Omega)/\mathbb{R}$ such that

$$\mathbf{v} = e^{q}\nabla\phi + e^{-q}\nabla^{\perp}\psi.$$

The logic leading to this representation is to choose ϕ so that $\nabla\cdot e^{2q}\nabla\phi = \nabla\cdot e^{q}\mathbf{v}$ and to similarly choose ψ to satisfy $\nabla\times e^{-2q}\nabla^{\perp}\psi = \nabla\times e^{-q}\mathbf{v}$. The choice of boundary conditions on ϕ and ψ enables us to eliminate the cross term that would otherwise appear below and to use a Poincaré–Friedrichs inequality. With this decomposition, we write

$$\begin{aligned}
\|\mathbf{w}\|_{0,\Omega} &= \sup_{\mathbf{v}\in(\mathrm{L}^2(\Omega))^2} \frac{\langle\mathbf{w},\mathbf{v}\rangle_{0,\Omega}}{\|\mathbf{v}\|_{0,\Omega}}\\
&= \sup_{\phi,\psi\in\mathrm{H}^1(\Omega)} \frac{\langle\mathbf{w},e^{q}\nabla\phi + e^{-q}\nabla^{\perp}\psi\rangle_{0,\Omega}}{\|e^{q}\nabla\phi + e^{-q}\nabla^{\perp}\psi\|_{0,\Omega}}\\
&\le \sup_{\phi\in\mathrm{H}^1(\Omega)} \frac{\langle\mathbf{w},e^{q}\nabla\phi\rangle_{0,\Omega}}{\|e^{q}\nabla\phi\|_{0,\Omega}} + \sup_{\psi\in\mathrm{H}^1(\Omega)} \frac{\langle\mathbf{w},e^{-q}\nabla^{\perp}\psi\rangle_{0,\Omega}}{\|e^{-q}\nabla^{\perp}\psi\|_{0,\Omega}}\\
&= \frac{\langle\mathbf{w},e^{q}\nabla\phi^*\rangle_{0,\Omega}}{\|e^{q}\nabla\phi^*\|_{0,\Omega}} + \frac{\langle\mathbf{w},e^{-q}\nabla^{\perp}\psi^*\rangle_{0,\Omega}}{\|e^{-q}\nabla^{\perp}\psi^*\|_{0,\Omega}}.
\end{aligned} \tag{2.32}$$

Applying Green's theorem to the first term, we have

$$\begin{aligned}
&\frac{\langle \mathbf{w}, e^q\nabla\phi^*\rangle_{0,\Omega}}{\|e^q\nabla\phi^*\|_{0,\Omega}} \\
&\quad\le \frac{|\langle e^{-q}\nabla\cdot e^q\mathbf{w}, e^q\phi^*\rangle_{0,\Omega}| + |\langle \mathbf{n}\cdot e^q\mathbf{w}, \phi^*\rangle_{0,\Gamma}|}{\|e^q\nabla\phi^*\|_{0,\Omega}} \\
&\quad\le \frac{\|e^{-q}\nabla\cdot e^q\mathbf{w}\|_{0,\Omega}\|e^q\phi^*\|_{0,\Omega} + \|\mathbf{n}\cdot e^q\mathbf{w}\|_{-1/2,\Gamma}\|\phi^*\|_{1/2,\Gamma}}{\|e^q\nabla\phi^*\|_{0,\Omega}} \\
&\quad\le \left(\|e^{-q}\nabla\cdot e^q\mathbf{w}\|_{0,\Omega}^2 + \|\mathbf{n}\cdot e^q\mathbf{w}\|_{-1/2,\Gamma}^2\right)^{1/2}\left(\frac{\|e^q\phi^*\|_{0,\Omega}^2 + \|\phi^*\|_{1/2,\Gamma}^2}{\|e^q\nabla\phi^*\|_{0,\Omega}^2}\right)^{1/2}. \qquad (2.33)
\end{aligned}$$

The term in (2.33) that does not involve $\mathbf{w}$ may be bounded by a constant depending on q and Ω. To see this, note that we can use the Poincaré–Friedrichs inequality [11] to write

$$\frac{\|e^q\phi^*\|_{0,\Omega}^2 + \|\phi^*\|_{1/2,\Gamma}^2}{\|e^q\nabla\phi^*\|_{0,\Omega}^2} \le \frac{\max_\Omega\{e^{2q},1\}}{\min_\Omega e^{2q}}\frac{(\|\phi^*\|_{0,\Omega}^2 + \|\phi^*\|_{1/2,\Gamma}^2)}{\|\nabla\phi^*\|_{0,\Omega}^2} \le c(q),$$

where

$$c(q) = c\frac{\max_\Omega\{e^{2q},1\}}{\min_\Omega e^{2q}}. \qquad (2.34)$$

Thus, (2.33) becomes

$$\frac{\langle \mathbf{w}, e^q\nabla\phi^*\rangle_{0,\Omega}}{\|e^q\nabla\phi^*\|_{0,\Omega}} \le c_1(q)\left(\|e^{-q}\nabla\cdot e^q\mathbf{w}\|_{0,\Omega}^2 + \|\mathbf{n}\cdot e^q\mathbf{w}\|_{-1/2,\Gamma}^2\right)^{1/2}. \qquad (2.35)$$

Similarly, there is a $c_2(q)$ such that

$$\frac{|\,\langle \mathbf{w}, e^{-q}\nabla^\perp\psi^*\rangle_{0,\Omega}\,|}{\|e^{-q}\nabla^\perp\psi^*\|_{0,\Omega}} \le c_2(q)\left(\|e^q\nabla\times e^{-q}\mathbf{w}\|_{0,\Omega}^2 + \|\boldsymbol{\tau}\cdot e^{-q}\mathbf{w}\|_{-1/2,\Gamma}^2\right)^{1/2}. \qquad (2.36)$$

Therefore, combining inequalities (2.35) and (2.36) and choosing c appropriately proves the lemma. ∎

We also make use of the next lemma to relate the FOSLS functional directly to the NtT map.

LEMMA 2.10. *Let* $\mathbf{w}\in \mathrm{W}_q(\Omega)$, *and suppose* $\mathbf{n}\cdot e^q\mathbf{w} = h$ *on* Γ. *If* $\mathbf{w}_h$ *satisfies*

$$\begin{aligned}
e^{-q}\nabla\cdot e^q\mathbf{w}_h &= 0 && \text{in } \Omega,\\
e^q\nabla\times e^{-q}\mathbf{w}_h &= 0 && \text{on } \Gamma,\\
\mathbf{n}\cdot e^q\mathbf{w}_h &= h && \text{on } \Gamma, \qquad (2.37)
\end{aligned}$$

then there exists a constant, c, depending only on Ω *and* $\max_\Omega e^q$, *such that* $\boldsymbol{\tau}\cdot e^{-q}(\mathbf{w}-\mathbf{w}_h)$ *satisfies*

$$\|\boldsymbol{\tau}\cdot e^{-q}(\mathbf{w}-\mathbf{w}_h)\|_{-1/2,\Gamma}^2 \le c\,\mathrm{F}_I(q,\mathbf{w}). \qquad (2.38)$$

Proof. We have from the trace theorem [11] that

$$\begin{aligned}
\|\boldsymbol{\tau}\cdot e^{-q}(\mathbf{w}-\mathbf{w}_h)\|_{-1/2,\Gamma}^2 &\le c\left(\|\nabla\times e^{-q}(\mathbf{w}-\mathbf{w}_h)\|_{0,\Omega}^2 + \|e^{-q}(\mathbf{w}-\mathbf{w}_h)\|_{0,\Omega}^2\right)\\
&\le c\left(\|\nabla\times e^{-q}(\mathbf{w}-\mathbf{w}_h)\|_{0,\Omega}^2 + \|e^{q}(\mathbf{w}-\mathbf{w}_h)\|_{0,\Omega}^2\right),
\end{aligned}$$

where c depends on Ω and $\max_\Omega e^q$. Since, by assumption, $\mathbf{n} \cdot e^q(\mathbf{w} - \mathbf{w}_h) = 0$ on the boundary, we may bound the L^2-term above using a Poincaré–Friedrichs-like inequality of the form [2]

$$\|e^q(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega}^2 \leq c \left(\|\nabla \times e^{-q}(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega}^2 + \|\nabla \cdot e^q(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega}^2 \right), \tag{2.39}$$

where c depends on Ω and $\max_\Omega e^q$. We therefore have

$$\begin{aligned} \|\boldsymbol{\tau} \cdot e^{-q}(\mathbf{w} - \mathbf{w}_h)\|_{-1/2,\Gamma}^2 &\leq c \left(\|\nabla \times e^{-q}(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega}^2 + \|\nabla \cdot e^q(\mathbf{w} - \mathbf{w}_h)\|_{0,\Omega}^2 \right) \\ &\leq c\, \mathrm{F}_I(q, \mathbf{w}), \end{aligned}$$

where, again, c depends on Ω and $\max_\Omega e^q$. ∎

The above lemmas now allow us to establish a lower bound for the FOSLS functional in terms of the defect in the ND map. In analogy to the equivalence established for the KV formulation, the line of proof follows for each NT pair. To begin, let $\{\mathbf{u}^*, h, \mathcal{P}_{q^*}(h)\}$ be an NT triple for unknown $\sigma^* = e^{2q^*}$. This implies that $\mathrm{F}(q^*, \mathbf{u}^*; h, \mathcal{P}_{q^*}(h)) = 0$. For arbitrary $q \in \mathrm{H}^1(\Omega)$, provided $q_{|\Gamma} = q^*_{|\Gamma}$, and arbitrary $\mathbf{u} \in \mathrm{W}_q(\Omega)$, we wish to interpret how $\mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h))$ expresses errors $q - q^*$ and $\mathbf{u} - \mathbf{u}^*$. To facilitate this analysis, we write the perturbed functional as

$$\mathrm{F}(q, \mathbf{u}^* + \delta\mathbf{u}; h, \mathcal{P}_{q^*}(h)),$$

where $\delta\mathbf{u} \in \mathrm{W}_q(\Omega)$. It is convenient to make the following definition.

Definition 2.11. *Denote*

$$\mathrm{S}_q := \{\mathbf{u} \in \mathrm{W}_q(\Omega) \,:\, \mathrm{F}_\mathrm{I}(q, \mathbf{u}) = 0\},$$

the space of F_I*-harmonic functions relative to* q*.*

Before addressing arbitrary $\mathbf{u} \in \mathrm{W}_q(\Omega)$, and thus arbitrary $\delta\mathbf{u} \in \mathrm{W}_q(\Omega)$, we first consider perturbations $\delta\mathbf{u}$ such that $\mathbf{u} \in \mathrm{S}_q$.

Lemma 2.12. *Let* $\mathbf{u} \in \mathrm{S}_q$*. Then*

$$\frac{1}{c} \|\mathcal{P}_q(h) - \mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma}^2 \leq \mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \qquad \forall\, h \in \mathrm{H}_0^{-1/2}(\Gamma),$$

where $c = 1 + \|\mathcal{P}_q\|^2_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma), \mathrm{H}_0^{-1/2}(\Gamma))}$.

Proof. Since $\mathbf{u} \in \mathrm{S}_q$, we immediately have that $\mathrm{F}_I(q, \mathbf{u}) = 0$. Also, writing $\mathbf{n} \cdot e^q \mathbf{u} = f$, we have that $\boldsymbol{\tau} \cdot e^{-q}\mathbf{u} = \mathcal{P}_q(f)$, and the remaining part of the functional is the boundary term:

$$\begin{aligned} \mathrm{F}_B(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) &= \|f - h\|_{-1/2,\Gamma}^2 + \|\mathcal{P}_q(f) - \mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma}^2 \\ &= \|f - h\|_{-1/2,\Gamma}^2 + \|\mathcal{P}_q(f) - \mathcal{P}_q(h) + \mathcal{P}_q(h) - \mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma}^2 \\ &= \|f - h\|_{-1/2,\Gamma}^2 + \|\mathcal{P}_q(f) - \mathcal{P}_q(h)\|_{-1/2,\Gamma}^2 + \|\mathcal{P}_q(h) - \mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma}^2 \\ &\quad + 2\langle \mathcal{P}_q(f) - \mathcal{P}_q(h), \mathcal{P}_q(h) - \mathcal{P}_{q^*}(h) \rangle_{-1/2,\Gamma}. \end{aligned} \tag{2.40}$$

Now, since $\mathcal{P}_q$ is linear, we have

$$\|\mathcal{P}_q\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma), \mathrm{H}_0^{-1/2}(\Gamma))} \geq \frac{\|\mathcal{P}_q(f) - \mathcal{P}_q(h)\|_{-1/2,\Gamma}}{\|f - h\|_{-1/2,\Gamma}}.$$

Denoting $\zeta = \|\mathcal{P}_q\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{-1/2}(\Gamma))}$, we may write

$$\|f-h\|_{-1/2,\Gamma} \geq \frac{1}{\zeta}\|\mathcal{P}_q(f)-\mathcal{P}_q(h)\|_{-1/2,\Gamma}$$

so that (2.40) becomes

$$\begin{aligned}\mathrm{F}_B(q,\mathbf{u};h,\mathcal{P}_{q^*}(h)) \geq \left(1+\frac{1}{\zeta^2}\right)\|\mathcal{P}_q(f)-\mathcal{P}_q(h)\|^2_{-1/2,\Gamma} + \|\mathcal{P}_q(h)-\mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}\\ -2\|\mathcal{P}_q(f)-\mathcal{P}_q(h)\|_{-1/2,\Gamma}\|\mathcal{P}_q(h)-\mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma}.\end{aligned}$$

Using an ϵ-inequality, $\forall\,\epsilon>0$ we have

$$\begin{aligned}&\mathrm{F}_B(q,\mathbf{u};h,\mathcal{P}_{q^*}(h))\\ &\quad\geq \left(1+\frac{1}{\zeta^2}-\epsilon\right)\|\mathcal{P}_q(f)-\mathcal{P}_q(h)\|^2_{-1/2,\Gamma} + \left(1-\frac{1}{\epsilon}\right)\|\mathcal{P}_q(h)-\mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}.\end{aligned}$$

Choosing $\epsilon = 1+\frac{1}{\zeta^2}$ leads to

$$\mathrm{F}(q,\mathbf{u};h,\mathcal{P}_{q^*}(h)) \geq \left(\frac{1}{1+\zeta^2}\right)\|\mathcal{P}_q(h)-\mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}. \qquad \square$$

We now have all the necessary tools and notation to establish the main theorem, associated with the lower bound in the equivalence we seek. We write an arbitrary $\mathbf{u}\in(\mathrm{H}^1(\Omega))^2$ as $\mathbf{u}^*+\delta\mathbf{u}$ and consider its projection onto space S_q.

THEOREM 2.13. *Assume that q^* and $\mathbf{u}^*$ are such that $\mathrm{F}(q^*,\mathbf{u}^*;h,\mathcal{P}_{q^*}(h))=0$, and consider $\mathrm{F}(q,\mathbf{u};h,\mathcal{P}_{q^*}(h))$, where q and $\mathbf{u}$ are arbitrary insofar as $q-q^*\in\mathrm{H}_0^1(\Omega)$ and $\mathbf{u}\in\mathrm{W}_q(\Omega)$. Also, let $\delta\mathbf{u}_s$ satisfy*

$$\mathbf{u}^*+\delta\mathbf{u}_s = \arg\min_{\mathbf{u}^*+\delta\mathbf{u}'\in\mathrm{S}_q}\mathrm{F}(q,\mathbf{u}^*+\delta\mathbf{u}';h,\mathcal{P}_{q^*}(h)), \tag{2.41}$$

and define $\delta\mathbf{u}=\mathbf{u}-\mathbf{u}^*$. Then there exists a positive constant, c, depending on Ω and $\|\mathcal{P}_q\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{-1/2}(\Gamma))}$, such that

$$\mathrm{F}(q,\mathbf{u};h,\mathcal{P}_{q^*}(h)) \geq c\left(\|\delta\mathbf{u}-\delta\mathbf{u}_s\|^2_{\mathrm{W}_q(\Omega)} + \|\mathcal{P}_q(h)-\mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}\right).$$

Proof. First, $\mathbf{u}^*+\delta\mathbf{u}_s\in\mathrm{S}_q$ implies that

$$\begin{aligned}\mathrm{F}(q,\mathbf{u}^*+\delta\mathbf{u};h,\mathcal{P}_{q^*}(h)) &= \mathrm{F}(q,\mathbf{u}^*+\delta\mathbf{u}_s+\delta\mathbf{u}-\delta\mathbf{u}_s;h,\mathcal{P}_{q^*}(h))\\ &= \mathrm{F}_I(q,\delta\mathbf{u}-\delta\mathbf{u}_s)+\mathrm{F}_B(q,\mathbf{u}^*+\delta\mathbf{u}_s+\delta\mathbf{u}-\delta\mathbf{u}_s;h,\mathcal{P}_{q^*}(h)).\end{aligned} \tag{2.42}$$

On the boundary, we denote

$$\begin{aligned}\mathbf{n}\cdot(e^q(\mathbf{u}^*+\delta\mathbf{u}_s)) &= h_s,\\ \boldsymbol{\tau}\cdot(e^{-q}(\mathbf{u}^*+\delta\mathbf{u}_s)) &= \mathcal{P}_q(h_s)\end{aligned}$$

and

$$\begin{aligned}\mathbf{n}\cdot(e^q(\delta\mathbf{u}-\delta\mathbf{u}_s)) &= \delta h,\\ \boldsymbol{\tau}\cdot(e^{-q}(\delta\mathbf{u}-\delta\mathbf{u}_s)) &= \delta f.\end{aligned} \tag{2.43}$$

Then (2.42) becomes

$$\begin{aligned}\mathrm{F}(q, \mathbf{u}^* + \delta\mathbf{u}; h, \mathcal{P}_{q^*}(h)) = \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) + \langle h_s - h + \delta h, h_s - h + \delta h\rangle_{-1/2,\Gamma} \\ + \langle \mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h) + \delta f, \mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h) + \delta f\rangle_{-1/2,\Gamma}.\end{aligned} \tag{2.44}$$

Next, expanding the terms in (2.44) leads to

$$\begin{aligned}&\mathrm{F}(q, \mathbf{u}^* + \delta\mathbf{u}; h, \mathcal{P}_{q^*}(h)) \\ &\quad = \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) + \|h_s - h\|^2_{-1/2,\Gamma} + \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma} \\ &\qquad + \|\delta h\|^2_{-1/2,\Gamma} + \|\delta f\|^2_{-1/2,\Gamma} + 2\langle h_s - h, \delta h\rangle_{-1/2,\Gamma} + 2\langle \mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h), \delta f\rangle_{-1/2,\Gamma}.\end{aligned} \tag{2.45}$$

To address the cross terms, we must use the orthogonality statement associated with (2.41). First, note that there exists a $\widehat{\delta\mathbf{u}}$ such that $\{\mathbf{u}^* + \widehat{\delta\mathbf{u}}, h + \delta h, \mathcal{P}_q(h + \delta h)\}$ is an NT triple. We may therefore speak of $\mathcal{P}_q(h + \delta h) = \mathcal{P}_q(h) + \mathcal{P}_q(\delta h)$ and of an orthogonality condition for $\mathbf{u}^* + \delta\mathbf{u}_s$:

$$\langle h_s - h, \delta h\rangle_{-1/2,\Gamma} + \langle \mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h), \mathcal{P}_q(\delta h)\rangle_{-1/2,\Gamma} = 0 \qquad \forall\ \delta h \in \mathrm{H}_0^{-1/2}(\Gamma). \tag{2.46}$$

We can now use (2.46) to add and subtract some helpful terms so that, for any $\epsilon \in (0, 1)$, equation (2.45) becomes

$$\begin{aligned}&\mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \\ &\quad = \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) + \|h_s - h\|^2_{-1/2,\Gamma} + \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma} + \|\delta h\|^2_{-1/2,\Gamma} \\ &\qquad + \|\delta f\|^2_{-1/2,\Gamma} + 2\epsilon\langle \mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h), \delta f - \mathcal{P}_q(\delta h)\rangle_{-1/2,\Gamma} \\ &\qquad + 2(1-\epsilon)\langle h_s - h, \delta h\rangle_{-1/2,\Gamma} + 2(1-\epsilon)\langle \mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h), \delta f\rangle_{-1/2,\Gamma} \\ &\quad \geq \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) + \|h_s - h\|^2_{-1/2,\Gamma} + \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma} + \|\delta h\|^2_{-1/2,\Gamma} \\ &\qquad + \|\delta f\|^2_{-1/2,\Gamma} - 2\epsilon\|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma}\|\delta f - \mathcal{P}_q(\delta h)\|_{-1/2,\Gamma} \\ &\qquad - 2(1-\epsilon)\left|\langle h_s - h, \delta h\rangle_{-1/2,\Gamma}\right| - 2(1-\epsilon)\left|\langle \mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h), \delta f\rangle_{-1/2,\Gamma}\right|.\end{aligned} \tag{2.47}$$

Focusing on the $(1-\epsilon)$ terms, by Cauchy–Schwarz we have

$$\begin{aligned}2\left|\langle h_s - h, \delta h\rangle_{-1/2,\Gamma}\right| &\leq \|h_s - h\|^2_{-1/2,\Gamma} + \|\delta h\|^2_{-1/2,\Gamma}, \\ 2\left|\langle \mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h), \delta f\rangle_{-1/2,\Gamma}\right| &\leq \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma} + \|\delta f\|^2_{-1/2,\Gamma}.\end{aligned}$$

Applying these bounds to (2.47), we have

$$\begin{aligned}&\mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \\ &\quad \geq \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) - 2\epsilon\|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma}\|\delta f - \mathcal{P}_q(\delta h)\|_{-1/2,\Gamma} \\ &\qquad + \epsilon\left(\|h_s - h\|^2_{-1/2,\Gamma} + \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma} + \|\delta h\|^2_{-1/2,\Gamma} + \|\delta f\|^2_{-1/2,\Gamma}\right).\end{aligned} \tag{2.48}$$

Now, as a direct application of Lemma 2.10, there exists a constant c_1 such that

$$\|\delta f - \mathcal{P}_q(\delta h)\|^2_{-1/2,\Gamma} \leq c_1\, \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s). \tag{2.49}$$

To see this, first note that $\exists\,\tilde{\mathbf{u}} \in \mathrm{W}_q(\Omega)$ such that $\{\tilde{\mathbf{u}}, h_s + \delta h, \mathcal{P}_q(h_s) + \mathcal{P}_q(\delta h)\}$ is an NT triple for q. Then observe that if we let $\mathbf{w}_h = \tilde{\mathbf{u}}$ and $\mathbf{w} = \mathbf{u}^* + \delta\mathbf{u}_s + \delta\mathbf{u} - \delta\mathbf{u}_s$, we may write

$$\boldsymbol{\tau} \cdot (e^{-q}(\mathbf{w} - \mathbf{w}_h)) = \delta f - \mathcal{P}_q(\delta h). \tag{2.50}$$

Therefore, applying Lemma 2.10, we have

$$\begin{aligned} \|\delta f - \mathcal{P}_q(\delta h)\|^2_{-1/2,\Gamma} &\le c\,\mathrm{F}_I(q, \mathbf{w} - \mathbf{w}_h) \\ &\le c\,\mathrm{F}_I(q, \mathbf{w}) \\ &\le c_1\,\mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s). \end{aligned} \tag{2.51}$$

As a direct result of (2.51), we may write (2.48) as

$$\begin{aligned} \mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \ge\ & \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) + \epsilon\|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma} \\ & - 2\epsilon c_1 \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma} \left(\mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s)\right)^{1/2} \\ & + \epsilon\left(\|h_s - h\|^2_{-1/2,\Gamma} + \|\delta h\|^2_{-1/2,\Gamma} + \|\delta f\|^2_{-1/2,\Gamma}\right). \end{aligned} \tag{2.52}$$

Next, we use an ϵ-inequality, with η in place of ϵ, to rewrite the first three terms: For any $\eta > 0$,

$$\begin{aligned} -2\epsilon c_1 \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|_{-1/2,\Gamma} (\mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s))^{1/2} & \\ \ge -\epsilon c_1^2 \eta \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) - \frac{\epsilon}{\eta}\|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}, & \end{aligned} \tag{2.53}$$

and (2.52) becomes

$$\begin{aligned} \mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \ge\ & (1 - \epsilon c_1^2\eta)\mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) + \epsilon\left(1 - \frac{1}{\eta}\right)\|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma} \\ & + \epsilon\left(\|h_s - h\|^2_{-1/2,\Gamma} + \|\delta h\|^2_{-1/2,\Gamma} + \|\delta f\|^2_{-1/2,\Gamma}\right). \end{aligned}$$

Choosing $0 < \epsilon < 1$ and $\eta > 1$ so that $(1 - \epsilon c_1^2\eta) = \epsilon(1 - 1/\eta)$ and ϵ is as large as possible, we have

$$\begin{aligned} \mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \ge\ & \epsilon\left(1 - \frac{1}{\eta}\right)\left(\mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) + \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}\right) \\ & + \epsilon\left(\|\delta h\|^2_{-1/2,\Gamma} + \|\delta f\|^2_{-1/2,\Gamma} + \|h_s - h\|^2_{-1/2,\Gamma}\right). \end{aligned} \tag{2.54}$$

Note that this is always possible since $\epsilon = \left(c_1^2\eta + (1 - 1/\eta)\right)^{-1}$. Now we use

$$\mathrm{F}(q, \delta\mathbf{u} - \delta\mathbf{u}_s; 0, 0) = \mathrm{F}_I(q, \delta\mathbf{u} - \delta\mathbf{u}_s) + \|\delta h\|^2_{-1/2,\Gamma} + \|\delta f\|^2_{-1/2,\Gamma}$$

so that we may write

$$\begin{aligned} \mathrm{F}(q, \mathbf{u}; h, & \mathcal{P}_{q^*}(h)) \\ & \ge c_2\left(\mathrm{F}(q, \delta\mathbf{u} - \delta\mathbf{u}_s; 0, 0) + \|h_s - h\|^2_{-1/2,\Gamma} + \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}\right), \end{aligned}$$

where $c_2 := \epsilon(1 - 1/\eta)$. Next, we recall that

$$\begin{aligned} \mathrm{F}(q, \mathbf{u}^* + \delta\mathbf{u}_s; h, \mathcal{P}_{q^*}(h)) &= \mathrm{F}_B(q, \mathbf{u}^* + \delta\mathbf{u}_s; h, \mathcal{P}_{q^*}(h)) \\ &= \|h_s - h\|^2_{-1/2,\Gamma} + \|\mathcal{P}_q(h_s) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}, \end{aligned}$$

since $\mathbf{u}^* + \delta\mathbf{u}_s \in \mathrm{S}_q$. Therefore, we may apply Lemma 2.12 to arrive at

$$\mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \;\geq\; c\left(\mathrm{F}(q, \delta\mathbf{u} - \delta\mathbf{u}_s; 0, 0) + \|\mathcal{P}_q(h) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma}\right).$$

Finally, by Lemma 2.9, we have the desired coercive bound. ∎

We now state an equivalence result for a fixed NT pair, $\{h, \mathcal{P}_{q^*}(h)\}$, which immediately follows from Theorem 2.13 and a simple derivation of the upper bound.

COROLLARY 2.14. *Let $h \in \mathrm{H}_0^{-1/2}(\Gamma)$, and recall that $\sigma = e^{2q}$ and $\sigma^* = e^{2q^*}$. Then there exists a positive constant, c, depending only on Ω and $\|\mathcal{P}_q\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{-1/2}(\Gamma))}$, such that*

$$\begin{aligned} c\|\mathcal{R}_\sigma(h) - \mathcal{R}_{\sigma^*}(h)\|^2_{1/2,\Gamma} & \\ &\leq \min_{\mathbf{u}\,\epsilon\,\mathrm{W}_q(\Omega)} \mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \\ &\leq \|\mathcal{R}_\sigma(h) - \mathcal{R}_{\sigma^*}(h)\|^2_{1/2,\Gamma}. \end{aligned} \tag{2.55}$$

Proof. The lower bound follows immediately from applying Lemma 2.8 to the statement in Theorem 2.13 and performing the minimization in (2.24). To show continuity, we simply restrict the space over which the minimization is conducted and use the definition of S_q to see that

$$\begin{aligned} \min_{\mathbf{u}\,\epsilon\,\mathrm{W}_q(\Omega)} \mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) &\leq \min_{\mathbf{u}\in\mathrm{S}_q} \mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \\ &\leq \min_{\{\mathbf{u}\in\mathrm{S}_q : \mathbf{n}\cdot\mathbf{u} = h\}} \mathrm{F}_B(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h)) \\ &= \|\mathcal{P}_q(h) - \mathcal{P}_{q^*}(h)\|^2_{-1/2,\Gamma} \end{aligned} \tag{2.56}$$

for any $h \in \mathrm{H}_0^{-1/2}(\Gamma)$. Hence, again appealing to Lemma 2.8 to relate $\mathcal{P}$ to $\mathcal{R}$ proves the upper bound in (2.55). ∎

The above corollary immediately leads to the equivalence of the OLS and FOSLS functionals, which we now state formally.

COROLLARY 2.15. *With $\mathcal{F}_{OSLS}$ defined in* (2.24)*, there exists a positive constant, c, depending only on Ω and $\|\mathcal{P}_q\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{-1/2}(\Gamma))}$, such that*

$$\begin{aligned} c\sum_{i=1}^{L} \|\mathcal{R}_\sigma(h_i) - \mathcal{R}_{\sigma^*}(h_i)\|^2_{1/2,\Gamma} & \\ &\leq \mathcal{F}_{OSLS}(q; \{h_i, \mathcal{P}_{q^*}(h_i)\}_{i=1}^{L}) \\ &\leq \sum_{i=1}^{L} \|\mathcal{R}_\sigma(h_i) - \mathcal{R}_{\sigma^*}(h_i)\|^2_{1/2,\Gamma}. \end{aligned} \tag{2.57}$$

As in the previous section, we can also state the equivalence of a sup version of the FOSLS functional and the operator norm of the defect in the NtD maps. For $c > 0$ depending only on Ω and $\|\mathcal{P}_q\|_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{-1/2}(\Gamma))}$, we may write

$$\begin{aligned} c\|\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*}\|^2_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{1/2}(\Gamma))} & \\ &\leq \sup_{h\in\mathrm{H}_0^{-1/2}(\Gamma)} \frac{\left(\min_{\mathbf{u}\,\epsilon\,\mathrm{W}_q(\Omega)} \mathrm{F}(q, \mathbf{u}; h, \mathcal{P}_{q^*}(h))\right)}{\|h\|_{-1/2,\Gamma}} \\ &\leq \|\mathcal{R}_\sigma - \mathcal{R}_{\sigma^*}\|^2_{\mathcal{L}(\mathrm{H}_0^{-1/2}(\Gamma),\mathrm{H}_0^{1/2}(\Gamma))}. \end{aligned}$$

3. Unifying FOSLS framework. Beyond their equivalence, another connection between the KV and FOSLS functionals can be made. Considering a single DN pair, the FOSLS interior functional may be written as

$$\left\| \begin{pmatrix} e^{-q}\nabla \cdot e^{q} \\ e^{q}\nabla \times e^{-q} \end{pmatrix} \mathbf{u} \right\|_{0,\Omega}^{2} := \|\mathcal{L}\mathbf{u}\|_{0,\Omega}^{2}.$$

Alternatively, one can arrive at the KV functional using a generalized Helmholtz decomposition for $\mathbf{u}$, leading to

$$\begin{aligned} \mathbf{u} &= e^{q}\nabla e^{-q} r - e^{-q}\nabla^{\perp} e^{q} t \\ &= \mathcal{L}^{*} \begin{pmatrix} r \\ t \end{pmatrix}, \end{aligned}$$

for some $r, t \in \mathrm{H}^{1}(\Omega)$. The KV functional in this context, for a single DN pair, may be written as

$$\|e^{q}\nabla p - e^{-q}\nabla^{\perp} s\|_{0,\Omega}^{2} = \left\| \mathcal{L}^{*} \begin{pmatrix} e^{q} p \\ e^{-q} s \end{pmatrix} \right\|_{0,\Omega}^{2}$$

and may thus be viewed as a first-order system LL* (FOSLL*) formulation [6]. Indeed, in being adjoint formulations, the $\mathcal{L}$ in FOSLS has overspecified boundary conditions, while the corresponding $\mathcal{L}^{*}$ has no boundary conditions. FOSLL* is in its formative stages as a viable methodology for the solution of PDEs, yet it may prove particularly effective in the forward solution of Maxwell's equations, for which regularity requirements of standard FOSLS are often too stringent. As such, the KV formulation represents an interesting connection between FOSLL* and the inverse problem.

4. Conclusion. We have presented a new FOSLS formulation for the reconstruction of conductivity or, equivalently, resistivity, given knowledge of the corresponding NtD map. Moreover, we have established its equivalence to two existing approaches, including the standard constrained minimization, OLS. Further analysis of many practical concerns is needed before this novel approach moves beyond theory to actually impact imaging technology. Here we have established the common ground from which to conduct such analysis and only begun to explore the role FOSLS can play in determining the sense in which the inverse problem of EIT can be posed well.

REFERENCES

[1] G. Backus and F. Gilbert, *The resolving power of gross earth data*, Geophysical Journal of the Royal Astronomical Society, 266 (1968), pp. 169–205.

[2] M. Berndt, T. Manteuffel, S. McCormick, and G. Starke, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients: Part* I, SIAM J. Numer. Anal., submitted.

[3] L. Borcea, *A nonlinear multigrid for imaging conductivity and permittivity at low frequency*, Inverse Problems, 17 (2001), pp. 329–360.

[4] B. Brown, *Medical impedance tomography and process impedance tomography: A brief review*, Measurement Science and Technology, 12 (2001), pp. 991–996.

[5] Z. Cai, T. A. Manteuffel, and S. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[6] Z. Cai, T. A. Manteuffel, S. McCormick, and J. Ruge, *First-order system $\mathcal{L}\mathcal{L}^{*}$ (FOSLL*): Scalar elliptic partial differential equations*, SIAM J. Numer. Anal., 39 (2001), pp. 1418–1445.

[7] A. Calderon, *On an inverse boundary problem*, in Seminar on Numerical Analysis and Its Applications to Continuum Physics, W. Meyer and M. Raupp, eds., 1980, Brazilian Math. Society, Rio de Janeiro, 1980, pp. 65–73.
[8] M. Cheney, D. Isaacson, and J. C. Newell, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.
[9] D. Dobson and F. Santosa, *An image-enhancement technique for electrical impedance tomography*, Inverse Problems, 10 (1993), pp. 317–334.
[10] *EIT.org*, *Academic Organization for Biomedical Applications*, http://www.eit.org.uk.
[11] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, New York, 1986.
[12] E. Haber and D. Oldenburg, *A gcv based method for nonlinear ill-posed problems*, Comput. Geosci., 4 (2000), pp. 41–63.
[13] E. Haber and L. Tenorio, *Learning regularization functionals: A supervised training approach*, Inverse Problems, 19 (2003), pp. 611–626.
[14] P. Holden, M. Wang, R. Mann, F. Dickin, and R. Edwards, *On detecting mixing pathologies inside a stirred vessel using electrical impedance tomography*, Chemical Engineering Research and Design, 77 (1999), pp. 709–712.
[15] *Industrial Tomography Systems, Inc.*, http://www.itoms.com.
[16] D. Isaacson, *Distinguishability of conductivities by electric current computed tomography*, IEEE Transactions on Medical Imaging, 5 (1986), pp. 91–95.
[17] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, New York, 1997.
[18] J. Kaipio, V. Kolehmainen, M. Vauhkonen, and E. Somersalo, *Inverse problems with structural prior information*, Inverse Problems, 15 (1999), pp. 713–729.
[19] J. P. Kaipio, V. Kolehmainen, E. Somersalo, and M. Vauhkonen, *Statistical inversion methods in electrical impedance tomography*, Inverse Problems, 16 (2000), pp. 1487–1522.
[20] R. Kohn and A. McKenney, *Numerical implementation of a variational method for electrical impedance tomography*, Inverse Problems, 6 (1990), pp. 389–414.
[21] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 37 (1984), pp. 289–298.
[22] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements* II. *Interior results*, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.
[23] R. Kohn and M. Vogelius, *Relaxation of a variational method for impedance computed tomograhy*, Comm. Pure Appl. Math., 40 (1987), pp. 745–777.
[24] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972.
[25] H. MacMillan, S. McCormick, and T. Manteuffel, *First-Order System Least Squares and Electrical Impedance Tomography: Numerical Results*, manuscript.
[26] K. Mosegaard and A. Tarantola, *Probabilistic Approach to Inverse Problems*, Academic Press, San Diego, 2002, pp. 100–180.
[27] A. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. of Math. (2), 143 (1996), pp. 71–96.
[28] G. A. Newman and G. M. Hoversten, *Solution strategies for two- and three-dimensional electromagnetic inverse problems*, Inverse Problems, 16 (2000), pp. 1357–1375.
[29] R. L. Parker, *The inverse problem of electrical conductivity in the mantle*, Geophysical Journal of the Royal Astronomical Society, 22 (1970), pp. 121–138.
[30] R. L. Parker, *Geophysical Inverse Theory*, Princeton Universty Press, Princeton, NJ, 1996.
[31] P. Pinheiro, W. W. Loh, M. Wang, R. Mann, and R. C. Waterfall, *Three-dimensional electrical resistance tomography in a stirred mixing vessel*, Chemical Engineering Communications, 175 (1999), pp. 25–38.
[32] A. Seagar, T. Yeo, and R. Bates, *Full-wave computed tomography: Resolution limits*, IEE Proceedings, Pt. A, Physical Science and Measurement, 131 (1984), pp. 616–622.
[33] A. D. Seagar, T. S. Yeo, and R. H. Bates, *Full-wave computed tomography: Low frequency electric current*, IEE Proceedings, Pt. A, Physical Science and Measurement, 132 (1985), pp. 455–466.
[34] H. Storz, W. Storz, and F. Jacobs, *Electrical resistivity tomography to investigate geological structures of the earth's upper crust*, Geophysical Prospecting, 48 (2000), pp. 455–471.
[35] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.
[36] A. Tikhonov and V. Ya, *Methods for Solving Ill-Posed Problems*, John Wiley and Sons, Des Moines, IA, 1977.
[37] A. Tripp, E. Cherkaeva, and J. Hulen, *Bounds on the complex conductivity of geophysical*

mixtures, Geophysical Prospecting, 46 (1998), pp. 589–601.

[38] M. Vauhkonen, J. P. Kaipio, E. Somersalo, and P. A. Karjalainen, *Electrical impedance tomography with basis constraints*, Inverse Problems, 13 (1997), pp. 523–530.

[39] M. Vauhkonen, D. Vadasz, P. Karjalainen, E. Somersalo, and J. Kaipio, *Tikhonov regularization and prior information in electrical impedance tomography*, IEEE Transactions on Medical Imaging, 17 (1998), pp. 197–219.

[40] P. Vauhkonen, M. Vauhkonen, T. Savolainen, and J. Kaipio, *Three-dimensional electrical impedance tomography on the complete electrode model*, IEEE Transactions on Biomedical Engineering, 46 (1999), pp. 1150–1160.

[41] A. Wexler, B. Fry, and M. Neumann, *Impedance-computed tomography algorithm and system*, Applied Optics, 24 (1985), pp. 3985–3992.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 484–504

AN EXPLICIT A PRIORI ESTIMATE FOR A FINITE VOLUME APPROXIMATION OF LINEAR ADVECTION ON NON-CARTESIAN GRIDS*

BRUNO DESPRÉS†

Abstract. We propose an elementary proof of strong convergence for a finite volume approximation of nonstationary linear advection on arbitrary grids in two space dimensions. This proof is elementary in the sense that the basic a priori estimate uses only some discrete integrations by parts and the Cauchy–Schwarz inequality. Numerical results show that the estimate is probably nonoptimal.

Key words. nonstationary linear advection, finite volume schemes, optimal rate of convergence

AMS subject classifications. 65M12, 65M15

DOI. 10.1137/S0036142901394558

1. Introduction.

Let us consider the following nonstationary linear advection model problem:

$$\begin{cases} \partial_t u + \vec{a}.\vec{\nabla} u = 0, & (t,\vec{x}) \in [0,T] \times \Omega, \\ u(t=0,\vec{x}) = u_0(\vec{x}), & \vec{x} \in \partial\Omega. \end{cases} \tag{1.1}$$

For the sake of simplicity we consider the two-dimensional case:

$$\vec{x} = (x_1, x_2) \in \Omega = [0,1] \times [0,1] \subset \mathbb{R}^2; \tag{1.2}$$

assume that $\vec{a} \neq 0$ is constant in space and time, and supplement (1.1) with periodic boundary conditions

$$\begin{cases} u(t,0,x_2) = u(t,1,x_2), & (t,x_2) \in [0,T] \times [0,1], \\ u(t,x_1,0) = u(t,x_1,1), & (t,x_1) \in [0,T] \times [0,1]. \end{cases} \tag{1.3}$$

Taking in account the periodicity of boundary conditions the exact solution is

$$u(t,\vec{x}) = u_0(\vec{x} - \vec{a}t). \tag{1.4}$$

In this work we address the standard finite volume numerical approximation of (1.1). Even if finite volume methods in conjunction with high order methods are widely used in practice for industrial problems (see [20, 12, 11, 4, 3, 21, 30] and the references therein), mathematical open problems still exist around these techniques. This has been reported, for instance, in [15] and [16]: in particular, explicit and simple estimates of the error between the exact solution of (1.1) and various lowest order finite volume approximations are not so simple to obtain. What we really intend to show in this work is that a complete understanding of the accuracy of nonstationary linear advection on non-Cartesian grids is still an open problem.

*Received by the editors August 29, 2001; accepted for publication (in revised form) August 22, 2003; published electronically March 11, 2004.
http://www.siam.org/journals/sinum/42-2/39455.html

†Commissariat à l'Énergie Atomique, 91680, Bruyères-le-Châtel BP 12, France (despres@bruyeres.cea.fr) and Laboratoire Jacques Louis Lions, 175 rue du Chevaleret, Université de Paris VI, 75013 Paris, France (despres@ann.jussieu.fr).

In their seminal work [24], Lesaint and Raviart address L^2-based approaches for transport with absorption such as

$$\partial_t u + \vec{a}.\vec{\nabla} u + cu - d\Delta u = 0, \quad \text{with } c > 0 \quad \text{and } d = 0. \tag{1.5}$$

Other L^2-based approaches with either diffusion or absorption (i.e., $c > 0$ or $d > 0$) may be found in [7, 8, 17, 18, 28] in the context of discontinuous Galerkin methods and also in [23] in the context of the Friedrichs systems.

Proof of convergence and estimate of convergence for finite volume methods applied to nonlinear problems is a subject that has already been considered by many authors. A nonexhaustive list of references is as follows: [31], based on [14], for a use of the measure value solution technique to give the first proof of convergence for transport on general meshes; [2], which uses the kinetic approach [22, 14]; [6, 15] about the Kuznetsov error approach [19]. Other references may be found in [15] and [16]. See also [13] for an approach with nonlinear schemes. Most proofs of convergence on arbitrary grids (at least all proofs we are aware of) are based on some mathematical notions which were designed mainly to tackle nonlinear problems or to tackle coercive-diffusive transport. The one proposed here, L^2-based, gives an explicit bound of the error in function of the solution for the case $c = d = 0$. To our knowledge, all L^2-based estimates given in [24, 7, 8] and [28] blow up when $c \to 0^+$ or $d \to 0^+$.

The central part of this paper addresses a new L^2-based a priori estimate for the numerical error concerning the problem (1.1) (i.e., (1.5) with $c = d = 0$): see inequality (4.21) of Theorem 2. The interest of this estimate is that it gives an explicit bound of the numerical error in terms of the solution only:

$$\|u^n - \mathbf{u}^n\|_2 \leq F(u(n\Delta t), u((n-1)\Delta t), \ldots, u(0)), \tag{1.6}$$

where $u(., j\Delta t)$ is the solution at $T = j\Delta t$, and F is some functional. A consequence is a proof of convergence for a standard upwind finite volume approximation of (1.1) on a regular (we will make this clearer in the following) arbitrary grid, due to

$$F(u(n\Delta t), u((n-1)\Delta t), \ldots, u(0)) \leq C(T, ||\nabla u_0||_{L^2}, ||\nabla^2 u_0||_{L^2})\Delta x^{\frac{1}{2}}, \tag{1.7}$$

where we assume that $u_0 \in H^2(\Omega)$ and that the mesh is a general triangular or quadrangular mesh: u^n is the numerical solution and $\mathbf{u}^n = \frac{1}{s_j}\int_{\Omega_j} u(n\Delta t, \vec{x})dx$ is the cell-averaged exact solution in cell Ω_j. A corollary is convergence of the upwind finite volume scheme for linear advection on regular grids in L^1 and L^2.

Estimate (1.7) displays a fractional order of convergence $C\Delta x^{\frac{1}{2}}$, even for a H^2 solution. It has to be compared with the standard estimate [29] of convergence $C\Delta x$, true for the numerical approximation on a Cartesian grid of the smooth solution of (1.1) (actually numerical results clearly indicate that the numerical rate of convergence is $C\Delta x$, even for non-Cartesian grids). It is still an open question to decide whether the fractional order of convergence $C\Delta x^{\frac{1}{2}}$ is optimal on arbitrary non-Cartesian grids or not. The reason why we obtain only a fractional order of convergence in our proof is that finite volume methods on general meshes are nonconsistent in the finite difference sense. To deal with the formal nonconsistency of finite volume techniques, which is here the real difficulty, we introduce an approximate and global cell-averaged solution $\mathbf{v}^n \approx \mathbf{u}^n$ and prove various but basic a priori estimates about this approximate solution. The interest of this approximate solution is that it is consistent, while the projection of the exact solution is nonconsistent. These a priori estimates give the $C\Delta x^{\frac{1}{2}}$ order of convergence. An important difference with previous works about the

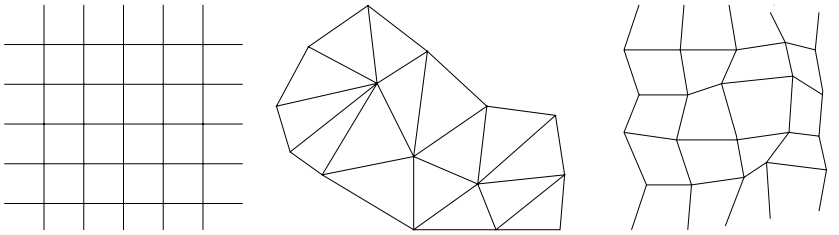


FIG. 1. *Some meshes.*

L^2 analysis (see, for example, [24, 26], in connection with analysis of discontinuous Galerkin methods) is that the keystone of our analysis is the construction of this approximate solution, which helps to give a direct analysis of the scheme in terms of the truncation error: actually we prove that the approximate solution is close to the exact solution and is consistent; the analysis of [26] is for the stationary equation, never explains anything about the consistency discrepancy of the scheme, is in some sense global, and does not help to get the error at time T, which is what one is really interested in for nonstationary problems.

The paper is organized as follows. In section 2, we introduce some notation and define the standard finite volume approximation of (1.1). Then, in section 3, we recall the basic definition of the stability of the finite volume discretization. In section 4, we introduce new material, define what we call the approximate solution $\mathbf{v}$, and prove various a priori estimates for this $\mathbf{v}$. Then, in section 5, we prove the convergence in L^2 of the finite volume scheme on arbitrary grids with a rate $C\Delta x^{\frac{1}{2}}$. In section 6, we give a counterexample of why $C\Delta x^{\frac{1}{2}}$ proofs of convergence via total variation bounded estimates are not possible. Finally, we give some numerical results in section 7.

2. Notation. From now on, we consider only the discretization of (1.1). Let $(\Omega_j)_{j\in[1,..,J]}$ be a finite mesh of Ω:

$$\left\{ \begin{array}{ll} \Omega_j \cap \Omega_k = \emptyset & \forall j,k, \quad j \neq k, \\ \bigcup\limits_{j\in[1,..,J]} \bar{\Omega}_j = \bar{\Omega} = \Omega. & \end{array} \right. \tag{2.1}$$

J is the number of cells. The shape of any cell is arbitrary. Most usual cases are square cells (finite difference) and triangle or distorted quadrangular cells (finite volume); see Figure 1. Other meshes can be constructed; see Figure 5. Two cells are neighboring cells if and only if they have an edge in common (taking into account periodic boundary conditions). Each cell has a finite number of neighbors: $I(j)$ is the set of the neighbors of cell j. The outgoing normal from Ω_j on the edge $\partial\Omega_j \cap \partial\Omega_k$ is denoted as $\vec{n}_{jk}$. Of course, the outgoing normal from Ω_j is the opposite of the outgoing normal from Ω_k for $k \in I(j)$:

$$\vec{n}_{jk} + \vec{n}_{kj} = 0. \tag{2.2}$$

We introduce some very natural notation:

$$\left\{ \begin{array}{ll} l_{jk} = l_{kj} = \mathbb{R}\text{-Lebesgue measure of } \partial\Omega_j \cap \partial\Omega_k, & \text{a length}, \\ s_j = \mathbb{R}^2\text{-Lebesgue measure of } \Omega_j, & \text{a surface}. \end{array} \right. \tag{2.3}$$

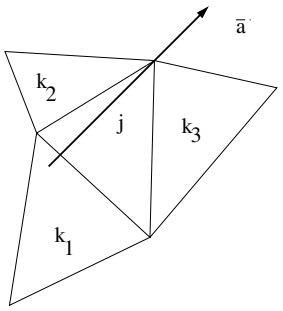


FIG. 2. $I^+(j) = \{k_2, k_3\}$, $I^-(j) = \{k_1\}$.

We also define

$$\left\{ \begin{array}{l} I^+(j) = \{k \in I(j); (\vec{a}, \vec{n}_{jk}) > 0\}, \\ I^0(j) = \{k \in I(j); (\vec{a}, \vec{n}_{jk}) = 0\}, \\ I^-(j) = \{k \in I(j); (\vec{a}, \vec{n}_{jk}) < 0\} \end{array} \right. \tag{2.4}$$

and

$$m_{jk} = m_{kj} = l_{jk} |(\vec{a}, \vec{n}_{jk})|. \tag{2.5}$$

$(.,.)$ denotes the standard scalar product. $I^+(j)$ (resp., $I^-(j)$) is the set of outgoing (resp., incoming) cells from Ω_j. An example with a triangular mesh is given in Figure 2. With all this notation the standard upwind finite volume-like method together with a constant mass initial condition is (2.6)–(2.7):

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + \sum_{k \in I^+(j)} m_{jk} u_j^n - \sum_{k \in I^-(j)} m_{jk} u_k^n = 0 \quad \forall j \in [1, ., J], \quad \forall n \geq 0, \tag{2.6}$$

$$u_j^0 = \frac{1}{s_j} \int_{\Omega_j} u_0(x) \ \ \forall j \in [1, ., J]. \tag{2.7}$$

The following formula will play an important role in the analysis.

LEMMA 1. *One has the equality*

$$\sum_{k \in I^+(j)} m_{jk} = \sum_{k \in I^-(j)} m_{jk} \ \ \forall j. \tag{2.8}$$

It is a well-known consequence of the divergence theorem:

$$0 = \int_{\Omega_j} \mathbf{div}\ \vec{a} = \int_{\partial \Omega_j} (\vec{a}, \vec{n}_{jk}) = \sum_{k \in I^+(j)} m_{jk} - \sum_{k \in I^-(j)} m_{jk}.$$

Taking into account periodic boundary conditions, the scheme is conservative since

$$\sum_j s_j u_j^{n+1} = \sum_j \left(s_j u_j^n - \Delta t \sum_{k \in I^+(j)} m_{jk} u_j^n - \sum_{k \in I^-(j)} m_{jk} u_k^n \right)$$

$$= \sum_j s_j u_j^n + \left(-\Delta t \sum_j \sum_{k \in I^+(j)} m_{jk} u_j^n + \Delta t \sum_j \sum_{k \in I^-(j)} m_{jk} u_k^n \right)$$
$$= \sum_j s_j u_j^n.$$

Since (2.8) proves that $\sum_j \sum_{k \in I^-(j)} m_{jk} u_k^n = \sum_k [\sum_{j \in I^+(k)} m_{jk}] u_k^n$, then the term between parentheses in the right-hand side of the above expression is zero. It shows the conservativity of the scheme.

3. Stability. In order to investigate the convergence of the scheme we equip the space $\mathbb{R}^J$ with either the standard L^2, L^1, or L^∞ norm.

DEFINITION 1. *The standard L^2, L^1, and L^∞ discrete norms are*

$$\|\mathbf{w}\|_2 = \left(\sum_j s_j (\mathbf{w}_j)^2 \right)^{\frac{1}{2}}, \quad \|\mathbf{w}\|_1 = \sum_j s_j |\mathbf{w}_j|, \quad \|\mathbf{w}\|_\infty = \max_j (|\mathbf{w}_j|). \tag{3.1}$$

Now we consider $\mathbf{w} = (\mathbf{w}_j)_{j \in [1,.,J]}$ an arbitrary vector $\mathbf{w} \in \mathbb{R}^J$ and define the iteration operator.

DEFINITION 2. *The linear iteration operator $A : \mathbb{R}^J \to \mathbb{R}^J$ extracted from* (2.6) *is defined by*

$$(A\mathbf{w})_j = \mathbf{w}_j - \frac{\Delta t}{s_j} \left(\sum_{k \in I^+(j)} m_{jk} \mathbf{w}_j - \sum_{k \in I^-(j)} m_{jk} \mathbf{w}_k \right) \quad \forall j \in [1,.,J]. \tag{3.2}$$

This iteration operator is stable in the following sense.

LEMMA 2. *Let us assume the CFL condition*

$$\frac{\sum_{k \in I^+(j)} m_{jk}}{s_j} \Delta t \leq 1 \quad \forall j \in [1,.,J]. \tag{3.3}$$

Then

$$\|A\mathbf{w}\|_2 \leq \|\mathbf{w}\|_2, \quad \|A\mathbf{w}\|_1 \leq \|\mathbf{w}\|_1, \quad \|A\mathbf{w}\|_\infty \leq \|\mathbf{w}\|_\infty. \tag{3.4}$$

The proof is completely standard.

A problem is that the scheme (2.6) is not consistent, at least in the finite difference sense and for general grids in several space dimensions: this has been reported, for instance, in [16, 2, 15, 6] and the references therein. To understand more precisely this property of nonconsistency, we consider the exact solution of (1.1), $u(t, \vec{x})$, and define $\mathbf{u}_j^n$ to be the mean value of the exact solution:

$$\mathbf{u}_j^n = \frac{1}{s_j} \int_{\Omega_j} u(n\Delta t, \vec{x}) dx \quad \forall j, n. \tag{3.5}$$

DEFINITION 3. *The truncation error $\mathbf{R}^n = (\mathbf{R}_j^n) \in \mathbb{R}^J$ is*

$$\mathbf{R}_j^n = \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t} + \frac{1}{s_j} \left(\sum_{k \in I^+(j)} m_{jk} \mathbf{u}_j^n - \sum_{k \in I^-(j)} m_{jk} \mathbf{u}_k^n \right). \tag{3.6}$$

The following result is referred to as the Lax theorem in the literature. We state it using the L^2 norm, but it is possible to use other norms as the L^1 or L^∞ norm.

THEOREM 1 (Lax theorem). *Assuming the CFL condition* (3.3)*, the numerical error* $\|u^n - \mathbf{u}^n\|_2$ *is bounded:*

$$\|u^n - \mathbf{u}^n\|_2 \leq \Delta t \sum_{p=0}^{n-1} \|\mathbf{R}^p\|_2. \tag{3.7}$$

The numerical error is $\mathbf{e}^n = u^n - \mathbf{u}^n$: using the definition of the scheme, the definition of the iteration operator A and the definition of the truncation error $\mathbf{R}$, one has

$$\mathbf{e}^{n+1} = A\mathbf{e}^n - \Delta t \mathbf{R}^n.$$

Due to the CFL condition we have $\|\mathbf{e}^{n+1}\|_2 \leq \|\mathbf{e}^n\|_2 + \Delta t \|\mathbf{R}^n\|_2$, which implies

$$\|\mathbf{e}^n\|_2 \leq \|\mathbf{e}^0\|_2 + \Delta t \sum_{p=0}^{n-1} \|\mathbf{R}^p\|_2.$$

Since $\mathbf{e}^0 = u^0 - \mathbf{u}^0 = 0$ by definition of u^0 (2.7), it gives (3.7). Assuming enough regularity of the initial data, and assuming that the mesh is Cartesian, simple Taylor expansions prove that

$$\|\mathbf{R}^n\|_2 = O(\Delta x, \Delta t). \tag{3.8}$$

Due to the CFL inequality (3.3), equation (3.8) implies $\|\mathbf{R}^n\|_2 = O(\Delta x)$.

However, the mesh might not have such a structure (a general triangular mesh with no particular symmetry is an example; see Figure 1). Then (3.9) is replaced by

$$\|\mathbf{R}^n\|_2 = O(1). \tag{3.9}$$

Using such a general mesh there is no way for $\mathbf{R}_j^n$ to be bounded like (3.8). The reason why we have only (3.9) is that

$$\frac{1}{s_j}\left(\sum_{k\in I^+(j)} m_{jk}\mathbf{u}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{u}_k^n\right) = \frac{1}{s_j}\int_{\Omega_j} \vec{a}.\nabla u(n\Delta t, \vec{x})dx + O(1) \tag{3.10}$$

on arbitrary grids for the exact solution. As a consequence of this lack of consistency (3.9), it is not possible to rely on the Lax theorem (3.7) to prove the convergence of the scheme.

4. The approximate solution. In order to circumvent this problem we consider that the difficulty comes from the numerical approximation of the gradient of the solution, which is expressed in (3.10). First we remark that the left-hand side of (3.10) may be rewritten as (4.1):

$$\begin{aligned}\frac{1}{s_j}&\left(\sum_{k\in I^+(j)} m_{jk}\mathbf{u}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{u}_k^n\right)\\ &= \frac{1}{s_j}\left(\sum_{k\in I^+(j)} m_{jk}\mathbf{u}_{jk}^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{u}_{jk}^n\right) + O(1),\end{aligned} \tag{4.1}$$

where $\mathbf{u}_{jk}^n$ is the mean value of the exact solution on the edges:

$$\mathbf{u}_{jk}^n = \mathbf{u}_{kj}^n = \frac{1}{l_{jk}} \int_{\partial\Omega_j \cap \partial\Omega_k} u(n\Delta t, \vec{x}) d\sigma \quad \forall j,k,n. \tag{4.2}$$

Second, it suggests that we may define a new approximate solution, denoted as $\mathbf{v}_j^n$, in order to get rid of this consistency problem.

DEFINITION 4. *Let $\sigma > 0$. Assuming that u is C^0 (so that (4.2) makes sense), the approximate solution $\mathbf{v}^n \in \mathbb{R}^J$ is defined as the solution of the following linear system:*

$$\begin{aligned} &\sigma s_j(\mathbf{v}_j^n - \mathbf{u}_j^n) + \sum_{k\in I^+(j)} m_{jk}\mathbf{v}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{v}_k^n \\ &\quad = \sum_{k\in I^+(j)} m_{jk}\mathbf{u}_{jk}^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{u}_{jk}^n \quad \forall j \in [1,.,J], \quad n \geq 0. \end{aligned} \tag{4.3}$$

Here σ is a kind of penalization parameter. If we take $\sigma = 0$, then the numerical flux of the approximate solution satisfies

$$\frac{1}{s_j}\left(\sum_{k\in I^+(j)} m_{jk}\mathbf{v}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{v}_k^n\right) = \frac{1}{s_j}\left(\sum_{k\in I^+(j)} m_{jk}\mathbf{u}_{jk}^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{u}_{jk}^n\right),$$

which is a better approximation of $\vec{a}.\nabla u(n\Delta t, \vec{x})$ than (4.1). For technical reasons which will appear clearly in (4.5), it is better to consider $\sigma > 0$. This parameter will be set to the optimal value $\sigma \approx 1$ at the end of the proof. In complete rigor, we should write this approximate solution as $\mathbf{v} = \mathbf{v}(u, \sigma)$. However, we prefer to save notation, still working with the simplified notation $\mathbf{v}$. Let us point to an originality of the approximate solution: $\mathbf{v}$ is defined globally; that is, $\mathbf{v}_j^n$ is a function of $\mathbf{u}_i^n$ and $\mathbf{u}_{ik}^n \; \forall$ $i \in [1,.,J]$. To our knowledge, such a global definition has rarely been considered before. It is a simple exercise to check that

$$\sum_j s_j \mathbf{v}_j^n = \sum_j s_j \mathbf{u}_j^n. \tag{4.4}$$

The terminology *approximate solution* is justified, at least by the fact that $\sigma = +\infty$ implies $\mathbf{v} = \mathbf{u}$. The interest of definition (4.3) relies principally on the following a priori estimate.

LEMMA 3. *Let $\mathbf{v}^n \in \mathbb{R}^J$ be a solution of (4.3). Then*

(a) *one has the a priori estimate*

$$\sigma \|\mathbf{v}^n - \mathbf{u}^n\|_2^2 \leq \frac{1}{2}\sum_j \sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_j^n)^2; \tag{4.5}$$

(b) *$\forall \sigma > 0$ the linear system (4.3) is invertible; that is, the approximate solution $\mathbf{v}^n \in \mathbb{R}^J$ is well defined and unique;*

(c) *the discrete derivative in time is bounded:*

$$\begin{aligned} &\sigma \left\| \frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{\Delta t} - \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} \right\|_2^2 \\ &\quad \leq \frac{1}{2}\sum_j \sum_{k\in I^+(j)} m_{jk}\left(\frac{\mathbf{u}_{jk}^{n+1} - \mathbf{u}_{jk}^n}{\Delta t} - \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t}\right)^2. \end{aligned} \tag{4.6}$$

In the next section, we prove that the right-hand sides of (4.5) and (4.6) are bounded by $C\Delta x$ for a smooth solution u. So (4.5) and (4.6) prove that $\mathbf{v}^n$ is indeed an approximate numerical solution of (1.1) at time step n. We split the proof of this lemma in three steps.

(a) The a priori estimate (4.5) is actually the key to the approach developed in this paper. The proof is simple and uses basic discrete integrations by parts and the Cauchy–Schwarz inequality. We define $\mathbf{w}_j^n = \mathbf{v}_j^n - \mathbf{u}_j^n$ and rewrite (4.3) as

$$\begin{aligned}(4.7)\qquad & \sigma s_j \mathbf{w}_j^n + \sum_{k\in I^+(j)} m_{jk}\mathbf{w}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{w}_k^n \\ & = \sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_j^n) - \sum_{k\in I^-(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_k^n) \ \ \forall j \in [1,.,J].\end{aligned}$$

So $\mathbf{w}^n$ is a solution of the nonhomogeneous system.

We multiply by $\mathbf{w}_j^n$ and sum with respect to j. Thus

$$\begin{aligned}(4.8)\qquad & \sigma\|\mathbf{w}^n\|_2^2 + \sum_j \left(\sum_{k\in I^+(j)} m_{jk}\mathbf{w}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{w}_k^n \right) \mathbf{w}_j^n \\ & = \sum_j \left(\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_j^n) - \sum_{k\in I^-(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_k^n) \right) \mathbf{w}_j^n.\end{aligned}$$

We use Lemma 1 to get

$$\begin{aligned}(4.9)\qquad & \left(\sum_{k\in I^+(j)} m_{jk}\mathbf{w}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{w}_k^n \right) \mathbf{w}_j^n \\ & = \sum_{k\in I^-(j)} m_{jk} \left(\frac{1}{2}(\mathbf{w}_j^n)^2 - \frac{1}{2}(\mathbf{w}_k^n)^2 + \frac{1}{2}(\mathbf{w}_j^n - \mathbf{w}_k^n)^2 \right).\end{aligned}$$

So we deduce from (4.9) that

$$\begin{aligned}(4.10)\qquad & \sigma\|\mathbf{w}^n\|_2^2 + \frac{1}{2}\sum_j \left(\sum_{k\in I^-(j)} m_{jk}(\mathbf{w}_j^n - \mathbf{w}_k^n)^2 \right) \\ & + \frac{1}{2}\sum_j \left(\sum_{k\in I^-(j)} m_{jk} \right) (\mathbf{w}_j^n)^2 - \frac{1}{2}\sum_j \left(\sum_{k\in I^-(j)} m_{jk}(\mathbf{w}_k^n)^2 \right) \\ & = \sum_j \left(\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_j^n) - \sum_{k\in I^-(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_k^n) \right) \mathbf{w}_j^n.\end{aligned}$$

We reorganize the last term in the left-hand side and get

$$(4.11)\qquad \sum_j \sum_{k\in I^-(j)} m_{jk}\mathbf{w}_k^2 = \sum_j \left(\sum_{k\in I^+(j)} m_{jk} \right) \mathbf{w}_j^2.$$

Thus one has

$$\sum_{j\in[1,.,J]}\left(\sum_{k\in I^-(j)} m_{jk}(\mathbf{w}_k^n)^2\right)=\sum_{k\in[1,.,J]}\left(\sum_{j;k\in I^-(j)} m_{jk}\right)(\mathbf{w}_k^n)^2 \tag{4.12}$$
$$=\sum_{k\in[1,.,J]}\left(\sum_{j;j\in I^+(k)} m_{kj}\right)(\mathbf{w}_k^n)^2=\sum_{j\in[1,.,J]}\left(\sum_{k;k\in I^-(j)} m_{jk}\right)(\mathbf{w}_j^n)^2.$$

So we rewrite (4.8) as

$$\sigma\|\mathbf{w}^n\|_2^2+\frac{1}{2}\sum_j\left(\sum_{k\in I^-(j)} m_{jk}(\mathbf{w}_j^n-\mathbf{w}_k^n)^2\right)$$
$$=\sum_j\left(\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n-\mathbf{u}_j^n)-\sum_{k\in I^-(j)} m_{jk}(\mathbf{u}_{jk}^n-\mathbf{u}_k^n)\right)\mathbf{w}_j^n.$$

Using a discrete integration by part we transform the right-hand side:

$$\sigma\|\mathbf{w}^n\|_2^2+\frac{1}{2}\sum_j\left(\sum_{k\in I^-(j)} m_{jk}(\mathbf{w}_j^n-\mathbf{w}_k^n)^2\right) \tag{4.13}$$
$$=\sum_j\left(\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n-\mathbf{u}_j^n)(\mathbf{w}_j^n-\mathbf{w}_k^n)\right).$$

The Cauchy–Schwarz inequality applied to this right-hand side gives

$$\left|\sum_j\left(\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n-\mathbf{u}_j^n)(\mathbf{w}_j^n-\mathbf{w}_k^n)\right)\right|$$
$$\le\left(\sum_j\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n-\mathbf{u}_j^n)^2\right)^{\frac{1}{2}}\times\left(\sum_j\sum_{k\in I^+(j)} m_{jk}(\mathbf{w}_j^n-\mathbf{w}_k^n)^2\right)^{\frac{1}{2}}$$
$$\le\frac{1}{2}\sum_j\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n-\mathbf{u}_j^n)^2+\frac{1}{2}\sum_j\left(\sum_{k\in I^+(j)} m_{jk}(\mathbf{w}_j^n-\mathbf{w}_k^n)^2\right)$$
$$\le\frac{1}{2}\sum_j\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n-\mathbf{u}_j^n)^2+\frac{1}{2}\sum_j\left(\sum_{k\in I^-(j)} m_{jk}(\mathbf{w}_j^n-\mathbf{w}_k^n)^2\right).$$

Together with (4.13) it turns into $\sigma\|\mathbf{w}^n\|_2^2\le\frac{1}{2}\sum_j\sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n-\mathbf{u}_j^n)^2$, which gives point (a) of the lemma.

(b) To prove the well posedness of the linear system (4.3) we consider the homogeneous linear system:

$$\sigma s_j\mathbf{v}_j+\sum_{k\in I^+(j)} m_{jk}\mathbf{v}_j-\sum_{k\in I^-(j)} m_{jk}\mathbf{v}_k=0\ \ \forall j\in[1,.,J].$$

Inequality (4.5) applied to this homogeneous linear system gives

$$\sigma\|\mathbf{v}\|_2^2 + \frac{1}{2}\sum_j \left(\sum_{k\in I^-(j)} m_{jk}(\mathbf{v}_j - \mathbf{v}_k)^2 \right) = 0. \tag{4.14}$$

Using now $\sigma > 0$ we obtain $\mathbf{v} = 0$. It proves the well posedness of the definition of $\mathbf{v}$.

(c) Finally, we turn to (4.6). Let us define

$$\mathbf{z}^n = \frac{\mathbf{w}^{n+1} - \mathbf{w}^n}{\Delta t}.$$

This definition is equivalent to

$$\mathbf{z}_j^n = \frac{\mathbf{w}_j^{n+1} - \mathbf{w}_j^n}{\Delta t} = \frac{\mathbf{v}_j^{n+1} - \mathbf{v}_j^n}{\Delta t} - \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t} \quad \forall j \in [1,.,J].$$

Combining (4.7) with the same definition at time step $^{n+1}$, we obtain

$$\begin{aligned} &\sigma s_j \mathbf{z}_j^n + \sum_{k\in I^+(j)} m_{jk}\mathbf{z}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{z}_k^n \\ &\quad = \sum_{k\in I^+(j)} m_{jk}\left(\frac{\mathbf{u}_{jk}^{n+1} - \mathbf{u}_{jk}^n}{\Delta t} - \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t} \right) \\ &\quad\quad - \sum_{k\in I^-(j)} m_{jk}\left(\frac{\mathbf{u}_{jk}^{n+1} - \mathbf{u}_{jk}^n}{\Delta t} - \frac{\mathbf{u}_k^{n+1} - \mathbf{u}_k^n}{\Delta t} \right) \quad \forall j \in [1,.,J]. \end{aligned} \tag{4.15}$$

This equality has exactly the same structure as (4.8). Thus we obtain a very similar inequality:

$$\sigma\|\mathbf{z}^n\|_2^2 \le \frac{1}{2}\sum_j \left(\sum_{k\in I^+(j)} m_{jk}\left(\frac{\mathbf{u}_{jk}^{n+1} - \mathbf{u}_{jk}^n}{\Delta t} - \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t} \right)^2 \right),$$

which is exactly (4.6).

DEFINITION 5. *We define the truncation error* $\mathbf{S}^n = (\mathbf{S}_j^n) \in \mathbb{R}^J$,

$$\mathbf{S}_j^n = \frac{\mathbf{v}_j^{n+1} - \mathbf{v}_j^n}{\Delta t} + \frac{1}{s_j}\left(\sum_{k\in I^+(j)} m_{jk}\mathbf{v}_j^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{v}_k^n \right), \quad n \ge 0, \tag{4.16}$$

and the truncation error $\mathbf{T}^n = (\mathbf{T}_j^n) \in \mathbb{R}^J$,

$$\mathbf{T}_j^n = \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t} + \frac{1}{s_j}\left(\sum_{k\in I^+(j)} m_{jk}\mathbf{u}_{jk}^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{u}_{jk}^n \right), \quad n \ge 0. \tag{4.17}$$

LEMMA 4. *The truncation error* $\mathbf{S}^n$ *is bounded by*

$$\|\mathbf{S}^n\|_2 \le \sigma\|\mathbf{v}^n - \mathbf{u}^n\|_2 + \left\| \frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{\Delta t} - \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} \right\|_2 + \|\mathbf{T}^n\|_2 \quad \forall n \ge 0. \tag{4.18}$$

The truncation error is $\mathbf{S}^n = a^n + b^n + c^n$, where

$$a_j^n = \frac{\mathbf{v}_j^{n+1} - \mathbf{v}_j^n}{\Delta t} - \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t}, \quad b_j^n = -\sigma(\mathbf{v}_j^n - \mathbf{u}_j^n),$$

and due to the definition of $\mathbf{v}$ in (4.3),

$$c_j^n = \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t} + \frac{1}{s_j}\left(\sum_{k\in I^+(j)} m_{jk}\mathbf{u}_{jk}^n - \sum_{k\in I^-(j)} m_{jk}\mathbf{u}_{jk}^n\right) = \mathbf{T}_j^n.$$

Thus $\|\mathbf{S}^n\|_2 \le \|a^n\|_2 + \|b^n\|_2 + \|c^n\|_2$, which is exactly (4.18).

It remains to express some bounds between the numerical solution and the approximate or exact solution. Inequality (4.20) has to be compared with (3.7): the interest is that even if the right-hand side of (3.7) is $O(1)$ on arbitrary grids, the right-hand side of (4.20) will be proved to be $O(\Delta x^{\frac{1}{2}})$ on arbitrary grids at least when the exact solution is H^2.

THEOREM 2. *Assuming the CFL condition* (3.3)*, one has some a priori estimates:*

$$\|u^n - \mathbf{v}^n\|_2 \le \|\mathbf{u}^0 - \mathbf{v}^0\|_2 + \Delta t \sum_{p=0}^{n-1} \|\mathbf{S}^p\|_2 \quad \forall n \ge 0 \tag{4.19}$$

and

$$\begin{aligned}\|u^n - \mathbf{u}^n\|_2 \le{}& \|\mathbf{v}^n - \mathbf{u}^n\|_2 + \|\mathbf{u}^0 - \mathbf{v}^0\|_2 + \sigma\Delta t \sum_{p=0}^{n-1} \|\mathbf{v}^p - \mathbf{u}^p\|_2 \\ &+ \Delta t \sum_{p=0}^{n-1} \left\| \frac{\mathbf{v}^{p+1} - \mathbf{v}^p}{\Delta t} - \frac{\mathbf{u}^{p+1} - \mathbf{u}^p}{\Delta t} \right\|_2 + \Delta t \sum_{p=0}^{n-1} \|\mathbf{T}^p\|_2 \quad \forall n \ge 0,\end{aligned} \tag{4.20}$$

which implies

$$\begin{aligned}\|u^n - \mathbf{u}^n\|_2 \le{}& \left(\frac{1}{2\sigma}\sum_j \sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_j^n)^2\right)^{\frac{1}{2}} \\ &+ \left(\frac{1}{2\sigma}\sum_j \sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^0 - \mathbf{u}_j^0)^2\right)^{\frac{1}{2}} \\ &+ \sigma\Delta t \sum_{p=0}^{n-1} \left(\frac{1}{2\sigma}\sum_j \sum_{k\in I^+(j)} m_{jk}(\mathbf{u}_{jk}^p - \mathbf{u}_j^p)^2\right)^{\frac{1}{2}} \\ &+ \Delta t \sum_{p=0}^{n-1} \left(\frac{1}{2\sigma}\sum_j \sum_{k\in I^+(j)} m_{jk}\left(\frac{\mathbf{u}_{jk}^{p+1} - \mathbf{u}_{jk}^p}{\Delta t} - \frac{\mathbf{u}_j^{p+1} - \mathbf{u}_j^p}{\Delta t}\right)^2\right)^{\frac{1}{2}} \\ &+ \Delta t \sum_{p=0}^{n-1} \|\mathbf{T}^p\|_2 \quad \forall n \ge 0,\end{aligned} \tag{4.21}$$

where T^p is explicitly given in (4.17) *in terms of the solution u. The interest of inequality* (4.21) *is that the numerical error is explicitly bounded by a discrete quantity which is a function of the solution u only.*

To prove (4.19) we consider the numerical error between the numerical solution u^n and the approximate solution $\mathbf{v}^n$: $\mathbf{f}^n = u^n - \mathbf{v}^n$. Using the definition of the scheme, the definition of the iteration operator A and the definition of the truncation error $\mathbf{S}$, one has $\mathbf{f}^{n+1} = A\mathbf{f}^n - \Delta t \mathbf{S}^n$. One uses the CFL condition $\|\mathbf{f}^{n+1}\|_2 \leq \|\mathbf{f}^n\|_2 + \Delta t \|\mathbf{S}^n\|_2$ which proves (4.19). Next, the triangular inequality gives

$$\|u^n - \mathbf{u}^n\|_2 \leq \|u^n - \mathbf{v}^n\|_2 + \|\mathbf{v}^n - \mathbf{u}^n\|_2 \leq \|\mathbf{u}^0 - \mathbf{v}^0\|_2 + \Delta t \sum_{p=0}^{p=n-1} \|\mathbf{S}^p\|_2 + \|\mathbf{v}^n - \mathbf{u}^n\|_2.$$

Using (4.18), (4.5), and (4.6) we obtain (4.20) and (4.21).

5. Convergence. We need some regularity assumptions on the mesh. In the following we assume that the mesh is triangular, but this is only for the sake of simplicity; many other meshes can be considered as in Figures 1 and 5. We also assume that there exists two constants $c_1 > 0$ and $c_2 > 0$ such that

$$c_1 \Delta x^2 \leq s_j \leq c_2 \Delta x^2, \tag{5.1}$$

where Δx is a characteristic length of the mesh. A well-known consequence of (5.1) is

$$\exists c_3 = c_3(c_2) > 0, \quad l_{jk} \leq c_3 \Delta x \quad \forall j, k \in [1, ., J]. \tag{5.2}$$

DEFINITION 6. *A sequence of triangular meshes such that $\Delta x \to 0$ and such that* (5.1) *is true with uniform c_1 and c_2 is called a uniformly regular sequence of triangular meshes. Since $c_3 = c_3(c_2)$, a consequence of this property of uniform regularity is that c_3 is also uniform. We refer the reader to* [5] *for more definitions and properties about such uniformly regular meshes.*

Using the above material, it is easy to get an explicit bound of the numerical error. This bound is a theorem of convergence.

THEOREM 3. *We assume that the initial condition u_0 of* (1.1) *is $H^2(\Omega)$, so the solution u is $H^2([0,T] \times \Omega)$. Consider the numerical solution given by* (2.6)–(2.7) *on a sequence of triangular uniformly regular meshes. Then $\exists C > 0$ such that $\forall T > 0$, $\forall n \leq \frac{T}{\Delta t}$, and $\forall \Delta x \leq 1$,*

$$\|u^n - \mathbf{u}^n\|_2 \leq \left(C(T+2) \max(||\nabla u_0||_{L^2}, ||\nabla^2 u_0||_{L^2})\right) \Delta x^{\frac{1}{2}}, \tag{5.3}$$

where $C = C(c_1, c_2)$ is a constant which depends only on the parameters of the mesh.

The proof is, of course, based on inequalities (4.20), (4.5), and (4.6). First, we need to estimate $\|\mathbf{v}^n - \mathbf{u}^n\|_2$. Due to (4.5), one has

$$\|\mathbf{v}^n - \mathbf{u}^n\|_2^2 \leq \frac{1}{2\sigma} \sum_j \sum_{k \in I^+(j)} m_{jk} (\mathbf{u}_{jk}^n - \mathbf{u}_j^n)^2. \tag{5.4}$$

We split the rest of the proof in five steps. The first step is very classical in the framework of finite elements and is given here for the sake of the completeness of this work.

Step 1: *Study of* $|\mathbf{u}_{jk}^n - \mathbf{u}_j^n|$**.** In the next we will drop the superscript ^{n} and use $|\mathbf{u}_{jk} - \mathbf{u}_j|$ instead of $|\mathbf{u}_{jk}^n - \mathbf{u}_j^n|$.

We define the affine mapping of the plane $F_j : \mathbb{R}^2 \to \mathbb{R}^2$ such that a given triangle (denoted as Ω_j) is transformed into the reference triangle (denoted as T); see Figure 3. This mapping is well known in the theory of finite elements [5].

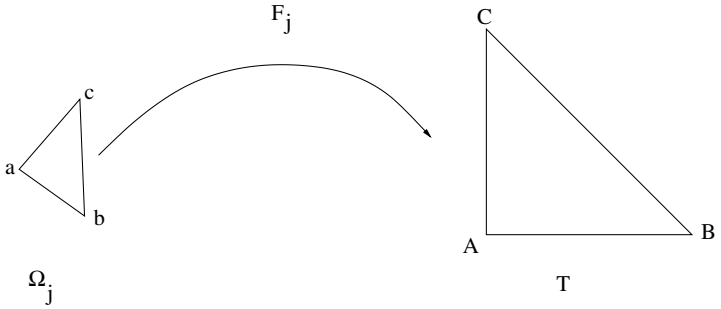


FIG. 3. *The mapping F_j.*

The coordinates of all these points are $a^j = (x_a^j, y_a^j)$, $b^j = (x_b^j, y_b^j)$, and $c^j = (x_c^j, y_c^j)$, together with $A = (0,0)$, $B = (1,0)$, and $C = (0,1)$. The mapping is $(\bar{x}, \bar{y}) = F_j(x,y)$:

$$\left\{ \begin{array}{l} \bar{x} = \alpha^j (x - x_a^j) + \beta^j (y - y_a^j), \\ \bar{y} = \gamma^j (x - x_a^j) + \delta^j (y - y_a^j), \end{array} \right. \tag{5.5}$$

with $D^j = (x_b^j - x_a^j)(y_c^j - y_a^j) - (x_c^j - x_a^j)(y_b^j - y_a^j)$, $\alpha^j = \frac{y_c^j - y_a^j}{D^j}$, $\beta^j = -\frac{x_c^j - x_a^j}{D^j}$, $\gamma^j = -\frac{y_b^j - y_a^j}{D^j}$, and $\delta^j = \frac{x_b^j - x_a^j}{D^j}$. The inverse mapping is defined by

$$\left\{ \begin{array}{l} x = \bar{\alpha}^j \bar{x} + \bar{\beta}^j \bar{y}, \\ y = \bar{\gamma}^j \bar{x} + \bar{\delta}^j \bar{y}, \end{array} \right. \tag{5.6}$$

where $\bar{\alpha}^j = x_b^j - x_a^j$, $\bar{\beta}^j = x_c^j - x_a^j$, $\bar{\gamma}^j = y_b^j - y_a^j$, and $\bar{\delta}^j = y_c^j - y_a^j$. Note that $|D^j| = \text{meas}(\Omega_j) = s_j$. Due to (5.1)

$$|D^j| \geq c_1 \Delta x^2. \tag{5.7}$$

Next, we define $u^j(\bar{x}, \bar{y}) = u(F_j^{-1}(x,y))$. Assuming now that $\mathbf{u}_{jk}$ (resp., $\mathbf{u}_j$) is the edge- (resp., cell-) averaged value of u on (a^j, b^j) (resp., Ω_j), we get that

$$\mathbf{u}_{jk} - \mathbf{u}_j = \int_{(A,B)} u^j(\bar{x}, \bar{y}) d\bar{x} - \int_T u^j(\bar{x}, \bar{y}) d\bar{x}\, d\bar{y}. \tag{5.8}$$

It is useful to introduce the reference square and to extend the function u^j by symmetry with respect to the (B,C) edge of the reference triangle T; see Figure 4. So (5.8) turns into

$$\mathbf{u}_{jk} - \mathbf{u}_j = \int_{(A,B)} u^j(\bar{x}, \bar{y}) d\bar{x} - \int_S u^j(\bar{x}, \bar{y}) d\bar{x}\, d\bar{y}. \tag{5.9}$$

It is then an easy matter to get an upper estimate of the right-hand side. It proves

$$\exists C > 0, \quad |\mathbf{u}_{jk} - \mathbf{u}_j|^2 \leq C ||\nabla u||^2_{L^2(\Omega_j)} \ \forall j, k. \tag{5.10}$$

This inequality is uniform with respect to the index of the triangle j and uniform with respect to the mesh size Δx.

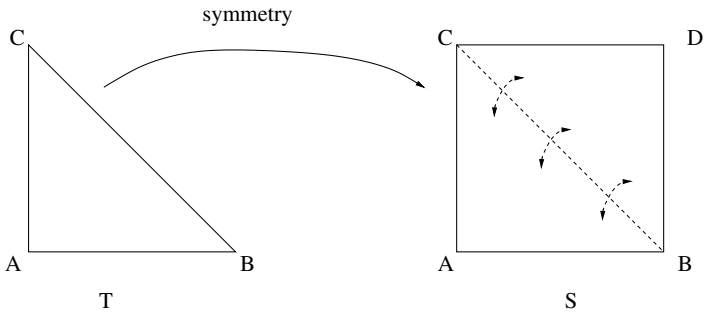


FIG. 4. *Symmetry.*

Step 2: *Bounds for* $\|\mathbf{v}^n - \mathbf{u}^n\|_2$**.** Since $m_{jk} = l_{jk}|(\vec{a}, \vec{n}_{jk})| \leq c_3|\vec{a}|\Delta x$, we get

$$\sum_j \sum_{k \in I^+(j)} m_{jk}(\mathbf{u}_{jk}^n - \mathbf{u}_j^n)^2 \leq (c_3|\vec{a}|\Delta x) \sum_j \left(3C \int_{\Omega_j} |\nabla u|^2 \right)$$
$$\leq (c_3|\vec{a}|3C\Delta x)||\nabla u||^2_{L^2(\Omega)}.$$

Here 3 is the maximal number of edges per triangle. So

$$\|\mathbf{v}^n - \mathbf{u}^n\|_2 \leq C_1||\nabla u||_{L^2(\Omega)} \left(\frac{\Delta x}{\sigma} \right)^{\frac{1}{2}}, \tag{5.11}$$

where the constant C_1 is a function only of (c_1, c_2, c_3).

Step 3: *Bounds for* $\|\frac{\mathbf{v}^{n+1}-\mathbf{v}^n}{\Delta t} - \frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}\|_2$**.** Of course, we use (4.6). With respect to what has already been done for (5.11), the extra work concerns only the evaluation of

$$\frac{\mathbf{u}_{jk}^{n+1} - \mathbf{u}_{jk}^n}{\Delta t} - \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t}. \tag{5.12}$$

Defining $w(n\Delta t) = \frac{u((n+1)\Delta) - u(n\Delta t)}{\Delta t}$, we remark that all terms in (5.12) may be reinterpreted as averages of w:

$$\frac{\mathbf{u}_{jk}^{n+1} - \mathbf{u}_{jk}^n}{\Delta t} - \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t} = \mathbf{w}_{jk}^n - \mathbf{w}_j^n, \tag{5.13}$$

where $\mathbf{w}_{jk}^n$ (resp., $\mathbf{w}_j^n$) is the edge (resp., cell) average of $w(n\Delta t)$. So if we prove that $\nabla w(n\Delta t)$ is bounded in $L^2(\Omega)$, then it is sufficient to use Step 2, applied to this $w(n\Delta t)$ function. Since $w(n\Delta t) = \int_{n\Delta t}^{(n+1)\Delta t} \frac{\partial u}{\partial t} dt = -\vec{a}(\int_{n\Delta t}^{(n+1)\Delta t} \nabla u dt)$, then $\nabla w(n\Delta t) = -(\int_{n\Delta t}^{(n+1)\Delta t} \nabla^2 u dt)\vec{a}$, so $\exists \tilde{c} > 0$ such that

$$||\nabla w(n\Delta t)||_{L^2(\Omega)} \leq \tilde{c}||\nabla^2 u(n\Delta t)||_{L^2(\Omega)} = \tilde{c}||\nabla^2 u_0||_{L^2(\Omega)}.$$

So we have

$$\left\| \frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{\Delta t} - \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} \right\|_2 \leq C_2||\nabla^2 u_0||_{L^2(\Omega)} \left(\frac{\Delta x}{\sigma} \right)^{\frac{1}{2}}, \tag{5.14}$$

where the constant $C_2 > 0$ is a function only of (c_1, c_2, c_3).

Step 4: *Estimate of* $\mathbf{T}^n$**.** Concerning $\mathbf{T}^n$, we have

$$\mathbf{T}_j^n = \frac{\mathbf{u}_j^{n+1} - \mathbf{u}_j^n}{\Delta t} + \frac{1}{s_j}\left(\sum_{k \in I^+(j)} m_{jk} \mathbf{u}_{jk}^n - \sum_{k \in I^-(j)} m_{jk} \mathbf{u}_{jk}^n \right).$$

By definition of $\mathbf{u}_{jk}^n$ in (4.2)

$$\begin{aligned}
\mathbf{T}_j^n &= \frac{1}{s_j} \int_{\Omega_j} \frac{u((n+1)\Delta t, \vec{x}) - u(n\Delta t, \vec{x})}{\Delta t} + \frac{1}{s_j} \int_{\Omega_j} \vec{a} \nabla u(n\Delta t, \vec{x}) \\
&= \frac{1}{s_j} \int_{\Omega_j} \int_{n\Delta t}^{(n+1)\Delta t} (\partial_t u(s, \vec{x}) - \partial_t u(n\Delta t, \vec{x})) ds\, dx \\
&= \frac{1}{\Delta t} \frac{1}{s_j} \int_{\Omega_j} \int_{n\Delta t}^{(n+1)\Delta t} \int_{n\Delta t}^{s} \partial_t^2 u(\tau, \vec{x}) d\tau\, ds\, dx.
\end{aligned}$$

Using the Cauchy–Schwarz inequality we get that

$$\begin{aligned}
|\mathbf{T}_j^n| &\le \frac{1}{\Delta t} \frac{1}{s_j} \left(\int_{n\Delta t}^{(n+1)\Delta t} \int_{n\Delta t}^{s} \int_{\Omega_j} \partial_t^2 u(\tau, \vec{x})^2 dx\, d\tau\, ds \right)^{\frac{1}{2}} \\
&\quad \times \left(\int_{n\Delta t}^{(n+1)\Delta t} \int_{n\Delta t}^{s} \int_{\Omega_j} dx\, d\tau\, ds \right)^{\frac{1}{2}} \\
&\le \frac{1}{s_j^{\frac{1}{2}}} \left(\int_{n\Delta t}^{(n+1)\Delta t} \int_{n\Delta t}^{s} \int_{\Omega_j} (\partial_t^2 u(\tau, \vec{x}))^2 dx\, d\tau\, ds \right)^{\frac{1}{2}}.
\end{aligned}$$

So finally we obtain the estimate

$$\|\mathbf{T}^n\|_2^2 = \sum_j s_j |\mathbf{T}_j^n|^2 \le \int_{n\Delta t}^{(n+1)\Delta t} \int_{n\Delta t}^{s} \left(\int_{\Omega} (\partial_t^2 u(\tau, \vec{x}))^2 dx \right) d\tau\, ds.$$

As a consequence $\exists \tilde{C}_3 > 0$ and $\exists C_3 > 0$ such that

$$\|\mathbf{T}^n\|_2^2 \le \tilde{C}_3 \Delta t^2 ||\partial_t^2 u||_{L^2(\Omega)}^2 \le C_3^2 \Delta x^2 ||\nabla^2 u_0||_{L^2(\Omega)}^2,$$

that is,

$$\|\mathbf{T}^n\|_2 \le C_3 \Delta x ||\nabla^2 u_0||_{L^2(\Omega)}. \tag{5.15}$$

Step 5: *End of the proof.* Finally, we use (5.11)–(5.15) in (4.20) and get

$$\begin{aligned}
\|u^n - \mathbf{u}^n\|_2 &\le (2 + \sigma n \Delta t) C_1 ||\nabla u_0|| \left(\frac{\Delta x}{\sigma} \right)^{\frac{1}{2}} \\
&\quad + (n\Delta t) C_2 ||\nabla^2 u_0|| \left(\frac{\Delta x}{\sigma} \right)^{\frac{1}{2}} + (n\Delta t) C_3 ||\nabla^2 u_0|| \Delta x.
\end{aligned}$$

The coefficient $2 + \sigma n \Delta t \le 2 + \sigma T$ is simply the multiplicative factor, equal to the number of terms like $\|\mathbf{v}^p - \mathbf{u}^p\|$ in the sum. Finally, we optimize the parameter σ and choose $\sigma = 1$. Be careful that we really need to optimize σ since it appears both in

the numerator $\sigma\frac{1}{\sigma^{\frac{1}{2}}}\times\ldots$ and in the denominator $\frac{1}{\sigma^{\frac{1}{2}}}\times\ldots$. Since $\Delta x\le 1$ we bound $\Delta x\le\Delta x^{\frac{1}{2}}$. It gives (5.3) and finishes the proof.

COROLLARY 1. *We assume that the initial condition u_0 of (1.1) is in $L^1(\Omega)$ (resp., $L^2(\Omega)$), so the solution u is $L^1([0,T]\times\Omega)$ (resp., $L^2([0,T]\times\Omega)$). Consider the numerical solution given by (2.6)–(2.7) on a sequence of triangular uniformly regular meshes, and assume a CFL inequality (3.3). Then $\forall T>0$ and $\forall n\le\frac{T}{\Delta t}$,*

$$u^n\to\mathbf{u}^n \text{ as } \Delta x\to 0 \text{ in } L^1(\Omega) \text{ (resp., } L^2(\Omega)). \tag{5.16}$$

This is straightforward and is a mere consequence of the stability of the iteration operator for the L^1 and L^2 norms.

6. A remark about convergence via bounded variation estimates. Nonlinear techniques applied to the study of the convergence of the scheme have been used in various works. Let us mention [4, 3, 15, 6], where a $\Delta x^{\frac{1}{4}}$ rate of convergence in L^1 for a initial data in $L^1\cap BV$ is proved; see also [25]. In [9, 10] it is proved that the scheme converges with a $\Delta x^{\frac{1}{2}}$ rate in L^1 for a initial data in $L^1\cap BV$, provided the numerical solution is bounded in $L^\infty([0,T]\times BV)$. Many numerical experiments show that this hypothesis—the numerical solution is bounded in $L^\infty([0,T]\times BV)$—is very reasonable. Unfortunately, it is false without additional hypotheses, as is shown in the proof of the next lemma.

LEMMA 5. *There exist meshes such that the numerical solution is not bounded in $L^\infty([0,T]\times BV)$ (with a initial data in $L^1\cap BV$, of course).*

Actually the total variation of these numerical solutions blows up like $\Delta x^{-\frac{1}{2}}$. The construction is very simple. Let us consider the almost Cartesian mesh represented in Figure 5. The mesh is made of squares, but the size of each square is either Δx or $\frac{\Delta x}{2}$. On each "line" all squares have the same size. So we distinguish all cells on "large" lines, referred to as $(2p,j)$ cells, and all cells on "small" line, referred to as $(2p+1,j)$ cells. Thus $2p$ (resp., $2p+1$) is the index of the line, while j is the index of the cell inside its line. Just be careful that there are twice as many cells on a $2p+1$ line than on a $2p$ line. The index of the column is $j=0$ for the first square on the positive half-plane $x>0$. Here P is the number of "large" lines: provided there are exactly the same number of large and small lines, one has

$$P=\frac{1}{\frac{3\Delta x}{2}}=\frac{2}{3\Delta x}. \tag{6.1}$$

Let us solve the problem with periodic boundary conditions:

$$\begin{cases}\partial_t u+\partial_x u=0, & -1\le x,\quad y\le 1,\\ u_0(x,y)=0 \text{ for } x<\frac{1}{2}, & u_0(x,y)=1 \text{ for } x>\frac{1}{2}.\end{cases} \tag{6.2}$$

This is a one-dimensional problem on each line. But since the size of each square is not the same, the CFL number is not the same on lines made with Δx "large" squares and lines made with $\frac{\Delta x}{2}$ "small" squares. Let us consider for the sake of simplicity that the CFL number is exactly one for the "small" squares $\Delta t=\frac{\Delta x}{2}$. It means that the scheme is the one-dimensional exact linear scheme on "small" lines,

$$u^{n+1}_{2p+1,j}=u^n_{2p+1,j-1},\quad u^0_{2p,j}=1 \quad\text{for}\quad j\ge 0,\quad u^0_{2p,j}=0 \quad\text{for}\quad j<0, \tag{6.3}$$

and is the one-dimensional nonexact linear scheme on "large" lines,

$$u^{n+1}_{2p,j}=\frac{1}{2}(u^n_{2p,j}+u^n_{2p,j-1}),\quad u^0_{2p,j}=1 \quad\text{for}\quad j\ge 0,\quad u^0_{2p,j}=0 \quad\text{for}\quad j<0. \tag{6.4}$$

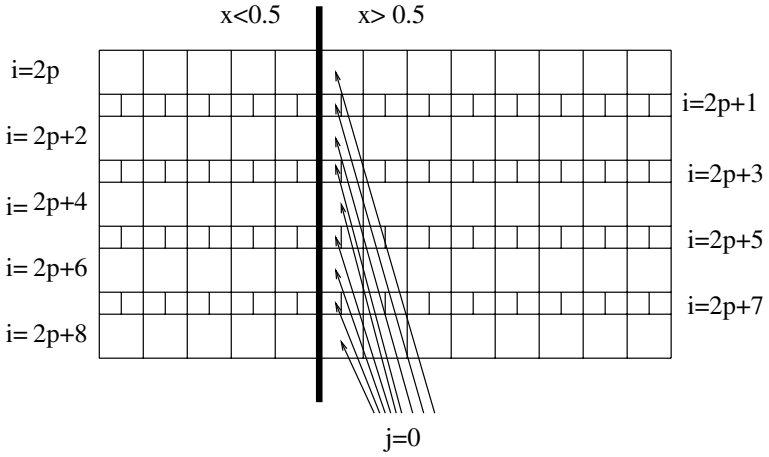


FIG. 5. *An almost Cartesian mesh.*

The solution of (6.3) is, of course, the exact solution $u^n_{2p+1,j} = u^0_{2p+1,j-n}$. It is a classroom exercise to check that the solution of (6.4) is $u^n_{2p,j} = \frac{1}{2^n}\sum_{q\le j}\binom{q}{n}$, where $\binom{q}{n} = \frac{n!}{q!(n-q)!}$ is the binomial coefficient: $\binom{q}{n} = 0$ for $q < 0$ and $q > n$. The L^1 difference between the exact the the numerical solution on "large" lines is

$$\sum_j |u^{2k}_{2p,j} - u_{\text{exact}}(2k\Delta t)| = \sum_{0\le j\le k-1} |u^{2k}_{2p,j}| + \sum_{k\le j\le 2k-1} |u^{2k}_{2p,j} - 1|.$$

Let us get a lower bound for $\sum_{0\le j\le k-1} |u^{2k}_{2p,j}|$. One has

$$\begin{aligned}\sum_{0\le j\le k-1} |u^{2k}_{2p,j}| &= \frac{1}{2^{2k}} \sum_{0\le j\le k-1}\sum_{q\le j}\binom{q}{2k} \\ &= \frac{1}{2^{2k}} \sum_{0\le q\le k-1} (k-q)\binom{q}{2k} = A - B.\end{aligned} \tag{6.5}$$

The right-hand side is the difference between

$$A = \frac{1}{2^{2k}}\frac{k}{2}\left(2\sum_{0\le q\le k-1}\binom{q}{2k}\right) = \frac{1}{2^{2k}}\frac{k}{2}\left(2^{2k} - \binom{k}{2k}\right) = \frac{k}{2}\left(1 - \frac{\binom{k}{2k}}{2^{2k}}\right)$$

and

$$\begin{aligned}B &= \frac{1}{2^{2k}}\sum_{0\le q\le k-1} q\binom{q}{2k} = \frac{k}{2^{2k}}\left(2\sum_{q\le k-1}\binom{q-1}{2k-1}\right) \\ &= \frac{k}{2^{2k}}\left(2^{2k-1} - 2\binom{k-1}{2k-1}\right) = \frac{k}{2}\left(1 - 4\frac{\binom{k-1}{2k-1}}{2^{2k}}\right) \\ &= \frac{k}{2}\left(1 - 2\frac{\binom{k}{2k}}{2^{2k}}\right).\end{aligned}$$

Thus

$$A - B = \frac{k}{2}\frac{\binom{k}{2k}}{2^{2k}} = \frac{\sqrt{k}}{2\sqrt{2\pi}} + o(\sqrt{k}) \tag{6.6}$$

from the standard Stirling approximation of the factorial $q! \approx \sqrt{2\pi q}\left(\frac{q}{e}\right)^q$; see [1]. Now going back to (6.5) we get that

$$\Delta x \sum_j |u^{2k}_{2p,j} - u_{\text{exact}}(2k\Delta t)| \geq \Delta x \left(\frac{\sqrt{k}}{2\sqrt{2\pi}} + o(\sqrt{k})\right).$$

But $2k$ is the iteration number and corresponds to the time $T = 2k\Delta t = k\Delta x$. Then $\Delta x\sqrt{k} = \sqrt{\Delta x}\sqrt{\Delta x k} = \sqrt{\Delta x}\sqrt{T}$. Then we get that $\Delta x \sum_j |u^{2k}_{2p,j} - u_{\text{exact}}(2k\Delta t)| \geq C\Delta x^{\frac{1}{2}} + o(\Delta x^{\frac{1}{2}})$, $C = \frac{\sqrt{T}}{2\sqrt{2\pi}}$. Let us now get a lower bound of the total variation in two dimensions of the numerical solution. Since the two-dimensional total variation is by definition the sum on all edges of the absolute value of the difference of the numerical solution on both sides, multiplied by the length of the edge, then we get that

$$\begin{aligned} |u^{2k}|_{BV} &\geq 2\Delta x \sum_{p=1}^{P} \left(\sum_j |u^{2k}_{2p,j} - u_{\text{exact}}(2k\Delta t)|\right) \\ &= P\frac{\sqrt{T}}{\sqrt{2\pi}}(\Delta x^{\frac{1}{2}} + o(\Delta x^{\frac{1}{2}})). \end{aligned} \tag{6.7}$$

Then we get from (6.1) that

$$|u^{2k}|_{BV} \geq \frac{2\sqrt{T}}{3\sqrt{2\pi}}\Delta x^{-\frac{1}{2}} + o(\Delta x^{-\frac{1}{2}}). \tag{6.8}$$

Here $T = 2k\Delta t$. It ends the proof of the lemma.

7. Numerical order of convergence. However, very simple numerical experiments show that the $\Delta x^{\frac{1}{2}}$ order of convergence proved in this work for twice differentiable initial data is probably not optimal. The initial data is

$$u_0(x, y) = \sin(2\pi(x + y)) \in H^2_{\text{per}}([0, 1]^2).$$

Indeed, the experimental order of convergence is clearly ≈ 1 $\forall L^p$ norms, $p = 1,\ 2,\ \infty$.

In Figure 6 is plotted the type of mesh one usually gets with a standard mesh generator. Here we have used the Freefem++ mesh generator [27], but it is very reasonable to think that similar conclusions can be drawn with other mesh generators. In Table 1 we give the experimental relative error which is defined by

$$e_p = \frac{||u(n\Delta t) - u^{n\Delta t}||_p}{||u(n\Delta t)||_p}, \tag{7.1}$$

where $u(n\Delta t)$ (resp., $u^{n\Delta t}$) is the exact (resp., numerical) solution at time $n\Delta t$. Here we took $n\Delta t = 1$. The equation is (1.1) with $\vec{a} = (1, 1)$. There are many possibilities to define the characteristic length of the mesh. We use

$$h = \max(\text{edge's length}). \tag{7.2}$$

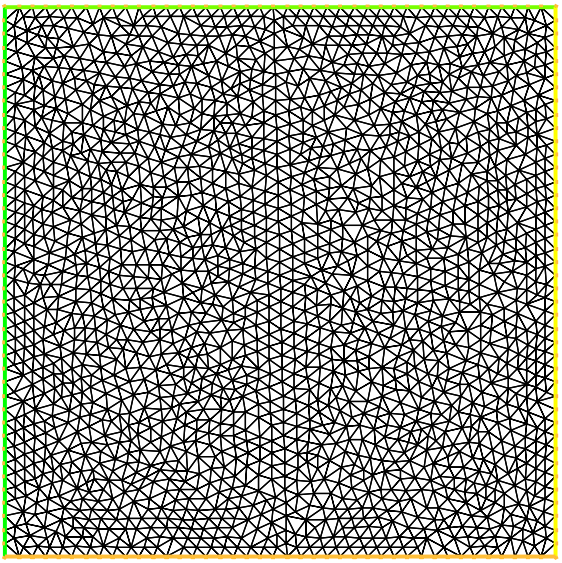


FIG. 6. *From Freefem++.*

TABLE 1
Errors with respect to the size of the mesh.

Triangular mesh	Points	Cells	h	L^1	L^2	L^∞
	2034	3902	0.0408	0.3967	0.3967	0.4102
	4488	8730	0.0276	0.2828	0.2828	0.2933
	7873	15420	0.0216	0.2173	0.2179	0.2308
	17756	35026	0.0147	0.1554	0.1553	0.1675
	31232	61818	0.0114	0.1207	0.1206	0.1290
Order				0.9939	0.9946	1.027

Cartesian mesh	Points	Cells	h	L^1	L^2	L^∞
	40^2	40^2	0.025	0.3964	0.3973	0.3961
	60^2	60^2	0.01666	0.2849	0.2852	0.2848
	80^2	80^2	0.0125	0.2187	0.2187	0.2187
	120^2	120^2	0.00833	0.1538	0.1539	0.1538
	160^2	160^2	0.00625	0.1160	0.1160	0.1160
Order				0.9818	0.9840	0.981

By inspection of the table one finds that the rate of convergence is very close to one. We use the approximate formula for the order of convergence:

$$\text{order}_{\text{triangle}} = \frac{\log \frac{e_p(0.0147)}{e_p(0.00114)}}{\log \frac{0.0147}{0.00114}} \quad \text{and} \quad \text{order}_{\text{square}} = \frac{\log \frac{e_p(0.025)}{e_p(0.00625)}}{\log \frac{0.025}{0.00625}}. \tag{7.3}$$

In each case, the formula is a function of the error for the two finest meshes.

8. Conclusion. In this paper we give an elementary proof of convergence for linear advection on arbitrary grids in several space dimensions, based on an explicit and

new a priori estimate. It is possible to take into account other boundary conditions. To our knowledge it is still an open problem to explain the one order of convergence for the linear advection discretized with finite volume methods on non-Cartesian meshes.

Acknowledgment. The author thanks both referees for their help in the improvement of the quality of this paper.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1966.
[2] R. Botchorishvili, B. Perthame, and A. Vasseur, *Schémas d'équilibre pour des lois de conservation scalaires avec des termes sources raides*, Report 3891, INRIA, France, 2000.
[3] C. Chainais-Hillairet, *First and second order schemes for a hyperbolic equation: Convergence and error estimate*, in Finite Volume for Complex Applications Problems and Perspectives, F. Benkhaldoun and R. Vilsmeier, eds., Hermes, Paris, 1997, pp. 137–144.
[4] S. Champier, T. Gallouët, and R. Herbin, *Convergence of an upstream finite volume scheme for a nonlinear hyperbolic equation on a triangular mesh*, Numer. Math., 66 (1993), pp. 139–157.
[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[6] B. Cockburn, F. Coquel, and P. LeFloch, *An error estimate for finite volume multidimensional conservation laws*, Math. Comp., 63 (1994), pp. 77–103.
[7] B. Cockburn and C. Dawson, *Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multi-dimensions*, in Proceedings of the Conference on the Mathematics of Finite Elements and Applications: MAFELAP X, J. R. Whiteman, ed., Elsevier, Oxford, 2000, pp. 225–238.
[8] B. Cockburn, *Devising discontinuous Galerkin methods for non-linear hyperbolic conservation laws*, J. Comput. Appl. Math., 128 (2001), pp. 187–204.
[9] B. Cockburn and P.-A. Gremaud, *Error estimates for finite element methods for scalar conservation laws*, SIAM J. Numer. Anal., 33 (1996), pp. 522–554.
[10] B. Cockburn, P.-A. Gremaud, and J. X. Yang, *A priori error estimates for numerical methods for scalar conservation laws Part* III: *Multidimensional flux-splitting monotone schemes on non-Cartesian grids*, SIAM J. Numer. Anal., 35 (1998), pp. 1775–1803.
[11] P. Collella, *Multidimensional upwind methods for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 171–200.
[12] H. Deconinck, R. Struijs, and G. Bourgeois, *Compact Advection Schemes on Unstructured Grids*, in VKI Lecture Series 1993–04 on CFD, von Karman Institute, Rhode-Saint-Genèse, Belgium, 1993.
[13] B. Després and F. Lagoutière, *Generalized Harten formalism and longitudinal variation diminishing schemes for linear advection on arbitrary grids*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 1159–1183.
[14] R. J. DiPerna, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223–270.
[15] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 713–1020.
[16] E. Godlewski and P. A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996.
[17] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.
[18] P. Houston and E. Süli, *hp-adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems*, SIAM J. Sci. Comput., 23 (2001), pp. 1226–1252.
[19] N. N. Kuznetzov, *Finite difference schemes for multidimensional first order quasi-linear equations in classes of discontinuous functions*, in Probl. Math. Phys. Vych. Mat., Nauka, Moscow, 1977, pp. 181–194.
[20] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, 1992.
[21] R. J. LeVeque, *High-resolution conservative algorithms for advection in incompressible flow*, SIAM J. Numer. Anal., 33 (1996), pp. 627–665.
[22] P. L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation laws and related equations*, J. Amer. Math. Soc., 7 (1994), pp. 169–191.

[23] P. Lesaint, *Sur la résolution des systèmes hyperboliques du premier ordre par des méthodes d'éléments finis*, Ph.D. thesis, Université de Paris VI, Paris, France, 1975.
[24] P. Lesaint and P. A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York, 1974, pp. 89–123.
[25] S. Nöelle, *Convergence of higher order finite volume schemes on irregular grids*, Adv. Comput. Math., 3 (1995), pp. 197–218.
[26] T. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.
[27] O. Pironneau and F. Hecht, http://www.freefem.org.
[28] Q. Lin, *Full convergence for hyperbolic finite elements*, in Proceedings of the First International Symposium on Discontinuous Galerkin Methods, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 167–177.
[29] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, Interscience, New York, 1957.
[30] P. L. Roe and D. Sidilkover, *Optimum positive linear schemes for advection in two and three dimensions*, SIAM J. Numer. Anal., 29 (1992), pp. 1542–1568.
[31] A. Szepessy, *Convergence of a streamline diffusion finite element method for conservation law with boundary conditions*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 749–782.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 505–529

ADAPTIVE APPROXIMATION OF YOUNG MEASURE SOLUTIONS IN SCALAR NONCONVEX VARIATIONAL PROBLEMS*

SÖREN BARTELS†

Abstract. This paper addresses the numerical approximation of Young measures appearing as generalized solutions to scalar nonconvex variational problems. We prove a priori and a posteriori error estimates for a macroscopic quantity, the stress. For a scalar three-well problem we show convergence of other quantities such as Young measure support and microstructure region. Numerical experiments indicate that the computational effort in the solution of the large optimization problem is significantly reduced by using an adaptive mesh refinement strategy based on a posteriori error estimates in combination with an active set strategy due to Carstensen and Roubíček [*Numer. Math.*, 84 (2000), pp. 395–414].

Key words. nonconvex variational problems, microstructure, finite elements, error estimation, adaptivity, multiple scales

AMS subject classifications. 65K10, 65N15, 49M40

DOI. 10.1137/S0036142902404091

1. Introduction.

A scalar model example in the context of phase transitions in crystalline solids reads

$$\text{(P)} \quad \begin{cases} \text{Seek } u \in \mathcal{A} := \{v \in W^{1,2}(\Omega) : v|_{\Gamma_D} = u_D\} \\ \text{such that } I(u) = \inf\limits_{v\in\mathcal{A}} I(v). \end{cases}$$

Here, $\Omega \subseteq \mathbb{R}^n$ is a bounded Lipschitz domain, $\Gamma_D \subseteq \partial\Omega$ a closed subset of $\partial\Omega$ with positive surface measure, and $u_D \in W^{1/2,2}(\Gamma_D)$ is the trace of some function $\tilde{u}_D \in W^{1,2}(\Omega)$. The energy functional $I : \mathcal{A} \to \mathbb{R}$ is for $v \in \mathcal{A}$ defined by

$$\begin{aligned} I(v) := &\int_\Omega W(\nabla v(x))dx + \alpha \int_\Omega |u_0(x) - v(x)|^2 dx \\ &- \int_\Omega f(x)v(x)dx - \int_{\Gamma_N} g(x)v(x)ds_x, \end{aligned}$$

where $u_0, f \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$ for $\Gamma_N := \partial\Omega\backslash\Gamma_D$, and $\alpha \geq 0$. An energy density W that can be derived from a three-dimensional model with one-dimensional symmetry [2] is given by $N+1$ wells $s_0, \dots, s_N \in \mathbb{R}^n$ and numbers $s_0^0, \dots, s_N^0 \in \mathbb{R}$ and reads

$$W(s) = \min_{j=0,\dots,N} \left(|s - s_j|^2 + s_j^0\right) \quad \forall s \in \mathbb{R}^n. \tag{1.1}$$

This function W serves as a model energy density, but more generally we will consider mappings $W : \mathbb{R}^n \to \mathbb{R}$ which are continuous and satisfy quadratic growth conditions.

The contributions in I which involve f and g represent outer body forces, while the integral of $W(\nabla v)$ measures the stored energy in Ω. A mechanical interpretation

*Received by the editors March 15, 2002; accepted for publication (in revised form) August 16, 2003; published electronically March 11, 2004. This research was supported by the Priority Program "Analysis, Modeling and Simulation of Multiscale Problems" and the Graduate School "Efficient Algorithms and Multiscale Methods" of the German Research Foundation (DFG).

http://www.siam.org/journals/sinum/42-2/40409.html

†Department of Mathematics, University of Maryland, College Park, MD 20742 (sba@math.umd.edu).

of the term $\alpha\|u_0 - v\|^2_{L^2(\Omega)}$ may be obtained from a model of a thin crystal plate glued to a rigid substrate [12]. Similar scalar minimization problems arise in optimal control theory [27]. For ease of presentation, we restrict the analysis to quadratic growth conditions ($p = 2$) but stress that the estimates can easily be generalized to other growth conditions ($2 \le p < \infty$).

It is well known that existence of solutions for (P) depends on convexity properties of W: If W is convex, then there exists a solution which is unique, provided W is strictly convex or $\alpha > 0$. In case that W fails to be convex, then I is not weakly lower semicontinuous and solutions may not exist. In the latter case, infimizing sequences are generically enforced to develop oscillations and therefore do not converge to a global minimizer of I. To be able to deal with this phenomenon, we will consider appropriate (weak*) limits of infimizing sequences for (P) which contain the most important information and which show where infimizing sequences develop oscillations. Those limits are measure-valued functions called Young measures and arise as solutions for an extended problem (EP). We refer the reader to [23, 26, 27] for details on the mathematical analysis of (P). The numerical approximation of the extended problem has been proposed in [24, 27, 17, 8, 25, 28] and in [18] for a nonconvex variational problem in the theory of micromagnetics. It is our aim to establish error estimates for the numerical treatment of the extended problem. We note, however, that our analysis is restricted to scalar problems. The practically more relevant case of nonconvex vectorial variational problems requires an efficient characterization of gradient Young measures and is excluded from our considerations. The alternative approach of directly minimizing I is discussed in [7, 9, 21].

The idea for the derivation of a priori and a posteriori error estimates is that the discretized extended problem may be regarded as a perturbation of a discretization of a relaxed (convexified) problem which has been analyzed in [7]. This perturbation consists of the difference between the convex hull of the energy density itself and the convex hull of a discrete approximation of the energy density. Employing the concept of subdifferentials in the theory of nonsmooth optimization we show that a dual variable, occurring in the discretized extended problem, converges to a macroscopic quantity of the relaxed problem and prove related error estimates. Moreover, we prove computable error estimates that allow for adaptive mesh refinement and which characterize a reliable relation between the two scales involved.

The "active set strategy" of [8] to solve a discretization of (EP) efficiently for a fixed triangulation of Ω is a multilevel scheme and depends on a good guess of a solution. Based on our error estimates we propose the embedding of that scheme into an adaptive mesh refining algorithm. We report the performance of the resulting algorithm for two examples. Our overall observation is that the algorithm performs very efficiently but depends on a good solver for large optimization problems. For a two-dimensional problem a numerical experiment indicates linear complexity of our solving strategy.

The outline of the rest of this paper is as follows. We state the extended problem in section 2 and proceed in section 3 with some notation, a construction of discrete Young measures, and the formulation of the discrete problem. Section 4 gives the announced error analysis as the main contribution of this work. Section 5 is devoted to the analysis of convergence of various quantities in a scalar three-well problem. The "active set strategy" of [8] and its embedding into an adaptive mesh refinement algorithm are given in section 6. Finally, in section 7, we report on numerical results for two specifications of (P) which illustrate the theoretical results of this article.

2. Young measures and the extended problem. In this section we recall the notion of Young measures which are mappings from Ω into the space of probability measures on $\mathbb{R}^n$ and allow for the computation of certain limits of weakly* convergent sequences in Lebesgue spaces.

DEFINITION 2.1. *Let $\mathcal{M}(\mathbb{R}^n)$ be the set of all signed Radon measures on $\mathbb{R}^n$, and let $PM(\mathbb{R}^n)$ be the subset of probability measures on $\mathbb{R}^n$, i.e., the set of all nonnegative Radon measures $\mu \in \mathcal{M}(\mathbb{R}^n)$ satisfying $\int_{\mathbb{R}^n} \mu(ds) = 1$. The set of L^2-*Young measures $\mathcal{Y}_2(\Omega;\mathbb{R}^n)$ *is defined as*

$$\mathcal{Y}_2(\Omega;\mathbb{R}^n) := \Big\{ \nu \in L^\infty_w(\Omega;\mathcal{M}(\mathbb{R}^n)) : \nu_x \in PM(\mathbb{R}^n) \text{ for a.a. } x \in \Omega, \int_\Omega \int_{\mathbb{R}^n} |s|^2 \nu_x(ds)dx < \infty \Big\}.$$

Here $\nu_x := \nu(x)$ for $x \in \Omega$, and $L^\infty_w(\Omega;\mathcal{M}(\mathbb{R}^n))$ consists of those mappings $\nu \in L^\infty(\Omega;\mathcal{M}(\mathbb{R}^n))$ for which the mapping $x \mapsto \int_{\mathbb{R}^n} v(s)\,\nu_x(ds)$ is measurable whenever $v \in C(\mathbb{R}^n)$ satisfies $\lim_{|s|\to\infty} v(s) = 0$.

Infimizing sequences for (P) generate Young measures in the sense of the following statement, which is a consequence of the fundamental theorem on Young measures [30, 1, 16, 27]. Throughout this paper we assume that there exist constants $c_1, c_2 > 0$ such that

$$c_1|s|^2 - c_2 \le W(s) \le c_2(1 + |s|^2) \quad \forall s \in \mathbb{R}^n. \tag{2.1}$$

LEMMA 2.2 (see [26, Lemma 4.3]). *Let $(u_j) \subseteq \mathcal{A}$ be an infimizing sequence for (P), i.e., $I(u_j) \to \inf_{v\in\mathcal{A}} I(v)$. Then there exist $u \in \mathcal{A}$, $\nu \in \mathcal{Y}_2(\Omega;\mathbb{R}^n)$, and a subsequence (u_k) such that $u_k \rightharpoonup u$ (weakly) in $W^{1,2}(\Omega)$,*

$$\int_\Omega W(\nabla u_k(x))dx \to \int_\Omega \int_{\mathbb{R}^n} W(s)\,\nu_x(ds)dx,$$

and, for almost all $x \in \Omega$, there holds $\nabla u(x) = \int_{\mathbb{R}^n} s\,\nu_x(ds)$.

The Young measure ν generated by the gradients of an infimizing sequence (u_j) for (P) describes oscillations in that sequence in a statistical way [1]. Together with the weak limit u, we obtain the most relevant information about (P). If we express the limit of $I(u_j)$ in terms of u and ν we obtain the extended problem (EP).

$$\text{(EP)} \quad \begin{cases} \text{Seek } (u,\nu) \in \mathcal{B} := \Big\{ (v,\mu) \in W^{1,2}(\Omega) \times \mathcal{Y}_2(\Omega;\mathbb{R}^n) : \\ \qquad v|_{\Gamma_D} = u_D, \nabla v(x) = \displaystyle\int_{\mathbb{R}^n} s\,\mu_x(ds) \text{ for a.a. } x \in \Omega \Big\} \\ \text{such that } \overline{I}(u,\nu) = \displaystyle\inf_{(v,\mu)\in\mathcal{B}} \overline{I}(v,\mu). \end{cases}$$

The extended energy functional $\overline{I}$ is for $(v,\mu) \in \mathcal{B}$ defined by

$$\overline{I}(v,\mu) := \int_\Omega \int_{\mathbb{R}^n} W(s)\,\mu_x(ds)dx + \alpha \int_\Omega |u_0 - v|^2 dx - \int_\Omega fv dx - \int_{\Gamma_N} gv ds_x.$$

The following theorem shows that (EP) is a correct extension of (P). Limits in $\mathcal{B}$ refer to the (weak, weak*)-topology in $W^{1,2}(\Omega) \times \mathcal{Y}_2(\Omega;\mathbb{R}^n)$ (cf. [27] for details).

Via the mapping $\iota : \mathcal{A} \to \mathcal{B}$, $u \mapsto (u, \delta_{\nabla u})$, where for almost all $x \in \Omega$ and all $v \in C(\mathbb{R}^n)$ with $\lim_{|s|\to\infty} v(s) = 0$ the Dirac measure $\delta_{\nabla u(x)} \in PM(\mathbb{R}^n)$ is defined by $\int_{\mathbb{R}^n} v(s)\,\delta_{\nabla u(x)}(ds) = v(\nabla u(x))$, $\mathcal{A}$ can be embedded continuously into $\mathcal{B}$.

THEOREM 2.3 (see [27, Proposition 5.2.1]). (i) (EP) *admits a solution.*

(ii) $\inf_{v\in\mathcal{A}} I(v) = \min_{(w,\mu)\in\mathcal{B}} \overline{I}(w,\mu)$.

(iii) *The embedding $\iota : \mathcal{A} \to \mathcal{B}$ of each infimizing sequence for* (P) *has a convergent subsequence whose limit is a solution to* (EP).

(iv) *Each solution to* (EP) *is the limit of the embedding $\iota : \mathcal{A} \to \mathcal{B}$ of an infimizing sequence for* (P).

Carathéodory's theorem implies that there exist solutions $(u, \nu) \in \mathcal{B}$ to (EP) such that for almost all $x \in \Omega$ the probability measure ν_x is a convex combination of at most $n + 1$ Dirac measures (cf. [27, Corollary 5.3.3]). This fact motivates the discretization of (EP) introduced in section 3 and the algorithm of [8] to efficiently approximate (EP).

3. Discretization of (EP). This section is devoted to the construction of a discrete subspace of $\mathcal{B}$.

3.1. Finite element spaces and notation. Let $\mathcal{T}$ be a regular triangulation of Ω into triangles ($n = 2$) or tetrahedra ($n = 3$) in the sense of [10], i.e., there are no hanging nodes; the domain is matched exactly, i.e., $\overline{\Omega} = \cup_{T\in\mathcal{T}} T$; and $\mathcal{T}$ satisfies the maximum angle condition. Therefore, $\partial\Omega$ is assumed to be polygonal. The extremal points of $T \in \mathcal{T}$ are called nodes, and $\mathcal{N}$ denotes the set of all such nodes. Let $\mathcal{K} := \mathcal{N} \setminus \Gamma_D$ be the subset of free nodes. The set of edges (respectively, faces if $n = 3$) $E = \operatorname{conv}\{z_1, \dots, z_n\} \subseteq \partial T$ for pairwise distinct $z_1, \dots, z_n \in \mathcal{N}$ and $T \in \mathcal{T}$ is denoted as $\mathcal{E}$. A partition $\mathcal{E} = \mathcal{E}_\Omega \cup \mathcal{E}_D \cup \mathcal{E}_N$ is given by $\mathcal{E}_N := \{E \in \mathcal{E} : E \subseteq \overline{\Gamma}_N\}$, $\mathcal{E}_D := \{E \in \mathcal{E} : E \subseteq \Gamma_D\}$, and $\mathcal{E}_\Omega := \mathcal{E} \setminus (\mathcal{E}_D \cup \mathcal{E}_N)$. The set

$$\mathcal{L}^k(\mathcal{T}) := \{v_h \in L^\infty(\Omega) : \forall T \in \mathcal{T},\ v_h|_T \in P_k(T)\}$$

consists of all (possibly discontinuous) $\mathcal{T}$-elementwise polynomials of degree at most k. Define

$$\mathcal{S}^1(\mathcal{T}) := \mathcal{L}^1(\mathcal{T}) \cap C(\overline{\Omega}) \quad \text{and} \quad \mathcal{S}^1_D(\mathcal{T}) := \{u_h \in \mathcal{S}^1(\mathcal{T}) : u_h|_{\Gamma_D} = 0\} \subseteq W^{1,2}_D(\Omega),$$

where $W^{1,2}_D(\Omega) := \{v \in W^{1,2}(\Omega) : v|_{\Gamma_D} = 0\}$. Let $(\varphi_z : z \in \mathcal{N})$ be the nodal basis of $\mathcal{S}^1(\mathcal{T})$; i.e., $\varphi_z \in \mathcal{S}^1(\mathcal{T})$ satisfies $\varphi_z(x) = 0$ if $x \in \mathcal{N} \setminus \{z\}$ and $\varphi_z(z) = 1$. A function $h_\mathcal{T} \in \mathcal{L}^0(\mathcal{T})$ is defined by $h_\mathcal{T}|_T = h_T := \operatorname{diam}(T)$ for all $T \in \mathcal{T}$. Moreover, let $h_\mathcal{E} \in L^\infty(\cup\mathcal{E})$ be defined by $h_\mathcal{E}|_E = h_E := \operatorname{diam}(E)$ for all $E \in \mathcal{E}$.

The nodal interpolation operator associated with a triangulation $\mathcal{T}$ is denoted by $P_\mathcal{T}$. If τ is a triangulation of a convex domain $\omega \subseteq \mathbb{R}^n$ and $v \in C(\overline{\omega})$ we extend $P_\tau v$ to $\mathbb{R}^n$ by setting $P_\tau v(s) = P_\tau v(\mathcal{P}_\omega(s))$, where $\mathcal{P}_\omega$ denotes the orthogonal projection onto $\overline{\omega}$.

Suppose $g \in L^2(\Gamma_N)$ is such that $g|_E \in W^{1,2}(E)$ for all $E \in \mathcal{E}_N$ and, for each node $z \in \mathcal{N} \cap \overline{\Gamma}_N$ where the outer unit normal n_{Γ_N} on Γ_N is continuous, g is continuous. We set

$$\mathcal{S}^1_N(\mathcal{T}, g) := \{\tau_h \in \mathcal{S}^1(\mathcal{T})^n : \forall E \in \mathcal{E}_N\ \forall z \in E \cap \mathcal{N},\ \tau_h(z) \cdot n_{\Gamma_N}|_E = g(z)\} \tag{3.1}$$

and note that $\mathcal{S}^1_N(\mathcal{T}, g) \neq \emptyset$ if $n = 2$. We will assume that $\mathcal{S}^1_N(\mathcal{T}, g) \neq \emptyset$ if $n = 3$.

Throughout this article $c, C > 0$ denote mesh-size independent, generic constants. For $1 \le p \le \infty$ and an integer $\ell > 0$, $\|\cdot\|_{L^p(\Omega)}$ stands for $\|\cdot\|_{L^p(\Omega;\mathbb{R}^\ell)}$, and $\|\cdot\|$ abbreviates $\|\cdot\|_{L^2(\Omega)}$. The operator $\partial_\mathcal{E}\cdot/\partial s$ denotes the edgewise derivative along (subsets of) $\partial\Omega$.

3.2. Discrete Young measures. We define a convex, discrete (i.e., finite-dimensional) subset of the set of L^2-Young measures $\mathcal{Y}_2(\Omega;\mathbb{R}^n)$ following ideas of [27, 8, 22].

DEFINITION 3.1 (see [27, Example 3.5.4]). *Given a convex polygonal set $\omega \subseteq \mathbb{R}^n$ and regular triangulations τ of ω with nodes $\mathcal{N}_\tau$ and $\mathcal{T}$ of Ω we set*

$$YM_{d,h}(\Omega;\mathbb{R}^n) := \Bigg\{ \nu_{d,h} \in \mathcal{Y}_2(\Omega;\mathbb{R}^n) : \forall z \in \mathcal{N}_\tau \ \exists a_z \in \mathcal{L}^0(\mathcal{T}),\, a_z \geq 0 \text{ and} \\ \sum_{z\in\mathcal{N}_\tau} a_z(x) = 1 \text{ a.e. in } \Omega,\, \nu_{d,h,x} = \sum_{z\in\mathcal{N}_\tau} a_z(x)\delta_z \text{ for a.e. } x \in \Omega \Bigg\},$$

where δ_z denotes the Dirac measure supported in the atom $z \in \mathbb{R}^n \cap \mathcal{N}_\tau$. *By d and h we denote the maximal mesh-size in τ and $\mathcal{T}$, respectively, and refer to τ and $\mathcal{T}$ through these quantities.*

3.3. Discretized extended problem. For regular triangulations $\mathcal{T}$ of Ω and τ of a convex Lipschitz domain $\omega \subseteq \mathbb{R}^n$ and an approximation $u_{D,h} \in \mathcal{S}^1(\mathcal{T})|_{\Gamma_D}$ of u_D we consider the following discrete problem ($\mathrm{EP}_{d,h}$):

$$(\mathrm{EP}_{d,h}) \quad \begin{cases} \text{Seek } (u_{d,h},\nu_{d,h}) \in \mathcal{B}_{d,h} := \Big\{ (v_h,\mu_{d,h}) \in \mathcal{S}^1(\mathcal{T}) \times YM_{d,h}(\Omega;\mathbb{R}^n) : \\ \qquad v_h|_{\Gamma_D} = u_{D,h}, \nabla v_h(x) = \displaystyle\int_{\mathbb{R}^n} s\,\mu_{d,h,x}(ds) \text{ for a.e. } x \in \Omega \Big\}, \\ \text{such that } \overline{I}(u_{d,h},\nu_{d,h}) = \displaystyle\inf_{(v_h,\mu_{d,h})\in\mathcal{B}_{d,h}} \overline{I}(v_h,\mu_{d,h}). \end{cases}$$

An existence result for ($\mathrm{EP}_{d,h}$) follows as for (EP).

PROPOSITION 3.2 (see [27, Proposition 5.5.1]). *If $\mathcal{B}_{d,h} \neq \emptyset$, then* ($\mathrm{EP}_{d,h}$) *admits a solution.*

Remarks. (i) There holds $\mathcal{B}_{d,h} \neq \emptyset$ if the diameter of ω is large enough.

(ii) For efficient approximations one has to assume a uniform bound on the gradient of a solution for (EP). Based on the optimality conditions stated below one may, however, enlarge ω successively to obtain a correct discrete solution. Therefore, no a priori bound on the gradient of an exact solution for (EP) will be assumed.

(iii) For a triangulation $\mathcal{T}$ of Ω with N^n free nodes and a triangulation τ of ω with N^n atoms the number of degrees of freedom in ($\mathrm{EP}_{d,h}$) is N^{2n} if $h \approx d \approx 1/N$.

3.4. Optimality conditions. The following lemma describes optimality conditions for ($\mathrm{EP}_{d,h}$) which are key ingredients for the subsequent analysis.

LEMMA 3.3 (see [8, Proposition 4.3]). *Assume $\omega = \mathbb{R}^n$. The pair $(u_{d,h},\nu_{d,h}) \in \mathcal{B}_{d,h}$ is a solution for* ($\mathrm{EP}_{d,h}$) *if and only if there exists $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^n$ such that, for almost all $x \in \Omega$, we have*

$$\max_{s\in\omega} \mathcal{H}_{\lambda_{d,h}}(x,s) = \int_{\mathbb{R}^n} \mathcal{H}_{\lambda_{d,h}}(x,s)\,\nu_{d,h,x}(ds),$$

where $\mathcal{H}_{\lambda_{d,h}}(x,s) := \lambda_{d,h}(x)\cdot s - P_\tau W(s)$, and, for all $v_h \in \mathcal{S}^1_D(\mathcal{T})$, there holds

$$\int_\Omega \lambda_{d,h}\cdot\nabla v_h dx = 2\alpha\int_\Omega (u_0 - u_{d,h})v_h dx + \int_\Omega f v_h dx + \int_{\Gamma_N} g v_h ds_x.$$

Remark. The elementwise constant function $\lambda_{d,h}$ is the Lagrange multiplier for the constraint $\nabla u_{d,h}|_T = \int_{\mathbb{R}^n} s\,\nu_{d,h}|_T(ds)$, $T \in \mathcal{T}$, in $(\mathrm{EP}_{d,h})$.

For the practical implementation a bounded domain ω and a finite discretization of ω has to be chosen. We formulate appropriate computable conditions that imply Lemma 3.3.

LEMMA 3.4. *Assume that ω is bounded and $B_{r_0}(0) := \{s \in \mathbb{R}^n : |s| < r_0\} \subseteq \omega$ for some $r_0 > 0$. Let $(u_{d,h}, \nu_{d,h}) \in \mathcal{B}_{d,h}$, $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^n$, and assume*

$$\int_\Omega \lambda_{d,h} \cdot \nabla v_h dx = 2\alpha \int_\Omega (u_0 - u_{d,h}) v_h dx + \int_\Omega f v_h dx + \int_{\Gamma_N} g v_h ds_x.$$

If for almost all $x \in \Omega$ the mapping $s \mapsto \lambda_{d,h}(x) \cdot s - P_\tau W(s)$, $s \in \overline{\omega}$, attains its maximum in some $s_x^* \in \overline{B_{r_0}(0)}$, if $2r_0 c_1 \geq \|\lambda\|_{L^\infty(\Omega)}$, and if for almost all $x \in \Omega$

$$r_0 \|\lambda\|_{L^\infty(\Omega)} - c_1 r_0^2 + c_2 \leq \lambda_{d,h}(x) \cdot s_x^* - W(s_x^*), \tag{3.2}$$

then the conditions of Lemma 3.3 are satisfied; i.e., $(u_{d,h}, \nu_{d,h})$ is a solution for $(\mathrm{EP}_{d,h})$.

Proof. It suffices to show that for almost all $x \in \Omega$, any extension $\tilde{\tau}$ of τ to $\mathbb{R}^n$, and all $s \in \mathcal{N}_{\tilde{\tau}} \setminus \overline{B_{r_0}(0)}$ there holds $\lambda_{d,h}(x) \cdot s - P_{\tilde{\tau}} W(s) \leq \mathcal{H}_{\lambda_{d,h}}(x, s_x^*)$ since then the optimality conditions of Lemma 3.3 are satisfied (with $\tilde{\omega} = \mathbb{R}^n$ and $\tilde{\tau}$). In view of (2.1) there holds

$$\lambda_{d,h}(x) \cdot s - P_{\tilde{\tau}} W(s) \leq \|\lambda_{d,h}\|_{L^\infty(\Omega)} |s| - c_1 |s|^2 + c_2.$$

Since $2r_0 c_1 \geq \|\lambda\|_{L^\infty(\Omega)}$ the mapping $t \mapsto \|\lambda_{d,h}\|_{L^\infty(\Omega)} t - c_1 t^2 + c_2$, $t \geq r_0$, is monotonically decreasing and since $\|\lambda_{d,h}\|_{L^\infty(\Omega)} r_0 - c_1 r_0^2 + c_2 \leq \mathcal{H}_{\lambda_{d,h}}(x, s_x^*)$ we have $\lambda_{d,h}(x) \cdot s - P_{\tilde{\tau}} W(s) \leq \mathcal{H}_{\lambda_{d,h}}(x, s_x^*)$ for all $s \in \mathcal{N}_{\tilde{\tau}} \setminus \overline{B_{r_0}(0)}$, which implies the same estimate for all $s \in \mathbb{R}^n \setminus \overline{B_{r_0}(0)}$. ∎

4. Error estimates for $(\mathrm{EP}_{d,h})$. We now turn to the formulation of error estimates for solutions for $(\mathrm{EP}_{d,h})$. We prove that the Lagrange multiplier $\lambda_{d,h}$ converges to a macroscopic quantity, the stress, that appears naturally in (P) and also in the convexified problem (P**). To estimate the distance between $\lambda_{d,h}$ and the exact stress we will regard $(\mathrm{EP}_{d,h})$ as a perturbation of a discretization of (P**).

$$(\mathrm{P}^{**}) \quad \text{Seek } u \in \mathcal{A} \text{ such that } I^{**}(u) = \inf_{v \in \mathcal{A}} I^{**}(v).$$

Here, the energy functional I^{**} is defined for $v \in \mathcal{A}$ and the convex envelope W^{**} of W by

$$I^{**}(v) := \int_\Omega W^{**}(\nabla v(x)) dx + \alpha \int_\Omega |u_0 - v|^2 dx - \int_\Omega f v dx - \int_{\Gamma_N} g v ds_x.$$

DEFINITION 4.1. *For a solution $u \in \mathcal{A}$ for* (P**) *we define the* stress $\sigma := DW^{**}(\nabla u) \in L^2(\Omega)^n$.

THEOREM 4.2 (see [7, Theorem 2]). (P**) *admits a solution $u \in \mathcal{A}$ such that*

$$\int_\Omega DW^{**}(\nabla u) \cdot \nabla v dx - 2\alpha \int_\Omega (u_0 - u) v dx - \int_\Omega f v dx - \int_{\Gamma_N} g v ds_x = 0 \tag{4.1}$$

for all $v \in W_D^{1,2}(\Omega)$. *If* DW^{**} *satisfies, for all* $F, G \in \mathbb{R}^n$,

$$|DW^{**}(F) - DW^{**}(G)|^2 \le C\,(DW^{**}(F) - DW^{**}(G)) \cdot (F - G), \tag{4.2}$$

then for two solutions $u, w \in \mathcal{A}$ *for* (P**) *there holds* $DW^{**}(\nabla u) = DW^{**}(\nabla w)$; *i.e.,* σ *is unique. If in addition to* (4.2) $\alpha > 0$ *or* W^{**} *is strictly convex, then* $u = w$.

Remarks. (i) If W is as in (1.1), then DW^{**} satisfies (4.2).

(ii) For a solution $(u, \nu) \in \mathcal{B}$ for (EP) and a solution w for (P**) we have, provided $W, W^{**} \in C^1(\mathbb{R}^n)$, for almost all $x \in \Omega$ [14, 16],

$$\int_{\mathbb{R}^n} DW(s)\,\nu_x(ds) = DW^{**}(\nabla w(x)).$$

(iii) A result in [6] shows $\sigma \in W^{1,2}_{loc}(\Omega)$.

In order to obtain a version of (4.1) in the discrete setting (EP$_{d,h}$) we need to differentiate the nonsmooth convexification of $P_\tau W$. To do this we apply the concept of subdifferentials.

DEFINITION 4.3. *For a convex function* $V : \mathbb{R}^n \to \mathbb{R}$ *and* $\varsigma \in \mathbb{R}^n$ *the* subdifferential *of* V *at* ς *is defined by*

$$\partial V(\varsigma) := \{\xi \in \mathbb{R}^n : V(\varsigma + \zeta) - V(\varsigma) \ge \zeta \cdot \xi \ \forall \zeta \in \mathbb{R}^n\}.$$

Remarks (see [11]). (i) V has a minimum in $\varsigma \in \mathbb{R}^n$ if and only if $0 \in \partial V(\varsigma)$.

(ii) If V is Gâteaux differentiable in $\varsigma \in \mathbb{R}^n$, then $\partial V(\varsigma) = \{\nabla V(\varsigma)\}$.

The following lemma shows that the finite-dimensional minimization problem (EP$_{d,h}$) may be seen as a perturbation of a discretization of (P**).

LEMMA 4.4. *Let* $W_d^{cx} := ((P_\tau W)|_{\overline{\omega}})^{**}$ *denote the convexification of the restriction of* $P_\tau W$ *to* $\overline{\omega}$. *Assume that* $(u_{d,h}, \nu_{d,h}) \in \mathcal{B}_{d,h}$ *and* $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^n$ *satisfy the conditions of Lemma* 3.4. *Then* $(u_{d,h}, \nu_{d,h})$ *minimizes the modified energy functional*

$$\begin{aligned}\overline{I}'(v_h, \mu_{d,h}) := &\int_\Omega \int_{\mathbb{R}^n} W_d^{cx}(s)\,\mu_{d,h,x}(ds)dx \\ &+ \alpha \int_\Omega |u_0 - v_h|^2 dx - \int_\Omega f v_h dx - \int_{\Gamma_N} g v_h ds_x,\end{aligned}$$

among all $(v_h, \mu_{d,h}) \in \mathcal{B}_{d,h}$. *Moreover,* $\lambda_{d,h}(x) \in \partial W_d^{cx}(\nabla u_{d,h}(x))$ *for a.e.* $x \in \Omega$.

Proof. For $s \in \omega$ we have by Carathéodory's theorem [27],

$$W_d^{cx}(s) = ((P_\tau W)|_{\overline{\omega}})^{**}(s) = \inf_{\substack{s_1,\dots,s_{n+1} \in \overline{\omega},\\ \theta_1,\dots,\theta_{n+1} \in [0,1],\\ \sum_{i=1}^{n+1} \theta_i = 1,\, \sum_{i=1}^{n+1} \theta_i s_i = s}} \sum_{i=1}^{n+1} \theta_i P_\tau W(s_i).$$

Since $P_\tau W|_{\overline{\omega}}$ is τ-elementwise affine, it suffices to use the nodal values of $P_\tau W$ in the calculation of W_d^{cx}, i.e.,

$$W_d^{cx}(s) = ((P_\tau W)|_{\overline{\omega}})^{**}(s) = \inf_{\substack{\theta_z \in [0,1],\\ \sum_{z \in \mathcal{N}_\tau} \theta_z = 1,\\ \sum_{z \in \mathcal{N}_\tau} \theta_z z = s}} \sum_{z \in \mathcal{N}_\tau} \theta_z P_\tau W(z). \tag{4.3}$$

Assume that there exists $s \in \operatorname{conv}\{z_1, \dots, z_{n+1}\} = t \in \tau$, $z_1, \dots, z_{n+1} \in \mathcal{N}_\tau$, such that $s = \sum_{i=1}^{n+1} \alpha_i z_i$ but $W_d^{cx}(s) \ne \sum_{i=1}^{n+1} \alpha_i W_d^{cx}(z_i)$ with $\alpha_i \in [0,1]$, $\sum_{i=1}^{n+1} \alpha_i = 1$.

If $W_d^{cx}(s) > \sum_{i=1}^{n+1} \alpha_i W_d^{cx}(z_i)$, then $W_d^{cx}(s)$ was not convex. In case that $W_d^{cx}(s) < \sum_{i=1}^{n+1} \alpha_i W_d^{cx}(z_i)$, then W_d^{cx} was not the largest convex function satisfying $W_d^{cx} \le P_\tau W|_{\overline{\omega}}$. Therefore, $W_d^{cx}(s) = \sum_{i=1}^{n+1} \alpha_i W_d^{cx}(z_i)$ so that W_d^{cx} is τ-elementwise affine and $P_\tau W_d^{cx}|_{\overline{\omega}} = W_d^{cx}$. To prove that $(u_{d,h}, \nu_{d,h})$ minimizes the functional $\overline{I}'$ it suffices to verify the optimality conditions from Lemma 3.4 with $P_\tau W$ replaced by $P_\tau W_d^{cx}$. For this it is sufficient to show that, for almost all $x \in \Omega$, there holds

$$\max_{s\in\omega}(\lambda_{d,h}(x)\cdot s - P_\tau W(s)) = \max_{s\in\omega}(\lambda_{d,h}(x)\cdot s - W_d^{cx}(s)) \tag{4.4}$$

and

$$\int_{\mathbb{R}^n} (\lambda_{d,h}(x)\cdot s - W_d^{cx}(s))\,\nu_{d,h,x}(ds) = \int_{\mathbb{R}^n} (\lambda_{d,h}(x)\cdot s - P_\tau W(s))\,\nu_{d,h,x}(ds). \tag{4.5}$$

Since $W_d^{cx} \le P_\tau W(s)|_{\overline{\omega}}$, we have to show only that

$$\max_{s\in\omega}(\lambda_{d,h}(x)\cdot s - P_\tau W(s)) \ge \max_{s\in\omega}(\lambda_{d,h}(x)\cdot s - W_d^{cx}(s))$$

and

$$\int_{\mathbb{R}^n} W_d^{cx}(s)\,\nu_{d,h,x}(ds) \ge \int_{\mathbb{R}^n} P_\tau(s)\,\nu_{d,h,x}(ds).$$

Let $\overline{s} \in \overline{\omega}$ be maximizing in the right-hand side of (4.4), i.e.,

$$\lambda_{d,h}(x)\cdot\overline{s} - W_d^{cx}(\overline{s}) = \max_{s\in\omega}(\lambda_{d,h}(x)\cdot s - W_d^{cx}(s)).$$

By definition of W_d^{cx} there exist $\theta_1,\dots,\theta_{n+1} \in [0,1]$, $\sum_{i=1}^{n+1}\theta_i = 1$, and $z_1,\dots,z_{n+1} \in \mathcal{N}_\tau$ such that $\sum_{i=1}^{n+1}\theta_i z_i = \overline{s}$ and $W_d^{cx}(\overline{s}) = \sum_{i=1}^{n+1}\theta_i P_\tau W(z_i)$. By linearity of $s \mapsto \lambda_{d,h}(x)\cdot s$ we have

$$\begin{aligned}
\lambda_{d,h}(x)\cdot\overline{s} - W_d^{cx}(\overline{s}) &= \sum_{i=1}^{n+1}\theta_i(\lambda_{d,h}(x)\cdot z_i - P_\tau W(z_i)) \\
&\le \sum_{i=1}^{n+1}\theta_i \max_{s\in\omega}(\lambda_{d,h}(x)\cdot s - P_\tau W(s)) \\
&= \max_{s\in\omega}(\lambda_{d,h}(x)\cdot s - P_\tau W(s)),
\end{aligned}$$

which proves (4.4). If $\int_{\mathbb{R}^n} W_d^{cx}(s)\,\nu_{d,h,x}(ds) < \int_{\mathbb{R}^n} P_\tau W(s)\,\nu_{d,h,x}(ds)$, the explicit representation of W_d^{cx} contradicts the fact that $(u_{d,h},\nu_{d,h})$ is minimal for $\overline{I}$. We have thus shown (4.5), which yields the optimality conditions. The maximum principle of Lemma 3.3, the convexity of the mapping $s \mapsto W_d^{cx}(s) - \lambda_{d,h}(x)\cdot s$ together with Jensen's inequality, and the identity $\nabla u_{d,h}(x) = \int_{\mathbb{R}^n} s d\nu_{d,h,x}(s)$ yield, for almost all $x \in \Omega$,

$$\begin{aligned}
\max_{s\in\omega}(\lambda_{d,h}(x)\cdot s - W_d^{cx}(s)) &= \int_{\mathbb{R}^n} (\lambda_{d,h}(x)\cdot s - W_d^{cx}(s))\,\nu_{d,h,x}(ds) \\
&\le \lambda_{d,h}(x)\cdot\nabla u_{d,h}(x) - W_d^{cx}(\nabla u_{d,h}(x)).
\end{aligned}$$

Therefore, for almost all $x \in \Omega$, we have $0 \in -\lambda_{d,h}(x) + \partial W_d^{cx}(\nabla u_{d,h}(x))$. ∎

Another definition is needed for the a priori and a posteriori error estimates. It concerns the approximation of DW^{**} by the multivalued mapping ∂W_d^{cx}.

DEFINITION 4.5. *For $A \subseteq \mathbb{R}^n$ and a multivalued mapping $S : A \to 2^{\mathbb{R}^n}$, where $2^{\mathbb{R}^n}$ denotes the power set of $\mathbb{R}^n$, let*

$$\|S\|_{L^\infty(A;2^{\mathbb{R}^n})} := \sup_{t\in A} \sup_{s\in S(t)} |s|.$$

4.1. A priori error estimates. The following theorem shows that the multiplier $\lambda_{d,h}$ for a solution $(u_{d,h}, \nu_{d,h}) \in \mathcal{B}_{d,h}$ for $(\mathrm{EP}_{d,h})$ approximates the unique quantity $\sigma = DW^{**}(\nabla u)$ for a solution $u \in \mathcal{A}$ for (P^{**}).

THEOREM 4.6. *Assume that DW^{**} satisfies (4.2) and $u \in \mathcal{A}$ solves (P^{**}). Assume that $(u_{d,h}, \nu_{d,h}) \in \mathcal{B}_{d,h}$ and $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^n$ satisfy the conditions of Lemma 3.3. There holds*

$$\begin{aligned}\|\sigma - \lambda_{d,h}\| + \alpha\|u - u_{d,h}\| \le C \inf_{(v_h,\mu_{d,h})\in\mathcal{B}_{d,h}} \big(\|\nabla(u - v_h)\| + \alpha\|u - v_h\|\big)\\ + C\|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}\\ + |\Omega|\sqrt{C'}\|\partial W_d^{cx} - DW^{**}\|^{1/2}_{L^\infty(\omega;2^{\mathbb{R}^n})}.\end{aligned}$$

Proof. The triangle inequality and estimate (4.2) show

$$\begin{aligned}\frac{1}{2}\|\sigma - \lambda_{d,h}\|^2 &\le \|\sigma - DW^{**}(\nabla u_{d,h})\|^2 + \|DW^{**}(\nabla u_{d,h}) - \lambda_{d,h}\|^2\\ &\le C\int_\Omega (DW^{**}(\nabla u) - DW^{**}(\nabla u_{d,h}))\cdot\nabla(u - u_{d,h})dx\\ &\quad + \|DW^{**}(\nabla u_{d,h}) - \lambda_{d,h}\|^2.\end{aligned}$$

Hölder's inequality yields

$$\begin{aligned}\frac{1}{2}\|\sigma - \lambda_{d,h}\|^2 &\le C\int_\Omega (DW^{**}(\nabla u) - \lambda_{d,h})\cdot\nabla(u - u_{d,h})dx\\ &\quad + C\int_\Omega (\lambda_{d,h} - DW^{**}(\nabla u_{d,h}))\cdot\nabla(u - u_{d,h})dx\\ &\quad + \|\lambda_{d,h} - DW^{**}(\nabla u_{d,h})\|^2\\ &\le C\int_\Omega (DW^{**}(\nabla u) - \lambda_{d,h})\cdot\nabla(u - u_{d,h})dx\\ &\quad + C\|\lambda_{d,h} - DW^{**}(\nabla u_{d,h})\|\|\nabla(u - u_{d,h})\|\\ &\quad + \|\lambda_{d,h} - DW^{**}(\nabla u_{d,h})\|^2.\end{aligned}$$

The Euler–Lagrange equations (4.1) for u and Lemma 3.3 yield, for all $w_h \in \mathcal{S}^1_D(\mathcal{T})$,

$$\int_\Omega (\sigma - \lambda_{d,h})\cdot\nabla w_h dx + 2\alpha\int_\Omega (u - u_{d,h})w_h dx = 0.$$

We thus have

$$\begin{aligned}\int_\Omega (\sigma-\lambda_{d,h})\cdot\nabla(u-u_{d,h})dx+2\alpha\int_\Omega (u-u_{d,h})^2dx\\ =\int_\Omega (\sigma-\lambda_{d,h})\cdot\nabla(u-u_{d,h}-w_h)dx+2\alpha\int_\Omega (u-u_{d,h})(u-u_{d,h}-w_h)dx\\ \le \|\sigma-\lambda_{d,h}\|\|\nabla(u-u_{d,h}-w_h)\|+2\alpha\|u-u_{d,h}\|\|u-u_{d,h}-w_h\|.\end{aligned}$$

The combination of the last two estimates shows, after absorption of $\|\sigma-\lambda_{d,h}\|$ and $\|u-u_{d,h}\|$,

$$\begin{aligned}\|\sigma-\lambda_{d,h}\|^2+\alpha\|u-u_{d,h}\|^2\le C\big(&\|\nabla(u-u_{d,h}-w_h)\|^2+\alpha\|u-u_{d,h}-w_h\|^2\\ &+\|\lambda_{d,h}-DW^{**}(\nabla u_{d,h})\|\|\nabla(u-u_{d,h})\|\\ &+\|\lambda_{d,h}-DW^{**}(\nabla u_{d,h})\|^2\big).\end{aligned}$$

Lemma 4.4 ensures $\lambda_{d,h}(x)\in\partial W_d^{cx}(\nabla u_{d,h}(x))$, and, by construction of $\mathcal{B}_{d,h}$ we have $\nabla u_{d,h}(x)\in\omega$ for almost all $x\in\Omega$. This implies

$$\begin{aligned}\|\lambda_{d,h}-DW^{**}(\nabla u_{d,h})\|^2&\le\int_\Omega \sup_{s\in\partial W_d^{cx}(\nabla u_{d,h}(x))-DW^{**}(\nabla u_{d,h})}|s|^2dx\\ &\le|\Omega|\sup_{t\in\omega}\sup_{s\in\partial W_d^{cx}(t)-DW^{**}(t)}|s|^2\\ &=|\Omega|\|\partial W_d^{cx}-DW^{**}\|^2_{L^\infty(\omega;2^{\mathbb{R}^n})}.\end{aligned}$$

Letting $w_h=v_h-u_{d,h}$ for arbitrary $(v_h,\mu_{d,h})\in\mathcal{B}_{d,h}$ and estimating $\|\nabla(u-u_{d,h})\|\le C$ (which follows from growth conditions (2.1)) we verify the assertion of the theorem. □

For a given energy density W and an appropriate triangulation τ of ω the term $\|\partial W_d^{cx}-DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}$ can be estimated by the mesh-size of the discretization τ of ω. We refer the reader to Theorem 5.1 below for an estimate for a three-well energy density.

Remarks. (i) Theorem 4.6, Theorem 5.1 below, and density of finite element spaces in $\mathcal{A}$ prove $\lambda_{d,h}\to\sigma$ in $L^2(\Omega)$ for $(d,h_{\mathcal{T}})\to 0$, and, if $\alpha>0$, we also have $u_{d,h}\to u$ in $L^2(\Omega)$. If $u\in C(\overline{\Omega})$ we may choose v_h in Theorem 4.6 as the nodal interpolant of u, and then we can estimate the error in powers of the mesh-size depending on smoothness properties of u. Since, in general, u has no higher regularity properties, computable error bounds are needed.

(ii) Owing to the nonuniqueness of u and degeneracy of (EP) we cannot expect strong convergence $u_{d,h}\to u$ in $W^{1,2}$.

4.2. A posteriori error estimates. In this section two a posteriori error estimates, which are computable bounds for the error $\|\sigma-\lambda_{d,h}\|$, are given. The first error estimate is similar to classical residual based a posteriori error estimates for elliptic partial differential equations [29] and employs jumps of normal components of $\lambda_{d,h}$ across edges. Recall from the definition of $(\mathrm{EP}_{d,h})$ that ω is a fixed convex subset of $\mathbb{R}^n$.

DEFINITION 4.7. *For $E\in\mathcal{E}_\Omega$ and $T_1,T_2\in\mathcal{T}$ with $E=T_1\cap T_2$ let n_E be the unit vector normal to E, pointing from T_1 into T_2. For $\lambda_{d,h}\in\mathcal{L}^0(\mathcal{T})^n$ define*

$$[\lambda_{d,h}\cdot n_E] := \begin{cases} (\lambda_{d,h}|_{T_2} - \lambda_{d,h}|_{T_1})\cdot n_E & \text{for } E\in\mathcal{E}_\Omega,\ T_1,T_2\in\mathcal{T},\ E=T_1\cap T_2,\\ g-\lambda_{d,h}|_T\cdot n_{\Gamma_N}|_E & \text{for } E\in\mathcal{E}_N,\ T\in\mathcal{T},\ E\subseteq\partial T.\end{cases}$$

THEOREM 4.8. *Assume that DW^{**} satisfies* (4.2) *and $u\in\mathcal{A}$ solves* (P^{**}). *Let $(u_{d,h},\nu_{d,h})\in\mathcal{B}_{d,h}$ and $\lambda_{d,h}\in\mathcal{L}^0(\mathcal{T})^n$ satisfy the conditions of Lemma* 3.3. *Then*

$$\begin{aligned}&\|\sigma-\lambda_{d,h}\|^2+\alpha\|u-u_{d,h}\|^2\\ &\quad\le C\Bigg\{\left(\sum_{T\in\mathcal{T}} h_T^2\|(f+\operatorname{div}\lambda_{d,h}+2\alpha(u_0-u_{d,h}))\|^2_{L^2(T)}\right)^{1/2}\\ &\qquad+\left(\sum_{E\in\mathcal{E}_\Omega\cup\mathcal{E}_N} h_E\|[\lambda_{d,h}\cdot n_E]\|^2_{L^2(E)}\right)^{1/2}+\|\partial W_d^{cx}-DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}\\ &\qquad+\left\|h_{\mathcal{E}}^{3/2}\partial^2_{\mathcal{E}}u_D/\partial s^2\right\|_{L^2(\Gamma_D)}\Bigg\}+|\Omega|\|\partial W_d^{cx}-DW^{**}\|^2_{L^\infty(\omega;2^{\mathbb{R}^n})}.\end{aligned}$$

Proof. Recall from the proof of Theorem 4.6 that, for $w\in W^{1,2}(\Omega)$ satisfying $w|_{\Gamma_D}=u_D-u_{D,h}$ and $v_h\in\mathcal{S}^1_D(\mathcal{T})$, there holds

$$\begin{aligned}C\|\sigma-\lambda_{d,h}\|^2+2\alpha\|u-u_{d,h}\|^2 \le{}& \int_\Omega (DW^{**}(\nabla u)-\lambda_{d,h})\cdot\nabla(u-u_{d,h}-w-v_h)dx\\ &+2\alpha\int_\Omega (u-u_{d,h})(u-u_{d,h}-w-v_h)dx\\ &+|\Omega|\|\partial W_d^{cx}-DW^{**}\|^2_{L^\infty(\omega;2^{\mathbb{R}^n})}\\ &+C|\Omega|\|\partial W_d^{cx}-DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}\\ &+\|\sigma-\lambda_{d,h}\|\|\nabla w\|+2\alpha\|u-u_{d,h}\|\|w\|.\end{aligned}$$

We employ the weak approximation operator $\mathcal{J}: W^{1,2}_D(\Omega)\to\mathcal{S}^1_D(\mathcal{T})$ of [4, 5] and set $v_h:=\mathcal{J}v$. We then have (cf. [5, Theorem 2.1])

$$\|\nabla v_h\|+\|h_{\mathcal{T}}^{-1}(v-v_h)\|+\|h_{\mathcal{E}}^{-1/2}(v-v_h)\|_{L^2(\cup\mathcal{E})}\le C\|\nabla v\| \tag{4.6}$$

The Euler–Lagrange equations (4.1) for u, an elementwise integration by parts, and (4.6) show for $v:=u-u_{d,h}-w\in W^{1,2}_D(\Omega)$

$$\begin{aligned}&\int_\Omega (DW^{**}(\nabla u)-\lambda_{d,h})\cdot\nabla(v-\mathcal{J}v)dx+2\alpha\int_\Omega (u-u_{d,h})(v-\mathcal{J}v)dx\\ &\quad=\sum_{T\in\mathcal{T}}\int_T (f+\operatorname{div}\lambda_{d,h})(v-\mathcal{J}v)dx+2\alpha\int_\Omega (u_0-u_{d,h})(v-\mathcal{J}v)dx\\ &\qquad+\sum_{E\in\mathcal{E}_\Omega\cup\mathcal{E}_N}\int_E[\lambda_{d,h}\cdot n_E](v-\mathcal{J}v)ds_x\\ &\quad\le C\Bigg(\left(\sum_{T\in\mathcal{T}} h_T^2\|(f+\operatorname{div}\lambda_{d,h}+2\alpha(u_0-u_{d,h}))\|^2_{L^2(T)}\right)^{1/2}\\ &\qquad+\left(\sum_{E\in\mathcal{E}_\Omega\cup\mathcal{E}_N} h_E\|[\lambda_{d,h}\cdot n_E]\|^2_{L^2(E)}\right)^{1/2}\Bigg)\|\nabla v\|.\end{aligned}$$

The combination of the last two estimates, together with the estimates $\|\nabla v\| \le \|\nabla(u-u_{d,h})\| + \|\nabla w\| \le C$, $\min_{w|_{\Gamma_D}=u_D-u_{D,h}} \|w\|_{W^{1,2}(\Omega)} \le C\|h_{\mathcal{E}}^{3/2}\partial_{\mathcal{E}}^2 u_D/\partial s^2\|_{L^2(\Gamma_D)}^2$ (cf. [3, Lemma 3.1]), shows the assertion after absorption of $\|\sigma-\lambda_{d,h}\|$ and $\|u-u_{d,h}\|$. ∎

Remarks. (i) The term $\|h_{\mathcal{E}}^{3/2}\partial_{\mathcal{E}}^2 u_D/\partial s^2\|_{L^2(\Gamma_D)}$ is of higher order.

(ii) The terms $\|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}^2$ and $\|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}$ are of higher order, provided $d \ll h_{\mathcal{T}}$ (cf. Theorem 5.1). It will be shown later in section 6 that the assumption $d \ll h_{\mathcal{T}}$ does not lead to inefficiency of our numerical schemes.

(iii) The a priori error estimate of Theorem 4.6 and the a posteriori error estimate of Theorem 4.8 yield a gap between reliability and efficiency of the error estimates with respect to the discretization parameter h. While the a priori estimate gives optimal convergence results (for smooth solutions) we face a loss of a factor $h_{\mathcal{T}}^{1/2}$ in the a posteriori estimate due to degeneracy of the problem.

Our second error estimate is related to Zienkiewicz–Zhu error estimators (see, e.g., [5]) for elliptic partial differential equations.

THEOREM 4.9. *Assume that DW^{**} satisfies* (4.2) *and $u \in \mathcal{A}$ solves* (P**). *Let $(u_{d,h},\nu_{d,h}) \in \mathcal{B}_{d,h}$ and $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^n$ satisfy the conditions of Lemma* 3.3. *If $\alpha = 0$ and $f \in W^{1,2}(\Omega)$, then*

$$\begin{aligned}\|\sigma-\lambda_{d,h}\|^2 \le C\Big\{ &\min_{\tau_h\in\mathcal{S}_N^1(\mathcal{T},g)} \|\lambda_{d,h}-\tau_h\| + \|h_{\mathcal{T}}^2\nabla f\| + \|h_{\mathcal{E}}^{3/2}\partial_{\mathcal{E}}^2 u_D/\partial s^2\|_{L^2(\Gamma_D)} \\ &+ \big\|h_{\mathcal{E}}^{3/2}\partial_{\mathcal{E}} g/\partial s\big\|_{L^2(\Gamma_N)} + \|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}\Big\} \\ &+ |\Omega|\|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}^2.\end{aligned}$$

Proof. As in the proof of Theorem 4.6 we have, for $w \in W^{1,2}(\Omega)$ with $w|_{\Gamma_D} = u_D - u_{D,h}$ and $v_h \in \mathcal{S}_D^1(\mathcal{T})$,

$$\begin{aligned}C\|\sigma-\lambda_{d,h}\|^2 \le &\int_\Omega (DW^{**}(\nabla u)-\lambda_{d,h})\cdot\nabla(u-u_{d,h}-w-v_h)dx \\ &+ |\Omega|\|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}^2 \\ &+ C|\Omega|\|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})} + \|\sigma-\lambda_{d,h}\|\|\nabla w\|.\end{aligned}$$

Letting $\tau_h \in \mathcal{S}_N^1(\mathcal{T},g)$ and writing $v := u-u_{d,h}-w \in W_D^{1,2}(\Omega)$ and $v_h := \mathcal{J}v \in \mathcal{S}_D^1(\mathcal{T})$ we verify, using $\operatorname{div}\lambda_{d,h}|_T = 0$, the Euler–Lagrange equation (4.1), an integration by parts, and Hölder's inequality,

$$\begin{aligned}&\int_\Omega (DW^{**}(\nabla u)-\lambda_{d,h})\cdot\nabla(v-\mathcal{J}v)dx \\ &\quad\le \int_\Omega f(v-\mathcal{J}v)dx + \sum_{T\in\mathcal{T}}\int_T \operatorname{div}(\tau_h-\lambda_{d,h})(v-\mathcal{J}v)dx \\ &\qquad + \int_{\Gamma_N}(g-\tau_h\cdot n_{\Gamma_N})(v-\mathcal{J}v)ds_x + \|\tau_h-\lambda_{d,h}\|\|\nabla(v-\mathcal{J}v)\|.\end{aligned}$$

The estimate (cf. [5, Theorem 2.1]) $\int_\Omega f(v-\mathcal{J}v)dx \le C\|\nabla v\|\,\|h_{\mathcal{T}}^2\nabla f\|$ and (4.6) yield

$$\begin{aligned}\int_\Omega &(DW^{**}(\nabla u)-\lambda_{d,h})\cdot\nabla(v-\mathcal{J}v)dx\\ &\le C\Bigg\{\|h_{\mathcal{T}}^2\nabla f\|+\bigg(\sum_{T\in\mathcal{T}}h_T^2\|\operatorname{div}(\tau_h-\lambda_{d,h})\|_{L^2(T)}^2\bigg)^{1/2}\\ &\qquad+\big\|h_{\mathcal{E}}^{3/2}\partial_{\mathcal{E}}g/\partial s\big\|_{L^2(\Gamma_N)}+\|\tau_h-\lambda_{d,h}\|\Bigg\}\|\nabla v\|.\end{aligned}$$

Choosing w as in [3] and employing elementary results about nodal interpolation on Γ_N we infer

$$\|w\|_{W^{1,2}(\Omega)}+\|g-\tau_h\cdot n_{\Gamma_N}\|_{L^2(\Gamma_N)}\le C\Big(\big\|h_{\mathcal{E}}^{3/2}\partial_{\mathcal{E}}^2u_D/\partial s^2\big\|_{L^2(\Gamma_D)}+\big\|h_{\mathcal{E}}^{3/2}\partial_{\mathcal{E}}g/\partial s\big\|_{L^2(\Gamma_N)}\Big),$$

and using an elementwise inverse estimate of the form

$$h_T\|\operatorname{div}(\tau_h-\lambda_{d,h})\|_{L^2(T)}\le C\|\tau_h-\lambda_{d,h}\|_{L^2(T)}\quad\forall T\in\mathcal{T}$$

we verify the assertion as in the proof of the preceding theorem. ∎

Remarks. (i) Terms including derivatives of u_D, g, or f are of higher order. Moreover, remarks (ii) and (iii) below Theorem 4.8 are valid here as well.

(ii) Theorem 4.9 shows, up to higher order terms, reliability of the error estimate $\|\lambda_{d,h}-\lambda_{d,h}^*\|$ for any choice of a smooth approximation $\lambda_{d,h}^*\in\mathcal{S}_N^1(\mathcal{T},g)$ to $\lambda_{d,h}$.

(iii) A triangle inequality proves an inverse, efficiency estimate of Theorem 4.9 which holds up to higher order terms, provided σ is smooth but with different exponents,

$$\min_{\tau_h\in\mathcal{S}_N^1(\mathcal{T},g)}\|\lambda_{d,h}-\tau_h\|\le\|\sigma-\lambda_{d,h}\|+\min_{\tau_h\in\mathcal{S}_N^1(\mathcal{T},g)}\|\sigma-\tau_h\|.$$

This efficiency estimate can be made rigorous but then without explicit constants.

5. Convergence of other quantities. In this section we present an estimate for $DW^{**}-\partial W_d^{cx}$ and results concerning the convergence behavior of other quantities such as Young measure support and microstructure region in a three-well problem. Ideas behind the proofs are adapted from [7, 14].

5.1. Approximation of DW^{**}. We first state an approximation result for W^{**}.

THEOREM 5.1. *For $W:\mathbb{R}^2\to\mathbb{R}$, $s\mapsto\min_{j=0,1,2}|s-s_j|^2$ with $s_0=(0,0)$, $s_1=(1,0)$, and $s_2=(0,1)$ and $\omega=(-m,m)^2$, $m\ge1$, there exists a triangulation τ of ω with maximal mesh-size $d=1/k$, k a positive integer, of ω such that*

$$\|\partial W_d^{cx}-DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^n})}\le Cd\,\|D^2W^{**}\|_{L^\infty(\omega)}.$$

Moreover, the mapping DW^{**} satisfies (4.2).

Proof. A careful analysis shows that $W^{**}\in C^1(\mathbb{R}^n)$ satisfies (4.2) and is for $F=(f_1,f_2)\in\mathbb{R}^2$ given by

$$W^{**}(F) = \begin{cases} 0, & F \in X_I, \\ W(F), & F \in X_{II} \cup X_{III} \cup X_{IV}, \\ f_2^2, & F \in X_V, \\ f_1^2, & F \in X_{VI}, \\ \frac{1}{2}(f_1+f_2-1)^2, & F \in X_{VII}. \end{cases}$$

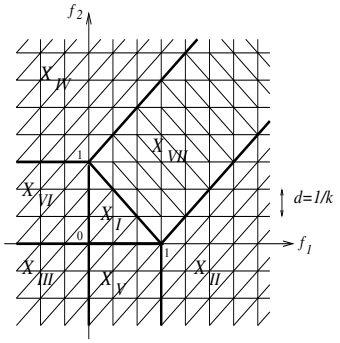


FIG. 1. *W^{**} and the triangulation of $\omega \subseteq \mathbb{R}^2$ to resolve the discontinuities of D^2W^{**}.*

For $d = 1/k$, k a positive integer, choose τ as in Figure 1. Since W_d^{cx} is affine on each $t \in \tau$ we have $\partial W_d^{cx}(s) = \operatorname{conv}\{DW_d^{cx}|_t : t \in \tau, s \in t\}$. Since DW^{**} is continuous and τ-elementwise differentiable it therefore suffices to show for each $t \in \tau$

$$\|DW_d^{cx} - DW^{**}\|_{L^\infty(t)} \le d\|D^2W^{**}\|_{L^\infty(t)}.$$

Letting $W_d^{**} = P_\tau W^{**}$ denote the nodal interpolant of W^{**} we have by standard interpolation results

$$\begin{aligned} \|DW_d^{cx} - DW^{**}\|_{L^\infty(t)} &\le \|DW_d^{cx} - DW_d^{**}\|_{L^\infty(t)} + \|DW_d^{**} - DW^{**}\|_{L^\infty(t)} \\ &\le \|DW_d^{cx} - DW_d^{**}\|_{L^\infty(t)} + Cd\,\|D^2W^{**}\|_{L^\infty(t)}. \end{aligned} \tag{5.1}$$

For each $k \in \tau$ we define an affine function $a_k : \mathbb{R}^2 \to \mathbb{R}$ such that, for all $x \in \mathbb{R}^2$, there holds

$$W_d^{cx}(x) = \sup_{k \in \tau} a_k(x) \tag{5.2}$$

and $W_d^{cx}|_k = a_k$. If $k \subseteq X_I \cup X_{II} \cup X_{III} \cup X_{IV} \cup X_V \cup X_{VI}$ we define a_k such that $a_k(z) = W^{**}(z)$ for all $z \in k \cap \mathcal{N}_\tau$. If $k \subseteq X_{VII}$ and there exists $y = (y_1, y_2) \in k$ with $y_1 + y_2 \in 1 + 2d[j, j+1)$, $j \ge 0$, then we define

$$\begin{aligned} a_k(x) = W(1+jd, jd) &+ (x_1 - 1 - jd, x_2 - jd) \cdot (1,1) \\ &\times \frac{W(1+(j+1)d, (j+1)d) - W(1+jd, jd)}{2d}. \end{aligned}$$

Then $\sup_{k\in\tau} a_k$ is convex, as it is the supreme of countably many affine functions. A proof for (5.2) then follows as above for the convexification of W. Note that W_d^{cx} is mesh dependent. We now prove the remaining estimates. For $k \subset X_I \cup X_{II} \cup \cdots \cup X_{VI}$ we have $DW_d^{cx}|_k = DW_d^{**}|_k$ so that the asserted estimate follows from (5.1). For $k \subseteq X_{VII}$ such that $k \subseteq A_j = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 - x_2 \in [-1, 1], x_1 + x_2 \in 1 + 2d[j, j+1)\}$, $j \ge 0$, there holds $W_d^{cx} = W^{**}$ on ∂A_j, and W_d^{cx} is affine on A_j. Therefore, W_d^{cx} interpolates W^{**} along each line segment in A_j parallel to $(1, 1)$. The estimate

$$\begin{aligned} \|DW_d^{cx} - DW_d^{**}\|_{L^\infty(t)} &\le \|DW_d^{cx} - DW^{**}\|_{L^\infty(t)} + \|DW^{**} - DW_d^{**}\|_{L^\infty(t)} \\ &\le Cd\,\|D^2W^{**}\|_{L^\infty(t)} \end{aligned}$$

follows from the fact that the line segments have a length d. ∎

5.2. Convergence of Young measure support. We employ the following definition in order to prove convergence for the support of the discrete Young measure solution to the support of an exact solution.

DEFINITION 5.2. *For $A, B \subseteq \mathbb{R}^n$ let* $\operatorname{dist}(A,B) := \inf_{(a,b)\in A\times B} |a-b|$. *We write* $\operatorname{Lim\,sup}_{\rho\to\rho_0} A_\rho \subseteq A$ *(i.e., A is the upper Kuratowski limit of A_ρ, cf.* [19, 20]*) if*

$$\forall \varepsilon > 0\, \exists \delta > 0\ \forall \rho, \quad |\rho - \rho_0| \le \delta\ \forall x \in A_\rho, \quad \operatorname{dist}(x, A) \le \varepsilon.$$

THEOREM 5.3. *Let W be as in Theorem* 5.1*, $u \in \mathcal{A}$ a solution for* (P**)*, and $(u_j)_{j>0}$ an infimizing sequence for* (P)*. Let $(u_{d,h}, \nu_{d,h}) \in \mathcal{B}_{d,h}$ and $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^n$ satisfy the conditions of Lemma* 3.3*. Assume that a subsequence of $(u_j)_{j>0}$ converges weakly to u and generates the Young measure ν. Then there exists a mapping $S : \mathbb{R}^2 \to 2^{\mathbb{R}^2}$ such that*

$$\operatorname{dist}(S(\lambda_{d,h}(x)), \operatorname{supp}\nu_x) \to 0$$

if $x \in \Omega$ and $\lambda_{d,h}(x) \to \sigma(x)$. If for all $T \in \mathbb{R}^2$ there holds

$$\begin{aligned}(5.3)\quad &\operatorname{Lim\,sup}_{d\to 0}\{F \in \mathbb{R}^2 : \exists G,\ S \in \mathbb{R}^2,\ \{S,T\} \subseteq \partial W_d^{cx}(G),\ S \in \partial W_d^{cx}(F)\}\\ &\qquad \subseteq \{F \in \mathbb{R}^2 : T = DW^{**}(F)\},\end{aligned}$$

then we also have, if $x \in \Omega$ and $\lambda_{d,h}(x) \to \sigma(x)$,

$$\operatorname{Lim\,sup}_{\lambda_{d,h}(x)\to\sigma(x)}\ \operatorname{conv}\operatorname{supp}\nu_{d,h,x} \subseteq \operatorname{conv} S(\sigma(x)).$$

Remark. If W_d^{cx} is continuously differentiable, then

$$\begin{aligned}&\{F \in \mathbb{R}^2 : \exists G,\ S \in \mathbb{R}^2,\ \{S,T\} \subseteq \partial W_d^{cx}(G),\ S \in \partial W_d^{cx}(F)\}\\ &\quad = \{F \in \mathbb{R}^2 : T = DW_d^{cx}(F)\}.\end{aligned}$$

Proof. Define $\mu : \mathbb{R}^2 \to PM(\mathbb{R}^2)$ by

$$F \mapsto \begin{cases} (1-f_1-f_2)\delta_{(0,0)} + f_1\delta_{(1,0)} + f_2\delta_{(0,1)} & \text{for } F \in X_I,\\ \delta_F & \text{for } F \in X_{II} \cup X_{III} \cup X_{IV},\\ (1-f_1)\delta_{(0,f_2)} + f_1\delta_{(1,f_2)} & \text{for } F \in X_V,\\ (1-f_2)\delta_{(f_1,0)} + f_2\delta_{(f_1,1)} & \text{for } F \in X_{VI},\\ \dfrac{1}{2}(f_1-f_2+1)\delta_{\frac{1}{2}(f_1+f_2+1,f_1+f_2-1)} & \\ \quad +\dfrac{1}{2}(1-f_1+f_2)\delta_{\frac{1}{2}(f_1+f_2-1,f_1+f_2+1)} & \text{for } F \in X_{VII}. \end{cases}$$

Since W^{**} is affine on $\operatorname{conv}\operatorname{supp}\nu_x$, $\int_{\mathbb{R}^2} s\,d\nu_x = \nabla u(x)$, and $\operatorname{supp}\nu_x \subseteq \{E \in \mathbb{R}^2 : W(E) = W^{**}(E)\}$ for almost all $x \in \Omega$ [7, 14], one can show $\nu_x = \mu(\nabla u(x))$ for almost all $x \in \Omega$. For $S : \mathbb{R}^2 \to 2^{\mathbb{R}^2}$ defined by

$$(t_1,t_2) \mapsto \begin{cases} \{(0,0),(1,0),(0,1)\} & \text{for } (t_1,t_2) = (0,0),\\ \{(t_1+2,t_2)/2\} & \text{for } t_1 > 0 \text{ and } t_2 < t_1,\\ \{(t_1,t_2)/2\} & \text{for } t_1 < 0 \text{ and } t_2 < 0,\\ \{(t_1,t_2+2)/2\} & \text{for } t_2 > 0 \text{ and } t_1 < t_2,\\ \{(0,t_2)/2,(2,t_2)/2\} & \text{for } t_1 = 0,\\ \{(t_1,0)/2,(t_1,2)/2\} & \text{for } t_2 = 0,\\ \{(t_1,t_2+2)/2,(t_1+2,t_2)/2\} & \text{for } t_1 = t_2 \text{ and } t_1 > 0, \end{cases}$$

the explicit representation of W^{**} shows

$$\operatorname{supp}\mu(F) = S(DW^{**}(F))$$

so that $\operatorname{supp}\nu_x = S(\sigma(x))$ for a.e. $x \in \Omega$. Hence

$$\operatorname{conv} S(T) = \{E \in \mathbb{R}^2 : T = DW^{**}(E)\}. \tag{5.4}$$

Moreover, for each $\Sigma \in \mathbb{R}^2$ the mapping $\operatorname{dist}(S(\cdot),\Sigma) : \mathbb{R}^2 \to \mathbb{R}$ is continuous, and therefore

$$\operatorname{dist}(S(\lambda_{d,h}(x)), \operatorname{supp}\nu_x) = \operatorname{dist}(S(\lambda_{d,h}(x)), S(\sigma(x))) \to 0$$

if $x \in \Omega$ and $\lambda_{d,h}(x) \to \sigma(x)$. Because of (5.3), (5.4) and since $\operatorname{Lim\,sup}_{\rho\to\rho_0} B_\rho \subseteq A$ if $\operatorname{Lim\,sup}_{\rho\to\rho_0} A_\rho \subseteq A$ and $B_\rho \subseteq A_\rho$ for all ρ, we have to show only that for $T = \lambda_{d,h}(x)$

$$\operatorname{conv}\operatorname{supp}\nu_{d,h,x} \subseteq \{F \in \mathbb{R}^2 : \exists G,\ S \in \mathbb{R}^2,\ \{S,T\} \subseteq \partial W_d^{cx}(G),\ S \in \partial W_d^{cx}(F)\}$$

in order to prove the second assertion. The set

$$M_1 := \{G \in \mathbb{R}^2 : \exists S \in \partial W_d^{cx}(\nabla u_{d,h}),\ S \in \partial W_d^{cx}(G)\}$$

contains each subset $A \subseteq \mathbb{R}^2$ with

$$W_d^{cx} \quad \text{affine on } A \text{ and } \nabla u_{d,h}(x) \in A.$$

Since W_d^{cx} is affine $\operatorname{conv}\operatorname{supp}\nu_{d,h,x}$ and $\nabla u_{d,h}(x) \in \operatorname{conv}\operatorname{supp}\nu_{d,h,x}$ we conclude that $\operatorname{conv}\operatorname{supp}\nu_{d,h,x} \subseteq M_1$. The inclusion $\lambda_{d,h}(x) \in \partial W_d^{cx}(\nabla u_{d,h}(x))$ and the choice $G = \nabla u_{d,h}(x)$ yield

$$M_1 \subseteq \{F \in \mathbb{R}^2 : \exists G,\ S \in \mathbb{R}^2,\ \{S,T\} \subseteq \partial W_d^{cx}(G),\ S \in \partial W_d^{cx}(F)\},$$

which concludes the proof. □

5.3. Convergence of the microstructure region. The microstructure region is that subset of Ω in which the exact Young measure solution is not a single Dirac measure. In this part of Ω infimizing sequences for (P) develop oscillations.

DEFINITION 5.4. *Let $\overline{M}$ denote the closure of $M := \{F \in \mathbb{R}^n : W(F) \neq W^{**}(F)\}$. For a solution $u \in \mathcal{A}$ for the convexified problem* (P**) *and a solution $(u_{d,h},\nu_{d,h}) \in \mathcal{B}_{d,h}$ for* (EP$_{d,h}$) *the* microstructure region $\Omega_{ms} \subseteq \Omega$ *and the* discrete microstructure region $\Omega_{ms,h} \subseteq \Omega$ *are defined by*

$$\Omega_{ms} := \{x \in \Omega : \nabla u(x) \in \overline{M}\} \qquad \textit{and} \qquad \Omega_{ms,h} := \{x \in \Omega : \nabla u_{d,h}(x) \in \overline{M}\},$$

respectively.

The following theorem shows that Ω_{ms} is uniquely defined and that an appropriate approximation $\tilde{\Omega}_{m,h}$ of $\Omega_{ms,h}$ converges to Ω_{ms}.

THEOREM 5.5. *Let W be as in Theorem* 5.1*, and let u solve* (P**)*. There exists a Lipschitz-continuous mapping $\xi : \mathbb{R}^2 \to \mathbb{R}$ such that, for almost all $x \in \Omega$, we have*

$$x \in \Omega_{ms} \Longleftrightarrow \xi(\sigma(x)) = 0.$$

If $v \in \mathcal{A}$ is another solution for (P**)*, then $\xi(DW^{**}(\nabla u)) = \xi(DW^{**}(\nabla v))$ a.e. in Ω. For a solution $(u_{d,h},\nu_{d,h}) \in \mathcal{B}_{d,h}$ for* (EP$_{d,h}$) *with multiplier $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^2$ let*

$$\tilde{\Omega}_{m,h} := \{x \in \Omega : \xi(\lambda_{d,h}(x)) = 0\}.$$

We then have

$$\|\xi(\sigma) - \xi(\lambda_{d,h})\| \le C\|\sigma - \lambda_{d,h}\| \tag{5.5}$$

and

$$x \in \tilde{\Omega}_{m,h} \quad \Longrightarrow \quad \operatorname{dist}(\nabla u_{d,h}(x), \overline{M}) \le C'\|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^2})}.$$

Conversely, there holds

$$x \in \Omega_{ms,h} \quad \Longrightarrow \quad |\xi(\lambda_{d,h}(x))| \le \|\partial W_d^{cx} - DW^{**}\|_{L^\infty(\omega;2^{\mathbb{R}^2})}.$$

Proof. The explicit representation of W^{**} in the proof of Theorem 5.1 shows, for almost all $x \in \Omega$, with $(s_1, s_2) = \sigma(x)$ and $F = \nabla u(x)$

$$\begin{aligned} x \in \Omega_{ms} &\iff F \in \overline{X}_I \cup \overline{X}_V \cup \overline{X}_{VI} \cup \overline{X}_{VII} \\ &\iff (s_1 = 0 \wedge s_2 \le 0) \vee (s_2 = 0 \wedge s_1 \le 0) \vee (s_1 = s_2 \wedge s_1 \ge 0). \end{aligned} \tag{5.6}$$

The mapping $\xi : \mathbb{R}^2 \to \mathbb{R}_{\ge 0}$ defined by

$$(s_1, s_2) \mapsto \min\{|s_1| - \min\{-s_2, 0\}, |s_2| - \min\{-s_1, 0\}, |s_1 - s_2| - \min\{s_1, 0\}\}$$

is Lipschitz continuous with bounded Lipschitz constant $C > 0$ and satisfies because of (5.6) the equivalence

$$\xi(\sigma(x)) = 0 \iff x \in \Omega_{ms}$$

for almost all $x \in \Omega$. Since the quantity $\sigma := DW^{**}(\nabla u)$ is independent of the choice of a solution (cf. Theorem 4.2) we have uniqueness of Ω_{ms}. The Lipschitz continuity of ξ implies the estimate (5.5). Let $x \in \Omega$ be such that $\xi(\lambda_{d,h}(x)) = 0$. The Lipschitz continuity of ξ and the inclusion $\lambda_{d,h}(x) \in \partial W_d^{cx}(\nabla u_{d,h}(x))$ show

$$\begin{aligned} \xi(DW^{**}(\nabla u_{d,h}(x))) &= |\xi(DW^{**}(\nabla u_{d,h}(x))) - \xi(\lambda_{d,h}(x))| \\ &\le C|DW^{**}(\nabla u_{d,h}(x)) - \lambda_{d,h}(x)| \\ &\le C\|DW^{**} - \partial W_d^{cx}\|_{L^\infty(\omega;2^{\mathbb{R}^2})}. \end{aligned}$$

To prove the asserted estimate for $\operatorname{dist}(\nabla u_{d,h}(x), \overline{M})$ it now suffices to prove

$$\operatorname{dist}(F, \overline{M}) \le c\xi(DW^{**}(F))$$

for a constant $c > 0$ and all $F \in \mathbb{R}^2$. The assertion is obvious if $F \in X_I \cup X_V \cup X_{VI} \cup X_{VII}$. We prove the case $F \in X_{II}$; the remaining cases $F \in X_{III}$, X_{IV} follow analogously. Let $F = (f_1, f_2) \in X_{II}$. Then $f_1 - 1 \ge 0$ and $f_1 - 1 \ge f_2$. A short calculation shows $\operatorname{dist}(F, \overline{M}) = \min\{f_1 - 1, (f_1 - f_2 - 1)/\sqrt{2}\}$. Since $DW^{**}(F) = 2(f_1 - 1, f_2)$ we have $\xi(DW^{**}(F)) = 2\min\{f_1 - 1 - \min\{-f_2, 0\}, |f_2| + f_1 - 1, f_1 - f_2 - 1\}$. If $f_2 \le 0$, then this term can be simplified to $\xi(DW^{**}(F)) = \min\{f_1 - 1, f_1 - f_2 - 1\}$, and the assertion follows. If $f_2 \ge 0$ we have

$$\begin{aligned} \xi(DW^{**}(F)) &= \min\{f_1 - 1 + f_2, f_1 - f_2 - 1\} = f_1 - f_2 - 1 \\ &\ge (f_1 - f_2 - 1)/\sqrt{2} = \min\{f_1 - 1, (f_1 - f_2 - 1)/\sqrt{2}\} \\ &= \operatorname{dist}(F, \overline{M}). \end{aligned}$$

To prove the inverse implication let $x \in \Omega_{ms,h}$, i.e., $\nabla u_{d,h}(x) \in X_I \cup X_V \cup X_{VI} \cup X_{VII}$. Since $\lambda_{d,h}(x) \in \partial W_d^{cx}(\nabla u_{d,h}(x))$ and since $\partial W_d^{cx}(\nabla u_{d,h}) = \operatorname{conv}\{DW_d^{cx}|_t : t \in \tau, \nabla u_{d,h}(x) \in t\}$, there exist $t_1, \dots, t_{n+1} \in \tau$ and $\varrho_i \in [0,1]$, $\sum_{i=1}^{n+1} \varrho_i = 1$, such that $\lambda_{d,h}(x) = \sum_{i=1}^{n+1} \varrho_i DW_d^{cx}|_{t_i}$. The identities

$$\lambda_{d,h}(x) = \sum_{i=1}^{n+1} \varrho_i DW_d^{cx}|_{t_i} = \sum_{i=1}^{n+1} \varrho_i (DW_d^{cx}|_{t_i} - DW^{**}(\nabla u_{d,h})) + DW^{**}(\nabla u_{d,h})$$

and $\xi(DW^{**}(\nabla u_{d,h})) = 0$ combined with the Lipschitz continuity of ξ show

$$\begin{aligned}
&|\xi(\lambda_{d,h}(x))| \\
&\quad = \left| \xi\left(\sum_{i=1}^{n+1} \varrho_i (DW_d^{cx}|_{t_i} - DW^{**}(\nabla u_{d,h})) + DW^{**}(\nabla u_{d,h}) \right) - \xi(DW^{**}(\nabla u_{d,h})) \right| \\
&\quad \le C \left| \sum_{i=1}^{n+1} \varrho_i (DW_d^{cx}|_{t_i} - DW^{**}(\nabla u_{d,h})) \right| \le C \|W^{**} - \partial W_d^{cx}\|_{L^\infty(\omega;2^{\mathbb{R}^2})}. \qquad \square
\end{aligned}$$

6. Combination of a multilevel scheme and adaptive mesh refinement. In this section we propose an iterative, adaptive strategy to efficiently approximate the extended problem (EP).

6.1. Active set strategy due to Carstensen and Roubíček. The identity

$$\max_{s\in\omega} \mathcal{H}_{\lambda_{d,h}}(x,s) = \int_{\mathbb{R}^n} \mathcal{H}_{\lambda_{d,h}}(x,s)\, \nu_{d,h,x}(ds)$$

in Lemma 3.3 for a solution $(u_{d,h}, \nu_{d,h}) \in \mathcal{B}_{d,h}$ for $(\mathrm{EP}_{d,h})$ with multiplier $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^n$ states that for almost each $x \in \Omega$ the probability measure $\nu_{d,h,x}$ is supported in those atoms $z \in \mathcal{N}_\tau$ for which $\mathcal{H}_{\lambda_{d,h}}(x,\cdot)$ attains its maximum. Typically, these are only a few atoms.

If the support of the Young measure $\nu_{d,h}$,

$$\operatorname{Supp}(\nu_{d,h}) := \{(x,z) \in \Omega \times \mathcal{N}_\tau : z \in \operatorname{supp}(\nu_{d,h,x})\},$$

where $\operatorname{supp}(\nu_{d,h,x}) \subseteq \mathbb{R}^n$ is the support of the Radon measure $\nu_{d,h,x}$, was known a priori, we could set $A := \operatorname{Supp}(\nu_{d,h})$ and seek $(u_{d,h}, \nu_{d,h})$ as a solution of the following lower-dimensional problem $(\mathrm{EP}_{d,h,A})$:

$$(\mathrm{EP}_{d,h,A}) \quad \begin{cases} \text{Seek } (u_{d,h}, \nu_{d,h}) \in \mathcal{B}_{d,h} \text{ such that } \operatorname{Supp}(\nu_{d,h}) \subseteq A \\ \text{and } \overline{I}(u_{d,h}, \nu_{d,h}) = \inf\limits_{(v_h,\mu_{d,h})\in\mathcal{B}_{d,h}} \overline{I}(v_h, \mu_{d,h}). \end{cases}$$

Proposition 5.4 in [8] gives a necessary condition on A which ensures that $(\mathrm{EP}_{d,h,A})$ is a correct reduction of $(\mathrm{EP}_{d,h})$. Conversely, Lemma 3.4 states a sufficient criterion for a solution of $(\mathrm{EP}_{d,h,A})$ to solve $(\mathrm{EP}_{d,h})$.

Given an approximation $\tilde{h}$ of $\mathcal{H}_{\lambda_{d,h}}$ we define a set of active atoms, called the *active set*, by

$$A = \Big\{(x,z) \in \Omega \times \mathcal{N}_\tau : \tilde{h}(x,z) \ge \max_{s\in\omega} \tilde{h}(x,s) - \varepsilon(x)\Big\}, \tag{6.1}$$

where $\varepsilon \in \mathcal{L}^0(\mathcal{T})$, $\varepsilon > 0$ a.e. in Ω, is a given tolerance. If ε is large enough, then any solution for $(\mathrm{EP}_{d,h,A})$ with A as in (6.1) is a solution for $(\mathrm{EP}_{d,h})$.

LEMMA 6.1. *Let $(u_{d,h}, \nu_{d,h})$ be a solution for $(\mathrm{EP}_{d,h})$ with corresponding multiplier $\lambda_{d,h}$ and $\mathcal{H}_{\lambda_{d,h}}(x,s) = \lambda_{d,h}(x)\cdot s - P_\tau W(s)$. Moreover, let $\tilde{h} : \Omega \times \mathbb{R}^n \to \mathbb{R}$ and $\varepsilon \in \mathcal{L}^0(\mathcal{T})$, $\varepsilon > 0$ a.e. in Ω, be such that, for each $T \in \mathcal{T}$,*

$$\|\mathcal{H}_{\lambda_{d,h}} - \tilde{h}\|_{L^\infty(T\times S_T)} \le \varepsilon|_T,$$

with $S_T \subseteq \mathbb{R}^n$ such that, for almost all $x \in T$, we have

$$\Big\{s \in \omega : \mathcal{H}_{\lambda_{d,h}}(x,s) = \max_{\tilde{s}\in\overline{\omega}} \mathcal{H}_{\lambda_{d,h}}(x,\tilde{s})\Big\} \cup \Big\{s \in \omega : \tilde{h}(x,s) = \max_{\tilde{s}\in\overline{\omega}} \tilde{h}(x,\tilde{s})\Big\} \subseteq S_T.$$

If A is defined by (6.1)*, then any solution for $(\mathrm{EP}_{d,h,A})$ is a solution for $(\mathrm{EP}_{d,h})$.*

Proof. The proof follows the arguments of [8]. □

The idea to guess the support of a Young measure solution in a multilevel scheme, together with Lemma 6.1, motivates the following algorithm, in which a sequence of refining triangulations, elementwise constant tolerances, and an initial guess $\tilde{h}_0$ for $\mathcal{H}_{\lambda_{d,h}}$, e.g., $\tilde{h}_0 = 0$, are given. Figure 2 includes a flow chart of the algorithm.

Algorithm ($A^{\textit{active set}}$). Let $\tau_1, \tau_2, \dots, \tau_J$ be triangulations of ω, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_J > 0$ be elementwise constant, and $\tilde{h}_0 \in L^1(\Omega; C(\mathbb{R}^n))$.

(1) Set $\varepsilon := \varepsilon_1$, $\tilde{h} := \tilde{h}_0$, $\tau := \tau_1$, and $j := 1$.

(2) Compute A from (6.1).

(3) Compute a solution $(u_{d,h}, \nu_{d,h}) \in \mathcal{B}_{d,h,A}$ for $(\mathrm{EP}_{d,h,A})$ and the multiplier $\lambda_{d,h} \in \mathcal{L}^0(\mathcal{T})^n$.

(4) If the conditions of Lemma 3.4 are satisfied, then go to (6); otherwise, proceed with (5).

(5) Increase m if necessary. Enlarge ε by $\varepsilon|_T := 2\varepsilon|_T$ if for some $x_T \in T$

$$\max_{z\in\mathcal{N}_\tau} \mathcal{H}_{\lambda_{d,h}}(x_T, z) > \int_{\mathbb{R}^n} \mathcal{H}_{\lambda_{d,h}}(x_T, s)\, \nu_{h,x_T}(ds),$$

and set $\varepsilon|_T := \varepsilon|_T$ otherwise. Go to (2).

(6) If $j < J$ proceed with (7); otherwise, terminate.

(7) Set $j := j + 1$, $\tilde{h}(x,s) := \lambda_{d,h}(x)\cdot s - P_\tau W(s)$, $\varepsilon := \varepsilon_j$, and go to (2).

Remarks. (i) The approximation $\tilde{h}_0$ may initially be chosen as $\tilde{h}_0 = 0$, and then all atoms are activated in (6.1) or $\tilde{h}_0$ is defined through the solution on a coarser triangulation $\mathcal{T}'$.

(ii) Since the tolerance ε is increased successively the optimality conditions of Lemma 6.1 are satisfied after a finite number of iterations.

6.2. Adaptive mesh refinement. Theorems 4.8 and 4.9 allow the introduction of local refinement indicators which may be used for automatic mesh refinement. Let $(u_{d,h}, \nu_{d,h})$ be a solution for $(\mathrm{EP}_{d,h})$ with corresponding multiplier $\lambda_{d,h}$.

Theorem 4.8 motivates the elementwise contributions, for $T \in \mathcal{T}$,

$$\eta_R(T)^2 := h_T^2 \|f + \operatorname{div}\lambda_{d,h} + 2\alpha(u_0 - u_{d,h})\|^2_{L^2(T)} + \sum_{\substack{E\in\mathcal{E}_\Omega\cup\mathcal{E}_N \\ E\subseteq\partial T}} h_E \|[\lambda_{d,h}\cdot n_E]\|^2_{L^2(E)}.$$

In regard to Theorem 4.9 we employ the operator $\bar{\mathcal{A}} : L^2(\Omega)^n \to \mathcal{S}^1_N(\mathcal{T}; g)$ of [5], which is for $\Gamma_N = \emptyset$ and $p \in L^2(\Omega)^n$ given by

$$\bar{\mathcal{A}}p = \sum_{z\in\mathcal{N}} p_z\varphi_z, \quad \text{for } p_z = \int_{\varphi_z>0} p dx \Big/ \int_{\varphi_z>0} 1 dx,$$

to define, for $T \in \mathcal{T}$,

$$\eta_Z(T) := \|\lambda_{d,h} - \bar{\mathcal{A}}\lambda_{d,h}\|_{L^2(T)}.$$

With these definitions we have

$$\|\sigma - \lambda_{d,h}\|^2 \le C\left(\sum_{T\in\mathcal{T}} \eta(T)^2\right)^{1/2} + \text{h.o.t.},$$

where $\eta(T) = \eta_Z(T)$ or $\eta(T) = \eta_R(T)$ and the higher order terms depend on the mesh-size of the triangulation τ which are of higher order, provided $d \ll h_{\mathcal{T}}$, and on the smoothness of given right-hand sides. We set

$$\eta_R := \left(\sum_{T\in\mathcal{T}} \eta_R(T)^2\right)^{1/4} \quad \text{and} \quad \eta_{Z,R} := \left(\sum_{T\in\mathcal{T}} \eta_Z(T)^2\right)^{1/4}.$$

Remark (iii) below Theorem 4.9 states

$$\eta_{Z,E} := \left(\sum_{T\in\mathcal{T}} \eta_Z(T)^2\right)^{1/2} \le \|\sigma - \lambda_{d,h}\| + \text{h.o.t.}$$

The following algorithm generates the triangulations in the numerical examples of the subsequent section. The parameter Θ allows us to use the algorithm for uniform mesh refinement, which corresponds to $\Theta = 0$, and adaptive mesh refinement, where $\Theta = 1/2$. For details on adaptive mesh refinement we refer the reader to [29]. A schematical flow chart for the combination of the active set strategy with the adaptive mesh refinement algorithm is shown in Figure 2.

Algorithm ($A_\Theta^{adaptive}$). (1) Start with a coarse triangulation $\mathcal{T}_1$ of Ω and set $\omega := (-m,m)^n$, $\ell = 1$, and $\tilde{\lambda}_\ell = 0$.

(2) Compute a discrete solution $(u_\ell, \nu_\ell, \lambda_\ell)$ with Algorithm ($A^{active\,set}$) and starting values $\tilde{h}_0(x,s) := \tilde{\lambda}_\ell(x)\cdot s - P_\tau W(s)$, $J = 2$, $d_j = 2^{j-1}/k$, $k = \lfloor 4m\, 2^{-J}\, \mathrm{card}(\mathcal{N}_{\mathcal{T}_\ell})^{3/2n} \rfloor$ ($\lfloor s \rfloor$ is the largest integer $\le s$), $\varepsilon_j := 2^{-\ell-j}\, 10^{-4}$ for $j = 1, \dots, J$ ($\varepsilon_1 := \infty$ if $\ell = 1$ to activate *all* atoms), and a triangulation τ_j of ω with maximal mesh-size d_j.

(3) For each $T \in \mathcal{T}_\ell$ compute refinement indicators $\eta_Z(T)$ and $\eta_R(T)$.

(4) Mark the element T for red-refinement if

$$\eta_R(T) \ge \Theta \max_{T'\in\mathcal{T}_\ell} \eta_R(T').$$

(5) Mark further elements (*red-blue-green-refinement*) to avoid hanging nodes. Terminate if the stopping criterion is satisfied, generate a new triangulation $\mathcal{T}_{\ell+1}$, define $\tilde{\lambda}_{\ell+1} := \lambda_\ell$, increment ℓ, and go to (2) otherwise.

Remarks. (i) We chose k such that $d \propto h^{3/2}$ so that $\|DW^{**} - \partial W_d^{cx}\|_{L^\infty(\Omega;2^{\mathbb{R}^n})}$ is of the same order as the presumed higher order terms involving g and u_D in Theorems 4.8 and 4.9.

(ii) Since $\lambda_\ell \to DW^{**}(\nabla u)$ in $L^2(\Omega)$ for a solution $u \in \mathcal{A}$ for (P**), λ_ℓ is a Cauchy sequence, and therefore λ_ℓ is a good approximation for $\lambda_{\ell+1}$ if ℓ is large enough.

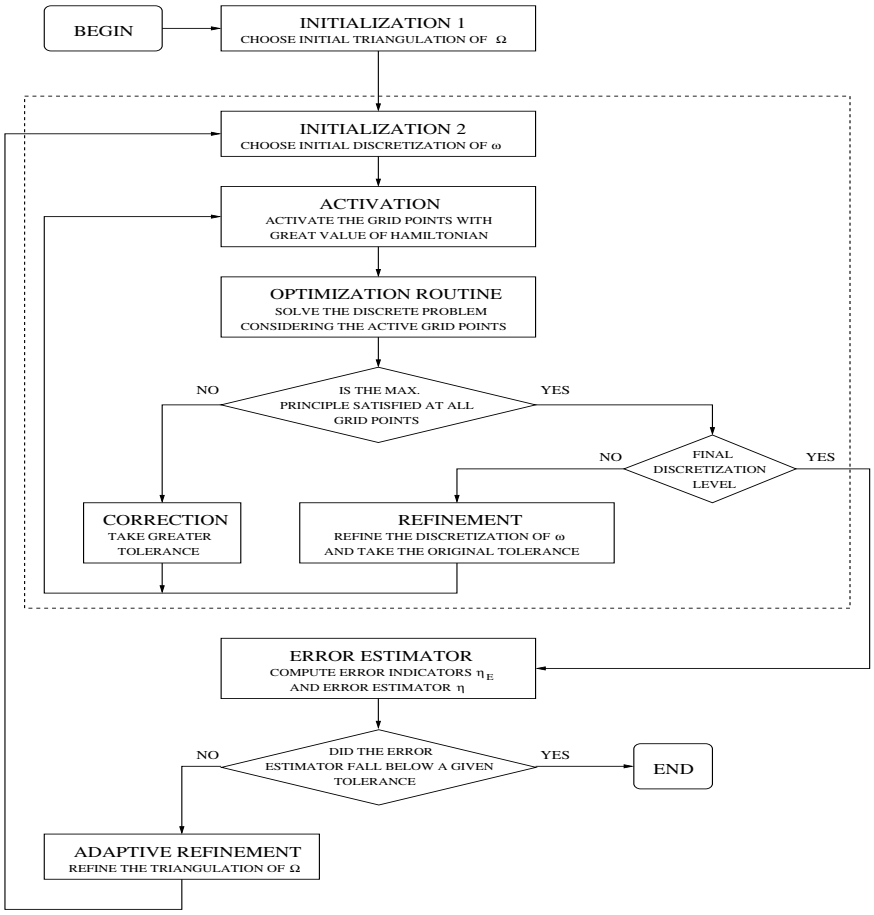


FIG. 2. *Flow chart for the combination of the active set strategy (as in* [8]*, inside the dashed box) with adaptive mesh refinement.*

7. Numerical experiments. In this section we present numerical results for two specifications of (P). The first example has been investigated in [8] and is modified here to obtain quadratic growth conditions. The second example is a two-dimensional problem that reveals limitations of our approach to solve (P) but thereby underlines the necessity of the design of efficient algorithms for the solution for ($\mathrm{EP}_{d,h}$).

The implementation of the algorithms was performed in Matlab as described in [8] for the part concerning the active set strategy. We solved the linear optimization problems with the interior point linear program solver HOPDM [15].

Example 7.1 (one-dimensional two-well problem). Let $n = 1$, $\Omega = (0,1)$, $\Gamma_D = \{0,1\}$, $\alpha = 0$, $\Gamma_N = \emptyset$, and $W(s) = \min\{(s-1)^2, (s+1)^2\}$. The right-hand sides are defined by

$$f(x) = \begin{cases} 0 & \text{for } x \le x_b, \\ \gamma(x - x_b)/2 & \text{for } x \ge x_b \end{cases}$$

and

$$u_D(0) = 3x_b^5/128 + x_b^3/3 \qquad \text{and} \qquad u_D(1) = \gamma(1 - x_b)^3/24 + 1 - x_b,$$

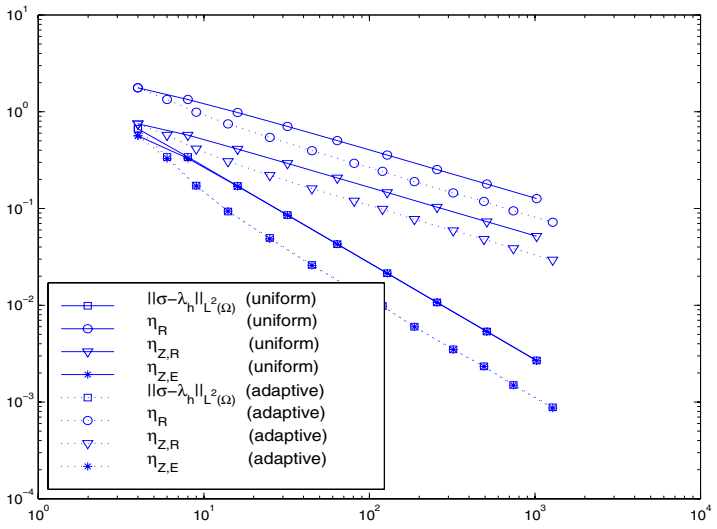


FIG. 3. *Error and error estimators in Example* 7.2 *for uniform and adaptive mesh refinement.*

where $\gamma = 100$ and $x_b = \pi/6$. A solution for (P**) is then given by

$$u(x) = \begin{cases} -3(x-x_b)^5/128 - (x-x_b)^3/3 & \text{for } x \le x_b, \\ \gamma(x-x_b)^3/24 + x - x_b & \text{for } x \ge x_b \end{cases}$$

and allows us to compute the unique quantity $\sigma := DW^{**}(u')$. The microstructure region is $(0, x_b)$ in which $\sigma = 0$ and u' lies between the wells -1 and 1; i.e., $u'(x) \in (-1, 1)$ for $x \in (0, x_b)$. A Young measure corresponding to u is given by

$$\nu_x = \begin{cases} \dfrac{1-u'(x)}{2}\delta_{-1} + \dfrac{1+u'(x)}{2}\delta_{+1} & \text{for } x \le x_b, \\ \delta_{u'(x)} & \text{for } x > x_b. \end{cases}$$

For Algorithm $(A_\Theta^{adaptive})$ we used $m = 4$ and

$$\mathcal{T}_1 = \{[0, 1/4], [1/4, 1/2], [1/2, 3/4], [3/4, 1]\}.$$

Note that the weighted jumps $h_E\|[\lambda_{d,h} \cdot n_E]\|^2_{L^2(E)}$ of $\lambda_{d,h}$ across edges $E \in \mathcal{E}_\Omega$ are in the one-dimensional situation given by

$$\max\{h_{T_1}, h_{T_2}\}(\lambda_{d,h}|_{T_1} - \lambda_{d,h}|_{T_2})^2$$

for $z \in \mathcal{K}$, $T_1, T_2 \in \mathcal{T}$ such that $z = T_1 \cap T_2$.

We ran Algorithm $(A_0^{adaptive})$ and $(A_{1/2}^{adaptive})$ in Example 7.2. The obtained error estimators η_R, $\eta_{Z,R}$, and $\eta_{Z,E}$ and the exact error $\|\sigma - \lambda_{d,h}\|$ for each triangulation are plotted against the degrees of freedom in $\mathcal{T}$ in Figure 3 with a logarithmic scaling used for both axes. Both uniform and adaptive refinement strategies yield the same experimental convergence rates, but the adaptive scheme yields a comparable error reduction at similar numbers of degrees of freedom. The error estimators η_R and $\eta_{Z,R}$ converge much slower than the error itself, while the efficient error estimator $\eta_{Z,E}$ approximates the error very well and converges with the same order.

TABLE 1
Possible and active atoms per element on uniform meshes.

Triangulation	3	4	5	6	7	8	9
# elements	16	32	64	128	256	512	1,024
# atoms	1,122	3,034	8,385	23,443	65,921	185,908	525,057
# active atoms	7.9	10.7	9.6	17.4	46.3	64.5	127.1

TABLE 2
Possible and active atoms per element on adapted meshes.

Triangulation	7	8	9	10	11	12	13
# elements	81	120	187	321	492	741	1,280
# atoms	11,881	21,297	41,244	92,450	175,143	323,390	733,574
# active atoms	7.4	11.6	38.9	31.2	29.6	27.4	30.7

In Tables 1 and 2 we displayed for uniform and adapted meshes, respectively, the number of possible atoms per element and the average number of active atoms per element selected by ($A^{active\ set}$). We observe that the numbers of atoms is significantly reduced by the active set strategy. Moreover, the average number of active atoms seems to be bounded or maybe grows very slowly on the adapted meshes, while on the uniform meshes the number of active atoms grows linearly.

Example 7.2 (two-dimensional, scalar three-well problem). Let $n = 2$, $\Omega = (0,1)^2$, W as in Theorem 5.1, $\alpha = 0$, $\Gamma_D = \partial\Omega$, and, for $(x,y) \in \overline{\Omega}$, $u_D(x,y) = v(x) + v(y)$, where, for $t \in [0,1]$,

$$v(t) = \begin{cases} (t-1/4)^3/6 + (t-1/4)/8 & \text{for } t \le 1/4, \\ -(t-1/4)^5/40 - (t-1/4)^3/8 & \text{for } t \ge 1/4. \end{cases}$$

Setting $f := -\operatorname{div} DW^{**}(\nabla u_D)$, i.e., for $(x,y) \in (0,1)^2$,

$$f(x,y) = \begin{cases} 0 & \text{for } x \le 1/4 \text{ and } y \le 1/4, \\ -2\,v''(y) & \text{for } x \le 1/4 \text{ and } 1/4 \le y, \\ -2\,v''(x) & \text{for } 1/4 \le x \text{ and } y \le 1/4, \\ -2\,(v''(x) + v''(y)) & \text{for } 1/4 \le x \text{ and } 1/4 \le y, \end{cases}$$

we have that $u = u_D$ is the weak limit of an infimizing sequence for (P). If u_x and u_y abbreviate $\partial u/\partial x$ and $\partial u/\partial y$, respectively, then for

$$\nu_{(x,y)} := \begin{cases} (1 - u_x(x,y) - u_y(x,y))\delta_{(0,0)} & \\ \quad + u_x(x,y)\delta_{(1,0)} + u_y(x,y)\delta_{(0,1)} & \text{for } x \le 1/4,\ y \le 1/4, \\ (1 - u_x(x,y))\delta_{(0,u_y(x,y))} + u_x(x,y)\delta_{(1,u_y(x,y))} & \text{for } x \le 1/4,\ 1/4 \le y, \\ (1 - u_y(x,y))\delta_{(u_x(x,y),0)} + u_y(x,y)\delta_{(u_x(x,y),1)} & \text{for } 1/4 \le x,\ y \le 1/4, \\ \delta_{\nabla u(x,y)} & \text{for } 1/4 \le x,\ 1/4 \le y, \end{cases}$$

the pair (u,ν) is a solution for (EP). The coarsest triangulation $\mathcal{T}_1$ consists of 32 triangles, which are halved squares, and we set $m = 1.5$.

Our numerical results in Example 7.2 are not as satisfying as those for Example 7.1. The Lagrange multiplier provided by the linear program solver did not satisfy the optimality conditions even when m was large and all atoms were activated. We suspect

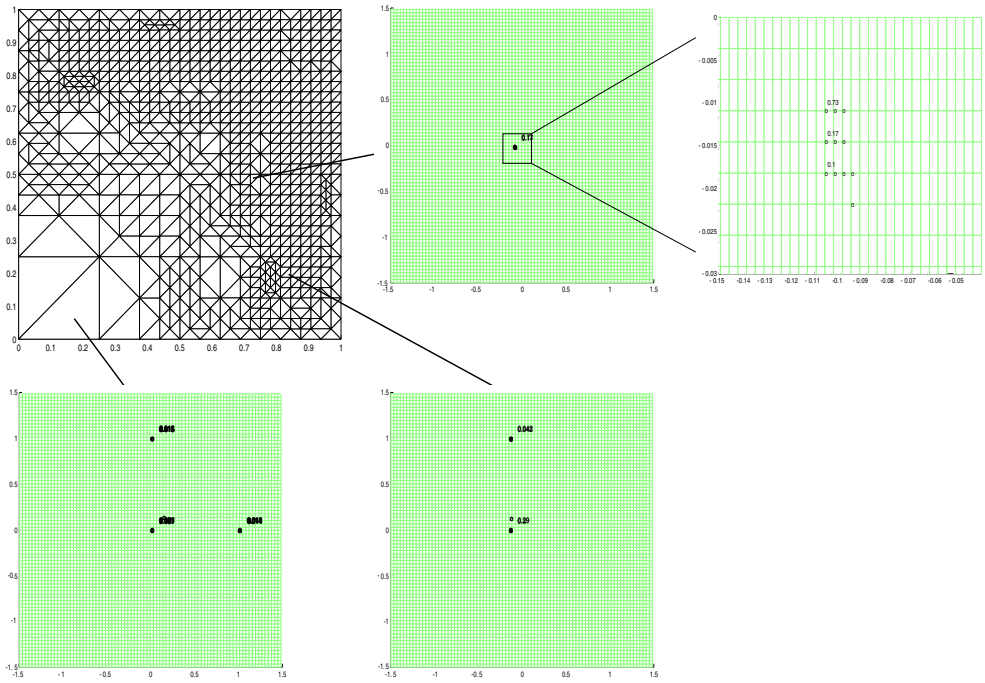


FIG. 4. *Adaptively generated mesh and Young measure restricted to three different elements in Example* 7.1.

TABLE 3
CPU-times for $(\mathrm{EP}_{d,h})$ *on adaptively refined meshes in Example* 7.1.

Dof	9	35	70	162	255	492
CPU-time [s]	11.7	269.0	830.7	5,792.7	9,797.4	24,317.5

that this is caused by the huge complexity of the problem. Other solvers for the linear programming problem did not find a solution when the problem became large. This indicates that efficient methods for the solution of $(\mathrm{EP}_{d,h})$ are very important. We found, however, that the quantity $DW^{**}(\nabla u_{d,h})$ satisfied the maximum principle and the equilibrium equation up to an absolute error of about 0.05 in Example 7.2 so that we used this quantity to activate atoms in Algorithm $(A^{active\,set})$ and to calculate error indicators η_R, $\eta_{Z,R}$, and $\eta_{Z,E}$ in order to refine the mesh and to estimate the error in Algorithm $(A^{adaptive}_{1/2})$.

Figure 4 shows the adaptively generated mesh $\mathcal{T}_6$ and the support of the discrete Young measure solution and the corresponding volume fractions restricted to three different elements. The three meshes show every tenth atom in τ, and circles indicate that an atom is active. Numbers next to circles are volume fractions, provided they are larger than 0.01. We observe that the discrete Young measure approximates the Young measure solution ν from Example 7.2 very well. Moreover, the adaptive algorithm refines the mesh in those regions where the stress is large. Since the error estimators and the active set strategy show the same behavior as in the previous example we omit the corresponding plots and tables here.

Table 3 displays the CPU-time needed to solve $(\mathrm{EP}_{d,h})$ in Example 7.2 on a

sequence of adaptively refined triangulations against the number of degrees of freedom in $\mathcal{T}_k$, $k = 1, \ldots, 6$. The numerical solutions were obtained on a SUN Enterprise with 14 processors and 14 GB RAM, and the numbers suggest that the CPU-time depends linearly on the number of degrees of freedom.

Acknowledgment. The author wishes to thank Professor C. Carstensen for stimulating discussions.

REFERENCES

[1] J.M. Ball, *A version of the fundamental theorem for Young measures*, in Partial Differential Equations and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Phys. 344, Springer, Berlin, 1989, pp. 207–215.

[2] J.M. Ball, P.J. Holmes, R.D. James, R.L. Pego, and P.J. Swart, *On the dynamics of fine structure*, J. Nonlinear Sci., 1 (1991), pp. 17–70.

[3] S. Bartels, C. Carstensen, and G. Dolzmann, *Inhomogeneous Dirichlet Conditions in A Priori and A Posteriori Finite Element Error Analysis*, Berichtsreihe des Mathematischen Seminars Kiel, Technical report, 02-1, Christian–Albrechts–Universität zu Kiel, Kiel, Germany, 2001.

[4] C. Carstensen, *Quasi interpolation and a posteriori error analysis in finite element method*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1187–1202.

[5] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. part* I: *Low order conforming, nonconforming, and mixed FEM*, Math. Comp., 71 (2002), pp. 945–969.

[6] C. Carstensen and S. Müller, *Local stress regularity in scalar nonconvex variational problems*, SIAM J. Math. Anal., 34 (2002), pp. 495–509.

[7] C. Carstensen and P. Plecháč, *Numerical solution of the scalar double-well problem allowing microstructure*, Math. Comp., 66 (1997), pp. 997–1026.

[8] C. Carstensen and T. Roubíček, *Numerical approximation of Young measures in non-convex variational problems*, Numer. Math., 84 (2000), pp. 395–414.

[9] M. Chipot and S. Müller, *Sharp energy estimates for finite element approximations of nonconvex problems*, in Variations of Domain and Free-Boundary Problems in Solid Mechanics, Solid Mech. Appl., 66, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997, pp. 317–325.

[10] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[11] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[12] C. Collins and M. Luskin, *Optimal order error estimates for the finite element approximation of the solution of a non-convex variational problem*, Math. Comp., 57 (1991), pp. 621–637.

[13] J.L. Ericksen, *Constitutive theory for some constrained elastic crystals*, J. Solids and Structures, 22 (1986), pp. 951–964.

[14] G. Friesecke, *A necessary and sufficient condition for non-attainment and formation of microstructure almost everywhere in scalar variational problems*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 437–471.

[15] J. Gondzio, *HOPDM (version* 2.12*) - A fast LP solver based on a primal-dual interior point method*, European J. Oper. Res., 85 (1995), pp. 221–225.

[16] D. Kinderlehrer and P. Pedregal, *Weak convergence of integrands and the Young measure representation*, SIAM J. Math. Anal., 23 (1991), pp. 1–19.

[17] M. Kružík, *Numerical approach to double well problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1833–1849.

[18] M. Kružík and A. Prohl, *Young measure approximation in micromagnetism*, Numer. Math., 90 (1998), pp. 291–307.

[19] K. Kuratowski, *Topology* I, Academic Press, New York, PWN, Warszawa, 1966.

[20] K. Kuratowski, *Topology* II, Academic Press, New York, PWN, Warszawa, 1968.

[21] M. Luskin, *On the computation of crystalline microstructure*, Acta Numerica, 5 (1996), pp. 191–257.

[22] A.-M. Matache, T. Roubíček, and C. Schwab, *Higher-order convex approximations of Young measures in optimal control*, Adv. Comput. Math., 19 (2003), pp. 73–97.

[23] S. Müller, *Variational models for microstructure and phase transitions*, in Calculus of Variations and Geometric Evolution Problems, Lecture Notes in Math. 1713, S. Hildebrandt and M. Struwer, eds., Springer, Berlin, 1999, pp. 85–210.
[24] R.-A. Nicolaides and N.J. Walkington, *Computation of microstructure utilizing Young measure representations*, in Transactions of the Tenth Army Conference on Applied Mathematics and Computing, U.S. Army Research Office, Research Triangle Park, NC, 1993, pp. 57–68.
[25] P. Pedregal, *Numerical approximation of parametrized measures*, Numer. Funct. Anal. Optim., 16 (1995), pp. 1049–1066.
[26] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser, Basel, 1997.
[27] T. Roubíček, *Relaxation in Optimization Theory and Variational Calculus*, De Gruyter, New York, 1997.
[28] T. Roubíček and M. Kružík, *Adaptive approximation algorithm for relaxed optimization problems*, in Fast Solution of Discretized Optimization Problems, K.-H. Hoffmann, R. H. W. Hoppe, and V. Schulz, eds., Birkhäuser, Basel, 2001, pp. 242–254.
[29] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Stuttgart, 1996.
[30] L.C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, Comptes Rendues de la Société des Sciences et des Lettres de Varsovie, classe III, 30 (1937), pp. 212–234.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 531–552

ON THE NUMERICAL INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS BY PROCESSED METHODS*

S. BLANES†, F. CASAS†, AND A. MURUA‡

Abstract. We provide a theoretical analysis of the processing technique for the numerical integration of ODEs. We get the effective order conditions for processed methods in a general setting so that the results obtained can be applied to different types of numerical integrators. We also propose a procedure to approximate the postprocessor such that its evaluation is virtually cost-free. The analysis is illustrated for a particular class of composition methods.

Key words. effective order, processing technique, cheap postprocessor, initial value problems

AMS subject classifications. 65L05, 65L70, 22E60

DOI. 10.1137/S0036142902417029

1. Introduction.

Given the ODE

$$x' = f(x), \qquad x_0 = x(t_0) \in \mathbb{R}^D, \tag{1.1}$$

with $f : \mathbb{R}^D \to \mathbb{R}^D$ and associated vector field (or Lie operator associated with f)

$$F = \sum_{i=1}^{D} f_i(x) \frac{\partial}{\partial x_i}, \tag{1.2}$$

a one-step numerical *integrator* for a time step h, $\psi_h : \mathbb{R}^D \to \mathbb{R}^D$, can be seen as a smooth family of maps with parameter h such that ψ_0 is the identity map. The integrator ψ_h is said to have order of consistency $\geq q$ (or, equivalently, to be of order $\geq q$) if

$$\psi_h = \varphi_h + \mathcal{O}(h^{q+1}), \tag{1.3}$$

where φ_h is the h-flow of the ODE (1.1). Then an approximation to the exact solution $x(h)$ is given by

$$x_h = \psi_h(x_0) = \varphi_h(x_0) + \delta_{h,q}(x_0),$$

where $\delta_{h,q}(x_0) = \mathcal{O}(h^{q+1})$ denotes the local truncation error. The efficiency of the integrator (when compared with methods of the same order and family) depends both on its computational cost and the magnitude of the error term.

In this work we discuss the class of methods obtained by enhancing an integrator ψ_h with processing. The idea of processing can be traced back to the work of Butcher

*Received by the editors October 31, 2002; accepted for publication (in revised form) September 8, 2003; published electronically April 14, 2004.
http://www.siam.org/journals/sinum/42-2/41702.html

†Departament de Matemàtiques, Universitat Jaume I, 12071 Castellón, Spain (sblanes@mat.uji.es, Fernando.Casas@uji.es). The work of these authors was partially supported by Fundació Caixa Castelló–Bancaixa. The work of the first author was also supported by Ministerio de Ciencia y Tecnología (Spain) through a contract in the Pogramme Ramón y Cajal 2001 and by the TMR programme through grant EC-12334303730.

‡Konputazio Zientziak eta A. A. saila, Informatika Fakultatea, EHU/UPV, Donostia/San Sebastián, Spain (ander@si.ehu.es).

[7] in 1969, where it is considered in the context of Runge–Kutta methods, and is summarized in [12, 19]. Essentially, it consists of obtaining a new (hopefully better) integrator of the form

$$\hat{\psi}_h = \pi_h \circ \psi_h \circ \pi_h^{-1}. \tag{1.4}$$

The method ψ_h is referred to as the *kernel* and the parametric map $\pi_h : \mathbb{R}^D \to \mathbb{R}^D$ as the *postprocessor* or *corrector*. Application of n steps of the integrator $\hat{\psi}_h$ leads to

$$\hat{\psi}_h^n = \pi_h \circ \psi_h^n \circ \pi_h^{-1},$$

which can be considered as a change of coordinates in phase space. Thus, it is not required that the kernel ψ_h used to propagate the numerical solution be a *good* integrator. It is sufficient, using dynamical system terminology, that ψ_h be *conjugate* to a good integrator.

Usually one is interested in the case where $\pi_0 = \mathrm{id}$, the identity map; i.e., π_h is also a near-identity map, although it is not intended to approximate the h-flow φ_h. The *preprocessor* π_h^{-1} is applied only once so that its computational cost may be ignored; then the kernel ψ_h acts once per step, and, finally, the action of the postprocessor π_h is evaluated only when output is required. Processing is advantageous if $\hat{\psi}_h$ is a more accurate method than ψ_h and the cost of π_h is negligible: it provides the accuracy of $\hat{\psi}_h$ at the cost of the less accurate method ψ_h.

Although initially intended for Runge–Kutta methods, the processing technique did not become significant in practice, probably due to the difficulties of coupling processing with classical strategies of variable step-sizes. It has been only recently that this idea has proved its usefulness in the context of geometric integration, where constant step-sizes are widely employed.

The aim of geometric integration is to construct numerical schemes for discretizing the differential equation (1.1) while preserving certain geometric properties of the vector field F. It is generally recognized that this class of numerical algorithms (the so-called *geometric integrators*) provide a better description of the system (1.1) than standard methods, both with respect to the preservation of invariants and also in the accumulation of numerical errors along the evolution [12, 22].

A typical procedure in geometric integration is to consider one or more low order methods and compose them with appropriately chosen weights to achieve higher order schemes. The resulting composition method inherits the relevant properties that the basic integrator shares with the exact solution, provided these properties are preserved by composition [16].

It has been precisely in this context where the application of processing has proved to be a very powerful tool, allowing one to build numerical schemes with both the kernel and the postprocessor taken as compositions of basic integrators. In particular, highly efficient processed composition methods have been proposed in the last few years, both in the separable case [3] (including families of Runge–Kutta–Nyström class of methods [5, 14, 15]) and also for slightly perturbed systems [4, 17, 24].

The method ψ_h is of *effective order* q if a postprocessor π_h exists for which $\hat{\psi}_h$ is of (conventional) order q [7], that is,

$$\pi_h \circ \psi_h \circ \pi_h^{-1} = \varphi_h + \mathcal{O}(h^{q+1}). \tag{1.5}$$

When analyzing the order conditions $\hat{\psi}_h$ has to verify to be a method of order q, it has been shown that many of them can be satisfied by using π_h [1, 3, 8] so that ψ_h must

fulfill a much reduced set of constraints. Furthermore, the error term $\delta_{h,q}(x_0)$ depends on both ψ_h and π_h, and additional conditions can be imposed on the postprocessor in order to reduce its magnitude. This allows one, on the one hand, to consider kernels involving fewer evaluations and, on the other hand, to analyze and obtain new and efficient composition methods of high order [5].

In this paper we develop a general theory of the processing technique as applied to the numerical integration of differential equations and derive, under very general assumptions, the conditions to be satisfied by the kernel and the postprocessor to attain a given order of consistency. The analysis can be directly applied to different types of numerical methods, including families of composition integrators and Runge–Kutta-type methods.

For processed methods whose postprocessor is itself constructed as a composition of basic integrators, it turns out that the computational cost of evaluating π_h is usually higher than of ψ_h so that their use is restricted (in sequential computer environments) to situations where intermediate results are not frequently required. Otherwise the overall efficiency of the methods is highly deteriorated.

Another goal of this work is precisely to show how to avoid this situation, i.e., how to obtain approximations to the postprocessor virtually cost-free and without loss of accuracy. The key point is a generalization of a procedure outlined in [14]: π_h is replaced by a new integrator $\hat{\pi}_h \simeq \pi_h$ obtained from the intermediate stages in the computation of ψ_h.

The plan of the paper is as follows. In section 2 we provide a general analysis of processed methods, obtaining the order conditions to be verified by the kernel and the postprocessor. In section 3 we propose a cheap alternative for approximating the postprocessor, study the corresponding order conditions, and examine the propagation of the error that results from replacing the optimal postprocessor by the cheap alternative. Section 4 is concerned with numerical examples, and section 5 contains some concluding remarks.

2. Analysis of processed methods.

2.1. Order of consistency of numerical integrators. Let ψ_h be an integrator that approximates the h-flow φ_h of the system (1.1). It is well known that, for each $g \in C^\infty(\mathbb{R}^D, \mathbb{R})$ (i.e, each infinitely differentiable map $g : \mathbb{R}^D \to \mathbb{R}$), $g(\varphi_h(x))$ admits an expansion of the form [21]

$$g(\varphi_h(x)) = \exp(hF)[g](x) = g(x) + \sum_{k\geq 1} \frac{h^k}{k!} F^k[g](x), \quad x \in \mathbb{R}^D,$$

where F is the vector field (1.2). Let us assume that, for each $g \in C^\infty(\mathbb{R}^D, \mathbb{R})$, $g(\psi_h(x))$ admits an expansion of the form

$$g(\psi_h(x)) = g(x) + h\Psi_1[g](x) + h^2\Psi_2[g](x) + \cdots,$$

where each Ψ_k is a linear differential operator, and let Ψ_h denote the series of differential operators

$$\Psi_h = I + \sum_{k\geq 1} h^k \Psi_k$$

so that, formally, $g \circ \psi_h = \Psi_h[g]$. Clearly, (1.3) is then equivalent to

$$\Psi_k = \frac{1}{k!} F^k, \quad 1 \leq k \leq q. \tag{2.1}$$

Alternatively, let us consider the series

$$F_h = \log(\Psi_h) = \sum_{m\geq 1} \frac{(-1)^{m+1}}{m}(\Psi_h - I)^m$$

so that, formally, $\Psi_h = \exp(F_h)$ and

$$F_h = \sum_{k\geq 1} h^k F_k, \quad \text{with } F_k = \sum_{m\geq 1} \frac{(-1)^{m+1}}{m} \sum_{j_1+\cdots+j_m=k} \Psi_{j_1}\cdots\Psi_{j_m}. \tag{2.2}$$

It can be shown that the algebraic properties of the linear differential operators Ψ_k imply that such F_h is a series of vector fields. This means the well-known fact that the integrator ψ_h can be formally interpreted as the exact 1-flow of the modified vector field F_h [12]. Then condition (2.1) is equivalent to

$$F_1 = F, \qquad F_k = 0 \quad \text{for } 2 \leq k \leq q. \tag{2.3}$$

It is worth noticing that characterizations (2.1) and (2.3) for the order conditions of the integrator ψ_h are written, in contrast with (1.3), in a way that it is straightforward to extend them to integrators on smooth manifolds so that we need not restrict ourselves to integrators on $\mathbb{R}^D$. In fact, the theory of the present paper remains true in a coordinate-free setting, where F_k are vector fields (sections of the tangent bundle) on a finite-dimensional smooth manifold.

2.2. Graded Lie algebra of vector fields. We have observed that numerical integrators can be expanded as exponentials of series of vector fields, and these can be used to compare with the exact flow of the system to be integrated numerically. In section 3 we will consider expansions of linear combinations of vector fields, which lie in the associative algebra $\mathcal{B}$ of linear differential operators generated by concatenation of smooth vector fields on $\mathbb{R}^D$, with the identity operator I as the unit element. At this point it seems appropriate to briefly review the main concepts of the theory of Lie algebras in this setting. They will prove to be very useful in the subsequent analysis.

As any associative algebra, the algebra $\mathcal{B}$ has structure of Lie algebra with the commutator $[a, b] = ab - ba$ as the Lie bracket. In other words, the commutator $[a, b]$ is a bilinear operator satisfying

- skew-symmetry: $[a, b] = -[b, a]$;
- the Jacobi identity: $[a, [b, c]] + [b, [c, a]] + [c, [a, b]] = 0$.

The vector fields on $\mathbb{R}^D$ form a subspace of $\mathcal{B}$ that is closed under commutation; i.e., $[F, G]$ is a smooth vector field, provided that both F, G are also smooth vector fields.

From (2.2) one has $F_h = hF_1 + h^2F_2 + h^3F_3 + \cdots$, where each F_k is a vector field and h is a symbol that corresponds to the parameter present in the definition of the integrator ψ_h. The set of series of this form inherits a Lie algebra structure from the Lie algebra structure of the set of vector fields if there is a sequence of vector subspaces $\mathcal{L}_k$, $k \geq 1$, of the Lie algebra of vector fields such that $F_k \in \mathcal{L}_k$ for each series $\sum_{k\geq 1} h^k F_k$ and

$$[\mathcal{L}_n, \mathcal{L}_m] \subset \mathcal{L}_{n+m} \quad \text{for each } n, m \geq 1. \tag{2.4}$$

In this way the concept of *graded Lie algebra* naturally arises. A graded Lie algebra $\mathcal{L}$ can be defined as a Lie algebra together with a sequence of subspaces $\{\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \ldots\}$ of $\mathcal{L}$ such that $\mathcal{L} = \bigoplus_{k\geq 1} \mathcal{L}_k$ and (2.4) holds. The vector spaces $\mathcal{L}_k$ in the graded Lie algebra $\mathcal{L}$ are called *homogeneous components* of $\mathcal{L}$.

Also the notion of *free Lie algebra* is very useful in this setting [23]. Roughly speaking, a Lie algebra $\mathcal{L}$ is free if there exists a set $S \subset \mathcal{L}$ such that (i) any element in $\mathcal{L}$ can be written as a linear combination of nested brackets of elements in S and (ii) the only linear dependencies among such nested brackets are due to the skew-symmetry property and the Jacobi identity of brackets (see [20] for more details on the theory of free Lie algebras in the context of numerical integration).

Given a Lie algebra $\mathcal{L}$ of vector fields one may consider the associative algebra generated by $\mathcal{L}$ (which is a subalgebra of $\mathcal{B}$). There exists an associative algebra $\mathcal{A} = U(\mathcal{L})$, called the *universal enveloping algebra* [23] of the Lie algebra $\mathcal{L}$ and a unique algebra homomorphism σ of $\mathcal{A}$ onto the algebra of linear differential operators generated by the vector fields in $\mathcal{L}$. That is, any such linear differential operator can be represented as an element of $\mathcal{A}$.

The Poincaré–Birkhoff–Witt theorem [23] allows one to construct a basis of the universal enveloping algebra $\mathcal{A}$ of $\mathcal{L}$ in terms of a basis of $\mathcal{L}$. More specifically, if $\{L_i\}$ denotes a basis of $\mathcal{L}$, each element of the basis of $\mathcal{A}$ is associated with a family $\{L_{i_1}, \dots, L_{i_k}\}$ of (possibly repeated) elements of the basis of $\mathcal{L}$, and it is the sum of all possible concatenations of basic vector fields $L_{j_1} \cdots L_{j_k}$ such that $(j_1, \dots, j_k)$ is obtained by reordering $(i_1, \dots, i_k)$. When $\mathcal{L}$ is a graded Lie algebra $\mathcal{L} = \bigoplus_{k\geq 1} \mathcal{L}_k$, then $\mathcal{A}$ also admits a graded associative algebra structure, with $\mathcal{A} = \bigoplus_{k\geq 0} \mathcal{A}_k$, where $\mathcal{A}_0 = \operatorname{span}(I)$ (that is, $\mathcal{A}_n \mathcal{A}_m \subset \mathcal{A}_{n+m}$). Given a basis $\{E_{k,j}\}_{j=1}^{n_k}$ in $\mathcal{L}_k$ for each $k \geq 1$ with $n_k = \dim \mathcal{L}_k$, this procedure leads to a basis $\{D_{k,j}\}_{j=1}^{m_k}$ in $\mathcal{A}_k$ for $k \geq 1$ with $m_k = \dim \mathcal{A}_k$. In particular, this allows one to obtain m_k in terms of the dimensions $n_1, \dots, n_k$.

2.3. Effective order conditions. Let us consider now a mapping π_h close to the identity as a postprocessor for the integrator ψ_h. Our aim is to obtain characterizations for the order of consistency of the resulting processed integrator (1.4).

As before, let

$$\Pi_h = I + \sum_{k\geq 1} h^k \Pi_k, \qquad \hat{\Psi}_h = I + \sum_{k\geq 1} h^k \hat{\Psi}_k$$

be the series of differential operators such that, formally, $g \circ \pi_h = \Pi_h[g]$ and $g \circ \hat{\psi}_h = \hat{\Psi}_h[g]$, respectively. Then $\hat{\Psi}_h = \Pi_h^{-1} \Psi_h \Pi_h$, where Π_h^{-1} can be expanded using the same differential operators as in Π_h, and the processed integrator $\hat{\psi}_h$ has order of consistency $\geq q$ if

$$\Psi_h \Pi_h = \Pi_h \exp(hF) + \mathcal{O}(h^{q+1}). \tag{2.5}$$

It is important to notice that different postprocessors may result in the same processed integrator so that it is useful to consider the following definition.

DEFINITION 2.1. *Two postprocessors π_h and $\bar{\pi}_h$ are said to be equivalent with respect to the kernel ψ_h if they give rise to the same processed integrator, i.e., if $\pi_h \circ \psi_h \circ \pi_h^{-1} = \bar{\pi}_h \circ \psi_h \circ \bar{\pi}_h^{-1}$ or, in terms of their respective series of differential operators, if*

$$\Pi_h^{-1} \Psi_h \Pi_h = \bar{\Pi}_h^{-1} \Psi_h \bar{\Pi}_h. \tag{2.6}$$

Remark. Clearly, Π_h and $\bar{\Pi}_h$ are equivalent with respect to the kernel $\Psi_h = \exp(F_h)$ if and only if the vector field $S_h = \log(\Pi_h \bar{\Pi}_h^{-1})$ commutes with F_h, for (2.6) can be written as $\exp(F_h) = \Pi_h \bar{\Pi}_h^{-1} \exp(F_h)(\Pi_h \bar{\Pi}_h^{-1})^{-1}$ or $\exp(F_h)\exp(S_h) =$

$\exp(S_h)\exp(F_h)$, and this is true if and only if $[F_h, S_h] = 0$. In particular, given a postprocessor Π_h and a kernel $\Psi_h = \exp(F_h)$, Π_h is equivalent to $\bar{\Pi}_h = \exp(\lambda F_h)\Pi_h$ for an arbitrary $\lambda \in \mathbb{R}$.

For a given family of integrators $\mathcal{G}$, the effective order conditions are equations on the parameters of the family that indicate the effective order of a particular integrator ψ_h in $\mathcal{G}$. Such effective order conditions can be directly derived from (2.5) for each family of integrators. For instance, for Runge–Kutta methods, (2.5) is equivalent to considering composition of B-series, which is the usual procedure to study the effective order conditions in that setting [8]. However, a general treatment, including the study of the generic number of order conditions, seems difficult with this approach: it would require making specific assumptions on the structure and properties of the series of linear differential operators Ψ_h and Π_h. Instead we propose an alternative based on the vector fields

$$F_h = \sum_{k\geq 1} h^k F_k = \log(\Psi_h), \qquad \hat{F}_h = \sum_{k\geq 1} h^k \hat{F}_k = \log(\hat{\Psi}_h),$$
$$P_h = \sum_{k\geq 1} h^k P_k = \log(\Pi_h).$$

In principle, given a kernel $\Psi_h = \exp(\sum h^k F_k)$, one might look for the best possible postprocessor $\Pi_h = \exp(P_h)$ among all possible series of vector fields $P_h = \sum h^k P_k$. However, if F_k is known to belong (for each $k \geq 1$) to a certain Lie algebra $\mathcal{L}$ of vector fields and it is desired that the vector fields $\hat{F}_k$ associated with $\hat{\psi}_h$ also belong to $\mathcal{L}$, then it seems natural to restrict oneself to the case $P_k \in \mathcal{L}$ (this is particularly true if no additional assumptions are made for F_k). We will say that a processed integrator $\hat{\psi}_h$ has order $p \geq q$ in $\mathcal{L}$ if there exist vector fields $P_k \in \mathcal{L}$, $k \geq 1$, such that (2.5) holds with $\Pi_h = \exp(\sum h^k P_k)$.

THEOREM 2.2. *An integrator ψ_h has effective order $p \geq q$ in $\mathcal{L}$ if and only if there exist vector fields $P_1, \ldots, P_{q-1} \in \mathcal{L}$ such that*

$$\begin{aligned} F_1 &= F, \\ [P_{k-1}, F] &= F_k + R_k(P_1, \ldots, P_{k-2}, F_1, \ldots, F_{k-1}), \quad 1 < k \leq q, \end{aligned} \tag{2.7}$$

holds, where

$$R_k = -\sum_{j=1}^{k-2} [P_j, F_{k-j}] + \sum_{l\geq 2} \frac{(-1)^l}{l!} \sum_{j_1+\cdots+j_{l+1}=k} [P_{j_1}, [P_{j_2}, \ldots [P_{j_l}, F_{j_{l+1}}] \ldots]]. \tag{2.8}$$

Proof. The equality $\hat{\Psi}_h = \Pi_h^{-1}\Psi_h\Pi_h$ can be written in terms of the respective vector fields as $\exp(\hat{F}_h) = \exp(-P_h)\exp(F_h)\exp(P_h)$. Formal application of the logarithm in both sides of this expression leads to [23]

$$\hat{F}_h = \exp(-P_h)F_h\exp(P_h) = \exp(\mathrm{ad}_{-P_h})F_h = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \mathrm{ad}_{P_h}^k F_h,$$

where $\mathrm{ad}_A B = [A, B]$. Therefore

$$\hat{F}_h = F_h - [P_h, F_h] + \frac{1}{2!}[P_h, [P_h, F_h]] - \frac{1}{3!}[P_h, [P_h, [P_h, F_h]]] + \cdots,$$

which implies

$$\begin{aligned} \hat{F}_1 &= F_1, \\ \hat{F}_k &= F_k + [F_1, P_{k-1}] + R_k, \quad k > 1, \end{aligned} \tag{2.9}$$

where $R_2 = 0$, and for $k > 2$, R_k is given by (2.8). Condition (2.5) reads $\hat{F}_1 = F$, $\hat{F}_k = 0$ for $2 \le k \le q$, which is equivalent to (2.7). ∎

In order to proceed further, we adopt the following assumption.

Assumption 1. The kernels ψ_h under consideration in this work are such that their associated vector fields $F_k \in \mathcal{L}_k$, $k \ge 1$, where $\{\mathcal{L}_n\}_{n\ge 1}$ is a sequence of subspaces of a certain graded Lie algebra $\mathcal{L}$ of vector fields satisfying (2.4).

In typical situations in numerical integration $\mathcal{L}$ is a graded free Lie algebra, and $n_k = \dim \mathcal{L}_k$ corresponds to the number of order conditions at order k for nonprocessed methods. The values of n_k, $k \ge 1$, can often be computed by using Witt's formula and their generalizations (see [19] and references therein).

Example 2.3. Let us now consider some particular cases which illustrate Assumption 1 and the context where the results of this paper can be applied.

(1.a) First assume that ODE (1.1) can be written as $x' = f_a(x) + f_b(x)$ and the vector field F is split accordingly as $F = F_a + F_b$. Suppose that the corresponding h-flows $\varphi_h^{[a]}$ and $\varphi_h^{[b]}$ can be exactly computed. Then it is useful to consider numerical integrators of the form

$$\psi_h = \varphi^{[b]}_{\alpha_{2s}h} \circ \varphi^{[a]}_{\alpha_{2s-1}h} \circ \cdots \circ \varphi^{[b]}_{\alpha_2 h} \circ \varphi^{[a]}_{\alpha_1 h}, \tag{2.10}$$

with $\alpha_i \in \mathbb{R}$; i.e., ψ_h is taken as a composition of basic flows. Now Assumption 1 holds for ψ_h with

$$\mathcal{L}_1 = \operatorname{span}(\{F_a, F_b\}), \qquad \mathcal{L}_k = \operatorname{span}\left(\bigcup_{l+m=k} [\mathcal{L}_l, \mathcal{L}_m] \right), \quad k \ge 2. \tag{2.11}$$

If one is interested in obtaining results that are valid for all pairs F_a, F_b of arbitrary vector fields, then one must assume that the only linear dependencies among nested commutators of F_a and F_b can be derived from the skew-symmetry and the Jacobi identity of commutators. In other words, $\mathcal{L} = \bigoplus_{k\ge 1} \mathcal{L}_k$ is the graded free Lie algebra generated by the symbolic vector fields F_a, F_b, where both have degree one. In particular, the dimensions n_k of the first homogeneous components $\mathcal{L}_k$ are $n_k = 2, 1, 2, 3, 6, 9, 18, 30, 56, 99$.

(1.b) Let us consider the generalized harmonic oscillator with Hamiltonian function

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^T M^{-1} \mathbf{p} + \frac{1}{2}\mathbf{q}^T S \mathbf{q}. \tag{2.12}$$

Here $\mathbf{q}, \mathbf{p} \in \mathbb{R}^d$ and M, S are constant symmetric matrices, M being invertible. This Hamiltonian (with $S = M^{-1}$) appears in the matrix representation of the time-dependent Schrödinger equation [10], where $\mathbf{q}$ and $\mathbf{p}$ represent the real and imaginary parts of the vector describing the state of the system. With $x = (\mathbf{q}, \mathbf{p})$, $D = 2d$, the corresponding equations of motion can be written as in (1.a) with $f_a(x) = (M^{-1}\mathbf{p}, \mathbf{0})$, $f_b(x) = (\mathbf{0}, -S\mathbf{q})$. Then the Hamiltonian vector field is decomposed as $F = F_a + F_b$, with

$$F_a = \sum_{i=1}^{d} (M^{-1}\mathbf{p})_i \frac{\partial}{\partial q_i}, \qquad F_b = \sum_{i=1}^{d} (-S\mathbf{q})_i \frac{\partial}{\partial p_i}.$$

Now Assumption 1 holds with $\mathcal{L}_k$ given by (2.11). In this case, not all the nested commutators are independent. For instance, $[F_a,[F_a,[F_a,F_b]]] = [F_b,[F_b,[F_b,F_a]]] = 0$. In fact, all nested commutators with an even number of operators F_a, F_b are either zero or a vector field F_C associated with the Hamiltonian $\mathbf{q}^T C\mathbf{p}$, where C is a polynomial matrix function of SM^{-1}. In consequence, $[F_a, F_C]$ is associated with a Hamiltonian function quadratic in $\mathbf{p}$, $[F_a,[F_a,F_C]] = 0$, and, similarly, $[F_b,[F_b,F_C]] = 0$. In addition, $[F_a,[F_b,F_C]] = [F_b,[F_a,F_C]]$ is also associated with a Hamiltonian function of the form $\mathbf{q}^T C_1\mathbf{p}$. As a result, $n_{2k} = 1$ and $n_{2k+1} = 2$ for all k.

(1.c) *Near-integrable system.* It corresponds to the problem $x' = f_a(x) + \varepsilon f_b(x)$ with $|\varepsilon| \ll 1$, which is a particular case of (1.a). The vector field associated with composition (2.10) takes the form $F_h = \sum_{k\geq 1}\sum_{i=1}^{k-1} h^k \varepsilon^i F_{k,i}$ so that we consider a bigraded Lie algebra with

$$F_a \in \mathcal{L}_{1,0}, \qquad F_b \in \mathcal{L}_{1,1}, \qquad [\mathcal{L}_{k,i}, \mathcal{L}_{m,j}] \subset \mathcal{L}_{k+m,i+j}, \tag{2.13}$$

and $\mathcal{L}_k = \bigoplus_{i=1}^{k-1} \mathcal{L}_{k,i}$ for $k \geq 2$. We denote $n_{k,i} = \dim \mathcal{L}_{k,i}$ so that obviously $n_k = \sum_{i=1}^{k-1} n_{k,i}$. An explicit formula for the $n_{k,i}$ can be found, in particular, in [16]: for instance, $n_{k,1} = n_{k,k-1} = 1$, $k > 1$, $n_{k,2} = n_{k,k-2} = \lfloor \frac{1}{2}(k-1)\rfloor$, $k > 2$, and $n_{k,3} = n_{k,k-3} = \lfloor \frac{1}{6}(k-1)(k-2)\rfloor$, $k > 3$ [19]. Here $\lfloor x\rfloor$ denotes the integer part of x.

(1.d) If $\mathcal{S}_h : \mathbb{R}^D \to \mathbb{R}^D$ is a second order time-symmetric integrator for (1.1), then we can consider integrators of the form [16]

$$\psi_h = \mathcal{S}_{\alpha_s h} \circ \cdots \circ \mathcal{S}_{\alpha_1 h}, \quad (\alpha_1, \ldots, \alpha_s) \in \mathbb{R}^s. \tag{2.14}$$

It can be shown (see Appendix A) that for such integrators Assumption 1 holds for the graded Lie algebra $\mathcal{L} = \bigoplus_{k\geq 1} \mathcal{L}_k$ generated by certain vector fields $\{Y_1, Y_3, Y_5, \ldots\}$ such that $Y_{2k-1} \in \mathcal{L}_{2k-1}$, $k \geq 1$. The dimensions n_k of the first homogeneous components $\mathcal{L}_k$ for $k \geq 1$ are $n_k = 1, 0, 1, 1, 2, 2, 4, 5, 8, 11, 18$ (see, for example, [19, 20]).

(1.e) *Runge–Kutta-type methods.* The set of rooted trees plays a fundamental role in the standard order theory of Runge–Kutta integrators applied to (1.1) (see, for instance, [6, 11, 12]). A similar role is played by certain sets of colored rooted trees in the case of other families of Runge–Kutta-type integrators such as Runge–Kutta–Nyström, partitioned Runge–Kutta, and additive Runge–Kutta methods. Let us generically denote as $\mathcal{T}$ the set of trees corresponding to a family of Runge–Kutta-type integrators and as $\mathcal{T}_k$ the set of trees in $\mathcal{T}$ with k vertices. For each family of methods, the parameters of any particular qth order integrator must satisfy $n_1 + \cdots + n_q$ algebraic equations, where n_k is the cardinal of $\mathcal{T}_k$. In the standard theory of order conditions, each tree $u \in \mathcal{T}$ is associated with an *elementary differential*, which is a map $F(u) : \mathbb{R}^D \to \mathbb{R}^D$ defined in terms of the map f in (1.1) and its partial derivatives. Now it can be seen that for each family of Runge–Kutta-type integrators considered above, Assumption 1 holds with

$$\mathcal{L}_k = \operatorname{span}\left(\sum_{i=1}^{D} (F(u))_i \frac{\partial}{\partial y_i} : u \in \mathcal{T}_k\right), \quad k \geq 1.$$

The dimensions n_k of the first homogeneous components $\mathcal{L}_k$ for $k \geq 1$ are $n_k = 1, 1, 2, 4, 9, 20, 48, 115, 286, 719$ [11].

As we have mentioned before, given a kernel of effective order q, the vector fields P_k satisfying (2.7) are not unique. This nonuniqueness is intimately related to the fact that the Lie subalgebra $\mathcal{L}^0 = \{G \in \mathcal{L} : [F, G] = 0\}$, i.e., the kernel of ad_F, is nonempty

(obviously, $F \in \mathcal{L}^0$). From this perspective, it is useful to choose a direct complement $\mathcal{L}^*$ of $\mathcal{L}^0$ with respect to $\mathcal{L}$ so that $\mathcal{L}$ is decomposed as a direct sum of two subspaces $\mathcal{L} = \mathcal{L}^0 \oplus \mathcal{L}^*$. For each k, we denote $\mathcal{L}_k^0 = \mathcal{L}^0 \cap \mathcal{L}_k$, $\mathcal{L}_k^* = \mathcal{L}^* \cap \mathcal{L}_k$, and n_k^* the dimension of $\mathcal{L}_k^*$ or, equivalently, $n_k^* = \dim[F, \mathcal{L}_k]$, where $[F, \mathcal{L}_k] = [F, \mathcal{L}_k^*]$ is a subspace of $\mathcal{L}_{k+1}$. In general, *if $\mathcal{L}$ is a graded free Lie algebra, then* $\dim[F, \mathcal{L}_k] = \dim \mathcal{L}_k$, $k > 1$, *i.e.,* $n_k^* = n_k$, $k > 1$, *and* $n_1^* = n_1 - 1$.

LEMMA 2.4. *Let $F_k, P_k \in \mathcal{L}_k$ for each $k \geq 1$, with $F_1 = F$. There exist unique $P_k^* \in \mathcal{L}_k^*$, $k \geq 1$, such that the postprocessors* $\exp(\sum_{k\geq 1} h^k P_k^*)$ *and* $\exp(\sum_{k\geq 1} h^k P_k)$ *are equivalent with respect to the kernel* $\Psi_h = \exp(\sum_{k\geq 1} h^k F_k)$.

Proof. By induction on n, it is sufficient to prove that if, in addition to the assumptions of Lemma 2.4, $P_1, \ldots, P_{n-1} \in \mathcal{L}^*$ and $P_n \notin \mathcal{L}_n^*$, then there exists a unique $P_n^* \in \mathcal{L}_n^*$ such that $\exp(hP_1 + \cdots + h^{n-1}P_{n-1} + h^n P_n^* + h^{n+1}Q_{n+1} + h^{n+2}Q_{n+2} + \cdots)$ is equivalent to $\exp(\sum h^k P_k)$ with certain $Q_k \in \mathcal{L}_k$, $k > n$.

One first proves that, for arbitrary $P_n^0 \in \mathcal{L}_n^0$, there exists a unique sequence $S_k^* \in \mathcal{L}_k^*$, $k \geq n+1$ such that $S_h = -h^n P_n^0 + \sum_{k\geq n+1} h^k S_k^*$ commutes with F_h. One considers $P_n^0 \in \mathcal{L}_n^0$, $P_n^* \in \mathcal{L}_n^*$ such that $P_n = P_n^0 + P_n^*$, and observe that, by choosing S_h as above, $\exp(\sum h^k P_k)$ is equivalent to

$$\exp\left(-h^n P_n^0 + \sum_{k\geq n+1} h^k S_k^*\right) \exp\left(\sum_{k\geq 1} h^k P_k\right) = \exp\left(\sum_{k=1}^{n-1} h^k P_k + h^n P_n^* + \cdots\right).$$

The uniqueness of P_n^* directly follows from (2.9). ∎

In other words, Lemma 2.4 shows that we can take into account only postprocessors such that $P_k \in \mathcal{L}_k^*$ without restricting the choice of the processed integrator. In addition, ψ_h has effective order $p \geq q$ in $\mathcal{L}$ if and only if there exist vector fields $P_k \in \mathcal{L}_k^*$, $k \leq q-1$, such that (2.7) hold. Moreover, such vector fields are unique in $\mathcal{L}^*$.

On the other hand, equations (2.9) lead directly to the following result.

LEMMA 2.5. *If the vector fields $F_k, \hat{F}_k \in \mathcal{L}_k$, $P_k \in \mathcal{L}_k^*$, $k \geq 1$, are associated with the kernel ψ_h, the processed method $\hat{\psi}_h$, and the postprocessor π_h, respectively, it follows that*

(a) *if ψ_h is a method of order d, then $\pi_h = \mathrm{id} + \mathcal{O}(h^d)$ or, equivalently,*

$$F_k = \hat{F}_k = 0, \quad 2 \leq k \leq d \quad \Longrightarrow \quad P_k = 0, \quad 1 \leq k \leq d-1;$$

(b) *provided the kernel is such that $\psi_{-h} = \psi_h^{-1} + \mathcal{O}(h^{2d+2})$, then it holds that $\hat{\psi}_{-h} = \hat{\psi}_h^{-1} + \mathcal{O}(h^{2d+2})$ if and only if $\pi_{-h} = \pi_h + \mathcal{O}(h^{2d+1})$. In terms of vector fields,*

$$F_{2k} = \hat{F}_{2k} = 0, \quad 1 \leq k \leq d \quad \Longleftrightarrow \quad F_{2k} = P_{2k-1} = 0, \quad 1 \leq k \leq d.$$

Next we rewrite the order conditions (2.7) for the processed integrator as a system of (polynomial) equations in the coefficients of the vector fields F_k in a basis $\{E_{k,i}\}_{i=1}^{n_k}$ of $\mathcal{L}_k$, $k \geq 1$. Such conditions take a very simple form if the basis of $\mathcal{L}_{k+1}$ ($k \geq 1$) includes a basis of $[F, \mathcal{L}_k] = [F, \mathcal{L}_k^*]$. This can be done, for instance, as follows. First, choose a basis $\{E_{k,i}^*\}_{i=1}^{n_k^*}$ of $\mathcal{L}_k^*$ (of course, such a basis of $\mathcal{L}_k^*$ can always be chosen as a subset of the basis $\{E_{k,i}\}_{i=1}^{n_k}$ of $\mathcal{L}_k$). Then take

$$E_{k+1,n_{k+1}-n_k^*+i} = [F, E_{k,i}^*], \quad \text{for } i = 1, \ldots, n_k^*, \tag{2.15}$$

and complete the basis of $\mathcal{L}_{k+1}$ by choosing $n_{k+1} - n_k^*$ elements of $\mathcal{L}_{k+1}$, say $E_{k+1,i}$, $i = 1, \ldots, n_{k+1} - n_k^*$, such that $\{E_{k+1,i}\}_{i=1}^{n_{k+1}}$ spans $\mathcal{L}_{k+1}$.

From now on, we assume that the bases of $\mathcal{L}_k$ and $\mathcal{L}_k^*$ have been constructed in such a way that (2.15) holds. Let us write

$$F_k = \sum_{i=1}^{n_k} f_{k,i} E_{k,i}, \qquad R_k = \sum_{i=1}^{n_k} r_{k,i} E_{k,i}, \qquad P_{k-1} = \sum_{i=1}^{n_{k-1}^*} p_{k-1,i} E_{k-1,i}^*. \tag{2.16}$$

The effective order conditions (2.7) are then expressed in terms of the coefficients $f_{k,i}$, $r_{k,i}$, and $p_{k-1,i}$ as follows.

THEOREM 2.6. *The scheme ψ_h, satisfying Assumption 1, has effective order $p \geq q$ if and only if*

$$f_{k,i} = -r_{k,i}, \qquad 1 \leq i \leq l_k := n_k - n_{k-1}^*, \tag{2.17}$$

$$p_{k-1,i} = -f_{k,l_k+i} - r_{k,l_k+i}, \quad 1 \leq i \leq n_{k-1}^*, \tag{2.18}$$

for $1 < k \leq q$. If, in addition, ψ_h is time-symmetric (i.e., if $\psi_{-h} \circ \psi_h = \mathrm{id}$), then, for even values of k, conditions (2.17) *are automatically satisfied and equations* (2.18) *reduce to $p_{k-1,i} = 0$.*

Proof. Assumption 1 implies that each R_k in (2.9) belongs to $\mathcal{L}_k$ and thus expressions (2.16) hold, where each $r_{k,i}$ is a polynomial in the coefficients $f_{l,j}$, $p_{l-1,j}$, $l = 2, \ldots, k-1$ (as R_k in (2.9) is a Lie polynomial in F_l, P_{l-1}, $l \leq k-1$). Conditions (2.7) together with (2.15) then lead to (2.17) and (2.18).

In the particular case of a time-symmetric kernel, $F_{2i} = 0$. The conclusion readily follows from Lemma 2.5. ∎

COROLLARY 2.7. *A total number of*

$$s(q) \equiv \sum_{k=1}^{q} n_k - \sum_{k=1}^{q-1} n_k^* = n_q + \sum_{k=1}^{q-1} (n_k - n_k^*)$$

conditions have to be satisfied by a given kernel ψ_h of effective order $p \geq q > 1$. If $\mathcal{L}$ is a graded free Lie algebra, this number is $s(q) = n_q + 1$. If ψ_h is time-symmetric, then the total number of effective order conditions reduces to $\bar{s}(2) = n_1$ and $\bar{s}(2n) = n_1 + \sum_{k=2}^{n} (n_{2k-1} - n_{2k-2}^*)$.

Proof. Equations (2.18) hold for any kernel, provided the postprocessor is appropriately chosen, and equations (2.17) give $l_k = n_k - n_{k-1}^*$ conditions for each $k = 2, \ldots, q$. This, together with the n_1 consistency conditions corresponding to $F_1 = F$, leads to $s(q)$ equations on the coefficients $f_{k,i}$. On the other hand, if the graded Lie algebra is free, then $n_k^* = n_k$, $k > 1$, and $n_1^* = n_1 - 1$. Finally, if ψ_h is time-symmetric, one has to count only the number of conditions for odd values of k. ∎

Remark. For any kernel, each $r_{k,i}$ is a polynomial in $f_{l,j}$, $p_{l-1,j}$, $l = 2, \ldots, k-1$. Recursive substitution of (2.17)–(2.18) in such polynomials $r_{k,i}$ leads to an equivalent system of equations of the form (2.17)–(2.18), where now each $r_{k,i}$ is a polynomial in the coefficients $f_{l,j}$, $l = 2, \ldots, k-1$, $j = n_l - n_{l-1}^* + 1, \ldots, n_l$.

Example 2.8. Next we provide the total number of order conditions for the particular cases collected in Example 2.3.

(2.a) For the composition methods of example (1.a), $\mathcal{L}^0 = \mathrm{span}(\{F\})$. Therefore $n_1^* = n_1 - 1 = 1$, and, among the different choices for $\mathcal{L}_1^*$, one can take, for instance, $\mathcal{L}_1^* = \mathrm{span}(\{F_a\})$, $\mathcal{L}_1^* = \mathrm{span}(\{F_b\})$, or $\mathcal{L}_1^* = \mathrm{span}(\{F_a - F_b\})$. For each $k \geq 2$, $\mathcal{L}_k \cap \mathcal{L}^0 = \{\emptyset\}$ so that $\mathcal{L}_k^* = \mathcal{L}_k$, $n_k^* = n_k$, and one can choose $E_{k,i}^* = E_{k,i}$. According to Corollary 2.7, the total number of effective order conditions is then $s(q) = n_q + 1$, a result already obtained in [3].

(2.b) For the harmonic oscillator (2.12) considered in example (1.b), the number of effective order conditions $s(q)$ is substantially reduced. As we have seen, $n_{2k-1} = 2$ and $n_{2k} = 1$ for each $k \geq 1$. The basis elements can recursively be built, for example, as follows: $E_{1,1} = F = F_a + F_b$, $E_{1,2} = F_a - F_b$, and for $k \geq 1$, $E_{2k,1} = [F, E_{2k-1,2}]$, $E_{2k+1,1} = [F_a - F_b, E_{2k,1}]$, $E_{2k+1,2} = [F, E_{2k,1}]$, with $\mathcal{L}_{2k} = \operatorname{span}(\{E_{2k,1}\})$ and $\mathcal{L}_{2k+1} = \operatorname{span}(\{E_{2k+1,1}, E_{2k+1,2}\})$. From example (1.b), we have that $[F_a, [F_a, E_{2k,1}]] = [F_b, [F_b, E_{2k,1}]] = 0$ and $[F, E_{2k+1,1}] = -[F, E_{2k+1,2}]$ so that

$$n^*_{2k} = \dim[F, \mathcal{L}_{2k}] = \dim \operatorname{span}(\{[F, E_{2k,1}]\}) = 1,$$
$$n^*_{2k+1} = \dim[F, \mathcal{L}_{2k+1}] = \dim \operatorname{span}(\{[F, E_{2k+1,1}], [F, E_{2k+1,2}]\}) = 1,$$

i.e., $n^*_k = 1$ for all k, and thus $s(q) = \lfloor (q+1)/2 \rfloor + 1$ (or $s(2n-1) = s(2n) = n+1$). Counting the number of effective order conditions and the number of variables from the composition (2.10) we observe that if the equations have real solutions, in principle methods of effective order $4s-2$ can be obtained. Furthermore, an interesting feature of schemes (2.10) applied to the generalized harmonic oscillator (2.12) is that for any kernel of the form (2.10), a postprocessor exists such that the processed integrator is time-symmetric. This is a consequence of the fact that $l_{2k} := n_{2k} - n^*_{2k-1} = 0$ for all k, and therefore $\hat{F}_{2k} = 0$ if the postprocessor is appropriately chosen (i.e., if $p_{2k-1,1} = -f_{2k,2} - r_{2k,2}$).

(2.c) Since the near-integrable problem is a particular case of (1.a), we can build a basis of $\mathcal{L}_k$ and then, by taking into account that $\mathcal{L}_k = \bigoplus_{i=1}^{k-1} \mathcal{L}_{k,i}$, obtain a basis of each $\mathcal{L}_{k,i}$. According to (2.a), $\mathcal{L}_k = \mathcal{L}^*_k$ and $\mathcal{L}_{k,i} = \mathcal{L}^*_{k,i}$ for $k > 1$, $i = 1, \ldots, k-1$. If we take $\mathcal{L}^0_1 = \operatorname{span}(\{F_a\})$, then $n_{1,0} = n_{1,1} = 1$ and $n^*_{1,0} = 0$, $n^*_{1,1} = 1$.

Usually, one is interested in designing methods such that [4]

$$F_h - F = \mathcal{O}(\varepsilon h^{s_1+1} + \varepsilon^2 h^{s_2+1} + \varepsilon^3 h^{s_3+1} + \cdots). \tag{2.19}$$

A method which satisfies this condition is said to be of order $(s_1, s_2, s_3, \ldots, s_q = q)$. We are interested in the case where $s_i \geq s_{i+1}$ and the list terminates with $\varepsilon^q h^{q+1}$, q being the standard order of consistency of the method. Observe that s_1 is the order of consistency the method would have in the limit $\varepsilon \to 0$.

To count the number of order conditions one has to consider each power of ε separately. In a nonprocessed method this number is $n_{1,0} + n_{1,1} + \sum_{i=1}^{q-1} \sum_{k=i+1}^{s_i} n_{k,i}$, whereas in the processed case this number reduces to (applying Corollary 2.7 to each power of ε separately)

$$s(s_1, \ldots, q) = 1 + \sum_{i=1}^{q-1} n_{s_i,i}.$$

Since $n_{s_1,1} = 1$, the number of order conditions is independent of s_1, and $(s_1, 2)$ methods can be obtained just with a consistent kernel (a first order method) [24]. If $s_1 = \cdots = s_q = q$, the result of Corollary 2.7 is recovered.

(2.d) For kernels constructed as compositions of a basic second order symmetric integrator (2.14), $\mathcal{L}^0 = \mathcal{L}_1 = \operatorname{span}(\{F\})$. Whence $n^*_1 = n_1 - 1 = 0$, and for each $k \geq 2$, $\mathcal{L}_k \cap \mathcal{L}^0 = \{\emptyset\}$ so that $\mathcal{L}^*_k = \mathcal{L}_k$, $n^*_k = n_k$. The total number of effective order conditions is then $s(q) = n_q + 1$.

(2.e) For the family of Runge–Kutta methods, the situation is very similar to (2.d). Now $n_1 = 1$, $n^*_1 = 0$, and $n^*_k = n_k$ for $k \geq 2$, and thus the number of conditions

to have effective order conditions q is $s(q) = n_q + 1$, that is, the number of rooted trees with q vertices plus one. This result was obtained by Butcher and Sanz-Serna in [8]. As for Runge–Kutta–Nyström methods, the situation is similar to (2.a), with $n_1 = 1$, $n_1^* = 0$, and $n_k^* = n_k$ for $k \geq 2$, and Corollary 2.7 again leads to $s(q) = n_q + 1$.

For a kernel of effective order q (i.e., satisfying equations (2.17) for $k \leq q$ but not for $k = q+1$), one could in principle determine a postprocessor such that (2.18) *holds also for all* $k > q$. From now on we shall refer to that postprocessor as *optimal*, as it causes many terms of each $\hat{F}_k = \sum_{i=1}^{n_k} \hat{f}_{k,i} E_{k,i}$ of the processed method $\hat{\psi}_h$ to cancel ($\hat{f}_{k,i} = 0$ for $i = n_k - n_{k-1}^* + 1, \ldots, n_k$).

Remark. This optimal postprocessor is not uniquely defined, and it depends on the way a basis of $[F, \mathcal{L}_{k-1}]$ ($k \geq 2$) is completed to get a basis of $\mathcal{L}_k$ (i.e., on the choice of the direct complement $\bar{\mathcal{L}}_k := \mathrm{span}(\{E_{k,i}\}_{i=1}^{n_k - n_{k-1}^*})$ of $[F, \mathcal{L}_{k-1}]$ with respect to $\mathcal{L}_k$). In fact, we are determining the optimal P_h by requiring that the vector field $\hat{F}_h$ belongs to $\bar{\mathcal{L}} := \bigoplus_{k \geq 1} \bar{\mathcal{L}}_k$ (i.e., that the projection onto $[F, \mathcal{L}]$ parallel to $\bar{\mathcal{L}}$ is canceled). This obviously depends on the choice of $\bar{\mathcal{L}}$. We will, nevertheless, still use the term "optimal postprocessor" by implicitly assuming that this refers to a prescribed decomposition $\mathcal{L} = \bar{\mathcal{L}} \oplus [F, \mathcal{L}]$.

DEFINITION 2.9. *We denote by* $\mathbb{P}_k$ *the set of maps* $\pi_h : \mathbb{R}^D \to \mathbb{R}^D$ *whose Taylor expansion is identical to the optimal postprocessor up to order* k *(i.e., their difference is* $\mathcal{O}(h^{k+1})$*).*

Thus, we have a qth order processed integrator $\hat{\psi}_h$ if the kernel ψ_h has effective order q and the postprocessor π_h is in $\mathbb{P}_{q-1}$. If, in addition, $\pi_h \in \mathbb{P}_q$, then the leading term of the resulting vector field $\hat{F}_h - hF$ coincides with the leading term of the optimal postprocessor.

3. Cheap postprocessing. In most cases the optimal postprocessor can be accurately approximated, but it usually turns into a scheme which is (at least) as expensive to evaluate as the kernel. Since the preprocessor is evaluated only once, it makes sense to use this (typically) expensive approximation. On the contrary, using the more accurate approximation to the postprocessor for obtaining intermediate results along the numerical integration process may deteriorate the efficiency of the method, especially if output is frequently required as occurs, for instance, in the calculation of Lyapunov exponents and the computation of Poincaré maps in dynamical systems. It is then reasonable to look for an approximation $\hat{\pi}_h$ to the optimal postprocessor as cheap to compute as possible. Usually, such a cheap postprocessor $\hat{\pi}_h$ will be a less accurate approximation to the optimal postprocessor, but the error $\hat{\pi}_h(y_n) - \pi_h(y_n)$ thus introduced will not be propagated: as we shall see in section 3.2, such an error eventually is overtaken by the global error of the underlying processed integrator in typical situations (where the global error grows at least linearly in time).

Computationally cheap approximations to the optimal postprocessor can be obtained by applying different techniques. Here we present an approach which can be considered cost-free. In essence, π_h is approximated by reusing intermediate calculations obtained in the evaluation of the kernel ψ_h.

More precisely, let $x(t_0) = x_0$ be the initial value of the problem, and $y_n = \psi_h^n(\pi_h^{-1}(x_0))$. Then we approximate $x_n = \pi_h(y_n)$ as the linear combination

$$x_n \approx \sum_{i=-s}^{s} w_i Y_i \tag{3.1}$$

of intermediate values $Y_i \in \mathbb{R}^D$ computed when evaluating $y_n = \psi_h(y_{n-1})$ and $y_{n+1} =$

$\psi_h(y_n)$. Here we consider only intermediate values from two steps, although more could also be used. There is no loss of generality though, since using $2m$ steps is equivalent to using two steps of the kernel ψ_h^m.

To proceed further, the existence of such intermediate values has to be guaranteed.

Assumption 2. After evaluating $y_{n+1} = \psi_h(y_n)$ with a kernel ψ_h satisfying Assumption 1, the intermediate values Y_i, $i = 1, \dots, s$, are available. These can be interpreted as $Y_i = \phi_h^{(i)}(y_n)$ for suitable integrators $\phi_h^{(i)}$ satisfying Assumption 1.

3.1. Conditions on the cheap postprocessor. Under Assumption 2, we consider (3.1) with $Y_0 = y_n$, and $Y_{-i} = \phi_h^{(s-i)}(y_{n-1})$, $i = 1, \dots, s$, that is, $Y_{-i} = \phi_h^{(-i)}(y_n)$, where $\phi_h^{(-i)} = \phi_h^{(s-i)} \circ \psi_h^{-1}$. Thus, (3.1) can be rewritten as

$$x_n \approx \hat{\pi}_h(y_n), \quad \text{where } \hat{\pi}_h(y) = \sum_{i=-s}^{s} w_i \phi_h^{(i)}(y) \tag{3.2}$$

and each $\phi_h^{(i)}(y)$, $-s \le i \le s$, is an integrator satisfying Assumption 1.

Example 3.1. We illustrate Assumption 2 in some particular cases.

(3.a) For kernels of the form (2.10), Assumption 2 holds for the intermediate values $Y_j = \phi_h^{(j)}(y_n)$ ($-2s \le j \le 2s$, because we have $2s$ intermediate stages per step), where

$$\begin{aligned} Y_{2i-1} &= \varphi^{[a]}_{\alpha_{2i-1}h} \circ \cdots \circ \varphi^{[a]}_{\alpha_1 h}(y_n), & Y_{2i} &= \varphi^{[b]}_{\alpha_{2i}h} \circ \cdots \circ \varphi^{[a]}_{\alpha_1 h}(y_n), \\ Y_{-2i+1} &= \varphi^{[a]}_{\alpha_{2s-2i+1}h} \circ \cdots \circ \varphi^{[a]}_{\alpha_1 h}(y_{n-1}), & Y_{-2i} &= \varphi^{[b]}_{\alpha_{2s-2i}h} \circ \cdots \circ \varphi^{[a]}_{\alpha_1 h}(y_{n-1}), \end{aligned}$$

and $-s \le i \le s$.

(3.b) For kernels of the form (2.14), Assumption 2 holds for the intermediate values $Y_i = \phi_h^{(i)}(y_n)$ ($-s \le i \le s$), where

$$Y_i = \mathcal{S}_{\alpha_i h} \circ \cdots \circ \mathcal{S}_{\alpha_1 h}(y_n), \qquad Y_{-i} = \mathcal{S}_{\alpha_{s-i} h} \circ \cdots \circ \mathcal{S}_{\alpha_1 h}(y_{n-1}). \tag{3.3}$$

(3.c) Recall that a Runge–Kutta integrator ψ_h for the system (1.1) reads as

$$\psi_h(y) = y + h \sum_{i=1}^{s} b_i f(Y_i), \quad Y_i = y + h \sum_{j=1}^{s} a_{ij} f(Y_j), \quad i = 1, \dots s, \tag{3.4}$$

where b_i, a_{ij} are parameters of the method. Clearly, Assumption 2 holds for the intermediate stages Y_i ($1 \le i \le s$), since each Y_i defines a Runge–Kutta scheme. The internal stages of other Runge–Kutta-type families of integrators can be similarly seen to satisfy Assumption 2.

Next we study the conditions the coefficients w_i must satisfy so that $\hat{\pi}_h \in \mathbb{P}_l$ with l as high as possible. In fact, this is guaranteed for a given $l \ge 1$ if

$$\hat{\Pi}_h := \sum_{i=-s}^{s} w_i \Phi_h^{(i)} = \Pi_h + \mathcal{O}(h^{l+1}), \tag{3.5}$$

where $\Phi_h^{(i)}$ ($-s < i \le s$) is the series of differential operators $\Phi_h^{(i)} = I + \sum_{j \ge 1} h^j \Phi_j^{(i)}$ such that, formally, $g \circ \phi_h^{(i)} = \Phi_h^{(i)}[g]$. From Assumption 2 we have that $\Phi_h^{(i)} = \exp(F_h^{(i)})$, where $F_h^{(i)} = \sum_{k \ge 1} h^k F_k^{(i)}$ and $F_k^{(i)} \in \mathcal{L}_k$, $k \ge 1$.

Observe that $\hat{\pi}_h$ cannot be interpreted as the exact 1-flow of a formal vector field in the Lie algebra $\mathcal{L}$, that is, $\log(\hat{\Pi}_h) \notin \mathcal{L}$. However, since $\hat{\Pi}_h$ is defined as a linear combination of exponentials of (formal) vector fields in $\mathcal{L}$, it is clear that $\hat{\Pi}_h$ is a formal series of elements in the associative algebra of linear differential operators generated by the vector fields in $\mathcal{L}$, and therefore, as noted in subsection 2.2, the series $\hat{\Pi}_h$ can be appropriately represented by using the universal enveloping algebra $\mathcal{A}$ of $\mathcal{L}$.

According to that discussion, Π_h and each Φ_h^i (hence $\hat{\Pi}_h$) can be expressed as

$$\Pi_k = \sum_{j=1}^{m_k} \pi_{k,j} D_{k,j}, \qquad \Phi_k^{(i)} = \sum_{j=1}^{m_k} \phi_{k,j}^{(i)} D_{k,j}, \quad -s \le i \le s, \tag{3.6}$$

where $\pi_{k,j}, \phi_{k,j}^{(i)} \in \mathbb{R}$, and $\{D_{k,j}\}_{j=1}^{m_k}$ is a basis of the kth homogeneous component $\mathcal{A}_k$ of $\mathcal{A}$, constructed, for instance, at the end of subsection 2.2. In particular, since $\Phi_h^{(i)} = \exp(F_h^{(i)})$ with $F_h^{(i)} = \sum_{k\ge 1} h^k F_k^{(i)}$, $F_k^{(i)} = \sum_{j=1}^{n_k} f_{k,j}^{(i)} E_{k,j}$, we have that each $\phi_{k,j}^{(i)}$ in (3.6) is a polynomial function of $f_{l,r}^{(i)}$, $l \le k$, $r \le n_l$. The same is true for the coefficients $\pi_{k,j}$ and the coefficients $p_{k,j}$ in $P_k = \sum_{j=1}^{n_k} p_{k,j} E_{k,j}$. Hence, (3.5) is equivalent to a system of linear equations on the unknowns w_i, i.e.,

$$\sum_{i=-s}^{s} w_i \phi_{k,j}^{(i)} = \pi_{k,j}, \quad 1 \le j \le m_k, \ 0 \le k \le l. \tag{3.7}$$

In particular, $\hat{\pi}_h \in \mathbb{P}_0$ is equivalent to $\sum_{i=-s}^{s} w_i = 1$, and the number of equations (3.7) required for $\hat{\pi}_h \in \mathbb{P}_l$ is then $1 + m_1 + \cdots + m_l$.

When the number of unknowns w_i in (3.1) is larger than the number of equations (3.7) required so that $\hat{\pi}_h \in \mathbb{P}_l$ for a given l, then one can use this freedom to minimize the difference with the optimal postprocessor at higher orders.

3.1.1. Cheap postprocessing for time-symmetric kernels. In the important case of time-symmetric kernels, $\Pi_k = 0$ for odd indices k. In addition, it is typically the case that $\Phi_h^{(-i)} = \Phi_{-h}^{(i)}$ for $-s \le i \le s$. The choice $w_{-i} = w_i$ for all i in (3.1) then makes sense, that is,

$$\hat{\pi}_h = w_0 \,\mathrm{id} + \sum_{i=1}^{s} w_i (\phi_h^{(i)} + \phi_h^{(-i)}), \tag{3.8}$$

so that ($w_0 = 1 - 2\sum_{i=1}^{s} w_i$)

$$\hat{\Pi}_h = w_0 I + \sum_{i=1}^{s} w_i (\Phi_h^{(i)} + \Phi_{-h}^{(i)}) = I + 2 \sum_{r\ge 1} h^{2r} \left(\sum_{i=1}^{s} w_i \Phi_{2r}^{(i)} \right).$$

This guarantees that equations (3.7) are automatically satisfied for odd values of k, and the equations for even values of k are of the form

$$2 \sum_{i=1}^{s} w_i \phi_{k,j}^{(i)} = \pi_{k,j}, \quad 1 \le j \le m_k. \tag{3.9}$$

Hence, the number of equations that remain to be satisfied by the unknowns $w_1, \ldots, w_s$ so that $\hat{\pi}_h \in \mathbb{P}_{2r-1}$ is $m_2 + \cdots + m_{2r-2}$.

Example 3.2. A kernel of the form (2.14) is time-symmetric if $\alpha_{s-i+1} = \alpha_i$ for each i. We already know that in that case, $f_{2i,j} = 0$, $p_{2i-1,j} = 0$. In addition, one has $\phi_h^{(-i)} = \phi_{-h}^{(i)}$ for the intermediate values (3.3) to be used for the cheap postprocessor. Hence, we take $w_{-i} = w_i$ $(1 \le i \le s)$ in (3.2). Thus, in particular, a total number of $m_2 + m_4 = 1 + 3 = 4$ linear equations (3.9) have to be satisfied in order that $\hat{\pi}_h \in \mathbb{P}_5$. In Appendix A these equations are written explicitly in terms of the coefficients α_i of the kernel.

3.1.2. Improved specialized cheap postprocessors. As we have seen, condition (3.5) is sufficient for a cheap postprocessor (3.2) to belong to $\mathbb{P}_l$. However, this is not necessary in general. In fact, (3.5) means that

$$\sum_{i=-s}^{s} w_i g(\phi^{(i)}(y)) = g(\pi_h(y)) + \mathcal{O}(h^{l+1}) \tag{3.10}$$

for any $g \in C^\infty(\mathbb{R}^D, \mathbb{R})$, $y \in \mathbb{R}^D$, but in order that $\hat{\pi}_h \in \mathbb{P}_l$, (3.10) has to be imposed only for $g = g_j$, $j = 1, \dots, D$, where g_j is the projection onto the jth coordinate. As we will see, this observation leads in certain cases to a reduction in the number of conditions required.

Example 3.3. Consider again the family of Runge–Kutta schemes (3.4). Recall that in that case, n_k is the number of rooted trees with k vertices, and it is not difficult to show that $m_k = n_{k+1}$ for each $k \ge 1$. Now the integrator $\hat{\pi}_h$ in (3.2) is itself a Runge–Kutta method (provided that $\sum w_i = 1$), and standard Runge–Kutta theory can be used to show that $1 + n_1 + \cdots + n_l$ conditions on the parameters w_i are sufficient for $\hat{\pi}_h \in \mathbb{P}_l$, instead of the $1 + m_1 + \cdots + m_l = 1 + n_1 + \cdots + n_{l+1}$ conditions obtained from (3.5).

One could also consider the use of cheap postprocessors with different sets of values of the parameters w_i for different components of y. In that case, one needs only to impose (3.10) for the projection onto the corresponding component. Under certain assumptions, this also leads to a reduction in the number of conditions to be satisfied by the coefficients w_i. To be more specific, let us consider the following assumptions.

Assumption 3. For a certain j, there exists $r_j \in C^\infty(\mathbb{R}^D, \mathbb{R})$ such that for any $k \ge 1$ and $\Phi_k \in \mathcal{A}_k$, $\Phi_k[g_j]$ can be written as a linear combination of elements in $\mathcal{A}_{k-1}$ acting on r_j.

One can show that, under Assumption 3, $2 + (m_1 + \cdots + m_{l-1})$ conditions on the parameters w_i guarantee that (3.10) holds for $g = g_j$ (such conditions are independent of the actual function r_j).

Assumption 3 holds, in particular, for every component for Runge–Kutta methods, that is, for the graded free Lie algebra associated with the set of rooted trees considered in example (1.e). It also holds for the case of integrators in example (1.d), provided the basic second order symmetric method $\mathcal{S}_h$ is the implicit trapezoidal rule.

Assumption 4. For a certain j, there exists $r_j \in C^\infty(\mathbb{R}^D, \mathbb{R})$ such that for any $k \ge 2$ and $\Phi_k \in \mathcal{A}_k$, $\Phi_k[g_j]$ can be written as a linear combination of elements in $\mathcal{A}_{k-2}$ acting on r_j.

In a similar way, it can be shown that, under Assumption 4, $1 + m_1 + (1 + m_1 + \cdots + m_{l-2})$ conditions on the parameters w_i are sufficient for (3.10) to hold with $g = g_j$.

It can be seen that when $\mathcal{L}$ is the Lie algebra corresponding to Runge–Kutta–Nyström methods (example (1.e)), Assumption 4 holds for the components corre-

sponding to positions, while Assumption 3 holds for the velocity components.

For the case of integrators in example (1.d), if the basic second order symmetric method $\mathcal{S}_h$ is the Störmer–Verlet method, then again Assumption 3 holds for velocities, while Assumption 4 holds for positions.

3.2. Error propagation. Our purpose now is to analyze the propagation of the global error when the postprocessor is approximated by the linear combination $\hat{\pi}_h$ of intermediate values obtained in the computation of the kernel. As a general rule, the precision of the final results is not conditioned by the use of a very accurate postprocessor, whereas the error introduced by replacing the preprocessor π_h^{-1} by $\hat{\pi}_h^{-1}$ can grow significantly along the integration.

To justify this assertion, let us consider a postprocessor π_h in $\mathbb{P}_l$, with $l \geq q$, and q is the order of the processed integrator $\hat{\psi}_h$. After n steps we have

$$x_n = \hat{\psi}_h^n(x_0) = \pi_h \circ \psi_h^n \circ \pi_h^{-1}(x_0) = x(t_n) + e_{h,q}(n, x_0),$$

where $t_n = nh$ and $e_{h,q}(n, x_0)$ is the global error of the method. If $\hat{\pi}_h \in \mathbb{P}_k$, with $k < q$, is used as the postprocessor, then

$$\tilde{x}_n \equiv \hat{\pi}_h \circ \psi_h^n \circ \pi_h^{-1}(x_0) = \hat{\pi}_h \circ \pi_h^{-1} \circ \hat{\psi}_h^n(x_0) = x(t_n) + e_{h,q}(n, x_0) + \hat{\delta}_{h,k}(x_n).$$

Here $\hat{\delta}_{h,k} \equiv \hat{\pi}_h \circ \pi_h^{-1} - \mathrm{id} = \mathcal{O}(h^{k+1})$ is an error of local nature which in general can be bounded independently of n, while the global error typically grows as n increases. On the other hand, if $\hat{\pi}_h^{-1}$ is used as the preprocessor, then

$$\begin{aligned}\hat{x}_n &\equiv \pi_h \circ \psi_h^n \circ \hat{\pi}_h^{-1}(x_0) = \hat{\psi}_h^n \circ \pi_h \circ \hat{\pi}_h^{-1}(x_0) = \hat{\psi}_h^n(x_0 + \tilde{\delta}_{h,k}(x_0)) \\ &= x(t_n) + e_{h,q}(n, x_0) + \tilde{e}_{h,k}(n, x_0),\end{aligned}$$

where $\tilde{e}_{h,k}$ corresponds to the propagation of the initial error $\tilde{\delta}_{h,k} \equiv \pi_h \circ \hat{\pi}_h^{-1} - \mathrm{id} = \mathcal{O}(h^{k+1})$. Now the error term $\tilde{e}_{h,k}$ is not of local character and can grow significantly as n increases.

It is important to notice that when the kernel approximately preserves an integral of motion I and $\hat{\pi}_h$ is used as the postprocessor, the accuracy in the value of I can be reduced. Nevertheless, one must keep in mind that this corresponds to a local error which is not propagated and that, if required, one can always use a more precise approximation to the postprocessor at selected times.

4. Numerical experiments. In this section we examine how the processing technique with a cheap postprocessor behaves in practice. Our purpose, rather than providing a complete analysis of different processed methods, is just to illustrate the previous theoretical analysis on some specific examples. We consider a kernel with effective order six of the form (2.14) with $s = 11$ constructed and studied in [18, 19]. Its coefficients $\alpha_i = \alpha_{12-i}$ are collected in Table 4.1. Next we construct an approximation $\pi_h^{(c)} \in \mathbb{P}_6$ to the postprocessor π_h also as a composition of the second order integrator $\mathcal{S}_h$ at different stages. In particular, we take

$$\pi_h^{(c)} = \omega_h \circ \omega_{-h} \simeq \pi_h, \quad \text{with } \omega_h = \mathcal{S}_{\gamma_6 h} \circ \cdots \circ \mathcal{S}_{\gamma_1 h} \tag{4.1}$$

and coefficients γ_i, $i = 1, \ldots, 6$, given in Table 4.1. Finally, we consider the intermediate values (3.3) and solve equations (A.2) for the cheap postprocessor $\hat{\pi}_h \in \mathbb{P}_5$. The corresponding solution obtained by taking w_1, w_5, w_6, w_7 (in addition to w_0) as the nonzero coefficients is also collected in Table 4.1.

TABLE 4.1

Coefficients for the sixth order processed method with kernel ψ_h of the form (2.14) ($s = 11$) and postprocessors π_h and $\hat{\pi}_h$ given by (4.1) and (3.8), respectively.

$P_{11}6$		
$\alpha_1 = 0.1705768865009222157$	$\gamma_6 = -0.1$	$w_0 = 1 - 2(w_1 + w_5 + w_6 + w_7)$
$\alpha_2 = \alpha_1$	$\gamma_5 = 0.24687306977659$	$w_1 = 0.35601475536028$
$\alpha_3 = \alpha_1$	$\gamma_4 = 0.09086982276241$	$w_5 = 0.12246549694690$
$\alpha_4 = \alpha_1$	$\gamma_3 = 0.23651387483203$	$w_6 = 0.00415291514453$
$\alpha_5 = -0.4233661408926580 48$	$\gamma_2 = -0.20621953139126$	$w_7 = -0.20658995116781$
$\alpha_6 = 1 - 2(\alpha_1 + \cdots + \alpha_5)$	$\gamma_1 = -(\gamma_2 + \cdots + \gamma_6)$	

We recall that both $\pi_h^{(c)}$ and $\hat{\pi}_h$ are approximations to the postprocessor π_h. The map $\pi_h^{(c)}$ is built as a composition of the basic integrator $\mathcal{S}_h$ so that $\log(\pi_h^{(c)}) \in \mathcal{L}$, and $\hat{\pi}_h$ is taken as a linear combination of intermediate values used in the calculation of the kernel. From (4.1) we observe that the computational cost of $\pi_h^{(c)}$ is similar to that of the kernel, whereas $\hat{\pi}$ can be considered cost-free.

We compare this sixth order test integrator with other standard nonprocessed composition methods of the same family. In particular, we consider the well-known sixth order seven stages method "A" ($Y_7 6$) built by Yoshida [25] and the optimized sixth order nine stages method ($SS, m = 9$) of McLachlan [16] ($M_9 6$) (similar results are obtained with the sixth order nine stages method proposed by Kahan and Li [13]).

Numerical example 1. To illustrate how the error is propagated along the evolution when different approximations to the postprocessor are considered, we take the simple Lotka–Volterra problem

$$u' = u(v-2), \qquad v' = v(1-u), \tag{4.2}$$

which admits as first integral $I(u,v) = \ln(uv^2) - (u+v)$. Using logarithmic scale ($q = \ln v,\ p = \ln u$) the system becomes Hamiltonian with $H = p - e^p + 2q - e^q = T(p) + V(q)$. Equations (4.2) can be written as $x' = f_a(x) + f_b(x)$ with $x = (u,v)$, $f_a = (u(v-2), 0)$, $f_b = (0, v(1-u))$ so that the corresponding h-flows $\varphi_h^{[a]}$ and $\varphi_h^{[b]}$ can be exactly computed. We choose as second order time-symmetric integrator the composition $\mathcal{S}_h = \varphi_{h/2}^{[a]} \circ \varphi_h^{[b]} \circ \varphi_{h/2}^{[a]}$.

In the region $0 < u, v$ the system has periodic trajectories around $(u,v) = (1,2)$. We take $(u_0, v_0) = (1,1)$, integrate up to $t = 100 \times 2\pi$, and get outputs at $t = i \times 2\pi$, $i = 1, \ldots, 100$. In Figure 4.1(a) we present the global error for the processed schemes both using the accurate postprocessor $\pi_h^{(c)}$ of (4.1) (method $P_{11}6$) and the cheap approximation $\hat{\pi}_h$ ($P_{11}6C$) only for output. The results obtained are compared with $Y_7 6$ and $M_9 6$. The time steps selected are $h = \frac{1}{14}, \frac{1}{11}, \frac{1}{9}$, for $Y_7 6$, $M_9 6$, and $P_{11}6$, respectively, so that all methods require approximately the same number of evaluations. Figure 4.1(b) shows the error in the first integral $I(u,v)$ for $P_{11}6$ and $P_{11}6C$. In this case, for $1.9 < \log(t) \le 2$ the cheap postprocessor $\hat{\pi}_h$ is replaced by $\pi_h^{(c)}$ just to clearly show that this higher accuracy can always be recovered. If $P_{11}6C$ is started with $(\hat{\pi}_h)^{-1}$ instead of $(\pi_h^{(c)})^{-1}$, this accuracy would not have been restored.

From the figures we observe the following: (a) the processed integrator is clearly more accurate; (b) the results for the global error obtained using $\hat{\pi}_h$ approach asymptotically those given by $\pi_h^{(c)}$; (c) the error in $I(u,v)$ is higher when $\hat{\pi}_h$ is used, but it does not grow with time, and the more accurate results can always be retrieved using $\pi_h^{(c)}$ when desired.

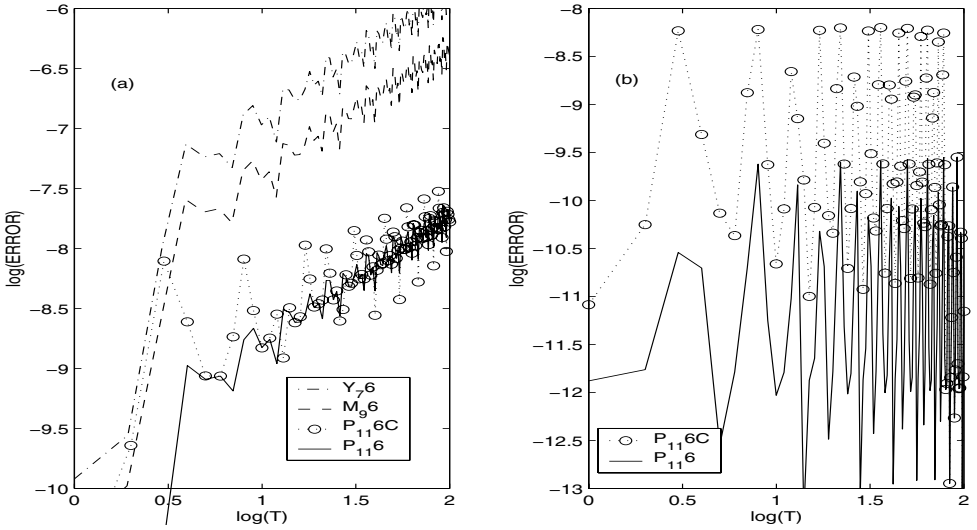


FIG. 4.1. (a) *Error in position and* (b) *error in the first integral* $I(u,v)$ *as functions of time for the Lotka–Volterra problem. The time step is chosen such that all methods require the same number of evaluations (this number corresponds to the kernel for the processed integrators).*

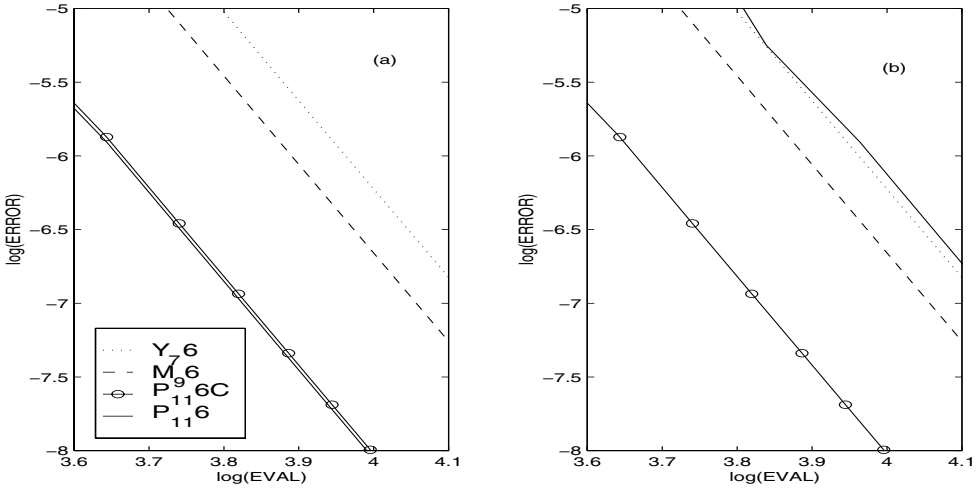


FIG. 4.2. *Average error in position versus number of evaluations for the first example* (a) *when the output is not frequent and* (b) *when the output is required at each step.*

Next we measure the average relative error in position versus the number of evaluations for different time steps and methods. Figure 4.2 shows the results (a) when the output is required only occasionally and (b) when it is required at each step. From this figure the importance of using a cheap postprocessor when the output is desired frequently is clear.

Numerical example 2. Let us consider now the ABC-flow [12], whose equations are given by

$$x' = B\cos y + C\sin z,$$

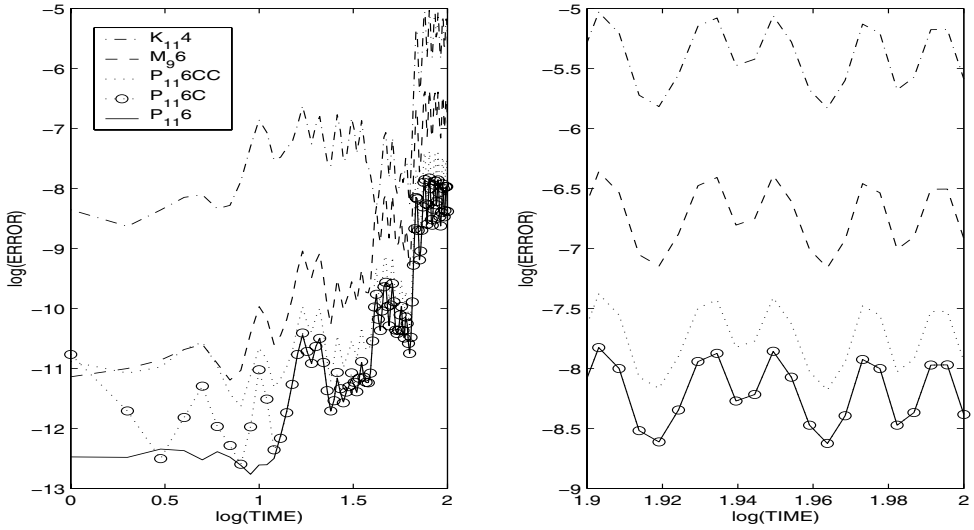


FIG. 4.3. *Error growth in position for the ABC-flow problem using a kernel* (2.14) *with coefficients in Table* 4.1 *and different pre- and postprocessors. The results obtained with the nonprocessed integrator* $\mathrm{M}_9 6$ *are also shown. The picture to the right is an enlargement of the rectangle* $[1.9, 2] \times [-9, -5]$ *in the left-hand picture.*

$$
\begin{aligned}
y' &= C\cos z + A\sin x,\\
z' &= A\cos x + B\sin y,
\end{aligned}
\tag{4.3}
$$

and the vector field is separable in three solvable parts, i.e.,

$$f = f_a + f_b + f_c = A(0, \sin x, \cos x) + B(\cos y, 0, \sin y) + C(\sin z, \cos z, 0).$$

We take as initial condition $(x_0, y_0, z_0) = (3.14, 2.77, 0)$, take as parameters $A = B = C = 1$, and integrate the system until $t = 100$. We choose as the basic symmetric second order integrator $\mathcal{S}_h = \chi_{h/2} \circ \chi^*_{h/2}$, where $\chi_h = \varphi_h^{[a]} \circ \varphi_h^{[b]} \circ \varphi_h^{[c]}$ and $\chi_h^* = \varphi_h^{[c]} \circ \varphi_h^{[b]} \circ \varphi_h^{[a]}$. In Figure 4.3 we show the error growth in the Euclidean norm when the following integrators based on $\mathrm{P}_{11}6$ are considered:

- ψ_h: only the kernel without the pre- and postprocessor (dash-dotted line, $\mathrm{K}_{11}4$);
- $\hat{\pi}_h \circ \psi_h \circ \hat{\pi}_h^{-1}$: the cheap pre- and postprocessors are employed (dotted line, $\mathrm{P}_{11}6\mathrm{CC}$);
- $\hat{\pi}_h \circ \psi_h \circ (\pi_h^{(c)})^{-1}$: we use the accurate preprocessor and the cheap postprocessor (circles joined by dotted lines, $\mathrm{P}_{11}6\mathrm{C}$);
- $\pi_h^{(c)} \circ \psi_h \circ (\pi_h^{(c)})^{-1}$: the accurate pre- and postprocessors are used (solid line, $\mathrm{P}_{11}6$).

We also include the results obtained using $\mathrm{M}_9 6$ (dashed line), choosing the time step such that the number of evaluations is the same as for the kernel. From the figure, it is clear that the kernel by itself is not good enough for giving accurate results (it is only a fourth order integrator). In addition we see that, at least for this problem, it is important to start the computation using a good preprocessor (some accuracy is lost when using $\hat{\pi}_h^{-1}$). Finally, we observe that after some time the results obtained using the cheap and the composition postprocessors agree up to drawing accuracy, but the former is faster to compute.

5. Concluding remarks. We have presented a general study of the processing technique which can be readily applied in several contexts. We obtain the number of order conditions and indicate how to find them explicitly in a systematic way. We have also presented a technique to find postprocessors virtually cost-free, just using intermediate results from the kernel. From the error propagation analysis we conclude that it is important to start the computation with an accurate preprocessor (even if it is expensive) and that, in general, a computationally cheap postprocessor can be safely used for ordinary intermediate output, although a more expensive postprocessor may be used, if required, to compute more accurate results at selected times.

An important application of the results contained in this paper is the construction of processed methods whose kernel is a composition of low order basic integrators. In that case, by analyzing the structure of the corresponding Lie algebra $\mathcal{L}$, it is possible to obtain approximations to the postprocessor either as a composition of basic methods or as a linear combination of intermediate stages of the kernel. In [2] this analysis is pursued in more detail for different families of composition methods, and new high order schemes are constructed which prove to be more efficient than other composition integrators available in the literature.

In practice, the efficient integration of systems of ODEs often requires the use of some step-size changing strategy. In principle, two possibilities can be contemplated. (i) Reparameterize the time variable in such a way that, with the new independent variable, a constant step-size can be used [12]. This is a familiar approach in geometric integration, and the theory developed here applies directly. (ii) Consider the problem of adapting the step-size in general terms, i.e., to construct processed methods whose step-size h changes to ρh, with $\rho \in [\rho_{\min}, \rho_{\max}]$ chosen according to some sort of local error estimation technique. This is the usual approach for general purpose integrators such as those based on explicit Runge–Kutta methods, and it is not suitable for geometric integration, as such standard variable step-size implementation destroys the geometric nature of the integration [22]. Although recently an adaptation of processing techniques to standard variable step-size strategies has been proposed in the Runge–Kutta context [9], this is largely an open problem which deserves further research.

Appendix A. Here we derive explicitly the effective order conditions up to order 6 for methods with kernel (2.14) and obtain the corresponding linear equations (3.9) for the cheap postprocessor (3.2). The series $S_h = I + \sum_{k\geq 1} h^k S_k$ of differential operators associated with the second order time-symmetric integrator $\mathcal{S}_h : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ for (1.1) can be written as $S_h = \exp(Y_h)$, where $Y_h = hY_1 + h^3 Y_3 + h^5 Y_5 + \cdots$, and $Y_1 = F$. Then

$$\Psi_h = \exp(Y_{h\alpha_1}) \cdots \exp(Y_{h\alpha_s}). \tag{A.1}$$

By repeated application of the Baker–Campbell–Hausdorff formula one arrives at an expansion of $F_h = \log(\Psi_h) = hF_1 + h^3 F_3 + h^4 F_4 + \cdots$, with $h^k F_k \in \mathcal{L}_k$ for the graded Lie algebra $\mathcal{L} = \bigoplus_{k\geq 1} \mathcal{L}_k$ generated by the vector fields $\{Y_1, Y_3, Y_5, \ldots\}$. Here $n_1 = 1$, $n_2 = 0$, $n_k^* = n_k$ for $k \geq 2$, whence, according to Lemma 2.5, $F_2 = 0$, $P_1 = P_2 = 0$. A basis of $\mathcal{L}$ is given in Table A.1 up to $k = 6$.

The order conditions for the kernel and postprocessor up to order six in this basis read as

$$\begin{gathered} f_{1,1} = 1, \qquad f_{3,1} = 0, \qquad f_{5,1} = 0, \\ p_{4,1} = -f_{5,2}, \qquad p_{1,1} = p_{3,1} = p_{5,1} = p_{5,2} = 0. \end{gathered}$$

TABLE A.1
Basis of $\mathcal{L} = \bigoplus_{k\geq 1} \mathcal{L}_k$, the free Lie algebra generated by $\{hY_1, h^3Y_3, h^5Y_5, \ldots\}$.

	Basis of $\mathcal{L}$	
$\mathcal{L}_1$	$E_{1,1} = Y_1 = F$	
$\mathcal{L}_3$	$E_{3,1} = Y_3$	
$\mathcal{L}_4$	$E_{4,1} = [F, E_{3,1}]$	
$\mathcal{L}_5$	$E_{5,1} = Y_5$	$E_{5,2} = [F, E_{4,1}]$
$\mathcal{L}_6$	$E_{6,1} = [F, E_{5,1}]$	$E_{6,2} = [F, E_{5,2}]$

The basis for $\mathcal{L}_k$ presented in Table A.1 leads to the following basis in the universal enveloping algebra:

$$\begin{aligned}
&\mathcal{A}_1 : D_{1,1} = E_{1,1}, \qquad \mathcal{A}_2 : D_{2,1} = \frac{1}{2}E_{1,1}^2,\\
&\mathcal{A}_3 : D_{3,1} = E_{3,1}, \quad D_{3,2} = \frac{1}{3!}E_{1,1}^3,\\
&\mathcal{A}_4 : D_{4,1} = E_{4,1}, \quad D_{4,2} = \frac{1}{4!}E_{1,1}^4, \quad D_{4,3} = \frac{1}{2}(E_{1,1}E_{3,1} + E_{3,1}E_{1,1}).
\end{aligned}$$

The series of vector fields Π_h corresponding to the optimal processor is

$$\Pi_h = \exp(P_h) = \exp(h^4 p_{4,1}E_{4,1} + \mathcal{O}(h^6)) = I + h^4 p_{4,1}D_{4,1} + \mathcal{O}(h^6).$$

For the intermediate stages of the cheap approximation we take (3.3), or, equivalently,

$$\Phi_h^{(i)} = \exp(Y_{h\alpha_1}) \cdots \exp(Y_{h\alpha_i}) = \exp(hf_{1,1}^{(i)}E_{1,1} + h^3 f_{3,1}^{(i)}E_{3,1} + h^4 f_{4,1}^{(i)}E_{4,1} + \mathcal{O}(h^5)).$$

Then $\Phi_h^{(i)} + \Phi_h^{(-i)} = 2(I + \Phi_2^{(i)}h^2 + \Phi_4^{(i)}h^4 + \mathcal{O}(h^6))$, with

$$\Phi_2^{(i)} = \phi_{2,1}^{(i)}D_{2,1}, \qquad \Phi_4^{(i)} = (\phi_{4,1}^{(i)}D_{4,1} + \phi_{4,2}^{(i)}D_{4,2} + \phi_{4,3}^{(i)}D_{4,3}).$$

Here

$$\phi_{2,1}^{(i)} = (f_{1,1}^{(i)})^2, \qquad \phi_{4,1}^{(i)} = f_{4,1}^{(i)}, \qquad \phi_{4,2}^{(i)} = (f_{1,1}^{(i)})^4, \qquad \phi_{4,3}^{(i)} = f_{1,1}^{(i)}f_{3,1}^{(i)},$$

and

$$f_{1,1}^{(i)} = \sum_{j=1}^{i} \alpha_j; \qquad f_{3,1}^{(i)} = \sum_{j=1}^{i} \alpha_j^3; \qquad f_{4,1}^{(i)} = \frac{1}{2}\left(\sum_{j=1}^{i-1} \alpha_j \sum_{k=1}^{j} \alpha_k^3 - \sum_{j=1}^{i-1} \alpha_j^3 \sum_{k=1}^{j} \alpha_k\right)$$

with $f_{4,1}^{(1)} = 0$. Finally, (3.9) for $k = 2, 4$ leads to the following linear system of equations:

$$\text{(A.2)} \quad \sum_{i=1}^{s} \phi_{2,1}^{(i)} w_i = 0; \quad \sum_{i=1}^{s} \phi_{4,1}^{(i)} w_i = \frac{1}{2}p_{4,1}; \quad \sum_{i=1}^{s} \phi_{4,2}^{(i)} w_i = 0; \quad \sum_{i=1}^{s} \phi_{4,3}^{(i)} w_i = 0$$

so that $\hat{\pi}_h \in \mathbb{P}_5$.

REFERENCES

[1] S. Blanes, *High order numerical integrators for differential equations using composition and processing of low order methods*, Appl. Numer. Math., 37 (2001), pp. 289–306.

[2] S. Blanes, F. Casas, and A. Murua, *Composition Methods for Differential Equations with Processing*, Preprint GIPS 2003-010, http://www.focm.net/gi/gips.

[3] S. Blanes, F. Casas, and J. Ros, *Symplectic integrators with processing: A general study*, SIAM J. Sci. Comput., 21 (1999), pp. 711–727.

[4] S. Blanes, F. Casas, and J. Ros, *Processing symplectic methods for near-integrable Hamiltonian systems*, Celestial Mech. Dynam. Astronom., 77 (2000), pp. 17–35.

[5] S. Blanes, F. Casas, and J. Ros, *High-order Runge-Kutta-Nyström geometric integrators with processing*, Appl. Numer. Math., 39 (2001), pp. 245–259.

[6] J. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley and Sons, Chichester, 1987.

[7] J. Butcher, *The effective order of Runge-Kutta methods*, in Conference on the Numerical Solution of Differential Equations, Lecture Notes in Math. 109, J.L. Morris, ed., Springer, Berlin, 1969, pp. 133–139.

[8] J. Butcher and J.M. Sanz-Serna, *The number of conditions for a Runge-Kutta method to have effective order p*, Appl. Numer. Math., 22 (1996), pp. 103–111.

[9] J.C. Butcher and T.M.H. Chan, *Variable stepsize schemes for effective order methods and enhanced order composition methods*, Numer. Algorithms, 26 (2001), pp. 131–150.

[10] S.K. Gray and D.E. Manolopoulos, *Symplectic integrators tailored to the time-dependent Schrödinger equation*, J. Chem. Phys., 104 (1996), pp. 7099–7112.

[11] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I, 2nd ed., Springer, Berlin, 1993.

[12] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer, Berlin, 2002.

[13] W. Kahan and R.C. Li, *Composition constants for raising the order of unconventional schemes for ordinary differential equations*, Math. Comp., 66 (1997), pp. 1089–1099.

[14] M.A. López-Marcos, J.M. Sanz-Serna, and R.D. Skeel, *Cheap enhancement of symplectic integrators*, in Proceedings of the Dundee Conference on Numerical Analysis, Dundee, Scotland, 1995, D.F. Griffiths and G.A. Watson, eds., Longman, Harlow, 1996.

[15] M.A. López-Marcos, J.M. Sanz-Serna, and R.D. Skeel, *Explicit symplectic integrators using Hessian-vector products*, SIAM J. Sci. Comput., 18 (1997), pp. 223–238.

[16] R.I. McLachlan, *On the numerical integration of ordinary differential equations by symmetric composition methods*, SIAM J. Sci. Comput., 16 (1995), pp. 151–168.

[17] R.I. McLachlan, *More on symplectic correctors*, in Integration Algorithms and Classical Mechanics, Fields Inst. Commun. 10, J.E. Marsden, G.W. Patrick, and W.F. Shadwick, eds., AMS, Providence, RI, 1996, pp. 141–149.

[18] R.I. McLachlan, *Families of high order composition methods*, Numer. Algorithms, 31 (2002), pp. 233–246.

[19] R.I. McLachlan and G.R.W. Quispel, *Splitting methods*, Acta Numer., 11 (2002), pp. 341–434.

[20] H. Munthe-Kaas and B. Owren, *Computations in a free Lie algebra*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 957–981.

[21] P.J. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed., Springer, New York, 1993.

[22] J.M. Sanz-Serna and M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.

[23] V.S. Varadarajan, *Lie Groups, Lie Algebras and Their Representations*, Springer, New York, 1984.

[24] J. Wisdom, M. Holman, and J. Touma, *Symplectic correctors*, in Integration Algorithms and Classical Mechanics, Fields Inst. Commun. 10, J.E. Marsden, G.W. Patrick, and W.F. Shadwick, eds., AMS, Providence, RI, 1996, pp. 217–244.

[25] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A, 150 (1990), pp. 262–268.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 553–575

DISCRETE DISPERSION RELATION FOR *hp*-VERSION FINITE ELEMENT APPROXIMATION AT HIGH WAVE NUMBER*

MARK AINSWORTH†

Dedicated to Professor Ian H. Sloan on the occasion of his sixty-fifth birthday

Abstract. The dispersive properties of high order finite element schemes are analyzed in the setting of the Helmholtz equation, and an explicit form of the discrete dispersion relation is obtained for elements of arbitrary order. It is shown that the numerical dispersion displays three different types of behavior, depending on the size of the order of the method relative to the mesh-size and the wave number. Quantitative estimates are obtained for the behavior and rates of decay of the dispersion error in the differing regimes. All estimates are fully explicit and are shown to be sharp. Limits are obtained, where transitions between the different regimes occur, and used to provide guidelines for the selection of the mesh-size and the polynomial order in terms of the wave number so that the dispersion error is controlled.

Key words. discrete dispersion relation, high wave number, *hp*-finite element method

AMS subject classifications. 65N50, 65N15, 65N30, 35A40, 35J05

DOI. 10.1137/S0036142903423460

1. Introduction.

Wave propagation phenomena arising in practical applications typically require large wave number (or frequency) ω. Accurate numerical simulation of such applications is thwarted by a number of issues, perhaps the most acute of which is numerical dispersion. This refers to the effect whereby the numerical scheme fails to propagate waves at the correct speed, resulting in a phase lead or lag in the numerical approximation. Numerical dispersion is often responsible not only for poor resolution but also for approximations that are even qualitatively incorrect.

Finite elements are often the method of choice for engineers interested in problems of continuum mechanics posed over complicated domains. It is therefore not surprising that finite elements are frequently used in numerical wave propagation [3]. The importance of numerical dispersion is widely recognized and is often used in assessing the quality of a numerical scheme and as a basis for ranking different finite element methods. For instance, Harari [16, 17], Harari and Avraham [18], and Harari and Hughes [19] consider the use of stabilized and Galerkin least squares finite element formulations for combatting the problem of dispersion in the solution of acoustic scattering problems, typically in the context of low order finite elements. Abboud and Pinsky [1] and Oberai and Pinsky [25] also study the discrete dispersive properties of various lower order finite element methods (such as the 8-node trilinear element, 20-node serendipity, 27-node triquadratic element) for the approximation of the scalar wave equation in three dimensions. More recently, Christon [6] considered the dispersive behavior of a variety of finite element schemes for the second order wave equation and performed a computational study of the discrete phase and group velocities.

*Received by the editors February 24, 2003; accepted for publication (in revised form) September 8, 2003; published electronically April 14, 2004.
http://www.siam.org/journals/sinum/42-2/42346.html

†Department of Mathematics, Strathclyde University, 26 Richmond Street, Glasgow G1 1XH, Scotland (M.Ainsworth@strath.ac.uk). The work of this author was supported in part by the Engineering and Physical Sciences Research Council of Great Britain under grant GR/M59426. This work was completed while the author was visiting the Newton Institute for Mathematical Sciences, Cambridge, UK.

The advantages of using higher order elements have also been widely recognized. For example, Harari and Avraham [18] compare the efficiency of first and second order elements for the solution of acoustic scattering problems and view their work as justifying the extension of the ideas to higher order (p-version) finite elements. Thompson and Pinsky [29] study the dispersive and attenuation properties of finite elements up to fifth order for the one-dimensional scalar Helmholtz equation and, on the basis of numerical evidence, conjecture that elements of degree p provide a $2p$th order accurate approximation of the dispersion relation in the limit where ωh tends to zero.

Applications are not confined to applications in acoustic scattering; for instance, Dyson [12] proposes the use of high (up to fifteenth) order schemes for propagation of waves for Euler equations. Cohen and Monk [7], Cohen [9], and Monk and Parrott [24] have considered the dispersive behavior of lower order finite element methods for Maxwell's equations. The use of high order finite element and spectral element schemes for the approximation of Maxwell's equations has attracted much interest [10, 20].

The first systematic study of the properties of finite element methods for high wave number applications was carried out in a series of papers by Babuška and Ihlenburg. In [4, 21], the convergence properties in the H^1-norm of first order finite elements for the one-dimensional model Helmholtz problem are studied working under the assumption that $\omega h < 1$. These ideas are extended to higher dimensions by Deraemaeker, Babuška, and Bouillard [11], who undertake a numerical study of the dispersive behavior for various finite element formulations in higher dimensions that allow one to include topological effects of the meshes, while Gerdes and Ihlenburg [14] study convergence of an h-version Galerkin finite element method for a three-dimensional problem of rigid scattering with mesh refinement in the radial direction and show that the error bound contains pollution effects similar to those observed in the one-dimensional analysis. A detailed study of the dispersion and approximation behavior of hp-finite elements for the Helmholtz equation in one dimension is undertaken by Ihlenburg and Babuška [22].

Despite extensive investigations, several important issues concerning the dispersive properties of standard finite element schemes remain unresolved, particularly in the context of high order elements. The aim of the present work is to give a sharp analysis of the dispersive properties of high order finite element schemes in the setting of the Helmholtz equation, to identify thresholds (relating the order p of the method and the mesh-size h to the wave number ω) where the dispersion error begins to decay, and to obtain sharp quantitative estimates on the rates of decay of the error in the differing regimes. A clear understanding of the dispersive properties of a scheme is not only of academic interest. Accurate quantitative information on the dispersion effects can serve as a practical guideline for the construction of a mesh and a polynomial order that will lead to a reasonable first approximation.

The analysis hinges on knowledge of an explicit form for the discrete dispersion relation valid for elements of arbitrary order. This is obviously a valuable tool for the study of numerical dispersion and, to the best of our knowledge, has not been obtained before. We derive a neat closed form expression for the discrete dispersion relation for elements of arbitrary order in terms of Padé approximants.

We study the behavior of the error in the discrete dispersion relation in two important limits: (i) in the small wave number limit, where $\omega h \ll 1$, where we provide a confirmation of the conjecture of Thompson and Pinsky [29]; and (ii) in the important practical case of high wave number, where $\omega h \gg 1$. The analysis provides a concrete guideline for choosing the order p of the elements and the mesh-size h in

order that the dispersion error is virtually eliminated:

$$p + \frac{1}{2} > \frac{\omega h}{2} + C(\omega h)^{1/3}, \tag{1.1}$$

where C is a constant which may in practice be chosen to be unity. In fact, when p is increased in this regime, the error decays at a superexponential rate, and an explicit expression for the error is given. In the limit where p is much larger than $\omega h/2$, the expression decays as $(\omega he/2(2p+1))^{2p+1}$, which is compatible with the upper bounds derived by Ihlenburg and Babuška [22]. More importantly, criterion (1.1) is shown to be sharp in the sense that if p does not satisfy (1.1), i.e., if

$$p + \frac{1}{2} < \frac{\omega h}{2} - o(\omega h)^{1/3},$$

then the dispersion error will not decay and may even increase significantly as the order p is increased. These results show that there is essentially no preasymptotic error reduction, which one might have expected based on the analysis for the positive definite case [28]. Strictly speaking, the error does begin to decay when p enters the transition zone:

$$p + \frac{1}{2} \in \left(\frac{\omega h}{2} - o(\omega h)^{1/3}, \frac{\omega h}{2} + o(\omega h)^{1/3} \right).$$

However, it is shown that the decay in this phase is only algebraic: $\mathcal{O}(p^{-1/3})$, and the comparatively narrow transition zone means that the preasymptotic decay is too short-lived to be of any real practical significance.

The results obtained here improve on the upper bounds given in [22]. In particular, all estimates are given explicitly and do not involve generic constants. This enables us to show that these estimates are the best ones possible. Our analysis is restricted to schemes of uniform order on tensor product meshes. Nevertheless, this type of scheme is used locally in regions remote from the scatterer, where the main issue is to control numerical dispersion. Equally well, our analysis deals only with the issue of numerical dispersion. The actual accuracy of the approximation is a separate issue which is considered by Ihlenburg and Babuška [22], where it is shown that the accuracy of the finite element approximation error in the H^1-norm will be quasi-optimal only if $\omega^{2p+1}h^{2p}$ is sufficiently small.

The remainder of this article is organized as follows. First, we review the standard framework leading to the derivation of the discrete dispersion relation in the setting of the wave equation in one dimension and describe the relevance of this to the multidimensional case, where tensor product meshes are used. The main results are outlined in the following section. The remaining sections deal with the technical details and proofs of the results.

2. The discrete dispersion relation.

2.1. The setting. It is well known that the general solution of the homogeneous wave equation in one space dimension,

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0,$$

can be expressed as a superposition of plane waves in the form

$$u(x,t) = \int_{\mathbb{R}} \left[a(k)e^{i(kx+\omega t)} + b(k)e^{i(kx-\omega t)} \right] \mathrm{d}k$$

for suitable functions a and b, where ω and k are related by the *dispersion relation*

$$\omega^2 = k^2.$$

Suppose that a uniform grid of size $h > 0$ is placed on the real line with nodes located at $h\mathbb{Z}$, and let V_h denote the set of continuous piecewise linear functions relative to the grid. By analogy with the continuous problem, we may seek solutions of the form

$$u_h(x,t) = e^{i\omega t} U_h(x) \tag{2.1}$$

so that $U_h \in V_h$ must satisfy

$$B_\omega(U_h, v_h) = 0 \quad \forall v_h \in V_h, \tag{2.2}$$

where

$$B_\omega(U_h, v_h) = (U_h', v_h') - \omega^2 (U_h, v_h)$$

and $(\cdot,\cdot)$ denotes the L_2-inner product on $\mathbb{R}$.

The invariance of the grid under translation by h prompts us to seek Bloch wave [26] solutions of the homogeneous (2.2) in the form

$$U_h(x) = \alpha \sum_{m\in\mathbb{Z}} e^{imkh} \theta_m(x), \tag{2.3}$$

where α and k are constants to be determined. Here, θ_m are the usual piecewise linear hat functions defined by

$$\theta_m(nh) = \delta_{mn}, \quad m, n \in \mathbb{Z}. \tag{2.4}$$

The translation invariance of the grid means that

$$\theta_m(x + nh) = \theta_{m-n}(x), \quad x \in \mathbb{R},\ n \in \mathbb{Z},$$

which in turn implies that U_h has the characteristic property of a Bloch wave: for each $n \in \mathbb{Z}$,

$$U_h(x + nh) = e^{iknh} U_h(x), \quad x \in \mathbb{R}. \tag{2.5}$$

This means that (2.2) is equivalent to the condition

$$B_\omega(U_h, \theta_0) = 0,$$

or, inserting the expression (2.3) for U_h,

$$\alpha \sum_{m\in\mathbb{Z}} e^{imkh} B_\omega(\theta_m, \theta_0) = 0.$$

Hence, a nontrivial Bloch wave exists, provided that

$$\sum_{m\in\mathbb{Z}} e^{imkh} B_\omega(\theta_m, \theta_0) = 0.$$

This expression may be simplified further using properties of the hat functions giving

$$e^{-ikh} B_\omega(\theta_{-1}, \theta_0) + B_\omega(\theta_0, \theta_0) + e^{ikh} B_\omega(\theta_1, \theta_0) = 0,$$

and, again by exploiting translation invariance of the grid, we obtain the *discrete dispersion relation*

$$2\cos(kh)B_\omega(\theta_1,\theta_0)+B_\omega(\theta_0,\theta_0)=0. \tag{2.6}$$

We refer to (2.6) as the dispersion relation for the following reason. For the first order standard Galerkin scheme considered above, we find that

$$B_\omega(\theta_0,\theta_0)=\frac{2}{h}\left(1-\frac{1}{3}\omega^2h^2\right),\qquad B_\omega(\theta_1,\theta_0)=-\frac{1}{h}\left(1+\frac{1}{6}\omega^2h^2\right),$$

and the discrete dispersion relation simplifies to give the well-known result,

$$kh=\cos^{-1}\left(\frac{6-2\omega^2h^2}{6+\omega^2h^2}\right)=\omega h-\frac{1}{24}(\omega h)^3+\cdots.$$

Here, we have adopted the common practice whereby kh is compared directly with ωh. However, examining (2.5) reveals that the value of $\exp(ikh)$ compared with $\exp(i\omega h)$ is actually of more physical relevance in the study of phase lag or lead. Of course, if both kh and ωh are small, then $\exp(ikh)-\exp(i\omega h)$ is closely related to $kh-\omega h$, and it makes sense to compare k and ω directly. However, we shall later be concerned with the regime $\omega h\gg 1$, and we shall therefore seek to compare $\exp(ikh)$ with $\exp(i\omega h)$. More precisely, we shall derive estimates for the difference $\cos(kh)-\cos(\omega h)$.

2.2. Extension to high order schemes. The discrete dispersion relation for higher order schemes may be obtained by modifying the previous arguments. Let V_{hp} denote the set of continuous piecewise polynomials of degree p on the grid $h\mathbb{Z}$, and let $V_{hp}^\flat$ denote the subspace

$$V_{hp}^\flat=\{v_{hp}\in V_{hp}: v_{hp}(mh)=0,\ m\in\mathbb{Z}\}.$$

As before, we seek a Bloch wave solution $U_{hp}\in V_{hp}$ satisfying

$$B_\omega(U_{hp},v_{hp})=0\quad\forall v_{hp}\in V_{hp}.$$

Accordingly, we write U_{hp} in the form

$$U_{hp}(x)=\sum_{m\in\mathbb{Z}}e^{ikmh}[\alpha\theta_m^{(p)}(x)+\beta\psi^{(p)}(x-mh)], \tag{2.7}$$

where α and β are constants, and $\psi^{(p)}\in V_{hp}^\flat$ is supported on $(0,h)$. By analogy with (2.4), the function $\theta_m^{(p)}\in V_{hp}$ has nodal values given by

$$\theta_m^{(p)}(nh)=\delta_{mn},\quad m,n\in\mathbb{Z}, \tag{2.8}$$

but is instead extended to the element interiors as a polynomial of degree p by requiring that

$$B_\omega(\theta_m^{(p)},v_{hp})=0\quad\forall v_{hp}\in V_{hp}^\flat. \tag{2.9}$$

This amounts to a decoupling of the nodal and interior degrees of freedom through Gaussian elimination or static condensation.

If ω does not correspond to a discrete eigenvalue, then $\theta_m^{(p)}$ is uniquely defined by this condition. It is not difficult to show that the function U_{hp} satisfies the Bloch wave condition: for all $n \in \mathbb{Z}$,

$$U_{hp}(x + nh) = e^{iknh} U_{hp}(x), \quad x \in \mathbb{R}.$$

Consequently, exploiting translation invariance of the grid, it suffices to require

$$\left.\begin{aligned} B_\omega(U_{hp}, \theta_0^{(p)}) &= 0 \\ B_\omega(U_{hp}, \psi^{(p)}) &= 0 \end{aligned}\right\}.$$

By inserting the expression (2.7) for U_{hp} into the latter equation, using (2.9) and the fact that ω does not correspond to a discrete eigenvalue, we conclude that $\psi^{(p)}$ must vanish identically. Consequently, the expression (2.7) collapses to the form considered in the case of first order schemes,

$$U_{hp}(x) = \alpha \sum_{m \in \mathbb{Z}} e^{ikmh} \theta_m^{(p)}(x), \tag{2.10}$$

and, by analogy with (2.6), the higher order discrete dispersion relation assumes the form

$$2\cos(kh) B_\omega(\theta_1^{(p)}, \theta_0^{(p)}) + B_\omega(\theta_0^{(p)}, \theta_0^{(p)}) = 0. \tag{2.11}$$

The same expression was obtained by Babuška and Ihlenburg [4] and Ihlenburg and Babuška [22]. A detailed study of the discrete dispersion relation for higher order elements is postponed until the next section.

2.3. Relevance to multidimensional problems. Information on the discrete dispersion relation for the scalar Helmholtz equation in one dimension may be used to derive explicit forms for the discrete dispersion relation for finite element approximation of problems in higher dimensions on tensor product meshes. Here, we describe the case of the wave equation in detail. An extension of the argument to the approximation of Maxwell equations using Nédélec elements may be found in [2].

Consider the wave equation in d-dimensions,

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} - \Delta u = 0,$$

and assume that a tensor product grid $h\mathbb{Z}^d$ is introduced on $\mathbb{R}^d$. Let V_{hp} denote the space of piecewise polynomials of total degree p in each variable on the grid; then seeking a discrete solution of the form (2.1) leads to the problem of determining $U_{hp} \in V_{hp}^{(d)}$ such that

$$\sum_{r=1}^{d} \left(\frac{\partial U_{hp}}{\partial x_r}, \frac{\partial v_{hp}}{\partial x_r} \right) - \kappa^2 (U_{hp}, v_{hp}) = 0 \quad \forall v_{hp} \in V_{hp}^{(d)}.$$

The tensor product structure prompts us to seek a solution of the form

$$U_{hp}(x_1, \ldots, x_d) = \alpha \prod_{r=1}^{d} X_p(k_r; x_r),$$

where α is a constant, and

$$X_p(k;s) = \sum_{m\in\mathbb{Z}} e^{imkh}\phi_m^{(p)}(s), \tag{2.12}$$

with $\phi_m^{(p)}$ defined as above. In particular, we recall that $X_p(k_r) \in V_{hp}$ satisfies

$$(X_p'(k_r), v') - \kappa_r^2(X_p(k_r), v) = 0 \quad \forall v \in V_{hp}, \tag{2.13}$$

where κ_r is related to k_r by the discrete dispersion relation (2.11).

Choosing the test function v_{hp} to be a product of one-dimensional functions $\prod v_r$ leads to the following necessary condition for the existence of nontrivial solutions:

$$\sum_{q=1}^{d}(\Phi_p'(k_q), v_q')\prod_{r\neq q}(X_p(k_r), v_r) - \kappa^2\prod_{r=1}^{d}(X_p(k_r), v_r) = 0.$$

Therefore, in view of (2.13), we obtain

$$\left(\sum_{r=1}^{d}\kappa_r^2 - \kappa^2\right)\prod_{q=1}^{d}(X_p(k_q), v_q) = 0,$$

and it follows that the discrete dispersion relation for the multidimensional scheme is given by

$$\sum_{r=1}^{d}\kappa_r^2 = \kappa^2, \tag{2.14}$$

where κ_r is related to k_r by the discrete dispersion relation (2.11). An alternative proof of this result for Gauss point mass lumped finite element schemes may be found in [9, p. 228]. The above argument extends immediately to these schemes.

3. Higher order discrete dispersion relation. An explicit expression for the higher order dispersion relation could, at least in principle, be derived by proceeding directly as in the first order case. Unfortunately, such a direct computation rapidly becomes intractable with increasing order, as pointed out in [9], and the general result might seem to be unattainable. In section 4, we prove that the discrete dispersion relation is given explicitly in terms of Padé approximants [5].

THEOREM 3.1. *Suppose $p \in \mathbb{N}$, and define set $N_e = \lfloor p/2\rfloor$ and $N_o = \lfloor (p+1)/2\rfloor$. Let $[2N_e + 2/2N_e]_{\kappa\tan\kappa}$ and $[2N_o/2N_o - 2]_{\kappa\cot\kappa}$ denote the Padé approximants to $\kappa\tan\kappa$ and $\kappa\cot\kappa$, respectively. Then the discrete dispersion relation is given by*

$$\cos(kh) = R_p(h\omega), \tag{3.1}$$

where R_p is the rational function

$$R_p(2\kappa) = \frac{[2N_o/2N_o - 2]_{\kappa\cot\kappa} - [2N_e + 2/2N_e]_{\kappa\tan\kappa}}{[2N_o/2N_o - 2]_{\kappa\cot\kappa} + [2N_e + 2/2N_e]_{\kappa\tan\kappa}}. \tag{3.2}$$

Despite the apparent simplicity, this result has been hitherto unknown in the literature, although special cases for lower order elements can be found in many sources. Table 3.1 shows the discrete dispersion relation and the leading term in the error in

TABLE 3.1
Discrete dispersion relation $\cos(kh) = R_p(\omega h)$ *for order* p *approximation given in Theorem* 3.1. *The leading term in the series expansion for the error when* $\Omega \ll 1$ *(see Theorem* 3.2*) is also indicated.*

Order p	$R_p(\Omega)$	$\cos^{-1} R_p(\Omega) - \Omega$
1	$\dfrac{-2\Omega^2+6}{\Omega^2+6}$	$-\dfrac{\Omega^3}{24}$
2	$\dfrac{3\Omega^4-104\Omega^2+240}{\Omega^4+16\Omega^2+240}$	$-\dfrac{\Omega^5}{1440}$
3	$\dfrac{-4\Omega^6+540\Omega^4-11520\Omega^2+25200}{\Omega^6+30\Omega^4+1080\Omega^2+25200}$	$-\dfrac{\Omega^6}{201600}$
4	$\dfrac{5\Omega^8-1800\Omega^6+134064\Omega^4-2378880\Omega^2+5080320}{\Omega^8+48\Omega^6+3024\Omega^4+161280\Omega^2+5080320}$	$-\dfrac{\Omega^7}{50803200}$

the dispersion relation for methods of order $p = 1$ to $p = 4$. For example, in the case of first order approximation $p = 1$,

$$\cos(kh) = R_1(\omega h) = \frac{6 - 2\omega^2 h^2}{6 + \omega^2 h^2},$$

which agrees with the result obtained for first order approximation in the previous section.

3.1. Accuracy at small wave numbers. The following general result for the leading term for the error in the dispersion relation, valid for wave numbers satisfying $\omega h \ll 1$, is proved in section 4.

THEOREM 3.2. *Let* $p \in \mathbb{N}$. *Then the error in the discrete dispersion relation is given by*

$$\cos kh - \cos \omega h = \frac{1}{2}\left[\frac{p!}{(2p)!}\right]^2 \frac{(\omega h)^{2p+2}}{2p+1} + \mathcal{O}(\omega h)^{2p+4} \tag{3.3}$$

or, if kh *is sufficiently small,*

$$kh - \omega h = -\frac{1}{2}\left[\frac{p!}{(2p)!}\right]^2 \frac{(\omega h)^{2p+1}}{2p+1} + \mathcal{O}(\omega h)^{2p+3}. \tag{3.4}$$

The result implies that

$$\frac{k}{\omega} - 1 = \mathcal{O}\left(\omega h\right)^{2p}, \tag{3.5}$$

meaning that the dispersion relation for a pth order scheme is accurate to order $2p$. This is consistent with the conjecture made by Thompson and Pinsky [29, eq. (41)] on the basis of numerical evidence in the particular cases of elements of order $p = 1$ to $p = 5$.

3.2. Accuracy at large wave numbers. While error estimates for small values of ωh are not without interest, the most interesting case in practice is the high wave number limit, where the practical limitations on the size of mesh means that product ωh is large even though h is small. The next result describes the behavior of the error as the order of approximation p is increased so that both p and ωh are large.

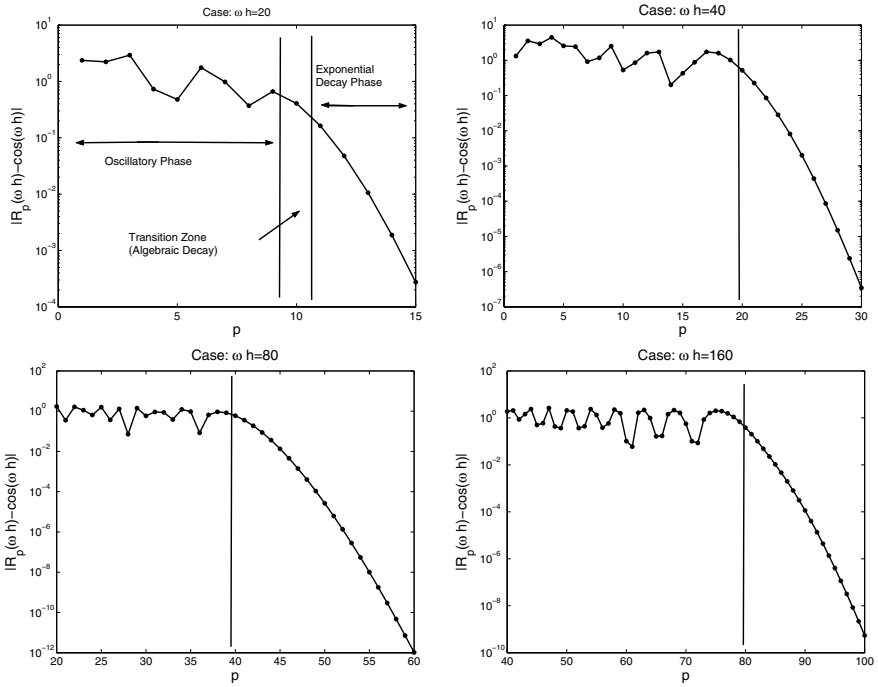


FIG. 3.1. *Behavior of error in the discrete dispersion relation for high wave numbers $\omega h \gg 1$ as the order p is increased. The transition region between the oscillatory phase and the superexponential decay of the error is indicated in each case (cf. Theorem* 3.3*).*

THEOREM 3.3. *Suppose that $\omega h \gg 1$. Then the error $\mathcal{E}^p = \cos kh - \cos \omega h$ in the discrete dispersion relation passes through three distinct phases as the order $p \in \mathbb{N}$ is increased*:

1. *Oscillatory phase: For $2p+1 < \omega h - o(\omega h)^{1/3}$, $\mathcal{E}^p$ oscillates, but does not decay, as p is increased.*

2. *Transition zone: For $\omega h - o(\omega h)^{1/3} < 2p+1 < \omega h + o(\omega h)^{1/3}$, the error $\mathcal{E}^p$ decays algebraically at a rate $\mathcal{O}(p^{-1/3})$.*

3. *Superexponential decay: For $2p+1 > \omega h + o(\omega h)^{1/3}$, $\mathcal{E}^p$ decreases at a superexponential rate:*

$$\mathcal{E}^p \approx \frac{\sin(\omega h)}{2} f(\sqrt{1-(\omega h/(2p+1))^2})^{p+1/2}, \tag{3.6}$$

where $f : w \to (1-w)/(1+w)\exp(2w)$, so that in the case where $2p+1 > \omega he/2$ with $e = \exp(1)$,

$$\mathcal{E}^p \approx \frac{\sin(\omega h)}{2}\left[\frac{\omega he}{2(2p+1)}\right]^{2p+1}. \tag{3.7}$$

Observe that the term appearing in parentheses in (3.6) has a magnitude less than unity, which follows from the fact that the function $f : w \to (1-w)/(1+w)\exp(2w)$ is nonnegative and monotonic decreasing on $[0,1]$ and the observation that $f(0) = 1$.

Figure 3.1 shows the behavior of the actual error as the order p is increased for a range of values of ωh. The oscillatory region and the transition to the superexponential decay in the error described in Theorem 3.3 can be clearly discerned.

Theorem 3.3 provides clear guidelines for the construction of meshes and the choice of order for the numerical resolution of waves using finite elements. In order to ensure that the dispersion error is for all practical purposes virtually eliminated, it is desirable that we work in the superexponential regime. For this reason, it is recommended to choose the order p and the mesh-size h so that

$$2p + 1 > \omega h + C(\omega h)^{1/3}, \tag{3.8}$$

where C is some fixed constant, which may be chosen to be unity in practice. The numerical results shown in Figure 3.1 support this criterion. It is interesting that the same type of criterion is arrived at for the choice of the number of terms to be used in the implementation of the fast multipole method for scattering problems [23, eq. (3.38)].

4. Proofs of the results.

4.1. Basic polynomials.

Let $\widehat{B}$ denote the bilinear form

$$\widehat{B}(u, v) = \int_{-1}^{1} \left(u'v' - \kappa^2 uv \right) \, \mathrm{d}s,$$

where $\kappa > 0$ is a constant. We introduce the basic polynomials, Φ_o^p and Φ_e^p, of degree at most $p \in \mathbb{N}$ satisfying

$$\Phi_e^p(\pm 1) = 1 : \quad \widehat{B}(\Phi_e^p, v) = 0 \quad \forall v \in \mathbb{P}_p \cap H_0^1(-1, 1) \tag{4.1}$$

and

$$\Phi_o^p(\pm 1) = \pm 1 : \quad \widehat{B}(\Phi_o^p, v) = 0 \quad \forall v \in \mathbb{P}_p \cap H_0^1(-1, 1). \tag{4.2}$$

Throughout, it will be assumed that κ does not coincide with an eigenvalue for this problem so that the polynomials are defined uniquely by these conditions.

It is easy to see that Φ_o^p and Φ_e^p are odd and even functions, respectively. The first result gives explicit closed forms for the expressions $\widehat{B}(\Phi_o^p, \Phi_o^p)$ and $\widehat{B}(\Phi_e^p, \Phi_e^p)$, which will be needed later.

THEOREM 4.1. *Let $p \in \mathbb{N}$ satisfy $p \geq 2$. Then*

1. *if $\kappa \neq (m + 1/2)\pi$ for all $m \in \mathbb{Z}$,*

$$\widehat{B}(\Phi_e^p, \Phi_e^p) = 2\kappa \frac{J_{2N+3/2}(\kappa) \cos \kappa + Y_{2N+3/2}(\kappa) \sin \kappa}{J_{2N+3/2}(\kappa) \sin \kappa - Y_{2N+3/2}(\kappa) \cos \kappa}, \tag{4.3}$$

where J and Y are cylindrical Bessel functions of the first and second kind, $N = \lfloor p/2 \rfloor$, and $\lfloor \cdot \rfloor$ denotes the integer part;

2. *if $\kappa \neq m\pi$ for all $m \in \mathbb{Z}$,*

$$\widehat{B}(\Phi_o^p, \Phi_o^p) = -2\kappa \frac{J_{2N+1/2}(\kappa) \sin \kappa - Y_{2N+1/2}(\kappa) \cos \kappa}{J_{2N+1/2}(\kappa) \cos \kappa + Y_{2N+1/2}(\kappa) \sin \kappa}, \tag{4.4}$$

where $N = \lfloor (p + 1)/2 \rfloor$.

Proof. Symmetry arguments reveal that the function Φ_e^p is an even order polynomial of degree $2N$, where $N = \lfloor p/2 \rfloor$. For the remainder of the proof superscripts will

be omitted since no confusion is likely to arise. Using definition (4.1) and integrating by parts shows that

$$\int_{-1}^{1} \left(\Phi_e'' + \kappa^2 \Phi_e\right) v \, \mathrm{d}x = 0 \quad \forall v \in \mathbb{P}_{2N} \cap H_0^1(-1,1).$$

The term in parentheses is a polynomial of degree $2N$, which may be written in the form

$$\Phi_e'' + \kappa^2 \Phi_e = \sum_{k=1}^{2N+1} \mu_k L_k'(x)$$

for suitable scalars μ_k, where L_k is the Legendre polynomial [15] of degree k. Inserting $v = (1-x^2)L_k'(x)$, with $j = 1, 2, \ldots, 2N-1$, and recalling the orthogonality property [15]

$$\int_{-1}^{1} (1-x^2) L_j'(x) \, L_k'(x) \, \mathrm{d}x = 0 \quad \text{for } j \neq k,$$

leads to the conclusion $\mu_j = 0$ for $j = 1, 2, \ldots, 2N-1$. Furthermore, the fact that L_{2N}' is an odd function and a parity argument shows that $\mu_{2N} = 0$. Hence,

$$\Phi_e'' + \kappa^2 \Phi_e = \mu_{2N+1} L_{2N+1}'. \tag{4.5}$$

It is not difficult to verify that the function w_{2N} defined by

$$w_{2N}(x) = \sum_{j=0}^{N} \left(-\frac{1}{\kappa^2}\right)^{j+1} L_{2N+1}^{(2j+1)}(x) \tag{4.6}$$

is a polynomial of degree $2N$ satisfying

$$w_{2N}'' + \kappa^2 w_{2N} = -L_{2N+1}'.$$

Consequently, Φ_e may be written in the form

$$\Phi_e(x) = \frac{w_{2N}(x)}{w_{2N}(1)},$$

provided that $w_{2N}(1)$ is nonzero, and, moreover, inserting this form into (4.5) reveals that

$$\mu_{2N+1} = -1/w_{2N}(1).$$

With the aid of these results, we obtain

$$\begin{aligned}
\widehat{B}(\Phi_e, \Phi_e) &= \left[\Phi_e' \Phi_e\right]_{-1}^{1} - \int_{-1}^{1} \Phi_e \left(\Phi_e'' + \kappa^2 \Phi_e\right) \mathrm{d}x \\
&= 2\Phi_e'(1) - \mu_{2N+1} \int_{-1}^{1} \Phi_e L_{2N+1}' \, \mathrm{d}x \\
&= 2\Phi_e'(1) - \mu_{2N+1} \left[\Phi_e L_{2N+1}\right]_{-1}^{1} \\
&= 2\left(\Phi_e'(1) - \mu_{2N+1}\right) \\
&= \frac{2}{w_{2N}(1)} \left(1 + w_{2N}'(1)\right),
\end{aligned}$$

where standard properties of Legendre polynomials have been used, such as the fact that L_{2N+1} is orthogonal to any polynomial of lower degree with respect to the $L_2(-1,1)$ inner product and that $L_{2N+1}(\pm 1) = \pm 1$. The values of w and its derivative at $x = 1$ are given by the following formula, which is obtained using $(8.910)_2$ in [15]:

$$L_n^{(d)}(\pm 1) = \begin{cases} \dfrac{(n+d)!}{d!(n-d)!}\dfrac{(\pm 1)^{d+n}}{2^d} & \text{for } d = 0, \dots, n, \\ 0 & \text{otherwise,} \end{cases} \tag{4.7}$$

giving, after some manipulation,

$$1 + w'_{2N}(1) = a_{2N+1} \tag{4.8}$$

and

$$w_{2N}(1) = -b_{2N+1}/\kappa. \tag{4.9}$$

Here, a_n and b_n are the expressions defined, for nonnegative integers n, by

$$a_n = \sum_{k=0}^{\lfloor n/2 \rfloor} \frac{(-1)^k}{(2k)!}\frac{(n+2k)!}{(n-2k)!}\frac{1}{(2\kappa)^{2k}} \tag{4.10}$$

and

$$b_n = \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \frac{(-1)^k}{(2k+1)!}\frac{(n+2k+1)!}{(n-2k-1)!}\frac{1}{(2\kappa)^{2k+1}}. \tag{4.11}$$

These series appear in formulas (8.461) and (8.465) of [15] and satisfy the identity

$$\begin{bmatrix} \sin(\kappa - \pi n/2) & \cos(\kappa - \pi n/2) \\ \cos(\kappa - \pi n/2) & -\sin(\kappa - \pi n/2) \end{bmatrix} \begin{bmatrix} a_n \\ b_n \end{bmatrix} = \sqrt{\frac{\pi\kappa}{2}} \begin{bmatrix} J_{n+1/2}(\kappa) \\ (-1)^{n-1} Y_{n+1/2}(\kappa) \end{bmatrix}, \tag{4.12}$$

where $J_{n+1/2}$ and $Y_{n+1/2}$ again denote cylindrical Bessel functions of the first and second kind, respectively. Equation (4.12) may be rearranged to obtain expressions for the series a_n and b_n, leading to the conclusion,

$$\widehat{B}(\Phi_e, \Phi_e) = -2\kappa\frac{a_{2N+1}}{b_{2N+1}} = 2\kappa\frac{J_{2N+3/2}(\kappa)\cos\kappa + Y_{2N+3/2}(\kappa)\sin\kappa}{J_{2N+3/2}(\kappa)\sin\kappa - Y_{2N+3/2}(\kappa)\cos\kappa}, \tag{4.13}$$

which completes the proof in the even case.

The proof of the odd order case follows similar lines, leading to the following analogue of (4.13):

$$\widehat{B}(\Phi_o, \Phi_o) = -2\kappa\frac{a_{2N}}{b_{2N}}, \tag{4.14}$$

where $N = \lfloor (p+1)/2 \rfloor$. Inserting expressions for the series a_{2N} and b_{2N} and simplifying leads to the result claimed. □

Equations (4.3) and (4.4) provide compact representations for the terms $\widehat{B}(\Phi_e^p, \Phi_e^p)$ and $\widehat{B}(\Phi_o^p, \Phi_o^p)$ but hide the fact that they are actually rational functions of κ. Interestingly enough, the expressions are actually certain types of Padé approximants [5].

THEOREM 4.2. *Let $p \in \mathbb{N}$ satisfy $p \geq 2$. Then we have the following:*

1. *$\widehat{B}(\Phi_e^p, \Phi_e^p)$ is the $[2N+2/2N]$-Padé approximant of $-2\kappa\tan\kappa$, where $N = \lfloor p/2 \rfloor$. Furthermore, if $\kappa \neq (m+1/2)\pi$, $m \in \mathbb{Z}$, then*

$$\begin{aligned} \mathcal{E}_e^p(\kappa) &= \widehat{B}(\Phi_e^p, \Phi_e^p) + 2\kappa\tan\kappa \\ &= \frac{1}{2}\left[\frac{(2N+1)!}{(4N+2)!}\right]^2 \frac{(2\kappa)^{4N+4}}{4N+3} + \mathcal{O}(\kappa^{4N+6}). \end{aligned} \tag{4.15}$$

2. *$\widehat{B}(\Phi_o^p, \Phi_o^p)$ is the $[2N/2N-2]$-Padé approximant of $2\kappa\cot\kappa$, where $N = \lfloor (p+1)/2 \rfloor$. If $\kappa \neq m\pi$, $m \in \mathbb{Z}$, then*

$$\begin{aligned} \mathcal{E}_o^p(\kappa) &= \widehat{B}(\Phi_o^p, \Phi_o^p) - 2\kappa\cot\kappa \\ &= 2\left[\frac{(2N)!}{(4N)!}\right]^2 \frac{(2\kappa)^{4N}}{4N+1} + \mathcal{O}(\kappa^{4N+2}). \end{aligned} \tag{4.16}$$

Proof. First, (4.13) shows that $\widehat{B}(\Phi_e^p, \Phi_e^p)$ is given by $-2\kappa a_{2N+1}/b_{2N+1}$, where a_{2N+1} and b_{2N+1} are defined in (4.10)–(4.11). It is not difficult to see that $\kappa^{2N+1}a_{2N+1}$ and $\kappa^{2N+1}b_{2N+1}$ are polynomials in κ of degree $2N+1$ and $2N$, respectively. Hence, $\widehat{B}(\Phi_e^p, \Phi_e^p)$ is a rational function of degree $[2N+2/2N]$. Straightforward manipulation beginning with the expression (4.3) gives

$$\widehat{B}(\Phi_e^p, \Phi_e^p) + 2\kappa\tan\kappa = -\frac{2\kappa}{\cos^2\kappa} Q_{2N+3/2}(\kappa)\left(1 - Q_{2N+3/2}(\kappa)\tan\kappa\right)^{-1}, \tag{4.17}$$

where

$$Q_{2N+3/2}(\kappa) = \frac{J_{2N+3/2}(\kappa)}{Y_{2N+3/2}(\kappa)}.$$

The behavior of $Q_{2N+3/2}(\kappa)$ is studied in the appendix, where the following estimate is proved in Lemma A.1:

$$Q_{2N+3/2}(\kappa) = -\frac{1}{2}\left[\frac{(2N+1)!}{(4N+2)!}\right]^2 \frac{(2\kappa)^{4N+3}}{4N+3} + \cdots.$$

With the aid of this estimate, we obtain that

$$\widehat{B}(\Phi_e^p, \Phi_e^p) + 2\kappa\tan\kappa = \frac{1}{2}\left[\frac{(2N+1)!}{(4N+2)!}\right]^2 \frac{(2\kappa)^{4N+4}}{4N+3} + \cdots$$

as claimed. Summarizing, we have shown that $\widehat{B}(\Phi_e^p, \Phi_e^p)$ is the rational function of degree $[2N+2/2N]$ which approximates $-2\kappa\tan\kappa$ to order $4N+4$. Consequently, $\widehat{B}(\Phi_e^p, \Phi_e^p)$ is the $[2N+2/2N]$-Padé approximant of $-2\kappa\tan\kappa$.

The assertions concerning $\widehat{B}(\Phi_o^p, \Phi_o^p)$ are proved in a similar fashion. In particular, using (4.14) it is easy to see that $\widehat{B}(\Phi_o^p, \Phi_o^p)$ is a rational function of type $[2N/2N-2]$. With the aid of (4.4), we derive

$$\widehat{B}(\Phi_o^p, \Phi_o^p) - 2\kappa\cot\kappa = -\frac{2\kappa}{\sin^2\kappa} Q_{2N+1/2}(\kappa)\left(1 + Q_{2N+1/2}(\kappa)\cot\kappa\right)^{-1}, \tag{4.18}$$

where

$$Q_{2N+1/2}(\kappa) = \frac{J_{2N+1/2}(\kappa)}{Y_{2N+1/2}(\kappa)},$$

and then, applying Lemma A.1, we deduce that

$$\widehat{B}(\Phi_o^p, \Phi_o^p) - 2\kappa \cot \kappa = 2 \left[\frac{(2N)!}{(4N)!}\right]^2 \frac{(2\kappa)^{4N}}{4N+1} + \cdots$$

as claimed. It follows that $\widehat{B}(\Phi_o^p, \Phi_o^p)$ is the $[2N/2N-2]$-Padé approximant of $2\kappa \cot \kappa$. ∎

4.2. Proof of Theorem 3.1. We are now in a position to present the proof of Theorem 3.1.

Proof. Fix $\kappa = \omega h/2$. First, we claim that for $x \in (0,h)$, the function $\theta_m^{(p)}$ defined in (2.8)–(2.9) may be expressed in terms of the basic polynomials as follows:

$$\theta_0^{(p)}(x) = \frac{1}{2}[\Phi_e^p(s) - \Phi_o^p(s)]$$

and

$$\theta_1^{(p)}(x) = \frac{1}{2}[\Phi_e^p(s) + \Phi_o^p(s)],$$

where $s = 2x/h - 1 \in (-1,1)$. It is easy to verify that the expression for $\theta_0^{(p)}$ takes the correct values at the endpoints $x = 0$ and $x = h$. Moreover, since Φ_e^p is a polynomial of degree p (in both x and s), it suffices to show that the orthogonality condition (2.9) is satisfied. Let $v_{hp} \in V_{hp}^\flat$ be supported on $(0,h)$, and define $V \in \mathbb{P} \cap H_0^1(-1,1)$ by $V(s) = v_{hp}(x)$, $x \in (0,h)$. A simple change of variable reveals that

$$B_\omega(\theta_0^{(p)}, v_{hp}) = h^{-1}\widehat{B}(\Phi_e^p - \Phi_o^p, V),$$

and conditions (4.1)–(4.2) show that this vanishes. Similar arguments may be applied in the case of $\theta_1^{(p)}$.

It is clear from symmetry considerations that $\theta_0^{(p)}$ is an even function. This fact, combined with a simple change of variable, shows that, since $\omega h = 2\kappa$,

$$B_\omega(\theta_0^{(p)}, \theta_0^{(p)}) = h^{-1}\widehat{B}(\Phi_e^p - \Phi_o^p, \Phi_e^p - \Phi_o^p);$$

then, exploiting the parities of Φ_e^p and Φ_o^p, we obtain

$$B_\omega(\theta_0^{(p)}, \theta_0^{(p)}) = h^{-1}[\widehat{B}(\Phi_e^p, \Phi_e^p) + \widehat{B}(\Phi_o^p, \Phi_o^p)].$$

Similar arguments reveal that

$$B_\omega(\theta_0^{(p)}, \theta_1^{(p)}) = (2h)^{-1}[\widehat{B}(\Phi_e^p, \Phi_e^p) - \widehat{B}(\Phi_o^p, \Phi_o^p)].$$

Substituting these results into (2.11) gives

$$\cos(kh) = \frac{\widehat{B}(\Phi_o^p, \Phi_o^p) + \widehat{B}(\Phi_e^p, \Phi_e^p)}{\widehat{B}(\Phi_o^p, \Phi_o^p) - \widehat{B}(\Phi_e^p, \Phi_e^p)}. \tag{4.19}$$

Theorem 4.2 identifies the terms appearing in this quotient as Padé approximants, and substituting for these expressions leads to the result claimed in Theorem 3.1. ∎

4.3. Error for small ωh. The general result for the leading term given in Theorem 3.2 for the error in the dispersion relation for the small wave number is proved using Theorem 4.2 as follows.

Proof. Fix $\kappa = \omega h/2 \ll 1$, and let $\mathcal{E}_e^p(\kappa)$ and $\mathcal{E}_o^p(\kappa)$ be defined as in Theorem 4.2. By writing $\widehat{B}(\Phi_e^p, \Phi_e^p)$ and $\widehat{B}(\Phi_o^p, \Phi_o^p)$ in terms of $\mathcal{E}_e^p$ and $\mathcal{E}_o^p$, respectively, substituting into (4.19), followed by a lengthy but otherwise straightforward computation, one arrives at the following expression for the error in the discrete dispersion relation:

$$\begin{aligned}(4.20)\qquad &\cos kh - \cos \omega h \\ &\quad = \frac{\sin \omega h}{\omega h}\left\{\mathcal{E}_o^p \sin^2\left(\frac{\omega h}{2}\right) + \mathcal{E}_e^p \cos^2\left(\frac{\omega h}{2}\right)\right\}\left\{1 + \frac{\sin \omega h}{2\omega h}(\mathcal{E}_o^p - \mathcal{E}_e^p)\right\}^{-1}.\end{aligned}$$

(Here, the argument κ of $\mathcal{E}_e^p$ and $\mathcal{E}_o^p$ has been suppressed.) In particular, for small argument, Theorem 4.2 implies that

$$\mathcal{E}_o^p = 2\left[\frac{(2N_o)!}{(4N_o)!}\right]^2 \frac{(\omega h)^{4N_o}}{4N_o + 1} + \cdots,$$

where $N_o = \lfloor (p+1)/2 \rfloor$, and

$$\mathcal{E}_e^p = \frac{1}{2}\left[\frac{(2N_e+1)!}{(4N_e+2)!}\right]^2 \frac{(\omega h)^{4N_e+4}}{4N_e + 3} + \cdots,$$

where $N_e = \lfloor p/2 \rfloor$. It then follows that

$$\cos kh - \cos \omega h = \left(\frac{\omega h}{2}\right)^2 \mathcal{E}_o^p + \mathcal{E}_e^p + \cdots,$$

where

$$\left(\frac{\omega h}{2}\right)^2 \mathcal{E}_o^p = \frac{1}{2}\left[\frac{(2N_o)!}{(4N_o)!}\right]^2 \frac{(\omega h)^{4N_o+2}}{4N_o + 1} + \cdots,$$

and $\mathcal{E}_e^p$ is given above. There are two cases, depending on the parity of the polynomial order p:

- If p is even, then $2N_e = 2N_o = p$, and the term involving $\mathcal{E}_o^p$ dominates the error, giving

$$\cos kh - \cos \omega h = \frac{1}{2}\left[\frac{p!}{(2p)!}\right]^2 \frac{(\omega h)^{2p+2}}{2p+1} + \cdots.$$

- If p is odd, then $2N_e = p - 1$ and $2N_o = p + 1$, and the term involving $\mathcal{E}_e^p$ now dominates the error, giving

$$\cos kh - \cos \omega h = \frac{1}{2}\left[\frac{p!}{(2p)!}\right]^2 \frac{(\omega h)^{2p+2}}{2p+1} + \cdots.$$

This concludes the proof of (3.3). Estimate (3.4) then follows immediately from (3.3) using the approximation

$$\cos kh - \cos \omega h = -(kh - \omega h)\sin \omega h + \cdots,$$

valid for small $kh - \omega h$. ∎

In examining this proof, we observe that the leading term in the remainder is the same, regardless of the parity of the polynomial order p. This effect occurs despite terms of different parities alternately dominating the error.

4.4. Error for large ωh. Now fix $\kappa = \omega h/2 \gg 1$. Equation (4.17) implies that, with $\kappa = \omega h/2$,

$$\mathcal{E}_e^p(\kappa)\frac{\cos^2(\omega h/2)}{\omega h} = -Q_{2N_e+3/2}(\kappa)\left\{1 - Q_{2N_e+3/2}(\kappa)\tan\kappa\right\}^{-1}, \tag{4.21}$$

while (4.18) implies that

$$\mathcal{E}_o^p(\kappa)\frac{\sin^2(\omega h/2)}{\omega h} = -Q_{2N_o+1/2}(\kappa)\{1 + Q_{2N_o+1/2}(\kappa)\cot\kappa\}^{-1}. \tag{4.22}$$

Theorem 3.3 is proved using the foregoing results along with estimates for the behavior of the quotient Q_m studied in Theorem A.2 of the appendix.

Proof. First, consider the preasymptotic regime where $2p+1 < \omega h - o(\omega h)^{1/3}$. For p in this range, neither $2N_e$ nor $2N_o$ exceeds $\kappa - o(\kappa^{1/3})$, where $\kappa = \omega h/2$. Therefore, we are in the situation covered by the first part of Theorem A.2, where both $Q_{2N_o+1/2}(\kappa)$ and $Q_{2N_e+3/2}(\kappa)$ oscillate but do not decay. Consequently, with the aid of the identities (4.21)–(4.22) and the expression for the error given in (4.20), we are led to the conclusion that $\mathcal{E}^p$ oscillates, but does not decay, as p is increased in this range.

For p in the transition region where $\omega h - o(\omega h)^{1/3} < 2p+1 < \omega h + o(\omega h)^{1/3}$, it follows that both $2N_e$ and $2N_o$ lie in the transition region $[\kappa - o(\kappa^{1/3}), \kappa + o(\kappa^{1/3})]$ dealt with in the second part of Theorem A.2. Here, the term appearing in the denominator of (4.20) is $\mathcal{O}(1)$ for $\omega h \gg 1$. The identities (4.21)–(4.22) show that the error in the dispersion relation is dictated by the behavior of the sum of $Q_{2N_e+3/2}(\kappa)$ and $Q_{2N_o+1/2}(\kappa)$. Applying Theorem A.2, we conclude that the error decays algebraically at a rate $\mathcal{O}(p^{-1/3})$.

The proof in the case where $2p+1 > \omega h + o(\omega h)^{1/3}$ follows along the same lines as the argument used in the transition region and will not be elaborated upon further. ∎

Appendix A. Behavior of $Q_m(\kappa)$. The quotient Q_m defined by

$$Q_m(\kappa) = \frac{J_m(\kappa)}{Y_m(\kappa)}, \quad m = \text{integer} + \frac{1}{2}, \tag{A.1}$$

appears in the expression for the error in the Padé approximants considered in section 4. The following estimate, valid for small values of κ, was used in the proof of Theorem 4.2.

LEMMA A.1. *Let $m = \text{integer} + 1/2$, and let Q_m be defined as above. Then, for $\kappa \ll 1$,*

$$Q_m(\kappa) = -\frac{1}{2}\left[\frac{(m-\frac{1}{2})!}{(2m-1)!}\right]^2 \frac{(2\kappa)^{2m}}{2m} + \cdots. \tag{A.2}$$

Proof. Write $m = n + 1/2$, where $n \in \mathbb{Z}$. For small κ, identity (8.440) of [15] gives

$$J_{n+1/2}(\kappa) = \frac{1}{\Gamma(3/2+n)}\left(\frac{\kappa}{2}\right)^{n+1/2} + \cdots,$$

while combining identities $(8.465)_1$ and (8.440) of [15] gives

$$Y_{n+1/2}(\kappa) = (-1)^{n-1}J_{-n-1/2}(\kappa) = \frac{(-1)^{n-1}}{\Gamma(1/2-n)}\left(\frac{\kappa}{2}\right)^{-n-1/2} + \cdots,$$

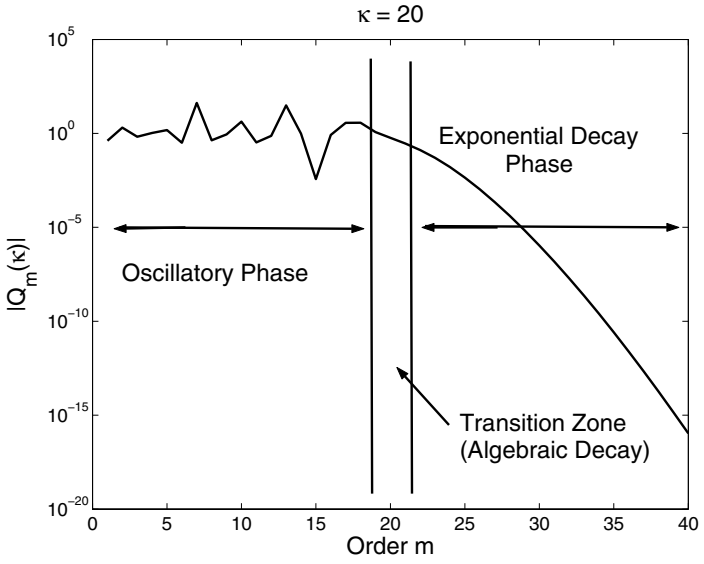


FIG. A.1. *Graph showing the three phases in the behavior of* $|Q_m(\kappa)|$ *for* $\kappa = 20$ *as the order* m *is increased.*

where Γ denotes the gamma function. Therefore, using formulas (8.339) of [15] gives, after some simplification,

$$Q_m(\kappa) = -\frac{1}{2}\left[\frac{n!}{(2n)!}\right]^2 \frac{(2\kappa)^{2n+1}}{2n+1} + \cdots,$$

and rewriting in terms of m gives the result claimed. □

Lemma A.1 shows that $Q_m(\kappa)$ decays algebraically as κ becomes small. However, it will be useful to consider the ratio in the regime $\kappa \gg 1$, with particular attention to the behavior as the order m of the Bessel functions becomes large. Figure A.1 shows the behavior of $Q_m(\kappa)$ when $\kappa = 20$ as the order m is increased. It is found that there are three distinct phases, depending on the size of the order m. Initially, $Q_m(\kappa)$ oscillates around unity. As the order m passes through κ, there is a relatively short-lived transition zone where $Q_m(\kappa)$ begins to decay at an algebraic rate as m is increased. Finally, as m is increased further, $Q_m(\kappa)$ decays at an exponential rate.

Our objective in the remainder of this section will be to show that the behavior observed in this particular case is typical. The following result provides sharp estimates for the values at which the different phases occur and quantifies the rates of decay.

THEOREM A.2. *Let* Q_m *be defined as above, and* $m = \text{integer} + 1/2$. *Then, as* m *is increased,* $Q_m(\kappa)$ *passes through three phases*:

1. *For* $m < \kappa - o(\kappa^{1/3})$, $Q_m(\kappa)$ *oscillates around unity but does not decay as* m *is increased.*

2. *For* $\kappa - o(\kappa^{1/3}) < m < \kappa + o(\kappa^{1/3})$, $Q_m(\kappa)$ *decays algebraically at a rate* $\mathcal{O}(m^{-1/3})$. *More precisely,*

$$Q_m(\kappa) \approx -\frac{1}{\sqrt{3}} - \frac{1}{\pi}\Gamma\left(\frac{2}{3}\right)^2 (\kappa - m)\left(\frac{6}{m}\right)^{1/3} + \cdots. \tag{A.3}$$

3. *For $m > \kappa + o(\kappa^{1/3})$, $Q_m(\kappa)$ decays at a superexponential rate:*

$$Q_m(\kappa) \approx -\frac{1}{2}\left[\frac{1-\sqrt{1-\kappa^2/m^2}}{1+\sqrt{1-\kappa^2/m^2}}e^{2\sqrt{1-\kappa^2/m^2}}\right]^m \tag{A.4}$$

so that, for $m \gg \kappa$,

$$Q_m(\kappa) \approx -\frac{1}{2}\left[\frac{\kappa e}{2m}\right]^{2m}. \tag{A.5}$$

The proof of this result is divided into four distinct cases covered in the following sections.

A.1. Preasymptotic regime: $m < \kappa$. We start by discussing the behavior of $Q_m(\kappa)$ in the preasymptotic regime, where the value of the argument κ exceeds the order m of the Bessel functions. Langer's formulas [13, sect. 7.13.4] provide uniform asymptotic expansions for Bessel functions of large order and large argument, and they give

$$Q_m(\kappa) = \frac{J_{1/3}(z)\cos(\pi/6) - Y_{1/3}(z)\sin(\pi/6) + \mathcal{O}(m^{-4/3})}{J_{1/3}(z)\sin(\pi/6) + Y_{1/3}(z)\cos(\pi/6) + \mathcal{O}(m^{-4/3})}, \tag{A.6}$$

where

$$z = m(w - \tan^{-1} w) \qquad \text{and} \qquad w = \sqrt{\kappa^2/m^2 - 1}.$$

Bounds on the accuracy of the approximation obtained when the higher order terms are dropped in (A.6) could be obtained using the uniform asymptotic expansions with the remainder given in Olver [27]. However, we shall content ourselves with making the approximation

$$Q_m(\kappa) \approx \frac{J_{1/3}(z)\cos(\pi/6) - Y_{1/3}(z)\sin(\pi/6)}{J_{1/3}(z)\sin(\pi/6) + Y_{1/3}(z)\cos(\pi/6)}. \tag{A.7}$$

A.1.1. Oscillatory phase: $m < \kappa - o(\kappa^{1/3})$. In the preasymptotic range where m is small relative to κ, the ratio $Q_m(\kappa)$ tends to oscillate and has magnitude of order unity. While it is difficult to make quantitative statements concerning the erratic behavior observed in Figure A.1, it is possible to give a qualitative explanation. When m is small relative to κ, the argument z of the Bessel functions appearing in (A.7) will be large and positive. Asymptotic expansions for Bessel functions of large argument are given in $(8.440)_1$ of [15]:

$$J_\nu(z) \sim \left(\frac{\pi z}{2}\right)^{-1/2}\cos\left(z - \frac{1}{2}\nu\pi - \frac{\pi}{4}\right)$$

and in $(8.440)_2$ of [15]:

$$Y_\nu(z) \sim \left(\frac{\pi z}{2}\right)^{-1/2}\sin\left(z - \frac{1}{2}\nu\pi - \frac{\pi}{4}\right).$$

Together with (A.7), these expressions show that $Q_m(\kappa)$ will tend to oscillate without a decay in the magnitude as m is increased. Indeed, inserting these expressions into the right-hand side of (A.7) and simplifying gives

$$\cot\left(z - \frac{\pi}{4}\right). \tag{A.8}$$

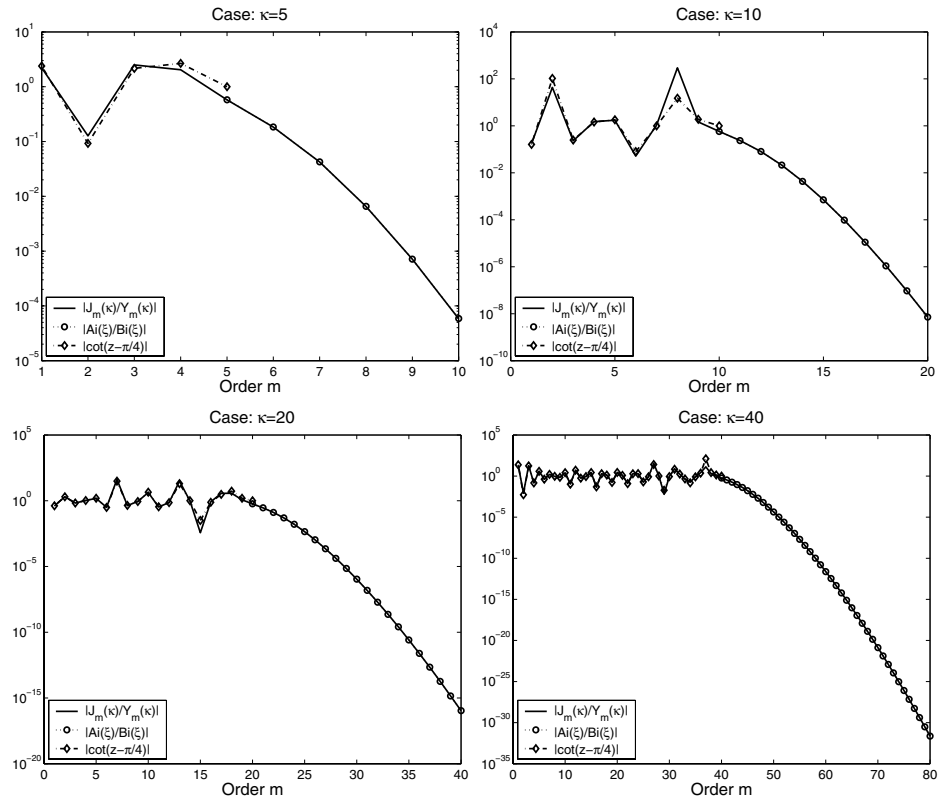


FIG. A.2. *Graphs of* $|Q_m(\kappa)|$ *in* (A.1) *for* $m = 1, \ldots, 2\kappa$, $|\cot(z - \pi/4)|$ *in* (A.8) *for* $m = 1, \ldots, \kappa$ *and* $|\operatorname{Ai}(\xi)/\operatorname{Bi}(\xi)|$ *in* (A.11) *for* $m = \kappa + 1, \ldots, 2\kappa$. *Values of* $\kappa = 5$, 10, 20, *and* 40 *are shown. Observe the oscillatory behavior of* $|Q_m|$ *and the good qualitative agreement provided by the cotangent in the preasymptotic regime* $m < \kappa$. *Furthermore, note the quantitative agreement between* $|Q_m|$ *and* $|\operatorname{Ai}(\xi)/\operatorname{Bi}(\xi)|$ *in the asymptotic regime* $m > \kappa$.

Of course, we would not expect this expression to necessarily agree closely with $Q_m(\kappa)$. Nevertheless, this expression actually provides a surprisingly good representation of the qualitative behavior in the preasymptotic regime even in the case of relatively modest values of κ, as shown in Figure A.2.

A.1.2. Transition zone: $\kappa - o(\kappa^{1/3}) < m < \kappa$. We consider the behavior in the zone where m approaches κ from below. For m in this range,

$$1 < \frac{\kappa}{m} < 1 + o(m^{-2/3})$$

so that

$$w \approx \left[\frac{\kappa - m}{(m/2)}\right]^{1/2} = o(1),$$

and therefore

$$z \approx \frac{1}{3} m w^3 \approx \frac{2}{3}\left[\frac{\kappa - m}{(m/2)^{1/3}}\right]^{3/2} = o(1).$$

Using the series representations for Bessel functions [15, eq. (8.440)], we obtain

$$Q_m(\kappa) \approx -\frac{1}{\sqrt{3}} - \frac{3}{\pi}\Gamma\left(\frac{2}{3}\right)^2\left(\frac{z}{2}\right)^{2/3} + \cdots,$$

and, by substituting for z and simplifying, we arrive at the conclusion:

$$Q_m(\kappa) \approx -\frac{1}{\sqrt{3}} - \frac{1}{\pi}\Gamma\left(\frac{2}{3}\right)^2(\kappa - m)\left(\frac{6}{m}\right)^{1/3} + \cdots, \tag{A.9}$$

valid for $\kappa - o(\kappa^{1/3}) < m < \kappa$. As a matter of fact, this result could also have been obtained formally using Nicholson's formulas [13, sect. 7.13.3].

A.2. Asymptotic regime: $m > \kappa$. We now study the behavior in the regime where the order of the Bessel functions exceeds the argument. Langer's formulas [13, sect. 7.13.4] imply that

$$Q_m(\kappa) = -\frac{\pi^{-1}K_{1/3}(z) + \mathcal{O}(m^{-4/3})}{I_{1/3}(z) + I_{1/3}(z) + \mathcal{O}(m^{-4/3})}, \tag{A.10}$$

where, in this case,

$$z = m\left(\tanh^{-1} w - w\right) \quad \text{and} \quad w = \sqrt{1 - \kappa^2/m^2}.$$

As before, it is possible to use the results of Olver [27] to obtain bounds on the accuracy of the approximation obtained when the higher order terms are dropped in (A.10), although we will not pursue this further here. Writing $z = \frac{2}{3}\xi^{3/2}$ and using formulas (11.1.04) and (11.1.12) from [27] gives

$$Q_m(\kappa) \approx -\frac{\pi^{-1}K_{1/3}(z)}{I_{1/3}(z) + I_{1/3}(z)} = -\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)}, \tag{A.11}$$

where Ai and Bi denote Airy functions of the first and second kinds, respectively [15]. The accuracy of this approximation is indicated in Figure A.2.

The behavior of the ratio of Airy functions for positive ξ may be deduced from the results quoted in Olver [27, pp. 392–393]. However, the following simple approximations will suffice for present purposes. For small argument $\xi \le 2$,

$$\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)} \approx \frac{1}{\sqrt{3}}\frac{1 - 3^{5/6}\Gamma\left(\frac{2}{3}\right)^2\frac{\xi}{2\pi} + \frac{\xi^3}{12}}{1 + 3^{5/6}\Gamma\left(\frac{2}{3}\right)^2\frac{\xi}{2\pi} + \frac{\xi^3}{12}}, \tag{A.12}$$

while for larger arguments where $\xi > 2$,

$$\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)} \approx \frac{e^{-2z}}{2}\frac{1 - \frac{15}{216z}}{1 + \frac{15}{216z}}, \quad z = \frac{2}{3}\xi^{3/2}. \tag{A.13}$$

Here Γ denotes the gamma function [15]. Together, these approximations provide an accurate picture of the behavior of the ratio throughout the full range of argument, as indicated in Figure A.3. In particular, it is observed that the ratio initially decays linearly,

$$\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)} \approx \frac{1}{\sqrt{3}} - \frac{3^{1/3}}{\pi}\Gamma\left(\frac{2}{3}\right)^2\xi, \tag{A.14}$$

before undergoing a rapid transition to an exponential rate of decay given by

$$\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)} \approx \frac{e^{-2z}}{2}. \tag{A.15}$$

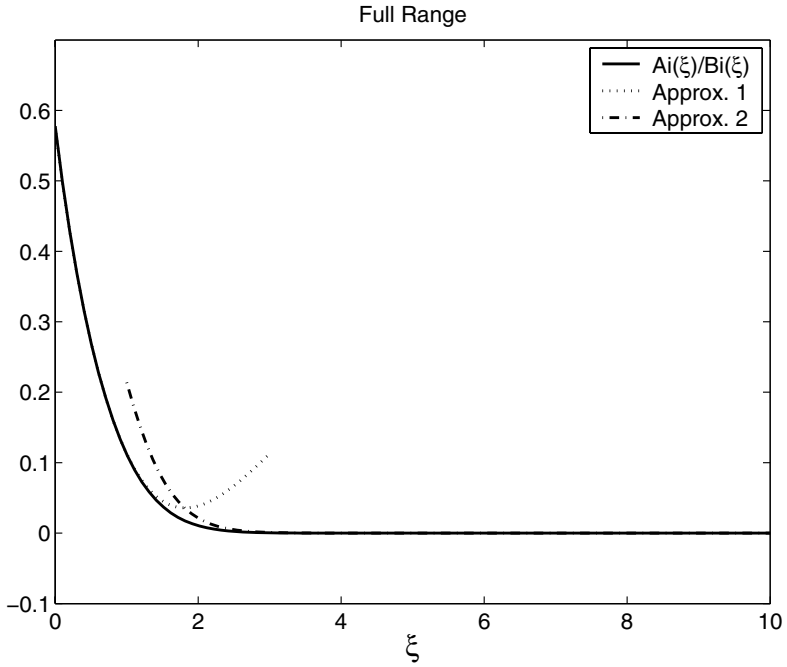


FIG. A.3. *Graph showing the ratio* $\mathrm{Ai}(\xi)/\mathrm{Bi}(\xi)$ *compared with the approximations for small and large arguments given in* (A.12) *and* (A.13)*, respectively.*

A.2.1. Transition zone: $\kappa < m < \kappa + o(\kappa^{1/3})$. As the order m passes through κ, we have

$$1 - o(m^{-2/3}) < \frac{\kappa}{m} < 1,$$

and, using similar arguments to those used before, we obtain

$$z \simeq \frac{1}{3} m w^3 \simeq \frac{2}{3} \left[\frac{m-\kappa}{(m/2)^{1/3}} \right]^{3/2}$$

or, equally well,

$$\xi \simeq \left(\frac{2}{m} \right)^{1/3} (m - \kappa).$$

Since $m - \kappa = o(\kappa^{1/3}) = o(m^{1/3})$, it follows that $\xi = o(1)$ and, using (A.14), we obtain

$$Q_m(\kappa) \approx -\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)} \simeq -\frac{1}{\sqrt{3}} + \frac{1}{\pi} \Gamma \left(\frac{2}{3} \right)^2 (m-\kappa) \left(\frac{6}{m} \right)^{1/3} + \cdots, \tag{A.16}$$

valid for $\kappa < m < \kappa + o(\kappa^{1/3})$. This form agrees with the result (A.9) obtained when m lies in the transition region to the left of κ.

A.2.2. Exponential decay phase: $m > \kappa + o(\kappa^{1/3})$. If m exceeds $\kappa + o(\kappa^{1/3})$, then w is no longer small, and, in turn, z and ξ will be large. By substituting for

z in terms of w in the expression (A.15) and applying elementary manipulations, we arrive at

$$\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)} \approx \frac{1}{2}\left[\frac{1-w}{1+w}e^{2w}\right]^m$$

or, substituting for w,

$$\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)} \approx \frac{1}{2}\left[\frac{1-\sqrt{1-\kappa^2/m^2}}{1+\sqrt{1-\kappa^2/m^2}}e^{2\sqrt{1-\kappa^2/m^2}}\right]^m.$$

The function $f : w \to (1-w)/(1+w)\exp(2w)$ is monotonic decreasing on $[0,1]$ from 1 to 0. Therefore, the term in parentheses is less than unity, and we have an exponential rate of decay when m is close to κ. This rate of decay increases as m becomes even larger relative to κ, and in the limiting case we find that

$$f\left(\sqrt{1-\kappa^2/m^2}\right) \simeq \left[\frac{\kappa e}{2m}\right]^2.$$

Therefore, when $m > \kappa e/2$, we obtain a superexponential rate of decay:

$$\frac{\mathrm{Ai}(\xi)}{\mathrm{Bi}(\xi)} \approx \frac{1}{2}\left[\frac{\kappa e}{2m}\right]^{2m}.$$

REFERENCES

[1] N.N. Abboud and P.M. Pinsky, *Finite element dispersion analysis for the* 3*-dimensional* 2*nd-order scalar wave-equation*, Internat. J. Numer. Methods Engrg., 35 (1992), pp. 1183–1218.

[2] M. Ainsworth, *Dispersive properties of high order Nédélec/edge element approximation of the time-harmonic Maxwell equations*, Philos. Trans. Roy. Soc. London Ser. A, to appear.

[3] J. Astley, K. Gerdes, D. Givoli, and I. Harari, *Special issue on finite elements for wave problems–Preface*, J. Comput. Acoust., 8 (2000), pp. vii–ix.

[4] I. Babuška and F. Ihlenburg, *Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation*, Internat. J. Numer. Methods Engrg., 38 (1995), pp. 3745–3774.

[5] G.A. Baker and P. Graves-Morris, *Padé Approximants*, 2nd ed., Encyclopedia Math. Appl. 59, Cambridge University Press, Cambridge, UK, 1995.

[6] M.A. Christon, *The influence of the mass matrix on the dispersive nature of the semi-discrete, second-order wave equation*, Comput. Methods Appl. Mech. Engrg., 173 (1999), pp. 146–166.

[7] G. Cohen and P. Monk, *Gauss point mass lumping schemes for Maxwell's equations*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 63–88.

[8] G. Cohen and P. Monk, *Mur-Nédélec finite element schemes for Maxwell's equations*, Comput. Methods Appl. Mech. Engrg., 169 (1999), pp. 197–217.

[9] G.C. Cohen, *Higher-Order Numerical Methods for Transient Wave Equations*, Springer Series in Scientific Computation, Springer-Verlag, Berlin, Heidelberg, New York, 2002.

[10] L. Demkowicz and L. Vardapetyan, *Modeling of electromagnetic absorption/scattering problems using hp-adaptive finite elements*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 103–124.

[11] A. Deraemaeker, I. Babuška, and P. Bouillard, *Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions*, Internat. J. Numer. Methods Engrg., 46 (1999), pp. 471–499.

[12] R.W. Dyson, *Technique for very high order nonlinear simulation and validation*, J. Comput. Acoust., 10 (2002), pp. 211–229.

[13] A. Erdélyi, W. Magnus, F. Oberhettinger, and F.G. Tricomi, *Higher Transcendental Functions*, Bateman Manuscript Project, McGraw-Hill, New York, London, 1953–55.

[14] K. Gerdes and F. Ihlenburg, *On the pollution effect in finite element solutions of the* 3*D Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 170 (1999), pp. 155–172.

[15] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products*, 5th ed., A. Jeffrey, ed., Academic Press, Boston, 1994.
[16] I. Harari, *Reducing spurious dispersion, anisotropy and reflection in finite element analysis of time-harmonic acoustics*, Comput. Methods Appl. Mech. Engrg., 140 (1997), pp. 39–58.
[17] I. Harari, *Finite element dispersion of cylindrical and spherical acoustic waves*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2533–2542.
[18] I. Harari and D. Avraham, *High-order finite element methods for acoustic problems*, J. Comput. Acoust., 5 (1997), pp. 33–51.
[19] I. Harari and T.J.R. Hughes, *Finite-element methods for the Helmholtz equation in an exterior domain-model problems*, Comput. Methods Appl. Mech. Engrg., 87 (1991), pp. 59–96.
[20] J.S. Hesthaven and T. Warburton, *Nodal high-order methods on unstructured grids,* I. *Time-domain solution of Maxwell's equations*, J. Comput. Phys., 181 (2002), pp. 186–221.
[21] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave-number. Part* 1. *The h-version of the finite element method*, Comput. Math. Appl., 30 (1995), pp. 9–37.
[22] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number Part* II: *The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.
[23] S. Koc, J. Song, and W.C. Chew, *Error analysis for the numerical evaluation of the diagonal forms of the scalar spherical addition theorem*, SIAM J. Numer. Anal., 36 (1999), pp. 906–921.
[24] P.B. Monk and A.K. Parrott, *A dispersion analysis of finite element methods for Maxwell's equations*, SIAM J. Sci. Comput., 15 (1994), pp. 916–937.
[25] A.A. Oberai and P.M. Pinsky, *A numerical comparison of finite element methods for the Helmholtz equation*, J. Comput. Acoust., 8 (2000), pp. 211–221.
[26] F. Odeh and J.B. Keller, *Partial differential equations with periodic coefficients and Bloch waves in crystals*, J. Math. Phys., 5 (1964), pp. 1499–1504.
[27] F.W.J. Olver, *Asymptotics and Special Functions*, Computer Science and Applied Mathematics, Academic Press, New York, London, 1974.
[28] C. Schwab and M. Suri, *The p and hp versions of the finite element method for problems with boundary layers*, Math. Comp., 65 (1996), pp. 1403–1429.
[29] L.L. Thompson and P.M. Pinsky, *Complex wavenumber Fourier analysis of the p-version finite element method*, Comput. Mech., 13 (1994), pp. 255–275.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 576–598

ANALYSIS OF A TWO-SCALE, LOCALLY CONSERVATIVE SUBGRID UPSCALING FOR ELLIPTIC PROBLEMS*

TODD ARBOGAST†

Abstract. We present a two-scale theoretical framework for approximating the solution of a second order elliptic problem. The elliptic coefficient is assumed to vary on a scale that can be resolved on a fine numerical grid, but limits on computational power require that computations be performed on a coarse grid. We consider the elliptic problem in mixed variational form over $W \times \mathbf{V} \subset L^2 \times H(\mathrm{div})$. We base our scale expansion on local mass conservation over the coarse grid. It is used to define a direct sum decomposition of $W \times \mathbf{V}$ into coarse and "subgrid" subspaces $W_c \times \mathbf{V}_c$ and $\delta W \times \delta \mathbf{V}$ such that (1) $\nabla \cdot \mathbf{V}_c = W_c$ and $\nabla \cdot \delta \mathbf{V} = \delta W$, and (2) the space $\delta \mathbf{V}$ is locally supported over the coarse mesh. We then explicitly decompose the variational problem into coarse and subgrid scale problems. The subgrid problem gives a well-defined operator taking $W_c \times \mathbf{V}_c$ to $\delta W \times \delta \mathbf{V}$, which is localized in space, and it is used to upscale, that is, to remove the subgrid from the coarse-scale problem. Using standard mixed finite element spaces, two-scale mixed spaces are defined. A mixed approximation is defined, which can be viewed as a type of *variational multiscale method* or a *residual-free bubble technique*. A numerical Green's function approach is used to make the approximation to the subgrid operator efficient to compute. A mixed method π-operator is defined for the two-scale approximation spaces and used to show optimal order error estimates.

Key words. second order elliptic, two-scale expansion, upscaling, subgrid, mixed method, variational multiscale method, numerical Green's function

AMS subject classifications. 65N15, 65N30, 35J20

DOI. 10.1137/S0036142902406636

1. Introduction.

Many mathematical models and numerical schemes have appeared in the literature that can capture fine-scale phenomena on coarse scales or grids. This is the essence of *upscaling*. The change-of-scale problem goes back to the beginning of mathematical modeling; however, research on it has recently seen a renewed and widespread resurgence.

Among many approaches, numerical techniques have been developed and exploited. For second order elliptic equations, a certainly not exhaustive list includes the multiscale finite element method [20], the residual-free bubble techniques [9], certain domain-decomposition techniques [27, 31], the two-grid techniques [30, 18], and a posteriori modeling techniques [25, 26]. A scheme related directly to the work here is the variational multiscale finite element method [21, 22, 23]. Each scheme can be viewed as a subgrid technique in the sense that each attempts to resolve scales below the coarse grid scale by incorporating local computations into a global problem defined only on a coarse grid.

A new subgrid technique for upscaling an elliptic partial differential equation based on a certain combination of low order mixed finite elements was introduced in [5] and [1]. It involves the decomposition of the solution operator into two parts, one representing the coarse scale and the other representing the subgrid scale. The method is described in general terms, and numerical tests are given that demonstrate

*Received by the editors April 29, 2002; accepted for publication (in revised form) September 22, 2003; published electronically April 14, 2004. This work was supported by the U.S. National Science Foundation under grants DMS-9707015 and SBR-9873326.

http://www.siam.org/journals/sinum/42-2/40663.html

†Department of Mathematics, University of Texas, 1 University Station C1200, Austin, TX 78712 and Institute for Computational Engineering and Sciences, University of Texas, 1 University Station C0200, Austin, TX 78712 (arbogast@ices.utexas.edu).

the overall speed and convergence properties of the method in [5, 1]. Applications to groundwater contaminant transport and petroleum simulation are given in [3, 4], wherein it is shown that the method has great potential to resolve fine-scale effects in practical problems. Complete details of implementation are presented in [2], as well as additional and more stringent numerical tests that apply the technique to two-phase porous medium problems with significant heterogeneity and wells. An advantage of this subgrid technique is that it needs no assumptions about the underlying physics. The data used in the simulation is to be provided directly on the fine scale.

The goals of this paper are threefold. First, we present a theoretical framework within which to understand the upscaling process. We achieve upscaling *without* the need for an explicit closure assumption or a restrictive assumption such as periodicity or the like. In general, it is difficult to analyze the errors introduced by a closure assumption; however, this problem does *not* arise here. Rather the ability of the upscaled model to capture fine-scale features in the solution becomes a question of approximation theory: how well do we approximate the upscaled model? Second, we generalize the mixed finite element technique of [5, 1] to essentially arbitrary choices of mixed spaces. Because of the upscaling framework, these methods can be implemented very efficiently and require the solution of a global problem defined only on the coarse grid. Finally, we provide an error analysis showing optimal order approximation.

Both an outline of the paper and a brief summary of results follow. After presenting in the next section the elliptic problem in mixed variational form posed in $W \times \mathbf{V} \subset L^2 \times H(\mathrm{div})$, we then proceed in section 3 to define our framework within which we upscale the differential problem. We define the coarse grid we can ultimately compute over and use it to define a direct sum decomposition of $W \times \mathbf{V}$ into coarse and subgrid subspaces $W_c \times \mathbf{V}_c$ and $\delta W \times \delta \mathbf{V}$ such that (1) the divergence constraints $\nabla \cdot \mathbf{V}_c = W_c$ and $\nabla \cdot \delta \mathbf{V} = \delta W$, needed for local mass conservation over the coarse and subgrid scales, and (2) the space $\delta \mathbf{V}$ is locally supported over the coarse mesh, which is needed for upscaling the subgrid. This then leads to a decomposition of the variational problem into coarse and subgrid scale problems, with solutions in $W_c \times \mathbf{V}_c$ and $\delta W \times \delta \mathbf{V}$, although the two problems remain coupled.

We define in section 4 the δ-solution operator as the solution of the subgrid problem. It is used to relate the subgrid to the coarse solution, and it is a well-defined operator that takes $W_c \times \mathbf{V}_c$ to $\delta W \times \delta \mathbf{V}$. Since this operator is localized in space, it can be used to control the fine scales. We use it in the coarse problem to remove direct reference to the subgrid, resulting in the upscaled problem involving only the coarse-scale solution.

In section 5 we exploit the two-scale structure of the solution to define an efficient mixed finite element method. We use any of the usual mixed elements to approximate the δ-solution operator and also any choice of mixed spaces to approximate the upscaled coarse solution. This defines many families of two-scale, mixed spaces. Our approximation can be viewed as a type of variational multiscale method [21, 22] or a residual-free bubble technique [9]. A numerical Green's function approach makes the approximation to the subgrid operator efficient to compute.

In section 6 we analyze the approximation error. We show optimal order a priori error estimates. Care must be taken, as the two-scale decomposition depends on the coarse grid. We therefore analyze the combined system, showing approximation of the full solution. The key development here is the definition of a suitable mixed method π-operator that preserves the L^2-projection of the discrete divergence and approximates

well in the two-scale context. Finally, in section 7 we apply the convergence theory to the special cases defined in [5] and [1].

2. A second order elliptic equation. Let $\Omega \subset \mathbb{R}^n$, $n = 2$ or 3, be a convex polygonal domain. Throughout the paper, for domain ω, we denote by $L^p(\omega)$ the usual Lebesgue space of index p, $1 \le p \le \infty$, and by $W^{k,p}(\omega)$ the usual Sobolev space of k weak derivatives in $L^p(\omega)$. We denote by $(\cdot,\cdot)_\omega$ the $L^2(\omega)$-inner product (i.e., Lebesgue integration over ω). Moreover, $\|\cdot\|_{k,\omega}$ is the norm of $H^k(\omega) \equiv W^{k,2}(\omega)$. In the notation we may suppress ω when it is Ω.

Decompose $\partial\Omega = \bar{\Gamma}_N \cup \bar{\Gamma}_R$, where Γ_N and Γ_R are disjoint open sets in $\partial\Omega$, and let ν be the outer unit normal vector. The problem is to find the unknown functions p (pressure) and $\mathbf{u}$ (velocity) satisfying

$$ap + \nabla \cdot \mathbf{u} = b \qquad \text{in } \Omega, \tag{2.1}$$

$$\mathbf{u} = -d(\nabla p - c) \quad \text{in } \Omega, \tag{2.2}$$

$$\mathbf{u} \cdot \nu = g_N \qquad \text{on } \Gamma_N, \tag{2.3}$$

$$\alpha \mathbf{u} \cdot \nu = p - g_R \quad \text{on } \Gamma_R, \tag{2.4}$$

wherein $a \in L^\infty(\Omega)$ is nonnegative, $b \in L^2(\Omega)$, $c \in (L^2(\Omega))^n$, and d is a second order uniformly positive definite symmetric tensor in $(L^\infty(\Omega))^{n\times n}$ (i.e., d and d^{-1} are both uniformly elliptic and uniformly bounded). The boundary conditions represent Neumann and Robin (and Dirichlet, if $\alpha = 0$) conditions for suitably nice functions g_N, g_R, and $\alpha \ge 0$. We assume that a unique and sufficiently regular solution to this system exists and that the coefficients are sufficiently regular for the error analysis to follow.

A special case arises if a vanishes identically on all of Ω and $\Gamma_N = \partial\Omega$. Then it is well known and follows from the divergence theorem that solvability requires the compatibility condition

$$\int_\Omega b(x)\,dx = \int_{\partial\Omega} g_N(x)\,ds(x). \tag{2.5}$$

In this case, we obtain p only up to an arbitrary constant.

To enforce conservation of mass (2.1) locally, we base our method on a mixed variational formulation. Let

$$H(\mathrm{div};\Omega) = \{\mathbf{v} \in (L^2(\Omega))^n : \nabla \cdot \mathbf{v} \in L^2(\Omega)\}$$

denote the usual space, with the inner product

$$(\mathbf{v}_1, \mathbf{v}_2)_{H(\mathrm{div})} = (\mathbf{v}_1, \mathbf{v}_2) + (\nabla \cdot \mathbf{v}_1, \nabla \cdot \mathbf{v}_2)$$

and norm $\|\mathbf{v}\|_{H(\mathrm{div})} = (\mathbf{v},\mathbf{v})^{1/2}_{H(\mathrm{div})}$, and let

$$\mathbf{V} = \{\mathbf{v} \in H(\mathrm{div};\Omega) : \mathbf{v} \cdot \nu = 0 \text{ on } \Gamma_N\},$$

which is a closed subspace. To impose the Neumann boundary condition, we need to extend g_N to some fixed vector $\mathbf{v}_{g_N} \in H(\mathrm{div};\Omega)$ such that

$$\mathbf{v}_{g_N} \cdot \nu = g_N \text{ on } \Gamma_N \quad \text{and} \quad \mathbf{v}_{g_N} \cdot \nu = 0 \text{ on } \Gamma_R.$$

Finally, let $W = L^2(\Omega)$, or let $W = L^2(\Omega)/\mathbb{R} = \{w \in L^2(\Omega) : \int_\Omega w(x)\,dx = 0\}$ if p will be defined only up to a constant.

The mixed variational problem equivalent to (2.1)–(2.4) is to find $\mathbf{u} \in \mathbf{V} + \mathbf{v}_{g_N}$ and $p \in W$ such that

$$(ap, w) + (\nabla \cdot \mathbf{u}, w) = (b, w) \qquad \forall\, w \in W, \tag{2.6}$$

$$(d^{-1}\mathbf{u}, \mathbf{v}) + (\alpha \mathbf{u} \cdot \nu, \mathbf{v} \cdot \nu)_{\Gamma_R} - (p, \nabla \cdot \mathbf{v}) = (c, \mathbf{v}) - (g_R, \mathbf{v} \cdot \nu)_{\Gamma_R} \quad \forall\, \mathbf{v} \in \mathbf{V}. \tag{2.7}$$

Note that (2.3) is imposed as an essential condition and (2.4) is imposed weakly as a natural boundary condition.

3. Separation of scales. We recall that a Hilbert space H is the direct sum of M and N if $H = M + N$ and M and N are closed subspaces that intersect only at the zero vector. We denote this fact by $H = M \oplus N$. In this case, given $x \in H$, there is some unique $m \in M$ and $n \in N$ such that $x = m + n$. We note the following result, which is an exercise in the application of the closed graph theorem [29].

PROPOSITION 3.1. *If H is a Hilbert space and $H = M \oplus N$, then the operator $\tilde{\mathcal{P}}_M : H \to M$ defined for $x \in H$ by $\tilde{\mathcal{P}}_M x = m$, where $x = m + n$, $m \in M$ and $n \in N$, is a bounded linear (but possibly nonorthogonal) projection.*

We expand functions in $W \times \mathbf{V}$ uniquely according to a direct sum decomposition of the spaces. We base our decomposition on our two primary requirements: that the finer (i.e., "subgrid") scales be localized and that mass conservation is maintained. To do so we choose a coarse mesh partition $\mathcal{T}_H$ of Ω of a finite number of convex elements over which we will decompose the solution into coarse and local (i.e., "subgrid") pieces. The choice is mostly arbitrary at this point, but later the mesh will be used as the coarse mesh we compute on. We do, however, need a nondegeneracy condition. We assume that there is some universal fixed constant $\gamma > 0$ such that any choice of $\mathcal{T}_H$ satisfies

$$\mathrm{msr}(E_c) \geq \gamma(\mathrm{diam}(E_c))^n \quad \forall E_c \in \mathcal{T}_H, \tag{3.1}$$

where $\mathrm{msr}(E_c)$ is the measure of E_c and $\mathrm{diam}(E_c)$ is its diameter.

3.1. A two-scale decomposition of $W \times \mathbf{V}$. As is well known, the divergence operator maps $\mathbf{V}$ onto W. The range of the divergence operator must be decomposed into a direct sum decomposition $W = W_c \oplus \delta W$ of closed subspaces. For our purposes, the decomposition is arbitrary, except that we must insist on two properties. First,

$$\delta W \subset (W_c^1)^\perp, \quad W_c^1 = \{w_c \in W : w_c \text{ is constant } \forall \text{ coarse elements } E_c \in \mathcal{T}_H\},$$

with respect to the $L^2(\Omega)$-inner product.

Second, we insist that there is a uniformity in the separation of W_c and δW. We define the possibly nonorthogonal projections

$$\tilde{\mathcal{P}}_{W_c} : W \to W_c \quad \text{and} \quad \tilde{\mathcal{P}}_{\delta W} : W \to \delta W \tag{3.2}$$

with respect to the direct sum decomposition. By Proposition 3.1, these operators are bounded but not necessarily uniformly so with respect to the coarse mesh $\mathcal{T}_H$ selected. Our requirement is that in fact these are bounded uniformly: there is some universal constant C, independent of the coarse mesh $\mathcal{T}_H$, such that

$$\|w_c\|_0 + \|\delta w\|_0 \leq C\|w\|_0, \tag{3.3}$$

where $w = w_c + \delta w \in W_c \oplus \delta W$. We can easily achieve this property if, for example, $W_c \subset W$ such that $W_c^1 \subset W_c$ is given arbitrarily and $\delta W = W_c^\perp$. However, we maintain flexibility by not assuming strict orthogonality.

To obtain a decomposition of $\mathbf{V}$, we first define

$$\mathbf{V}_c^1 = \{\mathbf{v} \in \mathbf{V} : \nabla \cdot \mathbf{v} \in W_c\},$$
$$\delta\mathbf{V}^1 = \{\delta\mathbf{v} \in \mathbf{V} : \nabla \cdot \delta\mathbf{v} \in \delta W \text{ and } \delta\mathbf{v} \cdot \nu = 0 \text{ on } \partial E_c \ \forall\ E_c \in \mathcal{T}_H\}.$$

PROPOSITION 3.2. *It follows that*
(a) $\mathbf{V}_c^1$ *and* $\delta\mathbf{V}^1$ *are closed subspaces of* $\mathbf{V}$*;*
(b) $\mathbf{V} = \mathbf{V}_c^1 + \delta\mathbf{V}^1$*;*
(c) $\mathbf{V}_c^1 \cap \delta\mathbf{V}^1 = \{\mathbf{v} \in \mathbf{V} : \nabla \cdot \mathbf{v} = 0$ *and* $\mathbf{v} \cdot \nu = 0$ *on* $\partial E_c \ \forall\ E_c \in \mathcal{T}_H\}$*;*
(d) $\nabla \cdot \mathbf{V}_c^1 = W_c$ *and* $\nabla \cdot \delta\mathbf{V}^1 = \delta W$.

Proof. For (a), first note that each space is a linear subspace. The divergence operator is continuous on $\mathbf{V}$, so $\mathbf{V}_c^1$ is closed. Finally, we note that $\delta\mathbf{V}^1$ is the intersection of a closed subspace of $\mathbf{V}$ (the vectors $\delta\mathbf{v}$ with $\nabla \cdot \delta\mathbf{v} \in \delta W$) and the kernel of a finite number of normal trace operators, so $\delta\mathbf{V}^1$ is also closed.

To see (d), we consider an auxiliary elliptic problem. Given $\delta w \in \delta W$, on each $E_c \in \mathcal{T}_H$ let $\varphi \in H^1(E_c)$ solve the linear problem

$$\Delta\varphi = \delta w \qquad \text{in } E_c, \tag{3.4}$$
$$\nabla\varphi \cdot \nu = 0 \qquad \text{on } \partial E_c. \tag{3.5}$$

This problem is solvable because $\delta w \perp W_c^1$, so δw satisfies the compatibility condition. Set $\delta\mathbf{v} = \nabla\varphi$. It is easy to conclude that $\delta\mathbf{v} \in \mathbf{V}$, since the normal traces match (in fact vanish) on each side of $\partial E_c \cap \Omega\ \forall E_c \in \mathcal{T}_H$. Thus we conclude that $\delta\mathbf{v} \in \delta\mathbf{V}^1$, and so $\delta W \subset \nabla \cdot \delta\mathbf{V}^1$. The opposite inclusion holds by definition, so $\nabla \cdot \delta\mathbf{V}^1 = \delta W$. Similarly, given $w_c \in W_c$, let $\psi \in H^1(\Omega)$ solve the linear problem

$$\Delta\psi = w_c \qquad \text{in } \Omega, \tag{3.6}$$
$$\nabla\psi \cdot \nu = 0 \qquad \text{on } \Gamma_N, \tag{3.7}$$
$$\psi = 0 \qquad \text{on } \Gamma_R. \tag{3.8}$$

Then $\mathbf{v}_c = \nabla\psi \in \mathbf{V}_c^1$ allows us to conclude that $\nabla \cdot \mathbf{V}_c^1 = W_c$.

For (b), we know that $\mathbf{V} \supset \mathbf{V}_c^1 + \delta\mathbf{V}^1$, so consider any $\mathbf{v} \in \mathbf{V}$ and decompose $\nabla \cdot \mathbf{v} = w_c + \delta w$ for $w_c \in W_c$ and $\delta w \in \delta W$. Construct φ and $\delta\mathbf{v} = \nabla\varphi$ from δw as in (3.4)–(3.5) above. Then we conclude that $\mathbf{v}_c = \mathbf{v} - \delta\mathbf{v} \in \mathbf{V}_c^1$, and so $\mathbf{V} \subset \mathbf{V}_c^1 + \delta\mathbf{V}^1$.

Finally, (c) follows trivially from (d), since $W_c \cap \delta W = \{0\}$. ∎

The proof above suggests the following Helmholtz decomposition. Let

$$\mathbf{V}_c^p = \{\mathbf{v}_c^p \in \mathbf{V}_c^1 : \mathbf{v}_c^p = \nabla\psi \text{ for some } w_c \in W_c \text{ and } \psi \text{ satisfying (3.6)–(3.8)}\},$$
$$\mathbf{V}^s = \{\mathbf{v}^s \in \mathbf{V} : \nabla \cdot \mathbf{v}^s = 0\} \subset \mathbf{V}_c^1.$$

These spaces are clearly closed subspaces, and we claim that $\mathbf{V}_c^p \cap \mathbf{V}^s = \{0\}$. Let $\mathbf{v}$ be a member of both. Then there is some scalar potential function ψ such that $\mathbf{v} = \nabla\psi$, and $\nabla \cdot \mathbf{v} = \Delta\psi = 0$. Moreover, the boundary conditions from $\mathbf{V}_c^p$ imply that ψ is constant (zero if $\Gamma_R \neq \emptyset$), and so $\mathbf{v} = 0$. Thus we conclude that in fact

$$\mathbf{V}_c^1 = \mathbf{V}_c^p \oplus \mathbf{V}^s$$

is a direct sum of potential and solenoidal vector fields. Similarly, we have the closed subspace

$$\delta\mathbf{V}^p = \{\delta\mathbf{v}^p \in \delta\mathbf{V}^1 : \delta\mathbf{v}^p = \nabla\varphi \text{ for some } \delta w \in \delta W \text{ and } \varphi \text{ satisfying (3.4)–(3.5)}\}$$

and the direct sum

$$\delta\mathbf{V}^1 = \delta\mathbf{V}^p \oplus \delta\mathbf{V}^s, \quad \text{where} \quad \delta\mathbf{V}^s = \mathbf{V}^s \cap \delta\mathbf{V}^1.$$

By similar reasoning, we conclude from $W_c \cap \delta W = \{0\}$ that $\mathbf{V}_c^p \cap \delta\mathbf{V}^p = \{0\}$, and thus

$$\mathbf{V} = \mathbf{V}_c^p \oplus \delta\mathbf{V}^p \oplus \mathbf{V}^s. \tag{3.9}$$

THEOREM 3.3. *There is some constant C, independent of $\mathcal{T}_H$, and there exist closed subspaces $\mathbf{V}_c$ and $\delta\mathbf{V}$ of $\mathbf{V}$ such that*

(a) $\mathbf{V} = \mathbf{V}_c \oplus \delta\mathbf{V}$;
(b) $\nabla \cdot \mathbf{V}_c = W_c$ *and* $\nabla \cdot \delta\mathbf{V} = \delta W$;
(c) $\delta\mathbf{V} \subset \delta\mathbf{V}^1 = \{\delta\mathbf{v} \in \mathbf{V} : \nabla \cdot \delta\mathbf{v} \in \delta W$ *and* $\delta\mathbf{v} \cdot \nu = 0$ *on* $\partial E_c \ \forall\ E_c \in \mathcal{T}_H\}$;
(d) *for* $\mathbf{v} = \mathbf{v}_c + \delta\mathbf{v} \in \mathbf{V}_c \oplus \delta\mathbf{V}$ *given,*

$$\|\mathbf{v}_c\|_{H(div)} + \|\delta\mathbf{v}\|_{H(div)} \le C\|\mathbf{v}\|_{H(div)}. \tag{3.10}$$

Moreover, a choice exists such that also the potential vector fields

(e) $\mathbf{V}_c^p \subset \mathbf{V}_c$ *and* $\delta\mathbf{V}^p \subset \delta\mathbf{V}$.

That is, (d) says that the projection operators defined by the direct sum,

$$\tilde{\mathcal{P}}_{\mathbf{V}_c} : \mathbf{V} \to \mathbf{V}_c \quad \text{and} \quad \tilde{\mathcal{P}}_{\delta\mathbf{V}} : \mathbf{V} \to \delta\mathbf{V}, \tag{3.11}$$

are bounded independently of $\mathcal{T}_H$.

Proof. The troublesome part of (3.9) are the solenoidal fields $\mathbf{V}^s$. With respect to the $H(\mathrm{div})$-inner product, let

$$\mathbf{V}_c^s = (\delta\mathbf{V}^s)^\perp \cap \mathbf{V}^s = \{\mathbf{v}_c^s \in \mathbf{V}^s : \mathbf{v}_c^s \perp \delta\mathbf{V}^s\}.$$

Then $\mathbf{V}^s = \mathbf{V}_c^s \oplus \delta\mathbf{V}^s$, and we can define

$$\mathbf{V}_c = \mathbf{V}_c^p \oplus \mathbf{V}_c^s \quad \text{and} \quad \delta\mathbf{V} = \delta\mathbf{V}^p \oplus \delta\mathbf{V}^s,$$

satisfying (a)–(c) and (e).

We need to examine the construction more carefully to conclude (d). Let $\mathbf{v} \in \mathbf{V}$ be given, and decompose

$$\nabla \cdot \mathbf{v} = w_c + \delta w = \tilde{\mathcal{P}}_{W_c} \nabla \cdot \mathbf{v} + \tilde{\mathcal{P}}_{\delta W} \nabla \cdot \mathbf{v} \in W_c \oplus \delta W.$$

We then construct $\delta\mathbf{v}^p = \nabla\varphi \in \delta\mathbf{V}^p$ from (3.4)–(3.5) using the given δw and note that standard elliptic energy estimates show that on each $E_c \in \mathcal{T}_H$,

$$\begin{aligned}\|\delta\mathbf{v}^p\|_{0,E_c}^2 &= \|\nabla\varphi\|_{0,E_c}^2 = (\tilde{\mathcal{P}}_{\delta W}\nabla \cdot \mathbf{v}, \varphi)_{E_c} \\ &\le \|\tilde{\mathcal{P}}_{\delta W}\nabla \cdot \mathbf{v}\|_{0,E_c}\|\varphi\|_{0,E_c} \le C\|\tilde{\mathcal{P}}_{\delta W}\nabla \cdot \mathbf{v}\|_{0,E_c}\|\nabla\varphi\|_{0,E_c},\end{aligned}$$

where C is the Poincaré inequality constant [17] for E_c, which is proportional to $\frac{\mathrm{diam}(E_c)^n}{\mathrm{msr}(E_c)^{1-1/n}}$ and therefore universally bounded by the nondegeneracy assumption (3.1). Thus, from (3.3),

$$\|\delta\mathbf{v}^p\|_{H(\mathrm{div})} \le C\|\tilde{\mathcal{P}}_{\delta W}\nabla \cdot \mathbf{v}\|_0 \le C\|\mathbf{v}\|_{H(\mathrm{div})}.$$

Similarly, we construct $\mathbf{v}_c^p = \nabla\psi \in \mathbf{V}_c^p$ from (3.6)–(3.8) using the given w_c and conclude that

$$\|\mathbf{v}_c^p\|_{H(\mathrm{div})} \le C\|\mathbf{v}\|_{H(\mathrm{div})}.$$

Now $\mathbf{v} = \mathbf{v}_c^p + \delta\mathbf{v}^p + \mathbf{v}^s \in \mathbf{V}_c^p \oplus \delta\mathbf{V}^p \oplus \mathbf{V}^s$ by (3.9), and $\mathbf{v}^s = \mathbf{v}_c^s + \delta\mathbf{v}^s \in \mathbf{V}_c^s \oplus \delta\mathbf{V}^s$, which is an orthogonal decomposition, so

$$\|\mathbf{v}_c^s\|_{H(\mathrm{div})}^2 + \|\delta\mathbf{v}^s\|_{H(\mathrm{div})}^2 = \|\mathbf{v}^s\|_{H(\mathrm{div})}^2 = \|\mathbf{v} - \mathbf{v}_c^p - \delta\mathbf{v}^p\|_{H(\mathrm{div})}^2 \le C\|\mathbf{v}\|_{H(\mathrm{div})}^2.$$

Finally,

$$\begin{aligned}\|\mathbf{v}_c\|_{H(\mathrm{div})} &= \|\mathbf{v}_c^p + \mathbf{v}_c^s\|_{H(\mathrm{div})} \le C\|\mathbf{v}\|_{H(\mathrm{div})},\\ \|\delta\mathbf{v}\|_{H(\mathrm{div})} &= \|\delta\mathbf{v}^p + \delta\mathbf{v}^s\|_{H(\mathrm{div})} \le C\|\delta\mathbf{v}\|_{H(\mathrm{div})},\end{aligned}$$

and the proof is complete. ∎

Thus (a) gives us a unique decomposition of vectors in $\mathbf{V}$, (b) allows us to enforce mass conservation over $\mathcal{T}_H$ on both the coarse and subgrid scales, (c) gives us a locality property of the space $\delta\mathbf{V}$ that we can exploit later, and (d) gives us a uniformity property of the decomposition independent of $\mathcal{T}_H$. The specific choice of decomposition appears to be unimportant for our purposes, although we will revisit this question later in section 5.1. In what follows, we fix a choice of $\mathbf{V}_c$ and $\delta\mathbf{V}$ satisfying the properties (a)–(d) of the theorem.

3.2. Separation of scales in the equations. Recall that $\mathbf{v}_{g_N} \in H(\mathrm{div};\Omega)$ satisfies the Neumann boundary condition. Decompose the solution

$$\begin{aligned}p &= p_c + \delta p \in W_c \oplus \delta W,\\ \mathbf{u} &= \mathbf{u}_c + \delta\mathbf{u} + \mathbf{v}_{g_N} \in \mathbf{V}_c \oplus \delta\mathbf{V} + \mathbf{v}_{g_N}.\end{aligned}$$

Then we decompose (2.6)–(2.7) by choosing test functions restricted to the spaces $W_c \times \mathbf{V}_c$ or $\delta W \times \delta\mathbf{V}$. This results in an equivalent system of the four equations (3.12)–(3.15) below. For convenience, let

$$\begin{aligned}b^* &= b - \nabla\cdot\mathbf{v}_{g_N},\\ c^* &= c - d^{-1}\mathbf{v}_{g_N}.\end{aligned}$$

Coarse-scale equations. Find $\mathbf{u}_c \in \mathbf{V}_c$ and $p_c \in W_c$ such that

$$(a(p_c+\delta p), w_c) + (\nabla\cdot(\mathbf{u}_c+\delta\mathbf{u}), w_c) = (b^*, w_c) \qquad \forall\ w_c \in W_c, \tag{3.12}$$

$$\begin{aligned}&(d^{-1}(\mathbf{u}_c+\delta\mathbf{u}), \mathbf{v}_c) + (\alpha\mathbf{u}_c\cdot\nu, \mathbf{v}_c\cdot\nu)_{\Gamma_R} - (p_c+\delta p, \nabla\cdot\mathbf{v}_c)\\ &\quad = (c^*, \mathbf{v}_c) - (g_R, \mathbf{v}_c\cdot\nu)_{\Gamma_R} \qquad \forall\ \mathbf{v}_c \in \mathbf{V}_c.\end{aligned} \tag{3.13}$$

Subgrid δ-scale equations. Find $\delta\mathbf{u} \in \delta\mathbf{V}$ and $\delta p \in \delta W$ such that

$$(a(p_c+\delta p), \delta w) + (\nabla\cdot(\mathbf{u}_c+\delta\mathbf{u}), \delta w) = (b^*, \delta w) \quad \forall\ \delta w \in \delta W, \tag{3.14}$$

$$(d^{-1}(\mathbf{u}_c+\delta\mathbf{u}), \delta\mathbf{v}) - (p_c+\delta p, \nabla\cdot\delta\mathbf{v}) = (c^*, \delta\mathbf{v}) \quad \forall\ \delta\mathbf{v} \in \delta\mathbf{V}. \tag{3.15}$$

4. The δ-solution operator and upscaling. The systems (3.12)–(3.13) and (3.14)–(3.15) are coupled together, and as written they do not allow us to exploit the locality of $\delta\mathbf{V}$. Our goal now is to rewrite (3.12)–(3.13) independently of δp and $\delta\mathbf{u}$. To do so, we need to write these quantities in terms of p_c and $\mathbf{u}_c$.

4.1. Solvability of the subgrid scale equations.

LEMMA 4.1. *Given* $(p_c, \mathbf{u}_c) \in W_c \times \mathbf{V}_c$, *there exists a unique solution* $(\delta p, \delta \mathbf{u}) \in \delta W \times \delta \mathbf{V}$ *to* (3.14)–(3.15). *Moreover, there is some constant* C, *independent of the coarse mesh* $\mathcal{T}_H$ *and the specific decomposition of* $W \times \mathbf{V}$ *selected, such that*

$$\|\delta p\|_0 + \|\delta \mathbf{u}\|_{H(div)} \leq C\{\|b\|_0 + \|c\|_0 + \|\mathbf{v}_{g_N}\|_{H(div)} + \|p_c\|_0 + \|\mathbf{u}_c\|_{H(div)}\}.$$

We can prove Lemma 4.1 using the theory of saddle point problems [6, 8, 13, 7]. We need a generalization of the theory developed in, e.g., [13]. Consider the following abstract problem: Find $\check{p} \in \check{W}$ and $\check{\mathbf{u}} \in \check{\mathbf{V}}$ such that

$$\check{c}(\check{p}, \check{w}) + (\nabla \cdot \check{\mathbf{u}}, \check{w}) = F(\check{w}) \quad \forall\ \check{w} \in \check{W}, \tag{4.1}$$

$$\check{a}(\check{\mathbf{u}}, \check{\mathbf{v}}) - (\check{p}, \nabla \cdot \check{\mathbf{v}}) = G(\check{\mathbf{v}}) \quad \forall\ \check{\mathbf{v}} \in \check{\mathbf{V}}, \tag{4.2}$$

where $\check{W} \subset L^2(\Omega)$ and $\check{\mathbf{V}} \subset H(\mathrm{div}; \Omega)$ are Hilbert spaces. The following result is a simple corollary of the more general theory [13, pp. 44–47].

THEOREM 4.2. *Suppose that* $\check{W} \subset L^2(\Omega)$ *and* $\check{\mathbf{V}} \subset H(\mathrm{div}; \Omega)$ *are Hilbert spaces such that* $\nabla \cdot \check{\mathbf{V}} = \check{W}$. *Suppose that* $\check{a}$ *and* $\check{c}$ *are continuous, symmetric, positive semidefinite bilinear forms on* $\check{\mathbf{V}} \times \check{\mathbf{V}}$ *and* $\check{W} \times \check{W}$, *respectively, and that* $\check{a}$ *is coercive on* $\check{\mathbf{V}} \cap \ker(\nabla \cdot)$, $G \in \check{\mathbf{V}}'$, $F \in \check{W}'$, *and there exists* $\beta > 0$ *such that*

$$\inf_{\check{w} \in \check{W}} \sup_{\check{\mathbf{v}} \in \check{\mathbf{V}}} \frac{(\nabla \cdot \check{\mathbf{v}}, \check{w})}{\|\check{\mathbf{v}}\|_{H(div)} \|\check{w}\|_0} \geq \beta > 0. \tag{4.3}$$

Then there exists a unique solution $(\check{p}, \check{\mathbf{u}}) \in \check{W} \times \check{\mathbf{V}}$ *to* (4.1)–(4.2), *and there is a constant* C *such that*

$$\|\check{p}\|_0 + \|\check{\mathbf{u}}\|_{H(div)} \leq C\{\|F\| + \|G\|\},$$

where C *is a nonlinear function of* $\|\check{a}\|$, $\|\check{c}\|$, *the reciprocal of the coercivity bound for* $\check{a}$, *and* $1/\beta$ *that is bounded on bounded subsets.*

The key result is to prove the celebrated inf-sup condition (4.3). This condition is known to hold over $W \times \mathbf{V}$, and the following corollary is well known and uses the fact that for $\mathbf{v} \in \mathbf{V}$,

$$(g_R, \mathbf{v} \cdot \nu)_{\Gamma_R} = (g_R, \mathbf{v} \cdot \nu)_{\partial\Omega} \leq C\|g_R\|_{1/2,\partial\Omega}\|\mathbf{v}\|_{H(\mathrm{div})},$$

where g_R on $\partial\Omega$ is any fixed bounded extension.

COROLLARY 4.3. *There exists a unique solution to* (2.6)–(2.7), *and there is some constant* C *depending on* a, d, *and the inf-sup bound such that*

$$\|p\|_0 + \|\mathbf{u}\|_{H(div)} \leq C\{\|b\|_0 + \|c\|_0 + \|\mathbf{v}_{g_N}\|_{H(div)} + \|g_R\|_{1/2,\partial\Omega}\}.$$

LEMMA 4.4. *The inf-sup condition holds over both* $W_c \times \mathbf{V}_c$ *and* $\delta W \times \delta \mathbf{V}$, *with constants independent of the coarse mesh* $\mathcal{T}_H$ *and the specific decomposition of* $W \times \mathbf{V}$ *selected.*

Proof. For $W_c \times \mathbf{V}_c$, we have the following argument. Given $w_c \in W_c$, solve for $\psi \in H^1(\Omega)$ satisfying (3.6)–(3.8), with $\int_\Omega \psi\, dx = 0$ if $\Gamma_N = \partial\Omega$. Set $\mathbf{v} = \nabla\psi$. Then

$$\|\mathbf{v}\|_0^2 = \|\nabla\psi\|_0^2 = (w_c, \psi) \leq \|w_c\|_0\|\psi\|_0 \leq C\|w_c\|_0\|\nabla\psi\|_0,$$

by Poincaré's inequality [17], so $\|\mathbf{v}\|_{H(\mathrm{div})} \le C\|w_c\|_0$. Let $\mathbf{v} = \hat{\mathbf{v}}_c + \delta\mathbf{v}$, where $\hat{\mathbf{v}}_c \in \mathbf{V}_c$ and $\delta\mathbf{v} \in \delta\mathbf{V}$. We note that $w_c = \nabla\cdot\mathbf{v} = \nabla\cdot\hat{\mathbf{v}}_c \in W_c$. Moreover,

$$\|\hat{\mathbf{v}}_c\|_{H(\mathrm{div})} = \|\tilde{\mathcal{P}}_{\mathbf{V}_c}\mathbf{v}\|_{H(\mathrm{div})} \le C\|\mathbf{v}\|_{H(\mathrm{div})} \le C\|w_c\|_0.$$

Thus

$$\begin{aligned}\inf_{w_c\in W_c}\sup_{\mathbf{v}_c\in\mathbf{V}_c}\frac{(\nabla\cdot\mathbf{v}_c, w_c)}{\|\mathbf{v}_c\|_{H(\mathrm{div})}\,\|w_c\|_0} &\ge \inf_{w_c\in W_c}\frac{(\nabla\cdot\hat{\mathbf{v}}_c, w_c)}{\|\hat{\mathbf{v}}_c\|_{H(\mathrm{div})}\,\|w_c\|_0}\\ &= \inf_{w_c\in W_c}\frac{\|w_c\|_0}{\|\hat{\mathbf{v}}_c\|_{H(\mathrm{div})}} \ge \frac{1}{C} > 0.\end{aligned}$$

The proof for $\delta W\times\delta\mathbf{V}$ is entirely similar and omitted. ∎

Proof of Lemma 4.1. We can rewrite the subgrid δ-scale equations (3.14)–(3.15) in the form of the abstract problem (4.1)–(4.2) by taking $\check{\mathbf{V}} = \delta\mathbf{V}$ and $\check{W} = \delta W$ and by defining

$$\begin{aligned}&\check{c}(\delta w_1, \delta w_2) = (a\delta w_1, \delta w_2),\\ &\check{a}(\delta\mathbf{v}_1, \delta\mathbf{v}_2) = (d^{-1}\delta\mathbf{v}_1, \delta\mathbf{v}_2),\\ &F(\delta w) = (b^* - ap_c - \nabla\cdot\mathbf{u}_c, \delta w),\\ &G(\delta\mathbf{v}) = (c^* - d^{-1}\mathbf{u}_c, \delta\mathbf{v}) + (p_c, \nabla\cdot\delta\mathbf{v}).\end{aligned}$$

Easily, the bilinear forms $\check{a}$ and $\check{c}$ are continuous, symmetric, and nonnegative on $\delta\mathbf{V}$, and $\check{a}$ is coercive on $\delta\mathbf{V}\cap\ker(\nabla\cdot)$, with constants depending on the coefficients a, c, and d. Moreover, F and G are continuous linear functionals. Lemma 4.4 gives us the inf-sup condition (4.3), so the hypotheses required by Theorem 4.2 are satisfied by the system, and so the conclusions follow. ∎

4.2. The δ-solution operator. Lemma 4.1 allows us to define the solution operator of the subgrid δ-scale equations (3.14)–(3.15) in terms of the coarse-scale solution. It is in fact an affine operator with constant and linear parts.

Constant part of the δ-solution operator. Find $\delta\bar{p}\in\delta W$ and $\delta\bar{\mathbf{u}}\in\delta\mathbf{V}$ such that

$$(a\delta\bar{p}, \delta w) + (\nabla\cdot\delta\bar{\mathbf{u}}, \delta w) = (b^*, \delta w) \quad \forall\ \delta w\in\delta W, \tag{4.4}$$

$$(d^{-1}\delta\bar{\mathbf{u}}, \delta\mathbf{v}) - (\delta\bar{p}, \nabla\cdot\delta\mathbf{v}) = (c^*, \delta\mathbf{v}) \quad \forall\ \delta\mathbf{v}\in\delta\mathbf{V}. \tag{4.5}$$

W_c-linear part of the δ-solution operator. For $w_c\in W_c$, find $\delta\tilde{p}\in\delta W$ and $\delta\tilde{\mathbf{u}}\in\delta\mathbf{V}$ such that

$$(a(w_c + \delta\tilde{p}), \delta w) + (\nabla\cdot\delta\tilde{\mathbf{u}}, \delta w) = 0 \quad \forall\ \delta w\in\delta W, \tag{4.6}$$

$$(d^{-1}\delta\tilde{\mathbf{u}}, \delta\mathbf{v}) - (w_c + \delta\tilde{p}, \nabla\cdot\delta\mathbf{v}) = 0 \quad \forall\ \delta\mathbf{v}\in\delta\mathbf{V}. \tag{4.7}$$

$\mathbf{V}_c$-linear part of the δ-solution operator. For $\mathbf{v}_c\in\mathbf{V}_c$, find $\delta\hat{p}\in\delta W$ and $\delta\hat{\mathbf{u}}\in\delta\mathbf{V}$ such that

$$(a\delta\hat{p}, \delta w) + (\nabla\cdot(\mathbf{v}_c + \delta\hat{\mathbf{u}}), \delta w) = 0 \quad \forall\ \delta w\in\delta W, \tag{4.8}$$

$$(d^{-1}(\mathbf{v}_c + \delta\hat{\mathbf{u}}), \delta\mathbf{v}) - (\delta\hat{p}, \nabla\cdot\delta\mathbf{v}) = 0 \quad \forall\ \delta\mathbf{v}\in\delta\mathbf{V}. \tag{4.9}$$

The theory of saddle point problems allows us to conclude the solvability and boundedness of each system, so we have the following result.

THEOREM 4.5. *There exist bounded linear operators*

$$\delta\tilde{p} : W_c \to \delta W \quad \text{and} \quad \delta\tilde{\mathbf{u}} : W_c \to \delta\mathbf{V},$$
$$\delta\hat{p} : \mathbf{V}_c \to \delta W \quad \text{and} \quad \delta\hat{\mathbf{u}} : \mathbf{V}_c \to \delta\mathbf{V},$$

bounded independent of the coarse mesh $\mathcal{T}_H$ and the specific decomposition of $W \times \mathbf{V}$ selected, defined by (4.6)–(4.9)*, and functions $\delta\bar{p} \in \delta W$ and $\delta\bar{\mathbf{u}} \in \delta\mathbf{V}$ defined by* (4.4)–(4.5) *such that*

$$\delta p = \delta\tilde{p}(p_c) + \delta\hat{p}(\mathbf{u}_c) + \delta\bar{p},$$
$$\delta\mathbf{u} = \delta\tilde{\mathbf{u}}(p_c) + \delta\hat{\mathbf{u}}(\mathbf{u}_c) + \delta\bar{\mathbf{u}}.$$

Moreover, there is some constant C such that

$$\|\delta\bar{p}\|_0 + \|\delta\bar{\mathbf{u}}\|_{H(div)} \le C\{\|b\|_0 + \|c\|_0 + \|\mathbf{v}_{g_N}\|_{H(div)}\}.$$

Because $\delta\mathbf{V} \cdot \nu = 0$ on each ∂E_c for $E_c \in \mathcal{T}_H$, $\delta\tilde{p}$, $\delta\tilde{\mathbf{u}}$, $\delta\hat{p}$, and $\delta\hat{\mathbf{u}}$ are locally defined operators. That is, the restriction to E_c of the result is given by evaluating the restricted operators, which are defined by restricting the integrals to E_c in (4.6)–(4.9). Symbolically, we might write

$$\delta\tilde{p}(p_c)|_{E_c} = \delta\tilde{p}|_{E_c}(p_c|_{E_c}) \quad \text{and} \quad \delta\tilde{\mathbf{u}}(p_c)|_{E_c} = \delta\tilde{\mathbf{u}}|_{E_c}(p_c|_{E_c}),$$
$$\delta\hat{p}(\mathbf{u}_c)|_{E_c} = \delta\hat{p}|_{E_c}(\mathbf{u}_c|_{E_c}) \quad \text{and} \quad \delta\hat{\mathbf{u}}(\mathbf{u}_c)|_{E_c} = \delta\hat{\mathbf{u}}|_{E_c}(\mathbf{u}_c|_{E_c}).$$

These operators are well defined, linear, and bounded uniformly with respect to $\mathcal{T}_H$ and the decomposition of $W \times \mathbf{V}$ selected.

In upscaling theory, results like Theorem 4.5 allow one to *close* the equations. That is, the fine scale is represented as an operator of the coarse scale. However, usually such a result is either *assumed* or additional assumptions are added to restrict the nature of the problem (such as assuming some kind of periodicity or ergodicity). Hence such results are often called *closure assumptions*. We have closed our system without the need of any additional assumptions.

4.3. The upscaled equation. If we substitute the δ-solution operator into the coarse-scale equations (3.12)–(3.13), we obtain the following problem.

Asymmetric upscaled equations. Find $p_c \in W_c$ and $\mathbf{u}_c \in \mathbf{V}_c$ such that

$$\begin{aligned}
&(a(p_c + \delta\tilde{p}(p_c) + \delta\hat{p}(\mathbf{u}_c)), w_c) \\
&\quad + (\nabla \cdot (\mathbf{u}_c + \delta\tilde{\mathbf{u}}(p_c) + \delta\hat{\mathbf{u}}(\mathbf{u}_c)), w_c) \\
&\quad = (b^* - a\delta\bar{p} - \nabla \cdot \delta\bar{\mathbf{u}}, w_c) \qquad \forall\ w_c \in W_c, \qquad (4.10)\\
&(d^{-1}(\mathbf{u}_c + \delta\tilde{\mathbf{u}}(p_c) + \delta\hat{\mathbf{u}}(\mathbf{u}_c)), \mathbf{v}_c) \\
&\quad + (\alpha\mathbf{u}_c \cdot \nu, \mathbf{v}_c \cdot \nu)_{\Gamma_R} - (p_c + \delta\tilde{p}(p_c) + \delta\hat{p}(\mathbf{u}_c), \nabla \cdot \mathbf{v}_c) \\
&\quad = (c^* - d^{-1}\delta\bar{\mathbf{u}}, \mathbf{v}_c) + (\delta\bar{p}, \nabla \cdot \mathbf{v}_c) - (g_R, \mathbf{v}_c \cdot \nu)_{\Gamma_R} \qquad \forall\ \mathbf{v}_c \in \mathbf{V}_c. \qquad (4.11)
\end{aligned}$$

This system is posed entirely with respect to coarse-scale functions, so we say that it has been *upscaled* from the fine scale. However, this system is not symmetric, even though the original fine-scale system is symmetric. We can remedy this by noting several equivalences. First, note that from (4.6) and then (4.7),

$$\begin{aligned}
(a(p_c + \delta\tilde{p}(p_c)), \delta\tilde{p}(w_c)) &= -(\nabla \cdot \delta\tilde{\mathbf{u}}(p_c), \delta\tilde{p}(w_c)) \\
&= -(d^{-1}\delta\tilde{\mathbf{u}}(w_c), \delta\tilde{\mathbf{u}}(p_c)) + (w_c, \nabla \cdot \delta\tilde{\mathbf{u}}(p_c)),
\end{aligned}$$

and similarly from (4.9) and then (4.8),

$$\begin{aligned}(d^{-1}(\mathbf{u}_c+\delta\hat{\mathbf{u}}(\mathbf{u}_c)),\delta\hat{\mathbf{u}}(\mathbf{v}_c)) &= (\delta\hat{p}(\mathbf{u}_c),\nabla\cdot\delta\hat{\mathbf{u}}(\mathbf{v}_c))\\ &= -(a\delta\hat{p}(\mathbf{v}_c),\delta\hat{p}(\mathbf{u}_c))-(\nabla\cdot\mathbf{v}_c,\delta\hat{p}(\mathbf{u}_c)).\end{aligned}$$

We also apply (4.9), (4.6), (4.7), and then (4.8) to obtain

$$\begin{aligned}(d^{-1}\delta\tilde{\mathbf{u}}(p_c),\mathbf{v}_c) &= (\delta\hat{p}(\mathbf{v}_c),\nabla\cdot\delta\tilde{\mathbf{u}}(p_c))-(d^{-1}\delta\hat{\mathbf{u}}(\mathbf{v}_c),\delta\tilde{\mathbf{u}}(p_c))\\ &= -(a(p_c+\delta\tilde{p}(p_c)),\delta\hat{p}(\mathbf{v}_c))-(d^{-1}\delta\hat{\mathbf{u}}(\mathbf{v}_c),\delta\tilde{\mathbf{u}}(p_c))\\ &= -(a(p_c+\delta\tilde{p}(p_c)),\delta\hat{p}(\mathbf{v}_c))-(p_c+\delta\tilde{p}(p_c),\nabla\cdot\delta\hat{\mathbf{u}}(\mathbf{v}_c))\\ &= -(ap_c,\delta\hat{p}(\mathbf{v}_c))-(p_c,\nabla\cdot\delta\hat{\mathbf{u}}(\mathbf{v}_c))+(\nabla\cdot\mathbf{v}_c,\delta\tilde{p}(p_c)).\end{aligned}$$

Combining, we obtain a symmetric form for our system.

Symmetric upscaled equations. Find $p_c\in W_c$ and $\mathbf{u}_c\in\mathbf{V}_c$ such that

$$\begin{aligned}&(a(p_c+\delta\tilde{p}(p_c)),w_c+\delta\tilde{p}(w_c))+(d^{-1}\delta\tilde{\mathbf{u}}(p_c),\delta\tilde{\mathbf{u}}(w_c))\\ &\quad+(\nabla\cdot(\mathbf{u}_c+\delta\hat{\mathbf{u}}(\mathbf{u}_c)),w_c)+(a\delta\hat{p}(\mathbf{u}_c),w_c)\\ &\quad=(b^*-a\delta\bar{p}-\nabla\cdot\delta\bar{\mathbf{u}},w_c)\qquad\forall\ w_c\in W_c,\end{aligned}\tag{4.12}$$

$$\begin{aligned}&(d^{-1}(\mathbf{u}_c+\delta\hat{\mathbf{u}}(\mathbf{u}_c)),\mathbf{v}_c+\delta\hat{\mathbf{u}}(\mathbf{v}_c))\\ &\quad+(a\delta\hat{p}(\mathbf{u}_c),\delta\hat{p}(\mathbf{v}_c))+(\alpha\mathbf{u}_c\cdot\nu,\mathbf{v}_c\cdot\nu)_{\Gamma_R}\\ &\quad-(p_c,\nabla\cdot(\mathbf{v}_c+\delta\hat{\mathbf{u}}(\mathbf{v}_c)))-(ap_c,\delta\hat{p}(\mathbf{v}_c))\\ &\quad=(c^*-d^{-1}\delta\bar{\mathbf{u}},\mathbf{v}_c)+(\delta\bar{p},\nabla\cdot\mathbf{v}_c)-(g_R,\mathbf{v}_c\cdot\nu)_{\Gamma_R}\qquad\forall\ \mathbf{v}_c\in\mathbf{V}_c.\end{aligned}\tag{4.13}$$

The final solution is given then by

$$p=p_c+\delta\tilde{p}(p_c)+\delta\hat{p}(\mathbf{u}_c)+\delta\bar{p},\tag{4.14}$$

$$\mathbf{u}=\mathbf{u}_c+\delta\tilde{\mathbf{u}}(p_c)+\delta\hat{\mathbf{u}}(\mathbf{u}_c)+\delta\bar{\mathbf{u}}+\mathbf{v}_{g_N}.\tag{4.15}$$

It remains to show that indeed (4.10)–(4.11) or, equivalently, (4.12)–(4.13) has a unique solution from which to construct the solution p and $\mathbf{u}$.

THEOREM 4.6. *There exists a unique solution to* (4.10)–(4.11) *or, equivalently, to* (4.12)–(4.13). *Moreover, there is some constant C, independent of the coarse mesh $\mathcal{T}_H$ and the specific decomposition of $W\times\mathbf{V}$ selected, such that*

$$\|p_c\|_0+\|\mathbf{u}_c\|_{H(div)}\le C\{\|b\|_0+\|c\|_0+\|\mathbf{v}_{g_N}\|_{H(div)}+\|g_R\|_{1/2,\partial\Omega}\}.$$

Proof. Rather than trying to show the inf-sup condition for the system (4.12)–(4.13) with its bilinear form $(\nabla\cdot(\mathbf{v}_c+\delta\hat{\mathbf{u}}(\mathbf{v}_c)),w_c)+(a\delta\hat{p}(\mathbf{v}_c),w_c)$, we use a more direct route. From Corollary 4.3, we have $(p,\mathbf{u})\in W\times(\mathbf{V}+\mathbf{v}_{g_N})$ solving the original system. We uniquely decompose $p=p_c+\delta p\in W_c\oplus\delta W$ and $\mathbf{u}-\mathbf{v}_{g_N}=\mathbf{u}_c+\delta\mathbf{u}\in\mathbf{V}_c\oplus\delta\mathbf{V}$. By construction, $(p_c,\mathbf{u}_c)\in W_c\times\mathbf{V}_c$ is a solution to (4.12)–(4.13).

To demonstrate the uniqueness of the solution, consider the difference of two solutions, which is equivalent to setting all constant terms to zero and showing that there is only the trivial solution. Take the test functions $w_c=p_c$ and $\mathbf{v}_c=\mathbf{u}_c$ and sum the equations to conclude that $\mathbf{u}_c+\delta\hat{\mathbf{u}}(\mathbf{u}_c)=\delta\tilde{\mathbf{u}}(p_c)=0$ and, by the uniqueness of the decomposition $\mathbf{V}_c\oplus\delta\mathbf{V}$, that $\mathbf{u}_c=0$ and thus also $\delta\hat{p}(\mathbf{u}_c)=0$. Since $\nabla\cdot\mathbf{V}=W$, equations (4.7) and (4.11) imply that $p_c+\delta\tilde{p}(p_c)=0$, and thus $p_c=0$ and uniqueness is established.

We use (3.3) and (3.10) to bound

$$\|p_c\|_0\le C\|p\|_0\quad\text{and}\quad\|\mathbf{u}_c\|_{H(\mathrm{div})}\le C\|\mathbf{u}\|_{H(\mathrm{div})}.$$

Finally, Corollary 4.3 bounds these terms as required. ∎

5. Numerical approximation. In the previous section, we demonstrated that the δ-problems (4.4)–(4.9) and the upscaled problem (4.12)–(4.13) are well-posed, uniformly with respect to $\mathcal{T}_H$. In this section, we construct an efficient computational algorithm, exploiting the structure exposed in the previous two sections. Namely, we exploit that the δ-problems are local and thus easily solved computationally and that the global upscaled problem on $\mathcal{T}_H$ is relatively small compared to the full fine-scale problem (2.6)–(2.7) itself.

We present a class of discretizations based on standard mixed spaces. Our class of discretizations includes the particularly pertinent low order discretization described in [1, 2] and later in section 7.

We now consider $\mathcal{T}_H$ as a coarse mesh. For approximation purposes, we assume that it is chosen of conforming simplexes, rectangular parallelepipeds, or prisms such that, for simplicity, $\bar{\Gamma}_N$ is the union of coarse edges or faces. Let

$$H = \max_{E_c \in \mathcal{T}_H} \operatorname{diam}(E_c).$$

On each $E_c \in \mathcal{T}_H$, let $\mathcal{T}_h(E_c)$ be a fine mesh sufficient to resolve the coefficients of the problem, and define

$$h = \max_{E_c \in \mathcal{T}_H} \max_{\delta E \in \mathcal{T}_h(E_c)} \operatorname{diam}(\delta E).$$

Then $\mathcal{T}_{H,h} = \cup_{E_c \in \mathcal{T}_H} \mathcal{T}_h(E_c)$ is the full fine mesh. The meshes need not match across boundaries of coarse elements.

5.1. Two-scale conforming approximation spaces. From among any of the usual mixed finite element spaces for second order elliptic equations, such as those of [28, 24, 12, 10, 11, 14, 13], we select the coarse space $W_H^* \times \mathbf{V}_H^* \subset W \times \mathbf{V}$ on the mesh $\mathcal{T}_H$, with $\mathbf{V}_H^*$ satisfying the homogeneous Neumann boundary condition on Γ_N. In all the usual spaces,

$$\nabla \cdot \mathbf{V}_H^* = W_H^*$$

and piecewise discontinuous constants $W_c^1 \subset W_H^*$.

On each coarse element $E_c \in \mathcal{T}_H$, we similarly select from among any of the usual mixed finite element spaces the δ-space $\delta W_h(E_c) \times \delta \mathbf{V}_h(E_c) \subset (W \times \mathbf{V})|_{E_c}$ on the mesh $\mathcal{T}_h(E_c)$, with $\delta W_h(E_c) \perp 1$ and $\delta \mathbf{V}_h(E_c)$ satisfying the homogeneous Neumann boundary condition on ∂E_c. Merging these spaces results in $\delta W_h \times \delta \mathbf{V}_h$ over the entire domain Ω. Then $\delta W_h \perp W_c^1$ and

$$\nabla \cdot \delta \mathbf{V}_h = \delta W_h.$$

For simplicity, we take the same mixed space for each coarse element, although this assumption could be relaxed.

The overall two-scale mixed spaces are then defined to be

$$W_{H,h} = W_H^* + \delta W_h \quad \text{and} \quad \mathbf{V}_{H,h} = \mathbf{V}_H^* + \delta \mathbf{V}_h.$$

However, it is possible for general combinations of mixed spaces that the coarse and δ-spaces are not linearly independent. The following construction suffices to rectify the problem. First, complete a basis for $\delta W_h \cap W_H^*$ to a basis for W_H^* and then define W_H as the span of the extra vectors. Similarly, we complete a basis for $\delta \mathbf{V}_h \cap \mathbf{V}_H^*$ to a basis for $\mathbf{V}_H^*$ and use the extra vectors to define $\mathbf{V}_H$.

To summarize our construction, our two-scale finite element spaces are and satisfy

$$W_{H,h} = W_H^* + \delta W_h = W_H \oplus \delta W_h \subset W,$$
$$\mathbf{V}_{H,h} = \mathbf{V}_H^* + \delta \mathbf{V}_h = \mathbf{V}_H \oplus \delta \mathbf{V}_h \subset \mathbf{V},$$

where

$$\nabla \cdot \delta \mathbf{V}_h = \delta W_h \quad \text{and} \quad \nabla \cdot \mathbf{V}_{H,h} = W_{H,h}.$$

Our spaces are conforming in the sense that both W_H and δW_h are subspaces of W, and $\mathbf{V}_H$ and $\delta \mathbf{V}_h$ are subspaces of $\mathbf{V}$ and $\delta \mathbf{V}^1$, respectively, and thus have the required $H(\mathrm{div})$ smoothness and satisfy the requisite boundary conditions.

However, it is not necessarily the case that $W_H \subset W_c$ and $\delta W_h \subset \delta W$ nor that $\mathbf{V}_H \subset \mathbf{V}_c$ and $\delta \mathbf{V}_h \subset \delta \mathbf{V}$. In section 3.1, we made a few arbitrary choices. We could, for example, have chosen $W_H = W_c$ and then defined δW in such a way that both $\delta W_h \subset \delta W$ and (3.3) hold, perhaps after assuming the restriction on the grid mentioned in section 6. We might similarly be able to decompose $\mathbf{V}^s$ in such a way that $\mathbf{V}_H \subset \mathbf{V}_c$ and $\delta \mathbf{V}_h \subset \delta \mathbf{V}$. Then the mixed spaces would be fully conforming in the two-scale sense. However, there appears to be no advantage to such a construction, so we will not attempt it here.

5.2. The discrete equations in computable form. The key to efficient implementation is to determine the δ-operators' actions only on the finite element basis for $\mathbf{V}_H$. We call such solutions *numerical Green's functions*, since they give the response of the system to a "unit" disturbance, which on the numerical level is a coarse-scale basis function.

Let $\{w_{H,i}\}_i$ and $\{\mathbf{v}_{H,j}\}_j$ be finite element bases for W_H and $\mathbf{V}_H$, respectively. One property of a finite element basis is that the support of any basis function is relatively small. Expand

$$p_H = \sum_i p_i w_{H,i} \quad \text{and} \quad \mathbf{u}_H = \sum_j u_j \mathbf{v}_{H,j}. \tag{5.1}$$

Then to compute, for example,

$$\delta \hat{\mathbf{u}}(\mathbf{u}_H) = \sum_j u_j \delta \hat{\mathbf{u}}(\mathbf{v}_{H,j})$$

requires only the numerical Green's functions $\delta \hat{\mathbf{u}}(\mathbf{v}_{H,j})$ for each j.

The numerical scheme has three main steps. The first step is to compute the solutions to the following problems.

Constant part of the approximate δ-solution operator. Find $\delta \bar{p}_h \in \delta W_h$ and $\delta \bar{\mathbf{u}}_h \in \delta \mathbf{V}_h$ such that

$$(a \delta \bar{p}_h, \delta w_h) + (\nabla \cdot \delta \bar{\mathbf{u}}_h, \delta w_h) = (b^*, \delta w_h) \qquad \forall\, \delta w_h \in \delta W_h, \tag{5.2}$$

$$(d^{-1} \delta \bar{\mathbf{u}}_h, \delta \mathbf{v}_h) - (\delta \bar{p}_h, \nabla \cdot \delta \mathbf{v}_h) = (c^*, \delta \mathbf{v}_h) \quad \forall\, \delta \mathbf{v}_h \in \delta \mathbf{V}_h. \tag{5.3}$$

W_H-linear part of the approximate δ-solution operator. For $w_{H,i}$ in a basis for W_H, find $\delta \tilde{p}_{h,i} \in \delta W_h$ and $\delta \tilde{\mathbf{u}}_{h,i} \in \delta \mathbf{V}_h$ such that

$$(a(w_{H,i} + \delta \tilde{p}_{h,i}), \delta w_h) + (\nabla \cdot \delta \tilde{\mathbf{u}}_{h,i}, \delta w_h) = 0 \quad \forall\, \delta w_h \in \delta W_h, \tag{5.4}$$

$$(d^{-1} \delta \tilde{\mathbf{u}}_{h,i}, \delta \mathbf{v}_h) - (w_{H,i} + \delta \tilde{p}_{h,i}, \nabla \cdot \delta \mathbf{v}_h) = 0 \quad \forall\, \delta \mathbf{v}_h \in \delta \mathbf{V}_h. \tag{5.5}$$

$\mathbf{V}_H$-linear part of the approximate δ-solution operator. For $\mathbf{v}_{H,j}$ in a basis for $\mathbf{V}_H$, find $\delta\hat{p}_{h,j} \in \delta W_h$ and $\delta\hat{\mathbf{u}}_{h,j} \in \delta\mathbf{V}_h$ such that

$$(a\delta\hat{p}_{h,j}, \delta w_h) + (\nabla\cdot(\mathbf{v}_{H,j} + \delta\hat{\mathbf{u}}_{h,j}), \delta w_h) = 0 \qquad \forall\, \delta w_h \in \delta W_h, \tag{5.6}$$

$$(d^{-1}(\mathbf{v}_{H,j} + \delta\hat{\mathbf{u}}_{h,j}), \delta\mathbf{v}_h) - (\delta\hat{p}_{h,j}, \nabla\cdot\delta\mathbf{v}_h) = 0 \quad \forall\, \delta\mathbf{v}_h \in \delta\mathbf{V}_h. \tag{5.7}$$

These problems are quick and efficient to solve, since they are relatively quite small due to their local nature. That is, we actually solve them on each coarse element independently. For example, we know that for the standard mixed spaces, $\mathbf{v}_{H,j}$ is supported on at most two coarse elements, E_c^1 and E_c^2. Thus, to evaluate $\delta\hat{\mathbf{u}}(\mathbf{v}_{H,j})$, we solve (5.6)–(5.7) twice, with all spaces and integrals restricted to E_c^k for $k = 1, 2$. Then in fact $\delta\hat{\mathbf{u}}(\mathbf{v}_{H,j})$ is the combination of the two solutions $\delta\hat{\mathbf{u}}|_{E_c^k}(\mathbf{v}_{H,j}|_{E_c^k})$ on E_c^k, $k = 1, 2$. On each coarse element, each linear system in (5.2)–(5.7) has the same matrix, and only the so-called right-hand side vector varies. Thus it is reasonable to use a direct solver for these problems. Moreover, they parallelize trivially. Since these are square linear systems, existence and uniqueness of a solution follow from uniqueness, which follow in the usual way from the fact that $\nabla\cdot\delta\mathbf{V}_h = \delta W_h$.

Then we have the implicit expressions

$$\delta p_h = \delta\tilde{p}_h(p_H) + \delta\hat{p}_h(\mathbf{u}_H) + \delta\bar{p}_h = \sum_i p_i\delta\tilde{p}_{h,i} + \sum_j u_j\delta\hat{p}_{h,j} + \delta\bar{p}_h, \tag{5.8}$$

$$\delta\mathbf{u}_h = \delta\tilde{\mathbf{u}}_h(p_H) + \delta\hat{\mathbf{u}}_h(\mathbf{u}_H) + \delta\bar{\mathbf{u}}_h = \sum_i p_i\delta\tilde{\mathbf{u}}_{h,i} + \sum_j u_j\delta\hat{\mathbf{u}}_{h,j} + \delta\bar{\mathbf{u}}_h, \tag{5.9}$$

since at this stage of the computation p_i and u_j are not known.

The second main step is to compute the solution to the upscaled equation. We approximate (4.12)–(4.13) in the symmetric case by restricting to the finite element basis: Find $p_H \in W_H$ and $\mathbf{u}_H \in \mathbf{V}_H$ such that

$$\begin{aligned}
&(a(p_H + \delta\tilde{p}(p_H)), w_H + \delta\tilde{p}(w_H)) + (d^{-1}\delta\tilde{\mathbf{u}}(p_H), \delta\tilde{\mathbf{u}}(w_H)) \\
&\quad + (\nabla\cdot(\mathbf{u}_H + \delta\hat{\mathbf{u}}(\mathbf{u}_H)), w_H) + (a\delta\hat{p}(\mathbf{u}_H), w_H) \\
&\quad = (b^* - a\delta\bar{p}_h - \nabla\cdot\delta\bar{\mathbf{u}}_h, w_H) \qquad \forall\, w_H \in W_H,
\end{aligned} \tag{5.10}$$

$$\begin{aligned}
&(d^{-1}(\mathbf{u}_H + \delta\hat{\mathbf{u}}(\mathbf{u}_H)), \mathbf{v}_H + \delta\hat{\mathbf{u}}(\mathbf{v}_H)) \\
&\quad + (a\delta\hat{p}(\mathbf{u}_H), \delta\hat{p}(\mathbf{v}_H)) + (\alpha\mathbf{u}_H\cdot\nu, \mathbf{v}_H\cdot\nu)_{\Gamma_R} \\
&\quad - (p_H, \nabla\cdot(\mathbf{v}_H + \delta\hat{\mathbf{u}}(\mathbf{v}_H))) - (ap_H, \delta\hat{p}(\mathbf{v}_H)) \\
&\quad = (c^* - d^{-1}\delta\bar{\mathbf{u}}_h, \mathbf{v}_H) - (\delta\bar{p}_h, \nabla\cdot\mathbf{v}_H) - (g_R, \mathbf{v}_H\cdot\nu)_{\Gamma_R} \quad \forall\, \mathbf{v}_H \in \mathbf{V}_H.
\end{aligned} \tag{5.11}$$

By following the computations in section 4.3, we easily see that a similar finite element approximation of the asymmetric formulation (4.10)–(4.11) is equivalent to (5.10)–(5.11). Either problem is the same size as a full finite element approximation of (2.6)–(2.7) over the *coarse* space $W_H \times \mathbf{V}_H$.

The final main step is to construct the solution using (5.8)–(5.9):

$$p_h = p_H + \delta p_h \in W_{H,h}, \tag{5.12}$$

$$\mathbf{u}_h = \mathbf{u}_H + \delta\mathbf{u}_h + \mathbf{v}_{g_N} \in \mathbf{V}_{H,h} + \mathbf{v}_{g_N}. \tag{5.13}$$

5.3. An equivalent form for the discrete equations. It should be noted that our procedure is an efficient implementation of the algebraically equivalent mixed

finite element method corresponding to (2.6)–(2.7), which is to find $\mathbf{u}_h \in \mathbf{V}_{H,h} + \mathbf{v}_{g_N}$ and $p_h \in W_{H,h}$ such that

$$(ap_h, w_h) + (\nabla \cdot \mathbf{u}_h, w_h) = (b, w_h) \qquad \forall\ w_h \in W_{H,h}, \tag{5.14}$$

$$\begin{aligned}(d^{-1}\mathbf{u}_h, \mathbf{v}_h) + (\alpha \mathbf{u}_h \cdot \nu, \mathbf{v}_h \cdot \nu)_{\Gamma_R} - (p_h, \nabla \cdot \mathbf{v}_h) \\ = (c, \mathbf{v}_h) - (g_R, \mathbf{v}_h \cdot \nu)_{\Gamma_R} \qquad \forall\ \mathbf{v}_h \in \mathbf{V}_{H,h}.\end{aligned} \tag{5.15}$$

Existence and uniqueness of a solution follow from uniqueness, which follows from the linear independence of $W_H \times \mathbf{V}_H$ and $\delta W_h \times \delta \mathbf{V}_h$ and the fact that $\nabla \cdot \mathbf{V}_{H,h} = W_{H,h}$.

From (5.14)–(5.15), we can conclude existence and uniqueness of (5.10)–(5.11), since $p_h = p_H + \delta p_h$ and $\mathbf{u}_h = \mathbf{u}_H + \delta \mathbf{u}_h$ give p_H and $\mathbf{u}_H$, which satisfy the system.

6. Analysis of the approximation error. We begin this section with some notation. For M, a subspace of L^2, we denote by $\mathcal{P}_M : L^2 \to M$ the orthogonal L^2-projection, based on the decomposition $L^2 = M \oplus M^\perp$. We contrast this with $\tilde{\mathcal{P}}_M : L^2 \to M$ from Proposition 3.1, which was based on a possibly nonorthogonal decomposition $L^2 = M \oplus N$.

At this point, we require some uniformity of the discrete decomposition. Let $\tilde{\mathcal{P}}_{W_H} : W_{H,h} \to W_H$ and $\tilde{\mathcal{P}}_{\delta W_h} : W_{H,h} \to \delta W_h$ be the projections associated with the decomposition $W_{H,h} = W_H \oplus \delta W_h$. We assume that there is a constant C such that

$$\|\tilde{\mathcal{P}}_{\delta W_h}\| \le C. \tag{6.1}$$

Note that then also

$$\|\tilde{\mathcal{P}}_{W_H}\| = \|I - \tilde{\mathcal{P}}_{\delta W_h}\| \le 1 + C. \tag{6.2}$$

It is not difficult to ensure (6.1). The simplest possibility is that $W_H \perp \delta W_h$ so that $\tilde{\mathcal{P}}_{\delta W_h} = \mathcal{P}_{\delta W_h}$ and we can take $C = 1$. This holds for certain choices of mixed spaces but not for others. Another possibility is to enforce uniformity on the two-scale mesh $\mathcal{T}_{H,h}$. Suppose that as $H \to 0$, we insist that H/h remains fixed. If we also assume that the coarse and fine element shapes remain fixed, then it is clear by a scaling argument that (6.1) will hold on each coarse element and thus globally. Moreover, we can even allow the element shapes to change as long as they do not change too badly, such as being the images of a reference element under a uniformly bounded family of affine maps with uniformly bounded inverses.

Let $K \ge 1$ and $L \ge 1$ denote the approximation orders of the coarse spaces $\mathbf{V}_H^*$ and W_H^*, respectively. That is, for some constant C and for any $\mathbf{v} \in \mathbf{V}$ and $w \in W$,

$$\inf_{\mathbf{v}_H \in \mathbf{V}_H^*} \|\mathbf{v} - \mathbf{v}_H\|_0 \le C \|\mathbf{v}\|_m\, H^m, \quad 0 \le m \le K, \tag{6.3}$$

$$\inf_{w_H \in W_H^*} \|w - w_H\|_0 \le C \|w\|_i\, H^i, \quad 0 \le i \le L. \tag{6.4}$$

For all the usual mixed spaces, $L = K$ or $L = K - 1$. It is also true that

$$\inf_{\mathbf{v}_H \in \mathbf{V}_H^*} \|(\mathbf{v} - \mathbf{v}_H) \cdot \nu\|_{0,\Gamma_R} \le C \|\mathbf{v} \cdot \nu\|_{m,\Gamma_R} H^m, \quad 0 \le m \le K. \tag{6.5}$$

Similarly, let $k \ge 1$ and $\ell \ge 1$ denote the approximation orders of the δ-spaces $\delta \mathbf{V}_h$ and δW_h, respectively.

LEMMA 6.1. *Given any $w \in W$,*

$$\|w - \mathcal{P}_{W_{H,h}} w\|_0 \le C \|w\|_{i+j} H^i h^j, \quad 0 \le i \le \max(0, L - j),\ 0 \le j \le \ell. \tag{6.6}$$

Proof. For each $E_c \in \mathcal{T}_H$, we note that in all the usual mixed spaces W_H^* restricted to E_c consists of polynomials of full degree at least $L-1$ that are discontinuous across ∂E_c. Thus from standard polynomial approximation results, we compute

$$\begin{aligned}\|w - \mathcal{P}_{W_{H,h}} w\|_{0,E_c} &= \inf_{w_H \in W_H^*} \inf_{\delta w_h \in \delta W_h} \|w - w_H - \delta w_h\|_{0,E_c} \\ &\le C \inf_{w_H \in W_H^*} \|w - w_H\|_{j,E_c} h^j \\ &\le C \|w\|_{i+j,E_c} H^i h^j,\end{aligned}$$

wherein $0 \le j \le \ell$ and then $0 \le i \le \max(0, L-j)$. ∎

Approximation in $\mathbf{V}_{H,h}$ is more delicate, as we need to preserve divergence properties.

6.1. A mixed method π-operator. All the usual mixed spaces $\check{W}_\eta \times \check{\mathbf{V}}_\eta$ have projection operators $\check{\pi} : \mathbf{V} \cap H^1(\Omega) \to \check{\mathbf{V}}_\eta$ such that

$$\begin{aligned}&\nabla \cdot \check{\pi}\mathbf{v} = \mathcal{P}_{\check{W}_\eta} \nabla \cdot \mathbf{v}, \\ &\|\mathbf{v} - \check{\pi}\mathbf{v}\| \le C \|\mathbf{v}\|_i \eta^i, \quad 1 \le i \le m,\end{aligned}$$

where C is a constant independent of the mesh spacing η and m is the approximation order of the space $\check{\mathbf{V}}_\eta$. Moreover, on $\partial\Omega$,

$$\check{\pi}\mathbf{v} \cdot \nu = \mathcal{P}_{\check{\mathbf{V}}_\eta \cdot \nu} \mathbf{v} \cdot \nu. \tag{6.7}$$

We have the associated operators

$$\begin{aligned}&\pi_H : \mathbf{V} \cap H^1(\Omega) \to \mathbf{V}_H^*, \\ &\delta\pi_{E_c,h} : \delta\mathbf{V}(E_c) \cap H^1(E_c) \to \delta\mathbf{V}_h(E_c) \quad \forall\, E_c \in \mathcal{T}_H,\end{aligned}$$

and also $\delta\pi_h$, defined by combining the $\delta\pi_{E_c,h}$. Then for any $\mathbf{v} \in \mathbf{V}$ and $\delta\mathbf{v} \in \delta\mathbf{V}^1$,

$$\begin{aligned}&\nabla \cdot \pi_H \mathbf{v} = \mathcal{P}_{W_H^*} \nabla \cdot \mathbf{v}, \\ &\nabla \cdot \delta\pi_h \delta\mathbf{v} = \mathcal{P}_{\delta W_h} \nabla \cdot \delta\mathbf{v}.\end{aligned}$$

Our goal now is to define a similar operator for the two-scale space $W_{H,h} \times \mathbf{V}_{H,h}$. Let $\mathbf{v} \in \mathbf{V} \cap H^1(\Omega)$. On each $E_c \in \mathcal{T}_H$, let

$$\delta w = \tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v}.$$

Define $\delta\mathbf{v}^p = \nabla\varphi$, where $\varphi \in H^1(E_c)$ satisfies (3.4)–(3.5) with the given δw. Then, because of (3.1), the Poincaré inequality constant is independent of H and h, so elliptic regularity [19] gives us the bound

$$\|\varphi\|_{2,E_c} \le C \|\tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v}\|_{0,E_c},$$

where a simple scaling argument shows that C depends on the shape of E_c but not on its size. Thus also

$$\|\delta\mathbf{v}^p\|_{1,E_c} \le C \|\tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v}\|_{0,E_c}, \tag{6.8}$$

and we conclude that we can apply $\delta\pi_h$ to $\delta\mathbf{v}^p$.

DEFINITION 6.2. *Let $\pi : \mathbf{V} \cap H^1(\Omega) \to \mathbf{V}_{H,h} = \mathbf{V}_H^* + \delta\mathbf{V}_h$ be defined by*

$$\pi\mathbf{v} = \pi_H(\mathbf{v} - \delta\mathbf{v}^p) + \delta\pi_h \delta\mathbf{v}^p,$$

where $\delta\mathbf{v}^p = \nabla\varphi$ and φ satisfies (3.4)–(3.5) *with $\delta w = \tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v}$.*

This operator, while well defined, is *not* a projection.

PROPOSITION 6.3. *It follows that*

$$\begin{aligned} &\nabla \cdot \pi \mathbf{v} = \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v} \quad &&\text{on } \Omega, \\ &\pi \mathbf{v} \cdot \nu = \pi_H \mathbf{v} \cdot \nu &&\text{on } \partial\Omega. \end{aligned}$$

Proof. Since $\mathcal{P}_{W_{H,h}} = (\tilde{\mathcal{P}}_{W_H} + \tilde{\mathcal{P}}_{\delta W_h})\mathcal{P}_{W_{H,h}}$ and $\mathcal{P}_{W_H^*}\mathcal{P}_{W_{H,h}} = \mathcal{P}_{W_H^*}$, we compute

$$\begin{aligned} \nabla \cdot \pi \mathbf{v} &= \nabla \cdot \pi_H(\mathbf{v} - \delta \mathbf{v}^p) + \nabla \cdot \delta \pi_h \delta \mathbf{v}^p \\ &= \mathcal{P}_{W_H^*} \nabla \cdot (\mathbf{v} - \delta \mathbf{v}^p) + \mathcal{P}_{\delta W_h} \nabla \cdot \delta \mathbf{v}^p \\ &= \mathcal{P}_{W_H^*} \nabla \cdot \mathbf{v} - \mathcal{P}_{W_H^*} \tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v} + \mathcal{P}_{\delta W_h} \tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v} \\ &= \mathcal{P}_{W_H^*} \nabla \cdot \mathbf{v} - \mathcal{P}_{W_H^*} (I - \tilde{\mathcal{P}}_{W_H}) \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v} + \tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v} \\ &= (\tilde{\mathcal{P}}_{W_H} + \tilde{\mathcal{P}}_{\delta W_h}) \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v} \\ &= \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v}. \end{aligned}$$

By (6.7), we see that

$$\pi \mathbf{v} \cdot \nu = \pi_H(\mathbf{v} - \delta \mathbf{v}^p) \cdot \nu + \delta \pi_h \delta \mathbf{v}^p \cdot \nu = \pi_H \mathbf{v} \cdot \nu,$$

since $\delta \mathbf{v}^p \cdot \nu = 0$ on $\partial\Omega$. ∎

LEMMA 6.4. *If* (3.1) *and* (6.1) *hold, then for* $\mathbf{v} \in \mathbf{V} \cap H^1(\Omega)$,

$$\begin{aligned} &\|\mathbf{v} - \pi \mathbf{v}\|_0 \le C \|\mathbf{v}\|_m H^m, \quad 1 \le m \le K, \\ &\|(\mathbf{v} - \pi \mathbf{v}) \cdot \nu\|_{0,\Gamma_R} \le C \|\mathbf{v} \cdot \nu\|_{m,\Gamma_R} H^m, \quad 0 \le m \le K. \end{aligned}$$

Proof. We construct $\delta \mathbf{v}^p \in \delta \mathbf{V}^p$ as in the definition of π. Then for $1 \le m \le K$,

$$\begin{aligned} \|\mathbf{v} - \pi \mathbf{v}\|_0 &= \|\mathbf{v} - \pi_H \mathbf{v} + \pi_H \delta \mathbf{v}^p - \delta \pi_h \delta \mathbf{v}^p\|_0 \\ &\le \|\mathbf{v} - \pi_H \mathbf{v}\|_0 + \|\pi_H \delta \mathbf{v}^p - \delta \mathbf{v}^p\|_0 + \|\delta \mathbf{v}^p - \delta \pi_h \delta \mathbf{v}^p\|_0 \\ &\le C \left\{ \|\mathbf{v}\|_m H^m + \sum_{E_c \in \mathcal{T}_H} (\|\delta \mathbf{v}^p\|_{1,E_c} H + \|\delta \mathbf{v}^p\|_{1,E_c} h) \right\} \\ &\le C \{ \|\mathbf{v}\|_m H^m + \|\tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v}\|_0 H \}, \end{aligned}$$

by (6.8). The Bramble–Hilbert lemma [15, 7] implies that for any $0 \le i \le L$,

$$\|\tilde{\mathcal{P}}_{\delta W_h} \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v}\|_0 = \|(I - \tilde{\mathcal{P}}_{W_H}) \mathcal{P}_{W_{H,h}} \nabla \cdot \mathbf{v}\|_0 \le C \|\nabla \cdot \mathbf{v}\|_i H^i,$$

since the operator $(I - \tilde{\mathcal{P}}_{W_H})\mathcal{P}_{W_{H,h}}$ is uniformly bounded by (6.1), and it preserves polynomials of the appropriate degree. With $i = m - 1$, i is in the range $0 \le i \le L$ ($L = K$ or $L = K - 1$), and we obtain the required first estimate.

The second estimate follows from Proposition 6.3 and its approximation properties, (6.7) and (6.5). ∎

6.2. Error analysis. The equation for the error is given by (2.6)–(2.7) with test functions in $W_{H,h} \times \mathbf{V}_{H,h}$ minus (5.14)–(5.15), which is

$$(a(p - p_h), w) + (\nabla \cdot (\mathbf{u} - \mathbf{u}_h), w) = 0 \qquad \forall\, w \in W_{H,h}, \tag{6.9}$$

$$(d^{-1}(\mathbf{u} - \mathbf{u}_h), \mathbf{v}) + (\alpha(\mathbf{u} - \mathbf{u}_h) \cdot \nu, \mathbf{v} \cdot \nu)_{\Gamma_R} = (p - p_h, \nabla \cdot \mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{V}_{H,h}. \tag{6.10}$$

THEOREM 6.5. *If* (3.1) *and* (6.1) *hold, and* $a \in W^{1,\infty}(\Omega)$*, then the two-scale approximation satisfies the error bounds*

$$\begin{aligned}
&\|\sqrt{a}(\mathcal{P}_{W_{H,h}}p - p_h)\|_0 + \|\mathbf{u} - \mathbf{u}_h\|_0 + \|\sqrt{\alpha}(\mathbf{u} - \mathbf{u}_h)\cdot\nu\|_{0,\Gamma_R} \\
&\quad \le C\{\|p\|_{i+j}H^i h^{j+1} + (\|\mathbf{u} - \mathbf{v}_{g_N}\|_m + \|(\mathbf{u} - \mathbf{v}_{g_N})\cdot\nu\|_{m,\Gamma_R})H^m\}, \\
&\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 \le C\{\|\mathbf{u} - \mathbf{u}_h\|_0 + \|\sqrt{\alpha}(\mathbf{u} - \mathbf{u}_h)\cdot\nu\|_{0,\Gamma_R}\}, \\
&\|p - p_h\|_0 \le C\{\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 + \|p\|_{i+j}H^i h^j\}, \\
&\|\nabla\cdot(\mathbf{u} - \mathbf{u}_h)\|_0 \le C\{\|\sqrt{a}(\mathcal{P}_{W_{H,h}}p - p_h)\|_0 + \|p\|_{i+j}H^i h^{j+1} + \|\nabla\cdot\mathbf{u}\|_{i+j}H^i h^j\},
\end{aligned}$$

wherein $0 \le i \le \max(0, L-j)$*,* $0 \le j \le \ell$*, and* $1 \le m \le K$*. Moreover, if* $\alpha = 0$ *or* $\Gamma_R = \emptyset$*, and if* h *is sufficiently small, then*

$$\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 \le C\{\|\nabla\cdot(\mathbf{u} - \mathbf{u}_h)\|_0 h + \|p\|_{i+j}H^i h^{j+1} + \|\mathbf{u} - \mathbf{u}_h\|_0 H\}.$$

Note that these are optimal order estimates, since $L = K$ or $L = K-1$. Moreover, $\|\mathcal{P}_{W_{H,h}}p - p_h\|_0$ is superconvergent if $\alpha = 0$ or $\Gamma_R = \emptyset$ and h is sufficiently small.

Proof. For notational convenience, let us define

$$\pi\mathbf{u} \equiv \pi(\mathbf{u} - \mathbf{v}_{g_N}) + \mathbf{v}_{g_N}.$$

The sum of the equations (6.9)–(6.10) with

$$w = \mathcal{P}_{W_{H,h}}p - p_h \in W_{H,h} \quad \text{and} \quad \mathbf{v} = \pi\mathbf{u} - \mathbf{u}_h \in \mathbf{V}_{H,h},$$

because of Proposition 6.3 and the fact that $\nabla\cdot\mathbf{V}_{H,h} = W_{H,h}$, results in

$$\begin{aligned}
&(a(\mathcal{P}_{W_{H,h}}p - p_h), \mathcal{P}_{W_{H,h}}p - p_h) + (d^{-1}(\mathbf{u} - \mathbf{u}_h), \mathbf{u} - \mathbf{u}_h) \\
&\qquad + (\alpha(\mathbf{u} - \mathbf{u}_h)\cdot\nu, (\mathbf{u} - \mathbf{u}_h)\cdot\nu)_{\Gamma_R} \\
&\quad = (a(\mathcal{P}_{W_{H,h}}p - p), \mathcal{P}_{W_{H,h}}p - p_h) + (d^{-1}(\mathbf{u} - \mathbf{u}_h), \mathbf{u} - \pi\mathbf{u}) \\
&\qquad + (\alpha(\mathbf{u} - \mathbf{u}_h)\cdot\nu, (\mathbf{u} - \pi\mathbf{u})\cdot\nu)_{\Gamma_R}.
\end{aligned}$$

If $\bar{a} \in W_{H,h}$ is the piecewise discontinuous constant average of a over the fine mesh $\mathcal{T}_{H,h}$, then

$$\begin{aligned}
&(a(\mathcal{P}_{W_{H,h}}p - p), \mathcal{P}_{W_{H,h}}p - p_h) = ((a - \bar{a})(\mathcal{P}_{W_{H,h}}p - p), \mathcal{P}_{W_{H,h}}p - p_h) \\
&\quad \le C\|a\|_{W^{1,\infty}(\Omega)}h\,\|\mathcal{P}_{W_{H,h}}p - p\|_0\|\mathcal{P}_{W_{H,h}}p - p_h\|_0.
\end{aligned}$$

Thus for any $\epsilon > 0$,

$$\begin{aligned}
&\|\sqrt{a}(\mathcal{P}_{W_{H,h}}p - p_h)\|_0 + \|\mathbf{u} - \mathbf{u}_h\|_0 + \|\sqrt{\alpha}(\mathbf{u} - \mathbf{u}_h)\cdot\nu\|_{0,\Gamma_R} \\
&\quad \le C_\epsilon\{\|p - \mathcal{P}_{W_{H,h}}p\|_0 h + \|\mathbf{u} - \pi\mathbf{u}\|_0 + \|(\mathbf{u} - \pi\mathbf{u})\cdot\nu\|_{0,\Gamma_R}\} + \epsilon\|\mathcal{P}_{W_{H,h}}p - p_h\|_0.
\end{aligned}$$

Standard elliptic lift arguments can be used to estimate $\mathcal{P}_{W_{H,h}}p - p_h$. That is, we solve (3.6)–(3.8) for ψ with w_c replaced by $\mathcal{P}_{W_{H,h}}p - p_h$ and take $\mathbf{v} = \pi\nabla\psi$. Then $\nabla\cdot\mathbf{v} = \mathcal{P}_{W_{H,h}}p - p_h$ and

$$\begin{aligned}
&\|\mathbf{v}\|_0 \le \|\nabla\psi\|_0 + \|\nabla\psi - \pi\nabla\psi\|_0 \le C\|\psi\|_2 \le C\|\mathcal{P}_{W_{H,h}}p - p_h\|_0, \\
&\|\mathbf{v}\cdot\nu\|_{0,\Gamma_R} \le \|\nabla\psi\cdot\nu\|_{0,\Gamma_R} + \|(\nabla\psi - \pi_H\nabla\psi)\cdot\nu\|_{0,\Gamma_R} \\
&\quad \le C\|\psi\|_2 \le C\|\mathcal{P}_{W_{H,h}}p - p_h\|_0,
\end{aligned}$$

using (6.7). Then (6.10) implies that

$$\begin{aligned}\|\mathcal{P}_{W_{H,h}}p - p_h\|_0^2 &= (d^{-1}(\mathbf{u}-\mathbf{u}_h),\mathbf{v}) + (\alpha(\mathbf{u}-\mathbf{u}_h)\cdot\nu,\mathbf{v}\cdot\nu)_{\Gamma_R}\\ &\le C\{\|\mathbf{u}-\mathbf{u}_h\|_0\|\mathbf{v}\|_0 + \|\sqrt{\alpha}(\mathbf{u}-\mathbf{u}_h)\cdot\nu\|_{0,\Gamma_R}\|\mathbf{v}\cdot\nu\|_{0,\Gamma_R}\\ &\le C\{\|\mathbf{u}-\mathbf{u}_h\|_0 + \|\sqrt{\alpha}(\mathbf{u}-\mathbf{u}_h)\cdot\nu\|_{0,\Gamma_R}\}\|\mathcal{P}_{W_{H,h}}p - p_h\|_0,\end{aligned}$$

and, with Lemmas 6.1 and 6.4, the first three estimates of the theorem follow.

If $\alpha = 0$ or $\Gamma_R = \emptyset$, we replace (3.6)–(3.8) by

$$a\psi - \nabla\cdot d\nabla\psi = \mathcal{P}_{W_{H,h}}p - p_h \qquad \text{in } \Omega, \tag{6.11}$$

$$-d\nabla\psi\cdot\nu = 0 \qquad \text{on } \Gamma_N, \tag{6.12}$$

$$\psi = 0 \qquad \text{on } \Gamma_R, \tag{6.13}$$

and we modify the argument as follows [16]:

$$\begin{aligned}\|\mathcal{P}_{W_{H,h}}p - p_h\|_0^2 &= (\mathcal{P}_{W_{H,h}}p - p_h, a\psi - \nabla\cdot d\nabla\psi)\\ &= (a(\mathcal{P}_{W_{H,h}}p - p_h),\psi) - (\mathcal{P}_{W_{H,h}}p - p_h, \nabla\cdot\pi d\nabla\psi),\end{aligned}$$

and, using (6.10),

$$\begin{aligned}(\mathcal{P}_{W_{H,h}}p - p_h, \nabla\cdot\pi d\nabla\psi) &= (d^{-1}(\mathbf{u}-\mathbf{u}_h),\pi d\nabla\psi)\\ &= (\mathbf{u}-\mathbf{u}_h,\nabla\psi) - (d^{-1}(\mathbf{u}-\mathbf{u}_h), d\nabla\psi - \pi d\nabla\psi)\\ &= -(\nabla\cdot(\mathbf{u}-\mathbf{u}_h),\psi) - (d^{-1}(\mathbf{u}-\mathbf{u}_h), d\nabla\psi - \pi d\nabla\psi),\end{aligned}$$

and, for $w \in W_{H,h}$ arbitrary, by (6.9),

$$\begin{aligned}(\nabla\cdot(\mathbf{u}-\mathbf{u}_h),\psi) &= (\nabla\cdot(\mathbf{u}-\mathbf{u}_h),\psi - w) - (a(\mathcal{P}_{W_{H,h}}p - p_h), w)\\ &\quad - (a(p - \mathcal{P}_{W_{H,h}}p), w)\\ &= (\nabla\cdot(\mathbf{u}-\mathbf{u}_h),\psi - w) - (a(\mathcal{P}_{W_{H,h}}p - p_h),\psi)\\ &\quad - (a(\mathcal{P}_{W_{H,h}}p - p_h), w - \psi) - ((a-\bar{a})(p - \mathcal{P}_{W_{H,h}}p), w).\end{aligned}$$

Since a good choice of w implies

$$\|w - \psi\|_0 \le C|\psi|_1 h \quad \text{and} \quad \|w\|_0 \le C|\psi|_1,$$

we have that

$$\begin{aligned}\|\mathcal{P}_{W_{H,h}}p - p_h\|_0^2 &\le C\{(\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_0 + \|\mathcal{P}_{W_{H,h}}p - p_h\|_0 + \|p - \mathcal{P}_{W_{H,h}}p\|_0)|\psi|_1 h\\ &\quad + \|\mathbf{u}-\mathbf{u}_h\|_0|\psi|_2 H\}\\ &\le C\{(\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_0 + \|\mathcal{P}_{W_{H,h}}p - p_h\|_0 + \|p - \mathcal{P}_{W_{H,h}}p\|_0)h\\ &\quad + \|\mathbf{u}-\mathbf{u}_h\|_0 H\}\|\mathcal{P}_{W_{H,h}}p - p_h\|_0,\end{aligned}$$

and the final result of the theorem follows for h sufficiently small.

Finally, (6.9) with $w = \nabla\cdot(\pi\mathbf{u} - \mathbf{u}_h) \in W_{H,h}$ implies that

$$\begin{aligned}\|\nabla\cdot(\pi\mathbf{u} - \mathbf{u}_h)\|_0^2 &= -(a(\mathcal{P}_{W_{H,h}}p - p_h), \nabla\cdot(\pi\mathbf{u} - \mathbf{u}_h))\\ &\quad - ((a-\bar{a})(p - \mathcal{P}_{W_{H,h}}p), \nabla\cdot(\pi\mathbf{u} - \mathbf{u}_h)).\end{aligned}$$

Since $\nabla\cdot(\mathbf{u} - \pi\mathbf{u}) = (I - \mathcal{P}_{W_{H,h}})\nabla\cdot\mathbf{u}$, which approximates as in (6.6), the divergence estimate of the theorem follows. ∎

In the special case $a = \alpha = 0$, we obtain optimality of the finite element approximation $\mathbf{u}_h$ to $\mathbf{u}$ in the energy norm subject to the appropriate divergence constraint.

THEOREM 6.6. *If $a = \alpha = 0$ and (3.1) and (6.1) hold, then*

$$\|d^{-1/2}(\mathbf{u}-\mathbf{u}_h)\|_0 \leq \inf_{\substack{\mathbf{v}_h\in\mathbf{V}_{H,h}\\ \nabla\cdot\mathbf{v}_h=\mathcal{P}_{W_{H,h}}\nabla\cdot(\mathbf{u}-\mathbf{v}_{g_N})}} \|d^{-1/2}(\mathbf{u}-\mathbf{v}_{g_N}-\mathbf{v}_h)\|_0 \leq C\|\mathbf{u}-\mathbf{v}_{g_N}\|_m H^m,$$

$$\nabla\cdot\mathbf{u}_h = \mathcal{P}_{W_{H,h}}\nabla\cdot\mathbf{u},$$

$$\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 \leq C\|\mathbf{u}-\mathbf{u}_h\|_0,$$

$$\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 \leq C\{\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_0 h + \|\mathbf{u}-\mathbf{u}_h\|_0 H\},$$

$$\|p-p_h\|_0 \leq C\{\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 + \|p\|_{i+j}H^i h^j\},$$

wherein $0 \leq i \leq \max(0, L-j)$, $0 \leq j \leq \ell$, and $1 \leq m \leq K$.

Proof. For any $\mathbf{v}_h \in \mathbf{V}_{H,h}$ such that $\nabla\cdot\mathbf{v}_h = \mathcal{P}_{W_{H,h}}\nabla\cdot(\mathbf{u}-\mathbf{v}_{g_N})$, take

$$w = \mathcal{P}_{W_{H,h}}p - p_h \in W_{H,h},$$
$$\mathbf{v} = (\mathbf{u}-\mathbf{u}_h) - (\mathbf{u}-\mathbf{v}_{g_N}-\mathbf{v}_h) \in \mathbf{V}_{H,h}$$

in (6.9)–(6.10). Then with $a = \alpha = 0$, the sum of the equations implies that

$$\begin{aligned}(d^{-1}(\mathbf{u}-\mathbf{u}_h), \mathbf{u}-\mathbf{u}_h) &= (d^{-1}(\mathbf{u}-\mathbf{u}_h), \mathbf{u}-\mathbf{v}_{g_N}-\mathbf{v}_h)\\ &\leq \|d^{-1/2}(\mathbf{u}-\mathbf{u}_h)\|_0 \|d^{-1/2}(\mathbf{u}-\mathbf{v}_{g_N}-\mathbf{v}_h)\|_0\end{aligned}$$

so that

$$\|d^{-1/2}(\mathbf{u}-\mathbf{u}_h)\|_0 \leq \inf_{\mathbf{v}_h}\|d^{-1/2}(\mathbf{u}-\mathbf{v}_{g_N}-\mathbf{v}_h)\|_0 \leq C\|\mathbf{u}-\mathbf{v}_{g_N}-\pi(\mathbf{u}-\mathbf{v}_{g_N})\|_0,$$

since $\pi(\mathbf{u}-\mathbf{v}_{g_N})$ satisfies the divergence constraint by Proposition 6.3. Lemma 6.4 gives the required approximation result for the first estimate of the theorem.

Now $a = 0$ and (6.9) imply that $\nabla\cdot\mathbf{u}_h = \mathcal{P}_{W_{H,h}}\nabla\cdot\mathbf{u}$, giving the second result of the theorem. The final estimates follow as in the previous proof. ∎

The underlying assumption in the error analysis, and the tacit assumption in all similar subgrid methods mentioned in the introduction, is that the finest grid, scale h, resolves the fine-scale details of the solution. We see this here in the Sobolev norms appearing in the error estimates. If h does not resolve the subgrid problems, then we cannot expect a good approximation.

Under this tacit assumption, we proved two main results. First, in Theorem 6.6, the solution is optimally approximated in the finite element space subject to the fine-scale divergence constraint. Thus the approximation is no worse than using only a coarse-scale approximation (up to questions of the scale of the divergence constraint). It is presumably much better, as numerical results show [1, 2, 4]. Second, the pressure is approximated on the finest scale, up to a higher order coarse H error term. This is a strict improvement over merely solving on the coarse scale. Moreover, these improvements are achieved for negligible additional numerical cost compared to the coarse-scale solution and much less cost than the fine-scale solution itself [1, 4].

The error estimates of this subsection appear to explain the numerical results that have appeared elsewhere: that they are amazingly accurate for reasonable problems where the fine grid resolves the coefficients of the problem [1, 2, 4]. The error estimates also explain why numerical results break down in certain difficult cases in which h does not resolve the problem well. If h and H are relatively large compared to the Sobolev norms of the solution, the method is not expected to work well. This appears to be the main factor limiting the utility of the method in practice. As an example, in [2], a numerical test was presented involving Darcy flow in a porous medium with a long thin high permeability streak. This is a very difficult problem to resolve numerically on a coarse grid. The streak introduces channelized flow, which means that the derivatives of the solution are large, and so by Theorem 6.5 or 6.6, we neither expect nor see good results. (It should be noted, however, that the upscaling technique of this paper is not entirely unsuccessful in resolving the solution of even this very difficult problem.)

7. Two special cases. A particularly pertinent choice of spaces was studied numerically in [1, 2]. On the coarse mesh, we approximate $(p_c, \mathbf{u}_c)$ in the two-dimensional BDM-1 or three-dimensional BDDF-1 mixed finite element spaces [12, 10]. The space of scalars W_H^* consists of piecewise discontinuous constants, and the space of vectors $\mathbf{V}_H^*$ is second order accurate and has linear fluxes on the edges or faces of coarse elements.

On each coarse element $E_c \in \mathcal{T}_H$, we approximate $(\delta\bar{p}, \delta\bar{\mathbf{u}})$, $(\delta\tilde{p}, \delta\tilde{\mathbf{u}})$, and $(\delta\hat{p}, \delta\hat{\mathbf{u}})$ in the RT-0 spaces [28]. These approximate with piecewise discontinuous constants for $\delta W_h(E_c)$ and with constant fluxes on each interior edge or face for $\delta\mathbf{V}_h(E_c)$.

In this case, $\mathbf{V}_H^* \cap \delta\mathbf{V}_h = \{0\}$, so $\mathbf{V}_H = \mathbf{V}_H^*$ requires no attention. Also, $W_{H,h} = W_h^*$, the space of piecewise discontinuous constants over the fine mesh $\mathcal{T}_{H,h}$. We note that $W_H^* \perp \delta W_h$, so $W_H = W_H^*$ and condition (6.1) holds. However, it is not so simple to compute with $\delta W_h(E_c)$, since such functions are supported in all of E_c. Fortunately, a careful implementation of the scheme [2] allows one to avoid working over $\delta W_h(E_c)$ and instead work over the full space $\delta W_h(E_c) + \operatorname{span}\{1\}$ of piecewise discontinuous constants over $\mathcal{T}_h(E_c)$. Now $K = 2$ and $L = k = \ell = 1$, so Theorem 6.5 takes the following simple form.

THEOREM 7.1. *If $a \in W^{1,\infty}(\Omega)$ and (3.1) holds, the BDDF-1(BDM-1)/RT-0 two-scale approximation satisfies the error bounds*

$$\begin{aligned}
&\|\sqrt{a}(\mathcal{P}_{W_{H,h}}p - p_h)\|_0 + \|\mathbf{u} - \mathbf{u}_h\|_0 + \|\sqrt{\alpha}(\mathbf{u} - \mathbf{u}_h)\cdot\nu\|_{0,\Gamma_R} \\
&\quad \le C\{\|p\|_1 h^2 + (\|\mathbf{u} - \mathbf{v}_{g_N}\|_2 + \|(\mathbf{u} - \mathbf{v}_{g_N})\cdot\nu\|_{2,\Gamma_R})H^2\} \le CH^2, \\
&\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 \le C\{\|\mathbf{u} - \mathbf{u}_h\|_0 + \|\sqrt{\alpha}(\mathbf{u} - \mathbf{u}_h)\cdot\nu\|_{0,\Gamma_R}\} \le CH^2, \\
&\|p - p_h\|_0 \le C\{\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 + \|p\|_1 h\} \le C(H^2 + h), \\
&\|\nabla\cdot(\mathbf{u} - \mathbf{u}_h)\|_0 \le C\{\|\sqrt{a}(\mathcal{P}_{W_{H,h}}p - p_h)\|_0 + \|p\|_1 h^2 + \|\nabla\cdot\mathbf{u}\|_1 h\} \le C(H^2 + h).
\end{aligned}$$

Moreover, if $\alpha = 0$ or $\Gamma_R = \emptyset$, and h is sufficiently small, then

$$\|\mathcal{P}_{W_{H,h}}p - p_h\|_0 \le C\{\|\nabla\cdot(\mathbf{u} - \mathbf{u}_h)\|_0 h + \|p\|_1 h^2 + \|\mathbf{u} - \mathbf{u}_h\|_0 H\} \le C(H^3 + h^2).$$

If $H^2 \sim h$ as $H \to 0$, then p and $\mathbf{u}$ are resolved on the fine scale to order h. Hence relatively good numerical approximation results have been obtained [1, 3, 4, 2].

A second special choice is to use RT-0 spaces on both scales. In this case, $K = L = k = \ell = 1$, and Theorem 6.5 would suggest that p and $\mathbf{u}$ are approximated only on the coarse scale to order H. Although we retain the optimality of the solution

in the energy norm under the appropriate conditions, the numerical approximation results are not nearly as good as in the previous special case (see [5]).

REFERENCES

[1] T. Arbogast, *Numerical subgrid upscaling of two-phase flow in porous media*, in Numerical Treatment of Multiphase Flows in Porous Media, Lecture Notes in Phys. 552, Z. Chen, R. E. Ewing, and Z.-C. Shi, eds., Springer, Berlin, 2000, pp. 35–49.

[2] T. Arbogast, *Implementation of a locally conservative numerical subgrid upscaling scheme for two-phase Darcy flow*, Comput. Geosci., 6 (2002), pp. 453–481.

[3] T. Arbogast and S. Bryant, *Efficient forward modeling for DNAPL site evaluation and remediation*, in Computational Methods in Water Resources XIII, L. R. Bentley, J. F. Sykes, C. A. Brebbia, W. G. Gray, and G. F. Pinder, eds., Balkema, Rotterdam, 2000, pp. 161–166.

[4] T. Arbogast and S. L. Bryant, *A two-scale numerical subgrid technique for waterflood simulations*, SPE J., 7 (2002), pp. 446–457.

[5] T. Arbogast, S. E. Minkoff, and P. T. Keenan, *An operator-based approach to upscaling the pressure equation*, in Computational Methods in Water Resources XII, Vol. 1, V. N. Burganos, G. P. Karatzas, A. C. Payatakes, C. A. Brebbia, W. G. Gray, and G. F. Pinder, eds., Computational Mechanics Publications, Southampton, UK, 1998, pp. 405–412.

[6] I. Babuška, *The finite element method with Lagrangian multipliers*, Numer. Math., 20 (1973), pp. 179–192.

[7] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[8] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 129–151.

[9] F. Brezzi, *Interacting with the subgrid world*, in Proceedings of the Conference on Numerical Analysis, Dundee, Scotland, 1999.

[10] F. Brezzi, J. Douglas, Jr., R. Durán, and M. Fortin, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.

[11] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

[12] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

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

[14] Z. Chen and J. Douglas, Jr., *Prismatic mixed finite elements for second order elliptic problems*, Calcolo, 26 (1989), pp. 135–148.

[15] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[16] J. Douglas, Jr., and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.

[17] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[18] V. Girault and J.-L. Lions, *Two-grid finite-element schemes for the transient Navier-Stokes problem*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 945–980.

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

[20] T. Y. Hou and X. H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.

[21] T. J. R. Hughes, *Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 387–401.

[22] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy, *The variational multiscale method—A paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.

[23] T. J. R. Hughes, A. A. Oberai, and L. Mazzei, *Large eddy simulation of turbulent channel flows by the variational multiscale method*, Phys. Fluids, 13 (2001), pp. 1784–1799.

[24] J. C. Nedélec, *Mixed finite elements in* $\mathbf{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[25] J. T. Oden and K. S. Vemaganti, *Adaptive hierarchical modeling of heterogeneous structures*,

Phys. D, 133 (1999), pp. 404–415.
[26] J. T. Oden and K. S. Vemaganti, *Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials.* I. *Error estimates and adaptive algorithms*, J. Comput. Phys., 164 (2000), pp. 22–47.
[27] M. Peszynska, M. F. Wheeler, and I. Yotov, *Mortar upscaling for multiphase flow in porous media*, Comput. Geosci., 6 (2002), pp. 73–100.
[28] R. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, New York, 1977, pp. 292–315.
[29] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1991.
[30] J. Xu, *Two-grid discretization techniques for linear and nonlinear PDEs*, SIAM. J. Numer. Anal., 33 (1996), pp. 1759–1777.
[31] J. Xu, *The method of subspace corrections*, J. Comput. Appl. Math., 128 (2001), pp. 335–362.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 599–620

A GRID-BASED BOUNDARY INTEGRAL METHOD FOR ELLIPTIC PROBLEMS IN THREE DIMENSIONS*

J. THOMAS BEALE†

Abstract. We develop a simple, efficient numerical method of boundary integral type for solving an elliptic partial differential equation in a three-dimensional region using the classical formulation of potential theory. Accurate values can be found near the boundary using special corrections to a standard quadrature. We treat the Dirichlet problem for a harmonic function with a prescribed boundary value in a bounded three-dimensional region with a smooth boundary. The solution is a double layer potential, whose strength is found by solving an integral equation of the second kind. The boundary surface is represented by rectangular grids in overlapping coordinate systems, with the boundary value known at the grid points. A discrete form of the integral equation is solved using a regularized form of the kernel. It is proved that the discrete solution converges to the exact solution with accuracy $O(h^p)$, $p < 5$, depending on the smoothing parameter. Once the dipole strength is found, the harmonic function can be computed from the double layer potential. For points close to the boundary, the integral is nearly singular, and accurate computation is not routine. We calculate the integral by summing over the boundary grid points and then adding corrections for the smoothing and discretization errors using formulas derived here; they are similar to those in the two-dimensional case given by [J. T. Beale and M.-C. Lai, *SIAM J. Numer. Anal.*, 38 (2001), pp. 1902–1925]. The resulting values of the solution are uniformly of $O(h^p)$ accuracy, $p < 3$. With a total of N points, the calculation could be done in essentially $O(N)$ operations if a rapid summation method is used.

Key words. boundary integral methods, nearly singular integrals, potential theory, Dirichlet problem, overlapping grids

AMS subject classifications. 65R20, 65D30, 31B10, 35J25

DOI. 10.1137/S0036142903420959

1. Introduction.

In this paper we develop a simple, direct numerical method of boundary integral type for the solution of an elliptic boundary value problem in a three-dimensional (3D) region with a smooth boundary. We emphasize the Dirichlet problem for an interior harmonic function, although a larger class of problems can be treated. We assume the boundary surface is represented by several overlapping grids, each rectangular in some coordinate system. Values of the solution are to be found at arbitrary points in space (typically, those on a regular 3D grid), some of which will be close to the surface. We do not assume any relationship between the coordinate grids on the surface and the interior grid. We use the classical formulation of the Dirichlet problem in potential theory: The solution can be written as a double layer potential, or dipole layer, with an unknown strength. This dipole strength is determined by a Fredholm integral equation of the second kind on the surface. We solve a discretized form of this integral equation at the coordinate grid points on the surface, replacing the Green's function in the layer potential with a regularized version and the integral with a sum over grid points. We prove that the solution of the numerical integral equation converges to the exact solution with order h^p, for any $p < 5$, depending on the choice of the smoothing parameter relative to h, and that it can be found by a simple iteration. Once the dipole strength is found, the value of the harmonic function at any point is given by the double layer potential. For points away from

*Received by the editors January 2, 2003; accepted for publication (in revised form) October 13, 2003; published electronically April 14, 2004. This research was supported by the NSF under grant DMS-0102356.
http://www.siam.org/journals/sinum/42-2/42095.html
†Department of Mathematics, Duke University, Durham, NC 27708-0320 (beale@math.duke.edu).

the boundary, the surface integral can be computed in a standard way, but, at points near the boundary, the integral is nearly singular, and accurate computation is not routine. We find the integral in this case by starting with a standard quadrature, again regularized, and then adding corrections. These corrections account for the errors introduced by the smoothing, or regularization, and by discretization. They are derived by local analysis near the singularity, extending the treatment for a two-dimensional (2D) region given in [3]. With grid spacing h on the boundary, the resulting values are accurate to $O(h^p)$ for $p < 3$, again depending on smoothing.

Boundary integral formulations of the Laplace or Helmholtz equation in three dimensions are widely used in engineering, especially electromagnetics. They are most often solved numerically by the boundary element method; the surface is triangulated, and the equation holds at collocation points (see, e.g., [2, 15]). The integration of the kernel times a basis function at a point in its support often uses a change of variables or a product integration rule. Special care is also needed for points near the support. High order accuracy can be achieved [2, section 9.2], and the method can be accurate for piecewise smooth surfaces. More direct quadrature, or Nyström, methods have not been widely used. Such a method for polyhedral surfaces, summing over centroids in a triangulation, was shown to converge by Rathsfeld [23]. In [24] he showed that a simple quadrature over a triangulation of a smooth surface gives almost $O(h^2)$ convergence with singularity subtraction, and this can be improved to higher order using product integration. Recently, practical methods of this type have been developed. Canino et al. [8] use a triangulation with quadrature based on a local correction method of Strain, an alternative to product integration. Bruno and Kunyansky [7] use overlapping grids and a partition of unity, as in the present work, but they integrate the singularity by a change of variables. Other current approaches include use of wavelets (see, e.g., [19]) or spherical polynomials (see, e.g., [13]). The difficulty of calculating nearly singular integrals for the field at points near the boundary is well recognized [2, section 7.2.1], but very few works have treated it; two recent approaches are given in [17, 25].

While there are many choices of numerical methods for elliptic problems, with boundary integrals or otherwise, the present approach has several advantages for some applications. The data structure needed to represent the surface is minimal, and calculations on the surface are done in the rectangular grids, with the dependence on the surface reduced to coordinate functions. The integration requires no extra work at the singularity while solving the integral equation. Finally, values of the field near the surface can be found by direct integration, adding the corrections presented here. These attributes may be especially important for a time-dependent calculation with a moving boundary. For viscous, incompressible fluid flow, pressure terms due to a boundary could be found as nearly singular integrals [3]. The linear system resulting from the integral equation is well conditioned. With n boundary points, the direct solution of the linear system would require about $O(n^2)$ operations, but this could be reduced to about $O(n)$ using a rapid summation method such as the fast multipole method of Greengard and Rokhlin [14]. Similarly, fast summation could be used to produce the values of the harmonic function. Another way to produce the interior solution, as observed by Mayo [21] (see also [1]), is to find values near the boundary, compute a discrete Laplacian, extend it by zero to a computational box including the original region, and then invert using a fast Poisson solver to produce values at grid points. This approach was used in the 2D case in [3, section 4]. Either way, the operation count with a total of N points could be reduced to about $O(N)$. We assume here that the surface and functions are smooth, i.e., have several derivatives

to justify various Taylor expansions; this method would not be valid for a boundary with corners and edges without further modification.

We now summarize the method and results. We consider the Dirichlet problem for a bounded domain $\Omega \subseteq \mathbf{R}^3$ with smooth boundary $\mathcal{S}$. Given a function g on $\mathcal{S}$ we want to find a function u on Ω such that

$$\Delta u = 0 \quad \text{on } \Omega, \qquad u = g \quad \text{on } \mathcal{S}. \tag{1.1}$$

The solution can be written as a double layer potential

$$u(y) = \int_{\mathcal{S}} \frac{\partial}{\partial n(x)} G(x-y) f(x)\, dS(x) \tag{1.2}$$

for some dipole strength f, determined by g, where $n(x)$ is the unit outward normal at $x \in \mathcal{S}$, G is the Green's function for Δ in $\mathbf{R}^3$, $G(x) = -1/4\pi|x|$, and

$$\frac{\partial}{\partial n(x)} G(x-y) \;=\; n(x) \cdot \nabla G(x-y) \;=\; \frac{n(x)\cdot(x-y)}{4\pi|x-y|^3}. \tag{1.3}$$

The unknown f is the solution of the integral equation on the surface $\mathcal{S}$:

$$\frac{1}{2} f(x) + \int_{\mathcal{S}} K(x,x') f(x')\, dS(x') \;=\; g(x), \qquad K(x,x') = \frac{\partial}{\partial n(x)} G(x'-x), \tag{1.4}$$

or $f + 2Tf = 2g$, where T is the integral operator with kernel K. It is known (see, e.g., [18, section 10.5]) that the iterates f^n defined by

$$f^{n+1} = (1-\beta) f^n - 2\beta T f^n + 2\beta g \tag{1.5}$$

converge to the solution of (1.4), provided $0 < \beta < 1$.

To describe the numerical formulation of the problem (1.4) for f, we suppose the surface $\mathcal{S}$ is covered by several coordinate patches $X^\sigma : U_\sigma \to \mathcal{S}$, where U_σ is an open subset of $\mathbf{R}^2$. We assume each $X^\sigma : U_\sigma \to \mathbf{R}^3$ is smooth and nondegenerate; i.e., $\partial X^\sigma(\alpha)/\partial\alpha$ has rank 2 at each point, with $\alpha = (\alpha_1, \alpha_2)$. In order to write the integral as a sum of integrals in coordinates, we introduce a partition of unity, that is, a collection of smooth functions $\{\psi^\sigma\} : \mathcal{S} \to \mathbf{R}$ such that ψ^σ is zero outside a compact subset of U_σ, called the support of ψ^σ, and $\sum_\sigma \psi^\sigma(x) = 1$ for each $x \in \mathcal{S}$. (See section 5 for typical choices.) We can now write the integral of a function F on $\mathcal{S}$ as

$$\int_{\mathcal{S}} F(x')\, dS(x') \;=\; \sum_\sigma \int_{U_\sigma} F(X^\sigma(\alpha)) \psi^\sigma(X^\sigma(\alpha)) A^\sigma(\alpha) d\alpha, \tag{1.6}$$

where $A^\sigma(\alpha)d\alpha$ is the element of surface area in the σth patch. In solving (1.4) it is helpful to use the familiar identity

$$\int_{\mathcal{S}} K(x,x')\, dS(x') = \frac{1}{2} \tag{1.7}$$

to reduce the order of the singularity; we rewrite (1.4) as

$$f(x) + \int_{\mathcal{S}} K(x,x')[f(x') - f(x)]\, dS(x') \;=\; g, \qquad x \in S. \tag{1.8}$$

Since $K = O(1/r)$ on $\mathcal{S}$, the resulting integrand is bounded, although not smooth. We will write the integral in (1.8) as in (1.6), replacing K by a regularized version K_δ, with a shape factor s and a smoothing parameter δ to be chosen:

(1.9)
$$K_\delta(x,x') = n(x') \cdot \nabla G_\delta(x'-x), \qquad \nabla G_\delta(x'-x) = \nabla G(x'-x) s(|x-x'|/\delta).$$

With the natural choice $G_\delta(x) = G(x)\,\mathrm{erf}(|x|/\delta) = -\,\mathrm{erf}(|x|/\delta)/4\pi|x|$, we have

$$s(r) = \mathrm{erf}(r) - (2/\sqrt{\pi}) r e^{-r^2}, \tag{1.10}$$

where erf is the usual error function

$$\mathrm{erf}(r) = \frac{2}{\sqrt{\pi}} \int_0^r e^{-t^2}\, dt. \tag{1.11}$$

In section 2 we find that this choice of smoothing leads to an $O(\delta^3)$ error to the integral in (1.8). Moreover, this can be improved to $O(\delta^5)$ with a slight change to

$$s(r) = \mathrm{erf}(r) - (2/\sqrt{\pi})(r - 2r^3/3) e^{-r^2}. \tag{1.12}$$

We can now give the discrete form of (1.8). Suppose a grid spacing h is chosen for the coordinate patches, and the σth patch has grid points $x_i^\sigma = X^\sigma(ih)$ for $i \in Z^2$ such that $x_i^\sigma \in V_\sigma$, where V_σ is the interior of the support of ψ^σ. We assume the specified function g is known at these points, say, $g_i^\sigma = g(x_i^\sigma)$. Then

$$f_i^\sigma + \sum_{j,\tau} K_{ij}^{\sigma\tau} \psi_j^\tau [f_j^\tau - f_i^\sigma] A_j^\tau\, h^2 \;=\; g_i^\sigma \tag{1.13}$$

is the discrete analogue of (1.8), where

$$K_{ij}^{\sigma\tau} = K_\delta(x_i^\sigma, x_j^\tau), \quad \psi_j^\tau = \psi^\tau(x_j^\tau), \quad A_j^\tau = A^\tau(jh). \tag{1.14}$$

We choose δ in terms of h so that $\delta \to 0$ as $h \to 0$, but, for some constant ρ_0,

$$\rho \;=\; \delta/h \geq \rho_0 > 0. \tag{1.15}$$

We can find the solution f_i^σ as the limit of the iteration corresponding to (1.5),

$$f_i^{\sigma,n+1} \;=\; (1-2\beta) f_i^{\sigma,n} - 2\beta \sum_{j,\tau} K_{ij}^{\sigma\tau} \psi_j^\tau [f_j^{\tau,n} - f_i^{\sigma,n}] A_j^\tau\, h^2 \;+\; 2\beta g_i^\sigma. \tag{1.16}$$

The following theorem assures the validity of this procedure.

THEOREM 1.1. *For the interior Dirichlet problem* (1.1)*, with $\mathcal{S}$ and g smooth, suppose that the grid size h and the smoothing radius δ are chosen sufficiently small subject to the condition* (1.15)*. Then the discrete integral equation* (1.13)*, using the regularization* (1.9)*,* (1.12)*, has a unique solution f_i^σ, which converges uniformly to the exact solution f of* (1.4) *or* (1.8) *as $h \to 0$, $\delta \to 0$, with the estimate*

$$|f_i^\sigma - f(x_i^\sigma)| \;\leq\; C_1 \delta^5 + C_2 h^2 e^{-c_0 (\delta/h)^2}. \tag{1.17}$$

Here c_0 depends only on the coordinate patches, but C_1, C_2 depend on bounds for derivatives of f. For $0 < \beta < 1$, the iterates defined by (1.16) *converge uniformly to the discrete solution f_i^σ, with a rate independent of h and δ.*

The convergence proof must take into account the agreement of function values where the grids overlap. This is done using a discrete Hölder norm. The error from the smoothing gives the first term in (1.17); the discretization contributes the second term, plus a remainder smaller than the first. The second term is dominated by the first if $\rho = \delta/h$ grows slowly as $h, \delta \to 0$. For example, if we choose $\delta = ch^q$, with $q < 1$, the error is $O(\delta^5) = O(h^{5q})$. Similarly, if we use the simpler choice of smoothing (1.10) we obtain almost $O(h^3)$ convergence. The constant c_0 is $\pi^2\gamma^2/2$, where γ, given by (3.19), depends on g^{ij}; see (3.25).

After solving the discrete integral equation for the dipole strength f at the coordinate grid points x_i^σ, we now want to find the values of the solution u of (1.1) from the double layer representation (1.2). For a point y away from the surface we can discretize the integral routinely, since the integrand is smooth. For y near the surface, however, the integral is nearly singular, and accurate calculation requires more care. In contrast to the case $y \in \mathcal{S}$, $n(x) \cdot \nabla G(x-y) = O(|x-y|^{-2})$. Given y near $\mathcal{S}$, there is a point x_0 on the surface so that y is on the normal line through x_0; i.e., $y = x_0 + bn_0$, for some b, where n_0 is the outward unit normal at x_0. We can use Green's identity to replace (1.2) with a subtracted form

$$u(y) = \int_{\mathcal{S}} \frac{\partial}{\partial n(x)} G(x-y)[f(x) - f(x_0)]\, dS(x) + f(x_0), \qquad y \in \Omega. \tag{1.18}$$

We can approximate the integral above by the sum

$$S = \sum_{\sigma,j} n(x_j^\sigma) \cdot \nabla G_\delta(x_j^\sigma - y)[f(x_j^\sigma) - f(x_0)]\psi_j^\sigma A_j^\sigma h^2. \tag{1.19}$$

We use the simpler regularization (1.10); it seems that higher order smoothing cannot be incorporated in the modified kernel for the nearly singular case. This sum typically has a large error for y near S. We can think of the error in two parts: the smoothing error from replacing ∇G by ∇G_δ and the discretization error from replacing the integral with ∇G_δ by the sum or, briefly,

$$\sum_\delta - \int = \left(\int_\delta - \int\right) + \left(\sum_\delta - \int_\delta\right). \tag{1.20}$$

The analysis of the following sections shows that these errors, uniform with respect to y near $\mathcal{S}$, are $O(\delta^2)$ and $O(h)$, respectively. However, the largest errors can be removed by corrections which we now describe. They are derived in sections 2 and 3 and are analogous to those found in the 2D case in [3].

The corrections involve the geometry of the surface near x_0. There is at least one σ so that x_0 is in the σth patch; i.e., $x_0 = X^\sigma(\alpha_0)$ for some $\alpha_0 \in U^\sigma$. Let T_i be the tangent vector $(\partial X^\sigma/\partial\alpha_i)(\alpha_0)$ at x_0, $i = 1, 2$. We use the metric tensor $g_{ij} = T_i \cdot T_j$; its determinant $g = \det(g_{ij})$; the inverse $g^{ij} = (g_{ij})^{-1}$; and the surface Laplacian

$$\Delta f = \sum_{i,j=1}^{2} \frac{1}{\sqrt{g}} \frac{\partial}{\partial\alpha_j} \left(\sqrt{g}\, g^{ij} \frac{\partial(f \circ X^\sigma)}{\partial\alpha_i}\right). \tag{1.21}$$

If the grid points $\{x_j^\sigma\}$ and values f_j^σ are known, these quantities can be found at α_0 by interpolation. The smoothing correction derived in section 2 is, with $\lambda = b/\delta$,

$$\mathcal{T}_1 = \delta^2(\Delta f(x_0))(\lambda/4)(|\lambda| \operatorname{erfc}|\lambda| - e^{-\lambda^2}/\sqrt{\pi}). \tag{1.22}$$

The discretization correction comes from the Poisson summation formula applied to the regularized kernel. It uses the function

$$E(p,q) = e^{2pq}\operatorname{erfc}(p+q) + e^{-2pq}\operatorname{erfc}(-p+q), \tag{1.23}$$

where erfc $= 1 -$ erf is the complementary error function. The point x_0 may be in more than one coordinate patch, and a correction term is needed for each. In the σth patch, $x_0 = X^\sigma(\alpha_0)$ for some α_0 depending on σ. We can write $\alpha_0 = ih + \nu h$ for some $i \in Z^2$ and $\nu = (\nu_1, \nu_2)$, with $0 \le \nu_s \le 1$, $s = 1, 2$. The correction for the σth patch is

$$\mathcal{T}_2^\sigma = -h \sum_{r=1}^{2} c_r \psi^\sigma(\alpha_0) \frac{\partial (f \circ X^\sigma)}{\partial \alpha_r}(\alpha_0), \tag{1.24}$$

$$c_r = \frac{\rho\lambda}{2} \sum_{s=1}^{2} \sum_{n \in Q} a(n,s) \sin(2\pi n \cdot \nu) \frac{g^{rs} n_s}{\|n\|} E(\lambda, \pi\rho\|n\|). \tag{1.25}$$

Here $Q = \{n = (n_1, n_2) \in \mathbf{Z}^2 : n_2 \ge 0, n \ne 0\}$; $\|n\| = \sqrt{g^{ij} n_i n_j}$; $a(n,s) = 1/2$ when $s = 1$ and $n_2 = 0$, and $a(n,s) = 1$ otherwise.

The following theorem summarizes the error estimates for the computed value of $u(y)$ in (1.2), starting with the sum (1.19) and adding corrections.

THEOREM 1.2. *For $y \in \Omega$, let $u(y)$ be the exact solution of* (1.1)*, given by* (1.2)*, assuming f and $\mathcal{S}$ are smooth. Let $\tilde{u}(y)$ be the value computed as the sum*

$$\tilde{u}(y) \;=\; S \;+\; f(x_0) \;+\; \mathcal{T}_1 \;+\; \sum_\sigma \mathcal{T}_2^\sigma, \tag{1.26}$$

with $S, \mathcal{T}_1, \mathcal{T}_2^\sigma$ given by (1.19)*,* (1.10)*,* (1.21)*–*(1.25)*, subject to* (1.15)*. Then the error has the form $\tilde{u}(y) - u(y) = \varepsilon_1 + \varepsilon_2$, where the smoothing error ε_1 and the discretization error ε_2 can be estimated as follows, uniformly for y near $\mathcal{S}$, with c_0 as before:*

$$|\varepsilon_1| \le C_1 \delta^3, \qquad |\varepsilon_2| \le C_2 h^2 e^{-c_0(\delta/h)^2} + C_3 h^3. \tag{1.27}$$

Again we can choose δ so that $\rho = \delta/h$ grows slowly and the error is almost $O(h^3)$. Thus, for $\delta = ch^q$ with $q < 1$, the error is $O(\delta^3) = O(h^{3q})$. The sum in (1.25) is infinite, but the terms decay at a Gaussian rate, independent of y, and only a few terms are needed; see Lemma 3.3. It decreases rapidly as ρ is increased, as does the main term in the estimate for ε_2; see (3.20).

Related problems for harmonic functions can be treated; correction formulas for the single layer potential in the nearly singular case will be given elsewhere. Similar methods should apply to the Helmholtz equation since the leading singularities are the same. This approach in two dimensions [3] was applied to the Stokes equations of viscous fluid flow by Cortez [10], and 3D applications should be possible as well. Regularized kernels have long been used for fluid flow in vorticity formulation (see, e.g., [16, 6]), as well as other physical contexts, and discretization corrections of the present type have also been used in fluid problems [20, 22, 4].

In section 2 we analyze the smoothing error, derive the correction (1.22), and show that (1.12) has $O(\delta^5)$ error on the surface. We begin section 3 with a general lemma about the quadrature of nearly singular integrals with a homogeneous kernel. This is used to derive the correction (1.24), (1.25) and the estimates for the remaining discretization error. Sections 2 and 3 together prove Theorem 1.2 and the consistency part of Theorem 1.1. The rest of the proof of Theorem 1.1 is given in section 4. In section 5 we present numerical examples, with $\mathcal{S}$ a sphere or ellipsoid, illustrating the general principles.

2. The smoothing error. In this section we find the first correction to the error in the double layer potential, evaluated at a point near the surface, resulting from the smoothing in the simplest regularized Green's function

$$G_\delta(x) = -\frac{1}{4\pi}\frac{\operatorname{erf}(r/\delta)}{r} = G(x)\operatorname{erf}(r/\delta), \qquad r = |x|. \tag{2.1}$$

The correction is $O(\delta^2)$ and leaves an error of $O(\delta^3)$. We also derive the fifth order kernel (1.9), (1.12) for points on the surface.

The smoothing error for a point y near the surface is localized, and we can assume the function f is zero outside one coordinate patch. We can write the error as an integral in this patch, regarding f as a function of α,

$$\varepsilon = \int [\nabla G_\delta(x(\alpha) - y) - \nabla G(x(\alpha) - y)] \cdot n(\alpha) f(\alpha)\, dS(\alpha). \tag{2.2}$$

For simplicity, we will assume that $x(0) = 0$ and y is along the normal line from $x(0)$ so that $y = bn_0$ for some b, where n_0 is the unit normal at $x(0)$. Since we always subtract out the principal singularity, we also assume $f(0) = 0$. Now

$$\frac{\partial}{\partial r}(G_\delta - G) = \frac{1}{4\pi r^2}\,\phi(r/\delta), \tag{2.3}$$

$$\phi(\rho) = -\operatorname{erfc}(\rho) + \rho \operatorname{erfc}'(\rho) = -\operatorname{erfc}(\rho) - (2/\sqrt{\pi})\rho e^{-\rho^2}. \tag{2.4}$$

Thus, with $r = |x(\alpha) - y|$,

$$\varepsilon = \frac{1}{4\pi}\int \frac{(x(\alpha) - y)\cdot n(\alpha)}{r^3}\phi(r/\delta) f(\alpha)\, dS(\alpha). \tag{2.5}$$

We will find the largest contribution to ε using Taylor expansions at $\alpha = 0$. To simplify the calculations we first assume the α-coordinate system is specially chosen and then extend the result to a general system. With $T_j = \partial x/\partial \alpha_j$ the tangent vectors to surface, $j = 1, 2$, we assume that the metric tensor $g_{ij} = T_i \cdot T_j$ is the identity at $\alpha = 0$ and, furthermore, that $\partial g_{ij}/\partial \alpha_k = 0$ at $\alpha = 0$, $i, j, k = 1, 2$. The latter is equivalent to assuming that the Christoffel symbols vanish at $\alpha = 0$. We also assume, rotating if necessary, that T_1, T_2 have the directions of principal curvature at $\alpha = 0$. We then have simple expansions for $x(\alpha)$ and $n(\alpha)$:

$$x(\alpha) = T_1(0)\alpha_1 + T_2(0)\alpha_2 + \tfrac{1}{2}\kappa_1 n_0 \alpha_1^2 + \tfrac{1}{2}\kappa_2 n_0 \alpha_2^2 + O(|\alpha|^3), \tag{2.6}$$

$$n(\alpha) = n_0 - \kappa_1 T_1(0)\alpha_1 - \kappa_2 T_2(0)\alpha_2 + O(|\alpha|^2), \tag{2.7}$$

where κ_1, κ_2 are the principal curvatures. Thus

$$r^2 = |x - bn_0|^2 = |\alpha|^2 + b^2 - b\kappa_1\alpha_1^2 - b\kappa_2\alpha_2^2 + O(|\alpha|^4) + O(|\alpha|^3 b).$$

We make a further coordinate change $\alpha \mapsto \xi$ to simplify the dependence of r in the integral. We define ξ_1, ξ_2 by requiring $\xi_j/|\xi| = \alpha_j/|\alpha|$ and $|\xi|^2 + b^2 = r^2$, or

$$|\xi|^2 = |\alpha|^2 - b\kappa_1\alpha_1^2 - b\kappa_2\alpha_2^2 + O(|\alpha|^4) + O(|\alpha|^3 b) \tag{2.8}$$

so that

$$|\xi| = |\alpha|\left(1 - \frac{1}{2}b\kappa_1\frac{\alpha_1^2}{|\alpha|^2} - \frac{1}{2}b\kappa_2\frac{\alpha_2^2}{|\alpha|^2}\right) + O(|\alpha|^3) + O(b^3). \tag{2.9}$$

For α near 0, we can solve for $|\alpha|$ to get $|\alpha| = |\xi|(1 + bq/2) + O(|\xi|^3 + b^3)$ and then

$$\alpha_j = (\xi_j/|\xi|)|\alpha| = (1 + bq/2)\xi_j + O(|\xi|^3) + O(b^3), \tag{2.10}$$

$$q = \kappa_1 \xi_1^2/|\xi|^2 + \kappa_2 \xi_2^2/|\xi|^2. \tag{2.11}$$

We can now write the smoothing error in the form

$$\varepsilon = \frac{1}{4\pi} \int \frac{\phi(\sqrt{|\xi|^2 + b^2}/\delta)}{(|\xi|^2 + b^2)^{3/2}} w(\xi, b)\, d\xi \tag{2.12}$$

with the nonradial factors combined into

$$w(\xi, b) = [(x - y) \cdot n] f \left| \frac{\partial \alpha}{\partial \xi} \right| |T_1 \times T_2|. \tag{2.13}$$

Next we approximate each of these factors. First, using the expressions above for $x(\alpha)$ and $n(\alpha)$ we find, with $y = bn_0$,

$$(x(\alpha) - y) \cdot n(\alpha) = -b - \tfrac{1}{2}\kappa_j \alpha_j^2 + O(|\alpha|^3) + O(b^3),$$

summed over j, and, since $\alpha_j = \xi_j + O(b)$,

$$(x - y) \cdot n = -b - \tfrac{1}{2} q|\xi|^2 + O(|\xi|^3) + O(b^3). \tag{2.14}$$

Next, with α-derivatives of f at $\alpha = 0$ denoted by f_i or f_{ij}, we have

$$f = f_j \alpha_j + \tfrac{1}{2} f_{ij} \alpha_i \alpha_j + O(|\alpha|^3)\,,$$

since $f(0) = 0$, with sums over i and j, or

$$f = f_j \left(1 + \frac{bq}{2}\right) \xi_j + \frac{1}{2} f_{ij} \xi_i \xi_j + O(|\xi|^3) + O(b^3). \tag{2.15}$$

For the Jacobian, since q depends only on $\xi/|\xi|$, we find from (2.10) that

$$\det(\partial \alpha/\partial \xi) = 1 + bq + O(|\xi|^2) + O(b^2). \tag{2.16}$$

Finally, from (2.6) we have $T_j(\alpha) = T_j(0) + \kappa_j \alpha_j n_0$, and, since $T_j(0) \perp n(0)$,

$$|T_1 \times T_2| = 1 + O(|\alpha|^2) = 1 + O(|\xi|^2). \tag{2.17}$$

To approximate (2.12), we now substitute (2.14)–(2.17) into (2.13) and obtain

$$w(\xi, b) = -b\xi_j f_j - (3q/2) b^2 \xi_j f_j - (q/2)|\xi|^2 \xi_j f_j - (b/2) f_{ij} \xi_i \xi_j + R(\xi, b), \tag{2.18}$$

where $R = O(|\xi|^4 + b^4)$. The first three terms are odd in ξ and will contribute zero to the integral (2.12). We check that the remainder R is negligible: With the change of variables $\xi = \delta\zeta$, $b = \delta\lambda$, we can write $R(\xi, b) = \delta^4 \tilde{R}(\zeta, \lambda)$ for some bounded function $\tilde{R}$. Then the contribution to (2.12) from the remainder R in w is

$$(4\pi)^{-1} \delta^{-3+4+2} \int \frac{\phi(\sqrt{|\zeta|^2 + \lambda^2})}{(|\zeta|^2 + \lambda^2)^{3/2}} \tilde{R}(\zeta, \lambda)\, d\xi \;=\; O(\delta^3).$$

We are now left with only the fourth term in (2.18). Because of symmetry, only the terms with $i = j$ contribute, and the integral with ξ_j^2 reduces to a radial one. Again with $\xi = \delta\zeta$, $b = \delta\lambda$, and $s = |\zeta|$, (2.12) is now

$$\varepsilon = -\delta^2 \frac{\lambda}{8}(f_{11} + f_{22}) \int_0^\infty \frac{\phi(\sqrt{s^2+\lambda^2})}{(s^2+\lambda^2)^{3/2}} s^3\, ds \,+\, O(\delta^3).$$

To evaluate the integral, let

$$I(\lambda) = \int_0^\infty (G_1(r) - G(r)) s^3\, ds, \qquad r^2 = s^2 + \lambda^2;$$

then $dI/d\lambda$ is similar to the above:

$$I'(\lambda) = \frac{\lambda}{4\pi} \int_0^\infty \frac{\phi(r)}{4\pi r^3} s^3\, ds.$$

Changing the variable of integration to r, we have

$$I(\lambda) = \int_{|\lambda|}^\infty (G_1(r) - G(r)) s^2 r\, dr = (4\pi)^{-1} \int_{|\lambda|}^\infty \operatorname{erfc}(r)(r^2 - \lambda^2)\, dr,$$
$$I'(\lambda) = -(2\pi)^{-1}\lambda \int_{|\lambda|}^\infty \operatorname{erfc}(r)\, dr = (2\pi)^{-1}\lambda\,(|\lambda| \operatorname{erfc}|\lambda| - e^{-\lambda^2}/\sqrt{\pi}),$$

and, finally,

$$\varepsilon = -\delta^2(f_{11} + f_{22})(\lambda/4)(|\lambda| \operatorname{erfc}|\lambda| - e^{-\lambda^2}/\sqrt{\pi}) \,+\, O(\delta^3). \tag{2.19}$$

We have now obtained the smoothing correction with derivatives expressed in the special coordinates. To restate the answer in a general coordinate system, we have only to note the invariant form of the Laplacian on the surface, given by (1.21). It reduces to $f_{11} + f_{22}$ at $\alpha = 0$ in the special case above. Therefore it is also the correct expression in arbitrary coordinates. Thus $\varepsilon = -\mathcal{T}_1 + O(\delta^3)$, with $\mathcal{T}_1$ given by (1.22), justifying the stated correction and smoothing error.

The fifth order kernel. In view of the above analysis, we can identify the source of the $O(\delta^3)$ error in the special case when the point y is on the surface. We can then modify the choice of G_δ to remove this error, resulting in smoothing that is $O(\delta^5)$ accurate on the surface. With y on the surface, the smoothing error ε is given by (2.12) with $b = 0$. From (2.18) we see that the terms in the expansion of $w(\xi, 0)$ up to order 3 make no contribution. The fourth order terms will give an error proportional to

$$\int \phi(|\xi|/\delta)|\xi|^{-3}|\xi|^4\, d\xi \;=\; \delta^3 \int \phi(|\zeta|)|\zeta| d\zeta \;=\; 2\pi\delta^3 \int_0^\infty \phi(r) r^2\, dr.$$

The fifth order terms will lead to odd integrands, and thus the remaining error is $O(\delta^5)$. To eliminate the $O(\delta^3)$ error, we simply have to change the choice of G_δ, and therefore ϕ, so that the last integral is zero. Our new choice of G_δ will be so that $\phi(r)$ is replaced by $\phi_5(r) = \phi(r) + ar\phi'(r)$ with a a constant. It is easy to see from an integration by parts that the integral above is zero, with ϕ_5 in place of ϕ, provided we take $a = 1/3$. Thus $\phi_5(r) = \operatorname{erf}(r) - 1 - (2/\sqrt{\pi})(r - 2r^3/3)e^{-r^2}$, and the modification of ∇G is $\nabla G_\delta(x) = \nabla G(x)[1 + \phi_5(|x|/\delta)]$ as given by (1.9), (1.12).

3. The discretization error. We begin with a general principle, Lemma 3.1, for the quadrature of nearly singular integrals. We apply this to a local approximation of the double layer potential in Lemma 3.3, obtaining the corrections (1.24), (1.25). We then compare with the exact potential, verifying the discretization estimate of Theorem 1.2, and find an improved estimate needed for Theorem 1.1.

Our treatment of the discretization error is based on a lemma describing the quadrature error for the integral of a regularized homogeneous function K_δ, times a smooth function f, over a plane displaced from the origin. This lemma is similar to Lemma 3.1 of [3], except that we do not assume here that the smoothing radius δ is proportional to the grid size h; we allow $\rho = \delta/h$ unbounded as $h \to 0$. Let K be a homogeneous function of $(x, y) \in \mathbf{R}^d \times \mathbf{R}$ with degree m, that is,

$$K(ax, ay) = a^m K(x, y)\,, \qquad a > 0\,, \quad (x, y) \neq 0. \tag{3.1}$$

We choose a regularization of the form

$$K_\delta(x, y) = K(x, y)s(x/\delta, y/\delta), \qquad (x, y) \in \mathbf{R}^d \times \mathbf{R}, \tag{3.2}$$

where s is a specified shape function such that $s \to 1$ rapidly at infinity. It follows that $K_\delta(x, y) = \delta^m K_1(x/\delta, y/\delta)$. The lemma concerns the integral of $K_\delta f$ over x with y fixed. We allow the singularity to be misaligned from the grid points in x-space.

LEMMA 3.1. *Let K_δ be a smooth function on $\mathbf{R}^{d+1}$ with the form* (3.2) *such that K and s are smooth for $(x, y) \neq 0$; K is homogeneous of degree m with $-d \leq m \leq 0$; $s(x, y) \to 1$ as $(x, y) \to \infty$; and $|D^\beta s(x, y)| \leq C_\beta |(x, y)|^{-|\beta|}$ for $|(x, y)| \geq 1$ and for each multi-index β. Let f be a smooth function on $\mathbf{R}^d$ such that f and its derivatives are rapidly decreasing. Suppose, for each $h > 0$, that δ is chosen so that $\rho = \delta/h \geq \rho_0$ for some fixed $\rho_0 > 0$. Assume $\nu \in \mathbf{R}^d$ with $0 \leq \nu_j \leq 1$ for $1 \leq j \leq d$. Let $\eta = y/h$. Then the error in replacing the integral I by the sum S,*

$$I = \int_{\mathbf{R}^d} K_\delta(x, y) f(x)\, dx = \int_{\mathbf{R}^d} K(x, y) s(x/\delta, y/\delta) f(x)\, dx, \tag{3.3}$$

$$S = \sum_{n \in \mathbf{Z}^d} K_\delta(nh - \nu h, y) f(nh - \nu h)\, h^d, \tag{3.4}$$

has the form

$$S - I = h^{d+m} \left(c_0 f(0) + C_1 h + C_2 h^2 + \cdots + C_\ell h^\ell \right) + O(\delta^{d+m+\ell+1}), \tag{3.5}$$

$$c_0 = (2\pi)^{d/2} \sum_{0 \neq n \in \mathbf{Z}^d} e^{-2\pi i n \nu} \hat{K}_\rho(2\pi n, \eta). \tag{3.6}$$

Here $\hat{K}_\rho(\cdot, \eta)$ is the Fourier transform of $K_\rho(\cdot, \eta)$, and ℓ depends on the smoothness of K, s, f. The leading constant c_0 is uniformly bounded in η, ρ, ν. The C_j, for $j \geq 1$, and the error term are uniformly bounded in η, ρ, ν, depending on f.

The value of c_0 is derived from the Poisson summation formula applied to K_ρ. The sum (3.6) converges rapidly. We write the Fourier transform as

$$\hat{F}(k) \,=\, (2\pi)^{-d/2} \int F(x) e^{-ikx}\, dx.$$

The proof is a modification of that for Lemma 3.1 in [3] and related to ones in [4, 12].

When we apply this lemma to the double layer potential, evaluated at a point y near the surface, the leading error comes from the part of the surface close to y,

and we begin with a simplified approximation for the potential. Suppose $y = bn_0$, where n_0 is the unit normal to a point $x(\alpha_0)$ on the surface. For convenience, we take $\alpha_0 = 0$ and $x(\alpha_0) = 0$. With T_j, g_{ij}, g^{ij} as defined earlier, at $\alpha = 0$, let $\tau = |T_1 \times T_2|$ so that $\det g_{ij} = \tau^2$. Also let $J = \partial x(0)/\partial\alpha$ so that $J\alpha = T_1\alpha_1 + T_2\alpha_2$. For α near 0, we can approximate $x(\alpha)$ by its projection $J\alpha$ in the tangent plane and the double layer kernel $n(\alpha) \cdot \nabla G_\delta(x(\alpha), y)$ by

$$K_\delta^{(0)}(\alpha, b) = n_0 \cdot \nabla G_\delta(J\alpha - bn_0) = -(\partial/\partial b)G_\delta(J\alpha - bn_0). \tag{3.7}$$

The surface area $dS(\alpha)$ is $\tau\, d\alpha$ to first approximation. We subtract out the singularity in the double layer potential, and the leading contribution to $f(\alpha) - f(0)$ will be $(\alpha_1\partial_1 f + \alpha_2\partial_2 f)\zeta(\alpha)$ with $\partial_r f = \partial_r f(0)$ for some cut-off function ζ, $\zeta = 1$ near $\alpha = 0$. Thus we actually apply the lemma to the kernel

$$K_\delta^{(r)}(\alpha, b) = K_\delta^{(0)}(\alpha, b)\alpha_r, \qquad r = 1, 2. \tag{3.8}$$

We first show that this simplified case gives the $O(h)$ error stated in (1.24), (1.25).

LEMMA 3.2. *With the notation above, let*

$$I = \int_{R^2} n_0 \cdot \nabla G_\delta(J\alpha - bn_0)(\alpha_1\partial_1 f + \alpha_2\partial_2 f)\zeta(\alpha)\tau\, d\alpha \tag{3.9}$$

and S be the corresponding sum with $\alpha = jh - \nu h$. Then $S - I = (c_1\partial_1 f + c_2\partial_2 f)h + O(h^p)$ for large p, where c_1, c_2 are given by (1.25).

Proof. Our main task is to find the Fourier transform of $K_\rho^{(r)}(\alpha, b)$ in α alone. We begin with the 3D Fourier transform of $G_\rho(x) = -\operatorname{erf}(|x|/\rho)/4\pi|x|$,

$$\hat{G_\rho}(k) = -(2\pi)^{-3/2}|k|^{-2}e^{-\rho^2|k|^2/4}, \qquad k = (k_1, k_2, k_3) \in \mathbf{R}^3. \tag{3.10}$$

We interpret $G_\rho(J\alpha - bn_0)$ as a composition: Since G_ρ is radial, it depends on α only through $|J\alpha|^2 = |B\alpha|^2$, where $B = (J^*J)^{1/2}$; note $|B\alpha|^2 = J^*J\alpha \cdot \alpha = \sum_{ij} g_{ij}\alpha_i\alpha_j$. Thus $G_\rho(J\alpha - bn_0) = (G_\rho \circ M)(\alpha, b)$, where $M(\alpha, b) = (B\alpha, b)$. The 3D transform of $G_\rho \circ M$, as a function of (α, b), is then

$$(G_\rho \circ M)\hat{}(k) = |\det M|^{-1}\hat{G_\rho}((M^*)^{-1}k) = \tau^{-1}\hat{G_\rho}(B^{-1}(k_1, k_2), k_3), \tag{3.11}$$

and, since $\hat{G_\rho}$ is radial and $B^{-2} = (J^*J)^{-1}$,

$$(G_\rho \circ M)\hat{}(k) = \tau^{-1}\hat{G_\rho}(\ell, k_3), \qquad \ell^2 = \sum_{i,j=1}^{2} g^{ij}k_ik_j. \tag{3.12}$$

Now

$$\hat{K}_\rho^{(0)}(k) = (-(\partial/\partial b)G_\rho \circ M)\hat{}(k) = -ik_3\,\tau^{-1}\hat{G_\rho}(\ell, k_3), \tag{3.13}$$

and the transform of $K_\rho^{(0)}(\alpha, b)$ in α alone is

$$\begin{aligned} K_\rho^{(0)}(\cdot, b)\hat{}(k_1, k_2) &= (2\pi)^{-1/2}\int_{-\infty}^{\infty} \hat{K}_\rho^{(0)}(k_1, k_2, k_3)e^{ik_3b}\, dk_3 \\ &= (2\pi)^{-2}\tau^{-1}\frac{\partial}{\partial b}\int_{-\infty}^{\infty}\frac{1}{\ell^2 + k_3^2}e^{-\rho^2(\ell^2+k_3^2)/4}e^{ik_3b}\, dk_3. \end{aligned} \tag{3.14}$$

This is similar to (3.29) in [3]; we find

$$K_\rho^{(0)}(\cdot,b)\hat{}(k_1,k_2) = (8\pi\tau)^{-1}\left[e^{\ell b}\operatorname{erfc}(b/\rho+\ell\rho/2) - e^{-\ell b}\operatorname{erfc}(-b/\rho+\ell\rho/2)\right]. \tag{3.15}$$

Next, since $K_\rho^{(r)} = K_\rho^{(0)}\alpha_r$, $\hat{K}_\rho^{(r)}(k_1,k_2,b)$ is given by

$$\hat{K}_\rho^{(r)} = i(\partial/\partial k_r)\hat{K}_\rho^{(0)} = i(\partial\ell/\partial k_r)(\partial/\partial\ell)\hat{K}_\rho^{(0)} = i(g^{rs}k_s/\ell)(\partial/\partial\ell)\hat{K}_\rho^{(0)}, \tag{3.16}$$

summed over s. After differentiating and canceling we get

$$\hat{K}_\rho^{(r)}(k_1,k_2,b) = \frac{ib}{8\pi\tau}\sum_{s=1}^{2}\frac{g^{rs}k_s}{\ell}E\left(\frac{b}{\rho},\frac{\ell\rho}{2}\right) \tag{3.17}$$

with $E(p,q)$ given by (1.23). We are now ready to apply Lemma 3.1 to the simplified integral (3.9). The kernel $K_\rho^{(r)}(\alpha,b)$ is homogeneous with degree $m=-1$, and, according to the lemma, $S-I = \varepsilon h + O(h^p)$, with

$$\varepsilon = 2\pi\tau\sum_{r=1}^{2}\sum_{n\neq 0}e^{-2\pi i n\nu}(\partial_r f)\hat{K}_\rho^{(r)}(2\pi n, b/h),$$

and, using (3.17), with $\lambda = b/\delta$ and $\|n\| = \sqrt{g^{ij}n_in_j}$,

$$\varepsilon = \frac{i\rho\lambda}{4}\sum_{r,s=1}^{2}\sum_{n\neq 0}e^{-2\pi i n\nu}(\partial_r f)\frac{g^{rs}n_s}{\|n\|}E(\lambda,\pi\rho\|n\|).$$

Combining terms for n and $-n$, we find $\varepsilon = c_1\partial_1 f + c_2\partial_2 f$, with c_1, c_2 expressed as in (1.25), and the lemma is proved. ∎

To understand the dependence of the errors on $\rho = \delta/h$, we will use the following lemma concerning the size of the function E of (1.23).

LEMMA 3.3. *With E as in* (1.23)*, we have for $q \geq q_0 > 0$ and any p,*

$$E(p,q) \leq C_0\exp(-q_0 p - q^2), \tag{3.18}$$

where C_0 depends on q_0.

Proof. Since E is even in p we assume $p \geq 0$. Call the two terms T_1 and T_2. Since $\operatorname{erfc} a \leq C\exp(-a^2)$ for $a \geq 0$, $T_1 \leq C\exp(2pq-(p+q)^2) = C\exp(-p^2-q^2)$. If $q \geq p$, then $T_2 \leq C\exp(-p^2-q^2)$ in the same way. If $q \leq p$, then $T_2 \leq 2\exp(-2pq) = 2\exp(-pq)\exp(-pq) \leq 2\exp(-pq)\exp(-q^2)$. Thus in any case $E \leq \left[C\exp(-p^2) + 2\exp(-q_0 p)\right]\exp(-q^2)$, and the result follows from this. ∎

Suppose now that we assume a lower bound for the matrix g^{ij},

$$\|k\|^2 \equiv \sum_{ij} g^{ij}k_ik_j \geq \gamma^2|k|^2, \tag{3.19}$$

for some $\gamma > 0$. Then applying (3.18) to ε above, we have $\rho|\lambda|E(\lambda,\pi\rho\|n\|) \leq C\rho\exp(-\pi^2\rho^2\gamma^2|n|^2)$, and, after summing,

$$|c_r| \leq C\rho\exp(-\pi^2\rho^2\gamma^2), \qquad r = 1,2. \tag{3.20}$$

We can now complete the discretization error estimate in Theorem 1.2. With $\alpha = 0$, $x(0) = 0$ as before, and assuming $f(0) = 0$ for simplicity, we compare

$$I = \int n(\alpha) \cdot \nabla G_\delta(x(\alpha) - y) f(\alpha) dS(\alpha), \tag{3.21}$$

$$S = \sum_j n(\alpha_j) \cdot \nabla G_\delta(x(\alpha_j) - y) f(\alpha_j) A_j \, h^2, \tag{3.22}$$

where $\alpha_j = jh - \nu h$ and $A_j \, d\alpha = dS(\alpha_j)$. We will show that, assuming (3.19),

$$S - I = (c_1 \partial_1 f + c_2 \partial_2 f)h + O(h^2 \rho^3 \exp(-\pi^2 \gamma^2 \rho^2)) + O(h^3) \tag{3.23}$$

with errors uniform in b, ρ, ν. The estimate in (1.27) follows, with $c_0 = \pi^2 \gamma^2 / 2$. We need to show that the error from replacing (3.21) by the simplified version (3.9) is bounded by the remainder in (3.23). This is the quadrature error for the integral of

$$n(\alpha) \nabla G_\delta(x(\alpha) - bn_0) f(\alpha) A(\alpha) - n_0 \nabla G_\delta(J\alpha - bn_0)[\alpha_1 \partial_1 f(0) + \alpha_2 \partial_2 f(0)] A(0). \tag{3.24}$$

We can suppose f is zero outside a neighborhood of $\alpha = 0$, since the outer part is smooth and gives a high order error. We can add and subtract to write (3.24) as a sum of terms, each a regularized homogeneous function in (α, b) of degree 0, times a smooth function, as in Lemma 3.1, plus a higher degree remainder. For the remainder we can show the quadrature error is $O(h^3)$ as in the end of the proof of Lemma 3.1 in [3]. The degree 0 terms are similar to the main term treated in Lemma 3.2; they involve ∇G_δ or $D^2 G_\delta$ and up to two more factors of α. By Lemma 3.1, the $O(h^2)$ errors resulting from these terms are given by expressions like those in Lemma 3.2. The Fourier transform of the kernel is very similar to that case, with a second α-derivative leading to a factor of k in the transform. An α-factor corresponds to a k-derivative, leading to an extra factor of ρ in the estimate like (3.20). Otherwise the $O(h^2)$ correction term is bounded as in (3.20), and the estimate (3.23) results.

Finally, we discuss the special case with y on the surface to obtain an improved estimate needed for the discretization error in (1.17) of Theorem 1.1. In this case $b = 0$, $\lambda = 0$, and the $O(h)$ correction term in (3.23) is zero. The expansion in the proof of (3.23) can be continued further, with successive regularized homogeneous terms of degree ≤ 2 and an $O(h^5)$ remainder. In this way we find, for y on the surface, $S - I = O(h^2 \rho^k \exp(-\pi^2 \gamma^2 \rho^2)) + O(h^5)$ for some integer k. The fifth order kernel used in solving the integral equation has an added term; it has Gaussian form, and the last estimate applies as well with this new term. Finally, since $\rho^k \exp(-\pi^2 \gamma^2 \rho^2 / 2)$ is bounded we can write

$$S - I = O(h^2 \exp(-\pi^2 \gamma^2 \rho^2 / 2)) + O(h^5). \tag{3.25}$$

This estimate is used in section 4 to prove (1.17) with $c_0 = \pi^2 \gamma^2 / 2$.

4. The integral equation. In this section we prove the assertions about the numerical solution of the integral equation in Theorem 1.1. The exact integral equation, for unknown f on the surface $\mathcal{S}$, with g prescribed on $\mathcal{S}$, has the form

$$\tfrac{1}{2} f + Tf = g, \qquad (Tf)(x) = \int_{\mathcal{S}} K(x, x') f(x') \, dS(x'), \tag{4.1}$$

where T is the double layer potential, with kernel $K(x,x') = n(x') \cdot \nabla G(x'-x)$. We can use the coordinate patches to write the integral operator as

$$Tf(x) = \sum_\sigma \int K(x, X^\sigma(\alpha)) \psi^\sigma(X^\sigma(\alpha)) f(X^\sigma(\alpha)) A^\sigma(\alpha) d\alpha. \tag{4.2}$$

Subtracting $f(x)$ inside the integral, we convert (4.1) to the equivalent form (1.8). We are concerned with the discrete version (1.13), (1.14) of (1.8).

We can regard T as a bounded operator on $L^p(\mathcal{S})$ for any p with $2 \le p < \infty$. From potential theory we know that $\frac{1}{2} + T$ has kernel $\{0\}$ in $C(\mathcal{S})$; the same is true in $L^2(\mathcal{S})$ (see, e.g., [11, Prop. 3.13]) and then also in $L^p(\mathcal{S})$, $p \ge 2$. Thus, from Fredholm theory, $\frac{1}{2} + T$ has a bounded inverse on $L^p(\mathcal{S})$ for $2 \le p < \infty$, and (4.1) has a unique solution $f \in L^p$ for each $g \in L^p$, with $\|f\|_p \le C\|g\|_p$. We will argue that the discrete problem can be solved by regarding it as a perturbation of the exact problem, making a sequence of steps from (4.1) to (1.13). This is the stability part of the proof; the consistency part has been done in sections 2 and 3. We always assume (1.15). When we say a discrete operator is bounded, we mean bounded uniformly with respect to h.

4.1. The smoothing. The regularized kernel K_δ has the form $K_\delta(x,x') = K(x,x')s(|x-x'|/\delta)$, where $s(x) \to 1$, $Ds(x) \to 0$ rapidly as $x \to \infty$. Using the smoothness of K_δ and s, we note the pointwise estimates

$$|K_\delta(x,x')| \le Cr^{-1}, \quad r \ge \delta; \qquad |K_\delta(x,x')| \le C\delta^{-1}, \quad r \le \delta; \tag{4.3}$$

$$|DK_\delta(x,x')| \le Cr^{-2}, \quad r \ge \delta; \qquad |DK_\delta(x,x')| \le C\delta^{-2}, \quad r \le \delta, \tag{4.4}$$

where $r = |x-x'|$ and D denotes an α-derivative, with either $x = X^\sigma(\alpha)$ or $x' = X^\sigma(\alpha)$. We assume here that x, x' are in a bounded set such as $\mathcal{S}$.

To see the effect of the smoothing, we can estimate, as in section 2,

$$\int_S |K_\delta(x,x') - K(x,x')|\, dS \le C_1 \int_0^\infty r^{-1} |s(r/\delta) - 1|\, r\, dr = C_2 \delta$$

uniformly for $x \in S$; the same holds with x, x' reversed. In general, for any integral operator $\mathcal{A}$ on $\mathcal{S}$ with kernel $\mathcal{K}$, if we have uniform estimates

$$\int_{\mathcal{S}} |\mathcal{K}(x,x')|\, dS(x') \le M, \qquad \int_{\mathcal{S}} |\mathcal{K}(x,x')|\, dS(x) \le M, \tag{4.5}$$

it follows that $\mathcal{A}$ is a bounded operator on $L^p(\mathcal{S})$ with $\|\mathcal{A}\| \le M$ (see, e.g., [11, Prop. 0.10]). In our case we conclude that the operator on $L^p(\mathcal{S})$ with kernel $K_\delta - K$ has norm $O(\delta)$. Thus, since the operator $\frac{1}{2} + T$ is invertible, the same is true for $\frac{1}{2} + T_\delta$, for δ small enough, where T_δ is the operator as in (4.1) with K_δ in place of K.

4.2. A discretized kernel. Next we modify the kernel K_δ so that it acts in a natural way on discrete functions. We work with grid functions f_i^σ, defined at each grid point $x_i^\sigma = X^\sigma(ih)$, $ih \in V_\sigma$, for each σ. If $x_i^\sigma = x_j^\tau$ for some i, j with $\sigma \ne \tau$, we require $f_i^\sigma = f_j^\tau$; this property is preserved by the discrete operators. Let Q_i be the grid square in the α-plane $Q_i = \{\alpha = (\alpha_1, \alpha_2) : -h/2 < \alpha_\nu - i_\nu h \le h/2, \nu = 1,2\}$, and let $\chi_i^\sigma(x) = 1$ for $x = X^\sigma(\alpha)$, with $\alpha \in Q_i$, and $\chi_i^\sigma(x) = 0$ otherwise; i.e., χ_i^σ is the characteristic function of the set $X^\sigma(Q_i)$. Given any grid function f_i^σ, we can associate a function Bf on $\mathcal{S}$ defined by

$$(Bf)(x) = \sum_{\sigma,i} \psi_i^\sigma \chi_i^\sigma(x) f_i^\sigma. \tag{4.6}$$

The part of Bf from one σ is piecewise constant, but Bf is not, since the parts overlap. Recalling (1.14), we define a modified kernel $\tilde{K}$ and operator $\tilde{T}$ on $L^p(\mathcal{S})$ as

$$\tilde{K}(x,x') = \sum_{i,\sigma,j,\tau} K_{ij}^{\sigma\tau}\psi_i^\sigma\psi_j^\tau\chi_i^\sigma(x)\chi_j^\tau(x')A_j^\tau \;=\; B\Big(\sum_{j,\tau} K_{ij}^{\sigma\tau}\psi_j^\tau\chi_j^\tau(x')A_j^\tau\Big), \tag{4.7}$$

$$\tilde{T}f(x) = \sum_\tau \int \tilde{K}(x, X^\tau(\alpha))f(X^\tau(\alpha))\,d\alpha. \tag{4.8}$$

We write $K_\delta(x,x')$ as $\sum_{\sigma,\tau} K_\delta(x,x')\psi^\sigma(x)\psi^\tau(x')$; for each σ,τ,i,j we have replaced $K^{\sigma\tau} \equiv K_\delta\psi^\sigma\psi^\tau A^\tau$ on $X^\sigma(Q_i)\times X^\tau(Q_j)$ with the constant $\tilde{K}^{\sigma\tau} \equiv K_\delta(x_i^\sigma, x_j^\tau)\psi_i^\sigma\psi_j^\tau A_j^\tau$. Then for each σ,τ and each $x \in X^\sigma(U^\sigma)$ we estimate

$$\int_{U_\tau} |\tilde{K}^{\sigma\tau}(x, X^\tau(\alpha)) - K^{\sigma\tau}(x, X^\tau(\alpha))|\,d\alpha \;\le\; Ch\sum_j M_{ij}^{\sigma\tau}\,h^2, \tag{4.9}$$

where i is such that $x \in X^\sigma(Q_i)$. Here $M_{ij}^{\sigma\tau}$ is a bound for $|DK^{\sigma\tau}|$ within radius $O(h)$ of (x_i^σ, x_j^τ). Using (4.4) we have $|M_{ij}^{\sigma\tau}| \le C|x_i^\sigma - x_j^\tau|^{-2}$ for $|x_i^\sigma - x_j^\tau| > C_0\delta$ and $|M_{ij}^{\sigma\tau}| \le C\delta^{-2}$ otherwise. Since $c_1|\alpha-\alpha'| \le |X^\tau(\alpha) - X^\tau(\alpha')| \le c_2|\alpha-\alpha'|$, we can bound (4.9) as follows, replacing the sum for $|x_i^\sigma - x_j^\tau| \ge O(\delta)$ by an α-integral:

$$C_1h\delta^{-2}(\delta/h)^2h^2 + C_2h\int_{C_3\delta}^{C_4} r^{-2}r\,dr \;\le\; C_5h(1+|\log\delta|) \;\le\; C_6h(1+|\log h|) \tag{4.10}$$

(cf. [5, Lemma 3.2] or [16, Lemma 5] for arguments of this type). This gives the first of two estimates of form (4.5), with $\mathcal{K} = \tilde{K} - K_\delta$, and the second is similar. Thus we can conclude that the norm of $\tilde{T} - T$, as an operator on $L^p(\mathcal{S})$, is $O(h|\log h|)$. Since the operator $\frac{1}{2} + T_\delta$ is invertible, the nearby operator $\frac{1}{2} + \tilde{T}$ is also invertible.

4.3. Discrete functions. Next we pass from functions of continuous $x \in \mathcal{S}$ to discrete functions. For a grid function f_i^σ we use the L_h^p norm, defined by

$$\|f\|_{L_h^p}^p = \sum_{\sigma,i} |f_i^\sigma|^p\,h^2. \tag{4.11}$$

The operator B of (4.6) is bounded from L_h^p to L^p. Now, given a grid function $g \in L_h^p$, we have $Bg \in L^p(\mathcal{S})$, and from the result above there is a unique $F \in L^p(\mathcal{S})$ depending boundedly on $Bg \in L^p(S)$ so that $\frac{1}{2}F + \tilde{T}F = Bg$, or, in view of (4.7), $F = -2Bw + 2Bg$, where

$$w_i^\sigma = \sum_{j,\tau} K_{ij}^{\sigma\tau}\psi_j^\tau A_j^\tau \int_{Q_j} F(X^\tau(\alpha))\,d\alpha. \tag{4.12}$$

Now we can define a discrete function f so that $F = Bf$:

$$f_i^\sigma = -2w_i^\sigma + 2g_i^\sigma. \tag{4.13}$$

This can be rewritten as

$$\tfrac{1}{2}f_i^\sigma + (T_h(\mathcal{M}f))_i^\sigma = g_i^\sigma, \qquad (\mathcal{M}f)_j^\tau = h^{-2}\int_{Q_j} F(X^\tau(\alpha))\,d\alpha, \qquad F = Bf, \tag{4.14}$$

where T_h is the discrete integral operator from L^p_h to L^p_h defined by

$$(T_h \bar f)^\sigma_i = \sum_{j,\tau} K^{\sigma\tau}_{ij} \psi^\tau_j \bar f^\tau_j A^\tau_j \, h^2. \tag{4.15}$$

Equation (4.14) resembles the desired discrete integral equation (1.13) but has $(\mathcal{M}f)^\tau_j$ rather than f^τ_j inside the integral operator. It is easy to check that the mapping $F \mapsto \mathcal{M}f$ is bounded from L^p to L^p_h, and thus $f \mapsto Bf = F \mapsto \mathcal{M}f$ is bounded from L^p_h to L^p_h. We can see that the discrete operator T_h is bounded on L^p_h using the pointwise estimate (4.3) and an argument similar to that for $\tilde K - K$ in (4.9), (4.10). We can now check that the discrete solution f of (4.14) depends boundedly on g: Given $g \in L^p_h$, $F \in L^p$, as determined by the equation $(\frac12 + \tilde T)F = Bg$, depends boundedly on $Bg \in L^p$ and therefore on $g \in L^p_h$. Then, using the boundedness of T_h, the discrete function f, defined by (4.12), (4.13), depends boundedly in L^p_h on g. The solution f of (4.14) is unique since the corresponding solution F of $(\frac12 + \tilde T)F = Bg$ is unique. We have shown that the operator $\mathcal{A}$ has a bounded inverse on L^p_h, where

$$\mathcal{A}f \;=\; \tfrac12 f + T_h \mathcal{M} f. \tag{4.16}$$

4.4. Boundedness in the Hölder norm. We will now improve the solvability estimate for $\mathcal{A}$ to a higher norm in order to measure the agreement of values on overlapping grids. For $0 < \lambda < 1$, we define the discrete Hölder norm as

$$\|f\|_{C^\lambda_h} = \sup_{\sigma,i} |f^\sigma_i| + \sup_{\sigma,i,\sigma',i'} |f^\sigma_i - f^{\sigma'}_{i'}| / |x^\sigma_i - x^{\sigma'}_{i'}|^\lambda, \tag{4.17}$$

excluding pairs with $x^\sigma_i = x^{\sigma'}_{i'}$ in the latter sup. Similarly, C^0_h will be the space with the supremum norm. We first consider the operator $\mathcal{A}$ on C^0_h. From (4.3) we see that, for $q < 2$, $\sum_j |K^{\sigma\tau}_{ij}|^q h^2 \le C$. Then, from Hölder's inequality, T_h is bounded from L^p_h to C^0_h, provided $p > 2$. Now if $\frac12 f + T_h \mathcal{M} f = g$, we have $\|T_h \mathcal{M} f\|_{C^0_h} \le C_1 \|g\|_{L^p_h} \le C_2 \|g\|_{C^0_h}$, and so $\|f\|_{C^0_h} \le C_3 \|g\|_{C^0_h}$. Thus $\mathcal{A}$ has a bounded inverse on C^0_h.

Next we check that T_h is bounded from C^0_h to C^λ_h. For arbitrary grid points x^σ_i and $x^{\sigma'}_{i'}$, let $d = |x^\sigma_i - x^{\sigma'}_{i'}|$. We need to verify that (with τ fixed)

$$d^{-\lambda} \sum_j \left[K^{\sigma\tau}_{ij} - K^{\sigma'\tau}_{i'j} \right] \psi^\tau_j f^\tau_j A^\tau_j \, h^2 \tag{4.18}$$

is bounded, assuming f bounded in C^h_0. We use the estimates (4.3), (4.4) to bound

$$K^{\sigma\tau}_{ij} - K^{\sigma'\tau}_{i'j} = K_\delta(x^\sigma_i, x^\tau_j) - K_\delta(x^{\sigma'}_{i'}, x^\tau_j). \tag{4.19}$$

As in classical arguments, we split the sum into two parts. For j in the set J_1 such that $|x^\sigma_i - x^\tau_j| \le 2d$, we apply (4.3) to the two terms separately. If $d < \delta$ we have

$$\sum_{j \in J_1} |K^{\sigma\tau}_{ij}| h^2 \le C_1 \delta^{-1} (d/h)^2 h^2 \le C_2 d,$$

and similarly for $K^{\sigma'\tau}_{i'j}$, whereas, for larger d, the sum is bounded by

$$C_3 \delta^{-1} (\delta/h)^2 h^2 + C_4 \int_{C_5\delta}^{C_6 d} r^{-1} r \, dr \le C_7 \delta + C_8 d \le C_9 d.$$

For the remaining sum over the set J_2 with $|x_i^\sigma - x_j^\tau| > 2d$, we bound (4.19) by d times $DK(z, x_j^\tau)$, with z along the line from x_i^σ to $x_{i'}^{\sigma'}$. With $r = |x_j^\tau - x_i^\sigma|$, we have $|x_j^\tau - z| \geq r - d \geq r/2$. Thus (4.19) is bounded by Cdr^{-2} for $r > \delta$ or $Cd\delta^{-2}$ otherwise, using (4.4). Now if $d \geq \delta$ we get

$$\sum_{j \in J_2} |K_{ij}^{\sigma\tau} - K_{i'j}^{\sigma'\tau}| h^2 \;\leq\; C_1 d \int_{C_2 d}^{C_3} r^{-2} r \, dr \;\leq\; C_4 d(1 + |\log d|),$$

and otherwise we have this plus an additional term $C_5 d\delta^{-2}(\delta/h)^2 h^2 \leq C_6 d$. We conclude that the Hölder quotient (4.18) is bounded by

$$C_1 d^{1-\lambda}(|\log d| + 1)\|f\|_{C_h^0} \leq C_2 \|f\|_{C_h^0}, \tag{4.20}$$

and, consequently, for any $f \in C_h^0$, $\|T_h f\|_{C_h^\lambda} \leq C_3 \|f\|_{C_h^0}$ as claimed.

We have already shown that the operator $\mathcal{A}$ has a bounded inverse on C_h^0; we can now show the same is true for C_h^λ. If $\frac{1}{2} f + T_h \mathcal{M} f = g$ with $g \in C_h^\lambda$, then $\|\mathcal{M} f\|_{C_h^0} \leq C_1 \|f\|_{C_h^0} \leq C_2 \|g\|_{C_h^0}$, and thus $\|T_h \mathcal{M} f\|_{C_h^\lambda} \leq C_3 \|g\|_{C_h^0} \leq C_3 \|g\|_{C_h^\lambda}$. Since $f = 2g - 2T_h \mathcal{M} f$, it follows that $\|f\|_{C_h^\lambda} \leq C \|g\|_{C_h^\lambda}$.

4.5. Correcting the discrete equation. Next, in order to compare the operator $T_h \mathcal{M}$ with T_h, we show that $\mathcal{M} f$, as defined in (4.14), is close to f in C_h^0, relative to the C_h^λ-norm of f, for small h. Given a grid function f and $x \in \mathcal{S}$, suppose $x \in X^\tau(Q_j)$ for some τ, j so that $|x - x_j^\tau| \leq Ch$. If also $x \in X^\sigma(Q_i)$, then $|x - x_i^\sigma| \leq Ch$ so that $|x_i^\sigma - x_j^\tau| \leq 2Ch$. Thus $|f_i^\sigma - f_j^\tau| \leq C_1 h^\lambda \|f\|_{C_h^\lambda}$. Similarly, $\psi_i^\sigma - \psi^\sigma(x) = O(h)$. Applying this to the definition of Bf, we find $|(Bf)(x) - f_j^\tau| \leq C_2 h^\lambda \|f\|_{C_h^\lambda}$ uniformly for $x \in X^\tau(Q_j)$. Then since $(\mathcal{M} f)_j^\tau$ is an average of Bf over this set,

$$|(\mathcal{M} f)_j^\tau - f_j^\tau| \leq C_2 h^\lambda \|f\|_{C_h^\lambda}. \tag{4.21}$$

The last result has an important consequence: Since $\mathcal{M}$ is close to the identity as an operator from C_h^λ to C_h^0, and since T_h is bounded from C_h^0 to C_h^λ, the operators $f \mapsto \frac{1}{2} f + T_h \mathcal{M} f$ and $f \mapsto \frac{1}{2} f + T_h f$, acting on C_h^λ, are close when h is small. Since the first has a bounded inverse on C_h^λ, the second does as well. That is, we have a bounded solution operator on C_h^λ for the equation $\frac{1}{2} f + T_h f = g$. This equation is the analogue of (1.13) without the subtracted term.

4.6. The subtraction. Now let $z_i^\sigma = T_h \cdot 1$. The sum z_i^σ approximates an integral which, according to (1.7), is identically $\frac{1}{2}$. We check that $\|z_i^\sigma - \frac{1}{2}\|_{C_h^\lambda} \to 0$ as $h, \delta \to 0$: The estimates of sections 2 and 3 show that $|z_i^\sigma - \frac{1}{2}| \leq C(h + \delta)$ uniformly in i, σ. As for the Hölder quotient, we see from (4.20) that it is small for d small, say, $d \leq d_0$, with d_0 independent of h. But for $d \geq d_0$ it is bounded by $Ch\, d_0^{-\lambda}$, which is small when h is small enough. As a consequence, the operators $f \mapsto \frac{1}{2} f + T_h f$ and $f \mapsto f + T_h f - zf$ on C_h^λ are close for h small. As before, we conclude that the new operator has a bounded inverse. This is exactly the statement that the discrete integral equation (1.13) is solvable, with solution bounded in C_h^λ, that is, $\|f\|_{C_h^\lambda} \leq C \|g\|_{C_h^\lambda}$.

We wish to show that the solution of (1.13) is also bounded on C_h^0, since this norm is more convenient for comparing with the exact solution. We can write the equation as $(1 - z) f + T_h f = g$ or $f + \zeta T_h f = \zeta g$, where $\zeta = 1/(1 - z)$. Since $z \approx \frac{1}{2}$ in C_h^λ, $\zeta \approx 2$ in C_h^λ. Rearranging the terms, we have $(1 - z)(f - \zeta g) + T_h(f - \zeta g) = -T_h(\zeta g)$.

Now, using our last result, with $f-\zeta g$ in place of f, and the bound for $T_h : C_h^0 \to C_h^\lambda$, we find $\|f-\zeta g\|_{C_h^\lambda} \le C_1\|T_h(\zeta g)\|_{C_h^\lambda} \le C_2\|g\|_{C_h^0}$ from which it follows that

$$\|f\|_{C_h^0} \le C\|g\|_{C_h^0}, \qquad f = ((1-z)I + T_h)^{-1}g. \tag{4.22}$$

4.7. The error estimate. Having established the bounded solvability (4.22) for the discrete integral equation (1.13), we can easily estimate the error when the exact solution is smooth. Suppose f, g are smooth functions on $\mathcal{S}$ satisfying (4.1). (If g is smooth, it follows that f is smooth.) Let g_i^σ be the grid function taking values $g_i^\sigma = g(x_i^\sigma)$, and let f_i^σ be the solution of (1.13). Define $e_i^\sigma = f_i^\sigma - f(x_i^\sigma)$. Subtracting, we get the error equation

$$e_i^\sigma + \sum_{j,\tau} K_{ij}^{\sigma\tau}\psi_j^\tau[e_j^\tau - e_i^\sigma]A_j^\tau\, h^2 + r_i^\sigma, \tag{4.23}$$

$$r_i^\sigma = \int K(x_i^\sigma, x')[f(x') - f(x_i^\sigma)]\,dx' - \sum_{j,\tau} K_{ij}^{\sigma\tau}\psi_j^\tau[f(x_j^\tau) - f(x_i^\sigma)]A_j^\tau\, h^2. \tag{4.24}$$

The $O(\delta^5)$ smoothing estimate of section 2 and the discretization estimate (3.25) apply to r_i^σ, since f is smooth, and we have $|r_i^\sigma| \le \varepsilon(h,\delta)$, uniformly in i, σ, where $\varepsilon(h,\delta)$ is the right side of (1.17). A similar estimate follows for e_i^σ, after applying (4.22) to the error equation (4.23), and we have $|f_i^\sigma - f(x_i^\sigma)| \le \varepsilon(h,\delta)$. Thus we have proved the error estimate (1.17) of Theorem 1.1.

4.8. The iteration. Finally, we discuss the convergence of the iterates (1.16) to the solution of the discrete integral equation (1.13). It is a fact of potential theory that the operator T has spectrum in the interval $-\frac{1}{2} < \lambda \le \frac{1}{2}$ (see, e.g., [18, section 10.5] or [9, section 5.1]). Then, for $0 < \beta < 1$, the related operator $(1-\beta)I - 2\beta T$ has spectral radius < 1, and the convergence of (1.5) follows. For the discrete case we can argue that the approximations to T above perturb this spectral radius only slightly, and thus the discrete iteration (1.16) also converges. We omit the details.

5. Numerical examples. We present the results of three test problems illustrating the solution of the integral equation and the subsequent calculation of the solution of the Dirichlet problem (1.1) at points near the surface. The results generally verify the predictions of the theory.

For our first two examples the surface $\mathcal{S}$ is the unit sphere $x_1^2 + x_2^2 + x_3^2 = 1$. We cover the sphere with two stereographic projections. The projection on the equatorial plane by rays through the south pole $(0,0,-1)$ gives the first coordinate system, $X^1 : \mathbf{R}^2 \to U^1 = S - \{(0,0,-1)\}$:

$$x_1 = \frac{2\alpha_1}{1+|\alpha|^2}, \quad x_2 = \frac{2\alpha_2}{1+|\alpha|^2}, \quad x_3 = \frac{1-|\alpha|^2}{1+|\alpha|^2}. \tag{5.1}$$

Similarly, projecting from the north pole gives a second system $X^2 : \mathbf{R}^2 \to U^2 = S - \{(0,0,1)\}$ as in (5.1) but with $x_3 \to -x_3$. Each system maps the unit disk in the plane to a hemisphere; we use disks of radius 1.25 so that the two systems overlap. To define the partition of unity $\{\psi^1, \psi^2\}$, we first set $\phi^\sigma(X^\sigma(\alpha)) = \exp(-1.25^2/(1.25^2 - |\alpha|^2))$ for $|\alpha| \le 1.25$ and $\phi^\sigma = 0$ otherwise; then ϕ^σ is a smooth function with support $\{X^\sigma(\alpha) : |\alpha| \le 1.25\}$. We then define $\psi^\sigma(x) = \phi^\sigma(x)/(\phi^1(x) + \phi^2(x))$. Grid points $\alpha_i = ih$ are introduced with $X^\sigma(\alpha_i) \in V^\sigma$, where V^σ is the interior of the support of ψ^σ. These sets are $V^1 = \mathcal{S} \cap \{x_3 > -9/41\}$, $V^2 = \mathcal{S} \cap \{x_3 < 9/41\}$, corresponding

TABLE 1
The integral equation with third order kernel.

		$\delta = .5h^{2/3}$			$\delta = .75h^{2/3}$			$\delta = 2h$	
$1/h$	Grid pts	δ/h	Rel err	Order	δ/h	Rel err	Order	Rel err	Order
8	610	1.00	1.7E-3		1.50	6.0E-3		1.4E-2	
16	2490	1.26	4.8E-4	1.9	1.89	1.6E-3	1.9	1.9E -3	2.9
32	10026	1.59	1.2E-4	2.0	2.38	4.0E-4	2.0	2.4E-4	3.0

TABLE 2
The integral equation with fifth order kernel.

		$\delta = .5h^{2/3}$			$\delta = .75h^{2/3}$			$\delta = 2h$	
$1/h$	Grid pts	δ/h	Rel err	Order	δ/h	Rel err	Order	Rel err	Order
8	610	1.00	6.0E-4		1.50	5.9E-4		4.8E-4	
16	2490	1.26	9.5E-5	2.7	1.89	1.5E-5	5.3	2.0E-5	4.6
32	10026	1.59	7.7E-6	3.6	2.38	1.7E-6	3.1	9.0E-7	4.5
64	40138	2.00	1.4E-7	5.7	3.00	1.7E-7	3.3	1.4E-7	2.6

to $|\alpha| < 1.25$. The metric tensor in each patch is $g_{ij} = 4(|\alpha|^2 + 1)^{-2}\delta_{ij}$, and the area factor $A = \sqrt{g}$ is $4(|\alpha|^2+1)^{-2}$. The surface Laplacian is $\frac{1}{4}(|\alpha|^2+1)^2(\partial_{11}+\partial_{22})$. These quantities could be computed from the points $\{X^\sigma(\alpha_i)\}$, but we used analytical values in our computations.

Our first example is based on a spherical harmonic, so that the solution for the integral equation is known, as well as for the boundary value problem. We define

$$f(x) = 1.75(y_1 - 2y_2)(7.5y_3^2 - 1.5), \qquad y = Mx,$$

with $M = (1/\sqrt{6})(\sqrt{2}(1,1,1)^T, \sqrt{3}(0,1,-1)^T, (-2,1,1)^T)$, an orthogonal matrix. We use M to avoid rectangular symmetry in the test problem. The functions $f(x/r)r^3$ and $f(x/r)/r^4$, $r = |x|$, are both harmonic. The double layer potential u due to f is determined by the jump in u and the fact that $\partial u/\partial n$ is continuous. It is

$$u(x) = (4/7)f(x/r)r^3, \quad r < 1; \qquad u(x) = (-3/7)f(x/r)/r^4, \quad r > 1. \tag{5.2}$$

Thus if we set $g(x) = (4/7)f(x)$ on $\mathcal{S}$, the solution of the Dirichlet problem (1.1) is u, as defined above for $r < 1$, and the solution of the integral equation (1.4) is f.

We solved the discrete form (1.13) of the exact integral equation (1.4) using 12 iterations of (1.16) with $\beta = .7$. We tested the third order kernel (1.10) as well as the fifth order one (1.12) to check the order of accuracy. The results are reported in Tables 1 and 2. A grid size h in the coordinate systems and the total number of grid points are displayed. Each coordinate patch is a disk with $5/2h$ points along a diameter. We choose the smoothing radius δ proportional to h^q, $q = 2/3$ or 1. The relative error displayed is the maximum error divided by $\max_{\mathcal{S}} f \approx 8.1$. The computed order of accuracy is found from two successive cases. With the third order kernel, the expected order of accuracy $3q$ is clearly visible in Table 1. The errors shown in Table 2 for the fifth order kernel are much smaller. With $q = 2/3$ the predicted order $10/3$ is less evident, but it appears to take over in the second case, with the larger δ. With $q = 1$, $\delta = 2h$, the order is at first near 5 but deteriorates as h is decreased, as we should expect from the analysis.

After solving the integral equation with a choice of h and δ, using the fifth order kernel, we then computed the solution u at points near $\mathcal{S}$ with the same h and δ. To select a set of points, we cover $\mathbf{R}^3$ with a 3D mesh of spacing h; it is an arbitrary

TABLE 3
Nearby points for the first problem.

	Irreg	$\delta = .5h^{2/3}$		$\delta = .75h^{2/3}$		$\delta = 2h$	
$1/h$	points	Rel err	Order	Rel err	Order	Rel err	Order
8	606	1.2E-2		1.4E-2		2.1E-2	
16	2546	6.2E-4	4.3	1.9E-3	2.9	2.3E-3	3.2
32	10470	1.4E-4	2.1	4.7E-4	2.0	2.8E-4	3.0
64	42282	3.5E-5	2.0	1.1E-4	2.1	3.5E-5	3.0

TABLE 4
Nearby points for the second problem.

	Irreg	$\delta = .5h^{2/3}$		$\delta = .75h^{2/3}$		$\delta = 2h$	
$1/h$	points	Rel err	Order	Rel err	Order	Rel err	Order
8	606	1.6E-2		3.2E-2		4.7E-2	
16	2546	4.2E-4	5.3	1.3E-3	4.6	1.6E-3	4.9
32	10470	1.1E-4	1.9	3.6E-4	1.9	2.2E-4	2.9
64	42282	2.7E-5	2.0	8.8E-5	2.0	2.7E-5	3.0

choice to use the same h in $\mathbf{R}^3$ as on $\mathcal{S}$. We select the set of "irregular" points (i_1h, i_2h, i_3h) inside $\mathcal{S}$ for which the stencil of the five-point discrete Laplacian crosses the surface; i.e., the two points obtained by displacement of $\pm h$ in some coordinate are on different sides of $\mathcal{S}$. All such points are within h of $\mathcal{S}$, and, for h small, all points within ch of $\mathcal{S}$ have this property if $c < 1/\sqrt{3}$. Thus they provided a good sample of 3D grid points near $\mathcal{S}$. These are the interior points needed to form the discrete Laplacian of u in order to recover the values elsewhere; cf. [21] or [3].

For each selected point y we find the closest point x_0 on $\mathcal{S}$. We need f and the first two derivatives at x_0 for the subtraction and corrections. We find these from the computed values of f at the grid points by Lagrange interpolation. We compute the sum (1.19) and add the smoothing and discretization corrections (1.21)–(1.25). Table 3 gives the number of irregular points; the relative error, found as the maximum error divided by $\max_{\mathcal{S}} g \approx 4.6$; and the computed order of accuracy. With $\delta = ch^q$, $q = 2/3$ or 1, we see the expected order $3q = 2$ or 3.

Our second test problem, still with the unit sphere, is the harmonic function

$$u(x) = \exp(y_1 + 2y_2)\cos\sqrt{5}y_3, \qquad y = Mx, \tag{5.3}$$

with M as before. We prescribe $g(x) = u(x)$ on $\mathcal{S}$ and solve the integral equation within an error tolerance, again with $\beta = .7$. In this case we do not know the solution f of the integral equation; however, we use the computed f to find u at the irregular points, as before, and compare with the exact solution in (5.3). Results are reported in Table 4. Again the predicted order of accuracy is evident as h decreases. The error displayed is the maximum error divided by $\max_{\mathcal{S}} g \approx 9.4$.

For our third problem the surface $\mathcal{S}$ is the ellipsoid $x_1^2 + x_2^2 + x_3^2/2 = 1$. We use coordinate systems similar to those for the sphere, with x_3 in (5.1) multiplied by $\sqrt{2}$. As the test problem we take the same harmonic function u in (5.3). The coordinate systems are symmetric, but the functions are not, because of the rotation by M. The needed geometric quantities can be found analytically in this case. The coordinates are not orthogonal, as they were for the sphere, so that the formulas are tested in a more general setting. We solve the integral equation as before. We then select the set of irregular points by the same criterion and compute the solution u on this set as a double layer potential, with corrections, using the solution f of the integral equation.

TABLE 5
Nearby points for the ellipsoid.

	Irreg	$\delta = .5h^{2/3}$		$\delta = .75h^{2/3}$		$\delta = 2h$	
$1/h$	points	Rel err	Order	Rel err	Order	Rel err	Order
8	798	1.7E-2		3.3E-2		4.7E-2	
16	3330	3.2E-4	5.7	1.0E-3	5.0	1.2E-3	5.3
32	13614	8.3E-5	2.0	2.7E-4	1.9	1.6E-4	2.9
64	54914	2.1E-5	2.0	6.6E-5	2.0	2.1E-5	3.0

TABLE 6
Nearby points with $h = 1/32$, varying δ.

δ/h	.01	.1	.5	1	2	5	10
Rel err	3.7E-4	3.7E-4	2.8E-4	6.8E-5	1.6E-4	2.2E-3	1.5E-2

For each irregular point y we need to find the point x_0 on the surface so that y is along the normal from x_0, as well as the distance; we do this using Newton's method. The results, displayed in Table 5, are similar to those for the sphere. Again the relative error is the maximum error divided by $\max_{\mathcal{S}} g \approx 9.4$.

For the last problem we show in Table 6 the results of computing the solution at the irregular points using various choices of the smoothing parameter δ. We first solve the integral equation with $h = 1/32$ and $\delta = 2h$. We then compute the values at the irregular points, chosen with $h = 1/32$, for various values of δ/h, to verify the effects of the two corrections. For small δ/h, the discretization correction is dominant; when δ/h is extremely small, more terms are needed in the sum (1.25). For larger δ/h the smoothing correction is important, and the discretization correction is negligible. As δ/h increases, the remaining smoothing error becomes significant.

REFERENCES

[1] C. ANDERSON, *A method of local corrections for computing the velocity field due to a distribution of vortex blobs*, J. Comput. Phys., 62 (1986), pp. 111–123.

[2] K. E. ATKINSON, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[3] J. T. BEALE AND M.-C. LAI, *A method for computing nearly singular integrals*, SIAM J. Numer. Anal., 38 (2001), pp. 1902–1925.

[4] J. T. BEALE, *A convergent boundary integral method for three-dimensional water waves*, Math. Comp., 70 (2001), pp. 977–1029.

[5] J. T. BEALE AND A. MAJDA, *Vortex methods,* I: *Convergence in three dimensions*, Math. Comp., 39 (1982), pp. 1–27.

[6] J. T. BEALE AND A. MAJDA, *High order accurate vortex methods with explicit velocity kernels*, J. Comput. Phys., 58 (1985), pp. 188–208.

[7] O. BRUNO AND L. KUNYANSKY, *A fast, high-order algorithm for the solution of surface scattering problems: Basic implementation, tests, and applications*, J. Comput. Phys., 169 (2001), pp. 80–110.

[8] L. CANINO, J. OTTUSCH, M. STALZER, J. VISHER, AND S. WANDZURA, *Numerical solution of the Helmholtz equation in 2D and 3D using a high-order Nyström discretization*, J. Comput. Phys., 146 (1998), pp. 627–663.

[9] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, Wiley, New York, 1983.

[10] R. CORTEZ, *The method of regularized Stokeslets*, SIAM J. Sci. Comput., 23 (2001), pp. 1204–1225.

[11] G. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1995.

[12] J. GOODMAN, T. Y. HOU, AND J. LOWENGRUB, *Convergence of the point vortex method for the 2-D Euler equations*, Comm. Pure Appl. Math., 43 (1990), pp. 415–430.

[13] I. Graham and I. Sloan, *Fully discrete spectral boundary integral methods for Helmholtz problems on smooth closed surfaces in R^3*, Numer. Math., 92 (2002), pp. 289–323.
[14] L. Greengard and V. Rokhlin, *A new version of the fast multipole method for the Laplace equation in three dimensions*, Acta Numer., 6 (1997), pp. 229–269.
[15] W. Hackbusch, *Integral Equations, Theory and Numerical Treatment*, Birkhäuser, Basel, 1995.
[16] O. H. Hald, *Convergence of vortex methods for Euler's equations.* II, SIAM J. Numer. Anal., 16 (1979), pp. 726–755.
[17] Q. Huang and T. A. Cruse, *Some notes on singular integral techniques in boundary element analysis*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 2643–2659.
[18] R. Kress, *Linear Integral Equations*, 2nd ed., Springer, New York, 1999.
[19] C. Lage and C. Schwab, *Wavelet Galerkin algorithms for boundary integral equations*, SIAM J. Sci. Comput., 20 (1999), pp. 2195–2222.
[20] J. S. Lowengrub, M. J. Shelley, and B. Merriman, *High-order and efficient methods for the vorticity formulation of the Euler equations*, SIAM J. Sci. Comput., 14 (1993), pp. 1107–1142.
[21] A. Mayo, *Fast high order accurate solution of Laplace's equation on irregular regions*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 144–157.
[22] M. Nitsche, *Axisymmetric vortex sheet motion: Accurate evaluation of the principal value integral*, SIAM J. Sci. Comput., 21 (1999), pp. 1066–1084.
[23] A. Rathsfeld, *Nyström's method and iterative solvers for the solution of the double-layer potential equation over polyhedral boundaries*, SIAM J. Numer. Anal., 32 (1995), pp. 924–951.
[24] A. Rathsfeld, *Quadrature methods for $2D$ and $3D$ problems*, J. Comput. Appl. Math., 125 (2000), pp. 439–460.
[25] C. Schwab and W. L. Wendland, *On the extraction technique in boundary integral equations*, Math. Comp., 68 (1999), pp. 91–122.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 621–640

A COERCIVE COMBINED FIELD INTEGRAL EQUATION FOR ELECTROMAGNETIC SCATTERING*

A. BUFFA† AND R. HIPTMAIR‡

Abstract. Many boundary integral equation methods used in the simulation of direct electromagnetic scattering of a time-harmonic wave at a perfectly conducting obstacle break down when applied at frequencies close to a resonant frequency of the obstacle. A remedy is offered by special indirect boundary element methods based on the so-called combined field integral equation. However, hitherto no theoretical results about the convergence of discretized combined field integral equations have been available.

In this paper we propose a new combined field integral equation, convert it into variational form, establish its coercivity in the natural trace spaces for electromagnetic fields, and conclude existence and uniqueness of solutions for any frequency. Moreover, a conforming Galerkin discretization of the variational equations by means of div_Γ-conforming boundary elements can be shown to be asymptotically quasi-optimal. This permits us to derive quantitative convergence rates on sufficiently fine, uniformly shape-regular sequences of surface triangulations.

Key words. electromagnetic scattering, combined field integral equations, coercivity, boundary element methods, Galerkin scheme

AMS subject classifications. 65N30, 65N38, 83N50

DOI. 10.1137/S0036142903423393

1. Introduction.

The numerical simulation of direct scattering at a perfect conductor, the so-called scatterer, is a central task in computational electromagnetism. The scatterer occupies a bounded domain $\Omega \subset \mathbb{R}^3$. In general, Ω will have Lipschitz-continuous boundary $\Gamma := \partial\Omega$, which can be equipped with an exterior unit normal vector field $\mathbf{n} \in \boldsymbol{L}^\infty(\Gamma)$. With boundary element methods in mind, we do not lose generality by admitting only scatterers Ω that are polyhedra with flat faces and a Lipschitz-continuous boundary. We emphasize that the extension of the results to curvilinear faces is straightforward.

Electromagnetic waves propagate outside the scatterer in the "air region" $\Omega' := \mathbb{R}^3 \setminus \Omega$. From an electrodynamic point of view, Ω' is filled with a homogeneous, isotropic, and linear material. Excitation is provided by the electric field $\mathbf{e}_i$ of an incident (plane) wave of angular frequency $\omega > 0$. Hence, we can switch to the frequency domain and are left with complex amplitudes (phasors) as unknown spatial functions. After suitable scaling, the complex amplitude $\mathbf{e}$ of the scattered field satisfies the following exterior Dirichlet problem for the *electric wave equation* [23, Chap. 6]:

$$\mathbf{curl}\,\mathbf{curl}\,\mathbf{e} - \kappa^2\mathbf{e} = 0 \quad \text{in } \Omega', \tag{1.1}$$

$$\mathbf{e} \times \mathbf{n} = g := \mathbf{e}_i \times \mathbf{n} \quad \text{on } \Gamma. \tag{1.2}$$

The constant $\kappa := \omega\sqrt{\epsilon_0\mu_0} > 0$ is called the *wave number*, because $\kappa/2\pi$ tells us the number of wavelengths per unit length. Henceforth, κ will stand for a fixed

*Received by the editors February 19, 2003; accepted for publication (in revised form) October 16, 2003; published electronically April 14, 2004.
http://www.siam.org/journals/sinum/42-2/42339.html
†Istituto di Matematica Applicate e Tecnologie Informatiche del CNR, Via Ferrata 1, 27100 Pavia, Italy (annalisa@imati.cnr.it).
‡Seminar für Angewandte Mathematik, ETH Zürich, CH-8092 Zürich, Switzerland (hiptmair@sam.math.ethz.ch). The research of this author was supported by the Newton Institute, Cambridge, UK.

positive wave number. These equations have to be supplemented with the *Silver–Müller radiation conditions*

$$\int_{\partial B_r} |\mathbf{curl}\,\mathbf{e} \times \mathbf{n} + i\kappa(\mathbf{n} \times \mathbf{e}) \times \mathbf{n}|^2 \, \mathrm{d}S \to 0 \quad \text{for } r \to \infty, \tag{1.3}$$

where B_r is a ball around 0 with radius $r > 0$. Existence and uniqueness of solutions of (1.1) and (1.3) can be inferred from Rellich's lemma [18, 42].

Integral equation methods are a natural choice for the discretization of the direct electromagnetic scattering problem, which is posed on an unbounded domain. Prominent examples are the electric field integral equation (EFIE) and magnetic field integral equation; see [42, sect. 5.6] or [18, Chap. 3]. These indirect methods display a worrisome instability when κ^2 coincides with a Dirichlet or Neumann eigenvalue (resonant frequency) of the **curl curl**-operator inside Ω; then the integral equation may not have a solution. After discretization this manifests itself in extreme ill-conditioning of the resulting linear systems of equations if κ is close to a resonant frequency [20].

Two classes of integral equation methods are known to avoid this difficulty. The first is the method proposed in [26] and examined for electromagnetism in [30]. However, it entails constructing an auxiliary surface and can be haunted by stability problems, too. The second, vastly more popular class of methods are approaches based on combined field integral equations (CFIEs), as introduced in [3] and [17]. A particular representative will be the focus of this paper.

CFIEs owe their name to the presence of both single and double layer potentials in the ansatz for the electric field in Ω'. As a theoretical tool they were pioneered for acoustic scattering in [3] and for electromagnetics in [43]. These methods are widely used in computational electromagnetism [47]. For acoustics, existence and uniqueness of solutions can be shown for smooth scatterers [23]. Yet, in the case of electromagnetism even this remains elusive. Hence, mainly for the sake of theoretical treatment, regularized formulations have been introduced by Kress in [36]. However, the idea is applicable only for scattering at smooth objects, and it is not suitable for numerical implementation.

In this article we hark back to the idea of regularization in a different way. Based on recent advances in the understanding of boundary integral operators of electromagnetic scattering achieved in [15, 32, 33], we apply regularization to the double layer part of the integral operator. Reformulation as a mixed problem and subsequent Galerkin discretization pave the way to a practical computational scheme. It is the first method based on an electromagnetic CFIE that can be proven to converge quasi-optimally in relevant trace norms. Related techniques for the Helmholtz equation of acoustic scattering are covered in [14].

The developments in this paper rest on a huge body of previous work. We will restate the most important results. However, in order to maintain a reasonable length we cannot elaborate on most of the existing theory of boundary integral equations for electromagnetic scattering. However, we will try to give comprehensive references for all results we rely upon.

The plan of the paper is as follows: the next section will give a concise survey of relevant function spaces and trace theorems and prove some new results which are needed in the rest of the paper. Then we briefly recall the crucial integral operators of electromagnetic scattering. In the fourth section we will present and analyze the new CFIE and the variational problem associated with it. The fifth section will be devoted to proving the asymptotic quasi optimality of a Galerkin discretization. Based on it, the final section will give quantitative convergence estimates.

2. Function spaces and traces. Let $\Omega \subseteq \mathbb{R}^3$ be any of the sets $\Omega, \Omega', \mathbb{R}^3$. We define the Fréchet space $\boldsymbol{L}^2_{\rm loc}(\Omega) = \{\mathbf{u}_{|\Omega} : \mathbf{u} \in \mathbf{L}^2_{\rm loc}(\mathbb{R}^3)\}$, where $\boldsymbol{L}^2_{\rm loc}(\mathbb{R}^3)$ is the space of complex, vector-valued, locally square-integrable functions on $\mathbb{R}^3$. In a similar way, we define the Sobolev spaces $\boldsymbol{H}^s_{\rm loc}(\Omega)$, $s \geq 0$ (see, e.g., [1] for definitions), with the convention $\boldsymbol{H}^0 \equiv \boldsymbol{L}^2$. The subscript $_{\rm loc}$ will be dropped whenever Ω is bounded: in this case, $\boldsymbol{H}^s(\Omega)$ is a Hilbert space endowed with the natural graph norm $\|\mathbf{u}\|_{\boldsymbol{H}^s(\Omega)}$ and seminorm $|\mathbf{u}|_{\boldsymbol{H}^s(\Omega)}$, respectively [1]. Parentheses will consistently be used to express inner products.

With D a first order differential operator, for any $s \geq 0$ we define

$$\boldsymbol{H}^s_{\rm loc}(D,\Omega) := \{\mathbf{u} \in \boldsymbol{H}^s_{\rm loc}(\Omega) :\ D\mathbf{u} \in \boldsymbol{H}^s_{\rm loc}(\Omega)\}, \tag{2.1}$$

$$\boldsymbol{H}^s_{\rm loc}(D0,\Omega) := \{\mathbf{u} \in \boldsymbol{H}^s_{\rm loc}(\Omega) :\ D\mathbf{u} = 0\}. \tag{2.2}$$

When $s = 0$, we simplify the notation by setting $\boldsymbol{H}^0 = \boldsymbol{H}$. If Ω is bounded, $\boldsymbol{H}^s_{\rm loc}(D,\Omega)$ is endowed with the graph norm $\|\cdot\|^2_{\boldsymbol{H}^s(D,\Omega)} := \|\cdot\|^2_{\boldsymbol{H}^s(\Omega)} + \|D\cdot\|^2_{\boldsymbol{H}^s(\Omega)}$ and seminorm $|\cdot|^2_{\boldsymbol{H}^s(D,\Omega)} := |\cdot|^2_{\boldsymbol{H}^s(\Omega)} + |D\cdot|^2_{\boldsymbol{H}^s(\Omega)}$. This defines the spaces $\boldsymbol{H}^s(\mathbf{curl},\Omega)$, $\boldsymbol{H}^s(\mathrm{div},\Omega)$ and $\boldsymbol{H}^s(\mathbf{curl}\,0,\Omega)$, $\boldsymbol{H}^s(\mathrm{div}\,0,\Omega)$, for which [28, Chap. 1] is the main reference.

The integration by parts formulae for the operators **curl** and div suggest that we define the tangential trace mapping $\gamma_{\mathbf{t}} : \mathbf{u} \mapsto \mathbf{u}_{|\Gamma} \times \mathbf{n}$ and the normal component trace $\gamma_{\mathbf{n}} : \mathbf{u} \mapsto \mathbf{u}_{|\Gamma} \cdot \mathbf{n}$. To begin with, they are defined for $\mathbf{u} \in C^\infty(\overline{\Omega})^3$.

The trace theorem for $\boldsymbol{H}^1(\Omega)$ [29, Thm. 1.5.1.1] shows that the tangential trace $\gamma_{\mathbf{t}} : \boldsymbol{C}^\infty(\bar{\Omega}) \mapsto \boldsymbol{L}^\infty(\Gamma)$ and the normal trace $\gamma_{\mathbf{n}} : \boldsymbol{C}^\infty(\bar{\Omega}) \mapsto L^\infty(\Gamma)$ are continuous as mappings $\boldsymbol{H}(\mathbf{curl};\Omega) \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\Gamma)$ and $\boldsymbol{H}(\mathrm{div};\Omega) \mapsto H^{-\frac{1}{2}}(\Gamma)$, respectively. Here, $H^{-\frac{1}{2}}(\Gamma)$ and $\boldsymbol{H}^{-\frac{1}{2}}(\Gamma)$ are the dual spaces of $H^{\frac{1}{2}}(\Gamma)$ and $\boldsymbol{H}^{\frac{1}{2}}(\Gamma) := (H^{\frac{1}{2}}(\Gamma))^3$, respectively, with respect to the pivot spaces $L^2(\Gamma)$ and $\boldsymbol{L}^2(\Gamma)$. Consequently, the traces can be extended to $\boldsymbol{H}(\mathbf{curl};\Omega)$ and $\boldsymbol{H}(\mathrm{div};\Omega)$, respectively. Moreover, if we define the antisymmetric pairing

$$\langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{\boldsymbol{\tau},\Gamma} := \int_\Gamma (\boldsymbol{\mu} \times \mathbf{n}) \cdot \boldsymbol{\eta}\, \mathrm{d}S, \qquad \boldsymbol{\mu}, \boldsymbol{\eta} \in \boldsymbol{L}^2_{\mathbf{t}}(\Gamma) := \{\mathbf{u} \in (L^2(\Gamma))^3,\ \mathbf{u} \cdot \mathbf{n} = 0\}, \tag{2.3}$$

then we can state the integration by parts formula for the **curl**-operator as [9, sect. 4]

$$\int_\Omega (\mathbf{curl}\,\mathbf{u} \cdot \mathbf{v} - \mathbf{u} \cdot \mathbf{curl}\,\mathbf{v})\, \mathrm{d}\mathbf{x} = \langle \gamma_{\mathbf{t}}\mathbf{v}, \gamma_{\mathbf{t}}\mathbf{u} \rangle_{\boldsymbol{\tau},\Gamma}. \tag{2.4}$$

A meaningful strong form of the electric wave equation (1.1) has to rely on yet another space: from the fact that a field $\mathbf{u}$ is a locally square-integrable function satisfying $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u} - \mathbf{u} = 0$ we can conclude that $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}$ is locally square-integrable, too. Hence, the space

$$\boldsymbol{H}_{\rm loc}(\mathbf{curl}^2,\Omega) := \{\mathbf{u} \in \boldsymbol{H}_{\rm loc}(\mathbf{curl};\Omega),\ \mathbf{curl}\,\mathbf{curl}\,\mathbf{u} \in \boldsymbol{L}^2_{\rm loc}(\Omega)\}$$

will play the role of the natural space for solutions of the electric wave equation with constant coefficients.

Trace spaces for electromagnetic fields are essential for stating the boundary integral equations and, in particular, their variational formulations. The corresponding results on nonsmooth boundaries are fairly recent: we refer the reader to [6, 9, 10] for the treatment of Lipschitz polyhedra. The issue of traces of $\boldsymbol{H}(\mathbf{curl};\Omega)$ for general Lipschitz domains was settled in [13]. The results are summarized in the survey article [7].

DEFINITION 2.1. *We introduce the Hilbert spaces* $\boldsymbol{H}^s_\times(\Gamma) := \gamma_{\mathbf{t}}(\boldsymbol{H}^{s+1/2}(\Omega))$, $s \in (0,1)$, *equipped with an inner product that renders* $\gamma_{\mathbf{t}} : \boldsymbol{H}^{s+1/2}(\Omega) \mapsto \boldsymbol{H}^s_\times(\Gamma)$ *continuous and surjective. For* $s = 0$ *we set* $\mathbf{H}^0_\times(\Gamma) := \mathbf{L}^2_t(\Gamma)$. *The dual spaces with respect to the pairing* $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$ *are denoted by* $\boldsymbol{H}^{-s}_\times(\Gamma)$.

REMARK 1. *When* $s = 1$, *the standard trace operator* γ *fails to map* $H^{3/2}(\Omega)$ *to* $H^1(\Gamma)$, *although* $H^1(\Gamma)$ *is well defined on the boundary* Γ. *In this case, we adopt the definition* $\mathbf{H}^1_\times(\Gamma) := \gamma_{\mathbf{t}}(\gamma^{-1}H^1(\Gamma)^3)$, *where* γ^{-1} *represents any continuous lifting from* $H^1(\Gamma)$ *to* $H^{3/2}(\Omega)$ *(see* [35]*).*

Next, we introduce the surface divergence operator div_Γ; cf. [9, sect. 2.1].

Let $\{\Gamma_1,\dots,\Gamma_P\}$, $P \in \mathbb{N}$, stand for the set of open flat faces of Γ, and write Σ_{ij} for the straight edge $\partial\Gamma_j \cap \partial\Gamma_i$. The vector $\boldsymbol{\nu}^{ij}$ lies in the plane of Γ_j, is perpendicular to Σ_{ij}, and points into the exterior of Γ_j. Then for $\mathbf{u} \in C^\infty(\overline{\Omega})$ we set

$$\mathrm{div}_\Gamma \gamma_{\mathbf{t}}\mathbf{u} := \begin{cases} \mathrm{div}_j(\gamma_{\mathbf{t}}\mathbf{u}_{|\Gamma_j}) & \text{on } \Gamma^j, \\ \left((\gamma_{\mathbf{t}}\mathbf{u}_{|\Gamma_j})\cdot\boldsymbol{\nu}^{ij} + (\gamma_{\mathbf{t}}\mathbf{u}_{|\Gamma_i})\cdot\boldsymbol{\nu}^{ji}\right)\delta_{ij} & \text{on } \overline{\Gamma^j}\cap\overline{\Gamma^i}, \end{cases} \tag{2.5}$$

where δ_{ij} is the delta distribution (in local coordinates) whose support is the edge $\overline{\Gamma^j}\cap\overline{\Gamma^i}$ and div_j denotes the two-dimensional divergence computed on the face Γ^j. By density, this differential operator can be extended to less regular distributions and, in particular, to functionals in $\boldsymbol{H}^{-\frac{1}{2}}_\times(\Gamma)$. We set

$$\boldsymbol{H}^s_\times(\mathrm{div}_\Gamma,\Gamma) := \{\boldsymbol{\mu} \in \boldsymbol{H}^s_\times(\Gamma),\ \mathrm{div}_\Gamma\boldsymbol{\mu} \in H^s(\Gamma)\}, \qquad s \in [-1/2, 0].$$

Finally, we denote by $\mathbf{curl}_\Gamma$ the operator adjoint to div_Γ with respect to the pairing $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$, i.e.,

$$\langle\mathbf{curl}_\Gamma q, \mathbf{p}\rangle_{\boldsymbol{\tau},\Gamma} = \langle\mathrm{div}_\Gamma\mathbf{p}, q\rangle_{\frac{1}{2},\Gamma}, \quad \mathbf{p} \in \boldsymbol{H}^{-\frac{1}{2}}_\times(\mathrm{div}_\Gamma,\Gamma), \quad q \in H^{\frac{1}{2}}(\Gamma). \tag{2.6}$$

It is known [7, sect. 1.2] that $\mathbf{curl}_\Gamma : H^s(\Gamma) \to \boldsymbol{H}^{s-1}_\times(\Gamma)$ is continuous for every s, $1/2 \le s \le 1$. The spaces just defined turn out to be the desired trace spaces; see [9, Prop. 1.7], [10, Thm. 5.4], and [13, sect. 2].

THEOREM 2.2. *The operator* $\gamma_{\mathbf{t}} : \boldsymbol{H}(\mathbf{curl};\Omega) \mapsto \boldsymbol{H}^{-\frac{1}{2}}_\times(\mathrm{div}_\Gamma,\Gamma)$ *is continuous, is surjective, and possesses a continuous right inverse.*

The following self-duality of the electromagnetic trace space will be the foundation of weak formulations. The result was first given in [10].

THEOREM 2.3. *The pairing* $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$ *can be extended to a continuous bilinear form on* $\boldsymbol{H}^{-\frac{1}{2}}_\times(\mathrm{div}_\Gamma,\Gamma)$. *With respect to* $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$ *the space* $\boldsymbol{H}^{-\frac{1}{2}}_\times(\mathrm{div}_\Gamma,\Gamma)$ *becomes its own dual.*

Piecewise smooth scatterers offer the possibility that some considerations can be done locally on the faces and thus become essentially two-dimensional. To provide a framework for such considerations we introduce the spaces $\boldsymbol{H}^s_\times(\Gamma_j)$, $s \ge 0$, defined locally on a face Γ_j in a straightforward fashion, and we denote by $\boldsymbol{H}^{-s}_\times(\Gamma_j)$, $s \in (0, \frac{1}{2})$, their duals (note that we adopt the notion introduced in [39] and not the one used in [29]).

In addition, we define the localized spaces $\boldsymbol{H}_{\times 0}(\mathrm{div}_\Gamma,\Gamma_j) := \{\mathbf{u} \in \boldsymbol{L}^2_{\mathbf{t}}(\Gamma_j) : \widetilde{\mathbf{u}} \in \boldsymbol{H}_\times(\mathrm{div}_\Gamma,\Gamma)\}$, where $\widetilde{\ }$ denotes the trivial extension by zero to all of Γ. These spaces will be combined to

$$\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma) := \prod_{j=1}^P \boldsymbol{H}_{\times 0}(\mathrm{div}_\Gamma,\Gamma_j).$$

LEMMA 2.4. *The space* $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$ *is dense in* $\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$.

Proof. Let us adopt the notation Σ for the skeleton of the polyhedron, that is, the union of all edges Σ_{ij}, $1 \le i,j \le P$. Then we recall that regular functions compactly supported in $\bar{\Omega} \setminus \Sigma$ are dense in $H^1(\Omega)$ [41]. Of course, also the inclusion $\boldsymbol{H}^1(\Omega) \subset \boldsymbol{H}(\mathbf{curl},\Omega)$ is dense. By continuity of the tangential trace operator $\gamma_\mathbf{t}$, we deduce that tangential vector fields in $\boldsymbol{H}_\times^{1/2}(\Gamma)$ compactly supported in $\Gamma \setminus \Sigma$ are dense in $\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$. Since the set of fields in $\boldsymbol{H}_\times^{1/2}(\Gamma)$ compactly supported in $\Gamma \setminus \Sigma$ is a subset of $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$, the statement is proved. ∎

LEMMA 2.5. *The embedding* $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma) \hookrightarrow \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ *is compact.*

Proof. To begin with, since $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma) \subset \boldsymbol{H}_\times(\mathrm{div}_\Gamma,\Gamma)$, we need to prove only that the injection $\boldsymbol{H}_\times(\mathrm{div}_\Gamma,\Gamma) \subset \boldsymbol{H}_\times^{-1/2}(\mathrm{div}_\Gamma,\Gamma)$ is compact.

Let $\{\mathbf{u}_n\}_{n\in\mathbb{N}} \subset \boldsymbol{H}_\times(\mathrm{div}_\Gamma,\Gamma)$ be a sequence such that $\|\mathbf{u}_n\|_{\boldsymbol{H}_\times(\mathrm{div}_\Gamma,\Gamma)} < 1$ for all n. Then, owing to the compact embedding $\mathbf{L}_\mathbf{t}^2(\Gamma) \hookrightarrow \boldsymbol{H}_\times^{-1/2}(\Gamma)$, there exists a subsequence $\mathbf{u}_{n_k}$ of $\mathbf{u}_n$ and a $\mathbf{u} \in \boldsymbol{H}_\times^{-1/2}(\Gamma)$ such that $\mathbf{u}_{n_k} \to \mathbf{u}$ strongly in $\boldsymbol{H}_\times^{-1/2}(\Gamma)$. The operator $\mathrm{div}_\Gamma : \boldsymbol{H}_\times^{-1/2}(\Gamma) \mapsto H^{-3/2}(\Gamma)$ is continuous (see [9] for a proof and the definition of $H^{-3/2}(\Gamma)$). Hence, $\mathrm{div}_\Gamma \mathbf{u}_{n_k} \to \mathrm{div}_\Gamma \mathbf{u}$ strongly in $H^{-3/2}(\Gamma)$.

On the other hand, we also know that $\|\mathrm{div}_\Gamma \mathbf{u}_{n_k}\|_{L^2(\Gamma)} < 1$, which implies that up to extraction of a subsequence $\mathrm{div}_\Gamma \mathbf{u}_{n_k}$ is strongly converging to an element in $H^{-1/2}(\Gamma)$. By uniqueness of the limit, we deduce that $\mathrm{div}_\Gamma \mathbf{u} \in H^{-1/2}(\Gamma)$, and, up to selecting a subsequence, $\mathbf{u}_{n_k} \to \mathbf{u} \in \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ strongly. ∎

When we want to examine the convergence of boundary element methods quantitatively, extra smoothness of the functions to be approximated is indispensable. A convenient gauge for smoothness is offered by scales of Sobolev spaces. Again, localization is a handy tool: for any $s > \frac{1}{2}$, we define $\mathbf{H}_-^s(\Gamma) := \{\mathbf{u} \in \mathbf{L}_\mathbf{t}^2(\Gamma) : \mathbf{u}_{|\Gamma^j} \in \boldsymbol{H}_\mathbf{t}^s(\Gamma^j)\}$. The corresponding space of scalar functions will be denoted by $H_-^s(\Gamma)$. Then, for $s \ge 1$, we set $\boldsymbol{H}_\times^s(\Gamma) := \boldsymbol{H}_\times^{\frac{1}{2}}(\Gamma) \cap \mathbf{H}_-^s(\Gamma)$.

To characterize extra smoothness of traces we resort to the family of Hilbert spaces

$$\boldsymbol{H}_\times^s(\mathrm{div}_\Gamma,\Gamma) := \begin{cases} \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) & \text{if } s = -\frac{1}{2}, \\ \{\boldsymbol{\mu} \in \boldsymbol{H}_\times^s(\Gamma),\ \mathrm{div}_\Gamma \boldsymbol{\mu} \in H^s(\Gamma)\} & \text{if } -\frac{1}{2} < s < \frac{1}{2}, \\ \{\boldsymbol{\mu} \in \boldsymbol{H}_\times^s(\Gamma),\ \mathrm{div}_\Gamma \boldsymbol{\mu} \in H_-^s(\Gamma)\} & \text{if } s \ge \frac{1}{2}. \end{cases}$$

The following trace theorem has been proved in the appendix of [8].

THEOREM 2.6. *Let* $\sigma \in \mathbb{R}$ *be the maximum real number such that* $\{p \in H^1(\Omega) : \Delta p \in L^2(\Omega),\ (\partial_\mathbf{n} p)_{|\Gamma} = 0\} \subset H^{1+\sigma'}(\Omega)$ *for all* $\sigma' < \sigma$. *For all* $0 \le s < \min\{\sigma, 1\}$ *the tangential trace mapping* $\gamma_\mathbf{t}$ *can be extended to a continuous and surjective mapping* $\gamma_\mathbf{t} : \boldsymbol{H}^s(\mathbf{curl},\Omega) \mapsto \boldsymbol{H}_\times^{s-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$, *which possesses a continuous right inverse.*

3. Potentials and integral operators. Here we define the boundary integral operators relevant for electromagnetic scattering and recall a few of their properties. More details can be found in [42, Chap. 5], [23, Chap. 6], [15, sect. 3], and [34].

DEFINITION 3.1. *A distribution* $\mathbf{e} \in \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2,\Omega)$ *is called a* Maxwell solution *on some generic domain* Ω, *if it satisfies* (1.1) *in* Ω, *and the Silver–Müller radiation conditions at* ∞ *if* Ω *is not bounded.*

As far as the differential operator $\mathbf{curl}\,\mathbf{curl} - \kappa^2\,\mathrm{Id}$ is concerned, the integration by parts formula (2.4) suggests the distinction between *Dirichlet trace* $\gamma_\mathbf{t}$ and *Neumann*

trace $\gamma_N := \kappa^{-1}\gamma_{\mathbf{t}} \circ \mathbf{curl}$. The trace γ_N can be labelled "magnetic," because it actually retrieves the tangential trace of the magnetic field solution. From the trace theorem, Theorem 2.2, we see that γ_N is meaningful on $\boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega \cup \Omega')$.

LEMMA 3.2. *The trace* $\gamma_N : \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega' \cup \Omega) \to \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ *is a continuous and surjective operator.*

The integral representation for Maxwell solutions relies on the famous Stratton–Chu representation formula for the electric field in $\Omega \cup \Omega'$ [46]. To state it we rely on the notion of a jump $[\cdot]_\Gamma$ across Γ defined by $[\gamma]_\Gamma := \gamma^+ - \gamma^-$ for some trace γ onto Γ. Here, superscripts $-$ and $+$ tag traces onto Γ from Ω and $\Omega' := \mathbb{R}^3 \setminus \bar{\Omega}$, respectively. For notational simplicity, it is also useful to resort to the average $\{\gamma\}_\Gamma = \frac{1}{2}(\gamma^+ + \gamma^-)$. Both operators can be applied only to functions defined in $\Omega \cup \Omega'$.

As elaborated in [23, sect. 6.2], [42, sect. 5.5], and [18, Chap. 3, sect. 1.3.2], any Maxwell solution in $\Omega \cup \Omega'$ satisfies

$$\mathbf{u}(\mathbf{x}) = -\boldsymbol{\Psi}^\kappa_{DL}([\gamma_{\mathbf{t}}]_\Gamma(\mathbf{u}))(\mathbf{x}) - \boldsymbol{\Psi}^\kappa_{SL}([\gamma_N]_\Gamma(\mathbf{u}))(\mathbf{x}), \qquad \mathbf{x} \in \Omega \cup \Omega', \tag{3.1}$$

where we have introduced the (electric) *Maxwell single layer potential*

$$\boldsymbol{\Psi}^\kappa_{SL}(\boldsymbol{\mu})(\mathbf{x}) := \kappa \boldsymbol{\Psi}^\kappa_{\mathbf{A}}(\boldsymbol{\mu})(\mathbf{x}) + \frac{1}{\kappa}\,\mathbf{grad}_{\mathbf{x}}\, \Psi^\kappa_V(\mathrm{div}_\Gamma \boldsymbol{\mu})(\mathbf{x}), \qquad \mathbf{x} \notin \Gamma, \tag{3.2}$$

and the (electric) *Maxwell double layer potential*

$$\boldsymbol{\Psi}^\kappa_{DL}(\boldsymbol{\mu})(\mathbf{x}) := \mathbf{curl}_{\mathbf{x}}\, \boldsymbol{\Psi}^\kappa_{\mathbf{A}}(\boldsymbol{\mu})(\mathbf{x}), \qquad \mathbf{x} \notin \Gamma. \tag{3.3}$$

Here, Ψ^κ_V and $\boldsymbol{\Psi}^\kappa_{\mathbf{A}}$ are the scalar and the vectorial single layer potentials for the Helmholtz kernel $E_\kappa(\mathbf{x}) := \exp(i\kappa|\mathbf{x}|)/4\pi|\mathbf{x}|$, whose integral representation is given by ($\mathbf{x} \notin \Gamma$)

$$\Psi^\kappa_V(\phi)(\mathbf{x}) := \int_\Gamma \phi(\mathbf{y}) E_\kappa(\mathbf{x} - \mathbf{y})\, dS(\mathbf{y}), \qquad \boldsymbol{\Psi}^\kappa_{\mathbf{A}}(\boldsymbol{\mu})(\mathbf{x}) := \int_\Gamma \boldsymbol{\mu}(\mathbf{y}) E_\kappa(\mathbf{x} - \mathbf{y})\, dS(\mathbf{y}).$$

Both potentials $\boldsymbol{\Psi}^\kappa_{SL}$ and $\boldsymbol{\Psi}^\kappa_{DL}$ are Maxwell solutions; that is, for $\boldsymbol{\mu} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$, they fulfill

$$(\mathbf{curl}\,\mathbf{curl} - \kappa^2\,\mathrm{Id})\boldsymbol{\Psi}^\kappa_{SL}(\boldsymbol{\mu}) = 0, \qquad (\mathbf{curl}\,\mathbf{curl} - \kappa^2\,\mathrm{Id})\boldsymbol{\Psi}^\kappa_{DL}(\boldsymbol{\mu}) = 0 \tag{3.4}$$

off the boundary Γ in a pointwise sense. In addition, they comply with the Silver–Müller radiation conditions.

From the well-known mapping properties of Ψ^κ_V and $\boldsymbol{\Psi}^\kappa_{\mathbf{A}}$ it is easy to get those for $\boldsymbol{\Psi}^\kappa_{SL}$ and $\boldsymbol{\Psi}^\kappa_{DL}$; see, e.g., [15, sect. 3].

THEOREM 3.3. *The following mappings are continuous:*

$$\begin{aligned}
\boldsymbol{\Psi}^\kappa_{SL} &: \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega \cup \Omega') \cap \boldsymbol{H}_{\mathrm{loc}}(\mathrm{div}\,0; \Omega \cup \Omega'),\\
\boldsymbol{\Psi}^\kappa_{DL} &: \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega \cup \Omega') \cap \boldsymbol{H}_{\mathrm{loc}}(\mathrm{div}\,0; \Omega \cup \Omega').
\end{aligned}$$

The fact that $\mathbf{curl} \circ \boldsymbol{\Psi}^\kappa_{SL} = \kappa\, \boldsymbol{\Psi}^\kappa_{DL}$ and $\mathbf{curl} \circ \boldsymbol{\Psi}^\kappa_{DL} = \kappa\, \boldsymbol{\Psi}^\kappa_{SL}$ implies

$$\gamma_N^\pm \boldsymbol{\Psi}^\kappa_{SL} = \gamma_{\mathbf{t}}^\pm \boldsymbol{\Psi}^\kappa_{DL}, \qquad \gamma_N^\pm \boldsymbol{\Psi}^\kappa_{DL} = \gamma_{\mathbf{t}}^\pm \boldsymbol{\Psi}^\kappa_{SL}. \tag{3.5}$$

This means that the following two *boundary integral operators* are sufficient for electromagnetic scattering:

$$\mathsf{S}_\kappa := \{\gamma_{\mathbf{t}}\}_\Gamma \circ \boldsymbol{\Psi}^\kappa_{SL} = \{\gamma_N\}_\Gamma \circ \boldsymbol{\Psi}^\kappa_{DL}, \qquad \mathsf{C}_\kappa := \{\gamma_{\mathbf{t}}\}_\Gamma \circ \boldsymbol{\Psi}^\kappa_{DL} = \{\gamma_N\}_\Gamma \circ \boldsymbol{\Psi}^\kappa_{SL}.$$

The continuity of S_κ and C_κ is immediate from Theorem 3.3, in conjunction with Lemma 3.2 and Theorem 2.2.

COROLLARY 3.4. *The operators* $\mathsf{S}_\kappa, \mathsf{C}_\kappa : \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ *are continuous.*

A fundamental tool for deriving boundary integral equations are *jump relations* describing the behavior of the potentials across Γ. For the Maxwell single and double layer potentials they closely resemble those for conventional single and double layer potentials for second order elliptic operators [40, Chap. 6]. For smooth domains these results are contained in [23, Thm. 6.11], [42, Thm. 5.5.1], and [45].

THEOREM 3.5. *The interior and exterior Dirichlet and Neumann traces of the potentials* $\boldsymbol{\Psi}_{SL}^\kappa$ *and* $\boldsymbol{\Psi}_{DL}^\kappa$ *are well defined and, on* $\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$*, satisfy*

$$[\gamma_{\mathbf{t}}]_\Gamma \circ \boldsymbol{\Psi}_{SL}^\kappa = [\gamma_N]_\Gamma \circ \boldsymbol{\Psi}_{DL}^\kappa = 0, \qquad [\gamma_N]_\Gamma \circ \boldsymbol{\Psi}_{SL}^\kappa = [\gamma_{\mathbf{t}}]_\Gamma \circ \boldsymbol{\Psi}_{DL}^\kappa = -\,\mathrm{Id}\,.$$

As auxiliary boundary integral operators, which supply building blocks for S_κ and C_κ, we introduce the two single layer boundary integral operators

$$\mathsf{V}_\kappa := \{\gamma\}_\Gamma \circ \Psi_V^\kappa, \qquad \mathsf{A}_\kappa := \{\gamma_{\mathbf{t}}\}_\Gamma \circ \boldsymbol{\Psi}_{\mathbf{A}}^\kappa.$$

By inspecting the potential $\boldsymbol{\Psi}_{SL}^\kappa$ and recalling $\gamma_{\mathbf{t}} \circ \mathbf{grad} = \mathbf{curl}_\Gamma \circ \gamma$, it is clear that we can write

$$\mathsf{S}_\kappa = \kappa\, \mathsf{A}_\kappa + \kappa^{-1} \mathbf{curl}_\Gamma \circ \mathsf{V}_\kappa \circ \mathrm{div}_\Gamma. \tag{3.6}$$

It is easy to see that the bilinear form associated with S_κ is given by

$$\langle \mathsf{S}_\kappa \boldsymbol{\mu}, \boldsymbol{\xi} \rangle_{\boldsymbol{\tau},\Gamma} = \frac{1}{\kappa} \langle \mathrm{div}_\Gamma \boldsymbol{\mu}, \mathsf{V}_\kappa \mathrm{div}_\Gamma \boldsymbol{\mu} \rangle_{\frac{1}{2},\Gamma} - \kappa \langle \boldsymbol{\mu}, \mathsf{A}_\kappa \boldsymbol{\xi} \rangle_{\boldsymbol{\tau},\Gamma}\,. \tag{3.7}$$

Obviously, it involves two parts of different order, neither of which is a compact perturbation of the other. In recent years a very successful approach to variational problems of this kind has emerged; see [31, sect. 5.1], [15], and [8]. The idea is to consider the above bilinear form separately on the components of a suitable splitting

$$\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) = \boldsymbol{\mathcal{X}}(\Gamma) \oplus \boldsymbol{\mathcal{N}}(\Gamma), \tag{3.8}$$

where $\boldsymbol{\mathcal{N}}(\Gamma) = \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma 0; \Gamma)$, and $\boldsymbol{\mathcal{X}}(\Gamma) \subset \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ is a closed subspace such that

1. the splitting (3.8) is direct, that is, $\boldsymbol{\mathcal{X}}(\Gamma) \cap \boldsymbol{\mathcal{N}}(\Gamma) = \emptyset$;
2. the embedding $\boldsymbol{\mathcal{X}}(\Gamma) \hookrightarrow \boldsymbol{H}_\times^{-\frac{1}{2}}(\Gamma)$ is compact.

Note that div_Γ has closed range in $\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$, and this implies that

$$\|\boldsymbol{\mu}\|_{\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \leq C \, \|\mathrm{div}_\Gamma \boldsymbol{\mu}\|_{H^{-\frac{1}{2}}(\Gamma)} \qquad \forall \boldsymbol{\mu} \in \boldsymbol{\mathcal{X}}(\Gamma).$$

By R^Γ and Z^Γ we denote the projectors onto $\boldsymbol{\mathcal{X}}(\Gamma)$ and $\boldsymbol{\mathcal{N}}(\Gamma)$, respectively, that are associated with the splitting (3.8). Examples of splittings satisfying these requirements are given by the "$\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$-orthogonal" Hodge decomposition [10] and the "projected regular splitting" [32, sect. 7].

To establish a generalized Gårding inequality for S_κ we employ the direct splitting (3.8) and two auxiliary lemmata; see [33, Lem. 3.2] and [12, Prop. 4.1].

LEMMA 3.6. *The integral operators* $\delta\mathsf{V}_\kappa := \mathsf{V}_\kappa - \mathsf{V}_0 : H^{-\frac{1}{2}}(\Gamma) \mapsto H^{\frac{1}{2}}(\Gamma)$ *and* $\delta\mathsf{A}_\kappa := \mathsf{A}_\kappa - \mathsf{A}_0 : \boldsymbol{H}_\times^{-\frac{1}{2}}(\Gamma) \mapsto \boldsymbol{H}_\times^{\frac{1}{2}}(\Gamma)$ *are compact.*

LEMMA 3.7. *The operators* V_0 *and* A_0 *are continuous, are self-adjoint with respect to the bilinear pairings* $\langle\cdot,\cdot\rangle_{\frac{1}{2},\Gamma}$ *and* $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$*, respectively, and satisfy*

$$\langle\mu, \mathsf{V}_0\overline{\mu}\rangle_{\frac{1}{2},\Gamma} \geq C\,\|\mu\|^2_{H^{-\frac{1}{2}}(\Gamma)} \qquad \forall\mu \in H^{-\frac{1}{2}}(\Gamma),$$

$$\langle\boldsymbol{\mu}, \mathsf{A}_0\overline{\boldsymbol{\mu}}\rangle_{\boldsymbol{\tau},\Gamma} \geq C\,\|\boldsymbol{\mu}\|^2_{\boldsymbol{H}_\times^{-\frac{1}{2}}(\Gamma)} \qquad \forall\boldsymbol{\mu} \in \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma 0;\Gamma),$$

with constants $C > 0$ *depending only on* Γ.

The main result will be a generalized Gårding inequality for S_κ that involves the isomorphism

$$\mathsf{X}_\Gamma = \mathsf{R}^\Gamma - \mathsf{Z}^\Gamma : \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \mapsto \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma). \tag{3.9}$$

LEMMA 3.8 (cf. [12, 33]). *There is a compact bilinear form* $c_\Gamma : \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \times \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \mapsto \mathbb{C}$ *and a constant* $C > 0$ *such that*

$$|\,\langle\mathsf{S}_\kappa\boldsymbol{\mu}, \mathsf{X}_\Gamma\overline{\boldsymbol{\mu}}\rangle_{\boldsymbol{\tau},\Gamma} + c_\Gamma(\boldsymbol{\mu},\overline{\boldsymbol{\mu}})| \geq C\,\|\boldsymbol{\mu}\|^2_{\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \qquad \forall\boldsymbol{\mu} \in \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma).$$

Remember that S_κ is the integral operator underlying the EFIE. Lemma 3.8 tells us that $\mathsf{S}_\kappa : \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \mapsto \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ is a Fredholm operator of index 0. This will ensure surjectivity as soon as injectivity holds. However, the very problem of instability at resonant frequencies is due to the failure of S_κ to be injective for certain discrete values of κ; see, e.g., [18], [42], or [15, sect. 5.2].

4. The CFIE. The CFIEs arise from an indirect approach which aims to exploit that both $\boldsymbol{\Psi}^\kappa_{SL}$ and $\boldsymbol{\Psi}^\kappa_{DL}$ yield Maxwell solutions; see (3.4). The crudest variant starts from the trial expression

$$\mathbf{e} = -i\eta\boldsymbol{\Psi}^\kappa_{SL}(\boldsymbol{\zeta}) - \boldsymbol{\Psi}^\kappa_{DL}(\boldsymbol{\zeta}), \tag{4.1}$$

with some parameter $\eta > 0$. By the jump relations, taking the exterior Dirichlet trace $\gamma^+_{\mathbf{t}}$ results in the boundary integral equation

$$-i\eta\mathsf{S}_\kappa(\boldsymbol{\zeta}) + (\tfrac{1}{2}\,\mathrm{Id} - \mathsf{C}_\kappa)(\boldsymbol{\zeta}) = \gamma^+_{\mathbf{t}}\mathbf{e}_i, \tag{4.2}$$

which is generically posed in $\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$. At least on smooth surfaces the operator $\mathsf{C}_\kappa : \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \mapsto \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ is compact [42, sect. 5.5] and a generalized Gårding inequality for the sum $-i\eta\mathsf{S}_\kappa + \frac{1}{2}\,\mathrm{Id}$ is available. However, on nonsmooth surfaces C_κ cannot be dismissed as compact perturbation.

The bottom line is that existence of solutions of (4.2) cannot be established on nonsmooth surfaces, let alone any theory about discrete approximations. This dire state led Kress to propose the introduction of a smoothing operator into (4.1) in [36]. His analysis was set in Hölder spaces, and he targeted the single layer potential $\boldsymbol{\Psi}^\kappa_{SL}$, because, working on smooth surfaces, he could rely on the compactness of C_κ.

We cannot make this assumption, but we are aware of Lemma 3.8. This means that the Fredholm operator S_κ is not the problem, but it is the innocent looking

identity Id in (4.2). Therefore, Kress's policy should be turned upside down, and regularization has to be aimed at the double layer potential $\mathbf{\Psi}^{\kappa}_{DL}$.

The crucial device for regularization is a *compact* "smoothing operator"

$$\mathsf{M} : \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma) \mapsto \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$$

that satisfies

$$\boldsymbol{\mu} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma) : \langle \mathsf{M}\boldsymbol{\mu}, \overline{\boldsymbol{\mu}} \rangle_{\boldsymbol{\tau},\Gamma} > 0 \quad \Leftrightarrow \quad \boldsymbol{\mu} \neq 0.$$

According to the strategy outlined above it will enter the contribution of the double layer potential to the representation formula: we get the trial expression

$$\mathbf{e} = -i\eta \mathbf{\Psi}^{\kappa}_{SL}(\boldsymbol{\zeta}) - \mathbf{\Psi}^{\kappa}_{DL}(\mathsf{M}\boldsymbol{\zeta}), \tag{4.3}$$

where $\boldsymbol{\zeta} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$, $\eta > 0$. By (3.4), this field is a Maxwell solution in $\Omega \cup \Omega'$. As above, the exterior Dirichlet trace applied to (4.3) results in the *new CFIE*

$$-i\eta \mathsf{S}_{\kappa}(\boldsymbol{\zeta}) + (\tfrac{1}{2}\,\mathrm{Id} - \mathsf{C}_{\kappa})(\mathsf{M}\boldsymbol{\zeta}) = \gamma_{\mathbf{t}}^{+}\mathbf{e}_i. \tag{4.4}$$

Since it is set in $\boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$, Theorem 2.3 hints at how to cast it into a variational form: find $\boldsymbol{\zeta} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$ such that for all $\boldsymbol{\mu} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$,

$$-i\eta \left\langle \mathsf{S}_{\kappa}(\boldsymbol{\zeta}), \boldsymbol{\mu} \right\rangle_{\boldsymbol{\tau},\Gamma} + \left\langle (\tfrac{1}{2}\,\mathrm{Id} - \mathsf{C}_{\kappa})(\mathsf{M}\boldsymbol{\zeta}), \boldsymbol{\mu} \right\rangle_{\boldsymbol{\tau},\Gamma} = \left\langle \gamma_{\mathbf{t}}^{+}\mathbf{e}_i, \boldsymbol{\mu} \right\rangle_{\boldsymbol{\tau},\Gamma}. \tag{4.5}$$

It shares the crucial uniqueness of solutions with other CFIEs.

THEOREM 4.1. *For all $\eta \neq 0$ and wave numbers $\kappa > 0$, the boundary integral equation* (4.5) *has a unique solution $\boldsymbol{\zeta} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$.*

Proof. To demonstrate uniqueness, we assume that $\boldsymbol{\zeta} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$ solves

$$-i\eta \mathsf{S}_{\kappa}(\boldsymbol{\zeta}) + (\tfrac{1}{2}\,\mathrm{Id} - \mathsf{C}_{\kappa})(\mathsf{M}\boldsymbol{\zeta}) = 0. \tag{4.6}$$

It is immediate from the jump relations that $\mathbf{e}$ given by (4.3) is an exterior Maxwell solution with $\gamma_{\mathbf{t}}^{+}\mathbf{e} = 0$. By their uniqueness we infer that $\mathbf{e} = 0$ in Ω'. Appealing to the jump relations from Theorem 3.5 once more, we find

$$\gamma_{\mathbf{t}}^{-}\mathbf{e} = -\mathsf{M}\boldsymbol{\zeta}, \qquad \gamma_{N}^{-}\mathbf{e} = -i\eta\boldsymbol{\zeta}.$$

Next, we use (2.4) and see that

$$i\mathbb{R} \ni i\eta \left\langle \boldsymbol{\zeta}, \overline{\mathsf{M}\boldsymbol{\zeta}} \right\rangle_{\boldsymbol{\tau},\Gamma} = \left\langle \gamma_{N}^{-}\mathbf{e}, \overline{\gamma_{\mathbf{t}}^{-}\mathbf{e}} \right\rangle_{\boldsymbol{\tau},\Gamma} = \int_{\Omega} \frac{1}{\kappa} |\,\mathbf{curl}\,\mathbf{e}|^2 \,\mathrm{d}\mathbf{x} - \kappa|\mathbf{e}|^2 \,\mathrm{d}\mathbf{x} \in \mathbb{R}.$$

Necessarily, $\left\langle \boldsymbol{\zeta}, \overline{\mathsf{M}\boldsymbol{\zeta}} \right\rangle_{\boldsymbol{\tau},\Gamma} = 0$ so that the requirements on M imply $\boldsymbol{\zeta} = 0$, which settles the issue of uniqueness.

Next, we know from Corollary 3.4 that $\mathsf{C}_{\kappa} : \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma) \mapsto \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$ is continuous so that $(\frac{1}{2}\,\mathrm{Id} - \mathsf{C}_{\kappa}) \circ \mathsf{M} : \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma) \mapsto \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_{\Gamma},\Gamma)$ turns out to be compact. Eventually, we conclude from Lemma 3.8 that the bilinear form of (4.5) satisfies a generalized Gårding inequality. Thus, a Fredholm alternative argument gives existence of a solution from its uniqueness. ∎

A simple eligible operator M can be introduced through a variational definition: for $\boldsymbol{\zeta} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ and all $\mathbf{q} \in \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$, $\mathsf{M}\boldsymbol{\zeta} \in \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$ is to satisfy

$$(\mathsf{M}\boldsymbol{\zeta},\mathbf{q})_{0;\Gamma} + (\mathrm{div}_\Gamma \mathsf{M}\boldsymbol{\zeta}, \mathrm{div}_\Gamma \mathbf{q})_{0;\Gamma} = \langle \mathbf{q}, \boldsymbol{\zeta}\rangle_{\boldsymbol{\tau},\Gamma}\,, \tag{4.7}$$

where $(\cdot,\cdot)_{0;\Gamma}$ denotes the standard $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$ scalar product. It becomes obvious that $\mathsf{M} : \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \mapsto \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$ is a continuous linear operator. To prove injectivity, let $\boldsymbol{\zeta}$ be such that $\mathsf{M}\boldsymbol{\zeta} = 0$, and let $\boldsymbol{\eta} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ be the vector verifying

$$\langle \boldsymbol{\eta}, \boldsymbol{\zeta}\rangle_{\boldsymbol{\tau},\Gamma} = \|\boldsymbol{\zeta}\|^2_{\boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}.$$

Due to Lemma 2.4, there exists a sequence $\{\boldsymbol{\eta}_\ell\}_{\ell\in\mathbb{N}} \subset \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$ converging to $\boldsymbol{\eta}$. Now choosing $\boldsymbol{\eta}_\ell$ as the test function in (4.7) and passing to the limit for $\ell \to \infty$, we obtain $\boldsymbol{\zeta} = 0$. The injectivity of M immediately implies

$$\big\langle \mathsf{M}\boldsymbol{\zeta}, \overline{\boldsymbol{\zeta}}\big\rangle_{\boldsymbol{\tau},\Gamma} = \|\mathsf{M}\boldsymbol{\zeta}\|^2_{\boldsymbol{H}_\times(\mathrm{div}_\Gamma,\Gamma)} > 0 \quad \Leftrightarrow \quad \boldsymbol{\zeta} \neq 0.$$

In addition, M inherits compactness from the embedding $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma) \hookrightarrow \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$; see Lemma 2.5: it meets all requirements listed above.

The composition of the integral operator C_κ and the smoothing operator M in (4.5) is not problematic. However, it cannot be handled in the context of Galerkin discretization, which we intend to apply; we have to find an equivalent weak form that can be discretized easily.

The usual trick to avoid operator products is to switch to a mixed formulation. Here, this amounts to introducing the new unknown $\mathbf{p} := \mathsf{M}\boldsymbol{\zeta}$. If we use the particular smoothing operator from (4.7), we get $\mathbf{p} \in \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$ and may simply incorporate (4.7) into the eventual mixed variational problem: find $\boldsymbol{\zeta} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$, $\mathbf{p} \in \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$ such that for all $\boldsymbol{\mu} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$, $\mathbf{q} \in \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$,

$$\begin{array}{rcrcl} -i\eta\,\langle \mathsf{S}_\kappa\boldsymbol{\zeta},\boldsymbol{\mu}\rangle_{\boldsymbol{\tau},\Gamma} & + & \big\langle (\tfrac{1}{2}\,\mathrm{Id} - \mathsf{C}_\kappa)\mathbf{p},\boldsymbol{\mu}\big\rangle_{\boldsymbol{\tau},\Gamma} & = & \big\langle \gamma_{\mathbf{t}}^+ \mathbf{e}_i,\boldsymbol{\mu}\big\rangle_{\boldsymbol{\tau},\Gamma}\,, \\ \langle \mathbf{q},\boldsymbol{\zeta}\rangle_{\boldsymbol{\tau},\Gamma} & - & (\mathbf{p},\mathbf{q})_{0;\Gamma} - (\mathrm{div}_\Gamma\mathbf{p},\mathrm{div}_\Gamma\mathbf{q})_{0;\Gamma} & = & 0. \end{array} \tag{4.8}$$

The next lemma tells us that we need not worry about Id in (4.8).

LEMMA 4.2. *The bilinear forms* $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$ *and* / $\langle \mathsf{C}_\kappa\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$ *are compact as mapping* $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma) \times \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \mapsto \mathbb{C}$.

Proof. It is enough to note that $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$ / $\langle \mathsf{C}_\kappa\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma} : \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)\times\boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \to \mathbb{C}$ are continuous and the injection $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma) \hookrightarrow \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ is compact due to Lemma 2.5. ∎

As an immediate consequence of this result we note that the off-diagonal terms in (4.8) represent compact bilinear forms. It remains to investigate the diagonal terms. First, $(\mathbf{p},\mathbf{q})_{0;\Gamma} + (\mathrm{div}_\Gamma\mathbf{p},\mathrm{div}_\Gamma\mathbf{q})_{0;\Gamma}$ is clearly elliptic in $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$, because it gives rise to its inner product. Second, the other bilinear form $\langle \mathsf{S}_\kappa\boldsymbol{\zeta},\boldsymbol{\mu}\rangle_{\boldsymbol{\tau},\Gamma}$ has been found to verify a generalized Gårding inequality; see Lemma 3.8.

Let us summarize what we know about the entire variational problem (4.8). For the sake of brevity we write $\mathfrak{V} := \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \times \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$ and denote by $\|\cdot\|_{\mathfrak{V}}$ its natural graph norm. We use the symbols $\mathfrak{u},\mathfrak{v},\mathfrak{w},\ldots$ for pairs of functions in $\mathfrak{V}$. Let $\widetilde{a} : \mathfrak{V}\times\mathfrak{V} \mapsto \mathbb{C}$ be the bilinear form associated with (4.8). As an immediate

consequence of the preceding considerations, it will also fulfill a generalized Gårding inequality. It can be stated using the isomorphism

$$\mathbb{X}_\Gamma : \mathfrak{V} \mapsto \mathfrak{V}, \qquad \mathbb{X}_\Gamma \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{q} \end{pmatrix} := \begin{pmatrix} \mathsf{X}_\Gamma \boldsymbol{\mu} \\ \mathbf{q} \end{pmatrix}.$$

COROLLARY 4.3. *There is a compact bilinear form* $\widetilde{c} : \mathfrak{V} \times \mathfrak{V} \mapsto \mathbb{C}$ *and a constant* $C_G > 0$ *such that*

$$|\widetilde{a}(\mathbb{X}_\Gamma \mathfrak{v}, \overline{\mathfrak{v}}) + \widetilde{c}(\mathfrak{v}, \overline{\mathfrak{v}})| \geq C_G \, \|\mathfrak{v}\|_{\mathfrak{V}}^2 \qquad \forall \mathfrak{v} \in \mathfrak{V}.$$

Since we have confirmed the uniqueness of solutions of (4.8), a Fredholm alternative argument shows that $\widetilde{a}$ induces an isomorphism, in particular that the inf-sup condition

$$\sup_{\mathfrak{v} \in \mathfrak{V}} \frac{|\widetilde{a}(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathfrak{V}}} \geq C_S \, \|\mathfrak{u}\|_{\mathfrak{V}} \tag{4.9}$$

holds with $C_S > 0$ independent of $\mathfrak{u} \in \mathfrak{V}$.

REMARK 2. *Many choices of smoothing operators* M *are conceivable. For the following reasons we opted for the definition* (4.7).

First, the operator M *is the inverse of* $-\mathbf{grad}_\Gamma \operatorname{div}_\Gamma + \mathrm{Id}$ *with Dirichlet boundary conditions on the skeleton* Σ. *We are anxious to use the inverse of a proper differential operator, because any nonlocal operator in the definition of* M *will be awkward to deal with in an implementation. We also aimed at making* M *local on each face of the polyhedron, which is satisfied by the concrete choice, since surface vector fields* $\boldsymbol{H}_\Sigma(\operatorname{div}_\Gamma, \Gamma)$ *have no flux across any edge in* Σ.

Second, we have to take great pains to ensure sufficient regularity of the solution for the new unknown $\mathbf{p}$. *If, in* (4.7), *we used* $\boldsymbol{H}_\times(\operatorname{div}_\Gamma, \Gamma)$ *trial and test function spaces instead of* $\boldsymbol{H}_\Sigma(\operatorname{div}_\Gamma, \Gamma)$, *then the regularity of* $\mathbf{p}$ *would be impaired, because Laplace–Beltrami singularities* [12, *sect.* 5.2.1] *would sneak into* $\mathbf{p}$ *through the associated smoothing operator. We are going to resume the discussion at the end of section* 6.

5. Galerkin discretization. We equip the piecewise smooth compact two-dimensional surface Γ with an oriented triangulation Γ_h. This means that all its edges are endowed with a direction. We assume a perfect resolution of Γ; that is, $\Gamma = \bar{K}_1 \cup \cdots \cup \bar{K}_N$, where $\mathcal{K}_h := \{K_1, \ldots, K_N\}$ is the set of mutually disjoint open cells of Γ_h. Moreover, no cell may straddle boundaries of the smooth faces Γ^j of Γ. We will admit triangular and quadrilateral cells only: for each $K \in \mathcal{K}_h$ there is a diffeomorphism $\boldsymbol{\Phi}_K : \widehat{K} \mapsto \bar{K}$, where $\widehat{K}$ is the "unit triangle" or unit square in $\mathbb{R}^2$, depending on the shape of K [19, sect. 5].

This paves the way for a parametric construction of boundary elements: to begin with, choose finite-dimensional local spaces $\boldsymbol{\mathcal{W}}(\widehat{K}) \subset (C^\infty(\widehat{K}))^2$ of polynomial vector fields, together with a dual basis of so-called local degrees of freedom (d.o.f.). Possible choices for $\boldsymbol{\mathcal{W}}(\widehat{K})$ and related d.o.f. abound [5, Chap. III]: we may use the classical triangular Raviart–Thomas (RT_p) elements of polynomial order $p \in \mathbb{N}_0$ [44] that use

$$\boldsymbol{\mathcal{W}}(\widehat{K}) := \{\mathbf{x} \mapsto \mathbf{p}_1(\mathbf{x}) + p_2(\mathbf{x}) \cdot \mathbf{x}, \, \mathbf{x} \in \widehat{K}, \, \mathbf{p}_1 \in (\mathcal{P}_p(\widehat{K}))^2, \, p_2 \in \mathcal{P}_p(\widehat{K})\},$$

where $\mathcal{P}_p(\widehat{K})$ is the space of two-variable polynomials of total degree $\leq p$. Possible alternatives are the triangular BDM_p elements of degree p [4], $p \in \mathbb{N}_0$, which rely

on $\boldsymbol{\mathcal{W}}(\widehat{K}) := (\mathcal{P}_{p+1}(\widehat{K}))^2$. In both cases, the usual d.o.f. involve certain polynomial moments of normal components on edges, together with interior vectorial moments for $p > 0$. For instance, in the case of RT_0, edge fluxes are the appropriate d.o.f.:

$$\boldsymbol{\mu}_h \in \boldsymbol{\mathcal{W}}(\widehat{K}) \mapsto \int_{\widehat{e}} \boldsymbol{\mu}_h \cdot \widehat{\mathbf{n}} \,\mathrm{d}S, \qquad \widehat{e} \text{ edge of } \widehat{K}.$$

Similar local spaces and d.o.f. are available for the unit square.

Using the pullback of 1-forms the local spaces can be lifted to the cells of Γ_h. In terms of vector fields this is equivalent to the *Piola transformation*

$$(\mathfrak{F}_K\boldsymbol{\mu})(\mathbf{x}) := \sqrt{\det(\mathrm{G})}\,\mathrm{G}^{-1}\, D\boldsymbol{\Phi}_K^T(\widehat{\mathbf{x}})\boldsymbol{\mu}(\widehat{\mathbf{x}}), \tag{5.1}$$

where $\mathrm{G} := D\boldsymbol{\Phi}(\widehat{\mathbf{x}})^T D\boldsymbol{\Phi}(\widehat{\mathbf{x}})$, $\mathbf{x} = \boldsymbol{\Phi}_K(\widehat{\mathbf{x}})$, $\widehat{\mathbf{x}} \in \widehat{K}$. Thus, we can introduce the global boundary element space

$$\boldsymbol{\mathcal{W}}_h := \{\boldsymbol{\mu} \in \boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma) : \boldsymbol{\mu}_{|K} \in \mathfrak{F}_K(\boldsymbol{\mathcal{W}}(\widehat{K}))\, \forall K \in \mathcal{K}_h\}. \tag{5.2}$$

In practice, $\boldsymbol{\mathcal{W}}_h \subset \boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$ is ensured by a suitable choice of d.o.f. Remember that d.o.f. have to be associated with individual edges of $\widehat{K}$ or the interior of $\widehat{K}$. It is crucial that the normal component of any $\widehat{\boldsymbol{\mu}}_h \in \boldsymbol{\mathcal{W}}(\widehat{K})$ on any edge $\widehat{e}$ of $\widehat{K}$ vanishes if and only if $\widehat{\boldsymbol{\mu}}_h$ belongs to the kernel of all local d.o.f. associated with $\widehat{e}$. In light of (2.5), this ensures $\boldsymbol{\mathcal{W}} \subset \boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$. In the rest of the paper $\boldsymbol{\mathcal{W}}_h$ will designate a generic $\boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$-conforming boundary element space. It may arise from the RT_p family of elements, $p \in \mathbb{N}_0$, the BDM_p family, or a combination of both.

Based on the d.o.f. we can introduce *local interpolation operators* $\Pi_h : \mathrm{Dom}(\Pi_h) \mapsto \boldsymbol{\mathcal{W}}_h$. They are projectors onto $\boldsymbol{\mathcal{W}}_h$ and enjoy the fundamental *commuting diagram property* [5, Chap. III, sect. 3]

$$\mathrm{div}_\Gamma \circ \Pi_h = \mathsf{Q}_h \circ \mathrm{div}_\Gamma \quad \text{on } \boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma) \cap \mathrm{Dom}(\Pi_h). \tag{5.3}$$

Here, Q_h is the $L^2(\Gamma)$-orthogonal projection onto a suitable space Q_h of Γ_h-piecewise polynomial discontinuous functions. It must be emphasized that the interpolation operators Π_h fail to be bounded on $\boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$; slightly more regularity of tangential vector fields in $\mathrm{Dom}(\Pi_h)$ is required [33, Lem. 5.1].

Next, we turn our attention to asymptotic properties of the boundary element spaces, in particular to estimates of interpolation errors and best approximation errors. We restrict ourselves to the h-version of boundary elements, which relies on uniformly shape-regular families $\{\Gamma_h\}_{h\in\mathbb{H}}$ of triangulations of Γ [22, Chap. 3, sect. 3.1]. Here, $\mathbb{H}$ stands for a decreasing sequence of meshwidths, and $\mathbb{H}$ is assumed to converge to zero.

By means of transformation to reference elements, the commuting diagram property, and Bramble–Hilbert arguments, interpolation error estimates can easily be obtained [5, Chap. III, sect. 3.3].

LEMMA 5.1 (interpolation error estimate). *For $0 < s \le p+1$ we find constants $C > 0$, depending only on the shape regularity of the meshes, s and p, such that for all $\boldsymbol{\mu} \in \boldsymbol{H}^s_\times(\Gamma) \cap \boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$, $h \in \mathbb{H}$,*

$$\|\boldsymbol{\mu} - \Pi_h\boldsymbol{\mu}\|_{\boldsymbol{L}^2(\Gamma)} \le Ch^s \left(\|\boldsymbol{\mu}\|_{\boldsymbol{H}^s_\times(\Gamma)} + \|\mathrm{div}_\Gamma\boldsymbol{\mu}\|_{L^2(\Gamma)}\right),$$

and such that for all $\boldsymbol{\mu} \in \boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$, $\mathrm{div}_\Gamma\boldsymbol{\mu} \in H^s_-(\Gamma)$,

$$\|\mathrm{div}_\Gamma(\boldsymbol{\mu} - \Pi_h\boldsymbol{\mu})\|_{L^2(\Gamma)} \le Ch^s \|\mathrm{div}_\Gamma\boldsymbol{\mu}\|_{H^s_-(\Gamma)}.$$

COROLLARY 5.2. *The union of all boundary element spaces* $\boldsymbol{\mathcal{W}}_h$, $h \in \mathbb{H}$, *is dense in* $\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$.

A particular variant of the above interpolation error estimate addresses vector fields with discrete surface divergence; cf. [33, Lem. 6.2].

LEMMA 5.3. *If* $\boldsymbol{\mu} \in \boldsymbol{H}_\times^s(\Gamma)$, $0 < s \leq 1$, *and* $\mathrm{div}_\Gamma \boldsymbol{\mu} \in Q_h$, *then*

$$\|\boldsymbol{\mu} - \Pi_h \boldsymbol{\mu}\|_{\boldsymbol{L}_\mathbf{t}^2(\Gamma)} \leq C h^s \|\boldsymbol{\mu}\|_{\boldsymbol{H}_\times^s(\Gamma)},$$

where the constant $C > 0$ *depends only on the shape regularity of the meshes and the polynomial degree* p.

From the interpolation error estimates we instantly get best approximation estimates in terms of the $\boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$-norm. Yet, what we actually need is a result about approximation in the "energy norm" (trace norm) of the form

$$\inf_{\boldsymbol{\xi}_h} \|\boldsymbol{\mu}_h - \boldsymbol{\xi}_h\|_{\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \leq C h^{s+\frac{1}{2}} \|\boldsymbol{\mu}\|_{\boldsymbol{H}_\times^s(\mathrm{div}_\Gamma,\Gamma)}. \tag{5.4}$$

The estimate in $\boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$ does not directly provide (5.4). The question of obtaining (5.4) has been addressed in [8, sect. 4.4.2], and the idea is to use the duality argument face by face (each one seen as a regular open manifold), relying on the continuity of the normal components of vector fields in $\boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$. At the end of a technical procedure we obtain the following result [8, Thm. 4.9].

THEOREM 5.4. *Let* $\mathcal{P}_h : \boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \to \boldsymbol{\mathcal{W}}_h$ *be the orthogonal projection with respect to the* $\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ *inner product. Then for any* $-\frac{1}{2} \leq s \leq p+1$ *we have*

$$\|\boldsymbol{\mu} - \mathcal{P}_h \boldsymbol{\mu}\|_{\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \leq C h^{s+\frac{1}{2}} \|\boldsymbol{\mu}\|_{\boldsymbol{H}_\times^s(\mathrm{div}_\Gamma,\Gamma)} \qquad \forall \boldsymbol{\mu} \in \boldsymbol{H}_\times^s(\mathrm{div}_\Gamma, \Gamma). \tag{5.5}$$

This theorem tells us that we can expect good approximation properties, much better than the estimates for the local interpolation error.

A finite-dimensional subspace of $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma, \Gamma)$ is easily obtained from $\boldsymbol{\mathcal{W}}_h$ by setting all d.o.f. associated with edges on the skeleton Σ to zero. Let us write $\boldsymbol{\mathcal{W}}_{\Sigma,h}$ for the resulting space. By construction the estimates of Lemma 5.1 will carry over to $\boldsymbol{H}_\times^s(\Gamma) \cap \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma, \Gamma)$ and $\boldsymbol{\mathcal{W}}_{\Sigma,h}$.

Based on the boundary element spaces $\boldsymbol{\mathcal{W}}_h$ and $\boldsymbol{\mathcal{W}}_{\Sigma,h}$, which are contained in $\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ and $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma, \Gamma)$, respectively, we pursue a standard Galerkin discretization of (4.8). Writing $\mathfrak{V}_h = \boldsymbol{\mathcal{W}}_h \times \boldsymbol{\mathcal{W}}_{\Sigma,h}$, we end up with the following discrete problem:

$$\text{Find } \mathfrak{u}_h \in \mathfrak{V}_h \ : \ \widetilde{a}(\mathfrak{u}_h, \mathfrak{v}_h) = \left\langle \begin{pmatrix} \gamma_\mathbf{t}^+ \mathbf{e} \\ 0 \end{pmatrix}, \mathfrak{v}_h \right\rangle_{\boldsymbol{\tau},\Gamma} \qquad \forall\, \mathfrak{v}_h \in \mathfrak{V}_h. \tag{5.6}$$

We aim at establishing a *uniform discrete inf-sup condition* of the following form: there exists $C_D > 0$ such that

$$\sup_{\mathfrak{v}_h \in \mathfrak{V}_h} \frac{|\widetilde{a}(\mathfrak{u}_h, \mathfrak{v}_h)|}{\|\mathfrak{v}_h\|_\mathfrak{V}} \geq C_D \|\mathfrak{u}_h\|_\mathfrak{V} \qquad \forall \mathfrak{u}_h \in \mathfrak{V}_h, \qquad h \in \mathbb{H}. \tag{5.7}$$

According to [48] this guarantees existence of discrete solutions $\mathfrak{u}_h := (\boldsymbol{\zeta}_h, \mathbf{p}_h) \in \mathfrak{V}_h$ of (5.6) and translates into their quasi-optimal behavior:

$$\|\mathfrak{u} - \mathfrak{u}_h\|_\mathfrak{V} \leq C_D^{-1} C_{\widetilde{a}} \inf_{\mathfrak{v}_h \in \mathfrak{V}_h} \|\mathfrak{u} - \mathfrak{v}_h\|_\mathfrak{V} \qquad \forall h \in \mathbb{H}, \tag{5.8}$$

where $C_{\widetilde{a}} > 0$ is the operator norm of $\widetilde{a}(\cdot,\cdot)$. We follow lines of reasoning laid out in [8, 15, 33]. As a first step towards a discrete inf-sup condition (5.7), we have to find a suitable candidate for $\mathfrak{v}$ in (4.9). To that end, introduce the operator $\mathsf{T} : \mathfrak{V} \mapsto \mathfrak{V}$ through

$$\widetilde{a}(\mathfrak{v}, \mathsf{T}\mathfrak{w}) = \widetilde{c}(\mathfrak{w}, \mathfrak{v}) \qquad \forall \mathfrak{v} \in \mathfrak{V},\, \mathfrak{w} \in \mathfrak{V},$$

where $\widetilde{c}$ is the compact bilinear form specified in Corollary 4.3. Owing to (4.9) this is a valid definition of a compact operator T. It is immediate from (4.9) and Lemma 3.8 that

$$|\widetilde{a}(\mathfrak{w}, (\mathbb{X}_\Gamma + \mathsf{T})\overline{\mathfrak{w}})| = |\widetilde{a}(\mathfrak{w}, \mathbb{X}_\Gamma\overline{\mathfrak{w}}) + c_\Gamma(\mathfrak{w}, \overline{\mathfrak{w}})| \geq C_G \, \|\mathfrak{w}\|^2_{\mathfrak{V}} \tag{5.9}$$

for all $\mathfrak{w} \in \mathfrak{V}_h$. Consequently, the choice $\mathfrak{v} := (\mathbb{X}_\Gamma + \mathsf{T})\overline{\mathfrak{w}}$ will make (4.9) hold with $C_S = C_G$.

Let $\mathfrak{w}_h \in \mathfrak{V}_h$, and $\mathfrak{v} := (\mathbb{X}_\Gamma + \mathsf{T})\mathfrak{w}_h$. In general, $\mathfrak{v} \notin \mathfrak{V}_h$ so that we have to use a projection. Write $\mathsf{P}_h : \boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma, \Gamma) \mapsto \boldsymbol{\mathcal{W}}_{\Sigma,h}$ for the $\boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$-orthogonal projection and introduce

$$\mathbb{P} : \mathfrak{V} \mapsto \mathfrak{V}_h, \qquad \mathbb{P}\begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{q}\end{pmatrix} := \begin{pmatrix}\mathcal{P}_h\boldsymbol{\mu}\\ \mathsf{P}_h\mathbf{q}\end{pmatrix}.$$

Then a promising candidate for the discrete inf-sup condition is the vector $\mathfrak{v}_h := \mathbb{P}_h\mathfrak{v} = (\mathbb{P}_h\mathbb{X}_\Gamma + \mathbb{P}_h\mathsf{T})\overline{\mathfrak{w}}_h$. The triangle inequality

$$|\widetilde{a}(\mathfrak{w}_h, \mathfrak{v}_h)| \geq |\widetilde{a}(\mathfrak{w}_h, \mathfrak{v})| - C_{\mathfrak{b}}\|\mathfrak{w}_h\|_{\mathfrak{V}}\|\mathfrak{v} - \mathfrak{v}_h\|_{\mathfrak{V}} \tag{5.10}$$

shows that (strong) convergence $\|\mathfrak{v} - \mathfrak{v}_h\|_{\mathfrak{V}} \to 0$ is needed. First, $\|(I - \mathbb{P}_h)\mathsf{T}\mathfrak{w}_h\|_{\mathfrak{V}} \to 0$ uniformly for all $\mathfrak{w}_h \in \mathfrak{V}_h$, since the composition of pointwise convergent and compact operators gives uniform convergence in operator norms [37, Cor. 10.4]. Second, it is important to note that X_Γ leaves the div_Γ of its argument function invariant, which means that the first component of $\mathbb{X}_\Gamma\mathfrak{w}_h$ has a surface divergence in a space of Γ_h-piecewise polynomials. This enables us to invoke Lemma 5.3, and we obtain (see [15, sect. 4.2]) that there exists an $s > 0$ such that

$$\|(I - \mathcal{P}_h)\mathbb{X}_\Gamma\mathfrak{w}_h\|_{\mathfrak{V}} \leq \|(\mathrm{Id} - \Pi_h)\mathsf{X}_\Gamma\boldsymbol{\mu}_h\|_{\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)} \leq Ch^s \, \|\mathrm{div}_\Gamma\mathsf{X}_\Gamma\boldsymbol{\mu}_h\|_{H^{-\frac{1}{2}}(\Gamma)},$$

where $\boldsymbol{\mu}_h \in \boldsymbol{\mathcal{W}}_h$ stands for the first component of $\mathfrak{w}_h$.

Using these estimates in (5.10) and recalling that, by definition of $\mathfrak{v}$, $|\widetilde{a}(\mathfrak{w}_h, \mathfrak{v})| \geq C_G\|\mathfrak{w}_h\|^2_{\mathfrak{V}}$, we easily deduce the following theorem.

THEOREM 5.5. *There is an $h^* > 0$, depending on the parameters of the continuous problem and the shape regularity of the triangulation, such that a unique solution $\mathfrak{u}_h \in \mathfrak{V}_h$ of the discretized problem* (5.6) *exists, provided that $h < h_*$. It supplies an asymptotically optimal approximation to the continuous solution $\mathfrak{u} = (\boldsymbol{\zeta}, \mathbf{p})$ of* (4.8) *in the sense of* (5.8).

After choosing local bases of $\boldsymbol{\mathcal{W}}_h$ and $\boldsymbol{\mathcal{W}}_{\Sigma,h}$, we end up with a linear system of equations of the form

$$\begin{pmatrix} i\eta S & \frac{1}{2}B - C\\ B^T & -D\end{pmatrix}\begin{pmatrix}\underline{\zeta}\\ \underline{\mathbf{p}}\end{pmatrix} = \begin{pmatrix}\underline{\mathbf{g}}\\ 0\end{pmatrix}. \tag{5.11}$$

Here S and C will be dense square matrices arising from the discretized boundary integral operators S_κ and C_κ. The sparse, skew-symmetric matrix B is related to

$\langle\cdot,\cdot\rangle_{\boldsymbol{\tau},\Gamma}$, whereas the s.p.d. matrix D corresponds to the $\boldsymbol{H}_\Sigma(\mathrm{div}_\Gamma,\Gamma)$-inner product. The other symbols have obvious meanings.

Note that D is even block-diagonal with one sparse block for each face Γ_j, $j = 1,\ldots,P$. Using advanced sparse Cholesky factorization techniques, it may be feasible to compute the application of D^{-1} to a vector directly. Then we face the linear system of equations

$$(i\eta S + (\tfrac{1}{2}B - C)D^{-1}B^T)\underline{\zeta} = \underline{\mathbf{g}}. \tag{5.12}$$

It can be solved only iteratively, because the actual matrix D^{-1} is not available. Besides, iterative solvers allow the use of fast summation techniques (multipole, $\mathcal{H}^2$-matrices) for the approximate application of S and C to a vector.

REMARK 3. *Of course, $\boldsymbol{\zeta}$ and $\mathbf{p}$ can be approximated in completely different boundary element spaces, as long as these are contained in $\boldsymbol{H}_\times(\mathrm{div}_\Gamma,\Gamma)$. The analysis can immediately be extended to this case.*

REMARK 4. *The iterative solution of* (5.12) *(e.g., by means of GMRES) requires a preconditioner, because the principal part of the related boundary integral operator is given by S_κ. As pointed out in* [21] *the condition number of S will deteriorate on fine meshes. Yet, the fact that S is related to the principal part also means that preconditioning needs only to target this matrix, which is the same matrix as in the Galerkin discretization of the EFIE. An elaborate preconditioning strategy has been devised in* [21].

Yet, if κ is close to a resonant frequency, S will become nearly singular, and preconditioning might suffer. This requires further investigation, which is beyond the scope of this paper.

REMARK 5. *The choice of η is another issue which has eluded theory so far. It is clear that η has a major impact on the spectral properties of the final linear system* (5.12)*, but it is not clear how to choose η to achieve good properties of the discrete problem. This situation is commonly faced with CFIE approaches. Some investigations in the case of two-dimensional acoustic scattering can be found in* [38]*; see also* [27, *sect.* 2.4.1] *and* [16].

REMARK 6. *For reasons explained in Remark* 2*, we have decided to use a localized version of M. One could argue that localization could be carried further by considering split faces. Of course, the theory will cover this, but it is important to keep in mind that the result of Theorem* 5.5 *is asymptotic in nature. The choice of M will affect the threshold h^*, and it may well be that certain choices of M will delay the onset of asymptotic convergence until unreasonably fine meshes. We acknowledge that this might also be true for our choice of M.*

6. Convergence estimates. In light of the asymptotic quasi optimality of the conforming Galerkin solutions expressed in Theorem 5.5, we have to investigate how well the solution $(\boldsymbol{\zeta},\mathbf{p})$ of (4.8) can be approximated in $\mathfrak{V}_h$. This entails knowledge about the regularity of both $\boldsymbol{\zeta}$ and $\mathbf{p}$.

Thanks to the localization of M onto the faces of Γ, studying the smoothness of $\mathbf{p}$ can chiefly rely on two-dimensional considerations.

LEMMA 6.1. *For a Lipschitz domain $\omega \subset \mathbb{R}^2$ we denote by α the maximum regularity exponent for the Laplace problem with Dirichlet or Neumann boundary conditions; i.e., if $\Delta u \in H^{\alpha-1}(\omega)$ and u verifies either the Dirichlet or Neumann homogeneous boundary condition, then $u \in H^{\alpha+1}(\omega)$.*

Let $\mathbf{f} \in (H^\sigma(\omega))^2$ and $\mathrm{curl}_{2D}\,\mathbf{f} \in H^\sigma(\omega)$, $\sigma \geq 0$. If $\mathbf{p} \in \boldsymbol{H}_0(\mathrm{div};\omega)$ satisfies

$$(\mathrm{div}\,\mathbf{p},\mathrm{div}\,\mathbf{v})_{0;\omega} + (\mathbf{p},\mathbf{v})_{0;\omega} = (\mathbf{f},\mathbf{v})_{0;\omega} \qquad \forall \mathbf{v} \in \boldsymbol{H}_0(\mathrm{div};\omega),$$

then $\mathbf{p} \in H^{\min\{\alpha,\sigma+1\}}(\omega)$ *and* $\operatorname{div}\mathbf{p} \in H^{\min\{\alpha+1,1+\sigma\}}(\omega)$.

Proof. It goes without saying that $\mathbf{p}$ is well defined. The main tool for the proof of the asserted regularity properties will be $\boldsymbol{L}^2(\omega)$-orthogonal Helmholtz decompositions (see [28, Chap. 1])

$$\begin{aligned}\boldsymbol{L}^2(\omega) &= \mathbf{curl}_{2D}\, H^1_0(\omega) \oplus \mathbf{grad}\, H^1(\omega),\\ \boldsymbol{H}_0(\operatorname{div};\omega) &= \mathbf{curl}_{2D}\, H^1_0(\omega) \oplus \mathbf{grad}\, H_0(\Delta,\omega),\end{aligned}$$

where

$$H_0(\Delta,\omega) := \left\{\psi \in H^1(\omega):\ \Delta\psi \in L^2(\omega),\ \frac{\partial\psi}{\partial\mathbf{n}} = 0 \text{ on } \partial\omega\right\}.$$

Accordingly, we decompose

$$\mathbf{p} = \mathbf{curl}_{2D}\,\varphi_1 + \mathbf{grad}\,\varphi_2, \qquad \mathbf{f} = \mathbf{curl}_{2D}\,\phi_1 + \mathbf{grad}\,\phi_2,$$

with $\varphi_1,\phi_1 \in H^1_0(\omega)$, $\varphi_2 \in H_0(\Delta,\omega)$, $\phi_2 \in H^1(\omega)$. A closer scrutiny reveals that

$$\operatorname{curl}_{2D}\mathbf{curl}_{2D}\,\phi_1 = -\Delta\phi_1 = \operatorname{curl}_{2D}\mathbf{f} \in H^\sigma(\omega) \quad\Rightarrow\quad \phi_1 \in H^{\min\{1+\alpha,2+\sigma\}}(\omega),$$

because of the $1+\alpha$-regularity of the Laplacian. Testing with $\mathbf{curl}_{2D}\,\nu$, $\nu \in H^1_0(\omega)$, in the definition of $\mathbf{p}$, we immediately see that $\varphi_1 = \phi_1$.

For $\nu_2 \in H_0(\Delta,\omega)$ we deduce from the variational equation that

$$(\operatorname{div}\mathbf{p}, \operatorname{div}\mathbf{grad}\,\nu_2)_{0;\omega} + (\mathbf{p}, \mathbf{grad}\,\nu_2)_{0;\omega} = (\mathbf{f}, \mathbf{grad}\,\nu_2)_{0;\omega}.$$

After integrating by parts, this means

$$(\operatorname{div}\mathbf{p}, \Delta\nu_2 - \nu_2)_{0;\omega} = (\mathbf{f}, \mathbf{grad}\,\nu_2)_{0;\omega}. \tag{6.1}$$

Now consider $\zeta \in H^1(\omega)$, solving

$$(\mathbf{grad}\,\zeta, \mathbf{grad}\,\nu)_{0;\omega} + (\zeta,\nu)_{0;\omega} = (\mathbf{f}, \mathbf{grad}\,\nu)_{0;\omega} \qquad \forall\nu \in H^1(\omega). \tag{6.2}$$

The regularity assumption implies that $\zeta \in H^{\min\{1+\alpha,1+\sigma\}}(\omega)$. We can pick $\nu \in H_0(\Delta,\omega)$ in this equation, carry out integration by parts, and subtract the result from (6.1). We end up with

$$(\operatorname{div}\mathbf{p} - \zeta, -\Delta\nu + \nu)_{0;\omega} = 0.$$

Since $(-\Delta + Id)(H_0(\Delta,\omega)) = L^2(\omega)$, we infer that $\operatorname{div}\mathbf{p} = \zeta$, i.e., $\operatorname{div}\mathbf{p} \in H^{\min\{1+\alpha,1+\sigma\}}(\omega)$. ∎

We point out that for a polygon ω the exponent α is directly related to the angles θ_i, $i = 1,\ldots,n_c$, at its corners:

$$\alpha = \min\{1, \pi/\theta_i, i = 1,\ldots,n_c\} \geq \tfrac{1}{2}.$$

This lemma can instantly be applied to all the smooth faces of Γ and supplies lifting properties of M, because there is no coupling between the faces.

COROLLARY 6.2. *If* $\boldsymbol{\mu} \in \boldsymbol{H}^\sigma_\times(\operatorname{div}_\Gamma,\Gamma)$, $\sigma \geq 0$, *then* $\mathsf{M}\boldsymbol{\mu} \in \boldsymbol{H}^{\min\{\alpha,\sigma+1\}}_\times(\operatorname{div}_\Gamma,\Gamma)$, *where* α *is the minimum of the* $\Delta_{\mathrm{Dir}}/\Delta_{\mathrm{Neu}}$*-regularity exponents on the flat faces* Γ_j, $j = 1,\ldots,P$.

Assume that $\boldsymbol{\zeta} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ is the unique solution of (4.5), and denote by $\mathbf{e} \in \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega \cup \Omega')$ the Maxwell solution according to (4.3):

$$\mathbf{e} = -i\eta \boldsymbol{\Psi}_{SL}^{\kappa}(\boldsymbol{\zeta}) - \boldsymbol{\Psi}_{DL}^{\kappa}(\mathsf{M}\boldsymbol{\zeta}).$$

To study the regularity it is essential to recall that by the jump relations

$$\gamma_{\mathbf{t}}^{-}\mathbf{e} = -\mathsf{M}\boldsymbol{\zeta} - \mathbf{g}, \qquad \gamma_{N}^{-}\mathbf{e} = i\eta\boldsymbol{\zeta} - \mathbf{h}, \tag{6.3}$$

where we wrote $\mathbf{h} := \gamma_N^+\mathbf{e} \in \boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ for the exterior Neumann data of the scattered field. As $\mathbf{g} := \gamma_{\mathbf{t}}^{+}\mathbf{e}_i$ is the tangential trace of an incident wave, it will belong to $\boldsymbol{H}_{\times}^{s}(\mathrm{div}_\Gamma, \Gamma)$ for all $s > 0$. Additional information can be gleaned only from lifting properties of the Maxwell operator. Its regularity theory, elaborated in [25], justifies the following assumption.

ASSUMPTION 6.2.1. *There are two regularity indices $\sigma^-, \sigma^+ > \frac{1}{2}$ such that*

1. *any field $\mathbf{u} \in \boldsymbol{H}(\mathbf{curl}^2, \Omega)$ solving*

$$\mathbf{curl}\,\mathbf{curl}\,\mathbf{u} - \mathbf{grad}\,\mathrm{div}\,\mathbf{u} - \kappa^2\mathbf{u} = \mathbf{f} \quad \text{in } \Omega, \quad \gamma_{\mathbf{t}}^{-}\mathbf{u} = 0, \quad \text{or} \quad \gamma_N^{-}\mathbf{u} = 0$$

   belongs to $\boldsymbol{H}^{\sigma}(\mathbf{curl}, \Omega)$ for all $\sigma \le \sigma^-$ if $\mathbf{f} \in H^{\sigma-1}(\Omega)$;
2. *any field $\mathbf{u} \in \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}^2, \Omega')$ satisfying the radiation condition and*

$$\mathbf{curl}\,\mathbf{curl}\,\mathbf{u} - \mathbf{grad}\,\mathrm{div}\,\mathbf{u} - \kappa^2\mathbf{u} = \mathbf{f} \quad \text{in } \Omega', \quad \gamma_{\mathbf{t}}^{+}\mathbf{u} = 0, \quad \text{or} \quad \gamma_N^{+}\mathbf{u} = 0$$

   lies in $\boldsymbol{H}_{\mathrm{loc}}^{\sigma}(\mathbf{curl}, \Omega')$ for all $\sigma < \sigma^+$ if $\mathbf{f} \in H^{\sigma-1}(\Omega')$.

Owing to the trace theorem, Theorem 2.6, this assumption implies that $\mathbf{h} \in \boldsymbol{H}_{\times}^{\sigma^+ - \frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$. Then we can resort to a "bootstrap argument."

Step 1. We remember a result by Costabel [24] confirming the existence of a constant $c > 0$ that depends only on Ω such that for all $\mathbf{u} \in \boldsymbol{H}(\mathrm{div}; \Omega) \cap \boldsymbol{H}(\mathbf{curl}; \Omega)$,

$$\left\|\gamma_{\mathbf{n}}^{-}\mathbf{u}\right\|_{L^2(\Gamma)} \le c\left\{\left\|\gamma_{\mathbf{t}}^{-}\mathbf{u}\right\|_{\boldsymbol{L}^2(\Gamma)} + \|\mathbf{u}\|_{\boldsymbol{L}^2(\Omega)} + \|\mathbf{curl}\,\mathbf{u}\|_{\boldsymbol{L}^2(\Omega)} + \|\mathrm{div}\,\mathbf{u}\|_{L^2(\Omega)}\right\}, \tag{6.4}$$

$$\left\|\gamma_{\mathbf{t}}^{-}\mathbf{u}\right\|_{\boldsymbol{L}^2(\Gamma)} \le c\left\{\left\|\gamma_{\mathbf{n}}^{-}\mathbf{u}\right\|_{L^2(\Gamma)} + \|\mathbf{u}\|_{\boldsymbol{L}^2(\Omega)} + \|\mathbf{curl}\,\mathbf{u}\|_{\boldsymbol{L}^2(\Omega)} + \|\mathrm{div}\,\mathbf{u}\|_{L^2(\Omega)}\right\}. \tag{6.5}$$

Since $\mathbf{e}$ is a Maxwell solution in Ω, these estimates combined with Theorem 3.3 give

$$\begin{aligned}\left\|\gamma_N^{-}\mathbf{e}\right\|_{\boldsymbol{L}^2(\Gamma)} &\le C\left\{\left\|\gamma_{\mathbf{n}}^{-}\,\mathbf{curl}\,\mathbf{e}\right\|_{L^2(\Gamma)} + \kappa^2\left\|\mathbf{e}\right\|_{\boldsymbol{L}^2(\Omega)} + \|\mathbf{curl}\,\mathbf{e}\|_{\boldsymbol{L}^2(\Omega)}\right\}\\ &\le C\left\{\left\|\mathrm{div}_\Gamma\gamma_{\mathbf{t}}^{-}\mathbf{e}\right\|_{L^2(\Gamma)} + \|\boldsymbol{\zeta}\|_{\boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} + \|\mathsf{M}\boldsymbol{\zeta}\|_{\boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}\right\}.\end{aligned}$$

Similarly, we can use (6.4) and get

$$\begin{aligned}\left\|\mathrm{div}_\Gamma\gamma_N^{-}\mathbf{e}\right\|_{L^2(\Gamma)} = \kappa^2\left\|\gamma_{\mathbf{n}}^{-}\mathbf{e}\right\|_{L^2(\Gamma)} &\le C\left\{\left\|\gamma_{\mathbf{t}}^{-}\mathbf{e}\right\|_{\boldsymbol{L}^2(\Gamma)} + \|\mathbf{curl}\,\mathbf{e}\|_{\boldsymbol{L}^2(\Omega)} + \|\mathbf{e}\|_{\boldsymbol{L}^2(\Omega)}\right\}\\ &\le C\left\{\left\|\gamma_{\mathbf{t}}^{-}\mathbf{e}\right\|_{\boldsymbol{L}^2(\Gamma)} + \|\boldsymbol{\zeta}\|_{\boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} + \|\mathsf{M}\boldsymbol{\zeta}\|_{\boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}\right\}.\end{aligned}$$

The generic constants $C > 0$ may depend on Ω, κ, and η. The combined estimate reads

$$\left\|\gamma_N^{-}\mathbf{e}\right\|_{\boldsymbol{H}_{\times}(\mathrm{div}_\Gamma,\Gamma)} \le C\left\{\left\|\gamma_{\mathbf{t}}^{-}\mathbf{e}\right\|_{\boldsymbol{H}_{\times}(\mathrm{div}_\Gamma,\Gamma)} + \|\boldsymbol{\zeta}\|_{\boldsymbol{H}_{\times}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}\right\},$$

which means $\gamma_N^- \mathbf{e} \in \boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$.

Step 2. Next, from (6.3) we infer that $\boldsymbol{\zeta} \in \boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$. Now, we can apply Corollary 6.2, we get $\mathsf{M}\boldsymbol{\zeta} \in \boldsymbol{H}_\times^{\min\{1,\alpha\}}(\mathrm{div}_\Gamma, \Gamma)$, and (6.3) gives us $\gamma_{\mathbf{t}}^- \mathbf{e} \in \boldsymbol{H}_\times^{\min\{1,\alpha\}}(\mathrm{div}_\Gamma, \Gamma)$. Now, since $\mathbf{e}$ is a Maxwell solution verifying $\gamma_{\mathbf{t}}^- \mathbf{e} \in \boldsymbol{H}_\times^{\min\{1,\alpha\}}(\mathrm{div}_\Gamma, \Gamma)$, using Theorem 2.6 and Assumption 6.2.1, we have that $\mathbf{e}, \mathbf{curl}\,\mathbf{e} \in \boldsymbol{H}^{\min\{\sigma^-,1\}}(\mathbf{curl}, \Omega)$, i.e., $\gamma_N^- \mathbf{e} \in \boldsymbol{H}_\times^{\min\{\sigma^- - \frac{1}{2}, \frac{1}{2}\}}(\mathrm{div}_\Gamma, \Gamma)$.

Step 3. Finally, we can conclude that $\boldsymbol{\zeta} \in \boldsymbol{H}_\times^{\min\{\sigma^- - \frac{1}{2}, \sigma^+ - \frac{1}{2}, \frac{1}{2}\}}(\mathrm{div}_\Gamma, \Gamma)$. On a polyhedron we can take for granted that either $\sigma^- < 1$ or $\sigma^+ < 1$. This gives us

$$\boldsymbol{\zeta} \in \boldsymbol{H}_\times^{\min\{\sigma^-, \sigma^+\} - \frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma).$$

Besides, we have already seen that $\mathbf{p} = \mathsf{M}\boldsymbol{\zeta} \in \boldsymbol{H}_\times^{\min\{\alpha,1\}}(\mathrm{div}_\Gamma, \Gamma)$.

Now we can employ the best approximation estimates for div_Γ-conforming elements from Lemma 5.1 and Theorem 5.4 and get quantitative asymptotic convergence estimates.

THEOREM 6.3. *If we rely on $\boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)$-conforming boundary elements for the discretization of both $\boldsymbol{\zeta}$ and $\mathbf{p}$, we are guaranteed to get*

$$\|\boldsymbol{\zeta} - \boldsymbol{\zeta}_h\|_{\boldsymbol{H}_\times^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)} + \|\mathbf{p} - \mathbf{p}_h\|_{\boldsymbol{H}_\times(\mathrm{div}_\Gamma, \Gamma)} \leq C(h^{\min\{\sigma^+, \sigma^-\}} + h^{\min\{\alpha,1\}}),$$

with a constant $C > 0$ independent of the meshwidth h, but may depend upon the wave number k.

REMARK 7. *Since we are solving an indirect boundary integral equation, it is not surprising that the convergence is limited by singularities of both the interior and the exterior Maxwell problem. On the other hand, the main observation is that one can always have $\alpha > 1$ since it is enough to split nonconvex faces into convex ones. Thus, the rate of convergence is not affected by the introduction of the auxiliary unknown $\mathbf{p}$; i.e., $\mathbf{p}$ is always much more regular than the primal unknown $\boldsymbol{\zeta}$.*

REMARK 8. *The above estimate relies on global regularity of the exact solutions. However, we know that $\boldsymbol{\zeta}$ is a combination of traces of Maxwell solutions. Besides, $\mathbf{p}$ emerges as patched-together solutions of Dirichlet boundary value problems for Δ_Γ on the flat faces. In both cases, results on singularities of solutions of boundary value problems on nonsmooth domains reveal much detail about the behavior of $\boldsymbol{\zeta}$ and $\mathbf{p}$ close to edges and corners. We can make use of this knowledge in order to obtain significantly faster convergence on meshes that feature algebraically graded refinement towards the edges of Γ [2, 11]. In this case, making use again of the regularity of $\mathbf{p}$, one might need only to "resolve" the singularities of $\boldsymbol{\zeta}$ by mesh grading. The use of different meshes, on which $\boldsymbol{\zeta}$ and $\mathbf{p}$ are approximated, seems to be advisable in this case; cf. Remark* 3.

REFERENCES

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

[2] T. APEL, A.-M. SÄNDIG, AND J. WHITEMAN, *Graded mesh refinement and error estimates for finite element solutions of elliptic boundary value problems in non-smooth domains*, Math. Methods Appl. Sci., 19 (1996), pp. 63–85.

[3] H. BRAKHAGE AND P. WERNER, *Über das Dirichletsche Außenraumproblem für die Helmholtzsche schwingungsgleichung*, Arch. Math., 16 (1965), pp. 325–329.

[4] F. BREZZI, J. DOUGLAS, AND D. MARINI, *Two families of mixed finite elements for 2nd order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[5] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.
[6] A. Buffa, *Hodge decompositions on the boundary of a polyhedron: The multiconnected case*, Math. Models Meth. Appl. Sci., 11 (2001), pp. 1491–1504.
[7] A. Buffa, *Traces theorems on non-smmoth boundaries for functional spaces related to Maxwell equations: An overview*, in Computational Electromagnetics, Lecture Notes Comput. Sci. Engrg. 28, C. Carstensen, S. Funken, W. Hackbusch, R. Hoppe, and P. Monk, eds., Springer, Berlin, 2003, pp. 23–34.
[8] A. Buffa and S. Christiansen, *The electric field integral equation on Lipschitz screens: Definition and numerical approximation*, Numer. Math., 94 (2003), pp. 229–267.
[9] A. Buffa and P. Ciarlet, *On traces for functional spaces related to Maxwell's equations. Part* I: *An integration by parts formula in Lipschitz polyhedra.*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.
[10] A. Buffa and P. Ciarlet, *On traces for functional spaces related to Maxwell's equations. Part* II: *Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 24 (2001), pp. 31–48.
[11] A. Buffa, M. Costabel, and M. Dauge, *Algebraic Convergence for Edge Elements in Polyhedral Domains*, Tech. report 28-PV, IMATI-CNR, Pavia, Italy, 2003.
[12] A. Buffa, M. Costabel, and C. Schwab, *Boundary element methods for Maxwell's equations on non-smooth domains*, Numer. Math., 92 (2002), pp. 679–710.
[13] A. Buffa, M. Costabel, and D. Sheen, *On traces for* $\mathbf{H}(\mathbf{curl}, \Omega)$ *in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.
[14] A. Buffa and R. Hiptmair, *Regularized Combined Field Integral Equations*, Tech. report 2003-06, SAM, ETH Zürich, Zürich, Switzerland, 2003.
[15] A. Buffa, R. Hiptmair, T. von Petersdorff, and C. Schwab, *Boundary element methods for Maxwell equations on Lipschitz domains*, Numer. Math., 95 (2003), pp. 459–485.
[16] A. Buffa and S. Sauter, *Stabilization of the Acoustic Single Layer Potential on Non-smooth Domains*, Tech. report 19-03, Institut für Mathematik, Universität Zürich, Zürich, Switzerland, 2003.
[17] A. Burton and G. Miller, *The application of integral methods for the numerical solution of boundary value problems*, Proc. Roy. Soc. London Ser. A, 323 (1971), pp. 201–210.
[18] M. Cessenat, *Mathematical Methods in Electromagnetism*, Ser. Adv. Math. Appl. Sci. 41, World Scientific, Singapore, 1996.
[19] S. Christiansen, *Mixed Boundary Element Method for Eddy Current Problems*, Research report 2002-16, SAM, ETH Zürich, Zürich, Switzerland, 2002.
[20] S. Christiansen, *Discrete Fredholm properties and convergence estimates for the electric field integral equation*, Math. Comp., 73 (2004), pp. 143–167.
[21] S. H. Christiansen and J.-C. Nédélec, *A preconditioner for the electric field integral equation based on Calderon formulas*, SIAM J. Numer. Anal., 40 (2002), pp. 1100–1135.
[22] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.
[23] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer, Heidelberg, 1998.
[24] M. Costabel, *A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains*, Math. Methods Appl. Sci., 12 (1990), pp. 365–368.
[25] M. Costabel and M. Dauge, *Singularities of Maxwell's equations on polyhedral domains*, in Analysis, Numerics and Applications of Differential and Integral Equations, Pitman Res. Notes Math. Ser. 379, M. Bach, ed., Longman, Harlow, 1998, pp. 69–76.
[26] G. Fairweather and A. Karageorghis, *The method of fundamental solutions for elliptic boundary value problems*, Adv. Comput. Math., 9 (1998), pp. 69–95.
[27] K. Giebermann, *Schnelle Summationsverfahren zur numerischen Lösung von Integralgleichungen für Streuprobleme im* $\mathbb{R}^3$ *(Fast Summation Methods for the Numerical Solution of Integral Equations for Scattering Problems in* $\mathbb{R}^3$*)*, Ph.D. thesis, Fakultät für Mathematik, Universtität Karlsruhe, Karlsruhe, Germany, 1997.
[28] V. Girault and P. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer, Berlin, 1986.
[29] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[30] C. Hazard and M. Lenoir, *On the solution of time-harmonic scattering problems for Maxwell's equations*, SIAM J. Math. Anal., 27 (1996), pp. 1597–1630.
[31] R. Hiptmair, *Finite elements in computational electromagnetism*, in Acta Numerica, Cambridge University Press, Cambridge, UK, 2002, pp. 237–339.
[32] R. Hiptmair, *Coupling of finite elements and boundary elements in electromagnetic scattering*,

SIAM J. Numer. Anal., 41 (2003), pp. 919–944.
[33] R. Hiptmair and C. Schwab, *Natural boundary element methods for the electric field integral equation on polyhedra*, SIAM J. Numer. Anal., 40 (2002), pp. 66–86.
[34] G. Hsiao, *Mathematical foundations for the boundary field equation methods in acoustic and electromagnetic scattering*, in Analytical and Computational Methods in Scattering and Applied Mathematics. A Volume in the Memory of Ralph Ellis Kleinman, Chapman Hall/CRC Res. Notes Math. 417, F. Santosa and I. Stakgold, eds., Chapman and Hall/CRC, Boca Raton, FL, 2000, pp. 149–163.
[35] D. Jerison and C. Kenig, *The inhomogeneous Dirichlet problem in Lipschitz domains*, J. Funct. Anal., 130 (1995), pp. 161–219.
[36] R. Kress, *On the boundary operator in electromagnetic scattering*, Proc. Roy. Soc. Edinburgh Sect. A, 103 (1986), pp. 91–98.
[37] R. Kress, *Linear Integral Equations*, Appl. Math. Sci. 82, Springer, Berlin, 1989.
[38] R. Kress and W. T. Spassov, *On the condition number of boundary integral operators for the exterior Dirichlet problem for the Helmholtz equation*, Numer. Math., 42 (1983), pp. 77–95.
[39] J. Lions and F. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, Springer, Berlin, 1972.
[40] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[41] J. Nečăs, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.
[42] J.-C. Nédélec, *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*, Appl. Math. Sci. 144, Springer, Berlin, 2001.
[43] O. Panich, *On the question of the solvability of the exterior boundary-value problems for the wave equation and maxwell's equations*, Uspekhi Mat. Nauk., 20 (1965), pp. 221–226 (in Russian).
[44] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, Springer, New York, 1977, pp. 292–315.
[45] M. Reissel, *On a transmission boundary-value problem for the time-harmonic Maxwell equations without displacement currents*, SIAM J. Math. Anal., 24 (1993), pp. 1440–1457.
[46] J. Stratton and L. Chu, *Diffraction theory of electromagnetic waves*, Phys. Rev., 56 (1939), pp. 99–107.
[47] D. Wilton, *Review of current status and trends in the use of integral equations in computational electromagnetics*, Electromagnetics, 12 (1992), pp. 287–341.
[48] J. Xu and L. Zikatanov, *Some observations on Babuška and Brezzi theories*, Numer. Math., 94 (2003), pp. 195–202.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 641–666

ERROR ESTIMATES TO SMOOTH SOLUTIONS OF RUNGE–KUTTA DISCONTINUOUS GALERKIN METHODS FOR SCALAR CONSERVATION LAWS*

QIANG ZHANG† AND CHI-WANG SHU‡

Abstract. In this paper we study the error estimates to sufficiently smooth solutions of scalar conservation laws for Runge–Kutta discontinuous Galerkin (RKDG) methods, where the time discretization is the second order explicit total variation diminishing (TVD) Runge–Kutta method. Error estimates for the $\mathbb{P}^1$ (piecewise linear) elements are obtained under the usual CFL condition $\tau \le \gamma h$ for general nonlinear conservation laws in one dimension and for linear conservation laws in multiple space dimensions, where h and τ are the maximum element lengths and time steps, respectively, and the positive constant γ is independent of h and τ. However, error estimates for higher order $\mathbb{P}^k (k \ge 2)$ elements need a more restrictive time step $\tau \le \gamma h^{4/3}$. We remark that this stronger condition is indeed necessary, as the method is linearly unstable under the usual CFL condition $\tau \le \gamma h$ for the $\mathbb{P}^k$ elements of degree $k \ge 2$. Error estimates of $O(h^{k+1/2} + \tau^2)$ are obtained for general monotone numerical fluxes, and optimal error estimates of $O(h^{k+1} + \tau^2)$ are obtained for upwind numerical fluxes.

Key words. discontinuous Galerkin, finite element, total variation diminishing Runge–Kutta method, error estimates

AMS subject classification. 65N12

DOI. 10.1137/S0036142902404182

1. Introduction.

In this paper, we present error estimates for the Runge–Kutta discontinuous Galerkin (RKDG) methods with smooth solutions of scalar conservation laws:

$$\partial_t u + \sum_{i=1}^{d} \partial_{x_i} f_i(u) = 0, \quad (x,t) \in \Omega \times (0, T^\star), \tag{1.1a}$$

$$u(t=0) = u_0, \quad x \in \Omega; \tag{1.1b}$$

here $x = (x_1, \ldots, x_d)$ and $f(u) = (f_1(u), \ldots, f_d(u))$. We do not pay attention to boundary conditions in this paper; hence the solution is considered to be either periodic or compactly supported. For simplicity of presentation, in most cases we will only give detailed analysis for the one-dimensional case; i.e., $\Omega = I = (0,1)$ is the unit interval. We will, however, point out any differences, both in the analysis and in the results, for the multidimensional cases. We assume that the flux $f(u)$ is smooth in the variable u; for example, $f \in C^3$ is enough. The analysis in this paper is for *smooth* solutions of (1.1). Discontinuous solutions with shocks are not considered.

The so-called RKDG method is introduced and developed by Cockburn and Shu [9, 8, 10], Cockburn, Lin, and Shu [7], and Cockburn, Hou, and Shu [6] for nonlinear

*Received by the editors March 16, 2002; accepted for publication (in revised form) October 22, 2003; published electronically May 17, 2004.
http://www.siam.org/journals/sinum/42-2/40418.html

†Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, China. Current address: School of Mathematical Science, Nankai University, Tianjing, China (qzh@nankai.edu.cn). The research of this author was supported by CNNSF grant 10028103.

‡Division of Applied Mathematics, Brown University, Providence, RI 02912 (shu@cfm.brown.edu). The research of this author was supported by CNNSF grant 10028103 while he was in residence at the University of Science and Technology of China, ARO grant DAAD19-00-1-0405, and NSF grants DMS-9804985 and DMS-0207451.

hyperbolic conservation laws, which uses DG discretization in space and combines it with an explicit total variation diminishing (TVD) Runge–Kutta time-marching algorithm [21]. In this paper, we concentrate on a second order TVD Runge–Kutta time discretization given in [21]. In practice, the third order TVD Runge–Kutta time discretization given in [21] is more popular, because of its higher order accuracy in time and better linear stability properties. However, the extension of the results in this paper to this third order TVD Runge–Kutta time discretization case is highly nontrivial, and this is still being investigated as an ongoing project.

In this paper, with the second order TVD Runge–Kutta time discretization and for $\mathbb{P}^k$ element space of piecewise kth degree polynomials, the a priori estimate in the usual $L^2(I)$-norm of the form

$$||u(t^n) - u_h^n||_{L^2(I)} \le C(h^{k+\sigma} + \tau^2), \quad \text{for any } t^n = n\tau \le T^\star, \tag{1.2}$$

is obtained, with $\sigma = 1/2$ for arbitrary monotone numerical fluxes and $\sigma = 1$ for the one-dimensional nonlinear case and for multidimensional tensor product ($\mathbb{P}^k$ replaced by $\mathbb{Q}^k$, the tensor product of one-dimensional degree k piecewise polynomials, in rectangular elements) linear cases when upwind numerical fluxes are used in the scheme. Here u_h^n denotes the numerical solution at the nth time level. The error estimates (1.2) for the $\mathbb{P}^1$ (piecewise linear) elements are obtained under the usual CFL condition $\tau \le \gamma h$ for general nonlinear conservation laws in one dimension and for linear conservation laws in multiple space dimensions, where h and τ are the maximum element lengths and time steps, respectively, and the positive constant γ is independent of h and τ. For the $\mathbb{P}^k$ elements with degree $k \ge 2$, a more restrictive time-step condition $\tau \le \gamma h^{4/3}$ is needed for the error estimates. We remark that this stronger condition is indeed necessary as the method is linearly unstable under the usual CFL condition $\tau \le \gamma h$ for the $\mathbb{P}^k$ elements of degree $k \ge 2$; see [12]. Here the positive constant C in (1.2) is independent of n, h, τ, and the numerical solution.

We now mention related results on error estimates of the DG methods in the literature. For smooth solutions of linear conservation laws, optimal a priori error estimates ($O(h^{k+1})$ for one-dimensional and some multidimensional cases and $O(h^{k+\frac{1}{2}})$ for other cases) have been given for the steady state solution or for a fully DG discretization by using space-time finite element spaces in, e.g., [17], [20], and [15], with the optimality in the general cases being proven in [19], and for the semidiscrete (continuous in time) DG method in [11]. For nonsmooth solutions of nonlinear conservation laws, Jiang and Shu [14] have proven a cell entropy inequality for the semidiscrete DG method for the square entropy, which implies that the numerical solutions, *if convergent*, will converge to an entropy solution of (1.1). Also, for nonsmooth solutions of nonlinear conservation laws, Cockburn et al. have proven an error estimate of order $O(h^{1/4})$ in the $L^1(\Omega)$-norm for the $\mathbb{P}^0$ (piecewise constant) finite element space, which is then a monotone finite volume scheme, and for higher order $\mathbb{P}^k$ elements but with additional "shock capturing" terms added to the method; see, e.g., [4]. Their result can be improved to $O(h^{1/2})$ if a uniform upper bound for the total variation can be found, which is the case for the explicit Lax–Friedrichs scheme defined on equilateral triangles [5]. Recently, there have also been works on the a posteriori error estimates; for example, see Adjerid et al. [1]. We have not attempted to address a posteriori error estimates in this paper.

We note that, unlike the semidiscrete or space-time DG methods, to the best of our knowledge there have been no high order error estimates for the RKDG methods in the literature up to now, even for linear problems. The results in this paper are thus

the first attempt to obtain optimal error estimates for the smooth solutions of RKDG methods. The main difficulty of the error estimates is the hybriding of a finite element spatial discretization with an explicit Runge–Kutta time stepping. For finite difference methods, such error estimates are typically obtained by combining local truncation error estimates with stability for linear PDEs and by using Strang's technique [22], which uses linear stability plus dissipation of the scheme and the smoothness of the solution to obtain error estimates for nonlinear PDEs. Unfortunately, such techniques rely heavily on the local truncation error being the same order as the global error, and hence they cannot be easily applied here to obtain optimal error estimates, because the DG spatial discretization, when viewed as a finite difference approximation, has a local truncation error which is lower order than the global error, which is the so-called supraconvergence phenomenon [16, 24]. The main techniques we use in this paper are Taylor expansion and energy estimates as in [23], which considers linear continuous finite elements and obtains error estimates under the rather restrictive time-step condition $\tau \le \gamma h^{4/3}$. We consider DG rather than continuous finite elements and are able to remove this restriction and replace it by the more natural condition $\tau \le \gamma h$ for linear finite elements. We would like to mention that the DG method is more complicated to analyze than the continuous finite element methods in this context, due to the extra interelement boundary terms. We would also like to mention in particular that the a priori assumption about the numerical solution is subtle, which plays an important role in the proof for the nonlinear cases; see section 5.

An overview of this paper is as follows. In section 2 we present the RKDG method with the second order TVD Runge–Kutta time discretization for the considered problem (1.1). In section 3 we introduce an important quantity related to the numerical flux and present the convergence theorem. In section 4, we derive the error equations of the RKDG scheme and a key lemma for the error estimates. We then prove the main result for general monotone numerical fluxes and for upwind numerical fluxes in section 5. Some of the more technical proofs of several lemmas are collected in the appendix.

2. RKDG method. We follow [8] and define the RKDG method for the problem (1.1) in one space dimension. The multidimensional case is similar. For each partition of the interval $I = (0,1)$, $\{x_{j+\frac{1}{2}}\}_{j=0}^{N}$, we set $I_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$, and $h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ for $j = 1, \dots, N$; we denote the quantity $\max_{1\le j\le N} h_j$ by h. For a given time step τ (which could actually change from step to step but is taken as a constant with respect to the time level n for simplicity), the solution of the scheme is denoted by $u_h^n(x) = u_h(x, n\tau)$, which belongs to the finite element space

$$V_h = V_h^k = \{\, v \in L^1(0,1) : v|_{I_j} \in \mathbb{P}^k(I_j), j = 1, \dots, N \,\}, \tag{2.1}$$

where $\mathbb{P}^k(I_j)$ denotes the space of polynomials in I_j of degree at most k. Note that the functions in V_h are allowed to have discontinuities across element interfaces.

In what follows, we will consider the standard L^2-projection of a function $p \in L^2(0,1)$ into the finite element space V_h, denoted by $\mathbb{P}_h p$, which is defined as the unique function in V_h such that

$$\int_0^1 (\mathbb{P}_h p(x) - p(x)) v_h(x)\, dx = 0 \quad \forall v_h \in V_h. \tag{2.2}$$

For notational convenience we would like to introduce the following operator related to the discontinuous Galerkin spatial discretization. For any functions $p, q \in$

$L^2(0,1)$, denote

$$\mathcal{H}_j(p,q) = \int_{I_j} f(p)\partial_x q(x)\,dx - \widehat{h}(p)_{j+\frac{1}{2}} q(x^-_{j+\frac{1}{2}}) + \widehat{h}(p)_{j-\frac{1}{2}} q(x^+_{j-\frac{1}{2}}), \tag{2.3}$$

where $\widehat{h}(p)_{j+1/2} \equiv \widehat{h}(p^-_{j+1/2}, p^+_{j+1/2})$ is a given monotone numerical flux that depends on the two values of the function p at the discontinuity point $x_{j+1/2}$, namely $p^\pm_{j+1/2} = p(x^\pm_{j+1/2})$. The numerical flux $\widehat{h}(a,b)$ satisfies the following conditions:

(a) It is locally Lipschitz continuous, so it is bounded in any bounded interval.
(b) It is consistent with the flux $f(p)$, i.e., $\widehat{h}(p,p) = f(p)$.
(c) It is a nondecreasing function of its first argument and a nonincreasing function of its second argument.

The best-known examples of monotone numerical fluxes are the Godunov flux, the Engquist–Osher flux, the Lax–Friedrichs flux, etc. For more details, see, for example, [18]. We use the usual notation $[p] = p^+ - p^-$ and $\bar{p} = (p^+ + p^-)/2$ to denote the jump and mean of the function p on each boundary point, respectively.

The approximate solution in V_h from time $n\tau$ to $(n+1)\tau$ given by the RKDG method with the second order TVD Runge–Kutta time discretization is defined as follows: find $w^n_h \equiv w^n_h(x) \in V_h$ and $u^{n+1}_h \equiv u^{n+1}_h(x) \in V_h$ such that, for any $v_h \equiv v_h(x) \in \mathbb{P}^k(I_j)$ and $1 \le j \le N$,

$$\int_{I_j} w^n_h v_h\,dx = \int_{I_j} u^n_h v_h\,dx + \tau\mathcal{H}_j(u^n_h, v_h), \tag{2.4a}$$

$$\int_{I_j} u^{n+1}_h v_h\,dx = \frac{1}{2}\int_{I_j} u^n_h v_h\,dx + \frac{1}{2}\int_{I_j} w^n_h v_h\,dx + \frac{\tau}{2}\mathcal{H}_j(w^n_h, v_h), \tag{2.4b}$$

with the initial value $u^0_h = \mathbb{P}_h u_0(x)$. This is an explicit time-marching method when a local orthogonal basis is chosen for polynomials on I_j or when a small local mass matrix on I_j is inverted. More details and numerical results of this scheme can be found in, e.g., [3] and [12].

3. The main results. We would like to present the main results on the error estimates in this section. To this end, we will introduce some notation for convenience and define an important quantity measuring the relationship between the numerical flux and the physical flux.

3.1. Notation for different constants. We will adopt the following convention for different constants. These constants may have a different value in each occurrence.

We will denote by C (or accompanied by lower indices) a positive constant independent of h and τ, which may depend on the solution of the considered conservation law (1.1). Especially, to emphasize the nonlinearity of the flux $f(u)$, we will denote by $C_\star$ a positive constant depending solely on the maximum of $|f''|$ or/and $|f'''|$. We remark that $C_\star = 0$ for a linear flux $f = cu$, where c is a constant.

We will denote by ε a small positive constant independent of h, τ, and u, the solution of conservation law (1.1). Meanwhile, by M and $M(\varepsilon)$ we will denote those constants depending only on the small constant ε.

3.2. Modification of the flux. To achieve uniform a priori error estimates for the RKDG method, we make the following customary modification on the flux $f(u)$. Suppose the initial solution $u_0(x)$ lies in $[m_0, M_0]$; then, by the maximum principle, the exact solution $u(x,t)$ is also in this range. Thus there is no harm in modifying the

flux function f on the set $\mathbb{R}\backslash[m_0, M_0]$, since the exact solution to (1.1) stays the same. We would choose the modified flux function $\tilde{f}$ to equal the original flux function f on $[m_0, M_0]$, belong to $C^3(\mathbb{R})$, and satisfy $\tilde{f}'(s) = \tilde{f}''(s) = 0$ for all $s \notin [m_0 - 1, M_0 + 1]$. For notational convenience this modified function is still denoted by f. Therefore, we can assume in this paper that the flux function $f(u)$ itself and up to the third order derivatives are all bounded on $\mathbb{R}$.

3.3. A quantity related to the numerical flux. For a general monotone numerical flux which is consistent with f, we introduce an important quantity $\alpha(\widehat{h}; p)$ to measure the difference between the numerical flux and the physical flux; cf. Harten [13]. The definition is given in the following lemma.

LEMMA 3.1. *For any piecewise smooth function $p \in L^2(0,1)$, on each boundary point we define*

$$\alpha(\widehat{h}; p) \equiv \alpha(\widehat{h}; p^-, p^+) \triangleq \begin{cases} [p]^{-1}(f(\bar{p}) - \widehat{h}(p)) & \text{if } [p] \neq 0, \\ |f'(\bar{p})| & \text{if } [p] = 0, \end{cases} \tag{3.1}$$

where $\widehat{h}(p) \equiv \widehat{h}(p^-, p^+)$ is a monotone numerical flux consistent with the given flux f. Then $\alpha(\widehat{h}; p)$ is nonnegative and bounded for any $(p^-, p^+) \in \mathbb{R}^2$. Moreover, we have

$$\frac{1}{2}|f'(\bar{p})| \leq \alpha(\widehat{h}; p) + C_\star|[p]|, \tag{3.2a}$$

$$-\frac{1}{8}f''(\bar{p})[p] \leq \alpha(\widehat{h}; p) + C_\star|[p]|^2, \tag{3.2b}$$

where the positive constant $C_\star$ depends solely on the maximum of $|f''|$ and/or $|f'''|$.

Proof. Obviously, the first conclusion $\alpha(\widehat{h}; p) \geq 0$ is true, because of the property of a monotone numerical flux or, more generally, an E-flux as defined by Osher [18]:

$$(f(q) - \widehat{h}(p^-, p^+))(p^+ - p^-) \geq 0 \tag{3.3}$$

for all q between p^- and p^+. The Lipschitz continuity of the monotone numerical flux $\widehat{h}$, together with the modification of the flux f, implies the bounded property of $\alpha(\widehat{h}; p)$.

We now prove the inequality (3.2a). If $[p] = 0$, the conclusion is obvious by the definition (3.1). Otherwise, we would like to consider the next two cases to get (3.2a): (i) if $f'(\bar{p}) \geq 0$, then from a simple Taylor expansion up to second order and property (3.3) we have

$$\alpha(\widehat{h}; p) = \frac{1}{[p]}(f(\bar{p}) - f(p^-)) + \frac{1}{[p]}(f(p^-) - \widehat{h}(p)) \geq \frac{1}{2}f'(\bar{p}) - \frac{C_\star}{8}|[p]|;$$

otherwise, (ii) if $f'(\bar{p}) < 0$, there similarly holds that

$$\alpha(\widehat{h}; p) = \frac{1}{[p]}(f(\bar{p}) - f(p^+)) + \frac{1}{[p]}(f(p^+) - \widehat{h}(p)) \geq -\frac{1}{2}f'(\bar{p}) - \frac{C_\star}{8}|[p]|,$$

where the positive constant $C_\star$ is the maximum of $|f''|$. This proves (3.2a). We can also prove the inequality (3.2b) along the same lines, wherein the Taylor expansion up to third order is used. ∎

Remark 3.1. The nonnegative property of the quantity $\alpha(\widehat{h}; p)$ is crucial for obtaining the L^2-stability of the RKDG scheme, especially for the case with the usual CFL condition for piecewise linear elements. Details can be seen in subsection 5.2.

Remark 3.2. Here and in what follows we will refer to the value $\bar{p} = (p^+ + p^-)/2$ as the *reference value* of the function p on each boundary point. It is used in the error analysis for general monotone numerical fluxes. Later on we will introduce another reference value p^* for upwind numerical fluxes; see subsection 5.3.

We would also like to use the following simplified notation. For any functions p and q, we denote

$$\alpha^m(\widehat{h};q)[p]^2 = \sum_{1\le j\le N} \left\{\alpha(\widehat{h};q)_{j+\frac{1}{2}}\right\}^m [p]^2_{j+\frac{1}{2}} \quad (m=1,2,3)$$

if there is no confusion.

3.4. The main results. We are now ready to state the main error estimates of the RKDG scheme (2.4). Detailed proof will be given in subsequent sections for different cases.

THEOREM 3.1 (the main results). *Let u be the exact solution of problem* (1.1)*, which is sufficiently smooth with bounded derivatives, and assume $f \in C^3$. Let u_h be the numerical solution of the RKDG scheme* (2.4) *with the second order TVD Runge–Kutta time discretization, and denote the corresponding numerical error by $e_u^n = u(t^n) - u_h^n$. For regular triangulations of $I = (0,1)$, if the finite element space V_h is of piecewise polynomials of degree $k \ge 1$, then for small enough h there holds the following error estimate:*

$$\max_{0\le n\le N_\tau} ||e_u^n||^2 + \sum_{0\le m<N_\tau} \alpha(\widehat{h};u_h^m)[e_u^m]^2\tau \le C(h^{2k+1}+\tau^4). \tag{3.4a}$$

Moreover, if an upwind numerical flux is used, then for small enough h there holds

$$\max_{0\le n\le N_\tau} ||e_u^n|| \le C(h^{k+1}+\tau^2). \tag{3.4b}$$

These estimates hold for $k \ge 2$ under the restrictive time-step condition $\tau \le \gamma h^{4/3}$ with any given positive constant γ; meanwhile, they hold for $k = 1$ under the usual CFL condition $\tau \le \gamma h$ with a suitable positive CFL number γ which is independent of τ and h. Here $||\cdot||$ is the norm in $L^2(0,1)$, the positive constant C is independent of n, h, τ, and the approximate solution u_h, and $N_\tau = [T^\star/\tau]$.

For the generalization of these results to multiple space dimensions, see the remarks in section 5.

4. Error equations, energy equality, and properties of the finite element spaces. In this section we derive the error equations of the RKDG scheme (2.4) and obtain the important energy equality for the error analysis. At the end of this section, we present some interpolation properties and inverse properties for the finite element spaces that will be used in the error analysis.

4.1. Error equations and energy equality. In order to obtain the error estimate to smooth solutions for the considered RKDG scheme (2.4), we need to first get the error equations. To do this, we follow the idea of Ying [23] and derive the truncation error in time.

LEMMA 4.1. *Assume that the solution u of the conservation law* (1.1) *is sufficiently smooth with bounded derivatives. Let*

$$w(x,t) = u(x,t) + \partial_t u(x,t)\tau; \tag{4.1}$$

then, for any $1 \le j \le N, n < N_\tau$, *and any function* $v(x) \in L^2(I_j)$, *the following equations hold:*

$$\int_{I_j} w(x,t^n)v(x)\,dx = \int_{I_j} u(x,t^n)v(x)\,dx + \tau\mathcal{H}_j(u(x,t^n),v(x)), \tag{4.2a}$$

$$\begin{aligned}\int_{I_j} u(x,t^{n+1})v(x)\,dx &= \frac{1}{2}\int_{I_j} u(x,t^n)v(x)\,dx + \frac{1}{2}\int_{I_j} w(x,t^n)v(x)\,dx \\ &\quad + \frac{\tau}{2}\mathcal{H}_j(w(x,t^n),v(x)) + \int_{I_j} E(x;n)v(x)\,dx,\end{aligned} \tag{4.2b}$$

where $E(x;n) = O(\tau^3)$.

Proof. By the Taylor expansion in time we have

$$u(x,t+\tau) - u(x,t) - \partial_t u(x,t)\frac{\tau}{2} - \partial_t u(x,t+\tau)\frac{\tau}{2} = O(\tau^3).$$

By the conservation law (1.1) and the Taylor expansion, it follows that

$$\begin{aligned}\partial_t u(x,t+\tau) &= -\partial_x f(u(x,t+\tau)) = -\partial_x f(u(x,t) + \partial_t u(x,t)\tau + O(\tau^2)) \\ &= -\partial_x f(u(x,t) + \partial_t u(x,t)\tau) + O(\tau^2) = -\partial_x f(w(x,t)) + O(\tau^2).\end{aligned}$$

We then substitute this into the former equation to get

$$u(x,t+\tau) = \frac{1}{2}u(x,t) + \frac{1}{2}w(x,t) - \partial_x f(w(x,t))\frac{\tau}{2} + O(\tau^3).$$

Therefore

$$\begin{aligned}w(x,t^n) &= u(x,t^n) - \partial_x f(u(x,t^n))\tau, \\ u(x,t^{n+1}) &= \frac{1}{2}u(x,t^n) + \frac{1}{2}w(x,t^n) - \partial_x f(w(x,t^n))\frac{\tau}{2} + O(\tau^3).\end{aligned}$$

The above analysis is the same as that in [23]. We multiply the above two equations by an arbitrary function $v(x)$ and integrate over I_j and get, after a simple integration by parts, their weak formulations, which are the conclusions of this lemma. Note that the consistency of the numerical flux $\widehat{h}(u,u) = f(u)$ and the continuity of the exact solution u is used in the derivation. This completes the proof. ∎

We would like to obtain error estimates for each time step, namely $e_u^n = u(t^n) - u_h^n$ and $e_w^n = w(t^n) - w_h^n$, where for notational convenience the argument x is suppressed. In what follows we will also denote $u^n = u(t^n)$ and $w^n = w(t^n)$.

As is customary in error analysis of finite element methods, we denote $\eta_p = \mathbb{Q}p - p$ and $\xi_p = \mathbb{Q}p - p_h$, where the projection $\mathbb{Q}$ is either the L^2-projection $\mathbb{P}_h$ or the Gauss–Radau projection $\mathbb{R}_h$ to be described in more detail in subsection 4.2. The errors of the RKDG scheme at each time step can then be written as $e_p^n = \xi_p^n - \eta_p^n$, where $p = u$ and w.

It is then easy to get, by Lemma 4.1 and the scheme (2.4), the error equations for the error variables ξ_u^n and ξ_w^n in the following form:

$$\int_{I_j} \xi_w^n v_h\,dx = \int_{I_j} \xi_u^n v_h\,dx + \mathcal{K}_j^n(v_h) \quad \forall v_h(x) \in \mathbb{P}^k(I_j), \tag{4.3a}$$

$$\begin{aligned}\int_{I_j} \xi_u^{n+1} v_h\,dx &= \int_{I_j} \left(\frac{1}{2}\xi_u^n + \frac{1}{2}\xi_w^n\right) v_h\,dx + \frac{1}{2}\mathcal{L}_j^n(v_h) \\ &= \int_{I_j} \xi_u^n v_h\,dx + \frac{1}{2}\mathcal{K}_j^n(v_h) + \frac{1}{2}\mathcal{L}_j^n(v_h) \quad \forall v_h(x) \in \mathbb{P}^k(I_j),\end{aligned} \tag{4.3b}$$

for any $1 \le j \le N$ and $n < N_\tau$, where

(4.3c)

$$\mathcal{K}_j^n(v_h) = \int_{I_j} (\eta_w^n - \eta_u^n) v_h \, dx + \tau \mathcal{H}_j(u^n, v_h) - \tau \mathcal{H}_j(u_h^n, v_h),$$

(4.3d)

$$\mathcal{L}_j^n(v_h) = \int_{I_j} (2\eta_u^{n+1} - \eta_w^n - \eta_u^n + 2E(x;n)) v_h \, dx + \tau \mathcal{H}_j(w^n, v_h) - \tau \mathcal{H}_j(w_h^n, v_h).$$

For notational convenience, we will denote $\mathcal{K}^n(v_h) = \sum_{1 \le j \le N} \mathcal{K}_j^n(v_h)$ and $\mathcal{L}^n(v_h) = \sum_{1 \le j \le N} \mathcal{L}_j^n(v_h)$.

Based on these error equations, we shall use energy estimates to analyze the error of the RKDG scheme with second order TVD Runge–Kutta time marching. By taking the test functions $v_h = \xi_u^n$ in (4.3a) and $v_h = \xi_w^n$ in (4.3b), after a simple calculation we obtain the important energy equality

$$||\xi_u^{n+1}||^2 - ||\xi_u^n||^2 = ||\xi_u^{n+1} - \xi_w^n||^2 + \mathcal{K}^n(\xi_u^n) + \mathcal{L}^n(\xi_w^n). \tag{4.4}$$

In the following sections we shall analyze each term on the right-hand side of this energy equation (4.4) to obtain the error estimates under different degrees of polynomials and different types of numerical fluxes. The key point in this process is to obtain a sharp estimate to the first term $||\xi_u^{n+1} - \xi_w^n||^2$, since the estimates to the other terms, namely $\mathcal{K}^n(\xi_u^n)$ and $\mathcal{L}^n(\xi_w^n)$, are a simple extension of that in the semidiscretized setting (cf. [3]).

4.2. Finite element spaces and projections. In this subsection we would like to introduce two interpolations $\mathbb{Q}p(\cdot, t^n)$ which we mentioned before. One is the standard L^2-projection $\mathbb{P}_h p(\cdot, t^n)$, which has been defined in section 2 and will be used for general monotone fluxes. The other is the Gauss–Radau projection $\mathbb{R}_h p(\cdot, t^n)$, which is defined below and will be used for upwind numerical fluxes.

The Gauss–Radau projection is defined following a standard trick in DG analysis. For functions at time level t^n, the projection actually depends on the exact solution u^n and is defined element by element as follows. If $f'(u^n)$ maintains its sign in element I_j, then the projection in this element is defined by

$$\mathbb{R}_h p(x_{j,\ell}) - p(x_{j,\ell}) = 0, \quad \ell = 0, 1, \ldots, k, \tag{4.5}$$

where the points $x_{j,\ell}$ are the Gauss–Radau quadrature points of the interval I_j; that is, one of the boundary points is within the quadrature points:

$$\begin{array}{ll} \text{(i)} & x_{j,k} = x_{j+\frac{1}{2}} \quad \text{if } f'(u^n) > 0 \text{ on the element } I_j, \\ \text{(ii)} & x_{j,0} = x_{j-\frac{1}{2}} \quad \text{if } f'(u^n) < 0 \text{ on the element } I_j. \end{array} \tag{4.6}$$

Otherwise, in case (iii) when $f'(u^n)$ has at least one zero point in the element I_j, we define $\mathbb{R}_h p(\cdot, t^n) = \mathbb{P}_h p(\cdot, t^n)$ to be the standard L^2-projection.

4.2.1. Interpolation properties. Before we start proving the main results for error estimates, we present some interpolation inequalities for these projections. The usual notation of norms and seminorms in Sobolev spaces will be used.

For both projections mentioned above, it is easy to show (cf. [2])

$$||\eta_p^n|| + h||\eta_p^n||_\infty + h^{\frac{1}{2}}||\eta_p^n||_{\Gamma_h} \le Ch^{k+1} \quad (p = u, w;\ 0 \le n \le N_\tau), \tag{4.7}$$

where the positive constant C, solely depending on u, is independent of n, h, and τ, and Γ_h denotes the set of boundary points of all elements I_j. Moreover, for $0 \le n < N_\tau$ we have that

$$\left| \int_{I_j} (p^n - \mathbb{Q}p^n)\partial_x v_h \, dx \right| \le Ch^{k+1}||v_h||_{I_j} \quad \forall v_h \in \mathbb{P}^k(I_j), \tag{4.8}$$

where p could be u and w, and the positive constant C is independent of h, τ, and v_h. This inequality (4.8) is easy to verify, the constant C depends solely on k and $||p||_{L^\infty(H^{k+2}(I_j))}$ for the Gauss–Radau projection $\mathbb{R}_h$, and $C = 0$ for the L^2-projection $\mathbb{P}_h$. For more details, see, for example, [3].

For the Gauss–Radau projection, it is important to mention that it is an exact collocation at one of the boundary points of each element I_j in cases (i) and (ii). If $f'(u^n) > 0$ on the element I_j (case (i)), we have $\eta_p^n(x_{j+1/2}^-) = 0$; if $f'(u^n) < 0$ on the element I_j (case (ii)), we have $\eta_p^n(x_{j-1/2}^+) = 0$, where p could be u or w.

Noticing the definition (4.1) of w, we can conclude, for both projections, that

$$||\eta_u^{n+1} - \eta_u^n|| + ||\eta_w^n - \eta_u^n|| \le Ch^{k+1}\tau, \quad n < N_\tau, \tag{4.9}$$

where the positive constant C solely depends on $\partial_t u$ and is independent of n, h, and τ. It is easy to get (4.9) for the L^2-projection $\mathbb{P}_h$ since the projection $\mathbb{P}_h$ is linear. For the Gauss–Radau projection $\mathbb{R}_h$ the conclusion (4.9) is valid due to the following observation: Notice that the different interpolation points in the Gauss–Radau projection $\mathbb{R}_h$ depend on the exact solution u^n, not on the numerical solution u_h^n. Hence if an element I_j is in case (iii) at the initial time $t = 0$, namely, there is a point inside I_j such that $f'(u_0) = 0$, then the characteristic line at this point is vertical and the element I_j would be in case (iii) for all future time levels. Since we assume in this paper that the exact solution is smooth, the characteristics will not intersect with each other; hence if a cell I_j is in one of the three cases (i), (ii), or (iii) initially, it will stay in that case for all future time. Therefore, we can easily conclude that the inequality (4.9) holds for the Gauss–Radau projection $\mathbb{R}_h$ as well.

4.2.2. Inverse properties. Finally, we list some inverse properties of the finite element space V_h that will be used in our error analysis. For any $v_h \in V_h$, there exists a positive constant C, independent of v_h and h, such that

$$\text{(i) } ||\partial_x v_h|| \le Ch^{-1}||v_h||; \quad \text{(ii) } ||v_h||_{\Gamma_h} \le Ch^{-1/2}||v_h||; \quad \text{(iii) } ||v_h||_\infty \le Ch^{-1/2}||v_h||.$$

For more details of these inverse properties, we refer the reader to [2].

5. Proof of the main results. In this section we present the main proof of Theorem 3.1, leaving some of the more technical details to the appendix. We shall prove the theorem by analyzing each term on the right-hand side of the energy equation (4.4) for different degrees of polynomials and different types of numerical fluxes. As we have mentioned before, the key point of the analysis is to give a sharp estimate to the first term $||\xi_u^{n+1} - \xi_w^n||$ on the right-hand side of (4.4).

5.1. Some lemmas. In this subsection we collect some lemmas for the estimate to the crucial term $||\xi_u^{n+1} - \xi_w^n||$ for both linear and higher order piecewise polynomials. This term comes from the time discretization of the second order TVD Runge–Kutta method and must be analyzed sharply in order to get global optimal error estimates.

We first consider the difference between the operator $\mathcal{L}^n$ and $\mathcal{K}^n$ at successive stages of the Runge–Kutta time marching. The main technique used is Taylor expansion, with special attention paid to the terms associated with the boundaries of each element. The inequalities (4.7), (4.8), and (4.9) will be used for the projection $\mathbb{Q}$.

By using Taylor expansions on $f(u)$ up to second order, and with some (tedious) manipulations, we can prove the following lemma.

LEMMA 5.1. *Let ε be any given small positive constant. Suppose that the interpolation properties* (4.7), (4.8), *and* (4.9) *are satisfied; then we have, for $n = 0, 1, \ldots, N_\tau - 1$ and any $v_h \in V_h$, that*

$$\begin{aligned}(\mathcal{L}^n - \mathcal{K}^n)(v_h) \le{}& \varepsilon||v_h||^2 + \frac{M(\varepsilon)\tau^2}{h}\alpha^2(\widehat{h}; w_h^n)[\xi_w^n]^2 + \frac{M(\varepsilon)\tau^2}{h}\alpha^2(\widehat{h}; u_h^n)[\xi_u^n]^2 \\ &+ \frac{C_\star\tau^2}{h^2}(||e_u^n||_\infty^2||\xi_u^n||^2 + ||e_w^n||_\infty^2||\xi_w^n||^2) + C\tau^2(||\xi_u^n||^2 + ||\xi_w^n||^2) \\ &+ C(\Xi(n)h^{2k+2}\tau + \tau^6) - \sum_{1\le j\le N}\tau f'(u_j^n)\int_{I_j} v_h\partial_x(\xi_w^n - \xi_u^n)\,dx, \end{aligned} \tag{5.1}$$

where $\Xi(n) = 1 + C_\star h^{-1}||e_u^n||_\infty^2 + C_\star h^{-1}||e_w^n||_\infty^2$, the positive constants C and $C_\star$ are independent of n, h, τ, and the approximate solution u_h, and the positive constant $M(\varepsilon) = O(\varepsilon^{-1})$ depends on ε solely. u_j^n denotes the value of the exact solution at the cell center, $u(x_j, t^n)$.

The rather technical proof of this lemma is left for the appendix. We now use this lemma to get an estimate to $||\xi_u^{n+1} - \xi_w^n||$. The next lemma follows from Lemma 5.1 by choosing suitable test functions v_h and setting the positive constant ε in (5.1) small enough.

LEMMA 5.2. *Suppose that the interpolation properties* (4.7), (4.8), *and* (4.9) *are satisfied; then we have, for any $n = 0, 1, \ldots, N_\tau - 1$,*

$$\begin{aligned}||\xi_u^{n+1} - \xi_w^n||^2 \le{}& C(\Xi(n)h^{2k+2}\tau + \tau^6) + \frac{M\tau^2}{h}\alpha^2(\widehat{h}; w_h^n)[\xi_w^n]^2 + \frac{M\tau^2}{h}\alpha^2(\widehat{h}; u_h^n)[\xi_u^n]^2 \\ &+ \frac{C_\star\tau^2}{h^2}(||e_u^n||_\infty^2||\xi_u^n||^2 + ||e_w^n||_\infty^2||\xi_w^n||^2) + C\tau^2(||\xi_u^n||^2 + ||\xi_w^n||^2) \\ &+ M\sum_{1\le j\le N}\tau^2|f'(u_j^n)|^2\int_{I_j}\{\partial_x(\xi_w^n - \xi_u^n)\}^2\,dx, \end{aligned} \tag{5.2}$$

where the positive constants C and $C_\star$ are independent of n, h, τ, and the approximate solution u_h, and the constant M is solely determined by a suitably given positive constant ε in Lemma 5.1.

Proof. By subtracting the error equation (4.3a) from (4.3b) we have that

$$\int_{I_j}(\xi_u^{n+1} - \xi_w^n)v_h\,dx = \frac{1}{2}(\mathcal{L}_j^n - \mathcal{K}_j^n)(v_h) \quad \forall v_h \in \mathbb{P}^k(I_j), \quad 1 \le j \le N, \tag{5.3}$$

which implies that $||\xi_u^{n+1} - \xi_w^n||^2 = \frac{1}{2}(\mathcal{L}^n - \mathcal{K}^n)(\xi_u^{n+1} - \xi_w^n)$ if the test function in (5.3) is taken as $v_h = \xi_u^{n+1} - \xi_w^n$. Also taking the test function $v_h = \xi_u^{n+1} - \xi_w^n$ in (5.1), and

letting the positive constant ε be small enough, we have that

$$\begin{aligned}||\xi_u^{n+1}-\xi_w^n||^2 \le C(\Xi(n)h^{2k+2}\tau+\tau^6)+\frac{M\tau^2}{h}\alpha^2(\widehat{h};w_h^n)[\xi_w^n]^2+\frac{M\tau^2}{h}\alpha^2(\widehat{h};u_h^n)[\xi_u^n]^2\\ +\frac{C_\star\tau^2}{h^2}(||e_u^n||_\infty^2||\xi_u^n||^2+||e_w^n||_\infty^2||\xi_w^n||^2)+C\tau^2(||\xi_u^n||^2+||\xi_w^n||^2)\\ -M\sum_{1\le j\le N}\tau f'(u_j^n)\int_{I_j}(\xi_u^{n+1}-\xi_w^n)\partial_x(\xi_w^n-\xi_u^n)\,dx,\end{aligned}\tag{5.4}$$

where the positive constant M is determined solely by ε.

We now need only to estimate the last integral term in (5.4). To this end, we take the test function $v_h\in V_h$ in (5.3), defined on each element I_j by $v_h=-\tau f'(u_j^n)\partial_x(\xi_w^n-\xi_u^n)$. Then we have, for $1\le j\le N$, that

$$-\tau f'(u_j^n)\int_{I_j}(\xi_u^{n+1}-\xi_w^n)\partial_x(\xi_w^n-\xi_u^n)\,dx=\frac{1}{2}(\mathcal{L}_j^n-\mathcal{K}_j^n)(-\tau f'(u_j^n)\partial_x(\xi_w^n-\xi_u^n)).$$

Thus we can use Lemma 5.1 to estimate the integral term in (5.4) by taking this test function in (5.1) and using the inverse property (i) and the general CFL condition $\tau\le Ch$ for a fixed constant C independent of h and τ. The proof of this lemma is thus completed by substituting this estimate into the inequality (5.4). ∎

The conclusion (5.2) is the key inequality for obtaining error estimates of the DG scheme with the second order TVD Runge–Kutta time marching, which holds for the finite element space with piecewise polynomials of any degree $k\ge 1$. In the next two subsections, we will use this key estimate to obtain error estimates for general numerical fluxes, i.e., to estimate the first term on the right-hand side of (4.4). We shall pay more attention to estimating sharply the last integral term in (5.2) for either linear or higher order piecewise polynomial finite element spaces.

We follow the same lines of analysis for the operator $(\mathcal{L}^n-\mathcal{K}^n)(v_h)$ in Lemma 5.1, using Taylor expansions of $f(u)$ up to second order and first order, respectively, to prove the following lemma.

LEMMA 5.3. *Let ε be any given small positive constant. Suppose that the interpolation properties* (4.7), (4.8), *and* (4.9) *are satisfied; then we have, for $n=0,1,\dots,N_\tau-1$ and any $v_h\in V_h$, that*

$$\begin{aligned}\mathcal{K}^n(v_h)\le\varepsilon||v_h||^2+\frac{M(\varepsilon)\tau^2}{h}\alpha^2(\widehat{h};u_h^n)[\xi_u^n]^2+\left(\frac{C_\star\tau^2}{h^2}||e_u^n||_\infty^2+C\tau^2\right)||\xi_u^n||^2\\ +(Ch^{-1}+C_\star h^{-1}||e_u^n||_\infty^2)h^{2k+2}\tau-\sum_{1\le j\le N}\tau f'(u_j^n)\int_{I_j}v_h\partial_x\xi_u^n\,dx\end{aligned}\tag{5.5}$$

and

$$\mathcal{K}^n(v_h)\le\varepsilon||v_h||^2+\frac{C\tau^2}{h^2}||\xi_u^n||^2+Ch^{2k}\tau^2,\tag{5.6}$$

where the positive constants C and $C_\star$ are independent of n,h,τ, and the approximate solution u_h, and the positive constant $M(\varepsilon)=O(\varepsilon^{-1})$ depends solely on ε.

Proof. The proof is technical but follows along the same lines as that for Lemma 5.1 and is thus omitted.

COROLLARY 5.1. *Under the assumptions of Lemma* 5.3, *if a general CFL condition $\tau\le Ch$ is satisfied, then we have*

$$||\xi_w^n||\le C||\xi_u^n||+Ch^k\tau\quad(n=0,1,\dots,N_\tau-1),\tag{5.7}$$

where the positive constant C is independent of n, h, τ, and the approximate solution u_h.

Proof. Since it follows from the error equations (4.3a) that $||\xi_w^n - \xi_u^n||^2 = \mathcal{K}^n(\xi_w^n - \xi_u^n)$, we can complete the proof by taking the test function $v_h = \xi_w^n - \xi_u^n$ and setting ε small enough in (5.6). Here the general CFL condition $\tau \le Ch$ is used. ∎

The estimates in Lemma 5.3 hold for any test function $v_h \in V_h$. However, for a special test function $\xi_u^n \in V_h$ we have another estimate to $\mathcal{K}^n(\xi_u^n)$ given in the lemma below, which corresponds to the second term on the right-hand side of the energy equation (4.4). The proof of this lemma will be given in the appendix.

LEMMA 5.4. *Suppose the interpolation properties* (4.7), (4.8), *and* (4.9) *are satisfied; then we have, for $n = 0, 1, \dots, N_\tau - 1$, that*

$$\mathcal{K}^n(\xi_u^n) \le \Phi(u^n)||\xi_u^n||^2\tau - \frac{\tau}{4}\alpha(\widehat{h}; u_h^n)\left[\xi_u^n\right]^2 + (C + C_\star h^{-1}||e_u^n||_\infty^2)h^{2k+1}\tau, \tag{5.8}$$

where $\Phi(u^n) = C + C_\star(||\xi_u^n||_\infty + h^{-1}||e_u^n||_\infty^2)$, and the positive constants C and $C_\star$ are independent of n, h, τ, and the approximate solution u_h.

By a similar analysis we have the following lemma to estimate the last term on the right-hand side of the energy equation (4.4). The proof is omitted.

LEMMA 5.5. *Under the assumptions in Lemma* 5.4, *we have, for* $n = 0, 1, \dots, N_\tau - 1$,

$$\mathcal{L}^n(\xi_w^n) \le \Phi(w^n)||\xi_w^n||^2\tau - \frac{\tau}{4}\alpha(\widehat{h}; w_h^n)\left[\xi_w^n\right]^2 + (C + C_\star h^{-1}||e_w^n||_\infty^2)h^{2k+1}\tau + C\tau^5, \tag{5.9}$$

where $\Phi(w^n) = C + C_\star(||\xi_w^n||_\infty + h^{-1}||e_w^n||_\infty^2)$, and the positive constants C and $C_\star$ are independent of n, h, τ, and the approximate solution u_h.

5.2. Error estimates for general monotone numerical fluxes. In this subsection we will prove Theorem 3.1 for arbitrary monotone numerical fluxes. To emphasize the main idea of the analysis we would like to present in detail the proof for the case of linear polynomials, i.e., the proof for (3.4a) under a suitable CFL condition for $k = 1$. We will then briefly indicate the additional difficulties for higher order piecewise polynomial cases.

To deal with the nonlinearity of the flux $f(u)$ we would like to make an a priori assumption that, for small enough h, there holds

$$||u^n - u_h^n|| \le h. \tag{5.10}$$

This assumption is obviously satisfied for $n = 0$ by $u_h^0 = \mathbb{P}_h u_0(x)$ defined in the DG scheme. We shall later verify the correctness of (5.10) by showing that (5.10) still holds true for $n+1$ if it holds true for a given n. For a linear flux $f = cu$, this a priori assumption is unnecessary.

COROLLARY 5.2. *Suppose that the interpolation property* (4.7) *is satisfied; then the a priori assumption* (5.10) *implies that*

$$||e_p^n||_\infty \le Ch^{\frac{1}{2}} \quad \textit{and} \quad ||\xi_p^n||_\infty \le Ch^{\frac{1}{2}} \quad (p = u, w). \tag{5.11}$$

Proof. This follows from the inverse property (iii) and the inequality (5.7). ∎

To obtain the error estimates shown in Theorem 3.1, we shall give separate estimates to each term on the right-hand side of (4.4). The analyses for $\mathcal{K}(\xi_u^n)$ and $\mathcal{L}(\xi_w^n)$ have been given in Lemma 5.4 and Lemma 5.5. Thus in what follows we will pay

special attention to the estimate of the first term on the right-hand side of (4.4) for linear and higher order piecewise polynomials, respectively. To this end, we need to obtain a sharp estimate to $\max_{1\le j\le N}|f'(u_j^n)|^2\cdot||\partial_x(\xi_w^n-\xi_u^n)||^2$ by virtue of Lemma 5.2, where $||\partial_x(\xi_w^n-\xi_u^n)||^2\triangleq\sum_{1\le j\le N}||\partial_x(\xi_w^n-\xi_u^n)||_{I_j}^2$.

5.2.1. Estimates for the linear polynomials. The inequality (5.5), together with (5.2), gives more information to estimate $||\partial_x(\xi_w^n-\xi_u^n)||$ sharply for piecewise linear polynomials than for higher order polynomials. We start by noting that any function g_h can be written as

$$g_h=\widetilde{g_h}+(g_h-\widetilde{g_h}), \tag{5.12}$$

where $\widetilde{g_h}$ is a piecewise constant function, defined by the corresponding average of g_h on each element I_j, i.e.,

$$\int_{I_j}(g_h-\widetilde{g_h})\,dx=0,\quad 1\le j\le N.$$

In the case when the finite element space V_h is of piecewise linear polynomials, it follows from $\xi_u^n\in V_h$ that its derivative $\partial_x\xi_u^n$ is a constant on each element I_j. Hence we also have, for any function g_h, that

$$\int_{I_j}(g_h-\widetilde{g_h})\partial_x\xi_u^n\,dx=0,\quad 1\le j\le N. \tag{5.13}$$

This property plays a key role in obtaining the estimates under the usual CFL condition. Unfortunately, this property holds only for piecewise linear polynomials, not for piecewise polynomials of higher order.

Let $g_h=\xi_w^n-\xi_u^n$; clearly both g_h and $g_h-\widetilde{g_h}\in V_h$. We remark that $g_h-\widetilde{g_h}\in V_h$ holds only for the DG method, not for the standard conforming finite element methods. After a simple calculation, it is easy to show by (4.3a) and (5.12) that

$$||g_h-\widetilde{g_h}||^2=(g_h,g_h-\widetilde{g_h})=(\xi_w^n-\xi_u^n,g_h-\widetilde{g_h})=\mathcal{K}^n(g_h-\widetilde{g_h}). \tag{5.14}$$

By setting the test function $v_h=g_h-\widetilde{g_h}$ in (5.5) we can see that the last integral term in (5.5) becomes 0. Choosing ε small enough, then we have, by (5.14), that

$$||g_h-\widetilde{g_h}||^2\le\frac{M\tau^2}{h}\alpha^2(\widehat{h};u_h^n)[\xi_u^n]^2+\left(C+\frac{C_\star}{h^2}||e_u^n||_\infty^2\right)||\xi_u^n||^2\tau^2+\Upsilon(n)h^3\tau, \tag{5.15}$$

where $\Upsilon(n)=C+C_\star||e_u^n||_\infty^2$ and the positive constants C and $C_\star$ are independent of n,h,τ, and the numerical solution u_h.

Noticing that for the DG method $\partial_x g_h=\partial_x(g_h-\widetilde{g_h})\in V_h$, we have, by virtue of the inverse property (i), a sharp estimate in the form

$$||\partial_x(\xi_w^n-\xi_u^n)||^2\le\frac{M\tau^2}{h^3}\alpha^2(\widehat{h};u_h^n)[\xi_u^n]^2+\left(\frac{C}{h^2}+\frac{C_\star}{h^4}||e_u^n||_\infty^2\right)||\xi_u^n||^2\tau^2+\Upsilon(n)h\tau. \tag{5.16}$$

To estimate the last integral term in (5.2), we multiply $\max_{1\le j\le N}|f'(u_j^n)|^2$ to both sides of the inequality (5.16) and analyze each term on the right-hand side separately. The first term can be bounded by using the following inequality:

$$\max_{1\le j\le N}|f'(u_j^n)|^2\le 8\max_{1\le j\le N}\alpha^2(\widehat{h};u_h^n)_{j+\frac12}+C_\star||e_u^n||_\infty^2,$$

which results from (3.2a) of Lemma 3.1, together with the smoothness of f and u; however, the other two terms can be bounded easily by using the boundedness of $\max_{1\le j\le N}|f'(u_j^n)|^2$. Thus, the last integral term in (5.2) for piecewise linear polynomials can be bounded by

$$\frac{M\tau^3}{h^3}\max_{1\le j\le N}\alpha^3(\widehat{h};u_h^n)_{j+\frac{1}{2}}\cdot\alpha(\widehat{h};u_h^n)[\xi_u^n]^2\tau+\left(\frac{C\tau^3}{h^2}+\frac{C_\star\tau^3}{h^4}||e_u^n||_\infty^2\right)||\xi_u^n||^2\tau+\Upsilon(n)h\tau^3,$$

where the inverse property (ii) and the boundedness of $\alpha(\widehat{h};u_h^n)$ (see Lemma 3.1) have been used.

We substitute this new estimate, together with (5.7), into (5.2). Then we get the following sharp estimate to the first term on the right-hand side of (4.4) for piecewise linear polynomials in the form

$$\begin{aligned}||\xi_u^{n+1}-\xi_w^n||^2\le{}&C(\Xi(n)h^4\tau+h^3\tau+\tau^6)+\delta_1(n)\alpha(\widehat{h};u_h^n)[\xi_u^n]^2\tau+\delta_2(n)\alpha(\widehat{h};w_h^n)[\xi_w^n]^2\tau\\&+\left\{\frac{C_\star\tau}{h^2}(||e_u^n||_\infty^2+||e_w^n||_\infty^2)+\frac{C_\star\tau^3}{h^4}||e_u^n||_\infty^2+C\tau\right\}||\xi_u^n||^2\tau,\end{aligned}\tag{5.17}$$

under the CFL condition $\tau\le\gamma h$ with a suitable CFL number γ (the exact bound of γ and the fact that it is bounded from below will be discussed later), where the positive constants C and $C_\star$ in (5.17) are independent of n,h,τ, and the approximate solution u_h, and $\Xi(n)=1+C_\star h^{-1}||e_u^n||_\infty^2+C_\star h^{-1}||e_w^n||_\infty^2$ has been defined in (5.1). In the above inequality (5.17),

$$\delta_1(n)=\frac{M\tau}{h}\max_{1\le j\le N}\alpha(\widehat{h};u_h^n)_{j+\frac{1}{2}}+\frac{M\tau^3}{h^3}\max_{1\le j\le N}\alpha^3(\widehat{h};u_h^n)_{j+\frac{1}{2}},\tag{5.18a}$$

$$\delta_2(n)=\frac{M\tau}{h}\max_{1\le j\le N}\alpha(\widehat{h};w_h^n)_{j+\frac{1}{2}},\tag{5.18b}$$

where the positive constant M is solely determined by the fixed constant ε and is independent of n,h,τ, the exact solution u^n, and the numerical solution u_h^n.

Therefore, by combining (5.8), (5.9), and (5.17), together with (5.7), and using the results (5.11) implied by the a priori assumption (5.10), under a suitable CFL condition $\tau\le\gamma h$, finally we obtain, for h small enough, that

$$\begin{aligned}&||\xi_u^{n+1}||^2-||\xi_u^n||^2+\frac{1}{4}\alpha(\widehat{h};u_h^n)[\xi_u^n]^2\tau+\frac{1}{4}\alpha(\widehat{h};w_h^n)[\xi_w^n]^2\tau\\&\le C(||\xi_u^n||^2\tau+h^3\tau+\tau^5)+\delta_1(n)\alpha(\widehat{h};u_h^n)[\xi_u^n]^2\tau+\delta_2(n)\alpha(\widehat{h};w_h^n)[\xi_w^n]^2\tau,\end{aligned}\tag{5.19}$$

where the positive constant C is independent of n,h,τ, and the approximate solution u_h^n.

Since the positive constant M in (5.18) is independent of n,h,τ, and the numerical solution u_h^n, there exist two positive constants r_1 and r_2, independent of n,h, and τ, such that

$$Mr_1+Mr_1^3\le\frac{1}{8}\quad\text{and}\quad Mr_2\le\frac{1}{8}.$$

Here we take r_1 and r_2 as large as possible under these restrictions. Then the time step τ^n for piecewise linear polynomials can be determined by $\tau^n\le\gamma^n h$, where the CFL number γ^n is given by

$$\gamma^n=\min\left\{r_1\left(\max_{1\le j\le N}\alpha(\widehat{h};u_h^n)_{j+\frac{1}{2}}\right)^{-1},r_2\left(\max_{1\le j\le N}\alpha(\widehat{h};w_h^n)_{j+\frac{1}{2}}\right)^{-1}\right\}.\tag{5.20}$$

Since the time step considered in this paper is assumed to be constant for convenience, we would like to write the above CFL condition in the uniform formulation $\tau \le \gamma h$, where the CFL number is taken as $\gamma = \min_{n \le N_\tau} \gamma^n$. We comment on the reason that γ is bounded below from zero in Remark 5.1 below.

Then, under this CFL condition, the inequality (5.19) shows that

$$||\xi_u^{n+1}||^2 - ||\xi_u^n||^2 + \frac{\tau}{8}\alpha(\widehat{h}; u_h^n)[\xi_u^n]^2 + \frac{\tau}{8}\alpha(\widehat{h}; w_h^n)[\xi_w^n]^2 \le C(||\xi_u^n||^2\tau + h^3\tau + \tau^5).$$

By Gronwall's inequality, we can get the following error estimate:

$$||\xi_u^{n+1}||^2 + \sum_{0 \le m \le n} \alpha(\widehat{h}; u_h^m)[\xi_u^m]^2\tau \le C_0 h^3, \quad \text{for any } \ n \le N_\tau, \tag{5.21}$$

where the positive constant C_0 is independent of n, h, τ, and the approximate solution u_h^n. Thus the conclusion (3.4a) for linear piecewise polynomials in Theorem 3.1 follows by triangle inequality and the interpolating property (4.7).

To complete the proof, let us verify the a priori assumption (5.10). If it is satisfied for a certain n, then it follows from (5.21) and the interpolation property (4.7) that

$$||u^{n+1} - u_h^{n+1}|| \le C_0 h^{3/2}.$$

It implies that $||u^{n+1} - u_h^{n+1}|| \le h$ for small enough h and the assumption (5.10) is also true for $n+1$. Thus the given a priori (5.10) is verified, and all of the estimates above based on this a priori (5.10) still hold for $n \le N_\tau$. This completes the proof of (3.4a) for the linear piecewise polynomials and for arbitrary monotone numerical fluxes in Theorem 3.1.

Remark 5.1. We notice that the condition (5.20) is the usual CFL condition for conservation laws. For example, by Lemma 3.1 and (5.11) we know that for any numerical flux $\alpha(\widehat{h}; u_h^n)$ is lower bounded by a constant times the maximum of $|f'|$ for h small enough. For example, the CFL number γ for the linear flux $f = cu$, determined by (5.20), depends solely on $|c|$. This also explains why the CFL constant γ is lower bounded away from zero during mesh refinements.

5.2.2. Estimates for the high order polynomials. As we have mentioned before, the property (5.13) is not true for high order piecewise polynomials. Then we can only get the following estimate to the first term on the right-hand side of (4.4):

$$\begin{aligned} ||\xi_u^{n+1} - \xi_w^n||^2 \le{}& C(\Xi(n)h^{2k+2}\tau + \tau^6) + \frac{M\tau^2}{h}\alpha^2(\widehat{h}; w_h^n)[\xi_w^n]^2 + \frac{M\tau^2}{h}\alpha^2(\widehat{h}; u_h^n)[\xi_u^n]^2 \\ &+ \left\{ \frac{C_\star \tau^2}{h^2}(||e_u^n||_\infty^2 + ||e_w^n||_\infty^2)||\xi_u^n||^2 + C\left(\tau^2 + \frac{\tau^4}{h^4}\right) \right\} ||\xi_u^n||^2 \end{aligned} \tag{5.22}$$

by combining (5.2) and (5.7) and using the inverse properties (i) and (ii). Here the constants C and $C_\star$ are independent of n, h, τ, and the approximate solution u_h, the constant M depends solely on the fixed positive constant ε, and $\Xi(n) = 1 + C_\star h^{-1}||e_u^n||_\infty^2 + C_\star h^{-1}||e_w^n||_\infty^2$ has been defined in (5.1).

Comparing this estimate (5.22) with (5.17), we can see that there is an additional factor τ^4/h^4 in front of the term $||\xi_u^n||^2$ for high order piecewise polynomials. It implies a need for the stronger time-step restriction $\tau \le \gamma h^{\frac{4}{3}}$. For high order ($k \ge 2$) piecewise polynomials, it is justified to have such a stronger time-step restriction because the

scheme (2.4) is linearly unstable under CFL condition $\tau \le \gamma h$ for any fixed $\gamma > 0$; see [12].

The remaining analysis for high order polynomials is almost identical to that for the linear piecewise polynomials, so we omit the details. Finally, under the stronger time-step restriction we also get the conclusion (3.4a).

Remark 5.2. We have only carried out the details of the error estimate for sufficiently smooth solutions of one-dimensional scalar conservation laws. However, for a linear flux $f = cu$ the results of Theorem 3.1 still hold for multidimensional conservation laws, because the a priori assumption (5.10) is unnecessary. Moreover, for d-dimensional general nonlinear problems the above analysis still works if we assume the degree of piecewise polynomials satisfies $k > \frac{1}{2}(d+1)$ to keep the a priori assumption (5.10) valid.

Remark 5.3. The assumption $f \in C^3$ is used in the above estimates. However, for the piecewise high order ($k \ge 2$) polynomials it is enough to assume $f \in C^2$ by using another a priori assumption, $||u^n - u_h^n|| \le h^{3/2}$, and using lower order Taylor expansions. The details of the analysis are omitted since they are very similar to what we have presented above.

5.3. Optimal error estimates for the upwind numerical fluxes. In this subsection we shall study the error estimates for RKDG schemes with upwind numerical fluxes. A numerical flux $\widehat{h}(p)$ is called upwind if it satisfies

$$\widehat{h}(p) = \begin{cases} f(p^-) & \text{if } f'(q) \ge 0 \quad \forall q \in [\min(p^-, p^+), \max(p^-, p^+)], \\ f(p^+) & \text{if } f'(q) < 0 \quad \forall q \in [\min(p^-, p^+), \max(p^-, p^+)]. \end{cases}$$

The best-known examples of upwind numerical fluxes are the Godunov flux, the Engquist–Osher flux, and the Roe flux with an entropy fix. In addition, we assume that, on each boundary point, the quantity $\alpha(\widehat{h}; p)$ for the upwind numerical flux depends on the value $|f'(p)|$ in a *local* interval including $p^{\pm}$; i.e., there are positive constants C and $C_\star$ independent of f and p such that

$$\alpha(\widehat{h}; p) \le C|f'(\bar{p})| + C_\star|[p]|. \tag{5.23}$$

This assumption (5.23), together with (3.2a), demonstrates the approximate equivalence between the quantity $\alpha(\widehat{h}; p)$ and $|f'(\bar{p})|$. It is easy to verify that all of the upwind numerical fluxes mentioned above satisfy this assumption (5.23).

To obtain the optimal error estimates for these upwind numerical fluxes, we will estimate again each term on the right-hand side of the energy equation (4.4), following along the lines of the analysis in subsection 5.2. Most of the analysis is similar to that in subsection 5.2; we thus point out only the main differences. One such difference is that the Gauss–Radau projection $\mathbb{R}_h$, which interpolates at one of the boundary points of each cell, is used here; see subsection 4.2. The other is that a new reference value $p_h^{n,\star}$ ($p_h^n = u_h^n$ or w_h^n) on each boundary point $x_{j+1/2}$ at time $t = t^n$ is introduced, corresponding to the sign of $f'(u^n)$ on the adjacent elements $I_j \cup I_{j+1}$. If $f'(u^n) < 0$ on $I_j \cup I_{j+1}$ (i.e., the wind blows to the left), we take the reference value $p_h^{n,\star} = p_h^{n,+}$; otherwise (i.e., either $f'(u^n)$ is always positive or it has zero point(s) in $I_j \cup I_{j+1}$), we take the reference value $p_h^{n,\star} = p_h^{n,-}$. This replaces the simple reference value $\bar{p}_h^n = \bar{p}_h$, the average of p_h, used for the case of general monotone numerical fluxes in subsection 5.2.

We now state conclusions for each term on the right-hand side of the energy equation (4.4), in parallel to those in subsection 5.2. We shall only present the results

for piecewise linear polynomials as an example, with proofs deferred to the appendix. In what follows we will use the notation $|f'(u^n)|[\xi_u^n]^2 = \sum_{1\le j\le N} |f'(u^n_{j+\frac{1}{2}})|[\xi_u^n]^2_{j+\frac{1}{2}}$, etc.

LEMMA 5.6. *If the Gauss–Radau projection $\mathbb{R}_h$ is used, then we have, for piecewise linear polynomials ($k = 1$),*

$$
\begin{aligned}
||\xi_u^{n+1} - \xi_w^n||^2 \le C\,(\Xi(n)h^4\tau + \tau^6) + \widetilde{\delta}_1(n)|f'(u^n)|[\xi_u^n]^2\tau + \widetilde{\delta}_2(n)|f'(w^n)|[\xi_w^n]^2\tau \\
+ \left\{ \frac{C_\star\tau^2}{h^2}(||e_u^n||_\infty^2 + ||e_w^n||_\infty^2) + \frac{C_\star\tau^4}{h^4}||e_u^n||_\infty^2 + C\tau^2 \right\} ||\xi_u^n||^2,
\end{aligned} \tag{5.24}
$$

under the general CFL condition $\tau \le Ch$, where the positive constants C and $C_\star$ are independent of n, h, τ, and the approximate solution u_h, and

$$
\widetilde{\delta}_1(n) = \frac{M\tau}{h}||f'(u^n)||_\infty + \frac{M\tau^3}{h^3}||f'(u^n)||_\infty^3 \quad \textit{and} \quad \widetilde{\delta}_2(n) = \frac{M\tau}{h}||f'(w^n)||_\infty, \tag{5.25}
$$

with the positive constant M independent of n, h, τ, u, and u_h.

Remark 5.4. For higher order piecewise polynomials, the estimate to the first term on the right-hand side of the energy equation (4.4), namely $||\xi_u^{n+1} - \xi_w^n||^2$, is almost the same as (5.24), except for an additional factor τ^4/h^4 in front of the last term $||\xi_u^n||^2$.

LEMMA 5.7. *If the Gauss–Radau projection $\mathbb{R}_h$ is used, then we have, for piecewise polynomials with any degree $k \ge 1$, that*

$$
\mathcal{K}^n(\xi_u^n) \le \Psi(u^n)||\xi_u^n||^2\tau - \frac{1}{2}|f'(u^n)|[\xi_u^n]^2\tau + (C + C_\star h^{-2}||e_u^n||_\infty^2)h^{2k+2}\tau, \tag{5.26a}
$$

$$
\mathcal{L}^n(\xi_w^n) \le \Psi(w^n)||\xi_w^n||^2\tau - \frac{1}{2}|f'(w^n)|[\xi_w^n]^2\tau + (C + C_\star h^{-2}||e_w^n||_\infty^2)h^{2k+2}\tau + C\tau^5, \tag{5.26b}
$$

where $\Psi(p^n) = C + C_\star(h^{-1}||e_p^n||_\infty + h^{-2}||e_p^n||_\infty^2)$ ($p = u, w$), and the positive constants C and $C_\star$ are independent of n, h, τ, and the approximate solution u_h.

To deal with the nonlinearity of the flux $f(u)$, we would also like to use here an a priori assumption that, for small enough h,

$$
||u^n - u_h^n|| \le h^{3/2}, \tag{5.27}
$$

which implies that $||e_u^n||_\infty \le Ch$ and $||e_w^n||_\infty \le Ch$. It is easy to verify that this a priori assumption is true.

Finally, by combining the above estimates (5.24), (5.26a), and (5.26b), we can use Gronwall's inequality to obtain the optimal error estimates for upwind numerical fluxes under a suitable CFL condition.

For piecewise linear polynomials, the usual CFL condition $\tau \le \gamma h$ is enough, where the CFL number γ can be determined by $\widetilde{\delta}_1(n) \le 1/4$ and $\widetilde{\delta}_2(n) \le 1/4$. The treatment is similar to that shown in subsection 5.2. However, for piecewise higher order polynomials, a stronger restriction on the time step is needed, as an additional factor τ^4/h^4 appears in front of the term $||\xi_u^n||^2$ in (5.24).

This finishes the proof of (3.4b) for upwind numerical fluxes and thus for all results in Theorem 3.1.

6. Appendix: Proof of several lemmas. We collect the rather technical proofs of Lemmas 5.1, 5.4, 5.6, and 5.7 in this appendix. The conclusions of these lemmas are used in the previous sections to prove Theorem 3.1.

6.1. Proof of Lemma 5.1. Denote $\Pi(n) = f(w^n) - f(w_h^n) - f(u^n) + f(u_h^n)$ for x inside each element and $\hat{\Pi}(n) = f(w^n) - \widehat{h}(w_h^n) - f(u^n) + \widehat{h}(u_h^n)$ for x on the boundary points of each element. Then by subtracting (4.3d) from (4.3e), we have, for any $v_h \in V_h$, the following result:

$$\begin{aligned}(\mathcal{L}_j^n - \mathcal{K}_j^n)(v_h) &= 2\int_{I_j} (\eta_u^{n+1} - \eta_w^n + E(x;n))v_h\,dx + \tau\int_{I_j} \Pi(n)\partial_x v_h\,dx \\ &\quad - \tau\hat{\Pi}(n)_{j+\frac{1}{2}} v_h(x_{j+\frac{1}{2}}^-) + \tau\hat{\Pi}(n)_{j-\frac{1}{2}} v_h(x_{j-\frac{1}{2}}^+) \\ &\triangleq \theta_{1,j}(v_h) + \theta_{2,j}(v_h) + \theta_{3,j}(v_h) + \theta_{4,j}(v_h). \end{aligned} \tag{6.1}$$

We denote the sum of $\theta_{i,j}(v_h)$ over all the elements I_j by $\Theta_i(v_h) = \sum_{1\le j\le N} \theta_{i,j}(v_h)$, where $i = 1, 2, 3, 4$.

We shall analyze each term on the right-hand side of (6.1) separately. The main tool used is Taylor expansion.

First, it is easy to show from Young's inequality $ab \le \varepsilon a^2 + b^2/(4\varepsilon)$, together with $E(x,n) = O(\tau^3)$ (see Lemma 4.1) and the approximation property (4.9), that

$$|\Theta_1(v_h)| \le C(h^{2k+2}\tau^2 + \tau^6) + \varepsilon||v_h||^2,$$

where ε is a small positive constant.

To estimate the second term $\Theta_2(v_h)$, we would like to split the term $\Pi(n)$ into five terms by the Taylor expansion and the definition (4.1), i.e.,

$$\begin{aligned}\Pi(n) &= \tau f''_{w,u} u'(t)(\xi_w^n - \eta_w^n) - \frac{1}{2} f''_w (\xi_w^n - \eta_w^n)^2 + \frac{1}{2} f''_u (\xi_u^n - \eta_u^n)^2 \\ &\quad - f'(u^n)(\eta_w^n - \eta_u^n) + f'(u^n)(\xi_w^n - \xi_u^n) \\ &\triangleq R_1 + R_2 + R_3 + R_4 + R_5; \end{aligned} \tag{6.2}$$

here f''_u, f''_w, and $f''_{w,u}$ are the mean values given by $f''_p = f''(\theta_p p^n + (1-\theta_p)p_h^n)$ and $f''_{u,w} = f''(\theta_{u,w} u^n + (1-\theta_{u,w})w^n)$ with $0 \le \theta_p \le 1$ $(p = u, w)$ and $0 \le \theta_{u,w} \le 1$. Thus, we can write, corresponding to each element I_j, the second term $\Theta_2(v_h)$ in the form

$$\theta_{2,j}(v_h) = \sum_{i=1}^5 \int_{I_j} \tau R_k \partial_x v_h\,dx = \sum_{i=1}^5 S_{i,j}(v_h).$$

Denote the sum of $S_{i,j}(v_h)$ over all elements I_j by $S_i(v_h)$; then $\Theta_2(v_h) = \sum_{i=1}^5 S_i(v_h)$. By Young's inequality and the inverse property (i), it is easy to get that

$$\begin{aligned} |S_1(v_h)| &\le \frac{C_\star \tau^4}{h^2}||\xi_w^n - \eta_w^n||^2 + \varepsilon||v_h||^2, \\ |S_2(v_h)| &\le \frac{C_\star \tau^2}{h^2}||e_w^n||_\infty^2||\xi_w^n - \eta_w^n||^2 + \varepsilon||v_h||^2, \\ |S_3(v_h)| &\le \frac{C_\star \tau^2}{h^2}||e_u^n||_\infty^2||\xi_u^n - \eta_u^n||^2 + \varepsilon||v_h||^2, \\ |S_4(v_h)| &\le \frac{C\tau^2}{h^2}||\eta_w^n - \eta_u^n||^2 + \varepsilon||v_h||^2. \end{aligned}$$

The last term $S_5(v_h)$, which is the sum of $S_{5,j}(v_h) = \tau \int_{I_j} f'(u^n)(\xi_w^n - \xi_u^n)\partial_x v_h \, dx$ for $1 \le j \le N$, remains to be estimated later, together with a few other terms.

Now we are going to estimate the last two terms in (6.1), namely $\theta_{3,j}(v_h)$ and $\theta_{4,j}(v_h)$, which are related to the numerical flux on the boundary points of each element I_j. Since their analysis is almost the same, we will only present the details of the estimate to the term $\theta_{3,j}(v_h) = -\tau\hat{\Pi}(n)_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}})$. To do this, we write $\hat{\Pi}(n)_{j+\frac{1}{2}}$ in the following way:

$$\hat{\Pi}(n) = \underbrace{f(w^n) - f(\bar{w}_h^n)}_{\Lambda_1} \underbrace{-f(u^n) + f(\bar{u}_h^n)}_{\Lambda_2} \underbrace{-f(\bar{u}_h^n) + \widehat{h}(u_h^n)}_{\Lambda_3} \underbrace{+f(\bar{w}_h^n) - \widehat{h}(w_h^n)}_{\Lambda_4}, \tag{6.3}$$

where the subscript $j + 1/2$ is omitted for convenience. As we have done for $\theta_{2,j}(v_h)$, we expand the first two terms, Λ_1 and Λ_2, by Taylor expansion in the form

$$\begin{aligned}\Lambda_1 + \Lambda_2 &= f''_{w,u} u'(t)\tau(\bar{\xi}_w^n - \bar{\eta}_w^n) - \frac{1}{2} f''_w(\bar{\xi}_w^n - \bar{\eta}_w^n)^2 + \frac{1}{2} f''_u(\bar{\xi}_u^n - \bar{\eta}_u^n)^2 \\ &\quad - f'(u^n)(\bar{\eta}_w^n - \bar{\eta}_u^n) + f'(u^n)(\bar{\xi}_w^n - \bar{\xi}_u^n) \\ &\triangleq Q_1 + Q_2 + Q_3 + Q_4 + Q_5, \end{aligned} \tag{6.4}$$

where f''_u, f''_w, and $f''_{w,u}$ are again the mean values. We remark that although we use the same notation here as that in (6.2), this notation may have different mean values. Therefore, we have the following representation for the terms $\theta_{3,j}(v_h)$ and $\theta_{4,j}(v_h)$:

$$\theta_{3,j} = -\tau(Q_1 + \cdots + Q_5 + \Lambda_3 + \Lambda_4)_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) \triangleq (T_{1,j} + \cdots + T_{5,j} + T_{6,j} + T_{7,j})(v_h),$$

$$\theta_{4,j} = +\tau(Q_1 + \cdots + Q_5 + \Lambda_3 + \Lambda_4)_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) \triangleq (\widetilde{T}_{1,j} + \cdots + \widetilde{T}_{5,j} + \widetilde{T}_{6,j} + \widetilde{T}_{7,j})(v_h).$$

Denote the sum of $T_{i,j}(v_h)$ (respectively, $\widetilde{T}_{i,j}(v_h)$) over all elements I_j by $T_i(v_h)$ (respectively, $\widetilde{T}_i(v_h)$). Next we will estimate one by one the terms listed above.

By Young's inequality and the inverse property (ii), it is easy to show that

$$\begin{aligned}|T_1(v_h)| &\le \frac{C_\star \tau^4}{h^2}||\xi_w^n||^2 + \frac{C_\star \tau^4}{h}||\eta_w^n||^2_{\Gamma_h} + \varepsilon||v_h||^2, \\ |T_2(v_h)| &\le \frac{C_\star \tau^2}{h^2}||e_w^n||^2_\infty||\xi_w^n||^2 + \frac{C_\star \tau^2}{h}||e_w^n||^2_\infty||\eta_w^n||^2_{\Gamma_h} + \varepsilon||v_h||^2, \\ |T_3(v_h)| &\le \frac{C_\star \tau^2}{h^2}||e_u^n||^2_\infty||\xi_u^n||^2 + \frac{C_\star \tau^2}{h}||e_u^n||^2_\infty||\eta_u^n||^2_{\Gamma_h} + \varepsilon||v_h||^2, \\ |T_4(v_h)| &\le \frac{C\tau^2}{h}||\eta_w^n - \eta_u^n||^2_{\Gamma_h} + \varepsilon||v_h||^2. \end{aligned}$$

By virtue of the definition of $\alpha(\widehat{h}; u_h^n)$ in Lemma 3.1, we have that

$$\begin{aligned}|\Lambda_3| &= \alpha(\widehat{h}; u_h^n)|[u_h^n]| = \alpha(\widehat{h}; u_h^n)|[u^n - u_h^n]| \le \alpha(\widehat{h}; u_h^n)(|[\xi_u^n]| + |[\eta_u^n]|), \\ |\Lambda_4| &= \alpha(\widehat{h}; w_h^n)|[w_h^n]| = \alpha(\widehat{h}; w_h^n)|[w^n - w_h^n]| \le \alpha(\widehat{h}; w_h^n)(|[\xi_w^n]| + |[\eta_w^n]|). \end{aligned}$$

Then Young's inequality with the inverse property (ii) implies that the terms $T_6(v_h)$ and $T_7(v_h)$ can be bounded in the form

$$\begin{aligned}|T_6(v_h)| &\le \frac{M(\varepsilon)\tau^2}{h}\alpha^2(\widehat{h}; u_h^n)[\xi_u^n]^2 + \frac{C\tau^2}{h}\alpha^2(\widehat{h}; u_h^n)[\eta_u^n]^2 + \varepsilon||v_h||^2, \\ |T_7(v_h)| &\le \frac{M(\varepsilon)\tau^2}{h}\alpha^2(\widehat{h}; w_h^n)[\xi_w^n]^2 + \frac{C\tau^2}{h}\alpha^2(\widehat{h}; w_h^n)[\eta_w^n]^2 + \varepsilon||v_h||^2. \end{aligned}$$

Thus we have estimated each term included in $\Theta_3(v_h)$, except the term $T_5(v_h)$, which is the sum of $T_{5,j}(v_h) = -\tau f'(u^n_{j+1/2})(\bar{\xi}^n_w - \bar{\xi}^n_u)_{j+1/2} v_h(x_{j+1/2})$ over each element I_j. This term will be estimated later, together with a few other terms.

Along the same lines we have almost the same estimate to the term $\Theta_4(v_h)$, except that the term $\widetilde{T}_5(v_h)$, which is the sum of $\tilde{T}_{5,j}(v_h) = \tau f'(u^n_{j-1/2})(\bar{\xi}^n_w - \bar{\xi}^n_u)_{j-1/2} v_h(x_{j-1/2})$ over each element I_j, is also left to be estimated later.

We now have obtained estimates for each term in the difference $(\mathcal{L}^n_j - \mathcal{K}^n_j)(v_h)$, except for three remaining terms, namely $S_5(v_h), T_5(v_h)$, and $\widetilde{T}_5(v_h)$. To estimate them sharply, we would like to consider their sum on each element I_j. After a simple integration by parts, we have that

$$\begin{aligned}
&S_{5,j}(v_h) + T_{5,j}(v_h) + \widetilde{T}_{5,j}(v_h) \\
&= -\tau \int_{I_j} f'(u^n_j) v_h \partial_x (\xi^n_w - \xi^n_u)\, dx - \tau \int_{I_j} \{f'(u^n) - f'(u^n_j)\} v_h \partial_x (\xi^n_w - \xi^n_u)\, dx \\
&\quad - \tau \int_{I_j} \partial_x f'(u^n)(\xi^n_w - \xi^n_u) v_h\, dx - \frac{\tau}{2} f'(u^n_{j+\frac{1}{2}})[\xi^n_w - \xi^n_u]_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}}) \\
&\quad + \frac{\tau}{2} f'(u^n_{j-\frac{1}{2}})[\xi^n_w - \xi^n_u]_{j-\frac{1}{2}} v_h(x^+_{j-\frac{1}{2}}) \\
&\triangleq G_{1,j}(v_h) + G_{2,j}(v_h) + G_{3,j}(v_h) + G_{4,j}(v_h) + G_{5,j}(v_h), \qquad (6.5)
\end{aligned}$$

where we have introduced the piecewise constant $u^n_j = u((x_{j+1/2} + x_{j-1/2})/2, t^n)$ on each element I_j. Also we denote the sum of $G_{i,j}(v_h)$ over all elements I_j by $G_i(v_h)$, and then

$$S_5(v_h) + T_5(v_h) + \widetilde{T}_5(v_h) = G_1(v_h) + G_2(v_h) + G_3(v_h) + G_4(v_h) + G_5(v_h). \qquad (6.6)$$

In what follows we will estimate each term on the right-hand side of (6.6) separately.

It follows that $|f'(u^n) - f'(u^n_j)| = O(h)$ on each element I_j from the smoothness of the solution u and the flux f. Then by the inverse property (i) and Young's inequality, it is easy to show that

$$|G_2(v_h)| \le C\tau^2 ||\xi^n_w - \xi^n_u||^2 + \varepsilon ||v_h||^2.$$

By Young's inequality we can show that

$$|G_3(v_h)| \le C\tau^2 ||\xi^n_w - \xi^n_u||^2 + \varepsilon ||v_h||^2.$$

For the last two terms on the right-hand side of (6.6) related to the numerical flux, namely $G_4(v_h)$ and $G_5(v_h)$, we need to estimate them more carefully. Our aim is to bound them by the product of $\alpha(\widehat{h}; u^n_h)$ or $\alpha(\widehat{h}; w^n_h)$ and the corresponding jump of ξ^n_u or ξ^n_w across element interfaces.

Since the estimates are very similar, we will only present the analysis to the term $G_4(v_h)$. Dropping the subscript $j + 1/2$ for convenience, we have the following representation:

$$\begin{aligned}
f'(u^n)[\xi^n_w - \xi^n_u] &= f'(w^n)[\xi^n_w] - f'(u^n)[\xi^n_u] + (f'(u^n) - f'(w^n))[\xi^n_w] \\
&= f'(\bar{w}^n_h)[\xi^n_w] - f'(\bar{u}^n_h)[\xi^n_u] + (f'(w^n) - f'(\bar{w}^n_h))[\xi^n_w] \\
&\quad - (f'(u^n) - f'(\bar{u}^n_h))[\xi^n_u] + (f'(u^n) - f'(w^n))[\xi^n_w] \\
&\triangleq \gamma_{1,j} + \gamma_{2,j} + \cdots + \gamma_{5,j}.
\end{aligned}$$

Then we have $G_4(v_h) = \sum_{i=1}^{5} \gamma_i(v_h)$, where $\gamma_i(v_h) = -\frac{\tau}{2} \sum_{1\le j\le N} \gamma_{i,j} v_h(x_{j+1/2})$. We would like to estimate each term separately. We use the conclusion (3.2a) in Lemma 3.1 to analyze the terms $\gamma_1(v_h)$ and $\gamma_2(v_h)$. Finally, by Young's inequality and the inverse property (ii), we have that

$$
\begin{aligned}
|\gamma_1(v_h)| &\le M(\varepsilon)\frac{\tau^2}{h}\alpha^2(\widehat{h}; w_h^n)[\xi_w^n]^2 + \frac{C_\star \tau^2}{h^2}||e_w^n||_\infty^2||\xi_w^n||^2 + \frac{\varepsilon}{5}||v_h||^2,\\
|\gamma_2(v_h)| &\le M(\varepsilon)\frac{\tau^2}{h}\alpha^2(\widehat{h}; u_h^n)[\xi_u^n]^2 + \frac{C_\star \tau^2}{h^2}||e_u^n||_\infty^2||\xi_u^n||^2 + \frac{\varepsilon}{5}||v_h||^2,\\
|\gamma_3(v_h)| &\le \frac{C_\star \tau^2}{h^2}||e_w^n||_\infty^2||\xi_w^n||^2 + \frac{\varepsilon}{5}||v_h||^2,\\
|\gamma_4(v_h)| &\le \frac{C_\star \tau^2}{h^2}||e_u^n||_\infty^2||\xi_u^n||^2 + \frac{\varepsilon}{5}||v_h||^2,\\
|\gamma_5(v_h)| &\le \frac{C_\star \tau^4}{h^2}||\xi_w^n||^2 + \frac{\varepsilon}{5}||v_h||^2.
\end{aligned}
$$

Therefore, we conclude, by summing up the above estimates, that

$$
\begin{aligned}
|G_4(v_h)| \le \varepsilon||v_h||^2 &+ \frac{C_\star \tau^2}{h^2}||e_u^n||_\infty^2||\xi_u^n||^2 + \frac{C_\star \tau^2}{h^2}||e_w^n||_\infty^2||\xi_w^n||^2 + \frac{C_\star \tau^4}{h^2}||\xi_w^n||^2\\
&+ M(\varepsilon)\frac{\tau^2}{h}\alpha^2(\widehat{h}; w_h^n)[\xi_w^n]^2 + M(\varepsilon)\frac{\tau^2}{h}\alpha^2(\widehat{h}; u_h^n)[\xi_u^n]^2,
\end{aligned}
$$

where the positive constant $M(\varepsilon)$ depends on ε solely, and the other positive constants C and $C_\star$ are independent of n, h, τ, and the numerical solution u_h.

Similarly, we can obtain the same estimate for the term $G_5(v_h)$ as that for $G_4(v_h)$.

Note that until now we have not yet analyzed the first term $G_1(v_h)$ in (6.6), which is the sum of $G_{1,j}(v_h) = -\tau \int_{I_j} f'(u_j^n) v_h \partial_x(\xi_w^n - \xi_u^n)\, dx$ over each element I_j. This term is left to be estimated differently for piecewise polynomials of different degrees; see subsection 5.2.

Finally, by collecting all the analysis presented above and using the interpolation property (4.7) and (4.9) and the general CFL condition $\tau \le Ch$, after some simple calculations we finish the proof of Lemma 5.1. ∎

6.2. Proof of Lemma 5.4. Noticing the periodic or zero (compactly supported) boundary conditions, after some simple calculation we have the sum of $\mathcal{K}_j^n(\xi_u^n)$ as follows:

$$
\begin{aligned}
\mathcal{K}^n(\xi_u^n) = (\eta_w^n - \eta_u^n, \xi_u^n) &+ \tau \sum_{1\le j\le N} \int_{I_j} \big(f(u^n) - f(u_h^n)\big)\partial_x \xi_u^n\, dx\\
&+ \tau \sum_{1\le j\le N} \big\{f(u^n) - f(\bar{u}_h^n)\big\}_{j+\frac{1}{2}} \big[\xi_u^n\big]_{j+\frac{1}{2}} + \tau \sum_{1\le j\le N} \big\{f(\bar{u}_h^n) - \widehat{h}(u_h^n)\big\}_{j+\frac{1}{2}} \big[\xi_u^n\big]_{j+\frac{1}{2}}\\
&\triangleq W_1 + W_2 + W_3 + W_4. \qquad (6.7)
\end{aligned}
$$

Note that we refer to $\bar{u}_h^n$ in the terms W_3 and W_4 above as the reference value on each boundary point. We would like to analyze each term on the right-hand side of (6.7) separately.

First, it follows from the interpolation property (4.9) that

$$
W_1 \le Ch^{2k+2}\tau + C\tau||\xi_u^n||^2. \qquad (6.8)
$$

Since the exact solution u^n of conservation laws (1.1) is continuous on each boundary point, we have that $[u_h^n] = -[e_u^n] = [\eta_u^n - \xi_u^n]$. Noticing the definition and boundedness of $\alpha(\widehat{h}; u_h^n)$ (see Lemma 3.1), we can easily show, by Young's inequality and the interpolation property (4.7), that

$$W_4 \le -\frac{3}{4}\alpha(\widehat{h}; u_h^n)[\xi_u^n]^2\tau + C\alpha(\widehat{h}; u_h^n)[\eta_u^n]^2\tau \le -\frac{3}{4}\alpha(\widehat{h}; u_h^n)[\xi_u^n]^2\tau + Ch^{2k+1}\tau. \tag{6.9}$$

To complete the proof of Lemma 5.4, we should pay more attention to the terms W_2 and W_3. Thus we would like to use the following Taylor expansions:

$$\begin{aligned} f(u^n) - f(u_h^n) &= f'(u^n)\xi_u^n - \frac{1}{2}f''(u^n)(\xi_u^n)^2 - f'(u^n)\eta_u^n + f''(u^n)\xi_u^n\eta_u^n \\ &\quad - \frac{1}{2}f''(u^n)(\eta_u^n)^2 - \frac{1}{6}f'''_u(\xi_u^n - \eta_u^n)^3 \triangleq \phi_1 + \cdots + \phi_6, \end{aligned} \tag{6.10a}$$

$$\begin{aligned} f(u^n) - f(\bar{u}_h^n) &= f'(u^n)\bar{\xi}_u^n - \frac{1}{2}f''(u^n)(\bar{\xi}_u^n)^2 - f'(u^n)\bar{\eta}_u^n + f''(u^n)\bar{\xi}_u^n\bar{\eta}_u^n \\ &\quad - \frac{1}{2}f''(u^n)(\bar{\eta}_u^n)^2 - \frac{1}{6}\tilde{f}'''_u(\bar{\xi}_u^n - \bar{\eta}_u^n)^3 \triangleq \psi_1 + \cdots + \psi_6, \end{aligned} \tag{6.10b}$$

where f'''_u and $\tilde{f}'''_u$ are again the mean values. These imply the following representation:

$$W_2 = X_1 + X_2 + \cdots + X_6 \quad \text{and} \quad W_3 = Y_1 + Y_2 + \cdots + \cdots + Y_6,$$

where X_i and Y_i, given by

$$X_i = \tau \sum_{1\le j\le N} \int_{I_j} \phi_i \partial_x \xi_u^n \, dx \quad \text{and} \quad Y_i = \tau \sum_{1\le j\le N} (\psi_i)_{j+\frac{1}{2}} [\xi_u^n]_{j+\frac{1}{2}} \quad (i = 1, 2, \ldots, 6),$$

will be estimated separately later.

It is easy to get, after a simple integration by parts, that

$$X_1 + Y_1 = -\frac{\tau}{2} \sum_{1\le j\le N} \int_{I_j} \partial_x f'(u^n)(\xi_u^n)^2 \, dx \le C||\xi_u^n||^2\tau \tag{6.11a}$$

and

$$X_2 + Y_2 = \frac{\tau}{24} \sum_{1\le j\le N} f''(u^n)_{j+\frac{1}{2}} \left[\xi_u^n\right]_{j+\frac{1}{2}}^3 + \frac{\tau}{6} \sum_{1\le j\le N} \int_{I_j} \partial_x f''(u^n)(\xi_u^n)^3 \, dx.$$

Notice that $f''(u^n)[\xi_u^n] = f''(\bar{u}_h^n)[\eta_h^n] - f''(\bar{u}_h^n)[u_h^n] + f'''[\xi_u^n]e_u^n$ by a simple Taylor expansion, where f''' denotes a mean value of the third derivative of the flux f. It implies, together with (3.2b) of Lemma 3.1 and the interpolation property (4.7), that on each boundary point $x_{j+1/2}$ there holds $\frac{1}{24}f''(u^n)[\xi_u^n] \le \frac{1}{3}\alpha(\widehat{h}; u_h^n) + C_\star(h + ||e_u^n||_\infty^2)$. Thus we have, by virtue of the inverse property (ii), that

$$X_2 + Y_2 \le \frac{1}{3}\alpha(\widehat{h}; u_h^n)[\xi_u^n]^2\tau + C_\star(C + ||\xi_u^n||_\infty + h^{-1}||e_u^n||_\infty^2)||\xi_u^n||^2\tau. \tag{6.11b}$$

As we have shown before, $|f'(u^n) - f'(u_j^n)| = O(h)$ on each element I_j. Then by the inverse property (i), together with the interpolation property (4.7) and (4.8), we have that

$$\begin{aligned} X_3 &\le \tau \sum_{1\le j\le N} \left| \int_{I_j} (f'(u^n) - f'(u_j^n))\eta_u^n \partial_x \xi_u^n \, dx \right| + \tau \sum_{1\le j\le N} \left| f'(u_j^n) \int_{I_j} \eta_u^n \partial_x \xi_u^n \, dx \right| \\ &\le Ch^{2k+2}\tau + C||\xi_u^n||^2\tau. \end{aligned}$$

It follows easily that $|f'(u^n_{j+1/2})| \le 2\alpha(\widehat{h}; u^n_h)_{j+1/2} + C_\star ||e^n_u||_\infty$ for any $1 \le j \le N$ from the conclusion (3.2a) in Lemma 3.1 and the smoothness of u and f. Hence, by Young's inequality and the boundedness of $\alpha(\widehat{h}; u^n_h)$ (see Lemma 3.1), we have that

$$Y_3 \le Ch^{2k+1}\tau + \frac{1}{6}\alpha(\widehat{h}; u^n_h)[\xi^n_u]^2\tau + \frac{C_\star\tau}{h}||e^n_u||^2_\infty ||\xi^n_u||^2.$$

Thus, we have that

$$X_3 + Y_3 \le \frac{1}{6}\alpha(\widehat{h}; u^n_h)[\xi^n_u]^2\tau + (C + C_\star h^{-1}||e^n_u||^2_\infty)||\xi^n_u||^2\tau + Ch^{2k+1}\tau. \tag{6.11c}$$

It is easy to show by Young's inequality and the inverse properties (i) and (ii) that

$$X_4 + Y_4 \le C_\star h^{-1}||\eta^n_u||_\infty ||\xi^n_u||^2\tau \le C_\star ||\xi^n_u||^2\tau, \tag{6.11d}$$

$$X_5 + Y_5 \le C_\star h^{-1}||\eta^n_u||_\infty (||\eta^n_u|| + h^{\frac{1}{2}}||\eta^n_u||_{\Gamma_h})||\xi^n_u||\tau \le C_\star ||\xi^n_u||^2\tau + C_\star h^{2k+2}\tau, \tag{6.11e}$$

$$X_6 + Y_6 \le C_\star h^{-1}||e^n_u||^2_\infty (||\xi^n_u||^2\tau + Ch^{2k+2}\tau). \tag{6.11f}$$

Therefore, by summing up the above estimates about W_2 and W_3, we have that

$$W_2 + W_3 \le \Phi(u^n)||\xi^n_u||^2\tau + \frac{1}{2}\alpha(\widehat{h}; u^n_h)\big[\xi^n_u\big]^2\tau + (C + C_\star h^{-1}||e^n_u||^2_\infty)h^{2k+1}\tau, \tag{6.12}$$

where $\Phi(u^n) = C + C_\star(||\xi^n_u||_\infty + h^{-1}||e^n_u||^2_\infty)$, and the positive constants C and $C_\star$ are independent of n, h, τ, and the numerical solution u_h.

Finally, we collect the above estimates (6.8), (6.12), and (6.9) to complete the proof of Lemma 5.4. ∎

6.3. Proof of Lemma 5.6. The proof of Lemma 5.6 follows the general lines of the proof of Lemmas 5.2 and 5.3, with some modifications. First, we use the assumption (5.23) and the fact that $[u^n_h] = -[e^n_u]$, since u^n is a continuous function, to obtain

$$\alpha(\widehat{h}; u^n_h)_{j+\frac{1}{2}} \le C|f'(u^n_{j+\frac{1}{2}})| + C_\star ||e^n_u||_\infty \tag{6.13}$$

for some positive constants C and $C_\star$ on each boundary point $x_{j+1/2}$. There is a similar inequality for the quantity $\alpha(\widehat{h}; w^n_h)$. We would also like, for the current upwind fluxes, to improve the expression $(Ch^{-1} + C_\star h^{-1}||e^n_u||^2_\infty)h^4\tau$ on the right-hand side of the inequality (5.5) in Lemma 5.3 to $(C + C_\star h^{-1}||e^n_u||^2_\infty)h^4\tau$. This improvement would need a sharp estimate to the term $f(u^n) - \hat{h}(u^n_h)$ on each boundary point, in parallel to the proof of Lemma 5.3. In what follows we will only present the key estimates different from the general treatments for monotone numerical fluxes.

On each boundary point $x_{j+1/2}$, we denote $G^n = f(u^{n,\star}_h) - \widehat{h}(u^n_h)$. Here and in what follows, the subscript $j + 1/2$ is omitted for notational convenience. We then have

$$f(u^n) - \hat{h}(u^n_h) = [f(u^n) - f(u^{n,\star}_h)] + G^n.$$

We would like to estimate the expression above for different cases of the sign variation of $f'(u^n)$ on each $I_j \cup I_{j+1}$.

The estimate to $f(u^n)-f(u_h^{n,\star})$ is similar as before, except for one term resulting from the Taylor expansion, namely $\pi \triangleq -f'(u^n)\eta_u^{n,\star}$, which needs more explanation. If the sign of $f'(u^n)$ is unchanged on $I_j \cup I_{j+1}$, then it is obvious that $\pi = 0$ by the definition of $\mathbb{R}_h$ and the setting of the reference value $u_h^{n,\star}$. Otherwise, if $f'(u^n)$ has at least one zero point on $I_j \cup I_{j+1}$, which implies $f'(u^n_{j+1/2}) = O(h)$, then we have $\pi = O(h^{k+3/2})$ by the interpolating approximate property (4.7). Therefore

$$\left| \sum_{1\le j\le N} \pi_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}})\tau \right| \le \varepsilon ||v_h||^2 + Ch^{2k+2}\tau^2. \tag{6.14}$$

Next we would like to estimate the second term G^n for different choices of the reference value $u_h^{n,\star}$ on each boundary point $x_{j+1/2}$. Since the analysis for each case is very similar, we will only present here the estimates for the case that $f'(u^n) < 0$ on $I_j \cup I_{j+1}$; i.e., the reference value is $u_h^{n,\star} = u_h^{n,+}$. To estimate G^n for this case, we consider the detailed setting of the upwind numerical flux $\widehat{h}(u_h^n)$, corresponding to the sign variation of $f'(\cdot)$ between $u_h^{n,-}$ and $u_h^{n,+}$. If $f'(\cdot)$ is negative, then it follows from the upwind property of the numerical flux that (a) $G^n = 0$. If $f'(\cdot)$ is positive, then $G^n = f(u_h^{n,+}) - f(u_h^{n,-})$. In this case, there certainly exists a zero point $u^\star$, such that $f'(u^\star) = 0$, in the interval covering the numerical solution $u_h^n(x^{\pm}_{j+\frac{1}{2}})$ and the exact solution u^n on $I_j \cup I_{j+1}$. Thus a simple manipulation shows (b) $|G^n| \le C_\star(h+||e_u^n||_\infty)|[u_h^n]|$. Otherwise, if the sign of $f'(\cdot)$ is changed, we also have the above inequality (b) by the assumption (5.23) for the upwind numerical fluxes. Thus, by summing up the above conclusions, we have

$$|G^n| \le C_\star(h + ||e_u^n||_\infty)|[u_h^n]|. \tag{6.15}$$

Noticing the inverse property (ii) and $[u_h^n] = [\eta_u^n] - [\xi_u^n]$, we thus have

$$\left| \sum_{1\le j\le N} G^n_{j+\frac{1}{2}} v_h(x^-_{j+\frac{1}{2}})\tau \right| \le \frac{C_\star\tau^2}{h^2}(h^2 + ||e_u^n||_\infty^2)(||\xi_u^n||^2 + ||\eta_u^n||^2_{\Gamma_h}) + \varepsilon||v_h||^2. \tag{6.16}$$

The estimates to the other terms are similar as before; thus we omit them here and conclude the proof for Lemma 5.6. □

6.4. Proof of Lemma 5.7. As indicated before, we will again use the reference value $u_h^{n,\star}$ on each boundary point instead of the simple reference value $\bar{u}_h$ used in the proof of Lemma 5.4, for example in the formulation (6.7), and denote the corresponding representation by the same notation. We will analyze again each term on the right-hand side of (6.7) separately.

The estimate to the term W_1 is the same as before. By the inequality (6.15) for upwind numerical fluxes, the term W_4 can be bounded by

$$W_4 \le C_\star(1 + h^{-1}||e_u^n||_\infty)||\xi_u^n||^2\tau + \frac{C_\star\tau}{h}(h^2 + ||e_u^n||_\infty^2)||\eta_u^n||^2_{\Gamma_h}. \tag{6.17}$$

To estimate the terms W_2 and W_3, we follow the proof of Lemma 5.4 and estimate each term in (6.10) with only minor modifications, arising from the replacement of $\bar{u}_h^n, \bar{\xi}_u^n$, and $\bar{\eta}_u^n$ in (6.10b) by $u_h^{n,\star}, \xi_u^{n,\star}$, and $\eta_u^{n,\star}$, respectively. We still denote these terms by X_i and Y_i separately and estimate them one by one.

After a simple integration by parts, it is easy to show that

$$X_1 + Y_1 = -\frac{\tau}{2} \sum_{1\le j\le N} \int_{I_j} \partial_x f'(u^n)(\xi_u^n)^2\, dx - \sum_{1\le j\le N} f'(u^n_{j+\frac12})(\bar\xi_u^n - \xi_u^{n,\star})_{j+\frac12} [\xi_u^n]_{j+\frac12}\tau.$$

If $f'(u^n) < 0$ on $I_j \cup I_{j+1}$, then the reference values are $u_h^{n,\star} = u_h^{n,+}$ and $\xi_u^{n,\star} = \xi_u^{n,+}$, respectively; otherwise, $u_h^{n,\star} = u_h^{n,-}$ and $\xi_u^{n,\star} = \xi_u^{n,-}$. If the sign of $f'(u^n)$ is unchanged on $I_j \cup I_{j+1}$, then it is easy to see that $f'(u^n_{j+1/2})(\bar\xi_u^n - \xi_u^{n,\star})_{j+1/2} = \frac12 |f'(u^n_{j+1/2})|[\xi_u^n]_{j+1/2}$; if the sign of $f(u^n)$ is changed on $I_j \cup I_{j+1}$, then a simple Taylor expansion implies that $|f'(u^n_{j+1/2}) - |f'(u^n_{j+1/2})|| \le C_\star h$, even if $f'(u^n_{j+1/2}) < 0$. Thus we have

$$X_1 + Y_1 \le C_\star ||\xi_u^n||^2 \tau - |f'(u^n)|[\xi_u^n]^2\tau. \tag{6.18a}$$

The inverse properties (i) and (ii) imply the estimate to $X_2 + Y_2$ in the form

$$X_2 + Y_2 \le Ch^{-1}||\xi_u^n||_\infty ||\xi_u^n||^2\tau. \tag{6.18b}$$

Noticing the estimate to the term π (see subsection 6.3), we have, by virtue of the inverse property (ii), that

$$X_3 + Y_3 \le Ch^{2k+2}\tau + C||\xi_u^n||^2\tau. \tag{6.18c}$$

The estimate to the remaining terms are the same as (6.11d), (6.11e), and (6.11f).

Thus, by combining the above inequalities we obtain the estimate (5.26a). Along the same line we can also get (5.26b). Thus we complete the proof of Lemma 5.7. □

REFERENCES

[1] S. Adjerid, K. D. Devine, J. F. Flaherty, and L. Krivodonova, *A posteriori error estimation for discontinuous Galerkin solutions of hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1097–1112.

[2] P. G. Ciarlet, *Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[3] B. Cockburn, *An introduction to the discontinuous Galerkin method for convection-dominated problems*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, B. Cockburn, C. Johnson, C.-W. Shu and E. Tadmor (Editor: A. Quarteroni), Lecture Notes in Math. 1697, Springer, Berlin, 1998, pp. 151–268.

[4] B. Cockburn and P.-A. Gremaud, *Error estimates for finite element methods for scalar conservation laws*, SIAM J. Numer. Anal., 33 (1996), pp. 522–554.

[5] B. Cockburn, P.-A. Gremaud, and J. X. Yang, *A priori error estimates for numerical methods for scalar conservation laws Part* III: *Multidimensional flux-splitting monotone schemes on non-Cartesian grids*, SIAM J. Numer. Anal., 35 (1998), pp. 1775–1803.

[6] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV: *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[7] B. Cockburn, S.-Y. Lin, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* III: *One dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[8] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws* II: *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[9] B. Cockburn and C.-W. Shu, *The Runge-Kutta local projection $\mathbb{P}^1$-discontinuous Galerkin method for scalar conservation laws*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 337–361.

[10] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin finite element method for conservation laws* V: *Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[11] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM. J. Numer. Anal., 35 (1998), pp. 2440–2463.

[12] B. Cockburn and C.-W. Shu, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[13] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[14] G.-S. Jiang and C.-W. Shu, *On cell entropy inequality for discontinuous Galerkin methods*, Math. Comp., 62 (1994), pp. 531–538.

[15] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[16] H.-O. Kreiss, T. A. Manteuffel, B. Swartz, B. Wendroff, and A. B. White, Jr., *Supra-convergent schemes on irregular grids*, Math. Comp., 47 (1986), pp. 537–554.

[17] P. LeSaint and P. A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–145.

[18] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM. J. Numer. Anal., 21 (1984), pp. 217–235.

[19] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[20] G. R. Richter, *An optimal-order error estimate for the discontinuous Galerkin method*, Math. Comp., 50 (1988), pp. 75–88.

[21] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[22] G. Strang, *Accurate partial difference methods* II. *Non-linear problems*, Numer. Math., 6 (1964), pp. 37–46.

[23] L.-A. Ying, *A second order explicit finite element scheme to multi–dimensional conservation laws and its convergence*, Sci. China Ser. A, 43 (2000), pp. 945–957.

[24] M. Zhang and C.-W. Shu, *An analysis of three different formulations of the discontinuous Galerkin method for diffusion equations*, Math. Models Methods Appl. Sci., 13 (2003), pp. 395–413.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 667–685

A STABILIZED DOMAIN DECOMPOSITION METHOD WITH NONMATCHING GRIDS FOR THE STOKES PROBLEM IN THREE DIMENSIONS*

FAKER BEN BELGACEM†

Abstract. We present and study a nonconforming domain decomposition method for the discretization of the *three-dimensional* Stokes problem in the velocity-pressure formulation. The approximation is based on some local mixed finite elements for nonmatching tetrahedral grids. The aim pursued is a systematic construction of the mortared discrete velocity space, the pressure being not subjected to any matching constraints across the interfaces. Using the bubble stabilization techniques, applied in Brezzi and Marini's paper to the three fields method [*Math. Comp.*, 70 (2001), pp. 911–934], allows us to define an algorithm which is easy to implement. The numerical analysis relies on the pressure-splitting argument of Boland and Nicolaides and allows us to establish an *inf-sup* condition with a constant that does not depend on the mesh size or on the total number of the subdomains. Then, by the Berger–Scott–Strang lemma written down for our saddle point system we derive optimal accuracy results.

Key words. mortar method, nonmatching meshes, bubble stabilization, mixed formulation, *inf-sup* condition, Boland and Nicolaides argument

AMS subject classifications. 65N12, 65N30, 65N55

DOI. 10.1137/S0036142900368824

1. Introduction and the functional framework.

The approximation of the Sobolev space H^1 in *three dimensions* by Lagrangian finite elements based on the concept of domain decomposition with nonmatching grids is a source of difficulties, especially for tetrahedral elements (see [5], [10], [21], [39]). Substantial research originated from the numerical modeling of the "weak continuity" versus the "strong continuity," to prevent numerical locking across the interfaces. Resorting to the nonconforming domain decomposition method is an opportunity for enhancing the numerical simulation of partial differential equations with nonregular coefficients and/or set on nonsmooth domains. The analysis of complex phenomena is often based on large scale numerical simulations using different environments and, then, can take advantage of any affordable tool to easily and efficiently match different meshes. Moreover, the high degree of parallelization achieved by the domain decomposition concept together with the permanent growth of performing iterative substructuring and multigrid procedures to compute nonconforming multidomain solutions (see [3], [2], [21], [47]) are encouraging us to study their extension to the higher dimension.

Among the class of existing multidomain techniques with nonmatching grids, the mortar element method has become very popular (see [8] for a *biography* of the method and its applications). Introduced by Bernardi, Maday, and Patera in [15]—we refer also to the earlier work of [13]—it was improved and widely analyzed for different variational discretizations such as finite elements, spectral elements, and wavelets for the Poisson problem in two dimensions. A nonexhaustive bibliography restricted to the second order elliptic problems is provided by [15], [33], [5], [6], [48], [21], [29],

*Received by the editors March 6, 2000; accepted for publication (in revised form) July 7, 2003; published electronically May 17, 2004. This work was completed during the author's stay for one year, as a CNRS researcher, at the Laboratoire Jacques-Louis Lions, Paris VI (UMR CNRS 7598).
http://www.siam.org/journals/sinum/42-2/36882.html

†Mathématiques pour l'Industrie et la Physique (UMR CNRS 5640), Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France (belgacem@mip.ups-tlse.fr).

[17], [18], [19]. The mortar finite element method is successful in three dimensions for hexahedral elements. An appropriate matching of hexahedral Lagrangian elements is described in [5] (see also [11]); the proof of the optimality of this approach and its implementation issues are detailed in [1], [44]. In contrast, when tetrahedral finite elements, which are of a high importance in practice, are employed, enforcing the matching constraints turns out to be more complicated. A suggestion is made in [10] (see also [22], [39] for further comments), where the construction of the matching functions—the Lagrange multipliers of the primal hybrid formulation—across the interfaces is not systematic. Even though the numerical analysis results in an optimal convergence rate for affine tetrahedral finite elements, developing a computing code taking into account this matching is not automatic. For instance, if a refining of the mesh or a remeshing is needed, then the practitioner has to update the reconstruction of the Lagrange multipliers.

An alternative allowing the use of more natural finite element multipliers in such a context is to combine the mortar geometrical features with the functional stabilization techniques used by Brezzi's team. The selected stabilization tool consists of adding to the local affine finite element discretization some boundary bubble functions as in [24] (see also [26], [28], [16]). The main difference is that the support of a bubble function is an entire triangle. This enables us to use piecewise constant Lagrange multipliers which are really easy to handle. The definition of the approximation spaces becomes clear and the implementation gains in facility. The modification proposed here on the bubble stabilization has a *major* impact as *it makes things coherent and conceptually easier* when we are involved with the extension to three dimensions of the nonconforming multidomain algorithms.

The core of this work is to deal with the *stabilized domain decomposition method applied to* the three-dimensional incompressible Stokes system. The local discrete spaces are obtained from some mixed finite elements satisfying the Babuška–Brezzi criterion (see [23], [25]). The global velocity finite element space is enriched by some stabilizing boundary bubble functions in order to enforce the matching conditions across the interfaces while the pressure space is free of any constraints. The verification of the *inf-sup* condition for the global discretization relies on a trick due to Boland and Nicolaides (see [20]). This contribution is an improvement of the results of [34], [7], and [9] (dealing with the two-dimensional mortar element method) since we prove that the constant involved in the *inf-sup* condition does not depend on the size of the meshes or on the total number of the subdomains.

A brief outline of the paper is as follows. Section 3 is a description of the stabilized mortared velocity and pressure finite elements spaces and a presentation of the discrete problem. In section 4 we use the Boland–Nicolaides argument to show an optimal Babuška–Brezzi condition. Section 5 is dedicated to the error estimate of the stabilized approximation and is based on abstract results of the linear saddle point theory.

Notation. We need to set up some notation and to recall some functional tools necessary to our analysis. Let $\Omega \subset \mathbb{R}^3$ be a Lipschitz domain of size $\mathcal{O}(1)$; a generic point of Ω is denoted $\boldsymbol{x}$. The symbol $L^2(\Omega)$ stands for the Lebesgue space. Throughout this work we make a constant use of the standard Sobolev space $H^m(\Omega)$, $m \geq 1$, provided with the norm

$$\|\psi\|_{H^m(\Omega)} = \left(\sum_{0 \leq |\alpha| \leq m} \|\partial^\alpha \psi\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}},$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ is a multi-index in $\mathbb{N}^3$ and ∂^α represents the corresponding partial derivative ($H^0(\Omega) = L^2(\Omega)$). The fractional order Sobolev space $H^\lambda(\Omega), \lambda \in \mathbb{R}_+ \setminus \mathbb{N}$, is defined by the norm

$$\|\psi\|_{H^\lambda(\Omega)} = \left(\|\psi\|^2_{H^m(\Omega)} + \sum_{|\alpha|=m} \int_\Omega \int_\Omega \frac{(\partial^\alpha \psi(\boldsymbol{x}) - \partial^\alpha \psi(\boldsymbol{y}))^2}{|\boldsymbol{x} - \boldsymbol{y}|^{3+2\theta}} \, d\boldsymbol{x} \, d\boldsymbol{y} \right)^{\frac{1}{2}},$$

where $\lambda = m + \theta$, m is the integer part of λ, and $\theta \in]0, 1[$ is the decimal part (see [4], [37]).

For any portion γ of the boundary $\partial\Omega$, with size $\mathcal{O}(1)$, the Hilbert space $H^{\frac{1}{2}}(\gamma)$ is associated with the norm

$$\|\psi\|_{H^{\frac{1}{2}}(\gamma)} = \left(\|\psi\|^2_{L^2(\gamma)} + \int_\gamma \int_\gamma \frac{(\psi(\boldsymbol{x}) - \psi(\boldsymbol{y}))^2}{|\boldsymbol{x} - \boldsymbol{y}|^3} \, d\Gamma d\Gamma \right)^{\frac{1}{2}}.$$

The symbol $d\Gamma$ denotes the surface measure on $\partial\Omega$. This space may be obtained by a Hilbertian interpolation argument between $H^1(\gamma)$ and $L^2(\gamma)$, i.e., $H^{\frac{1}{2}}(\gamma) = [H^1(\gamma), L^2(\gamma)]_{\frac{1}{2}}$ (see [40]). The space $H^{-\frac{1}{2}}(\gamma)$ stands for the dual space of $H^{\frac{1}{2}}(\gamma)$, and the duality pairing is denoted by $\langle \cdot, \cdot \rangle_{\frac{1}{2},\gamma}$. The special space $H^{\frac{1}{2}}_{00}(\gamma)$ results from the interpolation between $H^1_0(\gamma)$ and $L^2(\gamma)$. Finally, denoting by L a portion of the boundary $\partial\gamma$, then $H^1_0(\gamma, L)$ stands for $\{\psi \in H^1(\gamma), \psi_{|L} = 0\}$ and $H^{\frac{1}{2}}_{00}(\gamma, L) = [H^1_0(\gamma, L), L^2(\gamma)]_{\frac{1}{2}}$.

If a domain Ω is of size $\mathcal{O}(H)$, then $|\Omega| = \mathcal{O}(H^3)$; a more suitable norm contains a certain scaling factor. The local H^1-norms, which will be of permanent use in this paper, are defined by

$$\|\psi\|_{H^1(\Omega)} = (|\psi|^2_{H^1(\Omega)} + H^{-2}\|\psi\|^2_{L^2(\Omega)})^{\frac{1}{2}}.$$

Bold Latin letters like $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{f}, \ldots$, indicate vector valued functions, while the capital ones (e.g., $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{V}, \ldots$) are functional sets involving vector fields.

2. The variational Stokes problem. Let $\boldsymbol{f}$ be a given force in $(L^2(\Omega))^3$. We are interested in the steady Stokes problem with homogeneous Dirichlet boundary conditions which may be viewed as the "worst" case with respect to the analysis we carry out. Indeed, regarding the *inf-sup* condition, in both continuous and discrete levels, it is well known that Dirichlet-type conditions raise more technicalities than any other type of classical boundary conditions (Neumann, Robin, mixed, etc.). Then, the Stokes–Dirichlet system consists of *finding a velocity vector field* $\boldsymbol{u}$ *and a pressure scalar field* p *such that*

$$-\Delta \boldsymbol{u} + \nabla p = \boldsymbol{f} \quad \text{in } \Omega, \tag{2.1}$$

$$\operatorname{div} \boldsymbol{u} = 0 \quad \text{in } \Omega, \tag{2.2}$$

$$\boldsymbol{u} = 0 \quad \text{on } \Gamma. \tag{2.3}$$

The standard mixed variational formulation is based on a Green integration formula and is written as follows: find $(\boldsymbol{u}, p) \in H^1_0(\Omega)^3 \times L^2_0(\Omega)$ such that

$$\int_\Omega \nabla \boldsymbol{u} \cdot \nabla \boldsymbol{v} \, d\boldsymbol{x} + b(\boldsymbol{v}, p) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} \, d\boldsymbol{x} \quad \forall \boldsymbol{v} \in H^1_0(\Omega)^3, \tag{2.4}$$

$$b(\boldsymbol{u}, q) = 0 \quad \forall q \in L^2_0(\Omega). \tag{2.5}$$

The bilinear form $b(\cdot,\cdot)$ defined over $H_0^1(\Omega)^3 \times L_0^2(\Omega)$ is given by

$$b(\boldsymbol{v},q) = -\int_\Omega (\operatorname{div}\boldsymbol{v}) q \, d\boldsymbol{x}. \tag{2.6}$$

As is well known, $b(\cdot,\cdot)$ satisfies an *inf-sup* condition with a positive constant α (see [25], [36]), $\forall q \in L_0^2(\Omega)$:

$$\inf_{q\in L^2(\Omega)} \sup_{\boldsymbol{v}\in H_0^1(\Omega)^3} \frac{b(\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{H^1(\Omega)^3}\|q\|_{L^2(\Omega)}} \geq \alpha. \tag{2.7}$$

Using the saddle point theory of [23] (see also [36], [25]), problem (2.4)–(2.5) is well posed.

PROPOSITION 2.1. *The variational Stokes problem* (2.4)–(2.5) *has a unique solution* $(\boldsymbol{u},p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega)$*, and we also have the stability condition*

$$\|\boldsymbol{u}\|_{H^1(\Omega)^3} + \|p\|_{L^2(\Omega)} \leq C\|\boldsymbol{f}\|_{L^2(\Omega)^3}. \tag{2.8}$$

3. A mixed mortar finite element discretization. The framework of the stabilized domain decomposition algorithm proceeds by breaking up the domain Ω, where the partial differential equation is to be solved, into k^* nonoverlapping subdomains that are assumed to be polygonally shaped for simplicity:

$$\overline{\Omega} = \bigcup_{k=1}^{k^*} \overline{\Omega}_k \quad \text{with } \Omega_k \cap \Omega_\ell = \emptyset \text{ if } k \neq \ell,$$

where Ω_k is a polyhedron with planar polygonal faces. To prevent high technicalities, we shall analyze here only the case of conforming decompositions, meaning that when the subdomains are considered as macroelements, the intersection of two closed subdomains $\overline{\Omega}_k \cap \overline{\Omega}_\ell$ so as the intersection $\partial\Omega \cap \partial\Omega_k$ is either empty, or reduced to a common vertex, to a common edge, or to a common face. When two subdomains Ω_k and Ω_ℓ, $k < \ell$, are adjacent, $\Gamma_{k\ell}$ is the common interface. The unit outward normal of $\partial\Omega_k$ is $\boldsymbol{n}_k$. The index $k\ell$ is meaningless when Ω_k and Ω_ℓ do not share any common face. All the subdomains are assumed to be of comparable size $\mathcal{O}(H)$; that is, the length of any edge of any Ω_k is of order $\mathcal{O}(H)$ and therefore $|\Omega_k| = \mathcal{O}(H^3)$ and $|\Gamma_{k\ell}| = \mathcal{O}(H^2)$.

With each subdomain Ω_k we then associate a triangulation $\mathscr{T}_h^{\Omega_k}$ which is regular in the sense of [31] and made of tetrahedral elements with a maximum size h_k. We set $h = (h_k)_k$ to be the discretization parameter. The trace of $\mathscr{T}_h^{\Omega_k}$ on the boundary $\partial\Omega_k$ results in a regular (two-dimensional) triangulation $\mathscr{T}_h^{\partial\Omega_k}$ made of triangular elements (that are all entire faces of an element of the triangulation of Ω_k). The $\mathscr{T}_h^{\partial\Omega_k}$ is supposed compatible with the faces of Ω_k in the sense that each face $\Gamma_{k\ell}$ inherits from $\mathscr{T}_h^{\partial\Omega_k}$ a triangulation $\mathscr{T}_h^{\Gamma_{k\ell}}$. Note that, since the triangulations on two adjacent subdomains are independent, $\Gamma_{k\ell}$ is provided with two different and independent (two-dimensional) meshes $\mathscr{T}_h^{\Gamma_{k\ell}}$ (the trace of $\mathscr{T}_h^{\partial\Omega_k}$) and $\mathscr{T}_h^{\Gamma_{\ell k}}$ (the trace of $\mathscr{T}_h^{\partial\Omega_\ell}$). For some technical reasons that will appear later, we need to introduce some new notation taken from [14]. For any $T \in \mathscr{T}_h^{\partial\Omega_k}$, the macroelement Δ_T is the union of all elements of $\mathscr{T}_h^{\partial\Omega_k}$ which share at least a corner with T. Due to the regularity of $\mathscr{T}_h^{\partial\Omega_k}$, the number of elements T' contained in Δ_T is bounded by a constant independent of h_k. We assume also that on any face $\Gamma_{k\ell}$ the meshes $\mathscr{T}_h^{\partial\Omega_k}$ (or $\mathscr{T}_h^{\Gamma_{k\ell}}$) and $\mathscr{T}_h^{\partial\Omega_\ell}$ (or $\mathscr{T}_h^{\Gamma_{\ell k}}$) are locally comparable in the sense that (see Figure 3.1)

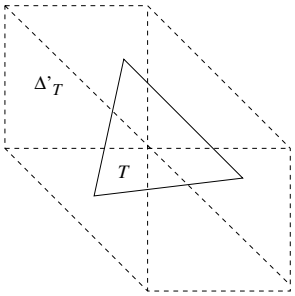


FIG. 3.1. Δ'_T—*dashed lines; for the triangle* T—*solid lines.*

($\mathrm{H_i}$) for any $T \in \mathscr{T}_h^{\partial\Omega_k}$, if Δ'_T is the "minimal" union of all elements of the opposite triangulation $\mathscr{T}_h^{\partial\Omega_\ell}$ such that $T \subset \Delta'_T$, then there exists a constant τ that does not depend on the parameters (h_k, h_ℓ) such that card $\Delta'_T \leq \tau$, where card is the cardinality;

($\mathrm{H_{ii}}$) there exists a constant τ' independent of (h_k, h_ℓ) such that $h_{T'} \leq \tau' h_T$ $\forall T' \in \Delta'_T$.

Remark 3.1. The hypotheses on the meshes $(\mathscr{T}_h^{\partial\Omega_k})_k$ are less stringent than those used in [26], where the authors assumed that these meshes are quasi-uniform and globally comparable, i.e., $\tau h_\ell \leq h_k \leq \tau' h_\ell$. Of course these kinds of meshes satisfy criteria ($\mathrm{H_i}$) and ($\mathrm{H_{ii}}$). Our hypothesis allows the use of some refined meshes—needed in adaptive methods—provided that the "opposite" new elements (in both sides), so to speak, generated by the refining process in both sides of each face $\Gamma_{k\ell}$ are of equivalent sizes so that criterion ($\mathrm{H_{ii}}$) is observed. As a matter of fact, our feeling is similar to that expressed in [26, Remark 1], *"... we believe that the above assumptions are only technical ones, and that the results (and the applicability of the whole method) are valid in more general circumstances."*

The present study is restricted to the configurations where the finite elements approximation of the velocity field within each subdomain is provided by continuous piecewise affine functions. This is the case for the Bercovier–Pironneau $\mathcal{P}_1\mathrm{iso}\mathcal{P}_2/\mathcal{P}1$ element or Arnold–Brezzi–Fortin bubble finite element (see [43], [25]). In order to explain the main ideas of our construction and analysis, we prefer to focus on the second finite element method even though, seemingly, the first one is currently more employed. However, the analysis extends in the same way to the Bercovier–Pironneau elements. For any $\kappa \in \mathscr{T}_h^{\Omega_k}$ and any $r \in \mathbb{N}$, let $\mathcal{P}_r(\kappa)$ stand for the set of polynomials of total degree $\leq r$; $(\boldsymbol{x}_{\kappa,i})_{1\leq i\leq 4}$ are the vertices of κ, and $\lambda_{\kappa,i}$ is the barycentric coordinate associated with $\boldsymbol{x}_{\kappa,i}$. We then define the standard volumic bubble function as

$$\varphi_\kappa(\boldsymbol{x}) = \prod_{i=1}^{4} \lambda_{\kappa,i}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \kappa$$

and set $\mathcal{P}_B(\kappa) = \mathcal{P}_1(\kappa) \oplus \mathbb{R}\varphi_\kappa$. Then we introduce the finite dimension vector valued spaces

$$\boldsymbol{Y}_h(\Omega_k) = \{\boldsymbol{v}_h \in \mathscr{C}(\overline{\Omega}_k)^3,\ \forall \kappa \in \mathscr{T}_h^{\Omega_k},\ \boldsymbol{v}_{h|\kappa} \in \mathcal{P}_B(\kappa)^3,\ \boldsymbol{v}_{h,k|\partial\Omega_k\cap\partial\Omega} = 0\}.$$

The notation $\boldsymbol{Y}_h(\partial\Omega_k)$ is used for the space of the functions defined on $\partial\Omega_k$ as the trace of all those in $\boldsymbol{Y}_h(\Omega_k)$.

The purpose is to choose the global velocity finite element approximation that consists of functions whose restriction over each Ω_k belongs to $\boldsymbol{Y}_h(\Omega_k)$. Since the

interface is provided with two independent meshes, the constraint of continuity of the global function over Ω is not compatible with good approximation properties of the discrete space as it would "block" all degrees of freedom over the interfaces $(\Gamma_{k\ell})_{k\ell}$ which results in a numerical locking phenomenon. To avoid this, a nonconforming matching, based on the mortar concept, is proposed in [10] and studied under the assumption that the interfacial meshes $(\mathscr{T}_h^{\Gamma_{k\ell}})_{k\ell}$ are structured and quasi-uniform. Despite the optimality of the results proven therein (see also [22]), the problem with that matching comes from the difficulty of constructing the Lagrange multipliers space. Indeed, for tetrahedral elements this construction is not systematic and may generate some complication for the programming work. The purpose of what follows is to combine the stabilizing techniques with the mortar concept to build-up a comprehensive matching while preserving the optimality to make sure that the method can be easily implemented.

Enforcing the matching conditions for tetrahedral finite elements across a given interface requires deciding which is the mortar side (master side); the other being the nonmortar side (slaved side). For clarity, assume that for any $\Gamma_{k\ell}$, $k<\ell$, the subdomain Ω_k provides the mortar and the subdomain Ω_ℓ the nonmortar. Thus, in $\Gamma_{k\ell}$, we have two grids; when associated with $\mathscr{T}_h^{\Gamma_{k\ell}}$ it is a mortar and if associated with $\mathscr{T}_h^{\Gamma_{\ell k}}$ it is nonmortar. Henceforth, we call $\mathscr{T}_h^{\Gamma_{k\ell}}$, $k<\ell$, the mortar mesh and $\mathscr{T}_h^{\Gamma_{\ell k}}$, $k<\ell$, the nonmortar mesh. The stabilization approach consists of enriching the discrete space $\boldsymbol{Y}_h(\Omega_k)$ by some boundary bubble functions defined on the nonmortar side of each interface $\Gamma_{k\ell}$, $k<\ell$, then endowed with the mesh $\mathscr{T}_h^{\Gamma_{\ell k}}$. Let a triangular element be $(T\in\mathscr{T}_h^{\Gamma_{\ell k}})$, and let $\kappa=\kappa_T$ be the unique tetrahedron of $\mathscr{T}_h^{\Omega_\ell}$ having T as a face. The vertices of κ that are also vertices of T are $\{\boldsymbol{x}_{\kappa,1},\boldsymbol{x}_{\kappa,2},\boldsymbol{x}_{\kappa,3}\}$. The boundary bubble function we need to use is defined by

$$\varphi_T(\boldsymbol{x})=\frac{60}{|T|}\lambda_{\kappa,1}(\boldsymbol{x})\lambda_{\kappa,2}(\boldsymbol{x})\lambda_{\kappa,3}(\boldsymbol{x})\quad\forall\boldsymbol{x}\in\kappa$$

and extended by zero elsewhere. $|T|$ is the surface of T, the modulation coefficient allows $\int_T\varphi_T\,d\Gamma=1$, and we have

$$|T||\varphi_T|^2_{H^1(\kappa_T)}+\|\varphi_T\|^2_{L^2(\kappa_T)}\le c|T|^{-\frac{1}{2}}. \tag{3.1}$$

The local velocities are taken in the stabilized finite dimension vector valued space

$$\boldsymbol{X}_h(\Omega_k)=\boldsymbol{Y}_h(\Omega_k)\oplus\left(\bigoplus_{\substack{T\in\mathscr{T}_h^{\Gamma_{k\ell}}\\ k>\ell}}\mathbb{R}^3\varphi_T\right),$$

and the local discrete pressure space is defined as follows:

$$Q_h(\Omega_k)=\{q_h\in\mathscr{C}(\overline{\Omega}_k),\ \forall\kappa\in\mathscr{T}_h^{\Omega_k},\ q_{h|\kappa}\in\mathcal{P}_1(\kappa)\}.$$

Setting $\tilde{\boldsymbol{X}}_h(\Omega_k)=\boldsymbol{X}_h(\Omega_k)\cap H_0^1(\Omega_k)^3$ and defining $\tilde{Q}_h(\Omega_k)$ as the subspace of $Q_h(\Omega_k)$ involving the null-averaged pressure, it is standard that the family $\{\tilde{\boldsymbol{X}}_h(\Omega_k),\tilde{Q}_h(\Omega_k)\}$ is div-stable; actually this result is issued from the standard div-stability of $(\boldsymbol{Y}_h(\Omega_k)\cap H_0^1(\Omega_k)^3, Q_h(\Omega_k)\cap L_0^2(\Omega))$. This means that the form $b(\cdot,\cdot)$ satisfies a uniform *inf-sup* condition that there exist a constant $\tilde{\alpha}_k>0$ such that for any $q_h\in\tilde{Q}_h(\Omega_k)$ we can

find $\tilde{\boldsymbol{v}}_h \in \tilde{\boldsymbol{X}}_h(\Omega_k)$ that verifies

$$\int_{\Omega_k} (\operatorname{div} \tilde{\boldsymbol{v}}_h) q_h \, d\boldsymbol{x} = \|q_h\|^2_{L^2(\Omega_k)} \qquad \text{and} \qquad \tilde{\alpha}_k \|\tilde{\boldsymbol{v}}_h\|_{H^1(\Omega_k)^3} \leq \|q_h\|_{L^2(\Omega_k)}. \tag{3.2}$$

The constant $\tilde{\alpha}_k$ is independent of h_k and H even though it does depend on the shape of Ω_k—this *inf-sup* condition is derived from a standard statement, proven for a domain with size $\mathcal{O}(1)$ (see [25]), then extended in our context through an appropriate scaling.

Given these local tools, velocities are taken locally in $\boldsymbol{X}_h(\Omega_k)$ and glued together through the interfaces $(\Gamma_{k\ell})_{k\ell}$ by suitable matching conditions. Expressing these conditions requires building some "gluing" functional spaces on the nonmortar side of $\Gamma_{k\ell}$ $(k < \ell)$, then endowed with the triangulation $\mathscr{T}_h^{\Gamma_{\ell k}}$,

$$\boldsymbol{M}_h(\Gamma_{k\ell}) = \{\Psi_h \in L^2(\overline{\Gamma}_{k\ell})^3, \ \forall T \in \mathscr{T}_h^{\Gamma_{\ell k}}, \ \Psi_{h|T} \in \mathcal{P}_0(T)^3\}.$$

By a duality argument, the following estimate holds: $\forall \Psi \in H^{\frac{1}{2}}(\Gamma_{k\ell})$,

$$\inf_{\Psi_h \in \boldsymbol{M}_h(\Gamma_{k\ell})} \|\Psi - \Psi_h\|_{H^{-\frac{1}{2}}(\Gamma_{k\ell})^3} \leq C h_\ell |\Psi|_{H^{\frac{1}{2}}(\Gamma_{k\ell})^3}. \tag{3.3}$$

The purpose of what follows is to use the stabilizing techniques applied to the mortar element method to build-up an easy matching for the velocity. The global velocity approximation space is then given by

$$\boldsymbol{X}_h(\Omega) = \Big\{ \boldsymbol{v}_h = (\boldsymbol{v}_{h,k})_k \in L^2(\Omega)^3 ; \boldsymbol{v}_{h,k} \in \boldsymbol{X}_h(\Omega_k), \, \forall (k,\ell), \, k < \ell, \\ \forall \Psi_h \in \boldsymbol{M}_h(\Gamma_{k\ell}), \int_{\Gamma_{k\ell}} (\boldsymbol{v}_{h,k} - \boldsymbol{v}_{h,\ell}) \cdot \Psi_h \, d\Gamma = 0 \Big\}.$$

Since it is not embedded in $H^1_0(\Omega)^3$, the space $\boldsymbol{X}_h(\Omega)$ is equipped with the Hilbertian broken seminorm (which is actually a norm)

$$\|\boldsymbol{v}_h\|_{\boldsymbol{X}} = \left(\sum_{k=1}^{k^*} |\boldsymbol{v}_{h,k}|^2_{H^1(\Omega_k)^3} \right)^{\frac{1}{2}}.$$

We will denote by C and c generic positive constants. These constants take different values in different occurrences but are always independent of the mesh parameter h and independent of the subdomain size H.

The discretization is nonconforming; nevertheless under assumptions $(\mathrm{H_i})$ and $(\mathrm{H_{ii}})$ the space $\boldsymbol{X}_h(\Omega)$ has suitable approximation properties, the proof of which is postponed to the appendix.

PROPOSITION 3.2. *For any $\boldsymbol{v} \in H^1_0(\Omega)^3$ with $\boldsymbol{v}_k = \boldsymbol{v}_{|\Omega_k} \in H^2(\Omega_k)^3$, it holds that*

$$\inf_{\boldsymbol{v}_h \in \boldsymbol{X}_h(\Omega)} \|\boldsymbol{v} - \boldsymbol{v}_h\|_{\boldsymbol{X}} \leq C \left(\sum_{k=1}^{k^*} h_k^2 |\boldsymbol{v}_k|^2_{H^2(\Omega_k)^3} \right)^{\frac{1}{2}}. \tag{3.4}$$

Remark 3.3. Different choices of the bubble function that leads to the same asymptotic convergence result of Proposition 3.2 are possible. For instance, the function $\varphi_T \in \mathcal{P}_3(\kappa_T)$ can be replaced by a continuous piecewise affine function constructed as follows: the triangle T is broken-up into three smaller ones $(t_{T,i})_{1 \leq i \leq 3}$

sharing the same vertex, the center of T. The tetrahedra determined by the basis $t_{T,i}$ and the internal vertex of κ_T are denoted by $\kappa_{T,i}$; the union of the three resulting tetrahedra $(\kappa_{T,i})_{1\le i\le 3}$ coincides with κ_T. Then, we can choose $\tilde{\varphi}_T$ to be continuous, to vanish on $\partial\kappa_T$, so that $\tilde{\varphi}_{T|\kappa_{T,i}} \in \mathcal{P}_1(\kappa_{T,i})$ and $\int_T \tilde{\varphi}_T \, d\Gamma = 1$. More general bubble functions could be used, provided that their integral on T is one.

Remark 3.4. In the three-dimensional case, a coherent construction of the stabilized finite element space $\boldsymbol{X}_h(\Omega)$ is made possible thanks to the mortaring concept applied to the nonmatching meshes. The original stabilization techniques developed in [24] (see also [26]) for two-dimensional three-field domain decomposition methods with nonmatching grids can hardly be extended to three dimensions. The main reason is that choosing completely independent meshes for $\boldsymbol{X}_h(\Omega)$ and $(\boldsymbol{M}_h(\Gamma_{k\ell}))_{k\ell}$, then merging them for bubble stabilization, would result in a very complicated new mesh with arbitrarily shaped elements that can hardly be used for finite element computations.

The discrete pressure is free of any matching constraints at the interfaces, and the pressure space is defined to be

$$Q_h(\Omega) = \{q_h = (q_{h,k})_k \in L^2_0(\Omega),\ q_{h,k} \in Q_h(\Omega_k)\}.$$

Endowed with the $L^2(\Omega)$-norm, $Q_h(\Omega)$ provides fairly good approximation of regular pressures; that is, the following optimal result holds (see [31]): $\forall q \in L^2_0(\Omega)$ with $q_k = q_{|\Omega_k} \in H^1(\Omega_k)$,

$$\inf_{q_h \in Q_h(\Omega)} \|q - q_h\|_{L^2(\Omega)} \le C \left(\sum_{k=1}^{k^*} h_k^2 |q_k|^2_{H^1(\Omega_k)} \right)^{\frac{1}{2}}. \tag{3.5}$$

Let us stress the fact that only the velocity space is mortared, while the pressure is not subjected to any particular constraints across the interfaces.

We are now in position to formulate and investigate the discrete problem, which consists of *finding* $(\boldsymbol{u}_h, p_h) \in \boldsymbol{X}_h(\Omega) \times Q_h(\Omega)$ *satisfying*

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, p_h) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}_h \, d\boldsymbol{x} \quad \forall \boldsymbol{v}_h \in \boldsymbol{X}_h(\Omega), \tag{3.6}$$

$$b(\boldsymbol{u}_h, q_h) = 0 \quad \forall q_h \in Q_h(\Omega), \tag{3.7}$$

where we have set

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h) = \sum_{k=1}^{k^*} \int_{\Omega_k} \nabla \boldsymbol{u}_{h,k} \cdot \nabla \boldsymbol{v}_{h,k} \, d\boldsymbol{x} \quad \forall \boldsymbol{u}_h, \boldsymbol{v}_h \in \boldsymbol{X}_h(\Omega),$$

$$b(\boldsymbol{v}_h, q_h) = -\sum_{k=1}^{k^*} \int_{\Omega_k} (\operatorname{div} \boldsymbol{v}_{h,k}) q_{h,k} \, d\boldsymbol{x} \quad \forall \boldsymbol{v}_h \in \boldsymbol{X}_h(\Omega),\ \forall q_h \in Q_h(\Omega).$$

A notation abuse is made here for $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ because the expression given here is an extension of (2.6). Deriving existence, uniqueness, and stability results is based on the saddle point theory. The bilinear form $a(\cdot,\cdot)$ is uniformly continuous on $\boldsymbol{X}_h(\Omega)$ and is positive-definite. The main remaining point is the verification of an *inf-sup* condition for $b(\cdot,\cdot)$ on the discrete spaces $\boldsymbol{X}_h(\Omega) \times Q_h(\Omega)$ with a constant that does not depend on h or on the total number k^* of the subdomains. In what follows, the notations C and c (indexed or not by k) are constants that do not depend on the size of the subdomains nor on the mesh size.

4. Boland–Nicolaides argument and the discrete *inf-sup* condition. The core of this work is to state a Babuška–Brezzi *inf-sup* condition on $b(\cdot,\cdot)$ restricted to $\boldsymbol{X}_h(\Omega)\times Q_h(\Omega)$. The proof is based on a tricky decomposition of the pressure space introduced by Boland and Nicolaides (see [20]). This reduces the problem to two simpler ones: the evaluation of the local *inf-sup* constants and the evaluation of a global *inf-sup* constant on reduced spaces. The same idea was used for the mortar element approximation of the Stokes equations for two-dimensional finite elements in [7] and for two- and three-dimensional spectral elements in [9]. The results given therein may be considered partially sharp; it is only proven that the *inf-sup* constant behaves like the local constants, but they do not answer the question of whether this constant grows with the total number k^* of the subdomains or not. Here we provide the answer: *it does not*. As already underlined in [8] for the Poisson problem, this point is of major importance for iterative preconditioned substructuring solvers (see [41], [42], [35]) as the iterations number increases with the *inf-sup* constant. Let us begin by proving an optimal result for $\boldsymbol{X}_h(\Omega)$ and the space of piecewise constant pressures:

$$\check{Q}(\Omega)=\left\{\check{q}=(\check{q}_k)\in\mathbb{R}^{k^*},\ (\check{q},1)_{L^2(\Omega)}=\sum_{k=1}^{k^*}\check{q}_k|\Omega_k|=0\right\}.$$

The main difficulty is inherent to the shape of the subdomains $(\Omega_k)_{1\le k\le k^*}$ which may be pretty complicated. For instance, in the case where the domain is decomposed into tetrahedral or hexahedral subdomains, the proof of the *inf-sup* condition is directly issued from standard conforming mixed finite element results (see [43], [25]). To handle the general case we first need a preliminary result.

LEMMA 4.1. *For any interface $\Gamma_{k\ell}$ $(k<\ell)$ there exists $\boldsymbol{v}_h^{k\ell}\in\boldsymbol{X}_h(\Omega)$ with a support contained in $\overline{\Omega}_k\cup\overline{\Omega}_\ell$ and such that*

$$\|\boldsymbol{v}_h^{k\ell}\|_{\boldsymbol{X}}\le\gamma_{k\ell}H^{-\frac{3}{2}},$$
$$\int_{\Gamma_{k\ell}}\boldsymbol{v}_{h,k}^{k\ell}\cdot\boldsymbol{n}_k\,d\Gamma=-\int_{\Gamma_{k\ell}}\boldsymbol{v}_{h,\ell}^{k\ell}\cdot\boldsymbol{n}_\ell\,d\Gamma=1.$$

The constant $\gamma_{k\ell}$ is independent of H.

Proof. Let $\Gamma_{k\ell}\,(k<\ell)$ be a fixed interface, and assume for a while that $|\Omega_k|\approx|\Omega_k|=\mathcal{O}(1)$. Consider the affine finite element trace space built on the mortar mesh $\mathscr{T}_h^{\Gamma_{k\ell}}$,

$$\boldsymbol{Y}_h(\Gamma_{k\ell})=\{\Phi_h\in\mathscr{C}(\overline{\Gamma}_{k\ell})^3,\ \forall T\in\mathscr{T}_h^{\Gamma_{k\ell}},\ \Phi_{h|T}\in\mathcal{P}_1(T)^3,\ \Phi_{h|\partial\Gamma_{k\ell}}=0\}.$$

Notice that $\prod_{k<\ell}(\boldsymbol{Y}_h(\Gamma_{k\ell}))_{k\ell}$ is the so-called mortar space; we refer to [15] for more details on the mortar terminology. In the same way we construct the space on the nonmortar mesh $\mathscr{T}_h^{\Gamma_{\ell k}}$, which is denoted by $\tilde{\boldsymbol{Y}}_h(\Gamma_{k\ell})$. For conciseness the index $k\ell$ is dropped except for $\Gamma_{k\ell}$ in the remaining part of the proof. If $\tilde{r}_h$ stands for the regularizing Clément operator (see [32] and [14] for the definition and the approximation results) mapping $L^2(\Gamma_{k\ell})^3$ into $\tilde{\boldsymbol{Y}}_h(\Gamma_{k\ell})$, it holds that $\forall\Phi\in H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})$, $\forall T\in\mathscr{T}_h^{\Gamma_{\ell k}}$,

$$\|\Phi-\tilde{r}_h\Phi\|_{L^2(T)}^2\le c|T|^{\frac{1}{2}}|\Phi|^2_{H_*^{\frac{1}{2}}(\Delta_T)},\tag{4.1}$$

$$\|\tilde{r}_h\Phi\|_{L^2(\Gamma_{k\ell})}\le c\|\Phi\|_{L^2(\Gamma_{k\ell})}\qquad\text{and}\qquad|\tilde{r}_h\Phi|_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})}\le c|\Phi|_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})}.\tag{4.2}$$

The space $H_*^{\frac{1}{2}}(\Delta_T)$ coincides with the standard $H^{\frac{1}{2}}(\Delta_T)$ if $\Delta_T \cap \partial\Gamma_{k\ell} = \emptyset$ and is $H_{00}^{\frac{1}{2}}(\Delta_T, \Delta_T \cap \partial\Gamma_{k\ell})$ otherwise. Then, choose $\Phi_h \in \boldsymbol{Y}_h(\Gamma_{k\ell})$ satisfying

$$\int_{\Gamma_{k\ell}} \Phi_h \cdot \boldsymbol{n}_k \, d\Gamma = 1,$$

with $\|\Phi_h\|_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})}$ independent of h. Obviously, such a $\Phi_h^{k\ell}$ exists; moreover, (4.2) implies that $\|\tilde{r}_h\Phi_h\|_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})} \leq C$. We introduce the trivial extensions of Φ_h to $\partial\Omega_k$ and of $(\tilde{r}_h\Phi_h)$ to $\partial\Omega_\ell$ that we still denote $\Phi_h \in \boldsymbol{Y}_h(\partial\Omega_k)$ and $\tilde{r}_h\Phi_h \in \boldsymbol{Y}_h(\partial\Omega_\ell)$. By stable extension (see [14]) we obtain $\boldsymbol{w}_{h,k} \in \boldsymbol{Y}_h(\Omega_k)$ and $\boldsymbol{w}_{h,\ell} \in \boldsymbol{Y}_h(\Omega_\ell)$ that are uniformly bounded:

$$\|\boldsymbol{w}_{h,k}\|_{H^1(\Omega_k)^3} \leq c\|\Phi_h\|_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})}, \tag{4.3}$$

$$\|\boldsymbol{w}_{h,\ell}\|_{H^1(\Omega_\ell)^3} \leq c\|\tilde{r}_h\Phi_h\|_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})} \leq c\|\Phi_h\|_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})}. \tag{4.4}$$

The construction of $\boldsymbol{v}_h$ follows from the correction added to $\boldsymbol{w}_{h,\ell}$,

$$\begin{aligned}\boldsymbol{v}_{h,\ell} &= \boldsymbol{w}_{h,\ell} + \sum_{T \in \mathscr{T}_h^{\Gamma_{k\ell}}} \left(\int_T (\boldsymbol{w}_{h,k} - \boldsymbol{w}_{h,\ell}) \, d\Gamma \right) \varphi_T \\ &= \boldsymbol{w}_{h,\ell} + \sum_{T \in \mathscr{T}_h^{\Gamma_{k\ell}}} \left(\int_T (\Phi_h - \tilde{r}_h\Phi_h) \, d\Gamma \right) \varphi_T,\end{aligned}$$

and we set $\boldsymbol{v}_{h,k} = \boldsymbol{w}_{h,k}$ and $\boldsymbol{v}_h = 0$ outside $\Omega_k \cup \Omega_\ell$. It is readily checked that $\boldsymbol{v}_h \in \boldsymbol{X}_h(\Omega)$ and

$$\int_{\Gamma_{k\ell}} \boldsymbol{v}_{h,k} \cdot \boldsymbol{n}_k \, d\Gamma = \int_{\Gamma_{k\ell}} \Phi_h \cdot \boldsymbol{n}_k \, d\Gamma = -\int_{\Gamma_{k\ell}} \boldsymbol{v}_{h,\ell} \cdot \boldsymbol{n}_\ell \, d\Gamma = 1.$$

What remains to prove is that the added term, henceforth denoted $\boldsymbol{t}_{k\ell}$, is bounded in $H^1(\Omega_\ell)^3$ independently of h_ℓ. Indeed, using (3.1) and (4.1), we have

$$\begin{aligned}|\boldsymbol{t}_{k\ell}|^2_{H^1(\Omega_\ell)^3} &= \sum_{T \in \mathscr{T}_h^{\Gamma_{k\ell}}} \left| \int_T (\Phi_h - \tilde{r}_h\Phi_h) \, d\Gamma \right|^2 |\varphi_T|^2_{H^1(\Omega_\ell)} \\ &\leq c \sum_{T \in \mathscr{T}_h^{\Gamma_{k\ell}}} |T|^{-\frac{3}{2}} \left| \int_T (\Phi_h - \tilde{r}_h\Phi_h) \, d\Gamma \right|^2 \\ &\leq c \sum_{T \in \mathscr{T}_h^{\Gamma_{k\ell}}} |T|^{-\frac{1}{2}} \|\Phi_h - \tilde{r}_h\Phi_h\|^2_{L^2(T)} \\ &\leq c \sum_{T \in \mathscr{T}_h^{\Gamma_{k\ell}}} |\Phi_h|^2_{H_*^{\frac{1}{2}}(\Delta_T)} \leq c|\Phi_h|^2_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})}.\end{aligned}$$

The uniform boundedness of $\boldsymbol{v}_h$ is achieved by (4.2) and is due to the uniform bound if we have a bound for $\|\Phi_h\|_{H_{00}^{\frac{1}{2}}(\Gamma_{k\ell})}$. The more general result of the lemma (i.e., $|\Omega_k| \approx |\Omega_\ell| = \mathcal{O}(H)$) is derived by an easy scaling argument. ∎

Now comes the main result of our study, the div-stability for the reduced family of spaces $(\boldsymbol{X}_h(\Omega), \check{Q}(\Omega))$.

PROPOSITION 4.2. *There exists a constant $\check{\alpha}$ depending neither on the discretization parameter h nor on the size H of the subdomains so that the following* inf-sup *condition holds*:

$$\inf_{\check{q}\in\check{Q}(\Omega)} \sup_{\boldsymbol{v}_h\in\boldsymbol{X}_h(\Omega)} \frac{b(\boldsymbol{v}_h,\check{q})}{\|\boldsymbol{v}_h\|_{\boldsymbol{X}}\|\check{q}\|_{L^2(\Omega)}} \geq \check{\alpha}.$$

Proof. As the *inf-sup* condition is available in the continuous setting (2.7), an equivalent statement of the proposition is to show that for any $\boldsymbol{v} \in H_0^1(\Omega)^3$ there exists $\boldsymbol{v}_h \in \boldsymbol{X}_h(\Omega)$ such that

$$b(\boldsymbol{v}-\boldsymbol{v}_h,\check{q}_h) = 0 \quad \forall \check{q}_h \in \check{Q}(\Omega), \tag{4.5}$$

$$\|\boldsymbol{v}_h\|_{\boldsymbol{X}} \leq \gamma \|\boldsymbol{v}\|_{H^1(\Omega)}, \tag{4.6}$$

where $\gamma > 0$ does not depend on h or H. Then, the constant $\check{\alpha}$ could be taken equal to the quotient $\frac{\alpha}{\gamma}$. Let us set

$$\boldsymbol{v}_h = \sum_{k\ell,\ k<\ell} \left(\int_{\Gamma_{k\ell}} \boldsymbol{v}\cdot\boldsymbol{n}_k \, d\Gamma \right) \boldsymbol{v}_h^{k\ell},$$

where $\boldsymbol{v}_h^{k\ell}$ is provided by Lemma 4.1. Noticing that $\boldsymbol{v}_{h,k|\Gamma_{k\ell}} = \boldsymbol{v}_{h,k|\Gamma_{k\ell}}^{k\ell}$ and $\boldsymbol{v}_{h,\ell|\Gamma_{k\ell}} = \boldsymbol{v}_{h,\ell|\Gamma_{k\ell}}^{k\ell}$, it becomes straightforward that $\forall k\ell$ $(k<\ell)$

$$\int_{\Gamma_{k\ell}} \boldsymbol{v}_{h,k}.\boldsymbol{n}_k \, d\Gamma = \left(\int_{\Gamma_{k\ell}} \boldsymbol{v}\cdot\boldsymbol{n}_k \, d\Gamma \right) \left(\int_{\Gamma_{k\ell}} \boldsymbol{v}_{h,k}^{k\ell}\cdot\boldsymbol{n}_k \, d\Gamma \right) = \int_{\Gamma_{k\ell}} \boldsymbol{v}\cdot\boldsymbol{n}_k \, d\Gamma,$$

$$\int_{\Gamma_{k\ell}} \boldsymbol{v}_{h,\ell}\cdot\boldsymbol{n}_\ell \, d\Gamma = -\int_{\Gamma_{k\ell}} \boldsymbol{v}\cdot\boldsymbol{n}_k \, d\Gamma = \int_{\Gamma_{k\ell}} \boldsymbol{v}\cdot\boldsymbol{n}_\ell \, d\Gamma.$$

This allows us to deduce that $\forall k$ $(1 \leq k \leq k^*)$

$$\int_{\partial\Omega_k} (\boldsymbol{v}-\boldsymbol{v}_{h,k})\cdot\boldsymbol{n}_k \, d\Gamma = 0,$$

and then statement (4.5) is checked by Green's formula. In order to bound $\|\boldsymbol{v}_h\|_{\boldsymbol{X}}$ we make the convention $\boldsymbol{v}_h^{k\ell} = \boldsymbol{v}_h^{\ell k}$; then a scaling argument yields $\forall k$ $(1 \leq k \leq k^*)$

$$\int_{\partial\Omega_k} |\boldsymbol{v}\cdot\boldsymbol{n}_k| \, d\Gamma \leq cH^{\frac{3}{2}} \|\boldsymbol{v}_k\|_{H^1(\Omega_k)^3},$$

which, together with Lemma 4.1, allows the following:

$$\|\boldsymbol{v}_{h,k}\|^2_{H^1(\Omega_k)^3} \leq c_k \sum_{\ell;\ k\ell \text{ is defined}} \|\boldsymbol{v}_{h,k}^{k\ell}\|^2_{H^1(\Omega_k)^3} \left| \int_{\Gamma_{k\ell}} \boldsymbol{v}\cdot\boldsymbol{n}_k \, d\Gamma \right|^2 \leq \gamma_k \|\boldsymbol{v}_k\|^2_{H^1(\Omega_k)^3}.$$

The constant $\gamma_k (= c \max_{k\ell} \gamma_{k\ell})$ is independent of h and of H. This yields the statement (4.6) and achieves the proof with $\gamma = \max_k \gamma_k$. ∎

Remark 4.3. Actually, the constant $\check{\alpha}$ is also independent of k^*, the total number of the subdomains, but it does depend on the shape of $(\Omega_k)_k$ and their distribution (*i.e., nearest neighbor relations*).

Back to the global pressure space $Q_h(\Omega)$, we intend to prove by the Boland–Nicolaides method an optimal *inf-sup* condition between $\boldsymbol{X}_h(\Omega)$ and $Q_h(\Omega)$.

THEOREM 4.4. *There exists a constant α' depending neither on the discretization parameter h nor on the size H of the subdomains so that the following* inf-sup *condition holds*:

$$\inf_{q_h\in Q_h(\Omega)}\ \sup_{\boldsymbol{v}_h\in\boldsymbol{X}_h(\Omega)}\frac{b(\boldsymbol{v}_h,q_h)}{\|\boldsymbol{v}_h\|_{\boldsymbol{X}}\|q_h\|_{L^2(\Omega)}}\geq\alpha'.$$

Proof. Let $q_h=(q_{h,k})_{1\leq k\leq k^*}$ be any function in $Q_h(\Omega)$; it may be decomposed as $\forall k(1\leq k\leq k^*)$,

$$q_{h,k}=\tilde{q}_{h,k}+\check{q}_{h,k},\quad\text{with}\quad\check{q}_{h,k}=\frac{1}{|\Omega_k|}\int_{\Omega_k}q_{h,k}(\boldsymbol{x})\,d\boldsymbol{x}.$$

Since the function $\tilde{q}_{h,k}\in Q_h(\Omega_k)\cap L^2_0(\Omega_k)$, there exists a local velocity field $\tilde{\boldsymbol{v}}_{h,k}\in\tilde{\boldsymbol{X}}_h(\Omega_k)=(\boldsymbol{X}_h(\Omega_k)\cap H^1_0(\Omega_k)^3)$ verifying (3.2). Then, we define the function $\tilde{\boldsymbol{v}}_h=(\tilde{\boldsymbol{v}}_{h,k})_{1\leq k\leq k^*}\in\boldsymbol{X}_h(\Omega)$. Furthermore, by Proposition 4.2 we construct $\check{\boldsymbol{w}}_h\in\boldsymbol{X}_h(\Omega)$ so that

$$b(\check{\boldsymbol{w}}_h,\check{q}_h)=\|\check{q}_h\|^2_{L^2(\Omega)}\qquad\text{and}\qquad\check{\alpha}\|\check{\boldsymbol{w}}_h\|_{\boldsymbol{X}}\leq\|\check{q}_h\|_{L^2(\Omega)}.$$

Next we choose $\check{\boldsymbol{v}}_h=(\check{\boldsymbol{v}}_{h,k})_{1\leq k\leq k^*}$ with $\check{\boldsymbol{v}}_{h,k}=\check{\boldsymbol{w}}_{h,k}+\beta_k\tilde{\boldsymbol{v}}_{h,k}$. The coefficients $(\beta_k)_{1\leq k\leq k^*}$ are computed so that $\{(\operatorname{div}\check{\boldsymbol{v}}_h,\tilde{q}_h)_{L^2(\Omega_k)},(1\leq k\leq k^*)\}$ vanish; thus

$$\beta_k=-\frac{(\operatorname{div}\check{\boldsymbol{w}}_{h,k},\tilde{q}_{h,k})_{L^2(\Omega_k)}}{\|\tilde{q}_{h,k}\|^2_{L^2(\Omega_k)}}.$$

Taking $\boldsymbol{v}_h=\check{\boldsymbol{v}}_h+\tilde{\boldsymbol{v}}_h$, it is readily checked that

$$b(\boldsymbol{v}_h,q_h)=\|\tilde{q}_h\|^2_{L^2(\Omega)}+\|\check{q}_h\|^2_{L^2(\Omega)}=\|q_h\|^2_{L^2(\Omega)},$$

and $\alpha'\|\boldsymbol{v}_h\|_{\boldsymbol{X}}\leq\|q_h\|_{L^2(\Omega)}$, where α' may be chosen so that

$$\alpha'=\frac{1}{2}\min(1,\check{\alpha})\min_{1\leq k\leq k^*}\frac{\tilde{\alpha}_k}{\sqrt{1+\tilde{\alpha}_k^2}}.\qquad\square$$

Remark 4.5. Observe that nowhere in the proof of the *inf-sup* condition were the assumptions ($\mathrm{H_i}$) or ($\mathrm{H_{ii}}$) employed. Theorem 4.4 is then valid for arbitrary meshes.

The immediate consequence is the well posedness of the discrete problem. Indeed, by the saddle point theory we have the following result.

PROPOSITION 4.6. *The discrete problem* (3.6)–(3.7) *has only one solution* $(\boldsymbol{u}_h,p_h)\in\boldsymbol{X}_h(\Omega)\times Q_h(\Omega)$*, and we have*

$$\|\boldsymbol{u}_h\|_{\boldsymbol{X}}+\|p_h\|_{L^2(\Omega)^2}\leq C\|\boldsymbol{f}\|_{L^2(\Omega)^3}.\tag{4.7}$$

5. The final error estimates. In order to perform the numerical analysis and to derive the convergence rate towards the exact solution, we adapt, as is done in [34], the approximation theory of the saddle point problems to our nonconforming discretization. Let us introduce the space

$$\boldsymbol{V}_h(\Omega)=\{\boldsymbol{v}_h\in\boldsymbol{X}_h(\Omega),\ b(\boldsymbol{v}_h,q_h)=0,\ \forall q_h\in Q_h(\Omega)\}.$$

Then we have the abstract error estimate which is considered as an extension of the Berger–Scott–Strang lemma to the Stokes problem (see [12], [34]).

LEMMA 5.1. *We have the following error estimate:*

$$\begin{aligned}&\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{X}}+\|p-p_h\|_{L^2(\Omega)}\\ &\quad\le C\left(\inf_{\boldsymbol{v}_h\in\boldsymbol{V}_h(\Omega)}\|\boldsymbol{u}-\boldsymbol{v}_h\|_{\boldsymbol{X}}+\inf_{q_h\in Q_h(\Omega)}\|p-q_h\|_{L^2(\Omega)}\right.\\ &\qquad\left.+\sup_{\boldsymbol{w}_h\in\boldsymbol{V}_h(\Omega)}\frac{1}{\|\boldsymbol{w}_h\|_{\boldsymbol{X}}}\sum_{k=1}^{k^*}\langle(\partial_{\boldsymbol{n}_k}\boldsymbol{u}+p),\boldsymbol{w}_{h,k}\rangle_{\frac12,\partial\Omega_k}\right).\end{aligned}\tag{5.1}$$

The first two error terms in (5.1) are known as the approximation error; the third term is the consistency error and is a consequence of the discontinuity of the elements of $\boldsymbol{V}_h(\Omega)$ through the interfaces. The main convergence result on u is given in the following theorem.

THEOREM 5.2. *Assume that the exact solution* $(\boldsymbol{u},p)\in H_0^1(\Omega)^3\times L_0^2(\Omega)$ *satisfies the regularities*

$$\boldsymbol{u}_k=\boldsymbol{u}_{|\Omega_k}\in H^2(\Omega_k)^3,\quad p_k=p_{|\Omega_k}\in H^1(\Omega_k)\quad\forall k\ (1\le k\le k^*);$$

then under the hypotheses ($\mathrm{H_i}$) *and* ($\mathrm{H_{ii}}$) *it holds that*

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{X}}+\|p-p_h\|_{L^2(\Omega)}\le C\left(\sum_{k=1}^{k^*}h_k^2(|\boldsymbol{u}_k|^2_{H^2(\Omega_k)^3}+|p_k|^2_{H^1(\Omega_k)})\right)^{\frac12}.\tag{5.2}$$

The *inf-sup* condition of Theorem 4.4 is independent of h; techniques used for standard h-finite element approximation work as well. Indeed, the best approximation rate of the velocity $\boldsymbol{u}$ by vector valued functions of $\boldsymbol{V}_h(\Omega)$ is asymptotically equivalent to that provided by the best fit of $\boldsymbol{u}$ by the functions of $\boldsymbol{X}_h(\Omega)$. Under the hypotheses ($\mathrm{H_i}$) and ($\mathrm{H_{ii}}$) this rate is optimal, as given in Proposition 3.2. Moreover, the approximation error is optimal for the pressure (3.5). The only remaining point is to evaluate the consistency error which is handled in a standard way using estimate (3.3) and turns out to be optimal.

LEMMA 5.3. *Under the regularity assumptions of Theorem* 5.2 *it holds that*

$$\begin{aligned}&\sup_{\boldsymbol{w}_h\in\boldsymbol{V}_h(\Omega)}\frac{1}{\|\boldsymbol{w}_h\|_{\boldsymbol{X}}}\sum_{k=1}^{k^*}\langle(\partial_{\boldsymbol{n}_k}\boldsymbol{u}+p\boldsymbol{n}_k),\boldsymbol{w}_{h,k}\rangle_{\frac12,\partial\Omega_k}\\ &\quad\le C\left(\sum_{k=1}^{k^*}h_k^2(|\boldsymbol{u}_k|^2_{H^2(\Omega_k)^3}+|p_k|^2_{H^1(\Omega_k)})\right)^{\frac12}.\end{aligned}$$

Proof. It is clear that $\forall\boldsymbol{v}_h\in\boldsymbol{V}_h(\Omega)$

$$\sum_{k=0}^{k^*}\langle(\partial_{\boldsymbol{n}_k}\boldsymbol{u}+p\boldsymbol{n}_k),\boldsymbol{w}_{h,k}\rangle_{\frac12,\partial\Omega_k}=\sum_{k<\ell}\int_{\Gamma_{k\ell}}(\partial_{\boldsymbol{n}_k}\boldsymbol{u}+p\boldsymbol{n}_k)(\boldsymbol{v}_{h,k}-\boldsymbol{v}_{h,\ell})\,d\Gamma.$$

On account of the definition of $\boldsymbol{X}_h(\Omega)$, which contains $\boldsymbol{V}_h(\Omega)$, we say that $\forall\psi_h\in\boldsymbol{M}_h(\Gamma_{k\ell})$,

$$\int_{\Gamma_{k\ell}}(\partial_{\boldsymbol{n}_k}\boldsymbol{u}+p\boldsymbol{n}_k)(\boldsymbol{v}_{h,k}-\boldsymbol{v}_{h,\ell})\,d\Gamma=\int_{\Gamma_{k\ell}}(\partial_{\boldsymbol{n}_k}\boldsymbol{u}+p\boldsymbol{n}_k-\psi_h)(\boldsymbol{v}_{h,k}-\boldsymbol{v}_{h,\ell})\,d\Gamma.$$

The proof is achieved in a standard way after noticing that Aubin–Nitsche duality yields

$$\inf_{\psi_h \in \boldsymbol{M}_h(\Gamma_{k\ell})} \|\partial_{\boldsymbol{n}_k} \boldsymbol{u} + p\boldsymbol{n}_k - \psi_h\|_{(H^{\frac{1}{2}}(\Gamma_{k\ell})^2)'} \leq C h_\ell (|\boldsymbol{u}_k|_{H^2(\Omega_\ell)^3} + |p_\ell|_{H^1(\Omega_\ell)}). \qquad \square$$

Proof of Theorem 5.2. It is a direct consequence of Lemma 5.1, Proposition 3.2, and Lemma 5.3.

Remark 5.4. It may occur that the solution $(\boldsymbol{u}, p)$ of the problem (2.1)–(2.3) is less regular than assumed in Theorem 5.2. Also, in this case the optimal error estimate holds. For instance, if $(\boldsymbol{u}, p) \in H_0^1(\Omega)^3 \times L^2(\Omega)$ such that $\boldsymbol{u}_k \in H^{1+\lambda_k}(\Omega_k)^3$ and $p_k \in H^{\lambda_k}(\Omega_k)$ with $0 < \lambda_k \leq 1$, we have

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{X}} + \|p - p_h\|_{L^2(\Omega)} \leq C \left(\sum_{k=1}^{k^*} h_k^{2\lambda_k} (\|\boldsymbol{u}_k\|^2_{H^{1+\lambda_k}(\Omega_k)^3} + \|p_k\|^2_{H^{\lambda_k}(\Omega_k)} + \|\boldsymbol{f}\|^2_{L^2(\Omega_k)^3}) \right)^{\frac{1}{2}}.$$

This convergence rate is proven directly for $\lambda_k > \frac{1}{2}$, following the analysis realized here, and the term $\|\boldsymbol{f}\|_{L^2(\Omega_k)^3}$ may disappear. It is more technical for $\lambda_k \leq \frac{1}{2}$ and is obtained as in [7] by Hilbertian interpolation of the stability estimates (2.8) and (4.7) in one side and from estimate (5.2) in the other side.

Remark 5.5. Using the bubble stabilized domain decomposition approach with higher order mixed $\mathcal{P}_2/\mathcal{P}_1$ Taylor–Hood finite elements also yields the optimal convergence rate (5.2) as long as the local solution $(\boldsymbol{u}_k, p_k)$ belongs to $H^{1+\lambda_k}(\Omega_k)^3 \times H^{\lambda_k}(\Omega_k)$ with $\lambda_k \leq \frac{3}{2}$.

6. Conclusion. For the numerical simulation by domain decomposition procedures of second order elliptic problems, the attempts already made to propose an efficient way to match (incompatible) *three-dimensional tetrahedral* finite element grids are based on a more or less heavy construction of Lagrange multiplier spaces (see [10], [48], [39]). Besides, using the current refining/unrefining processes for adaptivity turns out to be a pain in the neck due to the nonsystematic character of these constructions. Combining the (architectural) mortar features together with the bubble-stabilization techniques, as presented in this paper, allow an easy and systematic numerical modeling of the "*weak continuity*" across the interfaces. The mathematical analysis of the stabilized mortar finite element method highlights the expected optimality. Moreover, the implementation of this algorithm does not seem to be a difficult work! Some tips for the numerical programming are given in Appendix B.

Apart from the extension to three dimensions, this paper allows us to recover the optimality on the *inf-sup* constant for the Stokes system. Indeed, that constant turns out to be independent of h and of k^*, the mesh size and the total number of the subdomains. Earlier works in two dimensions (see [7], [9]) failed to prove that it does not grow with k^*.

Appendix A. The proof of Proposition 3.2. For simplicity, the proof is processed on each component of $\boldsymbol{v}$, henceforth denoted v. Let us associate with u the local (to subdomains) Lagrange interpolants $v_{h,k} = (\mathcal{I}_{h,k} v) \in Y_h(\Omega_k)$; then we have (see [31]) that $\forall k$, $\forall \kappa \in \mathscr{T}_h^{\Omega_k}$,

$$|v - v_{h,k}|_{H^1(\kappa)} + |\kappa|^{-\frac{1}{3}} \|v - v_{h,k}\|_{L^2(\kappa)} \leq C |\kappa|^{\frac{1}{3}} |v|_{H^2(\kappa)}; \tag{A.1}$$

note that $|\kappa|^{\frac{1}{3}} \leq ch_k$. The function $v_h = (v_{h,k})_k$ does not satisfy the matching condition across the interfaces. To cope with this, we have to add to v_h a correction term so as to obtain an approximation of v that belongs to $X_h(\Omega)$. Let us focus on a single face $\Gamma_{k\ell}$, $k < \ell$, which is common to Ω_k and Ω_ℓ with Ω_k the mortar side and Ω_ℓ the nonmortar side. Enforcing the matching across $\Gamma_{k\ell}$ requires the following modification of $v_{h,\ell}$ based on the bubble functions:

$$\tilde{v}_{h,\ell} = v_{h,\ell} + \sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} \left(\int_T (v_{h,k} - v_{h,\ell})\, d\Gamma \right) \varphi_T. \tag{A.2}$$

For convenience we also set $\tilde{v}_{h,k} = v_{h,k}$, $k \neq \ell$. Clearly, $\tilde{v}_h = (\tilde{v}_{h,k})_k$ fulfills the matching across $\Gamma_{k\ell}$. In addition, remark that $\tilde{v}_{h,\ell}$ coincides with the initial $v_{h,\ell}$ on the $\partial\Omega_\ell$ except, of course, the interior of $\Gamma_{k\ell}$. Bearing in mind that this process has to be iterated, this modification will be realized on all the faces; the aforementioned remark is very important and says that the correction made on $\Gamma_{k\ell}$ will not alter the matching on the other faces. The remaining work consists of bounding the correction term of (A.2), henceforth denoted $t_{k\ell}$. In view of (3.1) we have

$$\begin{aligned}
|t_{k\ell}|^2_{H^1(\Omega_\ell)} &= \sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} \left(\int_T (v_{h,k} - v_{h,\ell})\, d\Gamma \right)^2 |\varphi_T|^2_{H^1(\Omega_\ell)} \\
&\leq c \sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{3}{2}} \left(\int_T (v_{h,k} - v_{h,\ell})\, d\Gamma \right)^2 \\
&\leq c \sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}} \|v_{h,k} - v_{h,\ell}\|^2_{L^2(T)} \\
&\leq c \sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}} (\|v - v_{h,\ell}\|^2_{L^2(T)} + \|v - v_{h,k}\|^2_{L^2(T)}).
\end{aligned}$$

The first part of the sum is easily handled. Using (A.1) and the regularity of the mesh $\mathscr{T}_h^{\Gamma_{\ell k}}$ yields

$$\begin{aligned}
\sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}} \|v - v_{h,\ell}\|^2_{L^2(T)} &\leq c \sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}} |T|^{\frac{3}{2}} |v|^2_{H^{\frac{3}{2}}(T)} \\
&\leq Ch_\ell^2 |v|^2_{H^{\frac{3}{2}}(\Gamma_{k\ell})} \leq Ch_\ell^2 |v|^2_{H^2(\Omega_\ell)}.
\end{aligned}$$

Bounding the second part is performed using the technical assumptions on the trace meshes. It holds that

$$\begin{aligned}
\sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}} \|v - v_{h,k}\|^2_{L^2(T)} &\leq \sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}} \|v - v_{h,k}\|^2_{L^2(\Delta'_T)} \\
&\leq \sum_{T \in \mathscr{T}_h^{\Gamma_{\ell k}}} \sum_{T' \subset \Delta'_T} |T|^{-\frac{1}{2}} |T'|^{\frac{3}{2}} |v|^2_{H^{\frac{3}{2}}(T')}.
\end{aligned}$$

By assumption (H_{ii})—local comparison between adjacent meshes—we get

$$\sum_{T\in\mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}}\|v-v_{h,k}\|^2_{L^2(T)} \le c \sum_{T\in\mathscr{T}_h^{\Gamma_{\ell k}}} \sum_{T'\subset\Delta'_T} |T'||v|^2_{H^{\frac{3}{2}}(T')}$$
$$\le ch_k^2 \sum_{T\in\mathscr{T}_h^{\Gamma_{\ell k}}} |v|^2_{H^{\frac{3}{2}}(\Delta'_T)}.$$

Using hypothesis (H_i) yields the final result,

$$\sum_{T\in\mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}}\|v-v_{h,k}\|^2_{L^2(T)} \le ch_k^2|v|^2_{H^{\frac{3}{2}}(\Gamma_{k\ell})} \le Ch_k^2|v|^2_{H^2(\Omega_k)}.$$

Iterating the process on all the remaining faces completes the proof.

Remark A.1. The end of the proof becomes direct when the trace meshes $(\mathscr{T}_h^{\Gamma_{k\ell}})_{k\ell}$ are quasi-uniform and any couple of adjacent meshes are globally comparable (see Remark 3.2). Indeed, we may write that

$$\sum_{T\in\mathscr{T}_h^{\Gamma_{\ell k}}} |T|^{-\frac{1}{2}}\|v-v_{h,k}\|^2_{L^2(T)} \le ch_\ell^{-1} \sum_{T\in\mathscr{T}_h^{\Gamma_{\ell k}}} \|v-v_{h,k}\|^2_{L^2(T)}$$
$$\le ch_\ell^{-1}\|v-v_{h,k}\|^2_{L^2(\Gamma_{k\ell})} \le ch_\ell^{-1}h_k^3|v|^2_{H^{\frac{3}{2}}(\Gamma_{k\ell})}.$$

Appendix B. Some remarks on the implementation. We shall give some hints about the numerical implementation of the stabilized mortar finite element method. The goal we are assigned is to understand the issue of the matching constraints implementation in the algebraic form of the problem. Then we choose, for briefness, to address the Poisson problem consisting of *finding* $u_h \in X_h(\Omega)$ *such that*

$$\sum_{k=1}^{k^*} \int_{\Omega_k} \nabla u_{h,k}\cdot\nabla v_{h,k}\, d\boldsymbol{x} = \int_\Omega f v_h\, d\boldsymbol{x} \quad \forall v_h \in X_h(\Omega). \tag{B.1}$$

The presentation adopted here is pretty similar to that detailed in [27], even though the expository supplied there is based on the primal hybrid framework.

Let $v_h = (v_{h,k})_k$ be an arbitrary function of the approximated space $X_h(\Omega)$; then $v_{h,k}$ is split into

$$v_{h,k} = \tilde{v}_{h,k} + \sum_{k\ell,\ k>\ell} \sum_{T\in\mathscr{T}^{\Gamma_{k\ell}}} \beta_T(v_h)\varphi_T \quad \forall k\ (1\le kk^*),$$

where $\tilde{v}_{h,k} \in Y_h(\Omega_k)$ is piecewise linear and $\beta_T(v_h)$ is a real number for any T. The matching condition allows us to obtain the value of $\beta_T(v_h)$ so as to eliminate it from the variational formulation,

$$\beta_T(v_h) = \int_T (\tilde{v}_{h,\ell} - \tilde{v}_{h,k})\, d\Gamma = \int_T [\tilde{v}_h]_{k\ell}\, d\Gamma, \quad \forall T \in \mathscr{T}_h^{\Gamma_{k\ell}}. \tag{B.2}$$

Plugging it into the Poisson problem (B.1), we obtain a stabilized problem with

$\tilde{u}_h = (\tilde{u}_{h,k}) \in Y_h(\Omega) = \prod_{k=1}^{k^*} Y_h(\Omega_k)$ as the only unknown:

$$\begin{aligned}
&\sum_{k=1}^{k^*} \int_{\Omega_k} \nabla \tilde{u}_{h,k} \cdot \nabla \tilde{v}_{h,k}\, d\boldsymbol{x} + \sum_{k\ell,\, k>\ell} \sum_{T \in \mathscr{T}^{\Gamma_{k\ell}}} |\varphi_T|^2_{H^1(\kappa_T)} \beta_T(u_h)\beta_T(v_h) \\
&\qquad + \sum_{k\ell,\, k>\ell} \sum_{T \in \mathscr{T}^{\Gamma_{k\ell}}} \int_{\kappa_T} (\beta_T(u_h) \nabla \varphi_T \cdot \nabla v_{h,k} + \beta_T(v_h) \nabla u_{h,k} \cdot \nabla \varphi_T)\, d\boldsymbol{x} \\
&\qquad = \int_\Omega f \tilde{v}_h\, d\boldsymbol{x} + \sum_{k\ell, k>\ell} \sum_{T \in \mathscr{T}^{\Gamma_{k\ell}}} \beta_T(v_h) \int_{\kappa_T} f \varphi_T\, d\boldsymbol{x} \quad \forall v_h \in Y_h(\Omega).
\end{aligned}$$

The norm $|\varphi_T|_{H^1(T)}$ can be easily computed by hand. Equation (B.2) and an elementary integration by parts yield the following formulation: *find* $\tilde{u}_h \in Y_h(\Omega)$ *such that*

$$\begin{aligned}
&\sum_{k=1}^{k^*} \int_{\Omega_k} \nabla \tilde{u}_{h,k} \cdot \nabla \tilde{v}_{h,k}\, d\boldsymbol{x} + \sum_{k\ell,\, k>\ell} \sum_{T \in \mathscr{T}^{\Gamma_{k\ell}}} |\varphi_T|^2_{H^1(\kappa_T)} \left(\int_T [\tilde{u}_h]_{k\ell}\, d\Gamma \right) \left(\int_T [\tilde{v}_h]_{k\ell}\, d\Gamma \right) \\
&\qquad - \sum_{k\ell, k>\ell} \sum_{T \in \mathscr{T}^{\Gamma_{k\ell}}} \int_T ([\tilde{u}_h]_{k\ell} \partial_{\boldsymbol{n}} \tilde{v}_{h,k} + \partial_{\boldsymbol{n}} \tilde{u}_{h,k} [\tilde{v}_h]_{k\ell})\, d\Gamma \\
&\qquad = \int_\Omega f \tilde{v}_h\, d\boldsymbol{x} + \sum_{k\ell,\, k>\ell} \sum_{T \in \mathscr{T}^{\Gamma_{k\ell}}} \left(\int_T [\tilde{v}_h]_{k\ell}\, d\Gamma \right) \left(\int_{\kappa_T} f \varphi_T\, d\boldsymbol{x} \right) \quad \forall \tilde{v}_h \in Y_h(\Omega).
\end{aligned}$$

The bilinear form involved in the left-hand side of this variational equation is symmetric. As already indicated in [26], this formulation of the discrete problem (B.1) looks like the Nitsche stabilization introduced in [46] to handle in a variational way the Dirichlet condition. A similar stabilization is also used in [38] in the nonconforming domain decomposition context for the two-dimensional problems.

The most critical point pertains to the evaluation of $\int_T [v_h]_{k\ell}\, d\Gamma$. Indeed, $[v_h]_{k\ell}$ is piecewise linear on the triangle T. An exact computation of this term would be often expensive and sometimes inaccessible, in particular when the trace of the mortar mesh on T results in a complex partition of T. It is then necessary to resort to a numerical integration. The main effect is that the mortar integral matching is not exactly satisfied anymore. So far this difficulty, which has to be solved efficiently in practice, is common to all the domain decomposition methods with nonmatching grids. In [30] and in [45], a high number of numerical experiences realized in reasonable circumstances are discussed, and the authors come to the conclusion that a Gauss integration with a sufficient number of integrating points can be used. However, they recommend use of a different quadrature formula for the evaluation of $\int_T [u_h]_{k\ell}\, d\Gamma$ (the computation should be realized on the mesh $\mathscr{T}_h^{\Gamma_{k\ell}}$) and of $\int_T [v_h]_{k\ell}\, d\Gamma$ (rather $\mathscr{T}_h^{\Gamma_{\ell k}}$ has to be used), which results in a Petrov–Galerkin procedure. Although the symmetry of the system is definitely broken, they claim that this approach does not introduce any numerical instabilities and gives satisfactory results. Our belief is that these trends should be confirmed by numerical tests in three dimensions.

Acknowledgments. The author would like to address his deepest thanks to Y. Maday and C. Bernardi for fruitful discussions and for their help.

REFERENCES

[1] G. A. Abdoulaev, Y. Achdou, Y. Kusnetzov, and C. Prud'homme, *On a parallel implementation of the mortar element method*, M2AN Math. Modél. Numér. Anal., 33 (1999), pp. 245–259.

[2] Y. Achdou, Y. Kusnetzov, and O. Pironneau, *Substructuring preconditioners for the $\mathbb{Q}_1$ mortar element method*, Numer. Math., 71 (1995), pp. 419–449.

[3] Y. Achdou, Y. Maday, and O. B. Widlund, *Iterative substructuring preconditioners for mortar element methods in two dimensions*, SIAM J. Numer. Anal., 36 (1999) pp. 551–580.

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

[5] F. Ben Belgacem, *Discrétisations 3D nonconformes par la méthode de décomposition de domaines des éléments avec joints: Analyse mathématique et mise en œuvre pour le problème de Poisson*, Ph.D. thesis, Université Pierre et Marie Curie, Paris, France, Note technique EDF, ref. HI72/93017, 1993.

[6] F. Ben Belgacem, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[7] F. Ben Belgacem, *The mixed mortar finite element method for the incompressible Stokes problem: Convergence analysis*, SIAM J. Numer. Anal., 37 (2000), pp. 1085–1100.

[8] F. Ben Belgacem, *Sur le raccordement de maillages pour les méthodes des éléments finis pour les problèmes de contact unilatéral*. Dissertation d'Habilitation à Diriger des Recherches, Partie II, Université Paul Sabatier, Toulouse III, France, 2003.

[9] F. Ben Belgacem, C. Bernardi, N. Chorfi, and Y. Maday, *Inf-sup condition for the mortar spectral element discretization of the Stokes problem*, Numer. Math., 85 (2000), pp. 257–281.

[10] F. Ben Belgacem and Y. Maday, *The mortar element method for three dimensional finite elements*, Modél. Math. Anal. Numér., 31 (1997), pp. 289–302.

[11] F. Ben Belgacem and Y. Maday, *Coupling spectral and finite element for second order elliptic three dimensional equations*, SIAM J. Numer. Anal., 36 (1999), pp. 1234–1263.

[12] A. Berger, R. Scott, and G. Strang, *Approximate boundary conditions in the finite element method*, in Symposia Mathematica, Vol. X, Academic Press, London, 1972, pp. 295–313.

[13] C. Bernardi, N. Débit, and Y. Maday, *Coupling spectral and finite element and spectral methods: First results*, Math. Comp., 54 (1990), pp. 21–39.

[14] C. Bernardi and V. Girault, *A local regularization operator for triangular and quadrilateral finite elements*, SIAM J. Numer. Anal., 35 (1998), pp. 1893–1916.

[15] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, H. Brezis and J. L. Lions, eds., Pitman, New York, 1994, pp. 13–51.

[16] S. Bertoluzza, *Wavelet stabilization of the Lagrange multiplier method*, Numer. Math., 86 (2000), pp. 1–28.

[17] S. Bertoluzza and V. Perrier, *The mortar wavelet method*, in Numerical Mathematics and Advanced Applications (Jyväskylä, 1999), P. Neittaanmäki, T. Tiihonen, and P. Tarvanen, eds., World Scientific, River Edge, NJ, 2000, pp. 424–431.

[18] S. Bertoluzza and V. Perrier, *The mortar method in the wavelet context*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 647–673.

[19] S. Bertoluzza, S. Falletta, and V. Perrier, *Wavelet/*FEM *coupling by the mortar method*, in Recent Developments in Domain Decomposition Methods, Lect. Notes Comput. Sci. Eng. 23, L. F. Pavarino and A. Toselli, eds., Springer-Verlag, Berlin, 2002, pp. 119–132.

[20] J. Boland and R. Nicolaides, *Stability of finite elements under divergence constraints*, SIAM J. Numer. Anal., 20 (1983), pp. 722–731.

[21] D. Braess, W. Dahmen, and C. Wieners, *A multigrid algorithm for the mortar finite element method*, SIAM J. Numer. Anal., 37 (1999), pp. 48–69.

[22] D. Braess and W. Dahmen, *Stability estimates of the mortar finite element method for* 3*-dimensional problems*, East-West J. Numer. Math., 6 (1998), pp. 249–263.

[23] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, RAIRO Anal. Numér., 8-R2 (1974), pp. 129–151.

[24] F. Brezzi, L. P. Franca, D. Marini, and A. Russo, *Stabilization techniques for domain decomposition with nonmatching grids*, in Domain Decomposition Methods in Sciences and Engineering, P. Bjrstad, M. Espedal, and D. Keyes, eds., Domain Decomposition Press, Bergen, Norway, 1998, pp. 1–11.

[25] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[26] F. Brezzi and D. Marini, *Error estimates for the three-field formulation with bubble stabilization*, Math. Comp., 70 (2001), pp. 911–934.
[27] F. Brezzi and D. Marini, *Implementation of the stabilized three-field formulation*, in Recent Trends in Numerical Analysis, D. Trigiante, ed., Adv. Theory Comput. Math. 3, Nova Science, Commack, NY, 2000, pp. 59–70.
[28] A. Buffa, *Error estimate for a stabilised domain decomposition method with nonmatching grids*, Numer. Math., 90 (2002), pp. 617–640.
[29] A. Buffa, Y. Maday, and F. Rapetti, *A sliding mesh-mortar method for a two dimensional eddy current model of electric engines*, M2AN Math. Modél. Numér. Anal., 35 (2001), pp. 191–228.
[30] L. Cazabeau, Y. Maday, and C. Lacour, *Numerical quadratures and mortar methods*, in Computational Sciences for the 21st Century, Bristeau et al., eds., Wiley and Sons, New York, 1997, pp. 119–128.
[31] P.-G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[32] Ph. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 9 (1975), pp. 77–84.
[33] N. Débit, *La méthode des éléments avec joints dans le cas du couplage des méthodes spectrales et des éléments finis*, Ph.D. thesis, Université Pierre et Marie Curie, Paris, France, 1992.
[34] N. Débit and Y. Maday, *The coupling of spectral and finite element method for the approximation of the Stokes problem*, Computational Mathematics and Applications, 8th France–USSR–Italy Joint Symposium Proceedings, Pavia, 1989, pp. 129–163.
[35] C. Farhat and F. X. Roux, *A method of finite element tearing and interconnecting and its parallel solution algorithm*, Int. J. Numer. Methods Engrg., 32 (1991), pp. 1205–1227.
[36] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
[37] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[38] B. Henrish and S. Nicaise, *Nitsche Mortar Finite Element Method for Transmission Problems with Singularities*, preprint SFB 393/01-10, Technische Universität Chemnitz, Germany.
[39] C. Kim, R. D. Lazarov, J. E. Pasciak, and P. S. Vassilevski, *Multiplier spaces for the mortar finite element method in three dimensions*, SIAM J. Numer. Anal., 39 (2001), pp. 519–538.
[40] J.-L. Lions and E. Magenes, *Problèmes aux limites non-homogènes*, Vol. 1, Dunod, Paris, 1968.
[41] L. Pavarino and O. Widlund, *Iterative substructuring methods for spectral element discretizations of elliptic systems.* I: *Compressible linear elasticity*, SIAM J. Numer. Anal., 37 (2000), pp. 353–374.
[42] L. Pavarino and O. Widlund, *Iterative substructuring methods for spectral element discretizations of elliptic systems.* II: *Mixed methods for linear elasticity and Stokes flow*, SIAM J. Numer. Anal., 37 (2000), pp. 375–402.
[43] O. Pironneau, *Méthodes d'éléments finis pour les fluides*, Collection Recherche en Mathématiques Appliquées, Masson, Paris, 1988.
[44] C. Prud'Homme, *Décomposition de domaines, application aux équations de Navier–Stokes tridimensionnelles incompressibles.* Thèse de Doctorat de l'Université Pierre et Marie Curie, Paris VI, 2000.
[45] Y. Maday, F. Rapetti, and B. I. Wohlmuth, *The influence of quadrature formula in* $2D$ *and* $3D$ *mortar methods*, in Recent Developments in Domain Decomposition Methods, Lect. Notes Comput. Sci. Eng. 23, L. F. Pavarino and A. Toselli, eds., Springer-Verlag, Berlin, 2002, pp. 203–221.
[46] J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.
[47] B. I. Wohlmuth, *Multigrid methods for saddle point problems arising from mortar finite element discretizations*, Electron. Trans. Numer. Anal., 11 (2000), pp. 43–54.
[48] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 686–713

ON THE EQUIVALENCE OF SOFT WAVELET SHRINKAGE, TOTAL VARIATION DIFFUSION, TOTAL VARIATION REGULARIZATION, AND SIDEs*

GABRIELE STEIDL†, JOACHIM WEICKERT‡, THOMAS BROX‡, PAVEL MRÁZEK‡, AND MARTIN WELK‡

Abstract. Soft wavelet shrinkage, total variation (TV) diffusion, TV regularization, and a dynamical system called SIDEs are four useful techniques for discontinuity preserving denoising of signals and images. In this paper we investigate under which circumstances these methods are equivalent in the one-dimensional case. First, we prove that Haar wavelet shrinkage on a single scale is equivalent to a single step of space-discrete TV diffusion or regularization of two-pixel pairs. In the translationally invariant case we show that applying cycle spinning to Haar wavelet shrinkage on a single scale can be regarded as an absolutely stable explicit discretization of TV diffusion. We prove that space-discrete TV diffusion and TV regularization are identical and that they are also equivalent to the SIDEs system when a specific force function is chosen. Afterwards, we show that wavelet shrinkage on multiple scales can be regarded as a single step diffusion filtering or regularization of the Laplacian pyramid of the signal. We analyze possibilities to avoid Gibbs-like artifacts for multiscale Haar wavelet shrinkage by scaling the thresholds. Finally, we present experiments where hybrid methods are designed that combine the advantages of wavelets and PDE/variational approaches. These methods are based on iterated shift-invariant wavelet shrinkage at multiple scales with scaled thresholds.

Key words. soft wavelet shrinkage, total variation diffusion, total variation regularization, stabilized inverse diffusion equation

AMS subject classifications. 94A08, 65M06, 65T60, 49-99, 94A12

DOI. 10.1137/S0036142903422429

1. Introduction.

Image denoising is a field where one is typically interested in removing noise without sacrificing important structures such as edges. This goal cannot be achieved with linear filters. Consequently, a large variety of nonlinear strategies has been proposed including, among others, wavelet techniques [22, 23, 31], PDEs [2, 37, 41, 47], and variational methods [5, 6, 9, 36].

Although these method classes serve the same purpose, relatively few publications exist where their similarities and differences are juxtaposed and their mutual relations are analyzed. However, such an analysis is highly desirable, since it can help to transfer results from one of these classes to the others. Moreover, a deeper understanding of the differences between these classes might be helpful for designing novel hybrid methods that combine the advantages of the different classes.

The goal of the present paper is to address this problem by analyzing relations between four important representatives of discontinuity-preserving denoising methods:

- wavelet soft thresholding [22],
- space-discrete total variation (TV) diffusion [3, 4],

*Received by the editors February 4, 2003; accepted for publication (in revised form) September 16, 2003; published electronically May 17, 2004. This research was partially funded by DFG projects We 2602/2-1 and We 2602/1-1/So 363/9-1.

http://www.siam.org/journals/sinum/42-2/42242.html

†Faculty of Mathematics and Computer Science, University of Mannheim, 68131 Mannheim, Germany (steidl@math.uni-mannheim.de, http://kiwi.math.uni-mannheim.de).

‡Mathematical Image Analysis Group, Faculty of Mathematics and Computer Science, Saarland University, Building 27, 66041 Saarbrücken, Germany (weickert@mia.uni-saarland.de, brox@mia.uni-saarland.de, mrazek@mia.uni-saarland.de, welk@mia.uni-saarland.de; http://www.mia.uni-saarland.de).

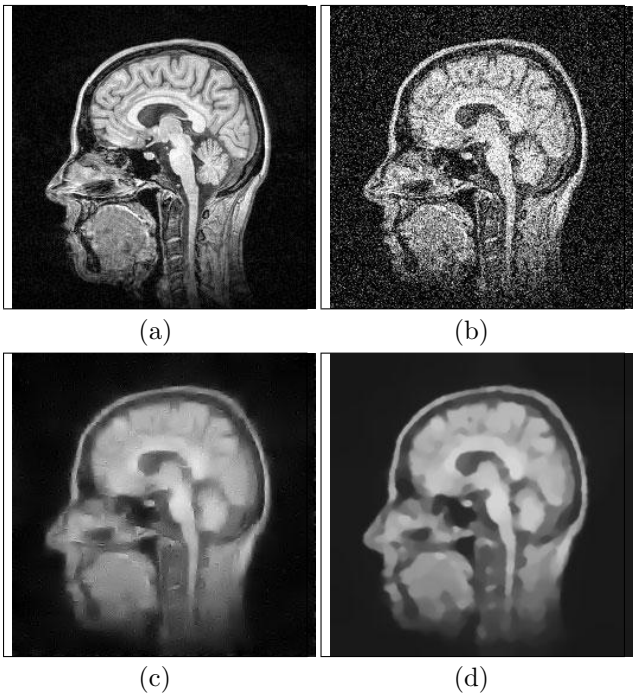


FIG. 1. (a) *Top left: original magnetic resonance image.* (b) *Top right: magnetic resonance image degraded with Gaussian noise with standard deviation* 50. (c) *Bottom left: wavelet denoising of* (b) *using translation-invariant soft shrinkage with Haar wavelets.* (d) *Bottom right: TV diffusion of* (b).

- discrete TV regularization [40, 1],
- SIDEs, a dynamical system that has been inspired from space-discrete stabilized inverse diffusion equations [38].

Figure 1 gives an illustration of the denoising properties of soft wavelet shrinkage and TV regularization methods. The original image is available from [31]. We observe that the results do not differ very much. Indeed, we shall prove in our paper that all four aforementioned methods are very closely related.

In order to keep things as simple as possible we base our analysis on the one-dimensional (1-D) case. Our basic strategy is to start with the simplest cases for which we can establish equivalence. Afterwards, we extend these results to more general situations. The higher-dimensional case is beyond the scope of the present paper, since it cannot be treated as a straightforward generalization of the 1-D ideas. For some preliminary results in two dimensions, we refer the reader to [34], where diffusion-inspired wavelet shrinkage with improved rotation invariance is introduced.

Our paper is organized as follows. In section 2 we give a very brief description of the general ideas behind wavelet shrinkage, nonlinear diffusion filtering, variational image denoising, and SIDEs. In section 3 we specify these paradigms to the simplest cases where equivalence can be shown. In this section we restrict ourselves to two-pixel signals, soft Haar wavelet shrinkage, TV diffusivity, and its corresponding regularizer. Under these circumstances we prove equivalence between wavelet shrinkage, TV diffusion, and TV regularization. These results are extended in section 4 to the translationally invariant case with N-pixel signals. In the wavelet setting, we use

a Haar wavelet-based technique on a single scale with cycle spinning. We show that it can be regarded as a single iteration of a stabilized explicit scheme for TV diffusion, and we prove that this TV diffusion is equivalent to both TV regularization and SIDEs with an appropriate force function. In section 5 we extend our wavelet results from a single scale to multiple scales. We show that multiple scale Haar wavelet soft shrinkage can be regarded as TV diffusion, TV regularization, or SIDEs applied to a Laplacian pyramid decomposition of the signal. Moreover, we propose and analyze a strategy for avoiding Gibbs-like artifacts by scaling the shrinkage thresholds. In section 6 we present experiments where we compare iterated single-scale filtering with noniterated and iterated multiscale filtering. The paper is concluded with a summary in section 7.

Related work. Recently, a number of interesting connections between wavelet shrinkage of functions, regularization methods, and PDEs has been established. A book by Meyer [33] presents a unified view on wavelets and nonlinear evolutions, and Shen and Strang [43] have included wavelets into the solution of the linear heat equation. Chambolle et al. [13] showed that one may interpret wavelet shrinkage of functions as regularization processes in suitable Besov spaces. In particular, Haar thresholding was considered in [18]. Furthermore, Cohen et al. [17] showed that the space of functions of bounded variation can be "almost" characterized by wavelet expansions. Chambolle and Lucier [15] considered iterated translationally invariant wavelet shrinkage and interpreted it as a nonlinear scale-space, which differs from other scale-spaces by the fact that it is not given in terms of PDEs.

There has also been a rapidly increasing interest in designing hybrid methods using both wavelet shrinkage and TV denoising methods. Durand and Froment [24] proposed to address the problem of pseudo-Gibbs artifacts in wavelet denoising by replacing the thresholded wavelet coefficients by coefficients that minimize the total variation. Their method is also close in spirit to approaches by Chan and Zhou [16], who postprocessed images obtained from wavelet shrinkage by a TV-like regularization technique. Coifman and Sowa [20] used functional minimization with wavelet constraints for postprocessing signals that have been degraded by wavelet thresholding or quantization. Candés and Guo [12] also presented related work, in which they combined ridgelets and curvelets with TV minimization strategies. Recently, Malgouyres [30] proposed a hybrid method that uses both wavelet packets and TV approaches. His experiments showed that it may restore textured regions without introducing ringing artifacts.

Regarding the relations between wavelet shrinkage denoising of discrete signals and TV reduction, not much research has been done so far. One notable exception is a recent paper by Coifman and Sowa [21], where they propose TV diminishing flows that act along the direction of Haar wavelets. Bao and Krim [7] addressed the problem of texture loss in diffusion scale-spaces by incorporating ideas from wavelet analysis. An experimental evaluation of the denoising capabilities of three-dimensional wavelet shrinkage and nonlinear diffusion filters is presented in a paper by Frangakis, Stoschek, and Hegerl [27].

This discussion shows that our paper differs from preceding work by the fact that we investigate conditions under which we can prove equivalence between wavelet shrinkage of discrete signals, space-discrete TV diffusion or regularization, and SIDEs. Some preliminary results in this paper have been presented at conferences [44, 10].

2. The basic methods. The goal of this section is to give a brief introduction to the methods that are considered in this paper: soft Haar wavelet shrinkage, TV diffusion, TV denoising, and SIDEs.

2.1. Wavelet shrinkage. During recent years wavelet methods have proved their use in various signal processing tasks. One of them is discontinuity-preserving denoising. The discrete wavelet transform represents a 1-D signal $f(x)$ in terms of shifted versions of a dilated lowpass scaling function $\varphi(x)$, as well as shifted and dilated versions of a bandpass wavelet function $\psi(x)$. In the case of orthogonal wavelets, this gives

$$f(x) = \sum_{i\in\mathbb{Z}} \langle f, \varphi_i^{j_e}\rangle\, \varphi_i^{j_e}(x) + \sum_{j=-\infty}^{j_e} \sum_{i\in\mathbb{Z}} \langle f, \psi_i^j\rangle\, \psi_i^j(x),$$

where $\psi_i^j(x) := 2^{-j/2}\psi(2^{-j}x - i)$ and where $\langle\cdot,\cdot\rangle$ denotes the inner product in $L_2(\mathbb{R})$. The wavelet representation employs scaling components only at one level j_e, and wavelet components at levels $j \le j_e$ add higher resolution details to the signal.

If the measurements f are corrupted by white Gaussian noise, then this noise is contained to a small amount in all wavelet coefficients $\langle f, \psi_i^j\rangle$, while the original signal is in general determined by few significant wavelet coefficients. Therefore wavelet shrinkage attempts to eliminate noise from the wavelet coefficients by the following three-step procedure:

- *Analysis.* Transform the noisy data f to the wavelet coefficients $d_i^j = \langle f, \psi_i^j\rangle$ and scaling function coefficients $c_i^{j_e} = \langle f, \varphi_i^{j_e}\rangle$.
- *Shrinkage.* Apply a shrinkage function S_τ with a threshold parameter τ related to the variance of the Gaussian noise to the wavelet coefficients, i.e., $S_\tau(d_i^j) = S_\tau(\langle f, \psi_i^j\rangle)$.
- *Synthesis.* Reconstruct the denoised version u of f from the shrunken wavelet coefficients

$$u(x) := \sum_{i\in\mathbb{Z}} \langle f, \varphi_i^{j_e}\rangle\, \varphi_i^{j_e}(x) + \sum_{j=-\infty}^{j_e} \sum_{i\in\mathbb{Z}} S_\tau(\langle f, \psi_i^j\rangle)\, \psi_i^j(x).$$

In the literature a number of different shrinkage functions have been considered. In this paper we focus on one of the most popular strategies, namely Donoho's soft shrinkage [22]. It uses the soft thresholding with threshold parameter $\tau > 0$:

$$S_\tau(x) = \begin{cases} x - \tau \operatorname{sgn}(x) & \text{if} \quad |x| > \tau, \\ 0 & \text{if} \quad |x| \le \tau, \end{cases} \tag{2.1}$$

which shrinks all coefficients towards zero. Other shrinkage functions will be considered in a forthcoming paper.

Furthermore, in this paper we restrict our attention to Haar wavelets. They are well suited for recovering piecewise constant signals with discontinuities. The Haar wavelet $\psi(x)$ and the corresponding scaling function $\varphi(x)$ are given by $\psi(x) := 1_{[0,\frac{1}{2})} - 1_{[\frac{1}{2},1)}$ and $\varphi(x) := 1_{[0,1)}$, where $1_{[a,b)}$ is the characteristic function of $[a,b)$:

$$1_{[a,b)}(x) \;:=\; \begin{cases} 1 & \text{if } x \in [a,b), \\ 0 & \text{else.} \end{cases}$$

Using the so-called two-scale relation of the wavelet and its scaling function, the coefficients c_i^j and d_i^j at higher level j can be computed from the coefficients c_i^{j-1} at lower level $j-1$ and vice versa. This results in fast algorithms for the analysis step

and the synthesis step. For the Haar wavelets, we obtain

$$c_i^j = \frac{c_{2i}^{j-1} + c_{2i+1}^{j-1}}{\sqrt{2}}, \qquad d_i^j = \frac{c_{2i}^{j-1} - c_{2i+1}^{j-1}}{\sqrt{2}}, \tag{2.2}$$

$$c_{2i}^{j-1} = \frac{c_i^j + d_i^j}{\sqrt{2}}, \qquad c_{2i+1}^{j-1} = \frac{c_i^j - d_i^j}{\sqrt{2}}. \tag{2.3}$$

2.2. Diffusion filtering. Let us now consider a function $f(x)$ on some interval $[a, b]$. The basic idea behind nonlinear diffusion filtering is to obtain a family $u(x, t)$ of filtered versions of the signal $f(x)$ as the solution of a suitable diffusion process with $f(x)$ as initial condition and homogeneous Neumann boundary conditions [37]:

$$\begin{aligned} u_t = (g(u_x^2)\, u_x)_x \quad & \text{on } (a, b) \times (0, \infty), \\ u(x, 0) = f(x) \quad & \text{for all } x \in [a, b], \\ u_x(a, t) = u_x(b, t) = 0 \quad & \text{for all } t \in (0, \infty), \end{aligned} \tag{2.4}$$

where subscripts denote partial derivatives, and the diffusion time t is a simplification parameter: larger values correspond to stronger filtering.

The diffusivity $g(u_x^2)$ is a nonnegative function that steers the amount of diffusion. Usually, it is decreasing in u_x^2. This ensures that strong edges are less blurred by the diffusion filter than noise and low-contrast details. In the present paper, we focus on the *TV diffusivity*

$$g(u_x^2) \;:=\; \frac{1}{|u_x|}. \tag{2.5}$$

The resulting *TV diffusion filter* (also called *TV flow*) has a number of interesting properties. It requires no additional parameters (besides t), it is well posed [3, 8, 25], it preserves the shape of some objects [8], and it leads to constant signals in finite time [4].

2.3. Regularization methods. Regularization methods constitute an alternative to diffusion filters when one is interested in a discontinuity-preserving denoising method for a continuous signal $f(x)$ with $x \in [a, b]$. Here the basic idea is to look for the minimizer u of the energy functional

$$E(u; \alpha, f) \;:=\; \int_a^b \Big((u - f)^2 \;+\; \alpha\, \Psi(u_x^2) \Big)\, dx. \tag{2.6}$$

The first term of this functional encourages similarity between the original signal $f(x)$ and its filtered version $u(x)$, while the second term penalizes deviations from smoothness. The increasing function Ψ is called the *penalizer (regularizer)*, and the nonnegative *regularization parameter* α serves as smoothness weight: larger values correspond to a more pronounced filtering.

As is explained in detail in [42], there are strong relations between regularization methods and diffusion filters: A minimizer of (2.6) satisfies necessarily the Euler–Lagrange equation

$$\frac{u - f}{\alpha} \;=\; (\Psi'(u_x^2)\, u_x)_x,$$

with homogeneous Neumann boundary conditions. This equation may be regarded as a fully implicit time discretization of the diffusion equation (2.4) with diffusivity

$g(u_x^2) = \Psi'(u_x^2)$, initial value $f(x)$, and stopping time α. Thus, one would expect that the minimizer of (2.6) *approximates* the diffusion filter (2.4) but is not identical to it.

In the present paper, we are interested in one of the most popular nonlinear regularization methods, namely *TV regularization* [40, 1]. It uses the penalizer $\Psi(u_x^2) := 2\,|u_x|$, which corresponds to the TV diffusivity (2.5). This regularization is well known for its good denoising capabilities and its tendency to create blocky, segmentation-like results. Well-posedness results can be found in [14].

2.4. SIDEs. A SIDE is a dynamical system that has been inspired from a stabilized limiting case of a space-discrete nonlinear diffusion filter [38]. The name SIDE is an acronym for *stabilized inverse diffusion equation*.

Let us consider a discrete signal $f = (f_i)_{i=0}^{N-1}$. Then its SIDE evolution produces a sequence of filtered images $u(t) = (u_i(t))_{i=0}^{N-1}$, with $u(0) = f$. Increasing the time t leads to a consecutive merging of regions. The evolution between two merging events is governed by a dynamical system with a discontinuous right-hand side.

Assume that at some time t_j a pixel with index i belongs to a constant region of size m_{i,t_j}; i.e., there exist $l \geq 1$ and $r \geq 0$ with $m_{i,t_j} = l + r$,

$$u_{i-l+1} = \cdots = u_i = u_{i+1} = \cdots = u_{i+r},$$
$$u_{i-l} \neq u_{i-l+1} \text{ if } i - l \geq 0, \qquad u_{i+r} \neq u_{i+r+1} \text{ if } i + r \leq N - 2.$$

Then the SIDEs algorithm proceeds as follows:

(i) *Initialization*. Start at time $t_0 = 0$ with the trivial segmentation, where each pixel i is regarded as a region of size $m_{i,0} = 1$:

$$u_i(0) = f_i.$$

(ii) *Evolution*. Given a segmentation at time t_j, the signal evolves according to

$$\dot{u}_i = \begin{cases} \frac{1}{m_{i,t_j}} F(u_{i+r+1} - u_{i+r}) & \text{if } i - l = -1, \\ \frac{-1}{m_{i,t_j}} F(u_{i-l+1} - u_{i-l}) & \text{if } i + r = N - 1, \\ \frac{1}{m_{i,t_j}} \left(F(u_{i+r+1} - u_{i+r}) - F(u_{i-l+1} - u_{i-l})\right) & \text{else,} \end{cases} \tag{2.7}$$

where $\dot{u}_i$ denotes the derivative of u_i with respect to t and F is a so-called *force function* that satisfies a number of formal requirements [38]. The first case in (2.7) describes the evolution of the region at the left signal boundary, the second case applies for the right boundary region, and the third case specifies the evolution of all inner regions.

In [38], only the third case has been specified. We have supplemented the other two cases here in order to be able to treat the boundary regions in a proper way as well. The evolution is stopped when two neighboring regions attain equal grey values. This determines the new merging time t_{j+1}.

(iii) *Merging*. Merge the neighboring regions with equal grey values.

(iv) *Loop control*. Stop if all regions are merged to one; else go back to step (ii).

We see that the stabilization in SIDEs is achieved by an additional definition that results in merging neighboring regions when they approach each other. This step is crucial for the performance of SIDEs, as it can be used for reducing the state variables of the dynamical system. The analytical solutions in the following sections will provide further theoretical justification for this region-merging step.

In [38] several theoretical results for SIDEs are proved, including a maximum principle, well-posedness properties, and a finite extinction time.

The dynamic system suggests that for the specific case $m_i = 1$ one may regard a 1-D SIDE as a space discretization of the PDE

$$u_t = (F(u_x))_x$$

with homogeneous Neumann boundary conditions. This is a diffusion equation with flux function F. Since we are specifically interested in the TV case, we do not consider the specific choice in [38] but restrict ourselves to the *TV force function*

$$F(v) := \begin{cases} 1 & \text{if} \quad v > 0, \\ -1 & \text{if} \quad v < 0. \end{cases}$$

Then it is evident that if $m_i = 1$ for all i, TV diffusion is approximated.

3. Two-pixel signals. In this section, we analyze relations between soft wavelet shrinkage, TV diffusion, TV regularization, and SIDEs for the simplest signals, namely discrete signals with only two pixels. We will see that the restriction to two pixels allows us to find analytical solutions for these degenerated nonlinear processes.

3.1. Soft Haar wavelet shrinkage of two-pixel signals. Let us now consider a discrete two-pixel signal $f = (f_0, f_1)$ and study its change under soft Haar wavelet shrinkage. The analysis step produces the coefficients $c = (f_0 + f_1)/\sqrt{2}$ and $d = (f_0 - f_1)/\sqrt{2}$ of the scaling function and the wavelet. For simplicity, we have dropped the sub- and superscripts for c and d. This step is followed by the shrinkage operation $S_\tau(d)$ with the soft shrinkage function (2.1). Then the synthesis step $u_0 = (c + S_\tau(d))/\sqrt{2}$, $u_1 = (c - S_\tau(d))/\sqrt{2}$ gives the final result:

$$u_0 = \begin{cases} f_0 + \frac{\tau}{\sqrt{2}} \operatorname{sgn}(f_1 - f_0) & \text{if} \quad \tau < |f_1 - f_0|/\sqrt{2}, \\ (f_0 + f_1)/2 & \text{else}, \end{cases} \tag{3.1}$$

$$u_1 = \begin{cases} f_1 - \frac{\tau}{\sqrt{2}} \operatorname{sgn}(f_1 - f_0) & \text{if} \quad \tau < |f_1 - f_0|/\sqrt{2}, \\ (f_0 + f_1)/2 & \text{else}. \end{cases} \tag{3.2}$$

This shows that by increasing the shrinkage threshold τ the grey values of both pixels approach each other. For $\tau = |f_1 - f_0|/\sqrt{2}$ they merge, and for larger τ they remain merged.

3.2. TV diffusion of two-pixel signals. Next, we are interested in the space-discrete diffusion of two-pixel signals (f_0, f_1). The homogeneous Neumann boundary conditions are discretized by setting flows over the signal boundary to zero. In this case a space-discrete version of the TV diffusion equation

$$u_t = \left(\frac{u_x}{|u_x|} \right)_x$$

can be written as

$$\dot{u}_0 = \frac{u_1 - u_0}{|u_1 - u_0|}, \qquad \dot{u}_1 = -\frac{u_1 - u_0}{|u_1 - u_0|}, \tag{3.3}$$

with initial conditions $u_0(0) = f_0$ and $u_1(0) = f_1$. Here the dot denotes again temporal differentiation, and the pixel size is set to 1. Setting $w(t) := u_1(t) - u_0(t)$ and

$\eta := f_1 - f_0$, and subtracting $\dot{u}_0$ from $\dot{u}_1$ in (3.3), we obtain the following initial value problem:

$$\dot{w} = -2\,\frac{w}{|w|}, \qquad w(0) = \eta. \tag{3.4}$$

The right-hand side of this differential equation is discontinuous for $w = 0$ and thus requires a generalization of the concept of solution. We say that w is a solution of (3.4) if it is an absolutely continuous function which fulfills

$$\dot{w} = -2\,\mathrm{sgn}\,(w), \qquad w(0) = \eta \tag{3.5}$$

almost everywhere, where

(I) $\mathrm{sgn}\,(w) := 1$ if $w > 0$,
$\mathrm{sgn}\,(w) := -1$ if $w < 0$
and may take any value in $[-1, 1]$ if $w = 0$.

Note that this definition is in agreement with the frequently used concept of differential inclusions for differential equations with discontinuous right-hand sides [26], where absolutely continuous solutions of

$$-\frac{1}{2}\,\dot{w} \in \begin{cases} \{1\} & \text{if } w > 0, \\ \{-1\} & \text{if } w < 0, \\ [-1, 1] & \text{if } w = 0 \end{cases}$$

were considered. The solution of (3.5) can be obtained as follows: If $\eta \neq 0$, then we have by straightforward computation for $t < |\eta|/2$ that $w(t) = \eta - 2\,t\,\mathrm{sgn}\,(\eta)$ and in particular, by continuity of w, that $w(|\eta|/2) = 0$. Assume that $w(t) \neq 0$ for some $t > |\eta|/2$. Let without loss of generality $w(t) > 0$. The opposite assumption $w(t) < 0$ can be handled in the same way. Then $w(t) = -2t + C$, where we get by continuity of w, if t approaches $|\eta|/2$, that $C = |\eta|$ and, consequently, $w(t) = 2(|\eta|/2 - t) < 0$. This contradicts our assumption. Thus $w(t) = 0$ for $t \geq |\eta|/2$. In summary, we obtain the solution

$$w(t) = \begin{cases} \eta \;-\; 2\,t\,\mathrm{sgn}\,(\eta) & \text{if } t < |\eta|/2, \\ 0 & \text{if } t \geq |\eta|/2. \end{cases} \tag{3.6}$$

This equation shows that the grey value difference $w(t) = u_1(t) - u_0(t)$ tends linearly to 0. Both pixels merge at time $t = |f_1 - f_0|/2$ and remain merged afterwards. Thus, already the simple two-pixel model indicates a finite extinction time for TV diffusion. Since $\dot{u}_0 + \dot{u}_1 = 0$ and $u_0(0) + u_1(0) = f_0 + f_1$, we see further that the average grey value is preserved:

$$u_0(t) + u_1(t) = f_0 + f_1 \qquad \forall\, t \geq 0. \tag{3.7}$$

Using (3.6) and (3.7), we obtain the analytical solution

$$u_0(t) = \begin{cases} f_0 \;+\; t\,\mathrm{sgn}\,(f_1 - f_0) & \text{if} \quad t < |f_1 - f_0|/2, \\ (f_0 + f_1)/2 & \text{else}, \end{cases} \tag{3.8}$$

$$u_1(t) = \begin{cases} f_1 \;-\; t\,\mathrm{sgn}\,(f_1 - f_0) & \text{if} \quad t < |f_1 - f_0|/2, \\ (f_0 + f_1)/2 & \text{else}. \end{cases} \tag{3.9}$$

Interestingly, this result is identical to the results (3.1)–(3.2) for soft Haar wavelet shrinkage if one identifies the diffusion time t with the threshold parameter $\tau = \sqrt{2}t$.

3.3. TV regularization of two-pixel signals. Let us now turn our attention to the regularization framework. Again we are interested only in the two-pixel model (f_0, f_1). We consider a space-discrete variant of (2.6) with a TV penalizer:

$$E(u_0, u_1; \alpha, f) \;=\; (f_0 - u_0)^2 + (f_1 - u_1)^2 + 2\alpha\,|u_1 - u_0|. \tag{3.10}$$

Straightforward computation results in the following minimizer of (3.10):

$$u_0 = \begin{cases} f_0 \;+\; \alpha\,\mathrm{sgn}\,(f_1 - f_0) & \text{if} \quad \alpha < |f_1 - f_0|/2, \\ (f_0 + f_1)/2 & \text{else,} \end{cases}$$

$$u_1 = \begin{cases} f_1 \;-\; \alpha\,\mathrm{sgn}\,(f_1 - f_0) & \text{if} \quad \alpha < |f_1 - f_0|/2, \\ (f_0 + f_1)/2 & \text{else.} \end{cases}$$

This result coincides with the outcome of a single Haar wavelet shrinkage step with shrinkage parameter $\tau = \sqrt{2}\,\alpha$. Moreover, it is identical to TV diffusion if one replaces the diffusion time t by the regularization parameter α. Thus, all three methods are equivalent by setting $\tau = \sqrt{2}\,t = \sqrt{2}\,\alpha$. It is remarkable that TV diffusion and TV regularization give identical evolutions in the two-pixel case. From the considerations in section 2.3 one would expect only that the processes approximate each other. In section 4.3 we will investigate if this equivalence also holds in the general space-discrete case with N pixels.

3.4. SIDEs for two-pixel signals. If we consider the SIDE evolution of a two-pixel signal (f_0, f_1), we obtain for the case of a TV force function the dynamical system

$$\dot{u}_0 = \frac{u_1 - u_0}{|u_1 - u_0|}, \qquad \dot{u}_1 = -\frac{u_1 - u_0}{|u_1 - u_0|},$$

with initial conditions $u_0(0) = f_0$ and $u_1(0) = f_1$.

This is the same evolution as in the TV diffusion case. Hence, its solution is given by (3.8)–(3.9), and there is a finite merging time $t = |f_1 - f_0|/2$.

4. *N*-pixel signals. So far we have focused on the two-pixel case. Let us now investigate which of the equivalences carry over to the general 1-D case with N pixels. To this end we will consider shift invariant wavelet shrinkage on a single scale, show that it performs a numerical approximation to TV diffusion, prove the equivalence of space-discrete TV diffusion and discrete TV regularization by deriving analytical solutions of both processes, and show that this solution coincides with SIDEs with TV force functions.

4.1. Shift invariant wavelet shrinkage on a single scale. Let us first reconsider the soft Haar wavelet shrinkage on a single scale with N pixels, where N is even. Figure 2 illustrates this computation as the two-channel filter bank. As usual we apply the z-transform notation $f(z) = \sum_{i=0}^{N-1} f_i z^{-i}$. Then $\boxed{H_i(z)}$ $(i = 0, 1)$ denotes the convolution of f with the lowpass filter $(i = 0)$ and the highpass filter $(i = 1)$, i.e., $f(z)H_i(z)$, $\boxed{2\downarrow}$ and $\boxed{2\uparrow}$ downsampling and upsampling by 2, respectively, and the circle soft thresholding by S_τ. Finally, $\bullet$ signifies addition.

The use of Haar wavelets creates a natural decomposition of the signal into two-pixel pairs of type (f_{2j}, f_{2j+1}) $(j = 0, \ldots, N/2-1)$. This two-pixel clustering, however, also causes a lack of translation invariance which may be responsible for visual artifacts. One method to improve the quality of the denoised signal considerably is to

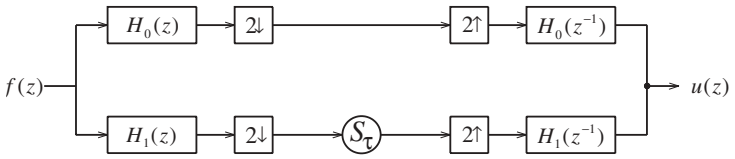


FIG. 2. *Two-channel filter bank with* $H_0(z) = \frac{1+z}{\sqrt{2}}$ *and* $H_1(z) = \frac{1-z}{\sqrt{2}}$.

"average out" the translation dependence. This method was termed *cycle spinning* by Coifman and Donoho [19]. For a single wavelet decomposition step, the basic idea of cycle spinning on a single scale reads as follows:

(a) perform wavelet shrinkage (3.1), (3.2) on successive pairs of the original signal;
(b) shift the signal one pixel to the right;
perform wavelet shrinkage on successive pairs of the shifted signal;
shift the resulting signal one pixel back to the left;
(c) average both results.

The shifting process requires the incorporation of boundary conditions for f. Again we mirror the signal f at its ends. Steps (a)–(c) are equivalent to denoising the signal using a nonsubsampled filter bank. More sophisticated material on oversampled filter banks, corresponding wavelet frames, and undecimated wavelet transforms can be found in [31].

4.2. Equivalence to a numerical scheme for TV diffusion. We have seen that, in order to improve the performance of wavelet shrinkage and to make wavelet-based denoising translationally invariant, cycle spinning can be used. Since there is an equivalence between Haar wavelet shrinkage and TV diffusion in the two-pixel case, it would be natural to ask if there is a TV diffusion scheme equivalent to translationally invariant soft Haar wavelet shrinkage on a single level. This leads us to an interesting novel scheme for TV diffusion.

Derivation of the scheme. We have been able to derive an analytical solution for TV diffusion in the two-pixel case. We can use this two-pixel solution to create a numerical scheme for N pixels. In order to derive such a scheme for some time step size $\triangle t$, we proceed in three steps that are inspired by the cycle spinning technique:

(a) perform TV diffusion with step size $2\triangle t$ on all pixel pairs (u_{2j}, u_{2j+1});
(b) perform TV diffusion with step size $2\triangle t$ on all pixel pairs (u_{2j-1}, u_{2j});
(c) average both results.

Obviously, one step of this iterative scheme is equivalent to a translationally invariant Haar wavelet shrinkage with threshold $\tau = 2\sqrt{2}\triangle t$ on a single level. So let us investigate this scheme in more detail.

At iteration level k, we assume that our signal is given by $(u_i^k)_{i=0}^{N-1}$. We denote the resulting signal of step (a) by $(v_i^{k+1})_{i=0}^{N-1}$ and the spatial grid size by h. From our analysis of the two-pixel situation, it follows that v_i in some even pixel $i = 2j$ is given by

$$v_i^{k+1} = \frac{u_i^k + u_{i+1}^k}{2} - \begin{cases} \max\left(\frac{u_{i+1}^k - u_i^k}{2} - \frac{2\triangle t}{h}, 0\right) & \text{if } u_{i+1}^k \geq u_i^k, \\ \min\left(\frac{u_{i+1}^k - u_i^k}{2} + \frac{2\triangle t}{h}, 0\right) & \text{if } u_{i+1}^k < u_i^k. \end{cases} \tag{4.1}$$

To simplify the notation, we assume only in this subsection instead of the third agreement in (I) that $\operatorname{sgn}(0) := 0$. It is not difficult to see that (4.1) can be rewritten

as

$$v_i^{k+1} \;=\; u_i^k + \frac{2\triangle t}{h}\,\mathrm{sgn}\,(u_{i+1}^k - u_i^k)\,\min\left(1,\,\frac{h}{4\triangle t}|u_{i+1}^k - u_i^k|\right). \tag{4.2}$$

Step (b) leads to a resulting signal $(w_i^{k+1})_{i=0}^{N-1}$. For $i = 2j$ it is given by

$$w_i^{k+1} \;=\; u_i^k - \frac{2\triangle t}{h}\,\mathrm{sgn}\,(u_i^k - u_{i-1}^k)\,\min\left(1,\,\frac{h}{4\triangle t}|u_i^k - u_{i-1}^k|\right). \tag{4.3}$$

Thus, the averaging step (c) gives the final scheme for TV diffusion:

$$\begin{aligned} u_i^{k+1} \;=\; u_i^k &+ \frac{\triangle t}{h}\,\mathrm{sgn}\,(u_{i+1}^k - u_i^k)\,\min\left(1,\,\frac{h}{4\triangle t}|u_{i+1}^k - u_i^k|\right) \\ &- \frac{\triangle t}{h}\,\mathrm{sgn}\,(u_i^k - u_{i-1}^k)\,\min\left(1,\,\frac{h}{4\triangle t}|u_i^k - u_{i-1}^k|\right). \end{aligned} \tag{4.4}$$

The same scheme can also be derived if i is odd, since the construction (a)–(c) in this subsection ensures that the result is translationally invariant. Hence it holds for every inner pixel $i \in \{1, \ldots, N-2\}$. It is even valid for the boundary pixels $i = 0$ and $i = N-1$ if we realize the homogeneous Neumann boundary conditions by introducing dummy values $u_{-1}^k := u_0^k$ and $u_N^k := u_{N-1}^k$.

Stability. Let us now investigate the stability properties of the explicit finite difference scheme (4.2). Since (4.2) satisfies

$$\min(u_i^k, u_{i+1}^k) \;\le\; v_i^{k+1} \;\le\; \max(u_i^k, u_{i+1}^k)$$

and (4.3) fulfills the estimate

$$\min(u_{i-1}^k, u_i^k) \;\le\; w_i^{k+1} \;\le\; \max(u_{i-1}^k, u_i^k),$$

we can conclude that

$$\min(u_{i-1}^k, u_i^k, u_{i+1}^k) \;\le\; u_i^{k+1} \;\le\; \max(u_{i-1}^k, u_i^k, u_{i+1}^k).$$

With the initial condition $u_j^0 = f_j$ for $j = 0, \ldots, N-1$, it follows that the two-pixel scheme (4.2) satisfies the discrete maximum-minimum principle

$$\min_j f_j \;\le\; u_i^{k+1} \;\le\; \max_j f_j$$

for all pixels $i \in \{0, \ldots, N-1\}$, all iteration levels $k = 0, 1, 2, \ldots$, and all time step sizes $\triangle t > 0$. In particular, this shows that the scheme is absolutely stable in the maximum norm.

We may regard (4.4) as a stabilization of the naive explicit scheme

$$u_i^{k+1} \;=\; u_i^k + \frac{\triangle t}{h}\,\mathrm{sgn}\,(u_{i+1}^k - u_i^k) - \frac{\triangle t}{h}\,\mathrm{sgn}\,(u_i^k - u_{i-1}^k), \tag{4.5}$$

which becomes unstable for arbitrary small time steps if neighboring values become arbitrarily close.

Consistency. The absolute stability in scheme (4.4) is at the expense that its consistency is no longer unconditional. This effect is typical for absolutely stable explicit schemes; see, for example, the DuFort–Frankel scheme for linear diffusion. In

our case, (4.4) is an $O(\triangle t + h^2)$ approximation to the continuous TV diffusion for $\triangle t \leq \frac{h}{4} \min\left(|u_{i+1}^k - u_i^k|, |u_i^k - u_{i-1}^k|\right)$, since it coincides with scheme (4.5) then. For larger time step sizes, the scheme performs averaging within the neighborhood of each pixel. By using small time step sizes, these averaging effects appear only in regions that are already almost flat such that the difference from real TV diffusion becomes invisible. This two-pixel scheme may be regarded as an alternative to classical finite difference schemes that are based on the regularized TV flow

$$u_t = \left(\frac{u_x}{\sqrt{\epsilon^2 + u_x^2}} \right)_x . \tag{4.6}$$

The ϵ-regularization is necessary for making the diffusivity bounded. It has an effect similar to the deviation from consistency in the two-pixel scheme (4.4): For small $|u_x|$, a PDE is approximated that differs from TV diffusion and has better stability properties. Indeed, in section 6 we shall see that both schemes give very similar results.

Related schemes. The idea to split up a diffusion process into pairwise interactions has also proved to be fruitful in other fields. In the context of fluid dynamic problems, related schemes have been formulated by Richardson, Ferrell, and Long [39]. These authors, however, use multiplicative splittings; i.e., they first compute the diffusion of the pairs of type (u_{2j}, u_{2j+1}), which is then used as the initial state for the subsequent diffusion of the shifted pairs. In a general nonlinear setting, such a scheme is not translationally invariant. Our approach computes the diffusion of the pairs and the shifted pairs in parallel and averages afterwards. This additive splitting guarantees translation invariance. The splitting into two-pixel interactions distinguishes scheme (4.4) from other additive operator splittings [29, 48]. They use directional splittings along the coordinate axis.

4.3. Equivalence of space-discrete TV diffusion and discrete TV regularization. The equivalence of TV diffusion and TV regularization in the two-pixel case gives rise to the question of whether this equivalence also holds in the N-pixel situation. In order to prove this, we now show that both processes have the same analytical solutions.

4.3.1. Space-discrete TV diffusion. We consider the following dynamical system designed to describe space-discrete TV flow on a 1-D signal with N pixels:

$$\begin{aligned} \dot{u}_0 &= \operatorname{sgn}(u_1 - u_0), \\ \dot{u}_i &= \operatorname{sgn}(u_{i+1} - u_i) - \operatorname{sgn}(u_i - u_{i-1}) \qquad (i = 1, \ldots, N-2), \\ \dot{u}_{N-1} &= -\operatorname{sgn}(u_{N-1} - u_{N-2}), \\ u(0) &= f. \end{aligned} \tag{4.7}$$

In the following, we further set $u_{-1} := u_0$ and $u_N := u_{N-1}$. Since the right-hand side of this system is discontinuous, we need again a more detailed specification of when a system of functions is said to satisfy these differential equations. A vector-valued function u is said to fulfill the system (4.7) over the time interval $[0, T]$ if the following holds true:

(II) u is an absolutely continuous vector-valued function which satisfies (4.7) almost everywhere, where sgn is defined by (I) in subsection 3.2.

(III) If $\dot{u}_i(t)$ and $\dot{u}_{i+1}(t)$ exist for the same t, and $u_{i+1}(t) = u_i(t)$ holds, then the expression $\operatorname{sgn}(u_{i+1}(t) - u_i(t))$ occurring in both the right-hand sides for $\dot{u}_i(t)$ and $\dot{u}_{i+1}(t)$ must take the same value in both equations.

With this notation we can establish the following results.

PROPOSITION 4.1 (properties of space-discrete TV diffusion). *The system* (4.7) *has a unique solution* $u(t)$ *in the sense of* (II) *and* (III). *This solution has the following properties:*

(i) Finite extinction time. *There exists a finite time* $T \geq 0$ *such that for all* $t \geq T$ *the signal becomes constant:*

$$u_i(t) = \frac{1}{N} \sum_{k=0}^{N-1} f_k \qquad \text{for all } \; i = 0, \ldots, N-1. \tag{4.8}$$

(ii) Finite number of merging events. *There exists a finite sequence* $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = T$ *such that the interval* $[0, T)$ *splits into subintervals* $[t_j, t_{j+1})$ *with the property that for all* $i = 0, \ldots, N-2$ *either* $u_i(t) = u_{i+1}(t)$ *or* $u_i(t) \neq u_{i+1}(t)$ *throughout* $[t_j, t_{j+1})$. *The absolute difference between neighboring pixels does not become larger for increasing* $t \in [t_j, t_{j+1})$.

(iii) Analytical solution. *In each of the subintervals* $[t_j, t_{j+1})$ *constant regions of* $u(t)$ *evolve linearly:*

For a fixed index i *let us consider a constant region given by*

$$\begin{gathered} u_{i-l+1} = \cdots = u_i = u_{i+1} = \cdots = u_{i+r} \qquad (l \geq 1, r \geq 0), \\ u_{i-l} \neq u_{i-l+1} \;\text{ if } i - l \geq 0, \qquad u_{i+r} \neq u_{i+r+1} \;\text{ if } i + r \leq N-1 \end{gathered} \tag{4.9}$$

for all $t \in [t_j, t_{j+1})$. *We call* (4.9) *a region of size* $m_{i,t_j} = l + r$. *For* $t \in [t_j, t_{j+1})$ *let* $\triangle t = t - t_j$. *Then* $u_i(t)$ *is given by*

$$u_i(t) = u_i(t_j) + \mu_{i,t_j} \frac{2 \triangle t}{m_{i,t_j}},$$

where μ_{i,t_j} *reflects the relation between the region containing* u_i *and its neighboring regions. It is given as follows:*

For inner regions (i.e., $i - l \geq 0$ *and* $i + r \leq N - 1$*) we have*

$$\mu_{i,t_j} = \begin{cases} 0 & \text{if } \; (u_{i-l}, u_i, u_{i+r+1}) \text{ is strictly monotonic}, \\ 1 & \text{if } \; u_i \text{ is minimal in } (u_{i-l}, u_i, u_{i+r+1}), \\ -1 & \text{if } \; u_i \text{ is maximal in } (u_{i-l}, u_i, u_{i+r+1}), \end{cases} \tag{4.10}$$

and in the boundary case ($i - l + 1 = 0$ *or* $i + r = N - 1$*) the evolution is half as fast:*

$$\mu_{i,t_j} = \begin{cases} 0 & \text{if } \; m = N, \\ \frac{1}{2} & \text{if } \; u_i \text{ is minimal in } (u_{i-l}, u_i, u_{i+r+1}), \\ -\frac{1}{2} & \text{if } \; u_i \text{ is maximal in } (u_{i-l}, u_i, u_{i+r+1}). \end{cases} \tag{4.11}$$

Proof. Let u be a solution of (4.7). We show that u is uniquely determined and satisfies the rules (i)–(iii). Our proof proceeds in four steps.

Step 1. If $\dot{u}(t)$ exists at a fixed time t and $u_i(t)$ lies at this time in some region

$$\begin{gathered} u_{i-l+1}(t) = \cdots = u_i(t) = \cdots = u_{i+r}(t) \qquad (l \geq 1, r \geq 0), \\ u_{i-l}(t) \neq u_{i-l+1}(t) \;\text{ if } i - l \geq 0, \qquad u_{i+r}(t) \neq u_{i+r+1}(t) \;\text{ if } i + r \leq N-1 \end{gathered}$$

of size $m_{i,t}$, then it follows by (4.7) and (III) in the nonboundary case $i - l \geq 0$ and $i + r \leq N - 1$ that

$$u_i(t) \;=\; \frac{1}{m_{i,t}} \sum_{k=-l+1}^{r} u_{i+k}(t),$$

and therefore

$$\dot{u}_i(t) = \frac{1}{m_{i,t}} \sum_{k=-l+1}^{r} \dot{u}_{i+k}(t) = \frac{1}{m_{i,t}} \left(\operatorname{sgn}(u_{i+r+1}(t) - u_i(t)) - \operatorname{sgn}(u_i(t) - u_{i-l}(t))\right)$$

$$= \mu_{i,t} \frac{2}{m_{i,t}}, \tag{4.12}$$

where $\mu_{i,t}$ describes the relation between the region containing u_i and its neighbors at time t as in (4.10). In the boundary case $i - l + 1 = 0$ or $i + r = N - 1$ we follow the same lines and obtain (4.12) with $\mu_{i,t}$ defined by (4.11).

Step 2. Let $\dot{u}(t)$ exist in some small interval (τ_0, τ_1), and assume that $u_i(t) \neq u_{i+1}(t)$ for some $i \in \{0, \dots, N-2\}$ and all $t \in (\tau_0, \tau_1)$. By continuity of u we may assume that $u_i(t) < u_{i+1}(t)$ throughout (τ_0, τ_1). The opposite case $u_i(t) > u_{i+1}(t)$ can be handled in the same way. Then we obtain by (4.12) and the definition of $\mu_{i,t}$ for all $t \in (\tau_0, \tau_1)$ that

$$\dot{u}_i(t) \geq 0 \quad \text{if} \quad i - l \geq 0, \tag{4.13}$$

$$\dot{u}_i(t) > 0 \quad \text{if} \quad i - l + 1 = 0, \tag{4.14}$$

$$\dot{u}_{i+1}(t) \leq 0 \quad \text{if} \quad i + r \leq N - 2, \tag{4.15}$$

$$\dot{u}_{i+1}(t) < 0 \quad \text{if} \quad i + r = N - 1. \tag{4.16}$$

Set $w(t) := u_{i+1}(t) - u_i(t)$. Then the mean value theorem yields

$$w(\tau_1) - w(\tau_0) = (\tau_1 - \tau_0)\, \dot{w}(t^*)$$

for some $t^* \in (\tau_0, \tau_1)$, and we get by (4.13)–(4.16) that

$$w(\tau_1) - w(\tau_0) \leq 0$$

with strict inequality in the boundary case. Consequently, the difference between pixels cannot become larger in the considered interval. In particular, by continuity of u, pixels cannot be split. Once merged they stay merged.

Step 3. Now we start at time $t_0 = 0$. Let t_1 be the largest time such that $\dot{u}(t)$ exists and no merging of regions appears in $(0, t_1)$. Then, for all $i \in \{0, \dots, N-1\}$, a function u_i is in the same region with the same relations to its neighboring regions throughout $[0, t_1)$. Thus, we conclude by (4.12) that

$$\dot{u}_i(t) = \mu_{i,0} \frac{2}{m_{i,0}} \qquad (t \in (0, t_1))$$

and, consequently,

$$u_i(t) = \mu_{i,0} \frac{2t}{m_{i,0}} + C_{i,0} = f_i + \mu_{i,0} \frac{2t}{m_{i,0}} \qquad (t \in [0, t_1]),$$

where the last equality follows by continuity of u_i if t approaches 0.

Step 4. We are now in the position to analyze the entire chain of merging events successively.

Next, we consider the largest interval (t_1, t_2) without merging events in the same way, where we take the initial setting $u(t_1)$ into account instead of f. Then we obtain

$$u_i(t) = \mu_{i,t_1} \frac{2t}{m_{i,t_1}} + C_{i,t_1},$$

where, by continuity of u_i, $u_i(t_1) = \mu_{i,t_1} \frac{2t_1}{m_{i,t_1}} + C_{i,t_1}$ and, consequently,

$$u_i(t) = u_i(t_1) + \mu_{i,t_1} \frac{2(t-t_1)}{m_{i,t_1}}.$$

Now we can continue in the same way by considering $[t_2, t_3)$ and so on. Since we have only a finite number N of pixels and some of these pixels merge at the points t_j, the process stops after a finite number of n steps with output (4.8). Conversely, it is easy to check that a function u with (i)–(iii) is a solution of the system (4.7). This completes the proof of the proposition. □

4.3.2. Discrete TV regularization. Next, we will prove that discrete TV regularization satisfies the same rules as space-discrete TV diffusion. For given initial data $f = (f_0, \dots, f_{N-1})$ discrete TV regularization consists of constructing the minimizer $u(\alpha) = \min_u E(u; \alpha, f)$ of the functional

$$E(u; \alpha, f) = \sum_{i=0}^{N-1} \left((u_i - f_i)^2 + 2\alpha |u_{i+1} - u_i| \right), \tag{4.17}$$

where we suppose again Neumann boundary conditions $u_{-1} = u_0$ and $u_N = u_{N-1}$.

For a fixed regularization parameter $\alpha \geq 0$, the minimizer of (4.17) is uniquely determined since $E(u; \alpha, f)$ is strictly convex in $u_0, \dots, u_{N-1}$. Further, $E(u, \alpha; f)$ is a continuous function in $u_0, \dots, u_{N-1}, \alpha$. Consequently, $u(\alpha)$ is a (componentwise) continuous function in α.

The following proposition implies, together with Proposition 4.1, the equivalence of space-discrete TV diffusion and discrete TV regularization.

PROPOSITION 4.2 (properties of discrete TV regularization). *The minimizing function $u(\alpha)$ of (4.17) is uniquely determined by the following rules:*

(i) Finite extinction parameter. *There exists a finite $A \geq 0$ such that for all $\alpha \geq A$ the signal becomes constant:*

$$u_i(\alpha) = \frac{1}{N} \sum_{k=0}^{N-1} f_k \qquad \text{for all } i = 0, \dots, N-1.$$

(ii) Finite number of merging events. *There exists a finite sequence $0 = a_0 < a_1 < \cdots < a_{n-1} < a_n = A$ such that the interval $[0, A)$ splits into subintervals $[a_j, a_{j+1})$ with the property that for all $i = 0, \dots, N-2$ either $u_i(\alpha) = u_{i+1}(\alpha)$ or $u_i(\alpha) \neq u_{i+1}(\alpha)$ throughout $[a_j, a_{j+1})$. The absolute difference between neighboring pixels does not become larger for increasing $\alpha \in [a_j, a_{j+1})$.*

(iii) Analytical solution. *In each of the subintervals $[a_j, a_{j+1})$ constant regions of $u(\alpha)$ evolve linearly:*

For a fixed index i let us consider a constant region given by

$$u_{i-l+1} = \cdots = u_i = u_{i+1} = \cdots = u_{i+r} \qquad (l \geq 1, r \geq 0), \tag{4.18}$$

$$u_{i-l} \neq u_{i-l+1} \text{ if } i - l \geq 0, \qquad u_{i+r} \neq u_{i+r+1} \text{ if } i + r \leq N-2 \tag{4.19}$$

for all $\alpha \in [a_j, a_{j+1})$. We call (4.18) a region of size $m_{i,a_j} = l + r$. For $\alpha \in [a_j, a_{j+1})$ let $\triangle\alpha = \alpha - a_j$. Then $u_i(\alpha)$ is given by

$$u_i(\alpha) = u_i(a_j) + \mu_{i,a_j} \frac{2\triangle\alpha}{m_{i,a_j}},$$

where μ_{i,a_j} reflects the relation between the region containing u_i and its neighboring regions. It is given as follows:

For inner regions (i.e., $i-l \geq 0$ and $i+r \leq N-2$) we have

$$\mu_{i,a_j} = \begin{cases} 0 & \text{if } (u_{i-l}, u_i, u_{i+r+1}) \text{ is strictly monotonic,} \\ 1 & \text{if } u_i \text{ is minimal in } (u_{i-l}, u_i, u_{i+r+1}), \\ -1 & \text{if } u_i \text{ is maximal in } (u_{i-l}, u_i, u_{i+r+1}), \end{cases} \tag{4.20}$$

and in the boundary case ($i-l+1=0$ or $i+r=N-1$) the evolution is half as fast:

$$\mu_{i,a_j} = \begin{cases} 0 & \text{if } m = N, \\ \frac{1}{2} & \text{if } u_i \text{ is minimal in } (u_{i-l}, u_i, u_{i+r+1}), \\ -\frac{1}{2} & \text{if } u_i \text{ is maximal in } (u_{i-l}, u_i, u_{i+r+1}). \end{cases} \tag{4.21}$$

Proof. A proof that is in complete analogy with our proof for the TV diffusion case is presented in [10, 45]. □

Similar results have also been established in a different way by Strong [46] for the case of continuous TV regularization methods with step functions as initializations. It should be noted that TV regularization by using the taut-string algorithm was also considered by Mammen and van de Geer [32]; see also [28].

4.4. Equivalence to SIDEs with TV force functions. In section 2.4 we have seen that 1-D SIDEs with region size 1 and TV force function are identical to space-discrete TV diffusion. Moreover, in section 4.3 we have derived analytical solutions of space-discrete TV diffusion and discrete TV regularization that show the same merging behavior as SIDEs with TV force functions. Consequently, 1-D SIDEs can be interpreted as an exact solution of space-discrete TV diffusion or regularization in general.

This also confirms that the merging steps in the SIDE evolution are much more than a heuristic stabilization that speeds up the evolution: They are a natural consequence of the degenerated diffusivities that are unbounded in 0. Last but not least, our considerations can be regarded as a theoretical justification of region merging in terms of variational and PDE-based techniques.

5. Multiple scales. So far we have considered only soft wavelet shrinkage on a single scale. In almost all practical applications, however, wavelet shrinkage is performed on multiple scales. In this section, we interpret *multiscale* soft shrinkage with Haar wavelets as the application of nonlinear TV-based diffusion to two-pixel groups of *hierarchical* signals. First, we consider the standard situation without shift invariance; then we discuss the shift-invariant case. Finally, we address a frequent problem that occurs with wavelet shrinkage on multiple scales: the presence of Gibbs-like artifacts. We analyze ways to circumvent this phenomenon by using scale-dependent thresholds.

Throughout this section we deal with signals of length $N = 2^n$ ($n \in \mathbb{N}$).

5.1. Standard case without shift invariance. Haar wavelet shrinkage on two scales is described by the filter bank in Figure 3. To obtain more than two scales we further split up the upper branch of the inner filter bank and so on until we arrive at scale $n = \log_2 N$, where the successive downsampling by 2 results in a one-pixel signal.

Next, we briefly recall the concept of *Gaussian* and *Laplacian pyramids* [11] with respect to the Haar filters. The Gaussian pyramid we are interested in is the sequence

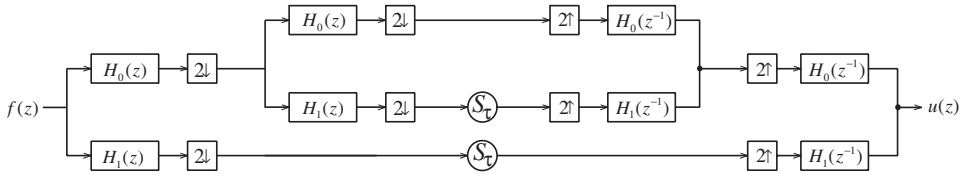


FIG. 3. *Two scales of Haar wavelet shrinkage with* $H_0(z) = \frac{1+z}{\sqrt{2}}$ *and* $H_1(z) = \frac{1-z}{\sqrt{2}}$.

of H_0-smoothed and downsampled versions of an initial signal f given by

$$f = f^{(0)} \longrightarrow f^{(1)} = Rf \longrightarrow \cdots \longrightarrow f^{(n)} = R^n f,$$

where R denotes the operator for H_0-smoothing and subsequent downsampling by 2, i.e.,

$$f_i^{(j+1)} = \big(Rf^{(j)}\big)_i = \big(f_{2i}^{(j)} + f_{2i+1}^{(j)}\big)/\sqrt{2} \qquad (j = 0, \ldots, n-1;\ i = 0, \ldots, N/2^{j+1} - 1).$$

Let $Pf^{(j)}$ denote the prolongated version of $f^{(j)}$ given by

$$\big(Pf^{(j)}\big)_{2i} = \big(Pf^{(j)}\big)_{2i+1} = f_i^{(j)}/\sqrt{2} \qquad (j = 1, \ldots, n;\ i = 0, \ldots, N/2^j - 1). \tag{5.1}$$

Then the corresponding Laplacian pyramid is the sequence

$$f - Pf^{(1)} \longrightarrow f^{(1)} - Pf^{(2)} \longrightarrow \cdots \longrightarrow f^{(n-1)} - Pf^{(n)} \longrightarrow f^{(n)}.$$

By

$$f^{(j)} = Pf^{(j+1)} + \big(f^{(j)} - Pf^{(j+1)}\big) \qquad (j = n-1, \ldots, 0)$$

we can reconstruct f from its Laplacian pyramid.

Let diff_t denote the operator of nonlinear diffusion with TV diffusivity and stopping time t applied to the successive two-pixel parts of a signal. By subsection 3.2 we know that diff_t performs like a single wavelet shrinkage step with soft threshold parameter $\tau = \sqrt{2}t$. In other words, the result of the filter bank in Figure 2 is $u = \mathrm{diff}_t(f)$. Further, we see that the upper branch of this filter bank produces $Pf^{(1)}$ so that the lower branch must produce $\mathrm{diff}_t(f) - Pf^{(1)}$. By (5.1) and (3.1) it is easy to check that the nonlinear operator diff_t fulfills $\mathrm{diff}_t(f) - Pf^{(1)} = \mathrm{diff}_t\big(f - Pf^{(1)}\big)$. Thus, one wavelet shrinkage step is given by $u = Pf^{(1)} + \mathrm{diff}_t\big(f - Pf^{(1)}\big)$. Now the multiscale Haar wavelet shrinkage up to scale n can be described by successive application of diff_t to the Laplacian pyramid:

$$u^{(n)} = f^{(n)}, \tag{5.2}$$

$$u^{(j)} = Pu^{(j+1)} + \mathrm{diff}_t\big(f^{(j)} - Pf^{(j+1)}\big) \qquad (j = n-1, \ldots, 0). \tag{5.3}$$

The result of the multiscale wavelet shrinkage is $u = u^{(0)}$.

5.2. Shift-invariant case. Now we consider translation-invariant multiscale wavelet shrinkage. In the multiscale setting we apply cycle spinning over the range of all N shifts of f. The filter bank which corresponds to two scales of translation-invariant Haar wavelet shrinkage is shown in Figure 4. Note that the inner filter bank uses z^2 instead of z in H_i $(i = 0, 1)$. In general we have to replace z by $z^{2^{j-1}}$ at

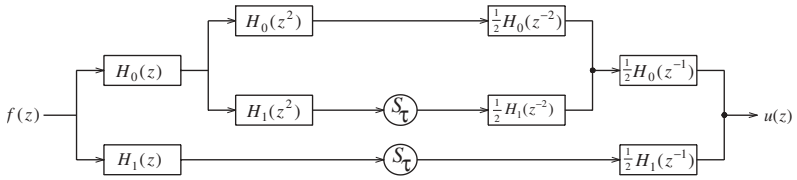


FIG. 4. *Two scales of shift-invariant Haar wavelet shrinkage with* $H_0(z) = \frac{1+z}{\sqrt{2}}$ *and* $H_1(z) = \frac{1-z}{\sqrt{2}}$.

scale j. While ordinary wavelet shrinkage requires $O(N)$ arithmetic operations, its translation-invariant version needs $O(N \log_2 N)$ arithmetic operations.

In subsection 4.2 we have deduced a numerical scheme for TV diffusion. Each iteration is given by (4.4). This coincides with a single translation-invariant Haar wavelet shrinkage step with threshold $\tau = 2\sqrt{2}t$. Using our operator diff. and the operator S, which shifts a signal one pixel to the right, the result u of the single-scale translation-invariant filter bank is given by

$$u = \frac{1}{2}\left(\mathrm{diff}_{2t}(f) + S^{-1}\mathrm{diff}_{2t}(Sf)\right).$$

Now the multiscale translation-invariant Haar wavelet shrinkage can be interpreted as application of diff. to a multiple Laplacian pyramid. We define a multiple Gaussian pyramid by

$$f^{(0,0)} \to \left(f^{(1,0)}, f^{(1,1)}\right) \to \left(f^{(2,0)}, f^{(2,1)}, f^{(2,2)}, f^{(2,3)}\right) \to \cdots \to \left(f^{(n,0)}, \ldots, f^{(n,2^n-1)}\right),$$

where $f = f^{(0,0)}$. Here $f^{(j,k)}$ is obtained by successive application of the operators R and RS on f as follows: Let 0 denote the application of R and 1 the application of RS; then these operators are applied to f in the order of the binary representation $(k_{j-1}, \ldots, k_0)_2$ of k, where we start from the left. For example, we get $f^{(2,1)} = f^{(2,(0,1)_2)} = RS\,Rf$ and $f^{(2,2)} = f^{(2,(1,0)_2)} = R\,RSf$. Then the multiple Laplacian pyramid is given by

$$\begin{aligned}&\left(f^{(0,0)} - Pf^{(1,0)}, Sf^{(0,0)} - Pf^{(1,1)}\right)\\ &\longrightarrow \left(f^{(1,0)} - Pf^{(2,0)}, Sf^{(1,0)} - Pf^{(2,1)}, f^{(1,1)} - Pf^{(2,2)}, Sf^{(1,1)} - Pf^{(2,3)}\right) \longrightarrow \cdots \longrightarrow\\ &\left(f^{(n,0)}, \ldots, f^{(n,2^n-1)}\right),\end{aligned}$$

and the translation-invariant version of (5.2)–(5.3) can be obtained from this multiple Laplacian pyramid by

$$\begin{aligned}u^{(n,k)} &= f^{(n,k)} \qquad (k = 0, \ldots, 2^n - 1),\\ u^{(j,k)} &= \frac{1}{2}\Big(Pu^{(j+1,2k)} + \mathrm{diff}_{2t}(f^{(j,k)} - Pf^{(j+1,2k)})\\ &\qquad + S^{-1}\left(Pu^{(j+1,2k+1)} + \mathrm{diff}_{2t}(Sf^{(j,k)} - Pf^{(j+1,2k+1)})\right)\Big)\end{aligned}$$

for $j = n-1, \ldots, 0$; $k = 0, \ldots, 2^j - 1$. The result is $u = u^{(0,0)}$.

5.3. Scale-dependent thresholds. Cycle spinning techniques can be used to make wavelet shrinkage not only translationally invariant, but they can also reduce artifacts. However, it is still possible that oscillatory (Gibbs-like) artifacts appear if

multiple scales are used. We want to demonstrate that the use of the scale-dependent thresholds

$$\tau_j = \tau/\sqrt{2}^{j-1} \qquad (j = 1, \dots, n) \tag{5.4}$$

suppresses oscillations in the shrinkage process.

In this subsection, we consider signals $f = (f_0, \dots, f_{N-1})$ with periodic boundary conditions. Note that mirror boundary conditions can easily be transferred into periodic ones by doubling the signal. The decimated Haar wavelet shrinkage with full n-scale decomposition and thresholds (5.4) consists of three operations. It starts with the linear transform (2.2) of f yielding the wavelet coefficients $(c^n, d^n, d^{n-1}, \dots, d^1)$, where $d^j := (d_0^j, \dots, d_{N/2^j-1}^j)$. The wavelet coefficients then undergo the soft wavelet thresholding $s_i^j := S_{\tau_j}(d_i^j)$ $(j = 1, \dots, n;\ i = 0, \dots, N/2^j - 1)$ followed by the inverse linear transform (2.3) of $(c^n, s^n, s^{n-1}, \dots, s^1)$ which gives the denoised signal $u(\tau)$. In particular we have $u(0) = f$. Note that by the semigroup property $S_{\tau+\tilde{\tau}}(x) = S_{\tilde{\tau}}\,(S_\tau(x))$ of our shrinkage function (2.1) the signal $u(K\tau)$ obtained by one n-scale wavelet shrinkage cycle with threshold $K\tau$ coincides with the signal which results from K times repeating one n-scale wavelet shrinkage cycle with smaller threshold τ. Of course, this is no longer true for the translation-invariant wavelet transform. In our examples in the next section we will consider iterated translation-invariant Haar wavelet shrinkage with small thresholds τ.

Since oscillatory (Gibbs-like) artifacts are characterized by the emergence of new local extrema, we study the behavior of local extrema of the signal under the shrinkage process. We call u_i an *extremal pixel* if either $u_{i-1} < u_i$, $u_i > u_{i+1}$ or $u_{i-1} > u_i$, $u_i < u_{i+1}$.

First, we consider the dynamics of "infinitesimal translation-invariant soft Haar wavelet shrinkage," i.e., the speed at which pixels of the signal evolve with respect to the threshold $\tau \in [0, T]$ in the limit case $T \to 0$.

PROPOSITION 5.1 (suppression of Gibbs-like artifacts by scaled thresholds). *Under infinitesimal translation-invariant soft Haar wavelet shrinkage, an extremal pixel f_i evolves as follows:*

(i) *The value of the extremal pixel decreases, i.e., $\dot{u}_i < 0$, if it is a maximum and increases, i.e., $\dot{u}_i > 0$, if it is a minimum. Here the dot denotes differentiation with respect to τ.*

(ii) *The absolute value of the difference of the extremal pixel to each of its two neighbors decreases; i.e., $\dot{u}_i - \dot{u}_{i\pm1} < 0$ for a maximum and $\dot{u}_i - \dot{u}_{i\pm1} > 0$ for a minimum.*

Statement (i) *holds also for the decimated Haar wavelet shrinkage, while statement* (ii) *cannot be established in that setting.*

Proof. For the decimated Haar wavelet shrinkage with full n-scale decomposition and thresholds (5.4) it is easy to check that the resulting signal $\tilde{u}_i$ is given by

$$\tilde{u}_i = \mu + \sum_{j=1}^{n} 2^{-j/2} \varepsilon_j(i) s^j_{\lfloor i/2^j \rfloor},$$

where $\lfloor x \rfloor$ denotes the largest integer $\le x$. Moreover, $\mu := \frac{1}{N} \sum_{i=0}^{N-1} f_i$ is the average value, and

$$\varepsilon_j(i) := \begin{cases} 1 & \text{if } \lfloor i/2^{j-1} \rfloor \text{ is even,} \\ -1 & \text{if } \lfloor i/2^{j-1} \rfloor \text{ is odd.} \end{cases}$$

For the translation-invariant Haar wavelet shrinkage, the sum on the right-hand side of this equation is replaced by the average of N sums of the same kind containing the back-shifted shrunken wavelet coefficients of N forward-shifted initial signals, i.e.,

$$u_i = \mu + \frac{1}{N}\sum_{\nu=0}^{N-1}\sum_{j=1}^{n} 2^{-j/2}\varepsilon_j(i+\nu)s^j_{\lfloor (i+\nu)/2^j\rfloor,\nu}, \tag{5.5}$$

where $s^j_{i,\nu}$ denotes the ith coefficient of the jth level of the ν-shifted initial signal, and the coefficients are treated $N/2^j$-periodic with respect to i. Of course, some coefficients coincide for different ν; more precisely,

$$s^j_{\lfloor (i+\nu)/2^j\rfloor,\nu} = s^j_{\lfloor (i+\nu+r2^j)/2^j\rfloor,\nu+r2^j} \qquad (\nu = 0,\dots,2^j-1; \quad r = 0,\dots,N/2^j-1).$$

This equation allows us to rewrite (5.5) as

$$\begin{aligned} u_i &= \mu + \frac{1}{2\sqrt{2}}\left(\varepsilon_1(i)s^1_{\lfloor i/2\rfloor,0} + \varepsilon_1(i+1)s^1_{\lfloor (i+1)/2\rfloor,1}\right) \\ &\quad + \frac{1}{N}\sum_{j=2}^{n} 2^{-j/2}\sum_{\nu=0}^{2^j-1}\sum_{r=0}^{N/2^j-1}\varepsilon_j(i+\nu+r2^j)s^j_{\lfloor (i+\nu+r2^j)/2^j\rfloor,\nu+r2^j} \\ &= \mu + \frac{s_{i,+}-s_{i,-}}{2\sqrt{2}} + \sum_{j=2}^{n} 2^{-3j/2}\sum_{\nu=0}^{2^j-1}\varepsilon_j(i+\nu)s^j_{\lfloor (i+\nu)/2^j\rfloor,\nu}, \end{aligned}$$

where $s_{i,+} := S_\tau(d_{i,+}) = S_\tau\big((f_i - f_{i+1})/\sqrt{2}\big)$ and $s_{i,-} := S_\tau(d_{i,-}) = S_\tau\big((f_{i-1} - f_i)/\sqrt{2}\big)$. Now the evolution of u_i under infinitesimal soft wavelet shrinkage is described by

$$\dot{u}_i = \frac{\dot{s}_{i,+}-\dot{s}_{i,-}}{2\sqrt{2}} + \sum_{j=2}^{n} 2^{-3j/2}\sum_{\nu=0}^{2^j-1}\varepsilon_j(i+\nu)\dot{s}^j_{\lfloor (i+\nu)/2^j\rfloor,\nu}, \tag{5.6}$$

where

$$\dot{s}^j_{\cdot} = \frac{\mathrm{d}S_{\tau_j}(d^j_{\cdot})}{\mathrm{d}\tau_j}\cdot\frac{\mathrm{d}\tau_j}{\mathrm{d}\tau} = \frac{-\operatorname{sgn}(d^j_{\cdot})}{\sqrt{2}^{j-1}}.$$

Inserting this into (5.6), we obtain

$$\dot{u}_i = \frac{-\operatorname{sgn}(d_{i,+}) + \operatorname{sgn}(d_{i,-})}{2\sqrt{2}} - A_i, \tag{5.7}$$

where

$$A_i := \sqrt{2}\sum_{j=2}^{n} 4^{-j}\sum_{\nu=0}^{2^j-1}\varepsilon_j(i+\nu)\operatorname{sgn}\left(d^j_{\lfloor (i+\nu))/2^j\rfloor,\nu}\right).$$

By the triangle inequality we can estimate

$$|A_i| \le \sqrt{2}\sum_{j=2}^{n} 2^{-j} < \frac{1}{\sqrt{2}}. \tag{5.8}$$

If f_i is an extremal pixel, then we have that $\operatorname{sgn}(d_{i,+}) = -\operatorname{sgn}(d_{i,-}) = 1$ for a maximum and -1 for a minimum. This implies by (5.7) and (5.8) that

$$\operatorname{sgn}(\dot{u}_i) = -\operatorname{sgn}(d_{i,+}), \tag{5.9}$$

proving statement (i) of the proposition.

By subtracting from (5.7) its counterpart for pixel u_{i+1}, we obtain by $d_{i,+} = d_{i+1,-}$ that

$$\dot{u}_i - \dot{u}_{i+1} = \frac{\operatorname{sgn}(d_{i,-}) - 2\operatorname{sgn}(d_{i,+}) + \operatorname{sgn}(d_{i+1,+})}{2\sqrt{2}} - (A_i - A_{i+1}). \tag{5.10}$$

In

$$\begin{aligned} &A_i - A_{i+1} \\ &= \sqrt{2}\sum_{j=2}^{n} 4^{-j} \sum_{\nu=0}^{2^j-1} \left(\varepsilon_j(i+\nu)\operatorname{sgn}\left(d^j_{\lfloor (i+\nu)/2^j \rfloor,\nu}\right) - \varepsilon_j(i+1+\nu)\operatorname{sgn}\left(d^j_{\lfloor (i+1+\nu)/2^j \rfloor,\nu}\right)\right) \end{aligned}$$

the values in the inner brackets cancel except for the two indices $\nu = \nu_k^j \in \{0,\dots,2^j - 1\}$ $(k = 0, 1)$ with $\nu_k^j + 1 + i \equiv 0 \bmod 2^{j-1}$. For these indices the signs of $\varepsilon_j(i + \nu_k)$ and $\varepsilon_j(i+1+\nu_k)$ are opposite. Consequently, for each j, the inner sum contains only four summands, and we can estimate

$$|A_i - A_{i+1}| \le \sqrt{2}\sum_{j=2}^{n} 4^{-j} \cdot 4 < \frac{\sqrt{2}}{3}. \tag{5.11}$$

By inserting this into (5.10), it becomes clear that for an extremal pixel f_i we get

$$\operatorname{sgn}(\dot{u}_i - \dot{u}_{i+1}) = -\operatorname{sgn}(d_{i,+}). \tag{5.12}$$

We have therefore proven that the difference of an extremal pixel to its right neighbor decreases under infinitesimal soft wavelet shrinkage. Analogous considerations apply to the left neighbor, which completes the proof of (ii). □

It follows particularly from Proposition 5.1 that under iterated infinitesimal soft wavelet shrinkage no oscillatory (Gibbs-like) artifacts can appear. Any artifact of this type would include at least one local extremum evolving from a flat region which would, for continuity, have to grow over a finite time interval in contradiction to Proposition 5.1.

It should be noted that a single step of infinitesimal shrinkage does not effectively change the signal any more since $T \to 0$. To investigate true changes of the signal by the shrinkage procedure, one has to consider iterated shrinkage. Summing up τ over all iteration steps, a "total evolution time" t is obtained; for fixed t, the number of iteration steps tends to infinity as τ goes to zero. Infinitesimal translation-invariant soft Haar wavelet shrinkage thus becomes a dynamic process parametrized by t, and Proposition 5.1 describes its behavior at a single point of time.

Of course, this analysis can be extended to a time interval. Then one has to take care of the discontinuity of sgn at 0. Similarly, as in the proof of Proposition 4.1, this can be done by splitting the time axis into intervals in which no sign changes of wavelet coefficients occur. However, since once-merged pixels can split again in the process considered here, sgn(0) will in most cases occur only in discrete time points.

Now we turn to consider finite-size shrinkage steps τ. The ideas used in the proof of Proposition 5.1 can also be applied to analyze soft wavelet shrinkage with finite threshold τ by simply replacing the derivatives $\dot{u}_i$, $\dot{s}^j_{i,\nu}$ by differences $\Delta u_i := u_i(\tau) - u_i(0)$ and $\Delta s^j_{i,\nu} := s^j_{i,\nu} - d^j_{i,\nu}$, respectively. Then we obtain instead of (5.7) that

$$\Delta u_i = \frac{(s_{i,+} - d_{i,+}) - (s_{i,-} - d_{i,-})}{2\sqrt{2}} + A_i,$$

where

$$A_i = \sum_{j=2}^{n} 2^{-3j/2} \sum_{\nu=0}^{2^j-1} \varepsilon_j(i+\nu) \Delta s^j_{\lfloor (i+\nu)/2^j \rfloor, \nu}.$$

By (5.4) and (2.1) we obtain instead of (5.8) the estimate

$$|A_i| \le \tau\sqrt{2} \sum_{j=2}^{n} 2^{-j} < \frac{\tau}{\sqrt{2}}.$$

However, the implication from inequality (5.8) to (5.9) can be transferred only if $|d_{i,+}| \ge \tau$ and $|d_{i,-}| \ge \tau$. Similarly, we conclude instead of (5.10) that

$$\Delta u_i - \Delta u_{i+1} = \frac{-\Delta s_{i,-} + 2\Delta s_{i,+} - \Delta s_{i+1,+}}{2\sqrt{2}} + (A_i - A_{i+1}) \tag{5.13}$$

and estimate the latter difference by

$$|A_i - A_{i+1}| \le \tau\sqrt{2} \sum_{j=2}^{n} 4^{-j}\, 4 < \frac{\tau\sqrt{2}}{3}. \tag{5.14}$$

However, the conclusion from (5.11) to (5.12) can be transferred only if $|\Delta s_{i,-} - 2\Delta s_{i,+} + \Delta s_{i+1,+}| \ge 4\tau/3$. The latter holds true if (but not only if) $|d_{i,+}| \ge \tau$ and $|d_{i,-}| \ge \tau$, i.e., if $f_i - f_{i\pm1} \ge \sqrt{2}\tau$. In this case we obtain by (5.13), (5.14), and their counterparts for the left neighbors of f_i that

$$-\frac{\tau\sqrt{2}}{3} \le u_i(\tau) - u_{i\pm1}(\tau) \le f_i - f_{i\pm1} - \frac{\tau\sqrt{2}}{6}$$

if f_i is a maximum. Analogous inequalities hold true if f_i is a minimum. We can therefore state the following corollary.

COROLLARY 5.2 (behavior of extrema under Haar wavelet shrinkage). *Under translation-invariant soft Haar wavelet shrinkage with thresholds* (5.4) *an extremal pixel f_i, which differs at least by $\sqrt{2}\tau$ from each of its neighbors, evolves as follows:*

(i) *The value of the extremal pixel decreases, i.e., $\Delta u_i < 0$, if it is a maximum and increases, i.e., $\Delta u_i > 0$, if it is a minimum.*

(ii) *The absolute value of the difference of the extremal pixel to each of its two neighbors decreases; in particular, one has $\Delta u_i - \Delta u_{i\pm1} < 0$ for a maximum and $\Delta u_i - \Delta u_{i\pm1} > 0$ for a minimum.*

It can be shown by examples that each of the statements (i) and (ii) of the corollary can be violated if the extremal pixel f_i differs from its neighbors by not more than $\sqrt{2}\tau$. In summary, it follows that Gibbs-like artifacts can in principle still occur under finite-size steps of soft Haar wavelet shrinkage but are restricted in amplitude.

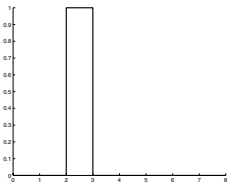


FIG. 5. *Test signal with $N = 8$ pixels.*

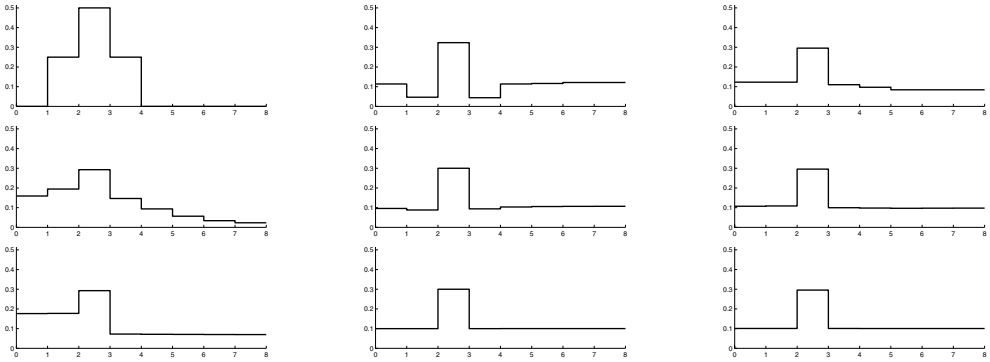


FIG. 6. *$K = 1, 20, 1000$ iterations (top to bottom) of translation-invariant soft Haar wavelet shrinkage with thresholds τ/K applied to the signal in Figure 5. Left column: single-scale wavelet shrinkage with $\tau = 1$. Center column: multiscale wavelet shrinkage ($m = 4$) with uniform threshold $\tau = 0.48$ on all scales. Right column: multiscale wavelet shrinkage ($m = 4$) with $\tau = 0.585$ and scale-adapted thresholds according to* (5.4).

6. Experiments. In this section we illustrate the interplay of iterations and multiscale soft Haar wavelet shrinkage by two examples. As in the previous section we consider initial signals $f = (f_0, \ldots, f_{N-1})$, where $N = 2^n$ is a power of 2. Furthermore, we restrict our attention to reflecting (Neumann) boundary conditions. Then we can perform multiscale wavelet shrinkage up to some assigned scale $m \leq n$.

We start with a simple example which demonstrates the influence of the interplay between iterations and multiscale wavelet shrinkage on Gibbs-like artifacts and its relation to TV diffusion. We consider the initial signal in Figure 5 and apply iterated translation-invariant single-scale and multiscale soft Haar wavelet shrinkage with various threshold parameters. The resulting signals are presented in Figure 6.

Consider the left column of Figure 6. In subsection 4.2 we have shown that translation-invariant soft Haar wavelet shrinkage corresponds to a stable numerical scheme for TV diffusion which represents real TV diffusion if the shrinkage parameter τ is small enough. The first row demonstrates the local effect of the single-scale wavelet shrinkage with threshold $\tau = 1$. The K-times iterated processes with thresholds $\tau = 1/K$ in the second and third rows spread the information globally over the signal. For $K = 1000$, the scheme is a very good approximation to TV diffusion.

The middle and the right columns of Figure 6 deal with multiscale wavelet shrinkage which does not fully correspond to TV diffusion. Already a single iteration results in global effects here. Iterating the multiplescale wavelet shrinkage flattens homogeneous regions, as desired also in TV diffusion. In the middle column, we can observe Gibbs-like phenomena. In the right column, they are avoided by scaling the thresholds.

In our second example we are concerned with the initial signal in Figure 7 ob-

tained using the *WaveLab* package [31]. Figure 8 presents the denoised signal, where the parameter of each method (threshold value or number of iterations) was chosen to optimize the signal-to-noise ratio on the output. We have applied the following techniques:

A. 1 level, regularized TV scheme (4.6) with $\epsilon = \frac{0.04}{2\sqrt{2}}$, iterated with $\tau = \frac{0.01}{2\sqrt{2}}$, $K = 53707$ iterations.

B. 1 level, two-pixel scheme (4.4), iterated with $\tau = 0.01$, $K = 53707$ iterations.

C. 13 levels, 1 iteration, uniform threshold $\tau = 37.4$.

D. 13 levels, iterated, $\tau = 0.01$, $K = 3244$ iterations.

E. 13 levels, 1 iteration, scaled thresholds, $\tau = 92.6$.

F. 13 levels, iterated, $\tau = 0.01$, $K = 7800$ iterations.

The best restoration results in terms of the signal-to-noise ratio are obtained using the regularized TV diffusion scheme (A, SNR=24.6dB), iterated single-scale wavelet shrinkage (B, SNR=24.5dB), or the iterated n-scale wavelet shrinkage with adapted thresholds (F, SNR=24.3dB). Although these methods are not exactly equivalent, they reveal a high level of visual similarity and provide a good piecewise constant approximation to the original signal. The single step multiscale wavelet shrinkage with scale-adapted threshold (E, SNR=21.9dB) performs slightly worse. The single step and iterated multiscale wavelet shrinkage techniques with a uniform threshold on all scales (C, SNR=18.3dB and D, SNR=21.3dB, respectively) are less satisfactory, also visually.

These experiments show that TV denoising outperforms many soft wavelet shrinkage strategies. On the other hand, this is at the expense of a relatively high numerical effort. In order to make wavelets competitive, the shrinkage should be shift invariant, iterative, and use multiple scales with scaled thresholds. In those cases where it is possible to reduce the number of iterations without severe quality degradations, one obtains a hybrid method that combines the speed of multiscale wavelet techniques with the quality of variational or PDE-based denoising methods. For more experiments on multiscale ideas versus iterations we refer the reader to [35].

7. Summary. The goal of the present paper was to investigate under which conditions one can prove equivalence between four discontinuity preserving denoising techniques in the 1-D case: soft wavelet thresholding, TV diffusion, TV regularization, and SIDEs. Starting from a simple two-pixel case we were able to derive analytical solutions. These two-pixel solutions have been used for the following purposes:

- They establish equivalence between soft Haar wavelet shrinkage with threshold parameter τ and TV diffusion of two-pixel signal pairs with diffusion time $t = \tau/\sqrt{2}$.
- They prove also equivalence to TV regularization of two-pixel pairs with regularization parameter $\alpha = \tau/\sqrt{2}$.
- They conjecture equivalence of space-discrete TV diffusion and discrete TV regularization for general N-pixel signals. This conjecture has been proven in subsection 4.3.
- They prove that space-discrete TV diffusion and discrete TV regularization are also equivalent to a SIDE evolution with a TV-based force function. This gives a sound theoretical justification for the heuristically introduced evolution rules for SIDEs.
- They design a novel numerical scheme for TV diffusion of N-pixel signals. It is based on an additive operator splitting into two-pixel interactions where analytical solutions exist for arbitrary large time step sizes. Thus, the numerical scheme is

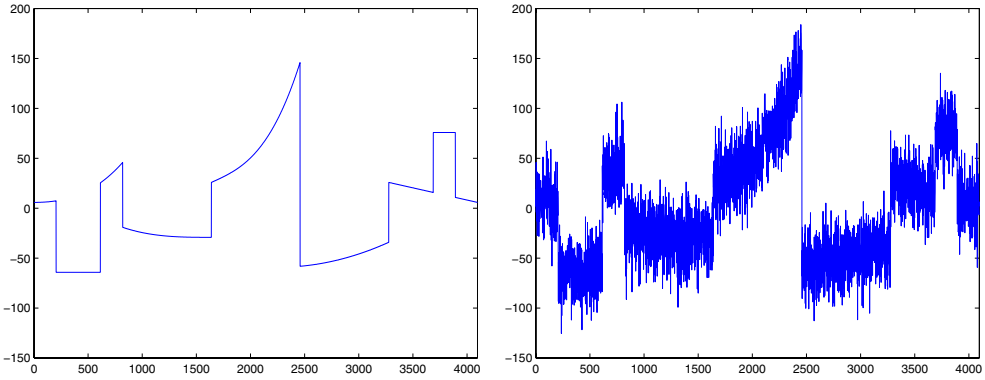


FIG. 7. *Piecewise polynomial signal. Left: original. Right: with additive Gaussian white noise (SNR = 8 dB) as input for the filtering procedures.*

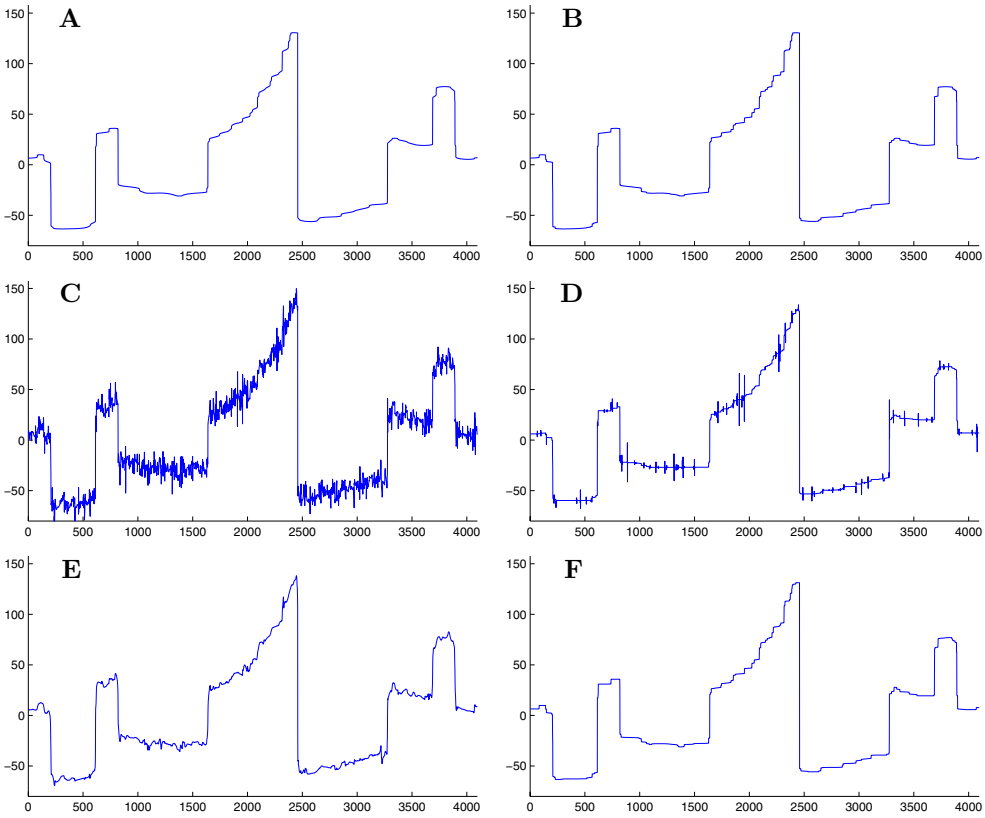


FIG. 8. *Optimal filtering results of several variants of procedures based on TV or wavelet filtering when run on the noisy data of Figure* 7.
A. *Iterated classical scheme for the regularized TV flow* (4.6).
B. *Iterated single-level shrinkage (equivalent to the scheme* (4.4) *for TV flow).*
C. *Multiple levels with a single threshold, single step (i.e., noniterated).*
D. *Iterated multiple level with a single threshold at each of the levels.*
E. *Multiple levels with thresholds scaled according to* (5.4)*, single step.*
F. *Iterated multiple level with scaled thresholds.*
See text for the explanation and numerical evaluation of the results.

explicit and absolutely stable.

We showed that wavelet shrinkage on multiple scales can also be regarded as two-pixel TV diffusion or regularization on the Laplacian pyramid of the signal. On the wavelet side, our experiments show that one can improve the denoising performance by rescaling the thresholds for each wavelet level and by iterating the translation-invariant wavelet shrinkage. On the PDE side, it is possible to achieve a speed-up without significant quality deterioration by using iterated multiple scales instead of iterated single-scale denoising. Thus, the resulting hybrid methods combine the advantages of wavelet and PDE-based denoising.

In our future work we intend to consider more advanced wavelet methods (other shrinkage functions, different wavelets) and to analyze the multidimensional case. In two dimensions, first results on diffusion-inspired wavelet shrinkage with improved rotation invariance are presented in [34]. We will also consider extensions of the numerical two-pixel schemes for TV diffusion.

Acknowledgment. Joachim Weickert thanks Stephen Keeling (Graz, Austria) for interesting discussions on two-pixel signals.

REFERENCES

[1] R. Acar and C. R. Vogel, *Analysis of bounded variation penalty methods for ill–posed problems*, Inverse Problems, 10 (1994), pp. 1217–1229.

[2] L. Alvarez, P.-L. Lions, and J.-M. Morel, *Image selective smoothing and edge detection by nonlinear diffusion.* II, SIAM J. Numer. Anal., 29 (1992), pp. 845–866.

[3] F. Andreu, C. Ballester, V. Caselles, and J. M. Mazón, *Minimizing total variation flow*, Differential Integral Equations, 14 (2001), pp. 321–360.

[4] F. Andreu, V. Caselles, J. I. Diaz, and J. M. Mazón, *Qualitative properties of the total variation flow*, J. Funct. Anal., 188 (2002), pp. 516–547.

[5] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*, Appl. Math. Sci. 147, Springer, New York, 2002.

[6] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[7] Y. Bao and H. Krim, *Towards bridging scale-space and multiscale frame analyses*, in Wavelets in Signal and Image Analysis, Comput. Imaging Vision 19, A. A. Petrosian and F. G. Meyer, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[8] G. Bellettini, V. Caselles, and M. Novaga, *The total variation flow in R^N*, J. Differential Equations, 184 (2002), pp. 475–525.

[9] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, Cambridge, MA, 1987.

[10] T. Brox, M. Welk, G. Steidl, and J. Weickert, *Equivalence results for TV diffusion and TV regularisation*, in Scale-Space Methods in Computer Vision, Lecture Notes in Comput. Sci. 2695, L. D. Griffin and M. Lillholm, eds., Springer, Berlin, 2003, pp. 86–100.

[11] P. J. Burt and E. H. Adelson, *The Laplacian pyramid as a compact image code*, IEEE Trans. Comm., 31 (1983), pp. 532–540.

[12] E. J. Candés and F. Guo, *New multiscale transforms, minimum total variation synthesis: Applications to edge-preserving image reconstruction*, Signal Process., 82 (2002), pp. 1519–1543.

[13] A. Chambolle, R. A. DeVore, N. Lee, and B. L. Lucier, *Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Process., 7 (1998), pp. 319–335.

[14] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[15] A. Chambolle and B. L. Lucier, *Interpreting translationally-invariant wavelet shrinkage as a new image smoothing scale space*, IEEE Trans. Image Process., 10 (2001), pp. 993–1000.

[16] T. F. Chan and H. M. Zhou, *Total variation improved wavelet thresholding in image compression*, in Proceedings of the Seventh International Conference on Image Processing, Vancouver, BC, Canada, 2000.

[17] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore, *Harmonic analysis in the space BV*,

Rev. Mat. Iberoamericana, 19 (2003), pp. 235–262.
[18] A. Cohen, R. DeVore, P. Petrushev, and H. Xu, *Nonlinear approximation and the space $BV(R^2)$*, Amer. J. Math., 121 (1999), pp. 587–628.
[19] R. R. Coifman and D. Donoho, *Translation invariant denoising*, in Wavelets in Statistics, A. Antoine and G. Oppenheim, eds., Springer, New York, 1995, pp. 125–150.
[20] R. R. Coifman and A. Sowa, *Combining the calculus of variations and wavelets for image enhancement*, Appl. Comput. Harmon. Anal., 9 (2000), pp. 1–18.
[21] R. R. Coifman and A. Sowa, *New methods of controlled total variation reduction for digital functions*, SIAM J. Numer. Anal., 39 (2001), pp. 480–498.
[22] D. L. Donoho, *De-noising by soft thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–627.
[23] D. L. Donoho and I. M. Johnstone, *Minimax estimation via wavelet shrinkage*, Ann. Statist., 26 (1998), pp. 879–921.
[24] S. Durand and J. Froment, *Reconstruction of wavelet coefficients using total variation minimization*, SIAM J. Sci. Comput., 24 (2003), pp. 1754–1767.
[25] X. Feng and A. Prohl, *Analysis of total variation flow and its finite element approximations*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 533–556.
[26] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.
[27] A. S. Frangakis, A. Stoschek, and R. Hegerl, *Wavelet transform filtering and nonlinear anisotropic diffusion assessed for signal reconstruction performance on multidimensional biomedical data*, IEEE Transactions on Biomedical Engineering, 48 (2001), pp. 213–222.
[28] W. Hinterberger, M. Hintermüller, K. Kunisch, M. von Oehsen, and O. Scherzer, *Tube methods for BV regularization*, J. Math. Imaging Vision, 19 (2003), pp. 219–235.
[29] T. Lu, P. Neittaanmäki, and X.-C. Tai, *A parallel splitting up method and its application to Navier–Stokes equations*, Appl. Math. Lett., 4 (1991), pp. 25–29.
[30] F. Malgouyres, *Mathematical analysis of a model which combines total variation and wavelet for image restoration*, Inverse Problems, 2 (2002), pp. 1–10.
[31] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, San Diego, 1999.
[32] E. Mammen and S. van de Geer, *Locally adaptive regression splines*, Ann. Statist., 25 (1997), pp. 387–413.
[33] Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, Univ. Lecture Ser. 22, AMS, Providence, RI, 2001.
[34] P. Mrázek and J. Weickert, *Rotationally invariant wavelet shrinkage*, in Pattern Recognition, Lecture Notes in Comput. Sci. 2781, B. Michaelis and G. Krell, eds., Springer, Berlin, 2003, pp. 156–163.
[35] P. Mrázek, J. Weickert, G. Steidl, and M. Welk, *On iterations and scales of nonlinear filters*, in Proceedings of the Eighth Computer Vision Winter Workshop, O. Drbohlav, ed., Valtice, Czech Republic, 2003, Czech Pattern Recognition Society, Peršlák, Czech Republic, 2003, pp. 61–66.
[36] N. Nordström, *Biased anisotropic diffusion – A unified regularization and diffusion approach to edge detection*, Image and Vision Computing, 8 (1990), pp. 318–327.
[37] P. Perona and J. Malik, *Scale space and edge detection using anisotropic diffusion*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (1990), pp. 629–639.
[38] I. Pollak, A. S. Willsky, and H. Krim, *Image segmentation and edge enhancement with stabilized inverse diffusion equations*, IEEE Trans. Image Process., 9 (2000), pp. 256–266.
[39] J. L. Richardson, R. C. Ferrell, and L. N. Long, *Unconditionally stable explicit algorithms for nonlinear fluid dynamics problems*, J. Comput. Phys., 104 (1993), pp. 69–74.
[40] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[41] G. Sapiro, *Geometric Partial Differential Equations and Image Analysis*, Cambridge University Press, Cambridge, UK, 2001.
[42] O. Scherzer and J. Weickert, *Relations between regularization and diffusion filtering*, J. Math. Imaging Vision, 12 (2000), pp. 43–63.
[43] J. Shen and G. Strang, *On wavelet fundamental solutions to the heat equation—Heatlets*, J. Differential Equations, 161 (2000), pp. 403–421.
[44] G. Steidl and J. Weickert, *Relations between soft wavelet shrinkage and total variation denoising*, in Pattern Recognition, Lecture Notes in Comput. Sci. 2449, L. Van Gool, ed., Springer, Berlin, 2002, pp. 198–205.
[45] G. Steidl, J. Weickert, T. Brox, P. Mrázek, and M. Welk, *On the Equivalence of Soft Wavelet Shrinkage, Total Variation Diffusion, Total Variation Regularization, and SIDEs*, Tech. Report 26, Series SPP-1114, Department of Mathematics, University of Bremen,

Bremen, Germany, 2003.
[46] D. M. Strong, *Adaptive Total Variation Minimizing Image Restoration*, Ph.D. thesis, Department of Mathematics, University of California, Los Angeles, CA, 1997.
[47] J. Weickert, *Anisotropic Diffusion in Image Processing*, Teubner, Stuttgart, 1998.
[48] J. Weickert, B. M. ter Haar Romeny, and M. A. Viergever, *Efficient and reliable schemes for nonlinear diffusion filtering*, IEEE Trans. Image Process., 7 (1998), pp. 398–410.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 714–737

A FINITE ELEMENT TECHNIQUE FOR SOLVING FIRST-ORDER PDEs IN L^P*

J. L. GUERMOND†

Abstract. An approximation technique for solving first-order PDEs in $L^p(\Omega)$, $1 \leq p < +\infty$, is proposed. The method is a generalization of the least-squares method to non-Hilbertian settings. A priori and a posteriori error estimates are proven. Numerical tests in $L^1(\Omega)$ show that this type of technique can handle discontinuities without resorting to limiting procedures.

Key words. finite elements, least-squares, first-order PDEs, nonsmooth optimization

AMS subject classifications. 65N35, 65N22, 65F05, 35J05

DOI. 10.1137/S0036142902417054

1. Introduction.

Given a linear, first-order PDE in a domain $\Omega \subset \mathbb{R}^d$,

$$Lu = f, \tag{1.1}$$

with suitable boundary conditions, the objective of this paper is to present an approximation technique that can handle right-hand sides in $L^1(\Omega)$ and, more generally, right-hand sides in $L^p(\Omega)$, $1 \leq p < +\infty$.

1.1. Introductory comments.

The number of attempts at approximating (1.1) directly in $L^1(\Omega)$ seem to be extremely few (see the series of papers by Lavery [29, 30, 28] and the iteratively reweighted least-squares method of Jiang [23] and [24, Chap. 9]) or seem to have encountered some theoretical difficulties (see [32]). This is in sharp contrast with the fact that an enormous amount of work has been dedicated to the study of first-order PDEs and their various nonlinear generalizations in $L^1(\Omega)$. The main difficulty is that when expressed directly in $L^1(\Omega)$ the discrete problems consist in minimizing nondifferentiable functionals; see, e.g., [23]. The lack of theory and of practical popular algorithms for minimizing this type of functional is responsible for the general preference of authors to seek an approximate solution in the $L^2(\Omega)$ framework where differentiability rules, and the force of habit has made this point of view an undisputed paradigm. The goal of the present work is to show that, as claimed in Jiang [23], when the right-hand side is really so rough as to not be in $L^2(\Omega)$ but in $L^1(\Omega)$ only or when the coefficients of the differential operator are so rough that the solution is only meaningful in $L^1(\Omega)$, then it really pays off to approximate the solution to (1.1) directly in $L^1(\Omega)$. In this case, the discontinuities of the solution are captured as sharply as the grid permits without resorting to adaptive refinement, and numerical tests reveal that the method is not plagued by spurious over- or undershootings. Contrary to standard stabilized $L^2(\Omega)$-based techniques, the direct $L^1(\Omega)$ approximation does not require additional ad hoc tunable coefficients or limiting procedures (see, e.g., Galerkin least-squares techniques [16, 25, 26], discontinuous

*Received by the editors October 31, 2002; accepted for publication (in revised form) October 20, 2003; published electronically June 4, 2004. This work was supported by the Centre National de la Recherche Scientifique (CNRS) and the Texas Institute for Computational and Applied Mathematics (TICAM), where the author completed the work while visiting from 08/2001–07/2003 under a TICAM Visiting Faculty Fellowship.
http://www.siam.org/journals/sinum/42-2/41705.html
†LIMSI (CNRS-UPR 3152), BP 133, 91403, Orsay, France (guermond@limsi.fr).

Galerkin methods [31, 17], bubble stabilization [13, 3, 14], or subgrid stabilization [20, 21, 15]).

The paper is organized as follows. In section 2, we introduce an abstract problem, together with its discrete counterpart, and we give an abstract convergence result. We reformulate this result in the $L^p(\Omega)$ setting in section 3, and we describe an algorithm for computing the approximate solution in this setting. We illustrate numerically the method in section 4, where we solve transport equations and advection-diffusion equations in mixed form in $L^1(\Omega)$. We record conclusions in section 5.

1.2. Notation. Let Ω be an open, bounded, connected Lipschitz domain in $\mathbb{R}^d$. We denote by $|\Omega|$ the measure of Ω. For every Lebesgue measurable function $v : \Omega \longrightarrow \mathbb{R}^m$, $m \geq 1$, we denote by $v \cdot w$ the Euclidean scalar product in $\mathbb{R}^m$. For $1 \leq p < +\infty$, we denote by $\|v\|_{\ell^p}$ the discrete ℓ^p-norm of v, i.e., $\|v\|_{\ell^p} = (\sum_{1\leq i\leq m} v_i^p)^{\frac{1}{p}}$. As usual, we denote by $L^p(\Omega)^m$ the real Banach space of $\mathbb{R}^m$-valued functions whose pth power is Lebesgue integrable, i.e., $\|v\|_{L^p(\Omega)^m} = (\int_\Omega \|v(x)\|_{\ell^p}^p dx)^{\frac{1}{p}}$. $W^{1,p}(\Omega)$ is the space of functions in $L^p(\Omega)$ whose partial derivatives in the distributional sense can be identified with functions in $L^p(\Omega)$. $L^\infty(\Omega)$ is the real Banach space of essentially bounded functions. Hereafter we identify the dual of $L^1(\Omega)$ with $L^\infty(\Omega)$.

Considering two real numbers A, B, we shall use the expression $A \lesssim B$ to say that there exists a generic positive constant c, independent of the discretization parameters, such that (s.t.) $A \leq cB$.

2. An abstract problem.

2.1. The continuous setting. Let E and F be two Banach spaces with norms $\|\cdot\|_E$ and $\|\cdot\|_F$, respectively. Let $L : E \longrightarrow F$ be a bounded linear operator, i.e., $L \in \mathcal{L}(E;F)$. We denote by $L^* : F' \longrightarrow E'$ its adjoint, where E' and F' are the duals of E and F, respectively. We assume also that L is bijective. Let us recall the following important consequence of Banach's closed range theorem and open mapping theorem (see, e.g., [12, p. 29] or [36, p. 205]).

LEMMA 2.1. *An operator $L \in \mathcal{L}(E;F)$ is bijective if and only if there is a constant $\alpha > 0$ s.t.*

$$\forall u \in E, \quad \alpha\|u\|_E \leq \|Lu\|_F, \tag{2.1}$$

$$\forall f' \in F', \quad (L^* f' = 0) \Rightarrow (f' = 0). \tag{2.2}$$

We want to solve the following problem: For $f \in F$,

$$\begin{cases} \text{find } u \in E \text{ s.t.} \\ Lu = f \quad \text{in } F. \end{cases} \tag{2.3}$$

This problem is well-posed, and (2.1) yields the following stability property:

$$\|u\|_E \leq \frac{1}{\alpha}\|f\|_F.$$

Let us now introduce an alternative formulation of problem (2.3). Let us define the functional $J : E \longrightarrow \mathbb{R}$ s.t. $J(v) = \|Lv - f\|_F$, and consider the following problem:

$$\begin{cases} \text{Find } u \in E \text{ s.t.} \\ J(u) \leq J(v) \quad \forall v \in E. \end{cases} \tag{2.4}$$

It is clear that problems (2.4) and (2.3) are equivalent in the sense that they have the same unique solution.

To gain more insight on the nature of problem (2.4), let us consider the case where F is a Hilbert space.

PROPOSITION 2.1. *If F is a Hilbert space (equipped with the scalar product $(v,w)_F = \frac{1}{2}(\|v+w\|_F^2 - \|v\|_F^2 - \|w\|_F^2)$), the solution to* (2.4) *is also the unique solution to the following problem:*

$$\begin{cases} \text{Find } u \in E \text{ s.t.} \\ (Lu, Lv)_F = (f, Lv)_F \quad \forall v \in E. \end{cases} \tag{2.5}$$

Proof. J and J^2 have the same minimum, J^2 is clearly differentiable, and (2.5) is the first-order condition for optimality. Owing to (2.1), the bilinear form $(Lu, Lv)_F$ is continuous and coercive, and $(f, Lv)_F$ is continuous; hence, existence and uniqueness of the solution are easy consequences of the Lax–Milgram theorem. ∎

Actually, (2.5) is the so-called least-squares formulation of (2.3), and it can also be interpreted as the Galerkin formulation of the problem

$$L^*Lu = L^*f.$$

Hence, (2.4) is a simple generalization of the least-squares method to non-Hilbertian settings.

2.2. The discrete setting. We now look for an approximate solution to (2.4). Let $(E_h)_{h>0}$ be a sequence of finite-dimensional spaces s.t. $E_h \subset E$. We assume that the sequence of spaces $(E_h)_{h>0}$ has some interpolation properties; that is, we assume that there is a dense normed subspace $W \subset E$ and a function $\epsilon(h)$, continuous at zero with $\epsilon(0) = 0$, s.t.

$$\forall v \in W, \quad \inf_{v_h \in E_h} \|v - v_h\|_E \lesssim \epsilon(h)\|v\|_W. \tag{2.6}$$

The discrete counterpart to (2.4) is as follows:

$$\begin{cases} \text{Find } u_h \in E_h \text{ s.t.} \\ J(u_h) = \min\limits_{v_h \in E_h} J(v_h). \end{cases} \tag{2.7}$$

The main result of this paper is stated in the following theorem.

THEOREM 2.1. (i) *Problem* (2.7) *has at least one global minimizer.*

(ii) *There are no local minimizers.*

(iii) *All minimizers satisfy the following stability property:*

$$\|u_h\|_E \lesssim \|f\|_F. \tag{2.8}$$

(iv) *All minimizers satisfy the a priori error bound*

$$\|u - u_h\|_E \lesssim \min_{v_h \in E_h} \|u - v_h\|_E, \tag{2.9}$$

and the following a posteriori error estimate holds:

$$\|u - u_h\|_E \lesssim \|f - Lu_h\|_F. \tag{2.10}$$

Proof. (i) Let $K_h \subset E_h$ be the ball of radius $\frac{2}{\alpha}\|f\|_F$ centered at 0. It is clear that

$$\inf_{v_h \in E_h} J((v_h) = \min\left(\inf_{v_h \in E_h \setminus K_h} J(v_h), \inf_{v_h \in K_h} J(v_h)\right).$$

But for all $v_h \in E_h \setminus K_h$ (i.e., $\|v_h\|_E > \frac{2}{\alpha}\|f\|_F$) we have

$$\begin{aligned} J(v_h) &\geq \|Lv_h\|_F - \|f\|_F \\ &\geq \alpha\|v_h\|_E - \|f\|_F \\ &> \|f\|_F \\ &> J(0), \end{aligned}$$

where we have used the stability condition (2.1). Since $0 \in K_h$, we infer that

$$\inf_{v_h \in E_h \setminus K_h} J(v_h) > J(0) \geq \inf_{v_h \in K_h} J(v_h).$$

That is to say,

$$\inf_{v_h \in E_h} J(v_h) = \inf_{v_h \in K_h} J(v_h).$$

As a result, the existence of a global minimizer is a simple consequence of the fact that J is continuous and K_h is compact (since E_h is finite-dimensional).

(ii) The functional $J(v_h) = \|Lu_h - f\|_F$ is obviously convex; hence, local minimizers of (2.7) are necessarily global.

(iii) From (i) we infer that any minimizer u_h is in K_h; hence, $\|u_h\|_E \lesssim \|f\|_F$.

(iv) The stability condition (2.1) yields

$$\begin{aligned} \alpha\|u - u_h\|_E &\leq \|Lu - Lu_h\|_F \\ &= \|f - Lu_h\|_F \\ &= \min_{v_h \in E_h} \|f - Lv_h\|_F \\ &= \min_{v_h \in E_h} \|Lu - Lv_h\|_F \\ &\leq \|L\|_{\mathcal{L}(E;F)} \min_{v_h \in E_h} \|u - v_h\|_E. \end{aligned}$$

The proof is complete. ∎

Remark 2.1. Note that the question of the uniqueness of u_h is open. Actually, it may happen that u_h is not unique. To gain some insight on this problem, let us consider $D = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_1 + x_2 \geq 1\}$, $x_0 = (0, 0)$, and let us define S to be the set of points in D that minimize the ℓ^1-distance to x_0. A simple calculation shows $S = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_1 + x_2 = 1,\, x_1 \geq 0,\, x_2 \geq 0\}$; that is, even though the functional and D are convex, the solution to this minimization problem is not unique. Of course, uniqueness would have been guaranteed if we had considered the Euclidean (Hilbertian) distance.

Now, using a standard density argument, we deduce the following corollary.

COROLLARY 2.1. *Under the hypotheses of Theorem* 2.1 *and* (2.6) *we have*

$$\lim_{h \to 0} \|u - u_h\|_E = 0, \tag{2.11}$$

and if $u \in W$*, the following a priori error estimate holds:*

$$\|u - u_h\|_E \lesssim \epsilon(h)\|u\|_W. \tag{2.12}$$

Remark 2.2.

(i) Note that the a priori error estimate (2.12) is optimal since it is bounded by the interpolation error up to a constant.

(ii) Note that the price paid for the approximation optimality when $F = L^1(\Omega)$ is the loss of differentiability. More precisely, the functional $J(v_h) = \|Lu_h - f\|_{L^1(\Omega)}$ is not differentiable; hence, no first-order optimality condition can be written. To better appreciate the difficulty we face here, think of the following two functionals: $\phi(x) = x^2$ and $\psi(x) = |x|$. It is clear that the minimum of ϕ is reached at x_0 when $\phi'(x_0) = 2x_0 = 0$, whereas no nice first-order optimality condition can be written for ψ except for the awkward statement that 0 is in the subdifferential of $\psi(x_0)$, i.e., $0 \in \partial\psi(x_0)$. We describe an algorithm in section 3.6 to solve this difficulty. ∎

3. The $L^p(\Omega)$ setting. We show in this section how the above abstract result can be reformulated in the $L^p(\Omega)$ setting for first-order PDEs.

3.1. Formulation of the problem. In the context of first-order PDEs, F is usually a space $L^p(\Omega)^m$, $1 \le p < \infty$ (or possibly a closed subspace of $L^p(\Omega)^m$), and E is the domain of an unbounded linear operator

$$L : D(L) = E \subset L^p(\Omega)^m \longrightarrow L^p(\Omega)^m = F$$

whose graph is closed in $L^p(\Omega)^m \times L^p(\Omega)^m$ and whose domain $D(L)$ is dense in $L^p(\Omega)^m$ so that when the vector space $E = D(L)$ is equipped with the graph norm $\|v\|_E = (\|v\|^p_{L^p(\Omega)^m} + \|Lv\|^p_{L^p(\Omega)^m})^{\frac{1}{p}}$ it becomes a Banach space.

In this setting, the abstract problem (2.3) is interpreted as follows: For $f \in L^p(\Omega)^m$,

$$\begin{cases} \text{find } u \in E \text{ s.t.} \\ Lu = f \quad \text{ in } L^p(\Omega)^m. \end{cases} \tag{3.1}$$

Owing to the Riesz representation theorem, which permits us to identify the dual of $L^p(\Omega)^m$ with $L^{p'}(\Omega)^m$, where $\frac{1}{p} + \frac{1}{p'} = 1$, this problem can be alternatively put into the following form:

$$\begin{cases} \text{Find } u \in E \text{ s.t.} \\ \displaystyle\int_\Omega \phi \cdot Lu = \int_\Omega f \cdot \phi \quad \forall \phi \in L^{p'}(\Omega)^m. \end{cases} \tag{3.2}$$

3.1.1. Example 1: Advection-reaction. Let us consider an advection-reaction problem. Let $\boldsymbol{\beta}$ be a smooth vector field in $\mathbb{R}^d$, say $\boldsymbol{\beta} \in L^\infty(\Omega)^d$ and $\nabla\cdot\boldsymbol{\beta} \in L^\infty(\Omega)$, and set

$$\partial\Omega^- = \{x \in \partial\Omega \mid \boldsymbol{\beta}(x) \cdot \mathbf{n}(\mathbf{x}) < \mathbf{0}\},$$
$$\partial\Omega^+ = \{x \in \partial\Omega \mid \boldsymbol{\beta}(x) \cdot \mathbf{n}(\mathbf{x}) > \mathbf{0}\}.$$

$\partial\Omega^-$ is the inflow boundary, $\partial\Omega^+$ is the outflow boundary, and $\mathbf{n}(\mathbf{x})$ is the unit exterior normal to $\partial\Omega$ at $x \in \partial\Omega$. It may happen that these two subsets of $\partial\Omega$ are empty if $\boldsymbol{\beta}$ is s.t. $\boldsymbol{\beta} \cdot \mathbf{n}(\mathbf{x}) = \mathbf{0}$ for all $x \in \partial\Omega$. Let μ be a function in $L^\infty(\Omega)$, and assume that there is a constant $\mu_0 > 0$ so that

$$\mu(x) \ge \mu_0 > 0 \quad \text{a.e. } x \text{ in } \Omega. \tag{3.3}$$

We introduce the differential operator

$$L(u) = \mu u + \nabla\cdot(u\boldsymbol{\beta}),$$

with domain

$$E = D(L) = \{w \in L^1(\Omega);\ \nabla\cdot(w\boldsymbol{\beta}) \in L^1(\Omega);\ \boldsymbol{\beta}\cdot\mathbf{n}|_{\partial\Omega^-} = 0\} \subset L^1(\Omega) = F.$$

It can be shown that L is an isomorphism from E to F; i.e., (2.1) and (2.2) hold.

Remark 3.1. If $\mu = 0$, the hypothesis (3.3) is not satisfied. Nevertheless, L is still an isomorphism if $\boldsymbol{\beta}$ is a smooth filling field, i.e., if for almost every x in Ω there is a characteristic of $\boldsymbol{\beta}$ that starts from x and reaches $\partial\Omega^-$ in finite time. The reader is referred to Azerad and Pousin [1] for other details on this problem.

3.1.2. Example 2: The Darcy equation. Let Ω be a porous medium characterized by the permeability tensor $\mathbf{K}(\mathbf{x})$. This tensor is assumed to be symmetric positive definite, and its smallest and largest eigenvalues are assumed to be bounded from below and from above uniformly in Ω. We consider the following problem:

$$\begin{cases} \mathbf{K^{-1}}\cdot\mathbf{u} + \nabla\mathbf{p} = \mathbf{f}, \\ \nabla\cdot\mathbf{u} + \alpha\mathbf{p} = \mathbf{g}, \\ p_{|\partial\Omega} = 0. \end{cases} \tag{3.4}$$

This problem is known as the Darcy problem. It is also the mixed form of the Poisson problem. Nonlinear versions of (3.4) play important roles in underground storage problems, hydrogeology, and the petroleum industry. It is very often coupled with a transport equation for the concentration of a chemical species or a phase fraction.

To formulate (3.4) in the $L^p(\Omega)$ setting, we introduce some definitions:

$$\begin{aligned} X &= \{\mathbf{v} \in \mathbf{L^p(\Omega)^d};\ \nabla\cdot\mathbf{v} \in \mathbf{L^p(\Omega)}\}, \\ \|\mathbf{v}\|_{\mathbf{X}} &= (\|\mathbf{v}\|^{\mathbf{p}}_{\mathbf{L^p(\Omega)^d}} + \|\nabla\cdot\mathbf{v}\|^{\mathbf{p}}_{\mathbf{L^p(\Omega)}})^{\frac{1}{\mathbf{p}}}, \\ Y &= \{q \in L^p(\Omega);\ \nabla q \in L^p(\Omega)^d,\ q_{|\partial\Omega} = 0\}, \\ \|q\|_Y &= \|q\|_{W^{1,p}(\Omega)} = (\|q\|^p_{L^p(\Omega)} + \|\nabla q\|^p_{L^p(\Omega)^d})^{\frac{1}{p}}. \end{aligned}$$

X and Y are Banach spaces. We set $E = X \times Y$ and $F = L^p(\Omega)^d \times L^p(\Omega)$, which we equip with the norms $\|(\mathbf{v},\mathbf{q})\|_{\mathbf{E}} = (\|\mathbf{v}\|^{\mathbf{p}}_{\mathbf{X}} + \|\mathbf{q}\|^{\mathbf{p}}_{\mathbf{Y}})^{\frac{1}{\mathbf{p}}}$ and $\|(\mathbf{v},\mathbf{q})\|_{\mathbf{F}} = (\|\mathbf{v}\|^{\mathbf{p}}_{\mathbf{L^p(\Omega)^d}} + \|\mathbf{q}\|^{\mathbf{p}}_{\mathbf{L^p(\Omega)}})^{\frac{1}{\mathbf{p}}}$, respectively. We now define the operator

$$\begin{aligned} &L : E \longrightarrow F, \\ &(\mathbf{v},\mathbf{q}) \longmapsto (\mathbf{K^{-1}}\mathbf{v} + \nabla\mathbf{q}, \nabla\cdot\mathbf{v} + \alpha\mathbf{q}). \end{aligned}$$

L is clearly continuous, and it can be shown that it is an isomorphism if $\alpha \geq 0$, for $1 < p < +\infty$, and if $\alpha > 0$ for $p = 1$ (see, e.g., [6, 11, 34]).

3.2. Friedrichs's systems. The above two examples are particular cases of Friedrichs's symmetric systems; see [19]. Most of what is said hereafter generalizes to this broad class of PDEs.

3.3. The discrete setting. Henceforth, we assume that F is a closed subspace of $L^p(\Omega)^m$. We assume also that we are given a sequence of regular finite element meshes $(\mathcal{T}_h)_{h>0}$ covering the domain Ω. With each mesh we associate a finite-dimensional space $E_h \subset E$ having some interpolation properties; that is, there is a dense normed subspace of smooth functions $W \subset E$ and a continuous function $\epsilon(h)$ with $\epsilon(0) = 0$ s.t. (2.6) holds. For $\mathbb{P}_k$ or $\mathbb{Q}_k$ Lagrange finite elements, we have $\epsilon(h) = h^k$, where h is the meshsize.

3.4. A brief review of some standard techniques. One of the standard ways of approximating (3.2) without invoking a minimization principle like (2.7) is the Galerkin technique. This method consists in replacing the solution space, E, and the test space, $L^{p'}(\Omega)^m$, by the same discrete space E_h as follows:

$$\begin{cases} \text{Find } u_h \in E_h \text{ s.t.} \\ \displaystyle\int_\Omega \phi_h \cdot Lu_h = \int_\Omega \phi_h \cdot f \quad \forall \phi_h \in E_h. \end{cases} \tag{3.5}$$

Note that using the same space for testing the equation and approximating the solution guarantees that the corresponding linear system has as many equations as unknowns. Even though it often happens that the discrete solution is unique, (3.5) does not yield stability in the E-norm in general. To better appreciate this point, let us consider the scalar problem $u' = f$ with $u(0) = 0$ in the one-dimensional (1D) domain $\Omega =]0,1[$, where we assume $f \in L^2(\Omega)$. For $N \in \mathbb{N}^\star$, let us set $h = 1/N$ and $x_i = ih$ for $i \in \{0, 1, \ldots, N\}$. We define

$$E_h = \{v_h \in \mathcal{C}^0(\overline{\Omega});\ v_{h|[x_i,x_{i+1}]} \in \mathbb{P}_1,\ 0 \le i \le N-1;\ v_h(0) = 0\}. \tag{3.6}$$

It is clear that $E_h \subset E = \{v \in H^1(\Omega);\ v(0) = 0\}$. The discrete Galerkin formulation of the problem is as follows:

$$\begin{cases} \text{Find } u_h \text{ in } E_h \text{ s.t.} \\ \displaystyle\int_0^1 v_h u_h' = \int_0^1 v_h f \qquad \forall v_h \in E_h, \end{cases} \tag{3.7}$$

and its stability constant (i.e., the counterpart of α in (2.1)) is

$$\alpha_h := \inf_{u_h \in E_h} \sup_{v_h \in E_h} \frac{\int_0^1 u_h' v_h}{\|u_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}.$$

The following negative result can be proved.

THEOREM 3.1. *There are two constants $c_1 > 0$ and $c_2 > 0$, independent of h, s.t.*

$$c_1 h \le \alpha_h \le c_2 h.$$

Proof. See, e.g., Ern and Guermond [18, pp. 197–199]. ∎

In other words, the stability constant for the approximate problem (3.7) goes to zero as the mesh is refined. This result is the main reason for the failure of the Galerkin technique to work properly for first-order PDEs in general.

An interesting alternative to the Galerkin formulation consists in the least-squares formulation. The origins of the least-squares technique can be traced back to Gauss (*Theoria Motus Corporum Coelestium* (1809)). As early papers in the numerical analysis literature we cite the series of papers by Bramble and Schatz [9, 10] published in 1970. Since then, it has been applied to a wide variety of problems (see, e.g., [2, 33, 24]). This method is clearly optimal in the $L^2(\Omega)$-graph norm, but it performs poorly when the source term is not in $L^2(\Omega)$ but in $L^1(\Omega)$ only or the boundary data are discontinuous (see the numerical tests in section 4).

The list of alternative techniques for solving (3.2) is quite long, and it is out of the question to make this list exhaustive, but among the most popular ones is the so-called Galerkin least-squares method [16, 25], which combines the accuracy of the Galerkin

method and the stability properties of the least-squares method. Other methods of interest are those based on discontinuous interpolation spaces (e.g., discontinuous Galerkin methods [31, 17]), on bubble functions (e.g., residual free bubble methods [13, 3, 14]), or on a hierarchical decomposition of the approximation space (e.g., subgrid stabilization [20, 21, 15] or spectral viscosity [35]). Although all these methods are quite efficient in general, they cannot cope with discontinuities and boundary layers without resorting to shock-capturing and nonlinear limiting techniques [26, 22] since they are all L^2-based; i.e., they rely on a priori L^2 estimates.

3.5. The discrete problem and regularization. Upon setting $J(v) = \|Lv - f\|_{L^p(\Omega)^m}$, the minimization problem we would like to solve is to find u_h in E_h such that $J(u_h) = \min_{v_h \in E_h} J(v_h)$. Actually, since $\mathbb{R}^+ \ni x \longmapsto x^p$ is an increasing function, an equivalent reformulation consists of setting

$$\mathcal{J}(v) = \|Lv - f\|^p_{L^p(\Omega)^m}$$

and considering the following problem:

$$\begin{cases} \text{Find } u_h \in E_h \text{ s.t.} \\ \mathcal{J}(u_h) = \min\limits_{v_h \in E_h} \mathcal{J}(v_h). \end{cases} \tag{3.8}$$

To handle this possibly nondifferentiable minimization problem by means of standard gradient techniques, we propose to regularize it as follows. Let us define $\varepsilon > 0$ and introduce

$$\varphi_\varepsilon(r) = r^2(r+\varepsilon)^{p-2}. \tag{3.9}$$

Then we regularize $\mathbb{R}^m \ni x \longmapsto \|x\|^p_{\ell^p}$ by replacing this function by

$$\psi_\varepsilon(x) = \sum_{i=1}^m \varphi_\varepsilon(|x_i|). \tag{3.10}$$

Upon denoting by $\mathrm{sg}(t)$ the sign function (i.e., $\mathrm{sg}(t) = t/|t|$ if $t \neq 0$ and $\mathrm{sg}(0) = 0$), we have

$$\forall v \in \mathbb{R}^m, \quad D\psi_\varepsilon(x) \cdot v = \sum_{i=1}^m \varphi'_\varepsilon(|x_i|)\mathrm{sg}(x_i)v_i, \tag{3.11}$$

$$\forall v, w \in \mathbb{R}^m, \quad w \cdot D^2\psi_\varepsilon(x) \cdot v = \sum_{i=1}^m \varphi''_\varepsilon(|x_i|)v_i w_i. \tag{3.12}$$

Note that φ''_ε is a decreasing function on $\mathbb{R}^+$ if $1 \le p \le 2$, and it is an increasing function if $p \ge 2$. More precisely, we have the following:

$$\text{if } 1 \le p \le 2, \quad \exists c > 0 \ \forall a > 0, \ \forall r \in [0,a], \quad c\,a^{p-2} \le \varphi''_\varepsilon(r) \le 2\,\varepsilon^{p-2}, \tag{3.13}$$

$$\text{if } 2 \le p, \quad \exists c > 0 \ \forall a > 0, \ \forall r \in [0,a], \quad 2\,\varepsilon^{p-2} \le \varphi''_\varepsilon(r) \le c\,a^{p-2}. \tag{3.14}$$

This, in turn, implies the following property:

$$\text{if } 1 \le p \le 2 \quad \begin{cases} \forall y \in \mathbb{R}^m, & c\,\|y\|^{p-2}_{\ell^\infty}\,\|y\|^2_{\ell^2} \le y \cdot D^2\psi_\varepsilon(x) \cdot y, \\ \forall y, z \in \mathbb{R}^m, & |z \cdot D^2\psi_\varepsilon(x) \cdot y| \le 2\,\varepsilon^{p-2}\,\|z\|_{\ell^2}\|y\|_{\ell^2}, \end{cases} \tag{3.15}$$

and an obvious similar property holds if $p \ge 2$.

Now we introduce the regularized functional

$$\mathcal{J}_\varepsilon(v_h) = \int_\Omega \psi_\varepsilon(Lu_h - f), \tag{3.16}$$

and we define u_h^ε to be a solution to the following minimization problem:

$$\begin{cases} \text{Find } u_h^\varepsilon \in E_h \text{ s.t.} \\ \mathcal{J}_\varepsilon(u_h^\varepsilon) = \min\limits_{v_h \in E_h} \mathcal{J}_\varepsilon(v_h). \end{cases} \tag{3.17}$$

It is clear that, owing to the regularization, $\mathcal{J}_\varepsilon$ is differentiable (in the Fréchet sense), and the first-order optimality condition for (3.17) is

$$\int_\Omega D\psi_\varepsilon(Lu_h^\varepsilon - f) \cdot Lv_h = 0 \quad \forall v_h \in E_h.$$

The algorithm that we propose in the next section consists in obtaining a solution to (3.8) as a limit of a sequence $(u_h^\varepsilon)_{\varepsilon>0}$ as $\varepsilon \to 0$.

Now we give a series of lemmas clarifying the stability of u_h^ε with respect to the data, the uniqueness of u_h^ε, the convergence of the sequence $(u_h^\varepsilon)_{\varepsilon>0}$ as $\varepsilon \to 0$, and, finally, the convergence of the sequence $(u_h^\varepsilon)_{\varepsilon>0,h>0}$ as both ε and h go to zero.

LEMMA 3.1. *Solutions to* (3.17) *satisfy the following stability estimate:*

$$\alpha \|u_h^\varepsilon\|_E \le \|f\|_{L^p(\Omega)^m} + (m\,\varepsilon^p|\Omega| + 2\|\psi_\varepsilon(f)\|_{L^1(\Omega)})^{\frac{1}{p}}.$$

Proof. Owing to the definition of φ_ε, it is clear that for all $g \in L^p(\Omega)^m$,

$$1 \le i \le m, \qquad \frac{1}{2}\int_{\{|g_i|\ge\varepsilon\}} |g_i|^p \le \int_\Omega \varphi_\varepsilon(|g_i|).$$

As a result,

$$\int_\Omega |g_i|^p = \int_{\{|g_i|<\varepsilon\}} |g_i|^p + \int_{\{|g_i|\ge\varepsilon\}} |g_i|^p \le \varepsilon^p|\Omega| + 2\int_\Omega \varphi_\varepsilon(|g_i|)$$

and

$$\int_\Omega \|g\|_{\ell^p}^p \le m\,\varepsilon^p|\Omega| + 2\int_\Omega \psi_\varepsilon(g).$$

Hence,

$$\int_\Omega \|Lu_h^\varepsilon - f\|_{\ell^p}^p \le m\,\varepsilon^p|\Omega| + 2\int_\Omega \psi_\varepsilon(Lu_h^\varepsilon - f) \le m\,\varepsilon^p|\Omega| + 2\int_\Omega \psi_\varepsilon(f).$$

The triangle inequality, together with (2.1), yields the result. ∎

LEMMA 3.2. *If $f \in L^\infty(\Omega)^m$, there is a unique function u_h^ε minimizing $\mathcal{J}_\varepsilon$.*

Proof. Let u_h^1 and u_h^2 be two functions in E_h. We have

$$\begin{aligned} D\mathcal{J}_\varepsilon(u_h^1)(v_h) - D\mathcal{J}_\varepsilon(u_h^2)(v_h) &= \int_\Omega [D\psi_\varepsilon(Lu_h^1 - f) - D\psi_\varepsilon(Lu_h^2 - f)] \cdot Lv_h \\ &= \int_\Omega L(u_h^1 - u_h^2) \cdot \int_0^1 D^2\psi_\varepsilon(R(s))ds \cdot Lv_h, \end{aligned}$$

where $R(s) = L(su_h^1 + (1-s)u_h^2) - f$. Now using $v_h = u_h^1 - u_h^2$ as a test function and making use of (3.15), we infer that

$$(D\mathcal{J}_\varepsilon(u_h^1) - D\mathcal{J}_\varepsilon(u_h^2))(u_h^1 - u_h^2) \geq \int_\Omega \alpha_\varepsilon(u_h^1, u_h^2, f)\|L(u_h^1 - u_h^2)\|_{\ell^2}^2,$$

where $\alpha_\varepsilon(u_h^1, u_h^2, f) = c \inf_{0\leq s\leq 1} \min(\|R(s)\|_{\ell^\infty}^{p-2})$ if $1 \leq p \leq 2$ and $\alpha_\varepsilon(u_h^1, u_h^2, f) = 2\varepsilon^{p-2}$ if $p \geq 2$.

If u_h^1 and u_h^2 both minimize $\mathcal{J}_\varepsilon$, then, owing to inverse inequalities, both of these functions are bounded. Since f is also assumed to be bounded, we necessarily have $\|Lu_h^i - f\|_{L^\infty(\Omega)^m} < +\infty$, $i = 1, 2$; that is to say, ess $\inf_\Omega \alpha_\varepsilon(u_h^1, u_h^2, f) > 0$ and

$$0 \geq \operatorname*{ess\,inf}_\Omega \alpha_\varepsilon(u_h^1, u_h^2, f) \int_\Omega \|L(u_h^1 - u_h^2)\|_{\ell^2}^2,$$

which yields $u_h^1 = u_h^2$ since L is injective. ∎

Since E_h is finite-dimensional (hence locally compact), a first consequence of Lemma 3.1 is that, up to a subsequence, $(u_h^\varepsilon)_{\varepsilon>0}$ converges to some u_h^0 in E_h.

LEMMA 3.3. *Every limit u_h^0 of $(u_h^\varepsilon)_{\varepsilon>0}$, up to a subsequence, is a solution to the unregularized minimization problem* (3.8).

Proof. First, let us observe that

$$\forall p \geq 1,\ \forall x \in \mathbb{R}, \qquad |\varphi_\varepsilon(|x|) - |x|^p| \leq (2+p)2^p\,\varepsilon\,(|x|^{p-1} + \varepsilon^{p-1}).$$

As a result, for all $x \in E$, we have

$$\begin{aligned}|\mathcal{J}_\varepsilon(x) - \mathcal{J}(x)| &\leq (2+p)2^p\,\varepsilon \int_\Omega \left(\sum_{i=1}^m |L_i x - f_i|^{p-1} + \varepsilon^{p-1}\right) \\ &\leq (2+p)2^p\,\varepsilon\,(m\,\varepsilon^{p-1}|\Omega| + m^{\frac{1}{p}}|\Omega|^{\frac{1}{p}}\mathcal{J}(x)^{\frac{p-1}{p}}).\end{aligned} \tag{3.18}$$

Furthermore, for all x^1 and x^2 in E, we have

$$|\mathcal{J}(x^1)^{\frac{1}{p}} - \mathcal{J}(x^2)^{\frac{1}{p}}| = |\|Lx^1 - f\|_F - \|Lx^2 - f\|_F| \leq \|L\|\,\|x^1 - x^2\|_E.$$

Combining these two results, we infer that if $x^\varepsilon \to x^0$ in E, then $\mathcal{J}_\varepsilon(x^\varepsilon) \to \mathcal{J}(x^0)$. Then we have

$$\mathcal{J}(u_h^0) = \lim_{\varepsilon\to 0} \mathcal{J}_\varepsilon(u_h^\varepsilon) = \lim_{\varepsilon\to 0} \min_{v_h\in E_h} \mathcal{J}_\varepsilon(v_h) = \min_{v_h\in E_h} \mathcal{J}(v_h),$$

which means that u_h^0 minimizes $\mathcal{J}$ in E_h. ∎

LEMMA 3.4. *Every solution to problem* (3.17) *is such that*

$$\|u - u_h^\varepsilon\|_E \lesssim \left(\varepsilon\,c(\|f\|_F) + \min_{v_h\in E_h} \|u - v_h\|_E^p\right)^{\frac{1}{p}},$$

where $c(\cdot)$ is a continuous function.

Proof. Owing to Lemma 3.1 and (3.18), we infer that

$$\begin{aligned}0 \leq \mathcal{J}(u_h^\varepsilon) - \mathcal{J}(u_h) &\leq \mathcal{J}(u_h^\varepsilon) - \mathcal{J}_\varepsilon(u_h^\varepsilon) + \mathcal{J}_\varepsilon(u_h^\varepsilon) - \mathcal{J}_\varepsilon(u_h) \\ &\quad + \mathcal{J}_\varepsilon(u_h) - \mathcal{J}(u_h) \\ &\leq \mathcal{J}(u_h^\varepsilon) - \mathcal{J}_\varepsilon(u_h^\varepsilon) + \mathcal{J}_\varepsilon(u_h) - \mathcal{J}(u_h) \\ &\lesssim \varepsilon\left(\varepsilon^{p-1} + \mathcal{J}(u_h^\varepsilon)^{\frac{p-1}{p}} + \mathcal{J}(u_h)^{\frac{p-1}{p}}\right) \\ &\lesssim \varepsilon\,c(\|f\|_F).\end{aligned}$$

Hence,

$$\begin{aligned}\|u-u_h^\varepsilon\|_E^p &\lesssim \|f-Lu_h^\varepsilon\|_F^p = \mathcal{J}(u_h^\varepsilon)-\mathcal{J}(u_h)+\mathcal{J}(u_h)\\ &\lesssim \varepsilon\, c(\|f\|_F)+\min_{v_h\in E_h}\|u-v_h\|_E^p.\end{aligned}$$

The proof is complete. ∎

Remark 3.2. Lemma 3.4 guarantees that if $\varepsilon^{1/p}$ is smaller than the interpolation error, then u_h^ε is as good an approximation of u as u_h. Note also that the smaller the p the smaller the error induced by regularization, and regularization is needed only if $1\le p<2$.

3.6. A simple algorithm for solving (3.8). We now present a simple algorithm for solving (3.8). The main idea is to set a sequence of regularization parameters $(\varepsilon^k)_{k\ge 0}$ tending to zero (or some numerically acceptable threshold) as k grows and then, for each parameter ε^k, to find a reasonable approximation of the minimizer of $\mathcal{J}_\varepsilon$ using Newton's algorithm and starting from the approximate minimizer evaluated at step $k-1$. More precisely, the algorithm we propose is as follows:

Step 1: Initialize ε^0 (say, $\varepsilon^0\sim h$) and compute some initial guess u_h^0 (use a crude $L^2(\Omega)$-stabilized technique; for instance, add a Laplace perturbation to the equation and evaluate the Galerkin solution, or evaluate the least-squares solution).

Step 2: Iterate on index k, starting from $k=0$.

Step 3$_k$: Set $u_h^{k,0}=u_h^k$.

Step 4$_k$: Iterate on index l, starting from $l=0$.

Step 5$_{k,l}$: Evaluate the gradient and the Hessian of $\mathcal{J}_{\varepsilon^k}(u_h^{k,l})$ as follows:

$$D\mathcal{J}_{\varepsilon^k}(u_h^{k,l})(v_h)=\int_\Omega D\psi_{\varepsilon^k}(Lu_h^{k,l}-f)\cdot Lv_h, \tag{3.19}$$

$$D^2\mathcal{J}_{\varepsilon^k}(u_h^{k,l})(v_h,w_h)=\int_\Omega Lw_h\cdot D^2\psi_{\varepsilon^k}(Lu_h^{k,l}-f)\cdot Lv_h. \tag{3.20}$$

Step 6$_{k,l}$: Deduce a descent direction, d_h, by solving the following problem:

$$\begin{cases}\text{Find } d_h\in X_h \text{ s.t.}\\ D^2\mathcal{J}_{\varepsilon^k}(u_h^{k,l})(v_h,d_h)=D\mathcal{J}_{\varepsilon^k}(u_h^{k,l})(v_h)\quad \forall v_h\in X_h.\end{cases} \tag{3.21}$$

Note that $D^2\psi_{\varepsilon^k}>0$; hence, in addition to being symmetric, the bilinear form $D^2\mathcal{J}_{\varepsilon^k}(u_h^{k,l})(v_h,w_h)$ is always positive definite; that is, (3.21) has always a unique solution.

Step 7$_{k,l}$: Make a line search of the minimum of $\mathcal{J}$ along the direction d_h. Call the corresponding solution $u_h^{k,l+1}$.

Step 8$_{k,l}$: If $\|u_h^{k,l+1}-u_h^{k,l}\|_E^p$ is smaller than ε^k, set $u_h^{k+1}=u_h^{k,l+1}$ and exit the l loop; otherwise continue iterations on l.

Step 9$_k$: If ε^k is smaller than some fixed tolerance, exit the k loop; otherwise divide ε^k by some fixed constant, say $\frac{3}{2}$, call the result ε^{k+1}, and continue iterations on k.

Step 10: Stop.

Remark 3.3.

(i) Note that at Step 7 the line search minimizes $\mathcal{J}$; hence, the algorithm always makes $\mathcal{J}$ decrease.

(ii) Note that in the above algorithm ε is not a tunable coefficient; i.e., this coefficient cannot be compared to any stabilizing parameter usually introduced by L^2-based stabilizing techniques (e.g., GaLS, residual free bubbles, subgrid viscosity, etc.). The sequence $(\varepsilon^k)_{k\geq 0}$ is meant to accelerate the convergence process, and it goes to zero as the number of iterations grows.

(iii) Note that the cost of one loop of the algorithm above is that of evaluating the Hessian and solving for the descent direction; however, (3.21) does not need to be solved very accurately. The computational cost per loop is identical to that of an approximate Galerkin solve. Hence, the total cost of the algorithm is that of an approximate Galerkin solve times the number of loops. In the examples reported below, the number of loops required to reach a reasonable convergence criterion was between 10 to 25 when using $u_h^0 = 0$. It is very likely that this crude algorithm is not optimal, and further research is needed to improve on this aspect of the problem. One can imagine, for instance, embedding the above algorithm within a multigrid strategy and/or some adaptive refinement strategy.

(iv) Using a stabilized L^2-based technique to compute u_h^0 significantly shortens the number of iterations in the above algorithm. In this context minimizing the residual in L^1 could be viewed as postprocessing for the L^2-based method.

(v) The above regularization-based iterative algorithm has some similarities with the so-called iteratively reweighted least-squares method of Jiang [24, Chap. 9].

(vi) When the operator L is nonlinear, the above algorithm still holds, provided formulas (3.19) and (3.20) defining the gradient and the Hessian of $\mathcal{J}_{\varepsilon^k}(u_h^{k,l})$ are modified accordingly. In this context, solving the problem in any $L^p(\Omega)$ does not cost more than solving the problem in the standard $L^2(\Omega)$ setting.

4. Numerical results. We report in this section on results of numerical tests meant to assess the theoretical a priori error estimates derived above and to illustrate the performance of the method when dealing with nonsmooth data and nonlinear problems. Unless stated explicitly otherwise, all the numerical tests reported herein have been performed in $L^1(\Omega)$.

4.1. Convergence tests.

4.1.1. A transport equation. We consider the two-dimensional (2D) domain $\Omega =]0,1[^2$ and the transport equation

$$\partial_x u = f, \quad u|_{x=0} = u_0, \tag{4.1}$$

with smooth data

$$f(x,y) = 2\pi\cos(2\pi(x+y)), \quad u_0(y) = \sin(2\pi y),$$

s.t. the exact solution is

$$u = \sin(2\pi(x+y)).$$

We approximate the solution using piecewise linear and piecewise quadratic triangular elements on unstructured Delaunay meshes. We compute the approximate solution of (4.1) in $L^1(\Omega)$. For the $\mathbb{P}_1$ solution we use meshes s.t. $\frac{1}{10} \leq h \leq \frac{1}{100}$, and for the $\mathbb{P}_2$ solution we take h in the range $\frac{1}{5} \leq h \leq \frac{1}{60}$. All the integrals are evaluated by using the 3 Gauss points quadrature rule for the $\mathbb{P}_1$ solution and the 7 Gauss points

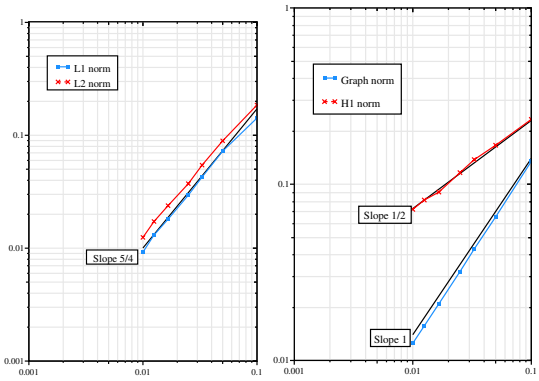


FIG. 1. *Convergence tests for $\mathbb{P}_1$ approximation. Left: Error in the L^1-norm and L^2-norm vs. the meshsize. Right: Error in the $L^1(\Omega)$-graph norm and H^1-norm vs. the meshsize.*

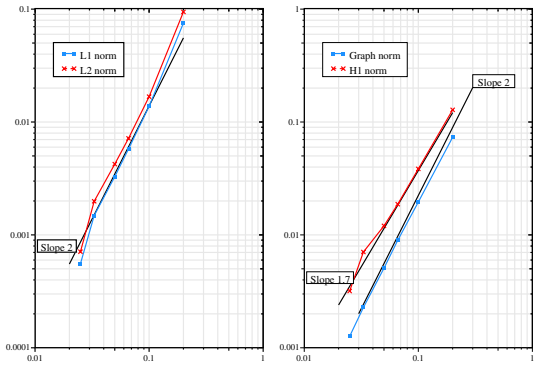


FIG. 2. *Convergence tests for $\mathbb{P}_2$ approximation. Left: Error in the L^1-norm and L^2-norm vs. the meshsize. Right: Error in the $L^1(\Omega)$-graph norm and H^1-norm vs. the meshsize.*

quadrature rule for the $\mathbb{P}_2$ solution. We evaluate the errors in the L^1-norm, the L^2-norm, the $L^1(\Omega)$-graph norm, and the H^1-norm. The results for the $\mathbb{P}_1$ approximation are displayed in Figure 1, and those for the $\mathbb{P}_2$ approximation are shown in Figure 2.

We note that the a priori error estimate (2.12) in the $L^1(\Omega)$-graph norm is fully confirmed: the slope is of order one with $\mathbb{P}_1$ finite elements and of order two with $\mathbb{P}_2$ finite elements. The error in the H^1-norm is of order $\frac{1}{2}$ for the $\mathbb{P}_1$ approximation and of order 1.7 for the $\mathbb{P}_2$ approximation. The fact that the convergence orders in the $L^1(\Omega)$-graph norm and the H^1-norm are different confirms that the method performs as expected and that it does not introduce excessive artificial cross-wind diffusion. The convergence rates of the error in the L^1-norm and the $L^2(\Omega)$-norm are slightly better than first-order in the $\mathbb{P}_1$ case and better than second-order in the $\mathbb{P}_2$ case. Note, however, that in both cases the rates are suboptimal. This result is not surprising since transport equations have no regularizing effects; that is, the Nitsche–Aubin duality argument that holds for elliptic equations does not hold here.

4.2. An elliptic operator. We now test the method on the Laplace operator. We use again $\Omega =]0,1[^2$. We solve

$$-\nabla^2 p = f, \quad p|_{\partial\Omega} = p_0,$$

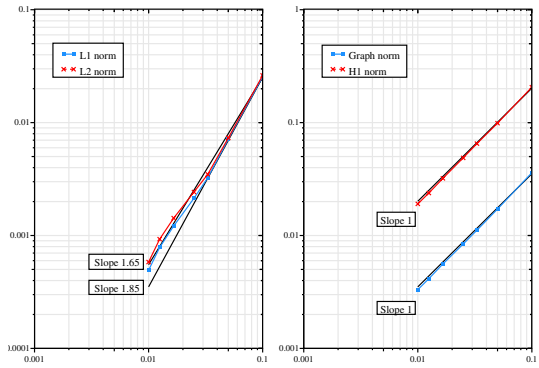


FIG. 3. *Convergence tests for $\mathbb{P}_1$ approximation of the Laplace operator. Left: Error in the L^1-norm and L^2-norm vs. the meshsize. Right: Error in the $L^1(\Omega)$-graph norm and H^1-norm vs. the meshsize.*

with the data being s.t. $p = x + 2y + \sin(2\pi x)\cos(2\pi y)$ is the exact solution.

The problem is rewritten in its first-order form (3.4) and solved in this form. We approximate the solution in $L^1(\Omega)$ using $\mathbb{P}_1$ finite elements on unstructured Delaunay meshes with $\frac{1}{10} \leq h \leq \frac{1}{100}$. For each mesh we measure the error in the $L^1(\Omega)$-norm, the $L^2(\Omega)$-norm, the $W^{1,1}(\Omega)$-norm, and the $H^1(\Omega)$-norm. The results are reported in Figure 3. We observe that the rate of convergence in the $W^{1,1}(\Omega)$-norm and in the $H^1(\Omega)$-norm are of first order; that is, they are optimal. Note that the error in the $W^{1,1}(\Omega)$-norm is almost eight times lower than that in the $H^1(\Omega)$-norm. For the $L^1(\Omega)$-norm there is not a clear rate, but we observe that the numerical results are bracketed by two lines of slope 1.65 and 1.85. Hence, the convergence is not second-order, but it is close to second-order. A similar conclusion holds for the convergence rate in the $L^2(\Omega)$-norm.

4.3. Transport equation with shock-like solutions. To illustrate the performance of the method when dealing with nonsmooth data, we consider again the 2D rectangular domain $\Omega =]0,1[^2$, and we solve the transport equation

$$\partial_x u = f, \quad u|_{x=0} = u_0, \tag{4.2}$$

with the two source terms

$$f_1(x,y) = \tfrac{1}{2\gamma}\Big[1 - \tanh^2\big(\tfrac{x-0.5}{\gamma}\big)\Big],$$
$$f_2(x,y) = \tfrac{1}{2\gamma}\Big[1 - \tanh^2\big(\tfrac{x-0.5(y+0.5)}{\gamma}\big)\Big],$$

for which the respective solutions are

$$u_1(x,y) = \tfrac{1}{2}\Big[1 + \tanh\big(\tfrac{x-0.5}{\gamma}\big)\Big],$$
$$u_2(x,y) = \tfrac{1}{2}\Big[1 + \tanh\big(\tfrac{x-0.5(y+0.5)}{\gamma}\big)\Big],$$

where $\gamma > 0$ is a small parameter. The source terms f_1 and f_2 are approximations of Dirac measures supported by the segments $x = \frac{1}{2}$ and $x - \frac{1}{2}(y - \frac{1}{2})$, respectively. These data mimic shock-like solutions.

To emphasize the capability of the method to perform well on unrefined and unstructured meshes, we show in Figure 4 the approximate solutions calculated on a

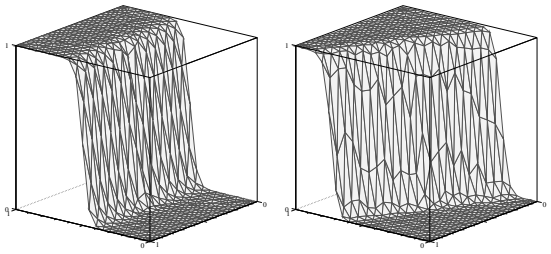


FIG. 4. *Piecewise linear L^1 approximations for test cases* (4.2). *Left: source term f_1; right: source term f_2.*

Delaunay mesh composed of 932 triangles of meshsize $h = \frac{1}{20}$. The parameter γ in the definition of the source terms is chosen to be $\gamma = h$ to guarantee that the inexact numerical integrations of the residuals are accurate enough.

Note that the solutions do not exhibit spurious over- or undershootings.

4.4. Transport equation with shear-layer-like solutions. In this section, we compare the performance of the method with that of the least-squares method on a transport equation with discontinuous boundary data.

4.4.1. 1D transport. We consider the 2D rectangular domain

$$\Omega =]0.2, 0.8[\times]0, 2[,$$

and we solve the following transport equation:

$$\partial_x u = 0, \quad u|_{x=0} = \begin{cases} 1 & \text{if } y \geq 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

The exact solution is

$$u(x, y) = \begin{cases} 1 & \text{if } y \geq 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

We perform the calculations on a Delaunay mesh with $h = \frac{1}{40}$. Due to the interpolation process, the boundary data is regularized for $0.475 \leq y \leq 0.525$.

We evaluate the least-squares solution (i.e., the $L^2(\Omega)$ approximation) and the $L^1(\Omega)$ approximation. The results are shown in Figure 5. The $L^1(\Omega)$ solution is shown at the top of the figure, and the $L^2(\Omega)$ one is shown at the bottom. For each solution we show contour lines in the left panels of the figure. Note that for both solutions there is some smearing in the transverse direction at the onset of the flow. This is due to the fact that the mesh is not aligned with the flow and the boundary data has been interpolated. In the right panels we show the projection of the graph of each solution onto the plane $x = 0$. It is clear that the least-squares solution is significantly more smeared than the $L^1(\Omega)$ one. The least-squares solution also exhibits over- and undershootings; i.e., it does not satisfy the maximum principle.

4.4.2. Curved transport. We consider the half disk

$$\Omega = \{(x, y);\ \sqrt{x^2 + y^2} < 1;\ y > 0\},$$

and let us set $\partial\Omega^- = \{-1 < x < 0;\ y = 0\}$. We want to solve the following transport problem:

$$\mathbf{v} \cdot \nabla u = 0, \quad u|_{\partial\Omega^-} = \begin{cases} 1 & \text{if } -1 < x < -0.74, \\ 0 & \text{if } -0.74 < x < 0, \end{cases} \tag{4.4}$$

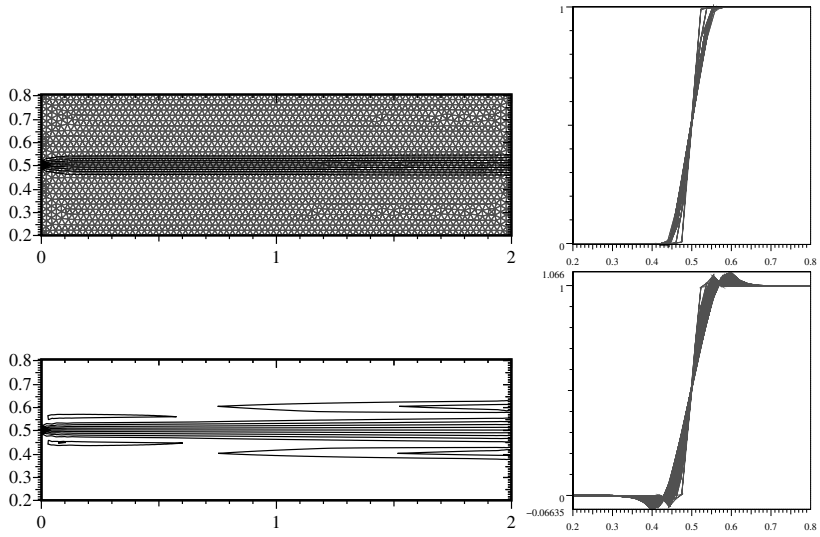


FIG. 5. *Advection equation* (4.3). *Top:* $L^1(\Omega)$ *solution and mesh; bottom:* $L^2(\Omega)$ *solution.*

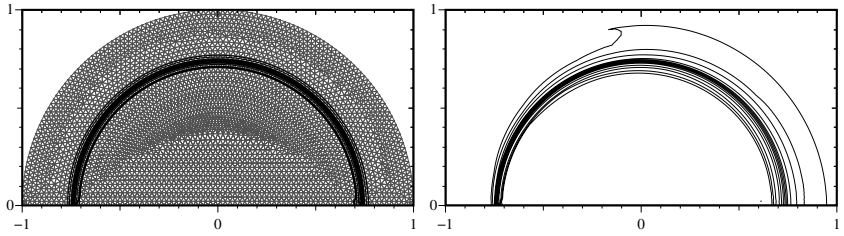


FIG. 6. *Advection equation* (4.4). *Left:* $L^1(\Omega)$ *solution with mesh; right:* $L^2(\Omega)$ *solution.*

with the curved flow field $\mathbf{v}(\mathbf{x}, \mathbf{y}) = (\sin\theta, -\cos\theta)$, where θ is the polar angle; i.e., $\theta = \arctan(y/x) \in [0, \pi[$ with the convention $\arctan(\pm\infty) = \pi/2$. The exact solution is

$$u(x, y) = \begin{cases} 1 & \text{if } \sqrt{x^2 + y^2} > 0.74, \\ 0 & \text{otherwise.} \end{cases}$$

We perform the calculations on a Delaunay mesh with $h = \frac{1}{40}$. Due to the interpolation process, the boundary condition is regularized for $-0.765 \le x \le -0.715$.

Contour lines of the $L^1(\Omega)$ and $L^2(\Omega)$ solutions are shown in Figure 6. We show about 13 contour lines for each solution. The $L^1(\Omega)$ solution is shown in the left panel of the figure, and the $L^2(\Omega)$ one is shown in the right panel. For both solutions there is some smearing in the transverse direction at the onset of the flow due to misalignment of the flow with the mesh and the interpolation of the data. It is clear, once again, that the least-squares solution is significantly more smeared than the $L^1(\Omega)$ solution and exhibits over- and undershootings.

Remark 4.1. The two test cases considered above show that for a given mesh the $L^1(\Omega)$ solution has better qualitative properties than the standard $L^2(\Omega)$ solution. In particular, discontinuities are less smeared by the L^1 approximation technique. We observe also that the $L^1(\Omega)$ solution satisfies the maximum principle. This numerical observation has yet to be fully explained mathematically.

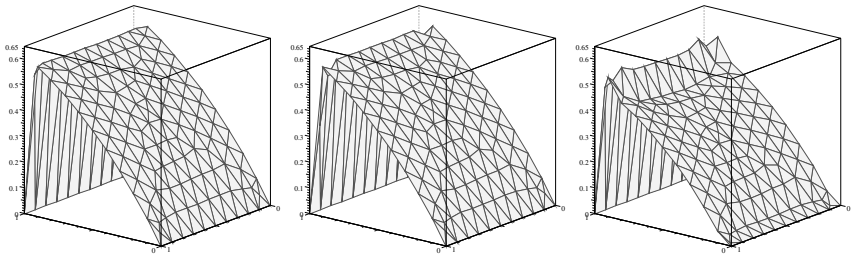


FIG. 7. *Advection equation* (4.5)*, Case* 1*,* $\nu = 0.02$*. Left:* $\mathbb{P}_1$ *Lagrange interpolate of the exact solution; center,* $L^1(\Omega)$ *solution; right,* $L^2(\Omega)$ *solution.*

Remark 4.2. Of course, if the meshes are adapted to the flow, results much sharper than those shown here can be easily obtained. We do not show these results here, for our objective in the present paper is rather to compare the performance of the $L^1(\Omega)$ and $L^2(\Omega)$ methods on arbitrary meshes than to show that each method can produce sharp results on adapted meshes.

4.5. Advection-diffusion equation. We conclude this series of tests on linear equations by solving an advection-diffusion equation in the vanishing viscosity regime.

For the sake of simplicity, we consider again the rectangular domain $\Omega =]0,1[^2$, and we denote

$$\partial\Omega_D = \{(x,y) \in \partial\Omega;\ x = 0 \text{ or } x = 1\},$$
$$\partial\Omega_N = \{(x,y) \in \partial\Omega;\ y = 0 \text{ or } y = 1\}.$$

We want to solve

$$\begin{cases} \alpha p + \boldsymbol{\beta} \cdot \nabla p + \sqrt{\nu}\, \nabla \cdot \mathbf{u} = \mathbf{f}, \\ \sqrt{\nu}\, \nabla p + \mathbf{u} = \mathbf{0}, \\ p_{|\partial\Omega_D} = p_D, \quad \mathbf{u} \cdot \mathbf{n}_{|\partial\boldsymbol{\Omega}_\mathbf{N}} = \mathbf{0}. \end{cases} \tag{4.5}$$

We set $\boldsymbol{\beta} = (1,0)$ s.t. the exact solution can be evaluated exactly.

4.5.1. Case 1. $\boldsymbol{\alpha \neq 0}$**.** We set

$$\alpha = 1, \quad f = 1, \quad p_D = 0. \tag{4.6}$$

The exact solution is

$$p(x,y) = \tfrac{1}{\alpha} + \mu^+ e^{\lambda^+ x} + \mu^- e^{\lambda^- x},$$
$$\lambda^{\pm} = \tfrac{-1 \pm \sqrt{1+4\alpha\nu}}{-2\nu}, \quad \mu^+ = -\tfrac{1}{\alpha} \tfrac{e^{\lambda^-} - 1}{e^{\lambda^-} - e^{\lambda^+}}, \quad \mu^- = -\tfrac{1}{\alpha} \tfrac{e^{\lambda^+} - 1}{e^{\lambda^+} - e^{\lambda^-}}. \tag{4.7}$$

We choose $\nu = 0.02$. We compute the $L^1(\Omega)$ and the least-squares approximations on a coarse grid with $h = \frac{1}{10}$. To the best of our knowledge, no finite element method is capable of producing a reasonable approximation to this problem with these parameters ($|\boldsymbol{\beta}|h/\nu = 5$) without resorting to some stabilization and/or nonlinear limiting technique. The results are shown in Figure 7. The $\mathbb{P}_1$ Lagrange interpolate of the exact solution is shown in the leftmost panel of the figure, the L^1 solution is in the center panel, and the least-squares solution is in the rightmost panel.

It is clear that the least-squares solution is far from the exact solution, whereas the L^1 one is a good approximation, considering the very low number of degrees of freedom used.

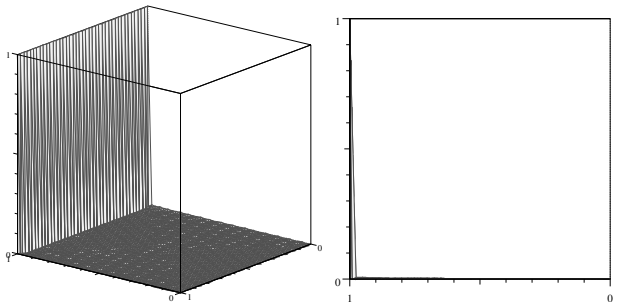


FIG. 8. *Advection equation* (4.5) *with* $\nu = 0.00125$, *Case* 2. *Left: side view of the graph of the approximate solution. Right: projection onto plane* $y = 0$ *of the graph of the approximate solution.*

4.5.2. Case 2. Now we set $\alpha = 0$. We choose the following data:

$$f = 0, \quad p_D = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x = 1. \end{cases} \tag{4.8}$$

The exact solution is

$$p(x, y) = \frac{e^{x/\nu} - 1}{e^{1/\nu} - 1}. \tag{4.9}$$

This case is frequently used in the literature to test the capability of numerical methods to solve advection-diffusion equations with dominant advection.

We set $\nu = 0.00125$, and we compute the $L^1(\Omega)$ solution on a grid of meshsize $h = \frac{1}{40}$. The result is shown in Figure 8.

It is clear that, within the capability of the mesh, the boundary layer is well-captured and the solution is not plagued by spurious oscillations.

4.6. Viscosity solutions of first-order PDEs. A striking property of the L^1 approximation technique is that it seemingly can select viscosity solutions of first-order PDEs (i.e., in the sense of Bardos, Leroux, and Nédélec [5] and Kružkov [27]).

4.6.1. Notion of viscosity solution. To illustrate this phenomenon, let Ω be a bounded domain of $\mathbb{R}^d$ with a smooth boundary. Let $\alpha > 0$, and let $\boldsymbol{\beta}$ be a vector field s.t. $\beta_i \in \mathcal{C}^1(\overline{\Omega})$, $1 \le i \le d$. Let u_0 be a smooth function on $\partial\Omega$, say $u_0 \in \mathcal{C}^2(\partial\Omega)$, and let $f \in W^{1,1}(\Omega)$. Following Bardos, Leroux, and Nédélec, [5], we say that u is a viscosity solution of

$$\alpha u + \nabla \cdot (\boldsymbol{\beta} u) = f, \quad u|_{\partial\Omega} = u_0, \tag{4.10}$$

if $u \in \mathrm{BV}(\Omega)$, u solves the PDE, and u satisfies the boundary condition in the following sense:

$$\int_{\partial\Omega} (\boldsymbol{\beta} \cdot \mathbf{n})(\mathbf{u} - \mathbf{k})(\mathrm{sg}(\mathbf{u} - \mathbf{k}) - \mathrm{sg}(\mathbf{u_0} - \mathbf{k})) \ge \mathbf{0} \quad \forall \mathbf{k} \in \mathbb{R}, \tag{4.11}$$

where $\mathrm{sg}(t)$ is the sign of t if $t \neq 0$ and $\mathrm{sg}(0) = 0$. In the present linear case, the boundary condition amounts to enforcing $u = u_0$ on $\partial\Omega^- = \{\mathbf{x} \in \partial\boldsymbol{\Omega} \,|\, \mathbf{x} \cdot \mathbf{n} < \mathbf{0}\}$. The interest of (4.11) is that it generalizes easily to nonlinear equations, whereas the notion of inflow and outflow boundary condition does not, since for nonlinear problems the inflow or outflow status of the boundary may depend on the solution

itself. For instance, if the linear term $\boldsymbol{\beta} u$ is replaced by $F(u, \mathbf{x})$, then the boundary condition $u|_{\partial\Omega} = u_0$ has to be understood in the sense

$$\int_{\partial\Omega} (F(u, \mathbf{x}) - \mathbf{F}(\mathbf{k}, \mathbf{x})) \cdot \mathbf{n}(\mathrm{sg}(\mathbf{u} - \mathbf{k}) - \mathrm{sg}(\mathbf{u_0} - \mathbf{k})) \geq \mathbf{0} \quad \forall \mathbf{k} \in \mathbb{R}. \tag{4.12}$$

Using arguments similar to those in [5] and [4], it is possible to prove that (4.10) has a unique viscosity solution, provided α is large enough. The bulk of the argument consists of proving that the solution to the following problem,

$$\alpha u_\epsilon + \nabla \cdot (\boldsymbol{\beta} u_\epsilon) - \epsilon \nabla^2 u_\epsilon = f, \quad u_\epsilon|_{\partial\Omega} = u_0, \tag{4.13}$$

converges in $\mathrm{BV}(\Omega)$ and the limit is the so-called viscosity solution; i.e., the limit satisfies the PDE in (4.10) and (4.11).

Despite the appearance, the problem (4.10) is not purely formal. It is typically this type of problem that arises when one tries to approximate (4.13) on meshes which are not refined enough. More precisely, when $\epsilon/h^2 \ll 1/h$ the second-order term in (4.13) is completely dominated by the first-order one, and solving (4.13) numerically amounts to trying to solve (4.10), where the boundary condition is understood in the classical sense instead of (4.11).

Once this point is understood, it becomes clear that the least-squares technique cannot work properly on the advection-diffusion equation (4.13) if the mesh is not refined enough.

PROPOSITION 4.1. *The H^1-conformal approximate least-squares solution to the linear problem* (4.13) *(written in mixed form) may not converge to the viscosity solution as $h \to 0$.*

Proof. As we want to build a counterexample, let us restrict ourselves to the 1D viewpoint, and let us take $\Omega =]0, 1[$, $\boldsymbol{\beta} = 1$, $u_0 = 0$, $\alpha = 1$, and $f = 1$. Let $E_h \subset H_0^1(\Omega)$ be a finite-dimensional finite element space. Thanks to crude a priori estimates in $L^2(\Omega)$ and standard inverse inequalities, it is clear that the least-squares approximation to (4.15) converges to the solution of the following problem as $\epsilon \to 0$:

$$\int_0^1 (u_h + u_h')(v_h + v_h') = \int_0^1 (v_h + v_h') \quad \forall v_h \in E_h.$$

Since test functions in E_h satisfy $v_h(0) = v_h(1) = 0$, we obtain

$$\int_0^1 (u_h + u_h')(v_h + v_h') = \int_0^1 v_h \quad \forall v_h \in E_h.$$

Then it is clear that, when $h \to 0$, the solution to the above problem converges in $H_0^1(]0, 1[)$ to the solution of the following PDE:

$$w - w'' = 1, \quad w(0) = w(1) = 0,$$

which is obviously different from the viscosity solution which solves

$$u + u' = 1, \quad u(0) = 0.$$

This completes the proof. ∎

This example shows that for a given mesh the L^2-based least-squares approximation technique does not select the right limit of (4.15) as $\epsilon \to 0$. The situation is quite different in $L^1(\Omega)$. For reasons not yet clear, numerical tests, reported in the next section, show that the solution that minimizes the L^1 distance is a reasonable approximation of the viscosity solution.

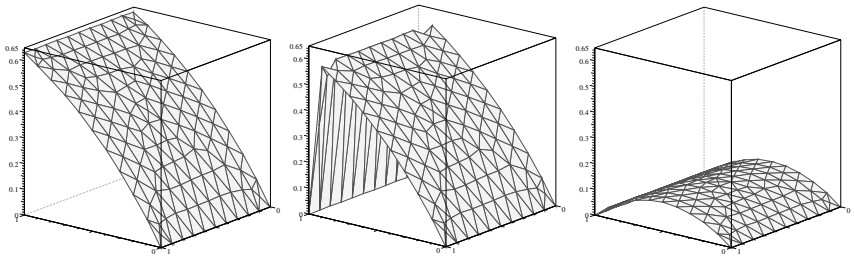


FIG. 9. *Viscosity solution to* (4.14). *Left:* $\mathbb{P}_1$ *Lagrange interpolate of the exact solution; center,* $L^1(\Omega)$ *solution; right,* $L^2(\Omega)$ *solution.*

4.6.2. Numerical experiments. let us consider the 2D rectangular domain $\Omega =]0,1[^2$ with $\partial\Omega_D = \{x = 0\} \cup \{x = 1\}$ and $\partial\Omega_N = \{y = 0\} \cup \{y = 1\}$. We want to solve the following scalar problem:

$$\alpha u + \partial_x u = f, \quad u|_{\partial\Omega_D} = u_0. \tag{4.14}$$

Of course, this problem is not well-posed in the standard sense since the outflow boundary condition is overspecified, but this problem is meaningful in the viscosity sense as defined above. Let E_h be a H^1-conformal finite element space s.t. for all v_h in E_h, $v_h|_{\partial\Omega_D} = 0$. It is clear that approximating the regularized problem

$$\alpha u_\epsilon + \partial_x u_\epsilon - \epsilon\nabla^2 u_\epsilon = f, \quad u_\epsilon|_{\partial\Omega_D} = u_0, \quad \partial_y u_\epsilon|_{\partial\Omega_N} = 0, \tag{4.15}$$

and taking the limit $\epsilon \to 0$, h being fixed, is equivalent to approximating (4.14) in E_h (recall that in E_h the Dirichlet boundary condition is enforced in the standard sense).

Let us set

$$\alpha = 1, \quad f = 1, \quad u_0 = 0. \tag{4.16}$$

We solve (4.14) in $L^1(\Omega)$ and in $L^2(\Omega)$, respectively, using continuous $\mathbb{P}_1$ finite elements and (3.8). To emphasize the capabilities of the L^1 approximation technique, we restrict ourselves to a very coarse mesh, $h = 1/10$. The results are shown in Figure 9.

In the left panel we show the $\mathbb{P}_1$ Lagrange interpolate of the viscosity solution, in the center panel we show the L^1 solution, and in the right panel we show the least-squares solution. Considering the mesh used, the $L^1(\Omega)$ approximation is a reasonable approximation, whereas the least-squares solution is completely wrong (thus confirming Proposition 4.1). Convergence tests, not reported here, show that the $L^1(\Omega)$ approximate solution converges in the $L^1(\Omega)$-norm to the viscosity solution as $h \to 0$.

Contrary to what it seems, the two horn-like spikes observable on the graph of the L^1 solution are not overshootings. These are perspective effects induced by the fact that the two corresponding $\mathbb{P}_1$-nodes are not aligned with the others. This is made clear by looking at the xz-projection of the graph of the L^1 solution shown in Figure 10.

Given that the least-squares method, together with its many variants, is a central part for the stabilization of the Galerkin technique (see, e.g., [16, 25, 26]), the above example gives new reasons why the Galerkin least-squares method cannot generally cope properly with shocks and boundary layers without the help of shock-capturing terms [26, 22].

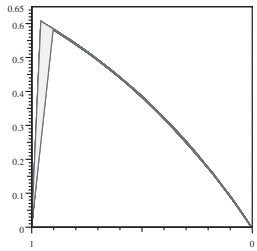


FIG. 10. *Projection in the xz-plane of the graph of the L^1 solutions in Figure 9.*

4.7. The Burgers equation. To finish this series of tests we propose to solve the Burgers-like equation

$$\nabla\cdot\left(\left(\boldsymbol{\beta}+\frac{u}{2}\frac{\mathbf{x}-\mathbf{x_0}}{\|\mathbf{x}-\mathbf{x_0}\|_{\ell^2}^{\mathbf{2}}}\right)u\right)=0,\quad u|_{\partial\Omega_1}=1,\quad u|_{\partial\Omega_0}=0, \tag{4.17}$$

in the following 2D domain:

$$\Omega=]0,1[^2\setminus\{\|\mathbf{x_0}-\mathbf{x}\|_{\ell^2}\leq\mathbf{0.2}\},$$

where $\partial\Omega_1=\{\|\mathbf{x_0}-\mathbf{x}\|_{\ell^2}=\mathbf{0.2}\}$, $\partial\Omega_0=\{x=0\}$, $\mathbf{x_0}=(\mathbf{0.5},\mathbf{0.5})$, and $\boldsymbol{\beta}=(v_0,0)$ with $v_0\geq 0$. This form of the Burgers equation retains the simplicity of its 1D counterpart and allows for more realistic 2D numerical tests.

We select an entropy solution to this problem by taking the limit as $t\to+\infty$ of the solution to the time-dependent version of (4.17), using as initial data the solution to the following problem:

$$\nabla^2 u_0=0,\quad u_0|_{\partial\Omega_1}=1,\quad u_0|_{\partial\Omega_0}=0,\quad \partial_n u_0|_{\partial\Omega_N}=0,$$

where $\partial\Omega_N$ is the complement of $\partial\Omega_0\cup\partial\Omega_1$.

The L^1 approximation is computed iteratively by solving the time-dependent problem, using the implicit Euler time-stepping. We test two configurations, $v_0=6$ and $v_0=4/3$, henceforth referred to as case 1 and case 2, respectively. To assess the accuracy of the method and its sensitivity to mesh refinement, we first do the computation on a uniform grid, $h=1/40$; then we redo it on a somewhat adapted grid with $1/10\leq h\leq 1/100$.

The results for case 1 are shown in Figure 11. We show the contour lines of the solution. Essentially, the exact solution consists of two regions where u is either equal to 1 or equal to 0, and these two regions are separated by a shock. Note that there is no shock in the upstream region and the $u=0$ solution reaches the cylinder. This means that the boundary condition $u|_{\partial\Omega_1}=1$ is satisfied in the entropy sense as defined in (4.12). We observe that the numerical solution satisfies the maximum principle and the contour lines are concentrated in the shock region. In this case the shock spreads over 2 to 3 elements; the reason for this is that the shock is almost aligned with the flow. This phenomenon is comparable to the smearing observed in the transport problem described in section 4.4. Smearing of oblique shocks is a common feature of techniques dealing with shocks. Note that the position of the shock does not change significantly as the mesh is refined, thus demonstrating that the coarse uniform mesh predicts quite well the position of the shock in question.

The contour lines of the numerical solution to case 2 are shown in Figure 12. Once more, the numerical solution is not plagued by spurious over- or undershootings. The

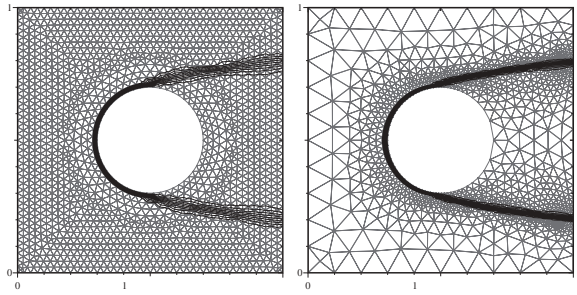


FIG. 11. *The Burgers equation with* $v_0 = 6$. *Contour lines of the solution. Left: uniform mesh* $h = 1/40$*; right: nonuniform mesh* $1/10 \leq h \leq 1/100$.

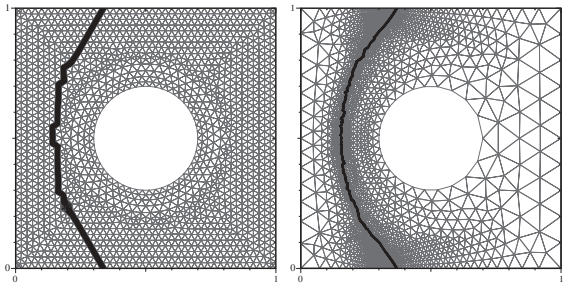


FIG. 12. *The Burgers equation with* $v_0 = 4/3$. *Contour lines of the solution. Left: uniform mesh* $h = 1/40$*; right: nonuniform mesh* $1/10 \leq h \leq 1/100$.

shock is almost perpendicular to the incoming flow; as a result, there is no smearing. The shock is a.e. contained within one element only.

The two above examples show that the L^1 technique is capable of selecting the entropy solution of Burgers-like equations. Moreover, the L^1 solution seems to satisfy a maximum principle. These two numerical observations are still to be understood and possibly proved mathematically.

5. Concluding remarks. One of the objectives driving the present work is to show that for solving first-order PDEs supplemented with nonsmooth data the ongoing debate pitting methods based on continuous interpolation against those based on discontinuous ones (e.g., H^1-conformal Galerkin vs. discontinuous Galerkin) is possibly pointless, insofar as the analysis is usually restricted to the $L^2(\Omega)$ setting. In the present paper, we have tried to promote the idea that working in a functional setting that provides for the right stability properties is as important as debating on the nature of the approximation (interpolation) space. Once the required stability property is guaranteed by the functional setting, the only requirement set for the discrete space is that it possesses good interpolation properties. As an illustration of this point of view, we have shown that, when working in $L^1(\Omega)$, the often despised continuous $\mathbb{P}_1$ finite element is capable of accurately approximating shocks, shear-layers, and boundary layers.

For reasons not yet completely clear, it seems that the $L^1(\Omega)$ approximation technique is capable of selecting viscosity solutions of first-order PDEs [5, 27] without resorting to any artificial artifact and/or tuning parameter, though this conjecture has yet to be substantiated mathematically.

All that has been said in this paper can be extended to spaces that are more exotic than the L^p's. For instance, we could consider Besov spaces or Radon measures, provided the corresponding norms can be computed efficiently in the discrete space E_h. While these spaces may provide better interpolation or approximation properties, they would require the use of wavelet bases or other hierarchical approximation spaces (for researches going in this direction, we refer the reader to, e.g., [7, 8]).

The generalization of the present work to evolution equations and conservation laws is under investigation and will be reported in a forthcoming paper.

Acknowledgments. The author acknowledges discussions with P. Azerad, J.T. Oden, B. Perthame, and L. Quartapelle that greatly improved the content of the present paper.

REFERENCES

[1] P. Azerad and G. Pousin, *Inégalité de Poincaré courbe pour le traitement variationnel de l'équation de transport*, C. R. Acad. Sci. Paris Sér. I Math, 322 (1996), pp. 721–727.

[2] A. K. Aziz, R. B. Kellogg, and A. B. Stephens, *Least-squares methods for elliptic systems*, Math. Comp., 44 (1985), pp. 53–70.

[3] C. Baiocchi, F. Brezzi, and L. P. Franca, *Virtual bubbles and Galerkin-Least-Squares type methods (GaLS)*, Comput. Methods Appl. Mech. Engrg., 105 (1993), pp. 125–141.

[4] C. Bardos, D. Brézis, and H. Brézis, *Perturbations singulières et prolongements maximaux d'opérateurs positifs*, Arch. Rational Mech. Anal., 53 (1973), pp. 69–100.

[5] C. Bardos, A. Y. Leroux, and J.-C. Nédélec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[6] P. Benilan, H. Brezis, and M. G. Crandall, *A semilinear equation in $L^1(R^N)$*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 2 (1975), pp. 523–555.

[7] J. H. Bramble, J. E. Pasciak, and P. S. Vassilevski, *Computational scales of Sobolev norms with application to preconditioning*, Math. Comp., 69 (2000), pp. 463–480.

[8] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[9] J. H. Bramble and A. H. Schatz, *Rayleigh-Ritz-Galerkin-methods for Dirichlet's problem using subspaces without boundary conditions*, Comm. Pure Appl. Math., 23 (1970), pp. 653–675.

[10] J. H. Bramble and A. H. Schatz, *Least squares for 2mth order elliptic boundary-value problems*, Math. Comp., 25 (1971), pp. 1–32.

[11] H. Brezis and W. A. Strauss, *Semi-linear second-order elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 565–590.

[12] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Applied Mathematics Series for the Master's Degree, Masson, Paris, 1983.

[13] F. Brezzi, M. O. Bristeau, L. P. Franca, M. Mallet, and G. Rogé, *A relationship between stabilized finite element methods and the Galerkin method with bubble functions*, Comput. Methods Appl. Mech. Engrg., 96 (1992), pp. 117–129.

[14] F. Brezzi, L. P. Franca, T. J. R. Hughes, and A. Russo, $b = \int g$, Comput. Methods Appl. Mech. Engrg., 145 (1997), pp. 329–364.

[15] F. Brezzi, P. Houston, D. Marini, and E. Süli, *Modeling subgrid viscosity for advection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 1601–1610.

[16] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov–Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[17] B. Cockburn, G. E. Karniadakis, and C. W. Shu, *Discontinuous Galerkin Methods - Theory, Computation and Applications*, Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin, 2000.

[18] A. Ern and J.-L. Guermond, *Elements finis: Théorie, applications, mise en œuvre*, Math. Appl. (Berlin) 36, Springer-Verlag, Paris, 2002.

[19] K. O. Friedrichs, *Symmetric positive linear differential equations*, Comm. Pure Appl. Math., 11 (1958), pp. 333–418.

[20] J.-L. Guermond, *Stabilisation par viscosité de sous-maille pour l'approximation de Galerkin*

des opérateurs linéaires monotones, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 617–622.

[21] J.-L. Guermond, *Stabilization of Galerkin approximations of transport equations by subgrid modeling*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1293–1316.

[22] T. J. R. Hughes and M. Mallet, *A new finite element formulation for computational fluid dynamics.* IV: *A discontinuity-capturing operator for multidimensional advective-diffusive systems*, Comput. Methods Appl. Mech. Engrg., 58 (1986), pp. 329–336.

[23] B.-N. Jiang, *Non-oscillatory and non-diffusive solution of convection problems by the iteratively reweighted least-squares finite element method*, J. Comput. Phys., 105 (1993), pp. 108–121.

[24] B.-N. Jiang, *The Least-Squares Finite Element Method*, Scientific Computation, Springer-Verlag, Berlin, 1998.

[25] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic equations*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[26] C. Johnson and A. Szepessy, *On the convergence of a finite element method for a nonlinear hyperbolic conservation law*, Math. Comp., 49 (1987), pp. 427–444.

[27] S. N. Kružkov, *First order quasilinear equations with several independent variables*, Mat. Sb. (N.S.), 81 (1970), pp. 228–255.

[28] J. E. Lavery, *Nonoscillatory solution of the steady-state inviscid Burgers' equation by mathematical programming*, J. Comput. Phys., 79 (1988), pp. 436–448.

[29] J. E. Lavery, *Solution of steady-state one-dimensional conservation laws by mathematical programming*, SIAM J. Numer. Anal., 26 (1989), pp. 1081–1089.

[30] J. E. Lavery, *Solution of steady-state, two-dimensional conservation laws by mathematical programming*, SIAM J. Numer. Anal., 28 (1991), pp. 141–155.

[31] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boors, ed., Academic Press, New York, 1974, pp. 89–123.

[32] R. B. Lowrie and P. L. Roe, *On the numerical solution of conservation laws by minimizing residuals*, J. Comput. Phys., 113 (1994), pp. 304–308.

[33] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[34] R. E. Showalter, *Monotone Operators in Banach Spaces and Nonlinear Partial Differential Equations*, Math. Surveys Monogr. 49, AMS, Providence, RI, 1996.

[35] E. Tadmor, *Convergence of spectral methods for nonlinear conservation laws*, SIAM J. Numer. Anal., 26 (1989), pp. 30–44.

[36] K. Yosida, *Functional Analysis*, Classics in Mathematics, Springer-Verlag, Berlin, 1995.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 738–772

FINITE ELEMENT APPROXIMATION OF A PHASE FIELD MODEL FOR VOID ELECTROMIGRATION*

JOHN W. BARRETT†, ROBERT NÜRNBERG†, AND VANESSA STYLES‡

Abstract. We consider a fully practical finite element approximation of the nonlinear degenerate parabolic system

$$\gamma\,\tfrac{\partial u}{\partial t} - \nabla.(\,b(u)\,\nabla[w+\alpha\,\phi]\,) = 0\,, \quad w = -\gamma\,\Delta u + \gamma^{-1}\,\Psi'(u)\,, \quad \nabla.(\,c(u)\,\nabla\phi) = 0$$

subject to an initial condition $u^0(\cdot) \in [-1,1]$ on u and flux boundary conditions on all three equations. Here $\gamma \in \mathbb{R}_{>0}$, $\alpha \in \mathbb{R}_{\geq 0}$, Ψ is a nonsmooth double well potential, and $c(u) := 1+u$, $b(u) := 1-u^2$ are degenerate coefficients. The degeneracy in b restricts $u(\cdot,\cdot) \in [-1,1]$. The above, in the limit $\gamma \to 0$, models the evolution of voids by surface diffusion in an electrically conducting solid. In addition to showing stability bounds for our approximation, we prove convergence, and hence existence of a solution to this nonlinear degenerate parabolic system in two space dimensions. Furthermore, an iterative scheme for solving the resulting nonlinear discrete system is introduced and analyzed. Finally, some numerical experiments are presented.

Key words. void electromigration, surface diffusion, phase field model, diffuse interface model, degenerate Cahn–Hilliard equation, fourth order degenerate parabolic system, finite elements, convergence analysis

AMS subject classifications. 65M60, 65M12, 35K55, 35K65, 35K35, 35Q60, 82C26, 65M50

DOI. 10.1137/S0036142902413421

1. Introduction.

Interconnect lines on microelectronic circuits usually contain small voids or cracks due to the extreme thermal stress that they are exposed to when cooled to room temperature during the production process. The applied electric field and interfacial tension cause surface diffusion; that is, atoms diffuse from one part of the void boundary to another. The void effectively "drifts" through the conductor, changing its shape as it does so. If the void becomes large enough to sever a line, it causes an open circuit. As producers try to reduce the dimensions of microchips further and further, these circuit failures become more and more frequent. Hence there is great interest in understanding the mechanism that leads to this phenomenon known as void electromigration. For further details, see, e.g., [28, 12] and the references therein. As the height of interconnect lines are extremely thin compared to the dimensions of the base, voids fully penetrate in this vertical direction. Hence a two dimensional model in the plane suffices.

Let $\Omega := (-L_1, L_1) \times (-L_2, L_2)$ be the rectangular domain in $\mathbb{R}^2$, representing the interconnect line, with boundary $\partial\Omega$. At any time $t \in [0,T]$, let the region occupied by the void be $\Omega^-(t) \subset\subset \Omega$ with boundary $\Gamma(t)$. Then the electric field in the conducting region, $\Omega^+(t) := \Omega \setminus \overline{\Omega^-(t)}$, is $E = -\nabla\phi$, where the potential ϕ at any time $t \in [0,T]$

*Received by the editors August 22, 2002; accepted for publication (in revised form) October 29, 2003; published electronically June 4, 2004.

http://www.siam.org/journals/sinum/42-2/41342.html

†Department of Mathematics, Imperial College, London, SW7 2AZ, UK (j.barrett@ic.ac.uk). The research of the second author was supported by the EPSRC and by the DAAD through a *Doktorandenstipendium*.

‡Centre for Mathematical Analysis and Its Applications, University of Sussex, Brighton, BN1 9QH, UK (v.styles@sussex.ac.uk). The research of this author was supported by a Leverhulme 2000 Fellowship.

satisfies

$$\Delta\phi = 0 \quad \text{in } \Omega^+(t), \qquad \tfrac{\partial\phi}{\partial\nu_{\Gamma(t)}} = 0 \quad \text{on } \Gamma(t)\,, \tag{1.1a}$$

$$\tfrac{\partial\phi}{\partial\nu} = 0 \quad \text{on } \partial_1\Omega, \qquad 2\,\tfrac{\partial\phi}{\partial\nu} + \phi = g^{\pm} := x_1 \pm 2 \quad \text{on } \partial_2^{\pm}\Omega\,, \tag{1.1b}$$

$\nu_{\Gamma(t)}$ being the unit normal to $\Gamma(t)$ pointing into $\Omega^-(t)$. In the above $\partial\Omega = \partial_1\Omega \cup \partial_2\Omega$, where $\partial_1\Omega \cap \partial_2\Omega = \emptyset$ and

$$\partial_2\Omega = \partial_2^-\Omega \cup \partial_2^+\Omega \quad \text{with} \quad \partial_2^{\pm}\Omega := \{\pm L_1\} \times [-L_2, L_2]\,,$$

and ν is the outward unit normal to $\partial\Omega$; see the sketch below. Hence $\partial_1\Omega$ is the insulated boundary of Ω, while the Robin boundary conditions on the ends $\partial_2^{\pm}\Omega$ model a uniform parallel electric field, $\phi \approx x_1$, as $L_1 \to \infty$. We note that one could alternatively model this with either (a) the Dirichlet condition $\phi = x_1$ or (b) the Neumann condition $\frac{\partial\phi}{\partial\nu} = \pm 1$ on $\partial_2^{\pm}\Omega$. However, in deriving energy bounds it is convenient to have a Robin condition; see (1.8) below. The motion of the void boundary, $\Gamma(t)$, then evolves according to the law

$$V = -\Delta_s\,[\,\alpha_1\,\kappa - \alpha_2\,\phi\,] \qquad \text{on } \Gamma(t)\,, \tag{1.2}$$

where V is the velocity of $\Gamma(t)$ in the direction of $\nu_{\Gamma(t)}$, Δ_s is the surface Laplacian $\equiv \frac{\partial^2}{\partial s^2}$, s being arc-length, and κ is the curvature of $\Gamma(t)$ (positive where $\Omega^-(t)$ is convex). Here $\alpha_1 \in \mathbb{R}_{>0}$ and, without loss of generality (see, e.g., [12, p. 101]), $\alpha_2 \in \mathbb{R}_{\geq 0}$ are given parameters depending on the conductor. The first term on the right-hand side of (1.2) is surface diffusion due to interfacial tension, which models atoms moving around the boundary to positions of large curvature, whereas the second term is surface diffusion due to the electric field. The void electromigration model is then the coupled system (1.1a), (1.1b) and (1.2).

If $\alpha_2 = 0$, then a local existence result for the motion (1.2) can be found in [14]. Moreover, they showed that a global solution exists if the initial curve, $\Gamma(0)$, is close to a circle and that it converges to a circle. For $\alpha_2 \geq 0$, the motion preserves the area enclosed by the closed curve $\Gamma(t)$ since

$$\tfrac{d}{dt}\,[\underline{m}(\Omega^-(t))] = -\int_{\Gamma(t)} V\,\mathrm{d}s = 0\,,$$

where $\underline{m}(D)$ is the measure of a domain D. In addition, for $\alpha_2 = 0$ this motion decreases the length of the interface since

$$\tfrac{d}{dt}\,[\underline{m}(\Gamma(t))] = -\int_{\Gamma(t)} V\,\kappa\,\mathrm{d}s = -\alpha_1 \int_{\Gamma(t)} [\tfrac{\partial\kappa}{\partial s}]^2\,\mathrm{d}s \leq 0\,.$$

A circular void moving at a constant speed is a solution of (1.1a), (1.1b) and (1.2) in the case of an infinite conductor: that is, for any $\alpha_i \in \mathbb{R}_{\geq 0}$, $R \in \mathbb{R}_{>0}$, and $z = (z_1, z_2) \in \mathbb{R}^2$,

$$\Gamma(t) := \{x \in \mathbb{R}^2 : (x_1 - z_1(t))^2 + (x_2 - z_2)^2 = R^2\}, \qquad z_1(t) := z_1 + \tfrac{2\,\alpha_2}{R}\,t, \tag{1.3a}$$

where the corresponding electric potential

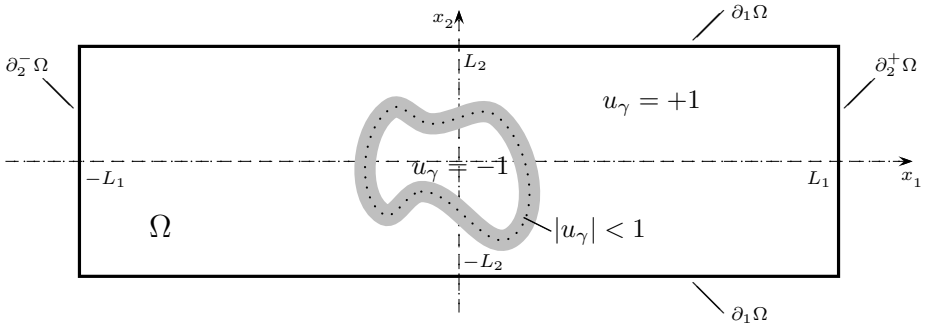
$$\phi(x,t) = [x_1 - z_1(t)]\left(1 + \tfrac{R^2}{(x_1 - z_1(t))^2 + (x_2 - z_2)^2}\right) \tag{1.3b}$$

solves (1.1a), (1.1b) and (1.2) with

(1.3c)
Ω in (1.1a) replaced by $\mathbb{R}^2$ and (1.1b) replaced by $\nabla\phi \to (1,0)^T$ as $|x| \to \infty$.

Observe that (1.2) reduces to $V = -\frac{2\,\alpha_2}{R^2}\,[x_1 - z_1(t)]$ on $\Gamma(t)$. The explicit solution (1.3a), (1.3b) was first noted in [18].

A number of authors (see, e.g., [9, 19, 28]) have considered a direct finite element approximation of (1.1a), (1.1b) and (1.2). This involves the explicit tracking of the approximate void boundary, the approximation of surface derivatives on it, and the remeshing of the approximation to $\Omega^+(t)$ in order to approximate ϕ. This direct approach breaks down at singularities, where there is a change in topology of the interface due to either the breakup or the coalescence of voids. In this paper we will consider a diffuse interface/phase field model of the original "sharp interface" void electromigration model (1.1a), (1.1b) and (1.2). The advantage of a phase field method is that the interface is implicitly embedded and is not tracked explicitly. Moreover, this approach can cope with the voids changing topology. One should also note that the phase field approach carries across unchanged to the three dimensional problem.



We introduce the interfacial parameter $\gamma \in \mathbb{R}_{>0}$ and the conserved order parameter $u_\gamma(\cdot,t) \in \mathcal{K} := [-1,1] \subset \mathbb{R}$, where at any time $t \in [0,T]$ $u_\gamma(\cdot,t) = -1$ denotes the void and $u_\gamma(\cdot,t) = +1$ denotes the conductor, while the void boundary is approximated by the $u_\gamma(\cdot,t) = 0$ contour line inside the $|u_\gamma(\cdot,t)| < 1$ interfacial region. We introduce also the chemical potential $w_\gamma(\cdot,t)$ and the electric potential $\phi_\gamma(\cdot,t)$. The sharp interface model, (1.1a), (1.1b) and (1.2), is then approximated by the following nonlinear degenerate parabolic system:

(P_γ) Find functions $u_\gamma : \Omega \times [0,T] \to \mathcal{K}$ and w_γ, $\phi_\gamma : \Omega \times [0,T] \to \mathbb{R}$ such that

(1.4a) $\gamma\,\frac{\partial u_\gamma}{\partial t} - \nabla.(\,b(u_\gamma)\,\nabla[w_\gamma + \alpha\,\phi_\gamma]\,) = 0 \quad \text{in } \Omega_T := \Omega \times (0,T],$

(1.4b) $w_\gamma = -\gamma\,\Delta u_\gamma + \gamma^{-1}\,\Psi'(u_\gamma) \quad \text{in } \Omega_T, \text{ where } |u_\gamma| < 1,$

(1.4c) $u_\gamma(x,0) = u_\gamma^0(x) \in \mathcal{K} \qquad \forall\, x \in \Omega,$

(1.4d) $\frac{\partial u_\gamma}{\partial \nu} = b(u_\gamma)\,\frac{\partial[w_\gamma + \alpha\,\phi_\gamma]}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0,T]\,,$

(1.4e) $\nabla.(\,c(u_\gamma)\,\nabla\phi_\gamma) = 0 \quad \text{in } \Omega_T,$

(1.4f) $c(u_\gamma)\,\frac{\partial \phi_\gamma}{\partial \nu} = 0 \text{ on } \partial_1\Omega \times (0,T],\ c(u_\gamma)\,\frac{\partial \phi_\gamma}{\partial \nu} + \phi_\gamma = g^\pm \text{ on } \partial_2^\pm\Omega \times (0,T]\,.$

In (1.4a)–(1.4f), $\gamma > 0$ and $\alpha \geq 0$ are given constants and

$$\Psi(s) := \begin{cases} \frac{1}{2}\left(1 - s^2\right) & \text{if } s \in \mathcal{K}, \\ \infty & \text{if } s \notin \mathcal{K} \end{cases} \tag{1.5}$$

is an obstacle-free energy which restricts $u_\gamma(\cdot,\cdot) \in \mathcal{K}$. In addition, we define the degenerate diffusion coefficients

$$c(s) := 1 + s, \qquad b(s) := 1 - s^2 = c(s)\, c(-s) \qquad \forall\ s \in \mathcal{K}. \tag{1.6}$$

If $\alpha = 0$, then (1.4a)–(1.4d) collapses to (Q_γ), the degenerate Cahn–Hilliard equation. Existence of a solution to (Q_γ), which is a fourth order degenerate parabolic equation for u_γ as $b(u_\gamma)$ can take on zero values, can be found in [13]. Degenerate parabolic equations of higher order exhibit some new characteristic features which are fundamentally different from those for second order degenerate parabolic equations. The key point is that there is no maximum or comparison principle for parabolic equations of higher order. This drastically complicates the analysis since a lot of results which are known for second order equations are proven with the help of comparison techniques. Related to this is the fact that there is no uniqueness result known for (Q_γ). Although there is no comparison principle, one of the main features of (Q_γ) is the fact that one can show existence of a solution with $|u_\gamma| \leq 1$ if given initial data $|u_\gamma^0| \leq 1$. This is in contrast to linear parabolic equations of fourth order.

Moreover, it is shown in [10] by using the techniques of formal asymptotic expansions that the zero level sets of u_γ, the solution to (Q_γ) for a fixed $\gamma > 0$, converge as $\gamma \to 0$ to an interface, $\Gamma(t)$, evolving according to the geometric motion (1.2) with $\alpha_1 = \frac{\pi^2}{16}$ and $\alpha_2 = 0$. Furthermore, on the zero level sets of u_γ the chemical potential w_γ tends to $-\frac{\pi}{4}\kappa$, where κ is the curvature of the limiting interface $\Gamma(t)$. This limiting motion of surface diffusion is a purely local geometric motion for the interface and is in contrast to the nonlocal Mullins–Sekerka motion, which is the limiting motion of (Q_γ) with a constant diffusion term b in place of the degenerate b, (1.6). It is a straightforward matter to extend the technique of formal asymptotic expansions in [10] for (Q_γ) to (P_γ), and one obtains that the zero level sets of u_γ, the solution to (P_γ) for a fixed $\gamma > 0$, converge as $\gamma \to 0$ to an interface, $\Gamma(t)$, evolving according to the modified motion (1.2) with $\alpha_1 = \frac{\pi^2}{16}$ and $\alpha_2 = \frac{\pi\,\alpha}{4}$; see [23] for details. Hence the limiting sharp interface motion of (P_γ) is the void electromigration model, (1.1a), (1.1b) and (1.2), for a suitable choice of α and on rescaling time. Note that (1.4e), (1.4f) with the choice (1.6) is the natural diffuse interface approximation of (1.1a), (1.1b). We remark that for both (P_γ) and (Q_γ) the formal asymptotics yield that the interface thickness is approximately $\gamma\,\pi$.

A phase field approximation of (1.1a), (1.1b) and (1.2), which is very similar to (P_γ), has been considered in [8]. The only difference is in the choice of mobility $b(s) = b_0 \in \mathbb{R}_{>0}$ for $|s| < 1$ and $b(s) = 0$ otherwise. An alternative phase field approximation of (1.1a), (1.1b) and (1.2), where the diffusion coefficient b is nondegenerate and depends on $|\nabla u|^2$ as opposed to u itself, is considered in [22, 21]. Finally, an alternative fixed mesh approximation of (1.1a), (1.1b) and (1.2) is considered in [20] and in [25]. Both are based on a local level set approach to approximate (1.2) and, for $\alpha > 0$, a modified immersed interface method for approximating (1.1a), (1.1b). The former requires approximating fourth order partial differential equations for the scalar level set variable.

We should stress that there is no analysis of any of the above numerical approaches to (1.1a), (1.1b) and (1.2). In this paper we introduce and analyze a finite element

approximation of the degenerate phase field model (P_γ), which approximates the sharp interface motion (1.1a), (1.1b) and (1.2) in the limit $\gamma \to 0$. There is very little work on the numerical analysis of degenerate parabolic equations of fourth order: for work on the thin film equation see [2, 29, 17]; for thin film flows in the presence of surfactants see [5]; and for work on degenerate Cahn–Hilliard systems see [3, 4, 1]. In all of these papers, although stability bounds were proved in space dimensions $d = 1$ and 2, the main convergence result was restricted to one space dimension. However, recently Grün [16] has proved convergence in two space dimensions of a finite element approximation to the thin film equation. This approach was extended in [7] to prove convergence in two space dimensions of a finite element approximation to the thin film equation in the presence of surfactants and repulsive van der Waals forces. It is the aim of this paper to adapt the techniques in [3, 4, 16] to propose and prove convergence of a finite element approximation of (P_γ), and hence prove existence of a solution to (P_γ).

The basic ingredients of our approach are some key energy estimates. First, we relate F to c and G to b by the identities

$$c(s)\,F''(s) = 1 \qquad \text{and} \qquad b(s)\,G''(s) = 1\,. \tag{1.7}$$

Knowing c and b (recall (1.6)), the above identities determine F and G up to a linear term. Furthermore, we have that F and G are convex. As the analysis in this paper is for a fixed γ, for the remainder of this paper we drop the γ subscripts in (P_γ) for notational convenience. One can then derive formally the following energy estimates for (P). Testing (1.4e) with ϕ yields that

$$\int_\Omega c(u)\,|\nabla\phi|^2\,\mathrm{d}x + \tfrac{1}{2}\int_{\partial_2\Omega}\phi^2\,\mathrm{d}s \le \tfrac{1}{2}\int_{\partial_2\Omega} g^2\,\mathrm{d}s\,, \tag{1.8}$$

where $g := g^\pm \equiv \pm(2 + L_1)$ on $\partial_2^\pm\Omega$. Testing (1.4e) with $F'(u)$ and noting (1.7) and (1.8) yield that

$$\left|\int_\Omega \nabla\phi\,.\,\nabla u\,\mathrm{d}x\right| = \left|\int_\Omega c(u)\nabla\phi\,.\,\nabla[F'(u)]\,\mathrm{d}x\right| \le 2\left[\int_{\partial_2\Omega} g^2\,\mathrm{d}s\right]^{\frac{1}{2}}\left[\int_{\partial_2\Omega}[F'(u)]^2\,\mathrm{d}s\right]^{\frac{1}{2}}. \tag{1.9}$$

Testing (1.4a) with w and (1.4b) with $\frac{\partial u}{\partial t}$, combining and noting (1.6) and (1.8) yields that

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}\int_\Omega\left[\tfrac{1}{2}\,\gamma\,|\nabla u|^2 + \gamma^{-1}\,\Psi(u)\right]\mathrm{d}x + \tfrac{1}{2}\,\gamma^{-1}\int_\Omega b(u)\,|\nabla w|^2\,\mathrm{d}x&\\ \le \tfrac{1}{2}\,\alpha^2\,\gamma^{-1}\int_\Omega b(u)\,|\nabla\phi|^2\,\mathrm{d}x \le \alpha^2\,\gamma^{-1}\int_\Omega c(u)\,|\nabla\phi|^2\,\mathrm{d}x&\\ \le \tfrac{1}{2}\,\alpha^2\,\gamma^{-1}\int_{\partial_2\Omega} g^2\,\mathrm{d}s\,.&\end{aligned} \tag{1.10}$$

Testing (1.4a) with $G'(u)$ and (1.4b) with $-\Delta u$, combining and noting (1.7), (1.5), and (1.9) yields that

$$\begin{aligned}\gamma\,\tfrac{\mathrm{d}}{\mathrm{d}t}\int_\Omega G(u)\,\mathrm{d}x + \gamma\int_\Omega|\Delta u|^2\,\mathrm{d}x \le \int_\Omega\nabla(\gamma^{-1}\,u - \alpha\,\phi)\,.\,\nabla u\,\mathrm{d}x&\\ \le \gamma^{-1}\int_\Omega|\nabla u|^2\,\mathrm{d}x + 2\,\alpha\left[\int_{\partial_2\Omega} g^2\,\mathrm{d}s\right]^{\frac{1}{2}}\left[\int_{\partial_2\Omega}[F'(u)]^2\,\mathrm{d}s\right]^{\frac{1}{2}}.&\end{aligned} \tag{1.11}$$

It is the goal of this paper to derive a finite element approximation of (P) that is consistent with the energy estimates (1.8)–(1.11). Following [3], we impose the $|\cdot| \leq 1$ constraint of the discrete approximation to u as a constraint and require equation (1.4b) only where $|u| < 1$. In addition, in order to derive a discrete analogue of the energy estimates (1.9) and (1.11) we adapt a technique introduced in [29, 17] for deriving a discrete entropy bound for the thin film equation.

This paper is organized as follows. In section 2 we formulate a fully practical finite element approximation of the degenerate system (P) and derive discrete analogues of the energy estimates (1.8)–(1.11). In section 3 we prove convergence, and hence existence of a solution to the system (P) in two space dimensions. In section 4 we introduce and prove convergence of a "Gauss–Seidel-type" iterative scheme for solving the nonlinear discrete system for the approximations of u and w at each time level. Finally, in section 5 we present some numerical experiments.

Notation and auxiliary results. For $D \subset \mathbb{R}^d$, $d = 1$ or 2, we adopt the standard notation for Sobolev spaces, denoting the norm of $W^{m,q}(D)$ ($m \in \mathbb{N}$, $q \in [1, \infty]$) by $\|\cdot\|_{m,q,D}$ and the seminorm by $|\cdot|_{m,q,D}$. We extend these norms and seminorms in the natural way to the corresponding spaces of vector and matrix valued functions. For $q = 2$, $W^{m,2}(D)$ will be denoted by $H^m(D)$ with the associated norm and seminorm written, respectively, as $\|\cdot\|_{m,D}$ and $|\cdot|_{m,D}$. For notational convenience, we drop the domain subscript on the above norms and seminorms in the case $D \equiv \Omega$. Throughout, $(\cdot, \cdot)$ denotes the standard L^2 inner product over Ω. In addition, we define

$$\fint \eta := \tfrac{1}{\mathcal{L}^2(\Omega)}\,(\eta, 1) \qquad \forall\ \eta \in L^1(\Omega)\,.$$

For later purposes, we recall the following compactness results. Let X, Y, and Z be Banach spaces with a compact embedding $X \hookrightarrow Y$ and a continuous embedding $Y \hookrightarrow Z$. Then the embeddings

$$\{\, \eta \in L^2(0,T;X) : \tfrac{\partial \eta}{\partial t} \in L^2(0,T;Z) \,\} \hookrightarrow L^2(0,T;Y) \tag{1.12a}$$

and

$$\{\, \eta \in L^\infty(0,T;X) : \tfrac{\partial \eta}{\partial t} \in L^2(0,T;Z) \,\} \hookrightarrow C([0,T];Y) \tag{1.12b}$$

are compact, and a generalized version of (1.12a), where the time derivative is replaced by a time translation, holds. That is, any bounded and closed subset E of $L^2(0,T;X)$ with

$$\lim_{\theta \to 0} \left\{ \sup_{\eta \in E} \|\eta(\cdot, \cdot + \theta) - \eta(\cdot, \cdot)\|_{L^2(0,T-\theta;Z)} \right\} = 0 \tag{1.12c}$$

is compact in $L^2(0,T;Y)$; see [26].

It is convenient to introduce the "inverse Laplacian" operator $\mathcal{G} : Y \to Z$ such that

$$(\nabla[\mathcal{G}z], \nabla\eta) = \langle z, \eta \rangle \qquad \forall\ \eta \in H^1(\Omega), \tag{1.13}$$

where $Y := \left\{ z \in (H^1(\Omega))' : \langle z, 1 \rangle = 0 \right\}$ and $Z := \{ z \in H^1(\Omega) : (z, 1) = 0 \}$. Here and throughout, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $(H^1(\Omega))'$ and $H^1(\Omega)$. The well

posedness of $\mathcal{G}$ follows from the generalized Lax–Milgram theorem and the Poincaré inequality

$$|\eta|_0 \leq C\,(\,|\eta|_1 + |(\eta,1)|\,) \qquad \forall\ \eta \in H^1(\Omega). \tag{1.14}$$

We note also for future reference Young's inequality

$$r\,s \leq \tfrac{\theta}{2}\,r^2 + \tfrac{1}{2\theta}\,s^2 \qquad \forall\ r,s \in \mathbb{R},\ \theta \in \mathbb{R}_{>0}\,. \tag{1.15}$$

Throughout, C denotes a generic constant independent of h, τ, and ε, the mesh and temporal discretization parameters, and the regularization parameter. In addition, $C(a_1,\ldots,a_I)$ denotes a constant depending on the arguments $\{a_i\}_{i=1}^I$. Furthermore, $\cdot^{(\star)}$ denotes an expression with or without the superscript $\star$. Finally, we define for any $s \in \mathbb{R}$

$$[s]_- := \min\{s,0\}, \quad [s]_+ := \max\{s,0\}, \quad [s]_{\mathcal{K}} := \max\{-1,\min\{s,1\}\}, \tag{1.16a}$$

and

$$\lfloor s \rfloor\ := \max\{z \in \mathbb{Z} : z \leq s\}. \tag{1.16b}$$

2. Finite element approximation. We consider the finite element approximation of (P) under the following assumptions on the mesh:

(A) Let Ω be a rectangular domain. Let $\{\mathcal{T}^h\}_{h>0}$ be a quasi-uniform family of partitionings of Ω into disjoint open simplices σ with $h_\sigma := \operatorname{diam}(\sigma)$ and $h := \max_{\sigma\in\mathcal{T}^h} h_\sigma$ so that $\overline{\Omega} = \cup_{\sigma\in\mathcal{T}^h}\overline{\sigma}$. In addition, it is assumed that all simplices $\sigma \in \mathcal{T}^h$ are right-angled.

We note that the right-angled simplices assumption is not a severe constraint, as there exist adaptive finite element codes that satisfy this requirement; see, e.g., [24].

Associated with $\mathcal{T}^h$ is the finite element space

$$S^h := \{\chi \in C(\overline{\Omega}) : \chi\,|_\sigma \ \text{ is linear } \forall\ \sigma \in \mathcal{T}^h\} \subset H^1(\Omega).$$

We introduce also

$$K^h := \{\chi \in S^h : |\chi| \leq 1 \text{ in } \Omega\} \subset K := \{\eta \in H^1(\Omega) : |\eta| \leq 1 \text{ a.e. in } \Omega\}\,.$$

Let J be the set of nodes of $\mathcal{T}^h$ and $\{p_j\}_{j\in J}$ the coordinates of these nodes. Let $\{\chi_j\}_{j\in J}$ be the standard basis functions for S^h; that is, $\chi_j \in S^h$ and $\chi_j(p_i) = \delta_{ij}$ for all $i,j \in J$. The right angle constraint on the partitioning is required for our approximations of $b(\cdot)$ and $c(\cdot)$ (see (2.12a), (2.12b) and (2.9a), (2.9b) below), but one consequence is that

$$\int_\sigma \nabla\chi_i\,.\,\nabla\chi_j\,\mathrm{d}x \leq 0, \qquad i \neq j \qquad \forall\ \sigma \in \mathcal{T}^h. \tag{2.1}$$

We introduce $\pi^h : C(\overline{\Omega}) \to S^h$, the interpolation operator, such that $(\pi^h\eta)(p_j) = \eta(p_j)$ for all $j \in J$. A discrete semi-inner product on $C(\overline{\Omega})$ is then defined by

$$(\eta_1,\eta_2)^h := \int_\Omega \pi^h(\eta_1(x)\,\eta_2(x))\,\mathrm{d}x = \sum_{j\in J} m_j\,\eta_1(p_j)\,\eta_2(p_j), \tag{2.2}$$

where $m_j := (1, \chi_j) > 0$. The induced discrete seminorm is then

$$|\eta|_h := [\,(\eta,\eta)^h\,]^{\frac{1}{2}} = \left(\int_\Omega \pi^h[\eta^2]\,\mathrm{d}x\right)^{\frac{1}{2}} \qquad \forall\ \eta \in C(\overline{\Omega}). \tag{2.3}$$

We introduce also the L^2 projection $Q^h : L^2(\Omega) \to S^h$ defined by

$$(Q^h\eta, \chi)^h = (\eta, \chi) \qquad \forall\ \chi \in S^h. \tag{2.4}$$

On recalling (1.6) and (1.7), we then define functions F and G such that $c(u)\,\nabla[F'(u)] = \nabla u$ and $b(u)\,\nabla[G'(u)] = \nabla u$; that is,

$$F''(s) = \tfrac{1}{c(s)} = \tfrac{1}{1+s} \quad \text{and} \quad G''(s) = \tfrac{1}{b(s)} = \tfrac{1}{c(s)\,c(-s)} = \tfrac{1}{1-s^2}\,. \tag{2.5}$$

We take F, $G \in C^\infty(-1,1)$ such that

$$F(s) = (1+s)\,\log(\tfrac{1+s}{2}) + (1-s) \qquad \text{and} \qquad G(s) = \tfrac{1}{2}\,[F(s) + F(-s)]\,; \tag{2.6}$$

and, for computational purposes, we replace F, G for any $\varepsilon \in (0,1)$ by the regularized functions F_ε, $G_\varepsilon : \mathbb{R} \to \mathbb{R}$ such that

$$F_\varepsilon(s) := \begin{cases} F(\varepsilon-1) + (s-\varepsilon+1)\,F'(\varepsilon-1) + \frac{(s-\varepsilon+1)^2}{2}\,F''(\varepsilon-1), & s \le \varepsilon-1, \\ F(s), & s \ge \varepsilon-1, \end{cases}$$

$$G_\varepsilon(s) := \tfrac{1}{2}\,[F_\varepsilon(s) + F_\varepsilon(-s)]\,. \tag{2.7}$$

Hence F_ε, $G_\varepsilon \in C^{2,1}(\mathbb{R})$ with the first two derivatives of F_ε given by

$$F_\varepsilon'(s) := \begin{cases} F'(\varepsilon-1) + (s-\varepsilon+1)\,F''(\varepsilon-1), & s \le \varepsilon-1, \\ F'(s), & s \ge \varepsilon-1, \end{cases}$$

$$\text{and} \quad F_\varepsilon''(s) := \begin{cases} F''(\varepsilon-1), & s \le \varepsilon-1, \\ F''(s), & s \ge \varepsilon-1\,, \end{cases}$$

respectively. We note for later purposes that for all $s \in \mathcal{K}$

$$\tfrac{1}{2} \le F_\varepsilon''(s) \le \varepsilon^{-1}, \qquad \tfrac{1}{2}\,F_\varepsilon''(s) \le G_\varepsilon''(s) \le [\varepsilon\,(2-\varepsilon)]^{-1} \le \varepsilon^{-1}\,; \tag{2.8a}$$

and for all $s_1, s_2 \in \mathcal{K}$ with $s_1 \ne s_2$

$$\tfrac{1}{2}\,\frac{F_\varepsilon'(s_1) - F_\varepsilon'(s_2)}{s_1 - s_2} \le \frac{G_\varepsilon'(s_1) - G_\varepsilon'(s_2)}{s_1 - s_2}\,. \tag{2.8b}$$

Similarly to the approach in [29, 17], we introduce $\Lambda_\varepsilon : S^h \to [L^\infty(\Omega)]^{2\times 2}$ such that for all $z^h \in S^h$ and a.e. in Ω

$$\Lambda_\varepsilon(z^h) \text{ is symmetric and positive semidefinite,} \tag{2.9a}$$

$$\Lambda_\varepsilon(z^h)\,\nabla\pi^h[F_\varepsilon'(z^h)] = \nabla z^h. \tag{2.9b}$$

We now give the construction of Λ_ε. Let $\{e_i\}_{i=1}^2$ be the orthonormal vectors in $\mathbb{R}^2$ such that the jth component of e_i is δ_{ij}, $i, j = 1 \to 2$. Given nonzero constants β_i, $i = 1 \to 2$, let $\widehat{\sigma}(\{\beta_i\}_{i=1}^2)$ be the reference open simplex in $\mathbb{R}^2$ with vertices $\{\widehat{p}_i\}_{i=0}^2$, where $\widehat{p}_0$ is the origin and $\widehat{p}_i = \beta_i\,e_i$, $i = 1 \to 2$. Given a $\sigma \in \mathcal{T}^h$ with vertices $\{p_{j_i}\}_{i=0}^2$ such that p_{j_0} is the right-angled vertex, then there exists a rotation matrix R_σ and

nonzero constants $\{\beta_i\}_{i=1}^2$ such that the mapping $\mathcal{R}_\sigma : \widehat{x} \in \mathbb{R}^2 \to p_{j_0} + R_\sigma \widehat{x} \in \mathbb{R}^2$ maps the vertex $\widehat{p}_i$ to p_{j_i}, $i = 0 \to 2$, and hence $\widehat{\sigma} \equiv \widehat{\sigma}(\{\beta_i\}_{i=1}^2)$ to σ. For any $z^h \in S^h$, we then set

$$\Lambda_\varepsilon(z^h)\,|_\sigma := R_\sigma\, \widehat{\Lambda}_\varepsilon(\widehat{z}^h)\,|_{\widehat{\sigma}}\, R_\sigma^T, \tag{2.10}$$

where $\widehat{z}^h(\widehat{x}) \equiv z^h(\mathcal{R}_\sigma \widehat{x})$ for all $\widehat{x} \in \overline{\widehat{\sigma}}$ and $\widehat{\Lambda}_\varepsilon(\widehat{z}^h)\,|_{\widehat{\sigma}}$ is the 2×2 diagonal matrix with diagonal entries, where $k = 1 \to 2$:

(2.11)
$$[\widehat{\Lambda}_\varepsilon(\widehat{z}^h)\,|_{\widehat{\sigma}}]_{kk} := \begin{cases} \frac{\widehat{z}^h(\widehat{p}_k) - \widehat{z}^h(\widehat{p}_0)}{F_\varepsilon'(\widehat{z}^h(\widehat{p}_k)) - F_\varepsilon'(\widehat{z}^h(\widehat{p}_0))} \equiv \frac{z^h(p_{j_k}) - z^h(p_{j_0})}{F_\varepsilon'(z^h(p_{j_k})) - F_\varepsilon'(z^h(p_{j_0}))} & \text{if } z^h(p_{j_k}) \neq z^h(p_{j_0}), \\ \frac{1}{F_\varepsilon''(\widehat{z}^h(\widehat{p}_0))} \equiv \frac{1}{F_\varepsilon''(z^h(p_{j_0}))} & \text{if } z^h(p_{j_k}) = z^h(p_{j_0}). \end{cases}$$

As $R_\sigma^T \equiv R_\sigma^{-1}$, $\nabla z^h \equiv R_\sigma\, \widehat{\nabla} \widehat{z}^h$, where $x \equiv (x_1, x_2)^T$, $\nabla \equiv (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2})^T$, $\widehat{x} \equiv (\widehat{x}_1, \widehat{x}_2)^T$, and $\widehat{\nabla} \equiv (\frac{\partial}{\partial \widehat{x}_1}, \frac{\partial}{\partial \widehat{x}_2})^T$, it easily follows that $\Lambda_\varepsilon(z^h)$ constructed in (2.10) and (2.11) satisfies (2.9a), (2.9b). It is this construction that requires the right angle constraint on the partitioning $\mathcal{T}^h$. In a similar fashion we introduce $\Xi_\varepsilon : S^h \to [L^\infty(\Omega)]^{2\times 2}$ such that for all $z^h \in S^h$ and a.e. in Ω

$$\Xi_\varepsilon(z^h) \text{ is symmetric and positive semidefinite,} \tag{2.12a}$$

$$\Xi_\varepsilon(z^h)\, \nabla \pi^h [G_\varepsilon'(z^h)] = \nabla z^h. \tag{2.12b}$$

We extend the construction (2.10)–(2.11) for Λ_ε to Ξ_ε.

In addition to $\mathcal{T}^h$, let $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$ be a partitioning of $[0, T]$ into possibly variable time steps $\tau_n := t_n - t_{n-1}$, $n = 1 \to N$. We set $\tau := \max_{n=1\to N} \tau_n$. For any given $\varepsilon \in (0, 1)$, we then consider the following fully practical finite element approximation of (P):

($\mathrm{P}_\varepsilon^{h,\tau}$) For $n \geq 1$ find $\{\Phi_\varepsilon^n, U_\varepsilon^n, W_\varepsilon^n\} \in S^h \times K^h \times S^h$ such that

$$(\Lambda_\varepsilon(U_\varepsilon^{n-1})\, \nabla \Phi_\varepsilon^n, \nabla \chi) + \int_{\partial_2 \Omega} \Phi_\varepsilon^n\, \chi \,\mathrm{d}s = \int_{\partial_2 \Omega} g\, \chi \,\mathrm{d}s \qquad \forall\ \chi \in S^h, \tag{2.13a}$$

$$\gamma \left(\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}, \chi \right)^h + (\Xi_\varepsilon(U_\varepsilon^{n-1})\, \nabla [W_\varepsilon^n + \alpha\, \Phi_\varepsilon^n], \nabla \chi) = 0 \qquad \forall\ \chi \in S^h, \tag{2.13b}$$

$$\gamma\, (\nabla U_\varepsilon^n, \nabla [\chi - U_\varepsilon^n]) \geq (W_\varepsilon^n + \gamma^{-1}\, U_\varepsilon^{n-1}, \chi - U_\varepsilon^n)^h \qquad \forall\ \chi \in K^h, \tag{2.13c}$$

where g is as in (1.8) and $U_\varepsilon^0 \in K^h$ is an approximation of $u^0 \in K$, e.g., $U_\varepsilon^0 \equiv Q^h u^0$ or $U_\varepsilon^0 \equiv \pi^h u^0$, if $u^0 \in C(\overline{\Omega})$.

Remark 2.1. We note that in the case $\alpha = 0$, (2.13b), (2.13c) collapses to an approximation of the degenerate Cahn–Hilliard equation, (1.4a)–(1.4d) with $\alpha = 0$, discussed in the multicomponent context in [4, pp. 731–734].

Below we recall some well-known results concerning S^h for any $\sigma \in \mathcal{T}^h$, $\chi, z^h \in S^h$, $m \in \{0, 1\}$, $p \in [1, \infty]$, and $q \in (2, \infty]$:

$$|\chi|_{1,\sigma} \leq C\, h_\sigma^{-1}\, |\chi|_{0,\sigma}\,; \tag{2.14}$$

$$|\chi|_{m,r,\sigma} \leq C\, h_\sigma^{-2\,(\frac{1}{p} - \frac{1}{r})}\, |\chi|_{m,p,\sigma} \qquad \text{for any } r \in [p, \infty]\,; \tag{2.15}$$

$$|(I - \pi^h)\eta|_m \leq C\, h^{2-m}\, |\eta|_2 \qquad \forall\ \eta \in H^2(\Omega)\,; \tag{2.16}$$

$$|(I - \pi^h)\eta|_{m,q} \leq C\, h^{1-m}\, |\eta|_{1,q} \qquad \forall\ \eta \in W^{1,q}(\Omega)\,; \tag{2.17}$$

$$\int_\sigma \chi^2 \,\mathrm{d}x \leq \int_\sigma \pi^h[\chi^2] \,\mathrm{d}x \leq 4 \int_\sigma \chi^2 \,\mathrm{d}x\,; \tag{2.18}$$

$$|(\chi, z^h) - (\chi, z^h)^h| \leq |(I - \pi^h)(\chi\, z^h)|_{0,1} \leq C\, h^{1+m}\, |\chi|_m\, |z^h|_1\,. \tag{2.19}$$

It follows from (2.4) that

$$(Q^h\eta)(p_j) = \frac{(\eta,\chi_j)}{(1,\chi_j)} \quad \forall\, j \in J \quad \Longrightarrow \quad |Q^h\eta|_{0,\infty} \leq |\eta|_{0,\infty} \quad \forall\, \eta \in L^\infty(\Omega). \tag{2.20}$$

Finally, as we have a quasi-uniform family of partitionings, it holds that

$$|(I - Q^h)\eta|_m \leq C\, h^{1-m}\, |\eta|_1 \qquad \forall\, \eta \in H^1(\Omega). \tag{2.21}$$

We define $Z^h := \{z^h \in S^h : (z^h, 1) = 0\} \subset Y^h := \{z \in C(\overline{\Omega}) : (z,1)^h = 0\}$. Then, similarly to (1.13), we introduce $\mathcal{G}^h : Y^h \to Z^h$ such that

$$(\nabla[\mathcal{G}^h z], \nabla\chi) = (z,\chi)^h \qquad \forall\, \chi \in S^h\,. \tag{2.22}$$

We introduce the "discrete Laplacian" operator $\Delta^h : S^h \to Z^h$ such that

$$(\Delta^h z^h, \chi)^h = -(\nabla z^h, \nabla\chi) \qquad \forall\, \chi \in S^h. \tag{2.23}$$

We note for future reference, as we have a quasi-uniform family of partitionings and as Ω is convex, that for all $z^h \in S^h$

$$|z^h|_{1,s} \leq C\, |\Delta^h z^h|_0 \quad \text{for any } s \in (1,\infty)\,; \tag{2.24}$$

see, for example, [6, Lemma 3.1].

We introduce, for all $\varepsilon \in (0,1)$, $c_\varepsilon : \mathcal{K} \to [\varepsilon, 2]$ and $b_\varepsilon : \mathcal{K} \to [\varepsilon\,(2-\varepsilon), 1]$, defined, on recalling (2.5), (2.7), (2.8a), and (1.16a), by

$$c_\varepsilon(s) := \ [c(s) - \varepsilon]_+ + \varepsilon = \tfrac{1}{F''_\varepsilon(s)} \geq \tfrac{1}{F''(s)} = c(s)\,, \tag{2.25a}$$

$$b_\varepsilon(s) := 2\, \tfrac{c_\varepsilon(s)\, c_\varepsilon(-s)}{c_\varepsilon(s) + c_\varepsilon(-s)} = \tfrac{1}{G''_\varepsilon(s)} \geq \tfrac{1}{G''(s)} = b(s)\,. \tag{2.25b}$$

LEMMA 2.2. *Let the assumptions* (A) *hold. Then for any given $\varepsilon \in (0,1)$ the functions Λ_ε, $\Xi_\varepsilon : S^h \to [L^\infty(\Omega)]^{2\times 2}$ satisfy for all $z^h \in K^h$, for all $\xi \in \mathbb{R}^2$, and for all $\sigma \in \mathcal{T}^h$*

$$\varepsilon\, \xi^T\xi \leq \min_{x\in\overline{\sigma}} c_\varepsilon(z^h(x))\, \xi^T\xi \leq \ \xi^T\, \Lambda_\varepsilon(z^h)\,|_\sigma\, \xi \leq \max_{x\in\overline{\sigma}} c_\varepsilon(z^h(x))\, \xi^T\xi \leq 2\, \xi^T\xi\,, \tag{2.26a}$$

$$\varepsilon\,(2-\varepsilon)\, \xi^T\xi \leq \min_{x\in\overline{\sigma}} b_\varepsilon(z^h(x))\, \xi^T\xi \leq \ \xi^T\, \Xi_\varepsilon(z^h)\,|_\sigma\, \xi \leq \max_{x\in\overline{\sigma}} b_\varepsilon(z^h(x))\, \xi^T\xi \leq \xi^T\xi\,, \tag{2.26b}$$

$$\xi^T\, \Xi_\varepsilon(z^h)\,|_\sigma\, \xi \leq 2\, \xi^T\, \Lambda_\varepsilon(z^h)\,|_\sigma\, \xi\,. \tag{2.26c}$$

Proof. The desired results (2.26a)–(2.26c) follow from the construction of Λ_ε and Ξ_ε; cf. (2.10) and (2.11), (2.25a), (2.25b), and (2.8a), (2.8b). ∎

LEMMA 2.3. *Let the assumptions* (A) *hold, and let $\|\cdot\|$ denote the spectral norm on $\mathbb{R}^{2\times 2}$. Then for any given $\varepsilon \in (0,1)$ the functions $\Lambda_\varepsilon : S^h \to [L^\infty(\Omega)]^{2\times 2}$ and $\Xi_\varepsilon : S^h \to [L^\infty(\Omega)]^{2\times 2}$ are such that for all $z^h \in K^h$ and for all $\sigma \in \mathcal{T}^h$*

$$\max_{x\in\sigma} \|\{\Lambda_\varepsilon(z^h) - c_\varepsilon(z^h)\, \mathcal{I}\}(x)\| \leq h_\sigma\, |\nabla[c_\varepsilon(z^h)]\,|_{0,\infty,\sigma} \leq h_\sigma\, |\nabla z^h\,|_\sigma\,|\,, \tag{2.27a}$$

$$\max_{x\in\sigma} \|\{\Xi_\varepsilon(z^h) - b_\varepsilon(z^h)\, \mathcal{I}\}(x)\| \leq h_\sigma\, |\nabla[b_\varepsilon(z^h)]\,|_{0,\infty,\sigma} \leq 2\, h_\sigma\, |\nabla z^h\,|_\sigma\,|\,, \tag{2.27b}$$

where $\mathcal{I}$ is the 2×2 identity matrix.

Proof. Adopting the notation of (2.10), we have from (2.11), (2.25a), (1.6), and the Lipschitz continuity of c_ε that

$$\begin{aligned}\max_{x\in\sigma}\|\{\Lambda_\varepsilon(z^h)-c_\varepsilon(z^h)\,\mathcal{I}\}(x)\| &= \max_{\widehat{x}\in\widehat{\sigma}}\|\{\widehat{\Lambda}_\varepsilon(\widehat{z}^h)-c_\varepsilon(\widehat{z}^h)\,\mathcal{I}\}(\widehat{x})\|\\ &= \max_{\widehat{x}\in\widehat{\sigma}}\max_{k=1\to 2}|\,[\widehat{\Lambda}_\varepsilon(\widehat{z}^h)]_{kk}-c_\varepsilon(\widehat{z}^h)(\widehat{x})\,|\le h_\sigma\,|\widehat{\nabla}[c_\varepsilon(\widehat{z}^h)]\,|_{0,\infty,\widehat{\sigma}}\\ &= h_\sigma\,|\nabla[c_\varepsilon(z^h)]\,|_{0,\infty,\sigma}\le h_\sigma\,|\nabla z^h\,|_\sigma\,|,\end{aligned}$$

where we have noted that $[\widehat{\Lambda}_\varepsilon(\widehat{z}^h)\,|_{\widehat{\sigma}}\,]_{kk}=c_\varepsilon(\widehat{z}^h(\widehat{\xi}^{(k)}))\equiv c_\varepsilon(z^h(\xi^{(k)}))$ with $\xi^{(k)}\equiv\mathcal{R}_\sigma\widehat{\xi}^{(k)}\in\overline{\sigma}$ for some point $\widehat{\xi}^{(k)}\in\overline{\widehat{\sigma}}$. Hence we obtain the desired result (2.27a). The desired result (2.27b) follows similarly to the above on noting the Lipschitz continuity of b_ε; see (2.25b). ∎

LEMMA 2.4. *Let the assumptions* (A) *hold and* $U_\varepsilon^{n-1}\in K^h$. *Then for all* $\varepsilon\in(0,1)$ *and for all* h, $\tau_n>0$ *there exists a solution* $\{\Phi_\varepsilon^n,U_\varepsilon^n,W_\varepsilon^n\}$ *to the* n*th step of* $(\mathrm{P}_\varepsilon^{h,\tau})$ *with* $\fint U_\varepsilon^n=\fint U_\varepsilon^{n-1}$. $\{\Phi_\varepsilon^n,U_\varepsilon^n\}$ *is unique. In addition,* W_ε^n *is unique if there exists* $j\in J$ *such that* $U_\varepsilon^n(p_j)\in(-1,1)$. *Moreover, it holds that*

$$(\Lambda_\varepsilon(U_\varepsilon^{n-1})\,\nabla\Phi_\varepsilon^n,\nabla\Phi_\varepsilon^n)+\tfrac12\,|\Phi_\varepsilon^n|^2_{0,\partial_2\Omega}\le\tfrac12\,|g|^2_{0,\partial_2\Omega},\tag{2.28}$$

$$|(\nabla\Phi_\varepsilon^n,\nabla U_\varepsilon^{n-1})|\le 2\,|g|_{0,\partial_2\Omega}\,|\pi^h[F_\varepsilon'(U_\varepsilon^{n-1})]|_{0,\partial_2\Omega},\tag{2.29}$$

and

$$\begin{aligned}&\mathcal{E}(U_\varepsilon^n)+\tfrac12\big[\,\gamma\,|U_\varepsilon^n-U_\varepsilon^{n-1}|_1^2+\gamma^{-1}\,|U_\varepsilon^n-U_\varepsilon^{n-1}|_h^2\,\big]\\ &\quad+\tfrac12\,\gamma^{-1}\,\tau_n\,|\,[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac12}\,\nabla W_\varepsilon^n|_0^2\le\mathcal{E}(U_\varepsilon^{n-1})+\tfrac12\,\alpha^2\,\gamma^{-1}\,\tau_n\,|g|^2_{0,\partial_2\Omega}\,,\end{aligned}\tag{2.30a}$$

where

$$\mathcal{E}(U_\varepsilon^n):=\tfrac12\,[\,\gamma\,|U_\varepsilon^n|_1^2-\gamma^{-1}\,|U_\varepsilon^n|_h^2\,]\,.\tag{2.30b}$$

Furthermore, it holds that

$$\begin{aligned}&\gamma\,(G_\varepsilon(U_\varepsilon^n)-G_\varepsilon(U_\varepsilon^{n-1}),1)^h+\gamma\,\tau_n\,|\Delta^h U_\varepsilon^n|_h^2\le\varepsilon^{-1}\,\gamma\,|U_\varepsilon^n-U_\varepsilon^{n-1}|_h^2\\ &\qquad+\tau_n\,(\nabla W_\varepsilon^n,\nabla[U_\varepsilon^n-U_\varepsilon^{n-1}]\,)+\tau_n\,(\nabla[\gamma^{-1}\,U_\varepsilon^n-\alpha\,\Phi_\varepsilon^n],\nabla U_\varepsilon^{n-1})\,.\end{aligned}\tag{2.31}$$

Proof. Given $U_\varepsilon^{n-1}\in K^h$, it follows immediately from (2.26a) and a Friedrich inequality that there exists a unique solution $\Phi_\varepsilon^n\in S^h$ to (2.13a). In order to prove existence of a solution $\{U_\varepsilon^n,W_\varepsilon^n\}\in K^h\times S^h$ to (2.13b), (2.13c), we introduce, similarly to (2.22), for $q^h\in K^h$ the discrete anisotropic Green's operator $\mathcal{G}_{q^h}^h:Z^h\to Z^h$ such that

$$(\Xi_\varepsilon(q^h)\,\nabla[\mathcal{G}_{q^h}^h z^h],\nabla\chi)=(z^h,\chi)^h\qquad\forall\ \chi\in S^h.\tag{2.32}$$

It follows immediately from (2.26b) and (1.14) that $\mathcal{G}_{q^h}^h$ is well-posed. It follows from (2.13b) and (2.32) that

$$W_\varepsilon^n\equiv-\alpha\,\Phi_\varepsilon^n-\gamma\,\mathcal{G}^h_{U_\varepsilon^{n-1}}\Big[\tfrac{U_\varepsilon^n-U_\varepsilon^{n-1}}{\tau_n}\Big]+\lambda^n,\tag{2.33}$$

where $\lambda^n\in\mathbb{R}$ is a constant. Hence (2.13b), (2.13c) can be restated as follows: Find $U_\varepsilon^n\in K^h(U_\varepsilon^{n-1}):=\{\chi\in K^h:\chi-U_\varepsilon^{n-1}\in Z^h\}$ and a Lagrange multiplier $\lambda^n\in\mathbb{R}$

such that for all $\chi \in K^h$

$$\gamma \left[(\nabla U_\varepsilon^n, \nabla(\chi - U_\varepsilon^n)) + \left(\mathcal{G}^h_{U_\varepsilon^{n-1}} \left[\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n} \right], \chi - U_\varepsilon^n \right)^h \right] \geq (\gamma^{-1}\, U_\varepsilon^{n-1} - \alpha\, \Phi_\varepsilon^n + \lambda^n, \chi - U_\varepsilon^n)^h. \tag{2.34}$$

It follows from (2.34) that $U_\varepsilon^n \in K^h(U_\varepsilon^{n-1})$ is such that for all $\chi \in K^h(U_\varepsilon^{n-1})$

$$\gamma \left[(\nabla U_\varepsilon^n, \nabla(\chi - U_\varepsilon^n)) + \left(\mathcal{G}^h_{U_\varepsilon^{n-1}} \left[\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n} \right], \chi - U_\varepsilon^n \right)^h \right] \geq (\gamma^{-1} U_\varepsilon^{n-1} - \alpha \Phi_\varepsilon^n, \chi - U_\varepsilon^n)^h. \tag{2.35}$$

There exists a unique $U_\varepsilon^n \in K^h(U_\varepsilon^{n-1})$ solving (2.35) since, on noting (2.32), this is the Euler–Lagrange variational inequality of the strictly convex minimization problem

$$\min_{z^h \in K^h(U_\varepsilon^{n-1})} \left\{ \tfrac{\gamma}{2} |z^h|_1^2 + \tfrac{\gamma}{2\,\tau_n} |[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac12}\, \nabla \mathcal{G}^h_{U_\varepsilon^{n-1}}(z^h - U_\varepsilon^{n-1})|_0^2 - (\gamma^{-1}\, U_\varepsilon^{n-1} - \alpha\, \Phi_\varepsilon^n, z^h)^h \right\}.$$

Existence of the Lagrange multiplier λ^n in (2.34) then follows from standard optimization theory; see, e.g., [11]. Therefore we have existence of a solution $\{U_\varepsilon^n, W_\varepsilon^n\} \in K^h \times S^h$ to (2.13b), (2.13c). If $|U_\varepsilon^n(p_j)| < 1$ for some $j \in J$, then $\pi^h[1 - (U_\varepsilon^n)^2] \not\equiv 0$ and choosing $\chi \equiv U_\varepsilon^n \pm \delta \pi^h[1 - (U_\varepsilon^n)^2]$ in (2.34) for $\delta > 0$ sufficiently small yields uniqueness of λ^n and, on noting (2.33), uniqueness of W_ε. Furthermore, choosing $\chi \equiv 1$ in (2.13b) yields $\fint U_\varepsilon^n = \fint U_\varepsilon^{n-1}$.

The bound (2.28) follows immediately from choosing $\chi \equiv \Phi_\varepsilon^n$ in (2.13a) and applying (1.15). Choosing $\chi \equiv \pi^h[F_\varepsilon'(U_\varepsilon^{n-1})]$ in (2.13a) and noting (2.9b) and (2.28) yield that

$$|(\nabla \Phi_\varepsilon^n, \nabla U_\varepsilon^{n-1})| = \left| \int_{\partial_2\Omega} (g - \Phi_\varepsilon^n)\, \pi^h[F_\varepsilon'(U_\varepsilon^{n-1})]\, \mathrm{d}s \right| \leq 2\, |g|_{0,\partial_2\Omega}\, |\pi^h[F_\varepsilon'(U_\varepsilon^{n-1})]|_{0,\partial_2\Omega},$$

and hence the desired result (2.29). Choosing $\chi \equiv W_\varepsilon^n$ in (2.13b) and $\chi \equiv U_\varepsilon^{n-1}$ in (2.13c) yields that

$$\gamma\, (U_\varepsilon^n - U_\varepsilon^{n-1}, W_\varepsilon^n)^h + \tau_n\, (\Xi_\varepsilon(U_\varepsilon^{n-1})\, \nabla[W_\varepsilon^n + \alpha\, \Phi_\varepsilon^n], \nabla W_\varepsilon^n) = 0, \tag{2.36a}$$

$$\gamma\, (\nabla U_\varepsilon^n, \nabla[U_\varepsilon^{n-1} - U_\varepsilon^n]) \geq (W_\varepsilon^n + \gamma^{-1}\, U_\varepsilon^{n-1}, U_\varepsilon^{n-1} - U_\varepsilon^n)^h. \tag{2.36b}$$

On noting the elementary identity

$$2\, r\, (r - s) = (r^2 - s^2) + (r - s)^2 \qquad \forall\ r, s \in \mathbb{R},$$

it follows from (2.36a), (2.36b), (2.30b), (1.15), and (2.26c) that

$$\begin{aligned} &\mathcal{E}(U_\varepsilon^n) + \tfrac12 \left[\gamma\, |U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2 + \gamma^{-1}\, |U_\varepsilon^n - U_\varepsilon^{n-1}|_h^2 \right] + \gamma^{-1}\, \tau_n\, |\, [\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac12}\, \nabla W_\varepsilon^n|_0^2 \\ &\quad \leq \mathcal{E}(U_\varepsilon^{n-1}) - \alpha\, \gamma^{-1}\, \tau_n\, (\Xi_\varepsilon(U_\varepsilon^{n-1})\, \nabla \Phi_\varepsilon^n, \nabla W_\varepsilon^n) \\ &\quad \leq \mathcal{E}(U_\varepsilon^{n-1}) + \gamma^{-1}\, \tfrac{\tau_n}{2} \left[|\, [\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac12}\, |\nabla W_\varepsilon^n|_0^2 + 2\, \alpha^2\, |\, [\Lambda_\varepsilon(U_\varepsilon^{n-1})]^{\frac12}\, \nabla \Phi_\varepsilon^n|_0^2 \right]. \end{aligned} \tag{2.37}$$

Hence the desired result (2.30a) follows from (2.37) and (2.28).

Choosing $\chi \equiv \pi^h[G'_\varepsilon(U_\varepsilon^{n-1})]$ in (2.13b) and noting (2.12b) yield that

$$\gamma\,(U_\varepsilon^n - U_\varepsilon^{n-1}, G'_\varepsilon(U_\varepsilon^{n-1}))^h + \tau_n\,(\nabla[W_\varepsilon^n + \alpha\,\Phi_\varepsilon^n], \nabla U_\varepsilon^{n-1}) = 0\,. \tag{2.38}$$

We now apply an argument similar to that in [4, Theorem 2.3]. From (2.13c) we have for all $j \in J$ on choosing $\chi \equiv U_\varepsilon^n + \delta\,\chi_j$, $U_\varepsilon^n \pm \delta\,\chi_j$, $U_\varepsilon^n - \delta\,\chi_j \in K^h$, respectively, for $\delta > 0$ sufficiently small, that

$$\gamma\,(\nabla U_\varepsilon^n, \nabla\chi_j) - (W_\varepsilon^n + \gamma^{-1}\,U_\varepsilon^{n-1}, \chi_j)^h \begin{cases} \geq 0 \\ = 0 \\ \leq 0 \end{cases} \quad \text{if} \quad U_\varepsilon^n(p_j) \quad \begin{cases} = -1 \\ \in (-1,1) \\ = 1 \end{cases}\,. \tag{2.39}$$

From (2.23), (2.2), and (2.1) it follows for all $j \in J$ that

$$U_\varepsilon^n(p_j) = \pm 1 \implies \pm U_\varepsilon^n(p_j) \geq \pm U_\varepsilon^n(p_i) \qquad \forall\, i \in J \implies \pm\,\Delta^h U_\varepsilon^n(p_j) \leq 0. \tag{2.40}$$

Combining (2.39) and (2.40) and noting (2.23) and (2.3) yield

$$\begin{aligned} \gamma\,|\Delta^h U_\varepsilon^n|_h^2 = -\gamma\,(\nabla U_\varepsilon^n, \nabla(\Delta^h U_\varepsilon^n)\,) &\leq -(W_\varepsilon^n + \gamma^{-1}\,U_\varepsilon^{n-1}, \Delta^h U_\varepsilon^n)^h \\ &= (\nabla[W_\varepsilon^n + \gamma^{-1}\,U_\varepsilon^{n-1}], \nabla U_\varepsilon^n). \end{aligned} \tag{2.41}$$

It follows from (2.38), (2.8a), and (2.41) that

$$\begin{aligned} &\gamma(G_\varepsilon(U_\varepsilon^n) - G_\varepsilon(U_\varepsilon^{n-1}), 1)^h + \gamma\tau_n|\Delta^h U_\varepsilon^n|_h^2 \\ &\quad\leq \gamma(U_\varepsilon^n - U_\varepsilon^{n-1}, G'_\varepsilon(U_\varepsilon^n))^h + \tau_n(\nabla[W_\varepsilon^n + \gamma^{-1}U_\varepsilon^{n-1}], \nabla U_\varepsilon^n) \\ &\quad\leq \gamma(U_\varepsilon^n - U_\varepsilon^{n-1}, G'_\varepsilon(U_\varepsilon^n) - G'_\varepsilon(U_\varepsilon^{n-1}))^h + \tau_n(\nabla W_\varepsilon^n, \nabla[U_\varepsilon^n - U_\varepsilon^{n-1}]) \\ &\qquad\qquad + \tau_n(\nabla[\gamma^{-1}U_\varepsilon^n - \alpha\Phi_\varepsilon^n], \nabla U_\varepsilon^{n-1}) \\ &\quad\leq \varepsilon^{-1}\gamma|U_\varepsilon^n - U_\varepsilon^{n-1}|_h^2 + \tau_n[(\nabla W_\varepsilon^n, \nabla[U_\varepsilon^n - U_\varepsilon^{n-1}]) + (\nabla[\gamma^{-1}U_\varepsilon^n - \alpha\Phi_\varepsilon^n], \nabla U_\varepsilon^{n-1})], \end{aligned}$$

and hence the desired result (2.31). ∎

Remark 2.5. We note that (2.28)–(2.31) are the discrete analogues of the energy estimates (1.8)–(1.11), respectively.

THEOREM 2.6. *Let the assumptions* (A) *hold and* $U_\varepsilon^0 \in K^h$. *Then for all* $\varepsilon \in (0,1)$, *for all* $h > 0$, *and for all time partitions* $\{\tau_n\}_{n=1}^N$, *the solution* $\{\Phi_\varepsilon^n, U_\varepsilon^n, W_\varepsilon^n\}_{n=1}^N$ *to* $(\mathrm{P}_\varepsilon^{h,\tau})$ *is such that* $\fint U_\varepsilon^n = \fint U_\varepsilon^0$, $n = 1 \to N$, *and*

$$\begin{aligned} &\gamma \max_{n=1\to N} \|U_\varepsilon^n\|_1^2 + \sum_{n=1}^N \left[\,\gamma\,|U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2 + \gamma^{-1}\,|U_\varepsilon^n - U_\varepsilon^{n-1}|_0^2\,\right] \\ &\quad + \gamma^{-1}\sum_{n=1}^N \tau_n\,|\,[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac12}\,\nabla W_\varepsilon^n|_0^2 \leq C\left[\,\gamma\,\|U_\varepsilon^0\|_1^2 + \gamma^{-1}\,(1 + T\,|g|_{0,\partial_2\Omega}^2)\,\right]. \end{aligned} \tag{2.42}$$

In addition,

$$\gamma \sum_{n=1}^{N} \tau_n \left| \mathcal{G}\left[\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}\right]\right|_1^2 + \gamma\, \tau^{-\frac{1}{2}} \sum_{n=1}^{N} |U_\varepsilon^n - U_\varepsilon^{n-1}|_0^2$$
$$\leq C\left[\, \gamma\, \|U_\varepsilon^0\|_1^2 + \gamma^{-1}\,(1 + T\,|g|_{0,\partial_2\Omega}^2)\,\right] \tag{2.43}$$

and
$$\gamma \max_{n=1\to N} (G_\varepsilon(U_\varepsilon^n), 1)^h + \gamma \sum_{n=1}^{N} \tau_n\, |\Delta^h U_\varepsilon^n|_h^2$$
$$\leq \gamma\,(G_\varepsilon(U_\varepsilon^0), 1)^h + \alpha^2 \sum_{n=1}^{N} \tau_n\, |\pi^h[F_\varepsilon'(U_\varepsilon^{n-1})]|_{0,\partial_2\Omega}^2$$
$$+ C(T)\,[\,1 + \gamma^{-2} + \varepsilon^{-1}\tau^{\frac{1}{2}}\,]\,\left[\,\gamma\,\|U_\varepsilon^0\|_1^2 + \gamma^{-1}\,(1 + T\,|g|_{0,\partial_2\Omega}^2)\,\right]. \tag{2.44}$$

Proof. Summing (2.30a) from $n = 1 \to k$ yields for any $k \leq N$ that

$$\mathcal{E}(U_\varepsilon^k) + \tfrac{1}{2} \sum_{n=1}^{k} \left[\gamma\, |U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2 + \gamma^{-1}\, |U_\varepsilon^n - U_\varepsilon^{n-1}|_h^2\,\right]$$
$$+ \tfrac{1}{2}\,\gamma^{-1} \sum_{n=1}^{k} \tau_n\, |\,[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac{1}{2}}\, \nabla W_\varepsilon^n|_0^2 \leq \mathcal{E}(U_\varepsilon^0) + \tfrac{1}{2}\,\alpha^2\,\gamma^{-1}\, t_k\, |g|_{0,\partial_2\Omega}^2\,. \tag{2.45}$$

The desired result (2.42) then follows from (2.45), (2.30b), (2.3), (2.18), and the fact that $U_\varepsilon^n \in K^h$, $n = 0 \to N$.

From (1.13), (2.4), (2.13b), (2.26b), (2.26c), and (2.21) we obtain for any $\eta \in H^1(\Omega)$ that

$$\gamma\left(\nabla \mathcal{G}[\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}], \nabla\eta\right) = \gamma\left(\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}, \eta\right) = \gamma\left(\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}, Q^h\eta\right)^h$$
$$= -(\Xi_\varepsilon(U_\varepsilon^{n-1})\,\nabla[W_\varepsilon^n + \alpha\,\Phi_\varepsilon^n], \nabla[Q^h\eta])$$
$$\leq \left[\, |[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla W_\varepsilon^n|_0 + \alpha\,|[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla\Phi_\varepsilon^n|_0\,\right] |Q^h\eta|_1$$
$$\leq C\left[\, |[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla W_\varepsilon^n|_0 + \alpha\,|[\Lambda_\varepsilon(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla\Phi_\varepsilon^n|_0\,\right] |\eta|_1\,. \tag{2.46}$$

The first bound in (2.43) then follows from (2.46), (2.28), and (2.42). Moreover, we have from (1.13) that

$$\sum_{n=1}^{N} |U_\varepsilon^n - U_\varepsilon^{n-1}|_0^2 \leq \tau^{\frac{1}{2}} \left[\sum_{n=1}^{N} |U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2\right]^{\frac{1}{2}} \left[\sum_{n=1}^{N} \tau_n \left|\mathcal{G}[\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}]\right|_1^2\right]^{\frac{1}{2}}.$$

The second bound in (2.43) then follows from the first and (2.42). Finally, summing (2.31) from $n = 1 \to k$ and noting (2.3), (2.18), and (2.26b) yield for any $k \leq N$ that

$$\gamma\,(G_\varepsilon(U_\varepsilon^k), 1)^h + \gamma \sum_{n=1}^{k} \tau_n\, |\Delta^h U_\varepsilon^n|_h^2 \leq \gamma\,(G_\varepsilon(U_\varepsilon^0), 1)^h$$
$$+ \sum_{n=1}^{k} \left[\, 4\,\varepsilon^{-1}\,\gamma\,|U_\varepsilon^n - U_\varepsilon^{n-1}|_0^2 + \alpha\,\tau_n\,|(\nabla\Phi_\varepsilon^n, \nabla U_\varepsilon^{n-1})|\,\right] + \gamma^{-1}\, t_k \max_{n=0\to k} \|U_\varepsilon^n\|_1^2$$
$$+ \left[\varepsilon^{-1} \sum_{n=1}^{k} \tau_n\, |\,[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac{1}{2}}\,\nabla W_\varepsilon^n|_0^2\right]^{\frac{1}{2}} \left[\sum_{n=1}^{k} \tau_n\, |U_\varepsilon^n - U_\varepsilon^{n-1}|_1^2\right]^{\frac{1}{2}}. \tag{2.47}$$

The desired result (2.44) then follows from (2.47), (2.29), (1.15), (2.42), and (2.43). ∎

LEMMA 2.7. *Let $u^0 \in K$ and the assumptions* (A) *hold. On choosing either $U_\varepsilon^0 \equiv Q^h u^0$ or, if $u^0 \in W^{1,p}(\Omega)$ with $p > 2$, $U_\varepsilon^0 \equiv \pi^h u^0$, it follows that $U_\varepsilon^0 \in K^h$ is such that for all $h > 0$*

$$\|U_\varepsilon^0\|_1^2 + (G_\varepsilon(U_\varepsilon^0), 1)^h \leq C. \tag{2.48}$$

Proof. The desired result (2.48) follows immediately from (2.20), (2.21), (2.17), (2.7), and (2.6). ∎

Remark 2.8. As an alternative to the approximation $(\mathrm{P}_\varepsilon^{h,\tau})$ of (P) one could consider $(\overline{\mathrm{P}}_\varepsilon^{h,\tau})$, which is the same as $(\mathrm{P}_\varepsilon^{h,\tau})$ but with $\Lambda_\varepsilon(U_\varepsilon^{n-1})$ in (2.13a) replaced by $\Lambda_\varepsilon(U_\varepsilon^n)$ and $\Xi_\varepsilon(U_\varepsilon^{n-1})$ in (2.13b) replaced by $\Xi_\varepsilon(U_\varepsilon^n)$. This is more in line with the approximation of the thin film equation in [17]. This has the advantage that to prove the key energy bound (2.44) one can choose $\chi \equiv \pi^h[F_\varepsilon'(U_\varepsilon^n)]$ and $\chi \equiv \pi^h[G_\varepsilon'(U_\varepsilon^n)]$ in the modified versions of (2.13a) and (2.13b), respectively. This would simplify the proof of (2.44) and in particular remove the term $\varepsilon^{-1}\tau^{\frac{1}{2}}$ on the right-hand side. The presence of this term for our chosen scheme $(\mathrm{P}_\varepsilon^{h,\tau})$ leads to the constraint $\tau \leq C\,\varepsilon^2$ for our convergence results; see Lemma 3.1. However, the scheme $(\overline{\mathrm{P}}_\varepsilon^{h,\tau})$ has the severe disadvantage that the well posedness and computation of $\{\Phi_\varepsilon^n, U_\varepsilon^n, W_\varepsilon^n\}$ is nontrivial, since they are coupled in a highly nonlinear system of equations.

Remark 2.9. Also in line with the approximation of the thin film equation in [17], one could remove the inequality constraint in (2.13c) for either of the approximations in Remark 2.8. In particular, it follows from (2.7) that

$$\varepsilon^{-1}\,(|s| - (1-\varepsilon))^2 \leq 4\,G_\varepsilon(s) \qquad \forall\ |s| \geq 1-\varepsilon. \tag{2.49}$$

On combining (2.49) with the energy bound (2.44), which still holds, one has control, in terms of ε, on the overshoot of U_ε^n from $\mathcal{K}$ in $|\cdot|_h$. As the inequality constraint in (2.13c) does not lead to any theoretical or computational complications, we prefer to impose it so one can clearly identify the three computational regions: conductor $U_\varepsilon^n = +1$, interface $|U_\varepsilon^n| < 1$, and void $U_\varepsilon^n = -1$.

Remark 2.10. The approximation $(\mathrm{P}_\varepsilon^{h,\tau})$ of (P) and all the variants mentioned in Remarks 2.8 and 2.9 require solving for $\{\Phi_\varepsilon^n, U_\varepsilon^n, W_\varepsilon^n\}$ over the whole domain Ω, due to the nondegeneracy of $\Lambda_\varepsilon(\cdot)$ and $\Xi_\varepsilon(\cdot)$; see (2.26a), (2.26b). For computational speed it would be more convenient to solve for Φ_ε^n just in the conductor and interfacial regions, $U_\varepsilon^{n-1} > -1$, and for $\{U_\varepsilon^n, W_\varepsilon^n\}$ just in the interfacial region, $|U_\varepsilon^{n-1}| < 1$. With this in mind, and adopting the notation (2.10) and (2.11), we introduce $\widetilde{\Lambda}_\varepsilon, \widetilde{\Xi}_\varepsilon : S^h \to [L^\infty(\Omega)]^{2\times 2}$ such that $\widetilde{\Lambda}_\varepsilon(z^h)\,|_\sigma := R_\sigma \widehat{\Lambda}_\varepsilon^\star(\widehat{z}^h)\,|_{\widehat{\sigma}}\ R_\sigma^T$ and $\widetilde{\Xi}_\varepsilon(z^h)\,|_\sigma := R_\sigma \widehat{\Xi}_\varepsilon^\star(\widehat{z}^h)\,|_{\widehat{\sigma}}\ R_\sigma^T$, where

$$[\widehat{\Lambda}_\varepsilon^\star(\widehat{z}^h)\,|_{\widehat{\sigma}}]_{kk} := \begin{cases} 0 & \text{if } \widehat{z}^h(\widehat{p}_k) = \widehat{z}^h(\widehat{p}_0) = -1, \\ [\widehat{\Lambda}_\varepsilon(\widehat{z}^h)\,|_{\widehat{\sigma}}]_{kk} & \text{otherwise;} \end{cases}$$

$$\text{and} \quad [\widehat{\Xi}_\varepsilon^\star(\widehat{z}^h)\,|_{\widehat{\sigma}}]_{kk} := \begin{cases} 0 & \text{if } \widehat{z}^h(\widehat{p}_k) = \widehat{z}^h(\widehat{p}_0) = \pm 1, \\ [\widehat{\Xi}_\varepsilon(\widehat{z}^h)\,|_{\widehat{\sigma}}]_{kk} & \text{otherwise.} \end{cases}$$

We note that the key identities, $\Lambda_\varepsilon(z^h)$ in (2.9a), (2.9b) replaced by $\widetilde{\Lambda}_\varepsilon(z^h)$ and $\Xi_\varepsilon(z^h)$ in (2.12a), (2.12b) replaced by $\widetilde{\Xi}_\varepsilon(z^h)$, still hold. We then introduce the approximation $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ of (P), which is the same as $(\mathrm{P}_\varepsilon^{h,\tau})$ but with $\Lambda_\varepsilon(U_\varepsilon^{n-1})$ in (2.13a) replaced by $\widetilde{\Lambda}_\varepsilon(U_\varepsilon^{n-1})$ and $\Xi_\varepsilon(U_\varepsilon^{n-1})$ in (2.13b) replaced by $\widetilde{\Xi}_\varepsilon(U_\varepsilon^{n-1})$. As $\widetilde{\Lambda}_\varepsilon(\cdot)$ and $\widetilde{\Xi}_\varepsilon(\cdot)$ are

now degenerate, existence of a solution $\{\Phi_\varepsilon^n, U_\varepsilon^n, W_\varepsilon^n\}$ to $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ does not appear to be trivial. However, this can easily be established by splitting the nodes into passive and active sets; see, e.g., [3]. Moreover, one can show that U_ε^n is unique, $\Phi_\varepsilon^n(p_j)$ is unique if $(\widetilde{\Lambda}_\varepsilon(U_\varepsilon^{n-1}), \chi_j) > 0$, and $W_\varepsilon^n(p_j)$ is unique if $(\widetilde{\Xi}_\varepsilon(U_\varepsilon^{n-1}), \chi_j) > 0$. Furthermore, one can establish analogues of the energy estimates (2.42) and (2.43). Unfortunately, it does not appear possible to establish an analogue of the key energy estimate (2.44) for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$.

3. Convergence. Let

$$U_\varepsilon(t) := \tfrac{t-t_{n-1}}{\tau_n}\, U_\varepsilon^n + \tfrac{t_n-t}{\tau_n}\, U_\varepsilon^{n-1}, \qquad t \in [t_{n-1}, t_n], \quad n \geq 1, \tag{3.1a}$$

$$U_\varepsilon^+(t) := U_\varepsilon^n, \qquad U_\varepsilon^-(t) := U_\varepsilon^{n-1}, \qquad t \in (t_{n-1}, t_n], \qquad n \geq 1. \tag{3.1b}$$

We note for future reference that

$$U_\varepsilon - U_\varepsilon^\pm = (t - t_n^\pm)\, \tfrac{\partial U_\varepsilon}{\partial t}, \qquad t \in (t_{n-1}, t_n), \quad n \geq 1, \tag{3.2}$$

where $t_n^+ := t_n$ and $t_n^- := t_{n-1}$. We introduce also

$$\bar\tau(t) := \tau_n, \qquad t \in (t_{n-1}, t_n], \quad n \geq 1. \tag{3.3}$$

Using the above notation and introducing analogous notation for W_ε^+ and Φ_ε^+, $(\mathrm{P}_\varepsilon^{h,\tau})$ can be restated as follows: Find $\{\Phi_\varepsilon^+, U_\varepsilon, W_\varepsilon^+\} \in L^\infty(0,T;S^h) \times C([0,T];K^h) \times L^\infty(0,T;S^h)$ such that for all $\chi \in L^\infty(0,T;S^h)$ and $z^h \in L^\infty(0,T;K^h)$

$$\int_0^T (\Lambda_\varepsilon(U_\varepsilon^-)\, \nabla\Phi_\varepsilon^+, \nabla\chi)\, \mathrm{d}t + \int_0^T \int_{\partial_2\Omega} \Phi_\varepsilon^+\, \chi\, \mathrm{d}s\, \mathrm{d}t = \int_0^T \int_{\partial_2\Omega} g\, \chi\, \mathrm{d}s\, \mathrm{d}t, \tag{3.4a}$$

$$\int_0^T \Big[\gamma\, \big(\tfrac{\partial U_\varepsilon}{\partial t}, \chi\big)^h + (\Xi_\varepsilon(U_\varepsilon^-)\, \nabla[W_\varepsilon^+ + \alpha\, \Phi_\varepsilon^+], \nabla\chi)\Big]\, \mathrm{d}t = 0, \tag{3.4b}$$

$$\gamma \int_0^T (\nabla U_\varepsilon^+, \nabla[z^h - U_\varepsilon^+])\, \mathrm{d}t \geq \int_0^T (W_\varepsilon^+ + \gamma^{-1}\, U_\varepsilon^-, z^h - U_\varepsilon^+)^h\, \mathrm{d}t. \tag{3.4c}$$

LEMMA 3.1. *Let $u^0 \in K$ with $\fint u^0 \in (-1,1)$. Let $\{\mathcal{T}^h, U_\varepsilon^0, \{\tau_n\}_{n=1}^N, \varepsilon\}_{h>0}$ be such that*

(i) *either $U_\varepsilon^0 \equiv Q^h u^0$ or $U_\varepsilon^0 \equiv \pi^h u^0$ if $u^0 \in W^{1,p}(\Omega)$ with $p > 2$;*
(ii) *Ω and $\{\mathcal{T}^h\}_{h>0}$ fulfill assumptions* (A)*, $\varepsilon \in (0,1)$ with $\varepsilon \to 0$ as $h \to 0$, and $\tau_n \leq C\, \tau_{n-1} \leq C\, \varepsilon^2$, $n = 2 \to N$.*

Then there exists a subsequence of $\{\Phi_\varepsilon^+, U_\varepsilon, W_\varepsilon^+\}_h$, where $\{\Phi_\varepsilon^+, U_\varepsilon, W_\varepsilon^+\}$ solve $(\mathrm{P}_\varepsilon^{h,\tau})$, and a function

$$u \in L^\infty(0,T;K) \cap H^1(0,T;(H^1(\Omega))') \tag{3.5}$$

with $u(\cdot,0) = u^0(\cdot)$ in $L^2(\Omega)$ and $\fint u(\cdot,t) = \fint u^0$ for a.a. $t \in (0,T)$ such that as $h \to 0$

$$U_\varepsilon,\ U_\varepsilon^\pm \to u \qquad \text{weak-}* \text{ in } L^\infty(0,T;H^1(\Omega)), \tag{3.6a}$$

$$\mathcal{G}\tfrac{\partial U_\varepsilon}{\partial t} \to \mathcal{G}\tfrac{\partial u}{\partial t} \qquad \text{weakly in } L^2(0,T;H^1(\Omega)), \tag{3.6b}$$

$$U_\varepsilon,\ U_\varepsilon^\pm \to u \qquad \text{strongly in } L^2(0,T;L^s(\Omega)), \tag{3.7a}$$

$$\Xi_\varepsilon(U_\varepsilon^-) \to b(u)\, \mathcal{I} \qquad \text{strongly in } L^2(0,T;L^s(\Omega)), \tag{3.7b}$$

$$\Lambda_\varepsilon(U_\varepsilon^-) \to c(u)\, \mathcal{I} \qquad \text{strongly in } L^2(0,T;L^s(\Omega)) \tag{3.7c}$$

for all $s \in [2,\infty)$. *If, in addition,* $u^0 \in H^2(\Omega)$ *with* $\frac{\partial u^0}{\partial \nu} = 0$ *on* $\partial\Omega$, $U_\varepsilon^0 \equiv \pi^h u^0$, *and*

$$\alpha^2 \int_0^T |\pi^h [F_\varepsilon'(U_\varepsilon^-)]|^2_{0,\partial_2\Omega} \,\mathrm{d}t \le C, \tag{3.8}$$

then u, *in addition to* (3.5), *satisfies*

$$u \in L^2(0,T;H^2(\Omega)) \tag{3.9}$$

and there exists a subsequence of $\{\Phi_\varepsilon^+, U_\varepsilon, W_\varepsilon^+\}_h$ *satisfying* (3.6a), (3.6b), (3.7a)–(3.7c) *and as* $h \to 0$

$$\Delta^h U_\varepsilon,\ \Delta^h U_\varepsilon^\pm \to \Delta u \qquad \text{weakly in } L^2(\Omega_T), \tag{3.10a}$$

$$U_\varepsilon,\ U_\varepsilon^\pm \to u \qquad \text{weakly in } L^2(0,T;W^{1,s}(\Omega)) \quad \text{for any } s \in [2,\infty), \tag{3.10b}$$

$$U_\varepsilon \to u \qquad \text{strongly in } L^2(0,T;C^{0,\beta}(\overline{\Omega})) \quad \text{for any } \beta \in (0,1). \tag{3.10c}$$

Proof. Noting the definitions (3.1a), (3.1b) and (3.3), the bounds in (2.28), (2.42), and (2.43), together with (1.14), (2.48) and our assumption (i), imply that

$$\begin{aligned} &\| [\Lambda_\varepsilon(U_\varepsilon^-)]^{\frac12} \nabla \Phi_\varepsilon^+ \|^2_{L^2(\Omega_T)} + \|\Phi_\varepsilon^+\|^2_{L^2(0,T;L^2(\partial_2\Omega))} + \|U_\varepsilon^{(\pm)}\|^2_{L^\infty(0,T;H^1(\Omega))} \\ &\quad + \|\bar\tau^{\frac12} \tfrac{\partial U_\varepsilon}{\partial t}\|^2_{L^2(0,T;H^1(\Omega))} + \| [\Xi_\varepsilon(U_\varepsilon^-)]^{\frac12} \nabla W_\varepsilon^+ \|^2_{L^2(\Omega_T)} + \|\mathcal{G} \tfrac{\partial U_\varepsilon}{\partial t}\|^2_{L^2(0,T;H^1(\Omega))} \\ &\quad + \tau^{-\frac12} \|\bar\tau^{\frac12} \tfrac{\partial U_\varepsilon}{\partial t}\|^2_{L^2(\Omega_T)} \le C. \end{aligned} \tag{3.11}$$

Furthermore, we deduce from (3.2) and (3.11) that

$$\|U_\varepsilon - U_\varepsilon^\pm\|^2_{L^2(0,T;H^1(\Omega))} \le \|\bar\tau \tfrac{\partial U_\varepsilon}{\partial t}\|^2_{L^2(0,T;H^1(\Omega))} \le C\,\tau\,. \tag{3.12}$$

Hence on noting (3.11), (3.12), $U_\varepsilon(\cdot,t) \in K^h$, and (1.12a) we can choose a subsequence $\{\Phi_\varepsilon^+, U_\varepsilon, W_\varepsilon^+\}_h$ such that the convergence results (3.5), (3.6a), (3.6b), and (3.7a) hold. Then (3.5) and Theorem 2.6 yield, on noting (1.12b), assumption (i), (2.21), and (2.17), that the subsequence satisfies the additional initial and integral conditions.

We now prove (3.7b). We have that

$$\begin{aligned} \|b(u)\,\mathcal{I} - \Xi_\varepsilon(U_\varepsilon^-)\|_{L^2(0,T;L^s(\Omega))} &\le \|b(u) - b(U_\varepsilon^-)\|_{L^2(0,T;L^s(\Omega))} \\ &\quad + \|b(U_\varepsilon^-) - b_\varepsilon(U_\varepsilon^-)\|_{L^2(0,T;L^s(\Omega))} + \|b_\varepsilon(U_\varepsilon^-)\,\mathcal{I} - \Xi_\varepsilon(U_\varepsilon^-)\|_{L^2(0,T;L^s(\Omega))}. \end{aligned} \tag{3.13}$$

Noting the Lipschitz continuity of b on $\mathcal{K}$, (2.27b), (2.15), and (3.11), we have that

$$\begin{aligned} &\|b(u) - b(U_\varepsilon^-)\|_{L^2(0,T;L^s(\Omega))} + \|b_\varepsilon(U_\varepsilon^-)\,\mathcal{I} - \Xi_\varepsilon(U_\varepsilon^-)\|_{L^2(0,T;L^s(\Omega))} \\ &\qquad \le 2\,\|u - U_\varepsilon^-\|_{L^2(0,T;L^s(\Omega))} + C\,h^{\frac{2}{s}}\,\|\nabla U_\varepsilon^-\|_{L^2(\Omega_T)} \\ &\qquad \le 2\,\|u - U_\varepsilon^-\|_{L^2(0,T;L^s(\Omega))} + C\,h^{\frac{2}{s}}. \end{aligned} \tag{3.14}$$

It follows from (2.25b) and (1.6) that

$$\|b(U_\varepsilon^-) - b_\varepsilon(U_\varepsilon^-)\|_{L^2(0,T;L^s(\Omega))} \le C\,b_\varepsilon(1) \le C\,\varepsilon\,. \tag{3.15}$$

Combining (3.13), (3.14), (3.15) and noting (3.7a) and our assumption (ii) on ε yield the desired result (3.7b). A similar argument to the above yields the desired result (3.7c).

We now prove the results (3.10a)–(3.10c). It follows from (2.3), (2.18), (2.23), (2.16), our assumptions on u^0, and (2.14) that

$$\begin{aligned}|\Delta^h U_\varepsilon^0|_0^2 &= |\Delta^h(\pi^h u^0)|_0^2 \leq |\Delta^h(\pi^h u^0)|_h^2 = -(\nabla(\pi^h u^0), \nabla(\Delta^h(\pi^h u^0)))\\ &= -(\nabla u^0, \nabla(\Delta^h(\pi^h u^0))) + (\nabla(I-\pi^h)u^0, \nabla(\Delta^h(\pi^h u^0)))\\ &\leq |\Delta u^0|_0\, |\Delta^h(\pi^h u^0)|_0 + C\, h\, |u^0|_2\, |\nabla(\Delta^h(\pi^h u^0))|_0 \leq C\, |u^0|_2^2 \leq C. \end{aligned} \tag{3.16}$$

Moreover, (2.44), (2.48), (3.16), (2.3), (2.18), (3.1a), (3.1b), and our assumption (ii) on $\{\tau_n\}_{n=1}^N$ yield that

$$\|\Delta^h U_\varepsilon^{(\pm)}\|_{L^2(\Omega_T)} \leq C\,. \tag{3.17}$$

From (3.17), (2.23), (2.17), (2.19), (3.11), and (3.6a) we have for any $\eta \in L^2(0,T; W^{1,q}(\Omega))$, $q > 2$, that

$$\begin{aligned}\int_0^T (\Delta^h U_\varepsilon^{(\pm)}, \eta)\,\mathrm{d}t &= \int_0^T (\Delta^h U_\varepsilon^{(\pm)}, (I-\pi^h)\eta)\,\mathrm{d}t\\ &\quad + \int_0^T \Big[\,(\Delta^h U_\varepsilon^{(\pm)}, \pi^h\,\eta) - (\Delta^h U_\varepsilon^{(\pm)}, \pi^h\,\eta)^h\,\Big]\,\mathrm{d}t\\ &\quad + \int_0^T (\nabla U_\varepsilon^{(\pm)}, \nabla(I-\pi^h)\eta)\,\mathrm{d}t - \int_0^T (\nabla U_\varepsilon^{(\pm)}, \nabla\eta)\,\mathrm{d}t\\ &\to -\int_0^T (\nabla u, \nabla\eta)\,\mathrm{d}t \qquad \text{as } h \to 0\,. \end{aligned} \tag{3.18}$$

Combining (3.17), (3.18), and the denseness of $L^2(0,T;W^{1,q}(\Omega))$ in $L^2(\Omega_T)$ yields (3.10a) and, in particular, $\Delta u \in L^2(\Omega_T)$. This, together with elliptic regularity, as Ω is a rectangle, and (3.5), proves (3.9). Furthermore, it follows from (3.10a) and (2.24) that (3.10b) holds on extracting a further subsequence. Finally, (3.10c) follows from (3.10b), (3.6b), (1.12a), and the compact embedding $W^{1,s}(\Omega) \hookrightarrow C^{0,\beta}(\overline{\Omega})$. ∎

Remark 3.2. The conditions $u^0 \in H^2(\Omega)$ with $\frac{\partial u^0}{\partial \nu} = 0$ on $\partial\Omega$ for the results (3.10a)–(3.10c) can be replaced by a restriction on τ_1 in terms of h (see [7, Lemma 3.1]), but they are not particularly restrictive. The assumption (3.8) holds if $U_\varepsilon(x,t) = 1$ for all $x \in \partial_2\Omega$ and $t \in [0,T]$, and this condition held in all our numerical experiments, provided $u^0 = 1$ on $\partial_2\Omega$ and either L_1 is chosen sufficiently large or T is chosen sufficiently small. This can be made rigorous for the approximation $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ (see Remark 2.10), as the degeneracy of $\widetilde{\Xi}_\varepsilon$ leads to finite speed of propagation of the numerical interfacial region; at each time level it can move locally at most one mesh point; see [3].

In addition to the above lemma, we need the following two lemmas in order to prove our main result, Theorem 3.6 below.

LEMMA 3.3. *Let all the assumptions of Lemma* 3.1 *hold. If, in addition,* $\tau_n = \tau$, $n = 1 \to N$, *then*

$$\int_0^{T-\theta} |U_\varepsilon^\pm(t+\theta) - U_\varepsilon^\pm(t)|_0^2\,\mathrm{d}t \leq C\,\theta \qquad \forall\ \theta \in (0,T). \tag{3.19}$$

Moreover, it holds that the subsequence of $\{\Phi_\varepsilon^+, U_\varepsilon, W_\varepsilon^+\}_h$ *in Lemma* 3.1 *is such that for any* $\beta \in (0,1)$

$$U_\varepsilon^\pm \to u \qquad \textit{strongly in } L^2(0,T;C^{0,\beta}(\overline{\Omega})) \qquad \textit{as } h \to 0; \tag{3.20a}$$

and, on extracting a further subsequence, it holds for a.a. $t \in (0,T)$ *that*

$$U_\varepsilon^\pm(\cdot,t) \to u(\cdot,t) \qquad \text{strongly in } C^{0,\beta}(\overline{\Omega}) \qquad \text{as } h \to 0. \tag{3.20b}$$

Proof. The proof is similar to the proofs in [16, Lemmas 7.1 and 8.8]; see also [7, Lemma 3.2]. It follows from (2.13b) for $m = 0 \to N-l$, $l \in \{1,\dots,N\}$ fixed, that

$$\begin{aligned} &\gamma \sum_{n=m+1}^{m+l} \tau_n \left(\tfrac{U_\varepsilon^n - U_\varepsilon^{n-1}}{\tau_n}, U_\varepsilon^{m+l} - U_\varepsilon^m\right)^h \\ &\qquad = -\sum_{n=m+1}^{m+l} \tau_n \left(\Xi_\varepsilon(U_\varepsilon^{n-1})\, \nabla[W_\varepsilon^n + \alpha\, \Phi_\varepsilon^n], \nabla(U_\varepsilon^{m+l} - U_\varepsilon^m)\right). \end{aligned} \tag{3.21}$$

Similarly to (2.46), we obtain from (3.21) and (2.26b), (2.26c) that

$$\gamma\, |U_\varepsilon^{m+l} - U_\varepsilon^m|_h^2 \le \sum_{n=m+1}^{m+l} \tau_n\, a_n\, |U_\varepsilon^{m+l} - U_\varepsilon^m|_1 = \sum_{k=1}^{l} \tau_{m+k}\, a_{m+k}\, |U_\varepsilon^{m+l} - U_\varepsilon^m|_1, \tag{3.22}$$

where $a_n := |[\Xi_\varepsilon(U_\varepsilon^{n-1})]^{\frac12}\, \nabla W_\varepsilon^n|_0 + 2^{\frac12}\, \alpha\, |[\Lambda_\varepsilon(U_\varepsilon^{n-1})]^{\frac12}\, \nabla \Phi_\varepsilon^n|_0$. Summing (3.22) for $m = 0 \to N-l$ and using the uniform time step assumption yield on noting (2.28), (2.42), and (2.48) that

$$\begin{aligned} \gamma \sum_{m=0}^{N-l} \tau\, |U_\varepsilon^{m+l} - U_\varepsilon^m|_h^2 &\le \sum_{k=1}^{l} \tau \sum_{m=0}^{N-l} \tau\, a_{m+k}\, |U_\varepsilon^{m+l} - U_\varepsilon^m|_1 \\ &\le \sum_{k=1}^{l} \tau \left[\sum_{m=0}^{N-l} \tau\, a_{m+k}^2\right]^{\frac12} \left[\sum_{m=0}^{N-l} \tau\, |U_\varepsilon^{m+l} - U_\varepsilon^m|_1^2\right]^{\frac12} \le C\, l\, \tau. \end{aligned} \tag{3.23}$$

Combining (3.23), (2.3), (2.18), and (3.1b) yields (3.19) for $\theta = l\,\tau$. For arbitrary $\theta \in (0,T)$ with $\theta = \mu\,\tau$, $\mu \in (0,N)$, we argue as follows. On recalling (1.16b), let $l = \lfloor \mu \rfloor$, $\vartheta = \mu - \lfloor \mu \rfloor \in [0,1)$ and $m \in \{0,\dots,N-l\}$ be such that $t \in (m\,\tau, (m+1)\,\tau]$. Hence

$$U_\varepsilon^\pm(t + \mu\,\tau) = \begin{cases} U_\varepsilon^\pm(t + l\,\tau) & \text{if } \ t \in (m\,\tau, m\,\tau + (1-\vartheta)\,\tau], \\ U_\varepsilon^\pm(t + (l+1)\,\tau) & \text{if } \ t \in (m\,\tau + (1-\vartheta)\,\tau, (m+1)\,\tau], \end{cases}$$

and we obtain on noting (3.23) that

$$\begin{aligned} &\gamma \int_0^{T-\mu\,\tau} |U_\varepsilon^\pm(t + \mu\,\tau) - U_\varepsilon^\pm(t)|_h^2 \,\mathrm{d}t \\ &\qquad \le \tau\,(1-\vartheta) \sum_{m=0}^{N-l} |U_\varepsilon^{m+l} - U_\varepsilon^m|_h^2 + \tau\,\vartheta \sum_{m=0}^{N-l-1} |U_\varepsilon^{m+l+1} - U_\varepsilon^m|_h^2 \\ &\qquad \le C\,[(1-\vartheta)\,l + \vartheta\,(l+1)]\,\tau = C\,\mu\,\tau. \end{aligned} \tag{3.24}$$

Combining (3.24), (2.3), and (2.18) yields (3.19) for all $\theta \in (0,T)$. It follows from (3.10b) and (3.19), on noting (1.12c) and the compact embedding $W^{1,s}(\Omega) \hookrightarrow C^{0,\beta}(\overline{\Omega})$, that (3.20a) holds. Finally, the desired result (3.20b) follows immediately from (3.20a). ∎

From (3.11), (2.26a), (2.26b), (2.25a), (2.25b), (1.6), and (3.20b) we see that we can only control $\nabla\Phi_\varepsilon^+$ and ∇W_ε^+ on the sets where $\Lambda_\varepsilon(U_\varepsilon^-)$ and $\Xi_\varepsilon(U_\varepsilon^-)$ are bounded below independently of ε, and hence h on noting (ii), i.e., on the sets where $u > -1$ and $|u| < 1$, respectively. Therefore in order to construct the appropriate limits as $h \to 0$, we introduce the following open subsets of $\overline{\Omega}$. For any $\delta \in (0,1)$, we define for a.a. $t \in (0,T)$

$$B_\delta(t) := \{\, x \in \overline{\Omega} : |u(x,t)| < 1-\delta \,\} \subset D_\delta(t) := \{\, x \in \overline{\Omega} : -1+\delta < u(x,t) \,\}\,. \tag{3.25}$$

From (3.20b) we have that there exist positive constants $C_x(t)$ such that

$$|u(y,t) - u(z,t)| \le C_x(t)\,|y-z|^\beta \qquad \forall\ y,\ z \in \overline{\Omega} \quad \text{for a.a. } t \in (0,T). \tag{3.26}$$

As $\fint u(\cdot,t) = \fint u^0 \in (-1,1)$ for a.a. $t \in (0,T)$, it follows that there exists a $\delta_0 \in (0, 1 - |\fint u^0|\,)$ such that $D_{\delta_0}(t) \supset B_{\delta_0}(t) \not\equiv \emptyset$ for a.a. $t \in (0,T)$. It immediately follows from (3.25) and (3.26) for a.a. $t \in (0,T)$ and for any δ_1, $\delta_2 \in (0,\delta_0)$ with $\delta_1 > \delta_2$ that

$$\begin{aligned} &\text{either} \quad y \in B_{\delta_1}(t) \text{ and } z \in \partial B_{\delta_2}(t) \quad \text{or} \quad y \in D_{\delta_1}(t) \text{ and } z \in \partial D_{\delta_2}(t) \quad \text{with } z \notin \partial\Omega \\ &\quad \Longrightarrow \qquad C_x(t)\,|y-z|^\beta \ge u(y,t) - u(z,t) > (\delta_1 - \delta_2), \end{aligned}$$

where $\partial B_\delta(t)$ and $\partial D_\delta(t)$ are the boundaries of $B_\delta(t)$ and $D_\delta(t)$, respectively. This implies that for a.a. $t \in (0,T)$ and any $\delta \in (0,\delta_0)$ there exists an $h_0(\delta,t)$ such that for all $h \le h_0(\delta,t)$ there exist collections of simplices $\mathcal{T}^h_{B,\delta}(t) \subset \mathcal{T}^h_{D,\delta}(t) \subset \mathcal{T}^h$ such that

$$B_\delta(t) \subset B^h_\delta(t) := \cup_{\sigma\in\mathcal{T}^h_{B,\delta}(t)}\,\overline{\sigma} \subset B_{\frac{\delta}{2}}(t)\,, \qquad D_\delta(t) \subset D^h_\delta(t) := \cup_{\sigma\in\mathcal{T}^h_{D,\delta}(t)}\,\overline{\sigma} \subset D_{\frac{\delta}{2}}(t)\,. \tag{3.27}$$

Clearly, we have from (3.25) that

$$\delta_2 < \delta_1 < \delta_0 \qquad \Longrightarrow \qquad h_0(\delta_2,t) \le h_0(\delta_1,t)\,.$$

For a.a. $t \in (0,T)$ and any fixed $\delta \in (0,\widehat{\delta}_0)$, where $\widehat{\delta}_0 := \min\{\delta_0, \frac{1}{2}\}$, it follows from (3.25), (3.20b), and our assumption (ii) of Lemma 3.1 that there exists an $\widehat{h}_0(\delta,t) \le h_0(\delta,t)$ such that for $h \le \widehat{h}_0(\delta,t)$

$$1 - 2\,\delta \le |U_\varepsilon^\pm(x,t)| \quad \forall\ x \notin B_\delta(t), \quad |U_\varepsilon^\pm(x,t)| \le 1 - \tfrac{\delta}{2} \qquad \forall\ x \in B_\delta(t)\,, \tag{3.28a}$$

$$U_\varepsilon^\pm(x,t) \le -1 + 2\,\delta \quad \forall\ x \notin D_\delta(t), \qquad -1 + \tfrac{\delta}{2} \le U_\varepsilon^\pm(x,t) \quad \forall\ x \in D_\delta(t), \tag{3.28b}$$

and

$$\varepsilon \leq \delta\,. \tag{3.29}$$

LEMMA 3.4. *Let all the assumptions of Lemma* 3.3 *hold. Then for a.a.* $t \in (0,T)$ *there exist functions*

$$\phi(\cdot,t) \in H^1_{loc}(\{u(\cdot,t) > -1\}),\ \ w(\cdot,t) \equiv -\gamma\,\Delta u(\cdot,t) - \gamma^{-1}\,u(\cdot,t) \in H^1_{loc}(\{|u(\cdot,t)| < 1\}), \tag{3.30}$$

where $\{u(\cdot,t) > -1\} := \{x \in \Omega : u(x,t) > -1\}$ *and* $\{|u(\cdot,t)| < 1\} := \{x \in \Omega : |u(x,t)| < 1\}$. *Moreover, on assuming that*

$$u(x,t) = 1 \qquad \forall\ x \in \partial_2\Omega, \quad \text{for a.a. } t \in (0,T)\,, \tag{3.31}$$

and extracting a further subsequence from the subsequence $\{\Phi^+_\varepsilon, U_\varepsilon, W^+_\varepsilon\}_h$ *in Lemma* 3.3, *it holds as* $h \to 0$ *that*

$$\Phi^+_\varepsilon \to \phi \qquad \text{weakly in } L^2(0,T;L^2(\partial_2\Omega))\,, \tag{3.32a}$$

$$\Lambda_\varepsilon(U^-_\varepsilon)\,\nabla\Phi^+_\varepsilon \to \mathcal{H}_{\{u>-1\}}\,c(u)\,\nabla\phi \qquad \text{weakly in } L^2(\Omega_T), \tag{3.32b}$$

$$\Xi_\varepsilon(U^-_\varepsilon)\,\nabla\Phi^+_\varepsilon \to \mathcal{H}_{\{|u|<1\}}\,b(u)\,\nabla\phi \qquad \text{weakly in } L^2(\Omega_T), \tag{3.32c}$$

$$\Xi_\varepsilon(U^-_\varepsilon)\,\nabla W^+_\varepsilon \to \mathcal{H}_{\{|u|<1\}}\,b(u)\,\nabla w \qquad \text{weakly in } L^2(\Omega_T), \tag{3.32d}$$

where $\mathcal{H}_{\{u>-1\}}$ *and* $\mathcal{H}_{\{|u|<1\}}$ *are the characteristic functions of the sets* $\{u > -1\} := \{(x,t) \in \Omega_T : u(x,t) > -1\}$ *and* $\{|u| < 1\} := \{(x,t) \in \Omega_T : |u(x,t)| < 1\}$, *respectively.*

Proof. It follows from (3.11) and (2.26a)–(2.26c) that

$$\|\Lambda_\varepsilon(U^-_\varepsilon)\nabla\Phi^+_\varepsilon\|^2_{L^2(\Omega_T)} + \|\Xi_\varepsilon(U^-_\varepsilon)\nabla\Phi^+_\varepsilon\|^2_{L^2(\Omega_T)} + \|\Xi_\varepsilon(U^-_\varepsilon)\nabla W^+_\varepsilon\|^2_{L^2(\Omega_T)} \leq C. \tag{3.33}$$

Hence (3.33) implies that there exist functions $\underline{z}_i \in L^2(\Omega_T)$, $i = 1 \to 3$, and, on extracting a further subsequence from the subsequence $\{\Phi^+_\varepsilon, U_\varepsilon, W^+_\varepsilon\}_h$ in Lemma 3.3, it holds as $h \to 0$ that

$$\Lambda_\varepsilon(U^-_\varepsilon)\nabla\Phi^+_\varepsilon \to \underline{z}_1\,,\ \Xi_\varepsilon(U^-_\varepsilon)\nabla\Phi^+_\varepsilon \to \underline{z}_2\,,\ \Xi_\varepsilon(U^-_\varepsilon)\nabla W^+_\varepsilon \to \underline{z}_3 \quad \text{weakly in } L^2(\Omega_T)\,. \tag{3.34}$$

We now identify the functions $\underline{z}_i$.

First, we consider a fixed $\delta \in (0, \delta_0)$. It follows from (1.6), (2.25a), (2.25b), (2.26a), (2.26b), (3.28a), (3.28b), and (3.11) that for a.a. $t \in (0,T)$ and for all $h \leq \widehat{h}_0(\delta, t)$

$$\begin{aligned}
\tfrac{\delta}{2}\,|\nabla\Phi_\varepsilon^+(\cdot,t)|^2_{0,D_\delta(t)} &= c(-1+\tfrac{\delta}{2})\,|\nabla\Phi_\varepsilon^+(\cdot,t)|^2_{0,D_\delta(t)} \\
&\leq c_\varepsilon(-1+\tfrac{\delta}{2})\,|\nabla\Phi_\varepsilon^+(\cdot,t)|^2_{0,D_\delta(t)} \\
&\leq |\,(\,[\Lambda_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla\Phi_\varepsilon^+)(\cdot,t)|^2_0 \leq C(t)\,, \qquad (3.35a)\\
\delta\,(1-\tfrac{\delta}{4})\,|\nabla W_\varepsilon^+(\cdot,t)|^2_{0,B_\delta(t)} &= b(1-\tfrac{\delta}{2})\,|\nabla W_\varepsilon^+(\cdot,t)|^2_{0,B_\delta(t)} \\
&\leq b_\varepsilon(1-\tfrac{\delta}{2})\,|\nabla W_\varepsilon^+(\cdot,t)|^2_{0,B_\delta(t)} \\
&\leq |\,(\,[\Xi_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla W_\varepsilon^+)(\cdot,t)|^2_0 \leq C(t)\,. \qquad (3.35b)
\end{aligned}$$

From (3.35a), (3.35b), (3.27), (2.26a)–(2.26c), (3.28a), (3.28b), and (3.29) we have for a.a. $t \in (0,T)$ and for all $h \leq \widehat{h}_0(\delta,t)$

$$\begin{aligned}
&|\,(\Lambda_\varepsilon(U_\varepsilon^-)\,\nabla\Phi_\varepsilon^+)(\cdot,t)|^2_{0,\Omega\setminus D_\delta(t)} \\
&\qquad\leq \max_{x\in\Omega\setminus D_{2\delta}(t)} c_\varepsilon(U_\varepsilon^-(x))\,|\,(\,[\Lambda_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla\Phi_\varepsilon^+)(\cdot,t)|^2_{0,\Omega\setminus D_\delta(t)} \\
&\qquad\leq C(t)\,c_\varepsilon(-1+4\,\delta) \leq C(t)\,\max\{4\delta,\varepsilon\} \leq C(t)\,\delta\,, \qquad (3.36a)\\
&|\,(\Xi_\varepsilon(U_\varepsilon^-)\,\nabla\Phi_\varepsilon^+)(\cdot,t)|^2_{0,\Omega\setminus B_\delta(t)} \\
&\qquad\leq \max_{x\in\Omega\setminus B_{2\delta}(t)} 2\,b_\varepsilon(U_\varepsilon^-(x))\,|\,(\,[\Lambda_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla\Phi_\varepsilon^+)(\cdot,t)|^2_{0,\Omega\setminus B_\delta(t)} \\
&\qquad\leq C(t)\,b_\varepsilon(1-4\,\delta) \leq C(t)\,\max\{4\delta,\varepsilon\} \leq C(t)\,\delta\,, \qquad (3.36b)\\
&|\,(\Xi_\varepsilon(U_\varepsilon^-)\,\nabla W_\varepsilon^+)(\cdot,t)|^2_{0,\Omega\setminus B_\delta(t)} \\
&\qquad\leq \max_{x\in\Omega\setminus B_{2\delta}(t)} b_\varepsilon(U_\varepsilon^-(x))\,|\,(\,[\Xi_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla W_\varepsilon^+)(\cdot,t)|^2_{0,\Omega\setminus B_\delta(t)} \\
&\qquad\leq C(t)\,b_\varepsilon(1-4\,\delta) \leq C(t)\,\delta\,. \qquad (3.36c)
\end{aligned}$$

From (3.35a) and appealing to the de Rham theorem (see, e.g., [27, p. 10]), we have that there exists a function $\phi \in L^2_{loc}(D_\delta(t))$ with $\nabla\phi \in L^2(D_\delta(t))$ and, on extracting a further subsequence, such that for a.a. $t \in (0,T)$

$$\nabla\Phi_\varepsilon^+(\cdot,t) \to \nabla\phi(\cdot,t) \quad \text{weakly in } L^2(D_\delta(t)) \quad \text{as } h \to 0. \qquad (3.37)$$

In order to show (3.32a) we proceed as follows. On recalling (3.20b) and (3.31) we define open sets $D_L^\pm(t) \subset D_\delta(t)$ with Lipschitz boundaries $\partial D_L^\pm(t)$ such that $\partial_2^\pm\Omega \subset \partial D_L^\pm(t)$. Since $D_L^\pm(t)$ is a Lipschitz domain, Friedrich's inequality, (3.11), and a corresponding $D_L^\pm(t)$ version of (3.35a) yield that for a.a. $t \in (0,T)$ and for all $h \leq \widehat{h}_0(\delta,t)$

$$\|\Phi_\varepsilon^+(\cdot,t)\|^2_{1,D_L^\pm(t)} \leq C(t)\,[\,|\nabla\Phi_\varepsilon^+(\cdot,t)|^2_{0,D_L^\pm(t)} + |\Phi_\varepsilon^+(\cdot,t)|^2_{0,\partial_2^\pm\Omega}\,] \leq C(t)\,. \qquad (3.38)$$

Combining (3.38) and (3.37) yields that

$$\Phi_\varepsilon^+(\cdot,t) \to \phi(\cdot,t) \quad \text{weakly in } H^1(D_L^\pm(t)) \quad \text{as } h \to 0. \qquad (3.39)$$

Then we define the following elliptic operators $\mathcal{F}^\pm : L^2(\partial_2^\pm\Omega) \to H^1(D_L^\pm(t))$:

$$\int_{D_L^\pm(t)} \left[\nabla[\mathcal{F}^\pm z].\nabla\eta + [\mathcal{F}^\pm z]\,\eta\right]\,\mathrm{d}x = \int_{\partial_2^\pm\Omega} z\,\eta\,\mathrm{d}s \qquad \forall\ \eta \in H^1(D_L^\pm(t)).$$

On noting that $\partial_2^\pm\Omega \subset \partial D_L^\pm(t)$, we have from (3.37) and (3.39) for a.a. $t \in (0,T)$ and any $\eta \in L^2(\partial_2^\pm\Omega)$ that

$$\int_{\partial_2^\pm\Omega} \Phi_\varepsilon^+(\cdot,t)\,\eta\,\mathrm{d}s = \int_{D_L^\pm(t)} \left[\nabla\Phi_\varepsilon^+(\cdot,t).\nabla[\mathcal{F}^\pm\eta] + \Phi_\varepsilon^+(\cdot,t)\,\mathcal{F}^\pm\eta\right]\,\mathrm{d}x$$

$$\to \int_{D_L^\pm(t)} \left[\nabla\phi(\cdot,t).\nabla[\mathcal{F}^\pm\eta] + \phi(\cdot,t)\,\mathcal{F}^\pm\eta\right]\,\mathrm{d}x = \int_{\partial_2^\pm\Omega} \phi(\cdot,t)\,\eta\,\mathrm{d}s\,. \tag{3.40}$$

Combining (3.11) and (3.40) yields the desired result (3.32a).

On noting (3.17) we have for a.a. $t \in (0,T)$ that

$$|\Delta^h U_\varepsilon^+(\cdot,t)|_0 \le C(t)\,.$$

Similarly to (3.18) this yields for a.a. $t \in (0,T)$ that as $h \to 0$

$$\Delta^h U_\varepsilon^+(\cdot,t) \to \Delta u(\cdot,t) \qquad \text{weakly in } L^2(\Omega). \tag{3.41}$$

Combining (2.39), (2.23), (3.1b), (3.28a), and (3.27) yields for a.a. $t \in (0,T)$, and for all $h \le \widehat{h}_0(\frac{\delta}{2},t)$ that

$$W_\varepsilon^+(\cdot,t) \equiv -\gamma\,\Delta^h U_\varepsilon^+(\cdot,t) - \gamma^{-1}\,U_\varepsilon^-(\cdot,t) \qquad \text{on } B_\delta(t)\,. \tag{3.42}$$

It follows from (3.42), (3.41), and (3.20b) for a.a. $t \in (0,T)$ that as $h \to 0$

$$W_\varepsilon^+(\cdot,t) \to -\gamma\,\Delta u(\cdot,t) - \gamma^{-1}\,u(\cdot,t) \quad \text{weakly in } L^2(B_\delta(t)).$$

This, together with (3.35b), yields

$$W_\varepsilon^+(\cdot,t) \to w(\cdot,t) \quad \text{weakly in } H^1(B_\delta(t)). \tag{3.43}$$

Combining (3.34), (3.37), (3.43), and (3.7b), (3.7c) yields for a.a. $t \in (0,T)$ that as $h \to 0$

$$(\Lambda_\varepsilon(U_\varepsilon^-)\nabla\Phi_\varepsilon^+)(\cdot,t) \to c(u(\cdot,t))\,\nabla\phi(\cdot,t) \qquad \text{weakly in } L^2(D_\delta(t)), \tag{3.44a}$$

$$(\Xi_\varepsilon(U_\varepsilon^-)\nabla\Phi_\varepsilon^+)(\cdot,t) \to b(u(\cdot,t))\,\nabla\phi(\cdot,t) \qquad \text{weakly in } L^2(B_\delta(t)), \tag{3.44b}$$

$$(\Xi_\varepsilon(U_\varepsilon^-)\nabla W_\varepsilon^+)(\cdot,t) \to b(u(\cdot,t))\,\nabla w(\cdot,t) \qquad \text{weakly in } L^2(B_\delta(t)). \tag{3.44c}$$

Repeating (3.35a), (3.35b)–(3.37) and (3.42)–(3.44a)–(3.44c) for all $\delta \in (0,\widehat{\delta}_0)$ yields, on recalling (3.20b), that (3.30) holds and, on noting (3.36a)–(3.36c) and (3.34), the desired results (3.32b)–(3.32d). ∎

Remark 3.5. The assumption (3.31) is similar to the assumption (3.8); see Remark 3.2.

THEOREM 3.6. *Let the assumptions of Lemma* 3.4 *hold. Then there exists a subsequence of* $\{\Phi_\varepsilon^+, U_\varepsilon, W_\varepsilon^+\}_h$, *where* $\{\Phi_\varepsilon^+, U_\varepsilon, W_\varepsilon^+\}$ *solve* $(\mathrm{P}_\varepsilon^{h,\tau})$, *and functions* $\{\phi,u,w\}$ *satisfying* (3.5), (3.9), *and* (3.30). *In addition, as* $h \to 0$ *the following hold:* (3.6a), (3.6b), (3.7a)–(3.7c), (3.10a)–(3.10c), (3.20a), (3.20b) *for* a.a. $t \in (0,T)$, *and* (3.32a)–(3.32d). *Furthermore, we have that* $\{\phi,u,w\}$ *fulfill* $u(\cdot,0) = u^0(\cdot)$ *in* $L^2(\Omega)$ *and satisfy for all* $\eta \in L^2(0,T;H^1(\Omega))$

$$\int_{\{u>-1\}} c(u)\,\nabla\phi\,.\,\nabla\eta\,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_{\partial_2\Omega} \phi\,\eta\,\mathrm{d}s\,\mathrm{d}t = \int_0^T \int_{\partial_2\Omega} g\,\eta\,\mathrm{d}s\,\mathrm{d}t\,, \tag{3.45a}$$

$$\gamma \int_0^T \langle \tfrac{\partial u}{\partial t}, \eta\rangle\,\mathrm{d}t + \int_{\{|u|<1\}} b(u)\,\nabla\left[w + \alpha\,\phi\right]\,.\,\nabla\eta\,\mathrm{d}x\,\mathrm{d}t = 0\,, \tag{3.45b}$$

where $w(\cdot,t) \equiv -\gamma\,\Delta u(\cdot,t) - \gamma^{-1}u(\cdot,t)$ *on the set* $\{|u(\cdot,t)| < 1\}$ *for a.a.* $t \in (0,T)$.

Proof. For any $\eta \in H^1(0,T;H^2(\Omega))$ we choose $\chi \equiv \pi^h \eta$ in (3.4a), (3.4b) and now analyze the subsequent terms. First, (2.19), the embedding $H^1(0,T;X) \hookrightarrow C([0,T];X)$, (3.11), and (2.16) yield that

$$\begin{aligned}
&\left| \int_0^T \left[\left(\tfrac{\partial U_\varepsilon}{\partial t}, \pi^h \eta\right)^h - \left(\tfrac{\partial U_\varepsilon}{\partial t}, \pi^h \eta\right) \right] \mathrm{d}t \right| \\
&\quad = \left| - \int_0^T \left(U_\varepsilon, \tfrac{\partial (\pi^h \eta)}{\partial t} \right)^h \mathrm{d}t + (U_\varepsilon(\cdot,T), \pi^h \eta(\cdot,T))^h - (U_\varepsilon(\cdot,0), \pi^h \eta(\cdot,0))^h \right. \\
&\qquad \left. + \int_0^T \left(U_\varepsilon, \tfrac{\partial (\pi^h \eta)}{\partial t} \right) \mathrm{d}t - (U_\varepsilon(\cdot,T), \pi^h \eta(\cdot,T)) + (U_\varepsilon(\cdot,0), \pi^h \eta(\cdot,0)) \right| \\
&\quad \le C\, h\, \|U_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))}\, \|\pi^h \eta\|_{H^1(0,T;H^1(\Omega))} \le C\, h\, \|\eta\|_{H^1(0,T;H^2(\Omega))}. \qquad (3.46)
\end{aligned}$$

Furthermore, it follows from (1.13), (3.11), and (2.16) that

$$\begin{aligned}
\left| \int_0^T \left(\tfrac{\partial U_\varepsilon}{\partial t}, (I - \pi^h)\eta \right) \mathrm{d}t \right| &\le C\, \|\mathcal{G} \tfrac{\partial U_\varepsilon}{\partial t}\|_{L^2(0,T;H^1(\Omega))}\, \|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\le C\, h\, \|\eta\|_{L^2(0,T;H^2(\Omega))}. \qquad (3.47)
\end{aligned}$$

Combining (3.46), (3.47), and (3.6b) yields that

$$\int_0^T \left(\tfrac{\partial U_\varepsilon}{\partial t}, \pi^h \eta \right)^h \mathrm{d}t \to \int_0^T \langle \tfrac{\partial u}{\partial t}, \eta \rangle \, \mathrm{d}t \quad \text{as } h \to 0. \qquad (3.48)$$

Moreover, it holds on noting (3.11), g as in (1.8), a trace inequality, and (2.16) that

$$\begin{aligned}
&\left| \int_0^T \int_{\partial_2 \Omega} (\Phi_\varepsilon^+ - g)\, (I - \pi^h)\eta \, \mathrm{d}s \, \mathrm{d}t \right| \\
&\quad \le \left[\|\Phi_\varepsilon^+\|_{L^2(0,T;L^2(\partial_2\Omega))} + \|g\|_{L^2(0,T;L^2(\partial_2\Omega))} \right] \|(I-\pi^h)\eta\|_{L^2(0,T;L^2(\partial_2\Omega))} \\
&\quad \le C\, \|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))} \le C\, h\, \|\eta\|_{L^2(0,T;H^2(\Omega))}. \qquad (3.49)
\end{aligned}$$

In view of (2.26a)–(2.26c), (3.11), and (2.16) we deduce that

$$\begin{aligned}
&\left| \int_0^T (\Xi_\varepsilon(U_\varepsilon^-)\, \nabla W_\varepsilon^+, \nabla (I - \pi^h)\eta) \, \mathrm{d}t \right| \\
&\qquad \le \|\Xi_\varepsilon(U_\varepsilon^-)\, \nabla W_\varepsilon^+\|_{L^2(\Omega_T)}\, \|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad \le \|[\Xi_\varepsilon(U_\varepsilon^-)]^{\frac{1}{2}}\, \nabla W_\varepsilon^+\|_{L^2(\Omega_T)}\, \|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))} \\
&\qquad \le C\, h\, \|\eta\|_{L^2(0,T;H^2(\Omega))} \qquad (3.50a)
\end{aligned}$$

and, similarly,

$$\begin{aligned}
&\left| \int_0^T (\Lambda_\varepsilon(U_\varepsilon^-)\, \nabla \Phi_\varepsilon^+, \nabla (I - \pi^h)\eta) \, \mathrm{d}t \right| + \left| \int_0^T (\Xi_\varepsilon(U_\varepsilon^-)\, \nabla \Phi_\varepsilon^+, \nabla (I - \pi^h)\eta) \, \mathrm{d}t \right| \\
&\qquad \le C\, \|(I-\pi^h)\eta\|_{L^2(0,T;H^1(\Omega))} \le C\, h\, \|\eta\|_{L^2(0,T;H^2(\Omega))}. \qquad (3.50b)
\end{aligned}$$

It follows from (3.50a), (3.50b) and (3.32b)–(3.32d) that as $h \to 0$

$$\int_0^T (\Lambda_\varepsilon(U_\varepsilon^-)\nabla\Phi_\varepsilon^+, \nabla(\pi^h\eta))\,\mathrm{d}t \to \int_{\{u>-1\}} c(u)\,\nabla\phi.\nabla\eta\,\mathrm{d}x\,\mathrm{d}t, \tag{3.51a}$$

$$\int_0^T (\Xi_\varepsilon(U_\varepsilon^-)\,\nabla\Phi_\varepsilon^+, \nabla(\pi^h\eta))\,\mathrm{d}t \to \int_{\{|u|<1\}} b(u)\,\nabla\phi.\nabla\eta\,\mathrm{d}x\,\mathrm{d}t, \tag{3.51b}$$

$$\int_0^T (\Xi_\varepsilon(U_\varepsilon^-)\,\nabla W_\varepsilon^+, \nabla(\pi^h\eta))\,\mathrm{d}t \to \int_{\{|u|<1\}} b(u)\,\nabla w.\nabla\eta\,\mathrm{d}x\,\mathrm{d}t. \tag{3.51c}$$

Combining (3.4a), (3.4b), (3.48), (3.49), (3.32a), (3.51a)–(3.51c), and the denseness of $H^1(0,T;\,H^2(\Omega))$ in $L^2(0,T;H^1(\Omega))$ yields the desired results (3.45a), (3.45b), on recalling (3.5) and (3.30). ∎

Remark 3.7. We note that it is possible to prove rigorously that the formally derived energy estimates (1.8), (1.10), and (1.11) are satisfied by the weak solution $\{\phi, u, w\}$. Using the techniques of the proof of Lemma 3.4, it is straightforward to derive from (3.11) and (3.7b), (3.7c) that as $h \to 0$

(3.52a)
$$[\Lambda_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla\Phi_\varepsilon^+ \to \underline{z}_1 \quad \text{weakly in } L^2(\Omega_T), \quad \text{with } \underline{z}_1 \equiv [c(u)]^{\frac12}\,\nabla\phi \text{ on } \{u>-1\};$$

(3.52b)
$$[\Xi_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla W_\varepsilon^+ \to \underline{z}_2 \quad \text{weakly in } L^2(\Omega_T), \quad \text{with } \underline{z}_2 \equiv [b(u)]^{\frac12}\,\nabla w \text{ on } \{|u|<1\}.$$

Combining (3.52a), (3.32a), and (2.28) then yields that

$$\begin{aligned}&\int_{\{u>-1\}} c(u)\,|\nabla\phi|^2\,\mathrm{d}x\,\mathrm{d}t + \tfrac12\int_0^T |\phi|^2_{0,\partial_2\Omega}\,\mathrm{d}t\\ &\le \liminf_{h\to0}\left\{\int_0^T |\,[\Lambda_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla\Phi_\varepsilon^+|_0^2\,\mathrm{d}t + \tfrac12\int_0^T |\Phi_\varepsilon^+|^2_{0,\partial_2\Omega}\,\mathrm{d}t\right\} \le \tfrac12\int_0^T |g|^2_{0,\partial_2\Omega}\,\mathrm{d}t\,.\end{aligned} \tag{3.53}$$

Similarly, it follows from (3.52b), (3.6a), (3.20b), (2.30a), and (2.16) that for a.a. $t \in (0,T)$

$$\begin{aligned}&\mathcal{E}(u(t)) + \gamma^{-1}\int_0^t \int_{\{|u(\cdot,\bar t)|<1\}} b(u)\,|\nabla w|^2\,\mathrm{d}x\,\mathrm{d}\bar t\\ &\qquad\le \liminf_{h\to0}\left\{\mathcal{E}(U_\varepsilon^+(t)) + \gamma^{-1}\int_0^t |\,[\Xi_\varepsilon(U_\varepsilon^-)]^{\frac12}\,\nabla W_\varepsilon^+|_0^2\,\mathrm{d}\bar t\right\}\\ &\qquad\le \lim_{h\to0}\mathcal{E}(U_\varepsilon^0) + \tfrac12\,\alpha^2\,\gamma^{-1}\,t\,|g|^2_{0,\partial_2\Omega} = \mathcal{E}(u^0) + \tfrac12\,\alpha^2\,\gamma^{-1}\,t\,|g|^2_{0,\partial_2\Omega}\,.\end{aligned} \tag{3.54}$$

We note that (3.53) and (3.54) correspond to the earlier formally derived energy estimates (1.8) and (1.10).

For the formally derived entropy estimate (1.11) we can argue as follows. First, it follows from (3.20b), (2.7), and assumption (ii) of Lemma 3.1 that for a.a. $t \in (0,T)$ as $h \to 0$

$$\begin{aligned}&|\,(G(u(\cdot,t)) - G_\varepsilon(U_\varepsilon^+(\cdot,t)),1)^h\,|\\ &\le |\,(G(u(\cdot,t)) - G(U_\varepsilon^+(\cdot,t)),1)^h\,| + |\,(G(U_\varepsilon^+(\cdot,t)) - G_\varepsilon(U_\varepsilon^+(\cdot,t)),1)^h\,| \;\to 0\,.\end{aligned} \tag{3.55}$$

Combining the convexity of G, (3.10a), (2.18), (2.44), (3.8), (3.55), (2.16), and assumption (ii) of Lemma 3.1 yields for a.a. $t \in (0,T)$ that

$$\begin{aligned}
\gamma\,(G(u(\cdot,t)),1) + \gamma \int_0^t |\Delta\, u|_0^2 \,\mathrm{d}\bar{t} &\le \gamma\,(G(u(\cdot,t)),1)^h + \gamma \int_0^t |\Delta\, u|_0^2 \,\mathrm{d}\bar{t} \\
&\le \lim_{h\to 0} \left\{ \gamma\,(G_\varepsilon(U_\varepsilon^+(\cdot,t)),1)^h + \gamma\,|\,(G(u(\cdot,t)) - G_\varepsilon(U_\varepsilon^+(\cdot,t)),1)^h\,| \,\right\} \\
&\quad + \gamma \liminf_{h\to 0} \int_0^t |\Delta^h\, U_\varepsilon^+|_h^2 \,\mathrm{d}\bar{t} \\
&\le \lim_{h\to 0} \{ \gamma\,(G_\varepsilon(U_\varepsilon^0),1)^h + \gamma\,|\,(G(u(\cdot,t)) - G_\varepsilon(U_\varepsilon^+(\cdot,t)),1)^h\,| \\
&\qquad + C(T)\,[1+\gamma^{-2}+\varepsilon^{-1}\,\tau^{\frac{1}{2}}\,]\,[\,\gamma\,\|U_\varepsilon^0\|_1^2 + \gamma^{-1}\,(1+T\,|g|_{0,\partial_2\Omega}^2)\,]\} \\
&= \gamma\,(G(u^0),1) + C(T)\,[1+\gamma^{-2}]\,[\,\gamma\,\|u^0\|_1^2 + \gamma^{-1}(\,1+T\,|g|_{0,\partial_2\Omega}^2)\,]\,.
\end{aligned}$$

This clearly corresponds to the formally derived estimate (1.11).

Remark 3.8. For our main convergence result, Theorem 3.6, we choose $U_\varepsilon^0 \equiv \pi^h u^0$. Therefore we require only the quasi-uniformity assumption in order to obtain (a) (2.43) via (2.46) and (2.21) and (b) (3.10b) via (3.10a) and (2.24). However, in case (a) we can replace the quasi uniformity with the far milder assumption that $\{\mathcal{T}^h\}_{h>0}$ is a regular partitioning at the expense of a minimum constraint on the uniform time step, similar to the argument in [2]. On recalling (1.13) and (2.22) it is easily established from $\{\mathcal{T}^h\}_{h>0}$ being a regular partitioning, elliptic regularity, as Ω is a rectangle, (1.14), and (2.19) that

$$\|(\mathcal{G}-\mathcal{G}^h)z^h\|_1 \le C\,h\,|z^h|_0 \qquad \forall\ z^h \in Z^h\,. \tag{3.56}$$

Then choosing $\chi \equiv \mathcal{G}^h \frac{\partial U_\varepsilon}{\partial t}$ in (2.13b) we obtain, similarly to (2.46), that

$$\|\mathcal{G}^h \tfrac{\partial U_\varepsilon}{\partial t}\|_{L^2(0,T;H^1(\Omega))} \le C. \tag{3.57}$$

Combining (3.56), (3.57) and noting (3.11), it follows that

$$\begin{aligned}
\|\mathcal{G}\tfrac{\partial U_\varepsilon}{\partial t}\|_{L^2(0,T;H^1(\Omega))} &\le \|(\mathcal{G}-\mathcal{G}^h)\tfrac{\partial U_\varepsilon}{\partial t}\|_{L^2(0,T;H^1(\Omega))} + \|\mathcal{G}^h \tfrac{\partial U_\varepsilon}{\partial t}\|_{L^2(0,T;H^1(\Omega))} \\
&\le C\,h\,\|\tfrac{\partial U_\varepsilon}{\partial t}\|_{L^2(\Omega_T)} + C \le C\,(\tau^{-\frac{1}{4}}\,h+1) \le C
\end{aligned}$$

if the mild time-step constraint $C\,h^4 \le \tau$ is satisfied. As for case (b), the obtained result $\|U_\varepsilon\|_{L^2(0,T;H^2(\Omega))} \le C$ is more than we need. For our main convergence result we really need only $\|U_\varepsilon\|_{L^2(0,T;W^{1,s}(\Omega))} \le C$, $s>2$. However, it does not appear possible to derive this bound without using the stronger result and the quasi-uniformity assumption.

4. Solution of the discrete system. We now discuss algorithms for solving the resulting system of algebraic equations for $\{\Phi_\varepsilon^n, U_\varepsilon^n, W_\varepsilon^n\}$ arising at each time level from the approximation ($\mathrm{P}_\varepsilon^{h,\tau}$). As (2.13a) in ($\mathrm{P}_\varepsilon^{h,\tau}$) is independent of $\{U_\varepsilon^n, W_\varepsilon^n\}$, we solve it first to obtain Φ_ε^n and then solve (2.13b), (2.13c) for $\{U_\varepsilon^n, W_\varepsilon^n\}$. Solving (2.13a) is straightforward, as it is linear.

Adopting the obvious notation, the system (2.13b), (2.13c) can be rewritten as follows: Find $\{\underline{U}_\varepsilon^n, \underline{W}_\varepsilon^n\} \in \mathcal{K}^{\mathcal{J}} \times \mathbb{R}^{\mathcal{J}}$ such that

$$\gamma\,M\,\underline{U}_\varepsilon^n + \tau_n\,A^{n-1}\,\underline{W}_\varepsilon^n = \underline{r}, \tag{4.1a}$$

$$\gamma\,(\underline{V}-\underline{U}_\varepsilon^n)^T\,B\,\underline{U}_\varepsilon^n - (\underline{V}-\underline{U}_\varepsilon^n)^T\,M\,\underline{W}_\varepsilon^n \ge (\underline{V}-\underline{U}_\varepsilon^n)^T\,\underline{s} \qquad \forall\ \underline{V} \in \mathcal{K}^{\mathcal{J}}, \tag{4.1b}$$

where M, B, and A^{n-1} are symmetric $\mathcal{J}\times\mathcal{J}$ matrices, $\mathcal{J}:=\#J$, with entries

$$M_{ij}:=(\chi_i,\chi_j)^h,\qquad B_{ij}:=(\nabla\chi_i,\nabla\chi_j),\qquad A^{n-1}_{ij}:=(\Xi_\varepsilon(U^{n-1}_\varepsilon)\,\nabla\chi_i,\nabla\chi_j)$$

$$\text{and}\qquad \underline{r}:=\gamma\,M\,\underline{U}^{n-1}_\varepsilon-\alpha\,\tau_n\,A^{n-1}\,\underline{\Phi}^n_\varepsilon\in\mathbb{R}^{\mathcal{J}},\qquad \underline{s}:=\gamma^{-1}\,M\,\underline{U}^{n-1}_\varepsilon\in\mathbb{R}^{\mathcal{J}}. \tag{4.2}$$

Let $A^{n-1}\equiv A_D-A_L-A_L^T$, with A_L and A_D being the lower triangular and diagonal parts of the matrix A^{n-1}, and similarly for B. We use this formulation in constructing our "Gauss–Seidel-type" iterative method to solve (2.13b), (2.13c).

Given $\{U^{n,0}_\varepsilon,W^{n,0}_\varepsilon\}\in K^h\times S^h$, for $k\geq 1$ find $\{U^{n,k}_\varepsilon,W^{n,k}_\varepsilon\}\in K^h\times S^h$ such that

$$\gamma\,M\,\underline{U}^{n,k}_\varepsilon+\tau_n\,(A_D-A_L)\,\underline{W}^{n,k}_\varepsilon=\underline{r}+\tau_n\,A_L^T\,\underline{W}^{n,k-1}_\varepsilon, \tag{4.3a}$$

$$(\underline{V}-\underline{U}^{n,k}_\varepsilon)^T\,(\gamma\,(B_D-B_L)\,\underline{U}^{n,k}_\varepsilon-M\,\underline{W}^{n,k}_\varepsilon)\geq(\underline{V}-\underline{U}^{n,k}_\varepsilon)^T\,(\underline{s}+\gamma\,B_L^T\,\underline{U}^{n,k-1}_\varepsilon)\qquad\forall\ \underline{V}\in\mathcal{K}^{\mathcal{J}}. \tag{4.3b}$$

A similar iterative method is used in [15] to solve a related linear system. They prove convergence of this approach for their linear system by analyzing the eigenvalues of the resultant iteration matrix. Below, we prove convergence of (4.3a), (4.3b) for our nonlinear system (2.13b), (2.13c) using an energy method.

THEOREM 4.1. *Let the assumptions* (A) *hold. Then for* $\{U^{n,0}_\varepsilon,W^{n,0}_\varepsilon\}\in K^h\times S^h$ *the sequence* $\{U^{n,k}_\varepsilon,W^{n,k}_\varepsilon\}_{k\geq0}$ *generated by the algorithm* (4.3a), (4.3b) *satisfies*

$$\|U^n_\varepsilon-U^{n,k}_\varepsilon\|_1\to0\quad\textit{and}\quad\int_\Omega\Xi_\varepsilon(U^{n-1}_\varepsilon)\,|\nabla(W^n_\varepsilon-W^{n,k}_\varepsilon)|^2\,\mathrm{d}x\to0\quad\textit{as } k\to\infty. \tag{4.4}$$

Proof. Let $\underline{Y}^{n,k}:=\underline{U}^n_\varepsilon-\underline{U}^{n,k}_\varepsilon$ and $\underline{Z}^{n,k}:=\underline{W}^n_\varepsilon-\underline{W}^{n,k}_\varepsilon$. Now subtracting (4.3a) from (4.1a) and testing the resulting equation with $\underline{Z}^{n,k}$ yield

$$\gamma\,[\underline{Z}^{n,k}]^T\,M\,\underline{Y}^{n,k}+\tau_n\,[\underline{Z}^{n,k}]^T\,(A_D-A_L)\,\underline{Z}^{n,k}=\tau_n\,[\underline{Z}^{n,k}]^T\,A_L^T\,\underline{Z}^{n,k-1}. \tag{4.5}$$

Choosing $\underline{V}\equiv\underline{U}^{n,k}_\varepsilon$ in (4.1b) and $\underline{V}\equiv\underline{U}^n_\varepsilon$ in (4.3b) yields

$$-\gamma\,[\underline{Y}^{n,k}]^T\,(B_D-B_L)\,\underline{Y}^{n,k}+[\underline{Y}^{n,k}]^T\,M\,\underline{Z}^{n,k}\geq-\gamma\,[\underline{Y}^{n,k}]^T\,B_L^T\,\underline{Y}^{n,k-1}. \tag{4.6}$$

Combining (4.5) and (4.6) yields that

$$\begin{aligned}&\gamma^2\,[\underline{Y}^{n,k}]^T\,(B_D-B_L)\,\underline{Y}^{n,k}+\tau_n\,[\underline{Z}^{n,k}]^T\,(A_D-A_L)\,\underline{Z}^{n,k}\\&\qquad\leq\gamma^2\,[\underline{Y}^{n,k}]^T\,B_L^T\,\underline{Y}^{n,k-1}+\tau_n\,[\underline{Z}^{n,k}]^T\,A_L^T\,\underline{Z}^{n,k-1}.\end{aligned} \tag{4.7}$$

We now split the diagonal matrix $A_D:=A_{D_1}+A_{D_2}$, where $(A_{D_1})_{ii}:=-\sum_{j=1}^{i-1}A_{ij}$ and $(A_{D_2})_{ii}:=-\sum_{j=i+1}^{\mathcal{J}}A_{ij}=A_{ii}-(A_{D_1})_{ii}$. Then we have that

$$\begin{aligned}[\underline{Z}^{n,k}]^T\,A_L^T\,\underline{Z}^{n,k-1}&=\sum_{i=1}^{\mathcal{J}}Z^{n,k}_i\sum_{j=1}^{\mathcal{J}}(A_L^T)_{ij}\,Z^{n,k-1}_j\\&\leq\tfrac12\sum_{i=1}^{\mathcal{J}}\sum_{j=1}^{\mathcal{J}}(A_L)_{ji}\,[(Z^{n,k}_i)^2+(Z^{n,k-1}_j)^2]\\&\leq\tfrac12\sum_{i=1}^{\mathcal{J}}(A_{D_2})_{ii}\,(Z^{n,k}_i)^2+\tfrac12\sum_{j=1}^{\mathcal{J}}(A_{D_1})_{jj}\,(Z^{n,k-1}_j)^2.\end{aligned} \tag{4.8}$$

Combining (4.7), (4.8), and a similar argument for B yields that

$$\tfrac{\gamma^2}{2}\,[\underline{Y}^{n,k}]^T B\,\underline{Y}^{n,k} + \tfrac{\gamma^2}{2}\,[\underline{Y}^{n,k}]^T B_{D_1}\,\underline{Y}^{n,k} + \tfrac{\tau_n}{2}\,[\underline{Z}^{n,k}]^T A^{n-1}\,\underline{Z}^{n,k} + \tfrac{\tau_n}{2}\,[\underline{Z}^{n,k}]^T A_{D_1}\,\underline{Z}^{n,k}$$

(4.9)

$$\leq \tfrac{\gamma^2}{2}\,[\underline{Y}^{n,k-1}]^T B_{D_1}\,\underline{Y}^{n,k-1} + \tfrac{\tau_n}{2}\,[\underline{Z}^{n,k-1}]^T A_{D_1}\,\underline{Z}^{n,k-1}.$$

Therefore, we have that $\{\, \frac{\gamma^2}{2}\,[\underline{Y}^{n,k}]^T\, B_{D_1}\,\underline{Y}^{n,k} + \frac{\tau_n}{2}\,[\underline{Z}^{n,k}]^T\, A_{D_1}\,\underline{Z}^{n,k}\,\}_{k\geq 0}$ is a decreasing sequence. Since it is bounded below the sequence has a limit. Combining this and (4.9) yields

(4.10)

$$|U_\varepsilon^n - U_\varepsilon^{n,k}|_1 \to 0 \quad \text{and} \quad \int_\Omega \Xi_\varepsilon(U_\varepsilon^{n-1})\,|\nabla(W_\varepsilon^n - W_\varepsilon^{n,k})|^2\,\mathrm{d}x \to 0 \quad \text{as } k\to\infty.$$

Furthermore, multiplying (4.3a) with $\underline{1}^T := (1,\dots,1)$, noting that $A^{n-1}\,\underline{1} = \underline{0}$, and recalling the splitting of A^{n-1} yields that

$$\begin{aligned}\gamma\,(U_\varepsilon^{n,k} - U_\varepsilon^{n-1}, 1)^h &= \tau_n\,\underline{1}^T A_L^T\,(\underline{W}_\varepsilon^{n,k-1} - \underline{W}_\varepsilon^{n,k}) = \tau_n\,\underline{1}^T A_{D_1}\,(\underline{W}_\varepsilon^{n,k} - \underline{W}_\varepsilon^{n,k-1})\\ &= \tau_n\,\underline{1}^T A_{D_1}\,\underline{Z}^{n,k-1} - \tau_n\,\underline{1}^T A_{D_1}\,\underline{Z}^{n,k} \quad \to\ 0\,,\end{aligned} \tag{4.11}$$

where we have again used the fact that $\{\,\tau_n\,[\underline{Z}^{n,k}]^T\,A_{D_1}\,\underline{Z}^{n,k}\,\}_{k\geq 0}$ has a limit. Combining (4.10), (4.11), (2.3), and (2.18) yields the desired result (4.4). ∎

Remark 4.2. We note that (4.3a), (4.3b) can be solved explicitly for $j = 1 \to \mathcal{J}$. In particular, let $\widehat{\underline{r}} := \underline{r} + \tau_n\,(A_L\,\underline{W}_\varepsilon^{n,k} + A_L^T\,\underline{W}_\varepsilon^{n,k-1})$ and $\widehat{\underline{s}} := \underline{s} + \gamma\,(B_L\,\underline{U}_\varepsilon^{n,k} + B_L^T\,\underline{U}_\varepsilon^{n,k-1})$. Then, on recalling (1.16a), we set for $j = 1 \to \mathcal{J}$

$$[\underline{U}_\varepsilon^{n,k}]_j = \left[\frac{M_{jj}\,\widehat{\underline{r}}_j + \tau_n\,A_{jj}^{n-1}\,\widehat{\underline{s}}_j}{\gamma\,[M_{jj}]^2 + \tau_n\,\gamma\,A_{jj}^{n-1}\,B_{jj}}\right]_{\mathcal{K}} \quad \text{and} \quad [\underline{W}_\varepsilon^{n,k}]_j = \frac{\widehat{\underline{r}}_j - \gamma\,M_{jj}\,[\underline{U}_\varepsilon^{n,k}]_j}{\tau_n\,A_{jj}^{n-1}}. \tag{4.12}$$

Remark 4.3. Although we have no convergence proof, in practice an overrelaxed version of (4.3a), (4.3b) performed better. To this end we replace (4.12) for a given $\omega \geq 1$ with

$$\begin{aligned}[\underline{U}_\varepsilon^{n,k}]_j &= \left[\omega\,\frac{M_{jj}\,\widehat{\underline{r}}_j - \tau_n\,A_{jj}^{n-1}\,\widehat{\underline{s}}_j}{\gamma\,[M_{jj}]^2 + \tau_n\,\gamma\,A_{jj}^{n-1}\,B_{jj}} + (1-\omega)\,[\underline{U}_\varepsilon^{n,k-1}]_j\right]_{\mathcal{K}}\\ \text{and} \quad [\underline{W}_\varepsilon^{n,k}]_j &= \frac{\widehat{\underline{r}}_j - \gamma\,M_{jj}\,[\underline{U}_\varepsilon^{n,k}]_j}{\tau_n\,A_{jj}^{n-1}}\,.\end{aligned} \tag{4.13}$$

5. Numerical results. In order to define the initial shape of the void we introduce the following function. Given $z \in \mathbb{R}^2$, $a \in \mathbb{R}^2$ with $\min\{a_1, a_2\} = 1$ and $R \in \mathbb{R}_{>0}$ we define

$$v(z,a,R;\,x) := \begin{cases} -1, & r(x) \leq R - \frac{\gamma\,\pi}{2},\\ \sin(\frac{r(x)-R}{\gamma}), & |r(x) - R| < \frac{\gamma\,\pi}{2}\,,\\ 1, & r(x) \geq R + \frac{\gamma\,\pi}{2}, \end{cases} \tag{5.1}$$

where $r(x) := ((\frac{x_1-z_1}{a_1})^2 + (\frac{x_2-z_2}{a_2})^2)^{\frac{1}{2}}$. Equation (5.1) represents a void with the shape of an ellipse with semiaxes $a_1 R$ and $a_2 R$. In line with the asymptotics of the phase field approach (see section 1), the interfacial thickness is not less than $\gamma\,\pi$. For the initial data u^0 to (P) we chose either (i) one ellipse or (ii) two ellipses; that is,

(5.2)

$$\text{(i)}\quad u^0(x) = v(z,a,R;\,x) \qquad \text{or} \quad \text{(ii)}\quad u^0(x) = v(z,a,R;\,x) + v(\tilde{z},\tilde{a},\widetilde{R};\,x) - 1\,.$$

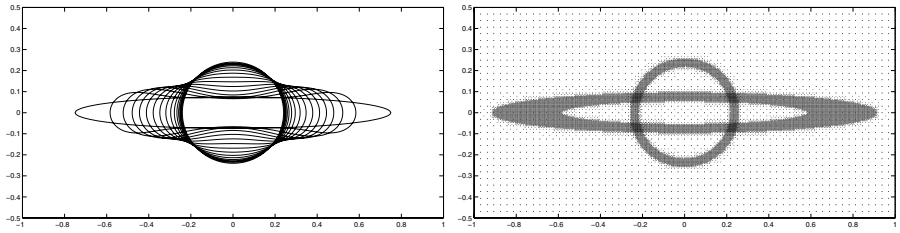


FIG. 1. *($\alpha = 0$). Zero level sets for solution $U_\varepsilon(x,t)$ of $(\mathrm{P}_\varepsilon^{h,\tau})$ at times $t = 0, 2\times10^{-4}, \ldots, 3\times10^{-3}, T = 10^{-2}$ and adaptive mesh for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ at times $t = 0, T$.*

In all the experiments below, the parameters above were chosen so that these ellipses lie in the interior of Ω; and hence the resulting u^0 satisfies all the assumptions of Lemma 3.1.

To solve for $(\mathrm{P}_\varepsilon^{h,\tau})$, the given domain $\Omega = (-L_1, L_1) \times (-L_2, L_2)$ was partitioned uniformly into right-angled isosceles triangles. Throughout, we chose the number of triangles such that there were at least approximately 6 mesh points across the interface in each direction; that is, $8(\frac{h}{\sqrt{2}}) \leq \gamma\,\pi$.

For the iterative algorithm (4.3a), (4.3b) we set, for $n \geq 1$, $\{U_\varepsilon^{n,0}, W_\varepsilon^{n,0}\} \equiv \{U_\varepsilon^{n-1}, W_\varepsilon^{n-1}\}$, where $U_\varepsilon^0 \equiv \pi^h u^0$ and $W_\varepsilon^0 \equiv -\gamma\,\Delta^h U_\varepsilon^0 - \gamma^{-1}\,U_\varepsilon^0$, and adopted the stopping criterion

$$|U_\varepsilon^{n,k} - U_\varepsilon^{n,k-1}|_{0,\infty} < tol,$$

with $tol = 10^{-7}$. Furthermore, we set $\{U_\varepsilon^n, W_\varepsilon^n\} \equiv \{\pi^h[p_{\mathcal{K}}^{tol}(U_\varepsilon^{n,k})], W_\varepsilon^{n,k}\}$, where

$$p_{\mathcal{K}}^{tol}(s) := \begin{cases} -1 & s \leq -1 + tol, \\ s & |s| < 1 - tol, \\ 1 & s \geq 1 - tol. \end{cases}$$

Our first experiment is for $\alpha = 0$ and shows the evolution of an ellipse to a circle due to surface diffusion. We chose the following parameters for $(\mathrm{P}_\varepsilon^{h,\tau})$: $L_1 = 1$, $L_2 = 0.5$, $\gamma = \frac{1}{16\pi}$, $\alpha = 0$, $T = 10^{-2}$, $\tau_n = \tau = 8\times10^{-7}$, $\varepsilon = 10^{-5}$. For the initial profile we chose (5.2)(i) with $z = (0,0)$, $a = (10,1)$, and $R = 0.075$ and used $\omega = 1.4$ for the iterative algorithm (4.13). We used a uniform 128×128 triangulation for each of the two unit squares. In Figure 1 we plot the zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 2\times10^{-4}, \ldots, 3\times10^{-3}, T$. We note the very good agreement with the direct finite element approximation of the sharp interface problem, (1.1a), (1.1b) and (1.2), in [28, Figure 3]. We repeated the above experiment for the scheme $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ (see Remark 2.10) and obtained graphically indistinguishable results. However, the scheme $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ was 2.2 times faster than solving the original approximation $(\mathrm{P}_\varepsilon^{h,\tau})$. Moreover, for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ one knows a priori that $U_\varepsilon^n(p_j) = U_\varepsilon^{n-1}(p_j)$ for all $j \in J$ with $(\widetilde{\Xi}_\varepsilon(U_\varepsilon^{n-1}), \chi_j) = 0$, the so-called passive nodes. A natural approach to utilize this fact is to use a fine mesh in the "interfacial region" only, while employing a coarser mesh elsewhere. We note that for $\alpha = 0$ this is equivalent to using a uniform fine mesh. However, for the above experiment the described adaptive mesh approach was 4.4 times faster than solving $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ on a uniform mesh. Hence, overall to solve for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ took 10% of the time it took to solve $(\mathrm{P}_\varepsilon^{h,\tau})$. Hence for all the remaining computations we used an adaptive mesh to solve for the approximation $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$.

We should note that for $\alpha > 0$ the adaptive mesh approach is not equivalent to solving $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ on a uniform mesh, since in the former case not all the active nodes with respect to $\widetilde{\Lambda}_\varepsilon(U_\varepsilon^{n-1})$, as opposed to $\widetilde{\Xi}_\varepsilon(U_\varepsilon^{n-1})$, are represented on the fine mesh. Hence the respective solutions Φ_ε^n can differ slightly in the interfacial region, yielding different solutions U_ε^n. But the electric potential is not rapidly varying away from the interfacial region, and hence is well approximated by the coarse mesh, so there is no need for the more costly fine mesh. Furthermore, we obtained virtually identical results in test runs, and hence we are satisfied that these differences are negligible.

In order to implement the desired mesh we used the adaptive finite element code Albert 1.0; see [24]. The code uses bisectioning, and its reversal, for refining and coarsening, respectively. Hence starting with an initial right-angled isosceles triangulation yields similar triangles throughout. We now describe our mesh refinement strategy for the physically relevant case of $L_1 \geq L_2$ and further assume without loss of generality that L_1 is an integer multiple of L_2.

Given the two parameters $N_c < N_f$ we set $h_c := \frac{2^{\frac{3}{2}} L_2}{N_c}$ and $h_f := \frac{2^{\frac{3}{2}} L_2}{N_f}$, respectively. Throughout, we choose our initial triangulation $\widetilde{\mathcal{T}}^0$ to be a uniform partitioning of Ω into triangles σ of diameter $h_\sigma = h_f$ and fix the parameters $\delta_f = tol \times 10^{-1}$ and $\delta_c = tol \times 10^{-3}$. Then, for $n \geq 1$, given U_ε^{n-1} and a triangulation $\widetilde{\mathcal{T}}^{n-1}$, a triangle $\sigma \in \widetilde{\mathcal{T}}^{n-1}$ is marked for refinement if it, or one of its neighboring elements, satisfies

$$\eta_\sigma := \left| \min_{x \in \overline{\sigma}} |U_\varepsilon^{n-1}(x)| - 1 \right| > \delta_f .$$

If a triangle that is marked for refinement satisfies $h_\sigma > h_f$, it is refined into two smaller triangles via a bisectioning of its longest edge. A triangle σ is marked for coarsening if it satisfies $h_\sigma < h_c$ and $\eta_\sigma < \delta_c$. A triangle that is marked for coarsening is coarsened only if all its neighboring elements are marked for coarsening as well. This cycle is repeated until no triangle has been refined or coarsened. We note that the maximum number of cycles is $\frac{N_f}{N_c}$. However, apart from the case $n = 1$ the number of cycles required will be 1, due to the fact that the region of active nodes can advance one mesh point per time step only; see Remark 3.2. The above process ensures that all active nodes are always within the fine part of the adaptive mesh.

For the remaining experiments we adopted the following strategy. Given a $\gamma > 0$ we chose N_f such that there were at least approximately 6 mesh points across the interface, set $N_c := \frac{1}{8} N_f$, and chose a suitable time step size τ. As the numerical interfacial region can only advance by one mesh point per time step one has to choose τ sufficiently small so that $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$ is capable of approximating the speed of propagation of the void; cf. [3, section 5.1]. On obtaining the desired experiment's solutions for these discretization parameters we halved γ, halved h_f, and quartered τ, while keeping $\varepsilon = 10^{-5}$ fixed throughout. In almost all instances we repeated the above procedure until the solutions for two consecutive choices of γ were graphically indistinguishable. Here we report on the converged experiments.

First, we repeated the previous experiment for $\alpha = 0$ for a smaller $\gamma = \frac{1}{32\pi}$, using a finer mesh. In particular, we chose the following parameters for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$: $L_1 = 1$, $L_2 = 0.5$, $T = 10^{-2}$, $\tau_n = \tau = 2 \times 10^{-7}$. For the initial profile we chose (5.2)(i) with $z = (0, 0)$, $a = (10, 1)$, and $R = 0.075$ and used $\omega = 1.5$ for the iterative algorithm (4.13). The refinement parameters were $N_f = 256$ and $N_c = 32$. The obtained results were virtually identical to the ones from our earlier computations; see Figure 1. On the right-hand side of this figure we plot the vertices of the adaptive mesh for the latter experiment at times $t = 0, T$.

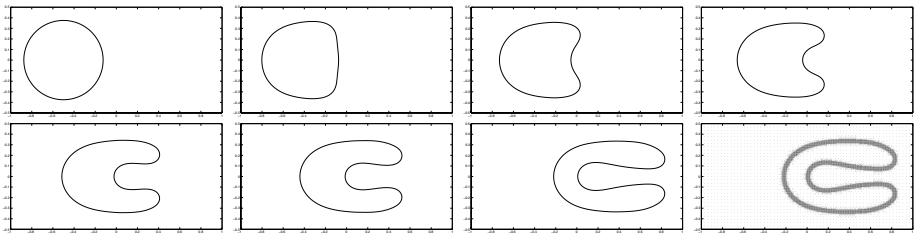


FIG. 2. *($\alpha \approx 114\,\pi$). Zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 4\times10^{-5}, 8\times10^{-5}, 1.2\times10^{-4}, 2\times10^{-4}, 2.4\times10^{-4}, T = 3.6\times10^{-4}$ and adaptive mesh at time $t = T$.*

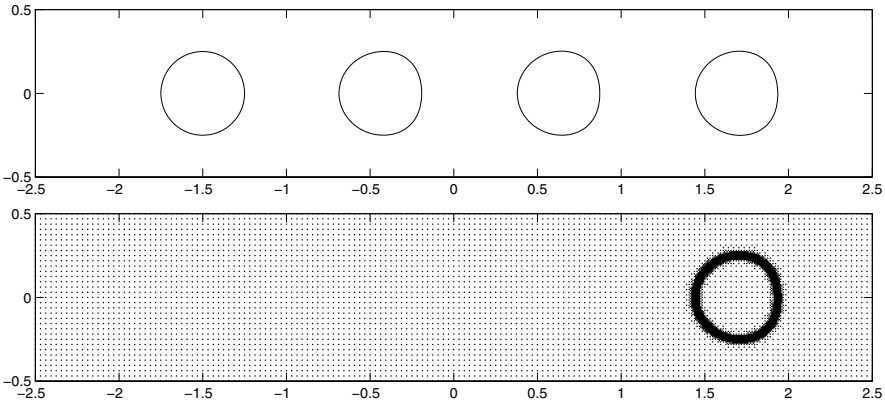


FIG. 3. *($\alpha = 40\,\pi$). Zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 1.25\times10^{-3}, \dots, T = 3.75\times10^{-3}$ and adaptive mesh at time $t = T$.*

In our first experiment for $\alpha > 0$ we chose the radius of the initially circular void to be relatively large compared to the width of the conductor, $2\,L_2$, in correspondence to [9, Figure 4]. We used the following parameters for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$: $L_1 = 1$, $L_2 = 0.5$, $\gamma = \frac{1}{32\pi}$, $\alpha = \frac{1024}{9}\,\pi \approx 114\,\pi$, $T = 3.6\times10^{-4}$, $\tau_n = \tau = 10^{-7}$. As initial data we chose (5.2)(i) with $z = (-0.5, 0)$, $a = (1, 1)$, and $R = 0.375$ and used $\omega = 1.9$ for the iterative algorithm (4.13). The refinement parameters were $N_f = 256$ and $N_c = 32$. In Figure 2 we plot the zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 4\times10^{-5}, 8\times10^{-5}, 1.2\times10^{-4}, 2\times10^{-4}, 2.4\times10^{-4}, T$ and the vertices of the adaptive mesh at time $t = T$. We note the good agreement with the direct finite element approximation of the sharp interface problem, (1.1a), (1.1b) and (1.2), in [9, Figure 4].

The next experiment corresponds to [20, Figure 5] and [28, Figure 6]. We chose the following parameters for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$: $L_1 = 2.5$, $L_2 = 0.5$, $\gamma = \frac{1}{32\pi}$, $\alpha = 40\,\pi$, $T = 3.75\times10^{-3}$, $\tau_n = \tau = 6\times10^{-7}$. As initial data we chose (5.2)(i) with $z = (-1.5, 0)$, $a = (1, 1)$, and $R = 0.25$ and used $\omega = 1.9$ for the iterative algorithm (4.13). The refinement parameters were $N_f = 256$ and $N_c = 32$. In Figure 3 we plot the zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 1.25\times10^{-3}, \dots, T$ and the vertices of the adaptive mesh at time $t = T$. One can observe that the circular void, with a slightly flattened front, stably propagates through the conductor.

However, for larger α this is no longer the case. We repeated the above experiment for $\alpha = 120\,\pi$ in correspondence to [28, Figure 7]. In particular, we chose $L_1 = 2.5$, $L_2 = 0.5$, $\gamma = \frac{1}{64\pi}$, $\alpha = 120\,\pi$, $T = 1.184\times10^{-3}$, $\tau_n = \tau = 5\times10^{-8}$. We used the same initial data as in Figure 3 and chose $\omega = 1$ for the iterative algorithm (4.13). The

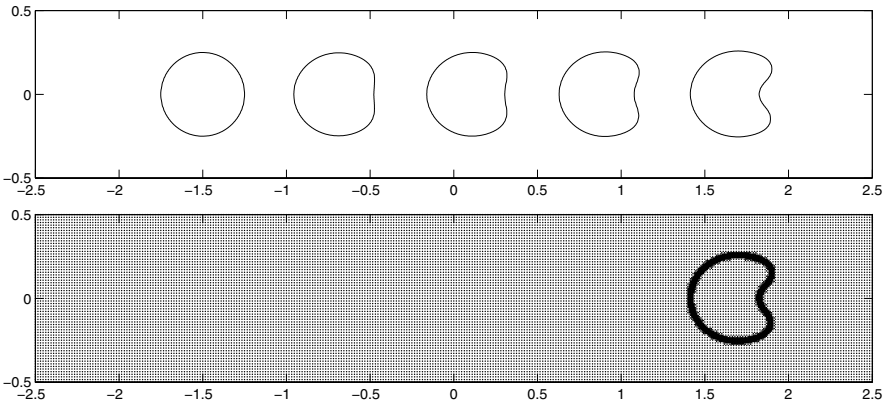


FIG. 4. *($\alpha = 120\,\pi$). Zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 3\times10^{-4}, \ldots, T = 1.184\times10^{-3}$ and adaptive mesh at time $t = T$.*

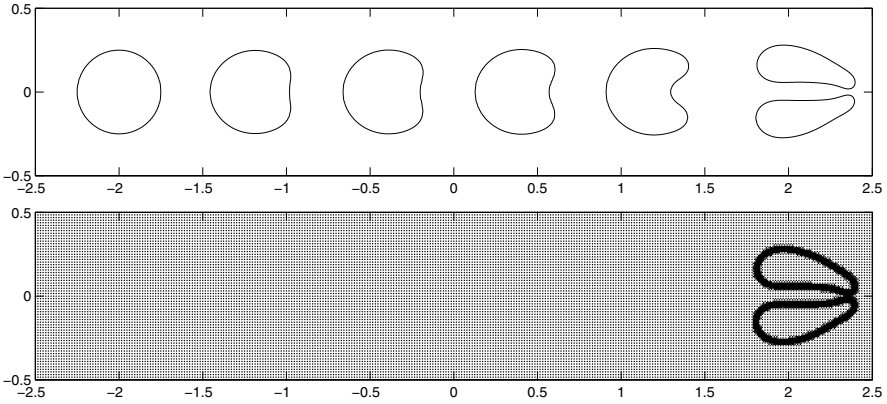


FIG. 5. *($\alpha = 120\,\pi$). Zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 3\times10^{-4}, \ldots, T = 1.5\times10^{-3}$ and adaptive mesh at time $t = T$.*

refinement parameters were $N_f = 512$ and $N_c = 64$. In Figure 4 we plot the zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 2.96\times10^{-4}, \ldots, T$ and the vertices of the adaptive mesh at time $t = T$. We repeated the last experiment with the same parameters but started with the initial void more to the left and integrated for a longer time; in particular, we set $z = (-2, 0)$ and $T = 1.5\times10^{-3}$. This allows the void to further change its shape; see Figure 5. The above three experiments are very sensitive to the choice of γ and τ. Although our results show similarities with the cited ones, there is no strong agreement between these different types of simulations.

The next experiment corresponds to [20, Figure 9]. We chose the following parameters for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$: $L_1 = 0.5$, $L_2 = 0.5$, $\gamma = \frac{1}{32\pi}$, $\alpha = 12\,\pi$, $T = 1.5\times10^{-4}$, $\tau_n = \tau = 2\times10^{-7}$. As initial data we chose (5.2)(ii) with $z = (-0.15, 0)$, $a = (1.12, 1.6)$, $R = 0.125$, $\tilde{z} = (0.15, 0)$, $\tilde{a} = (0.96, 1.92)$, and $\widetilde{R} = 0.125$ and used $\omega = 1.8$ for the iterative algorithm (4.13). The refinement parameters were $N_f = 256$ and $N_c = 32$. In Figure 6 we plot the zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 3.04\times10^{-5}, 3.8\times10^{-5}, 4.56\times10^{-5}, T$ and the vertices of the adaptive mesh at time $t = T$. We note that the time the two ellipses are merging is sensitive to the choice of γ.

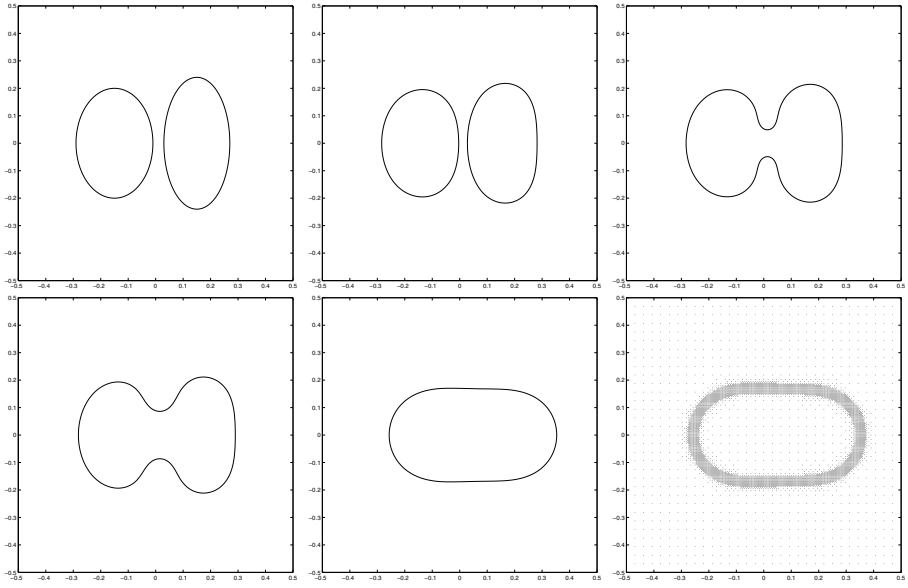


FIG. 6. *($\alpha = 12\,\pi$). Zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 3.04\times10^{-5}, 3.8\times10^{-5}, 4.56\times10^{-5}, T = 1.5\times10^{-4}$ and adaptive mesh at time $t = T$.*

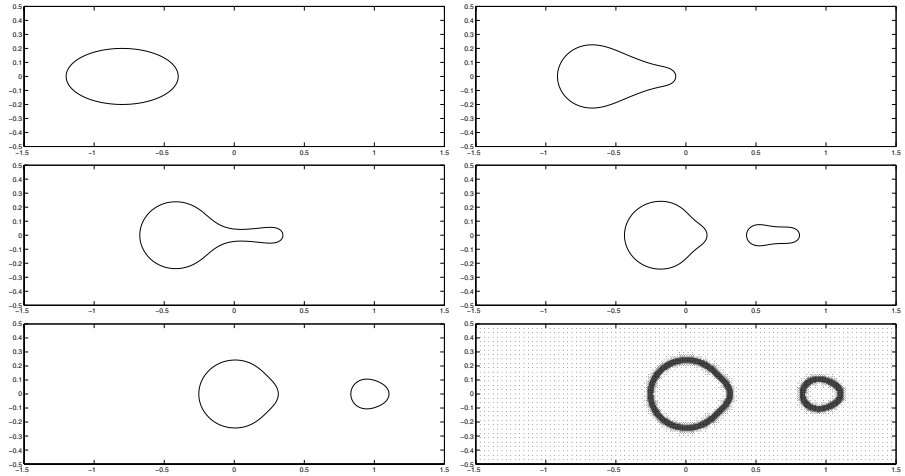


FIG. 7. *($\alpha = 120\,\pi$). Zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 8.75\times10^{-5}, 1.75\times10^{-4}, 2.625\times10^{-4}, T = 3.32\times10^{-4}$ and adaptive mesh at time $t = T$.*

The next experiment corresponds to [20, Figure 11]. We chose the following parameters for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$: $L_1 = 1.5$, $L_2 = 0.5$, $\gamma = \frac{1}{32\pi}$, $\alpha = 120\,\pi$, $T = 3.32\times10^{-4}$, $\tau_n = \tau = 2.5\times10^{-8}$. As initial data we chose (5.2)(i) with $z = (-0.8, 0)$, $a = (2, 1)$, and $R = 0.2$ and used $\omega = 1.3$ for the iterative algorithm (4.13). The refinement parameters were $N_f = 256$ and $N_c = 32$. In Figure 7 we plot the zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 8.75\times10^{-5}, 1.75\times10^{-4}, 2.625\times10^{-4}, T$ and the vertices of the adaptive mesh at time $t = T$.

Our final experiment corresponds to [20, Figure 10]. We chose the following parameters for $(\widetilde{\mathrm{P}}_\varepsilon^{h,\tau})$: $L_1 = 1.5$, $L_2 = 0.5$, $\gamma = \frac{1}{32\pi}$, $\alpha = 64\,\pi$, $T = 7.91\times10^{-4}$,

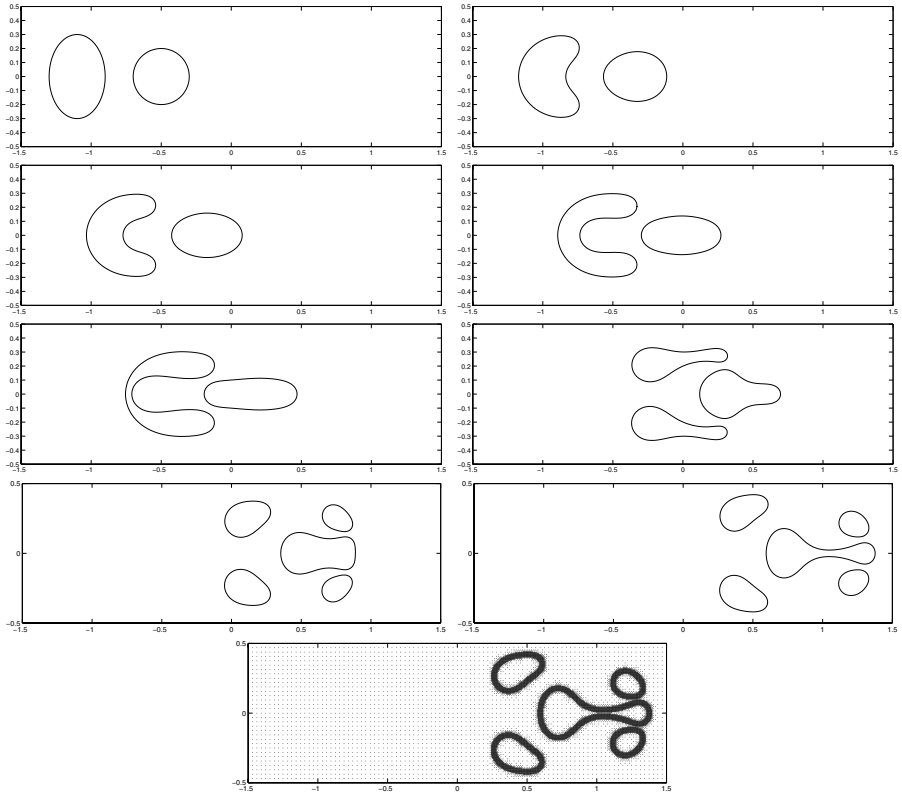


FIG. 8. *($\alpha = 64\,\pi$). Zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 1.13\times10^{-4}, \ldots, T = 7.91\times10^{-4}$ and adaptive mesh at time $t = T$.*

$\tau_n = \tau = 2.5\times10^{-8}$. As initial data we chose (5.2)(ii) with $z = (-1.1, 0)$, $a = (1, 1.5)$, $R = 0.2$, $\tilde{z} = (-0.5, 0)$, $\tilde{a} = (1, 1)$, and $\widetilde{R} = 0.2$ and used $\omega = 1.3$ for the iterative algorithm (4.13). The refinement parameters were $N_f = 256$ and $N_c = 32$. In Figure 8 we plot the zero level sets for $U_\varepsilon(x,t)$ at times $t = 0, 1.13\times10^{-4}, \ldots, T$ and the vertices of the adaptive mesh at time $t = T$. We note that for the last two experiments there is good agreement for different values of γ but only partial agreement with the cited results in [20]. However, one should note that the fixed mesh for their level set approach of the sharp interface model, (1.1a), (1.1b) and (1.2), is rather coarse.

REFERENCES

[1] J. W. BARRETT AND J. F. BLOWEY, *Finite element approximation of a degenerate Allen–Cahn/Cahn–Hilliard system*, SIAM J. Numer. Anal., 39 (2001), pp. 1598–1624.

[2] J. W. BARRETT, J. F. BLOWEY, AND H. GARCKE, *Finite element approximation of a fourth order nonlinear degenerate parabolic equation*, Numer. Math., 80 (1998), pp. 525–556.

[3] J. W. BARRETT, J. F. BLOWEY, AND H. GARCKE, *Finite element approximation of the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Numer. Anal., 37 (1999), pp. 286–318.

[4] J. W. BARRETT, J. F. BLOWEY, AND H. GARCKE, *On fully practical finite element approximations of degenerate Cahn–Hilliard systems*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 713–748.

[5] J. W. BARRETT, H. GARCKE, AND R. NÜRNBERG, *Finite element approximation of surfactant*

spreading on a thin film, SIAM J. Numer. Anal., 41 (2003), pp. 1427–1464.

[6] J. W. Barrett, S. Langdon, and R. Nürnberg, *Finite element approximation of a sixth order nonlinear degenerate parabolic equation*, Numer. Math., 96 (2004), pp. 401–434.

[7] J. W. Barrett and R. Nürnberg, *Convergence of a finite element approximation of surfactant spreading on a thin film in the presence of van der Waals forces*, IMA J. Numer. Anal., 24 (2004), pp. 323–363.

[8] D. N. Bhate, A. Kumar, and A. F. Bower, *Diffuse interface model for electromigration and stress voiding*, J. Appl. Phys., 87 (2000), pp. 1712–1721.

[9] A. F. Bower and L. B. Freund, *Finite element analysis of electromigration and stress induced diffusion in deformable solids*, in Materials Reliability in Microelectronics V, Mater. Res. Soc. Sympos. Proc. 391, A. S. Oates, W. F. Filter, R. Rosenberg, A. L. Greer, and K. Gadepally, eds., Materials Research Society, Warrendale, PA, 1995, pp. 177–188.

[10] J. W. Cahn, C. M. Elliott, and A. Novick-Cohen, *The Cahn–Hilliard equation with a concentration dependent mobility: Motion by minus the Laplacian of the mean curvature*, European J. Appl. Math., 7 (1996), pp. 287–301.

[11] P. G. Ciarlet, *Numerical Linear Algebra and Optimisation*, Cambridge University Press, Cambridge, UK, 1988.

[12] L. J. Cummings, G. Richardson, and M. B. Amar, *Models of void electromigration*, European J. Appl. Math., 12 (2001), pp. 97–134.

[13] C. M. Elliott and H. Garcke, *On the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Math. Anal., 27 (1996), pp. 404–423.

[14] C. M. Elliott and H. Garcke, *Existence results for diffusive surface motion laws*, Adv. Math. Sci. Appl., 7 (1997), pp. 465–488.

[15] C. M. Elliott and A. R. Gardiner, *One dimensional phase field computations*, in Numerical Analysis 1993, D. F. Griffiths and G. A. Watson, eds., Longman, Harlow, 1994, pp. 56–74.

[16] G. Grün, *On the convergence of entropy consistent schemes for lubrication type equations in multiple space dimensions*, Math. Comp., 72 (2003), pp. 1251–1279.

[17] G. Grün and M. Rumpf, *Nonnegativity preserving numerical schemes for the thin film equation*, Numer. Math., 87 (2000), pp. 113–152.

[18] P. S. Ho, *Motion of inclusion induced by a direct current and temperature gradient*, J. Appl. Phys., 41 (1970), pp. 64–68.

[19] O. Kraft and E. Arzt, *Electromigration mechanisms in conductor lines: Void shape changes and slit-like failure*, Acta Mater., 45 (1997), pp. 1599–1611.

[20] Z. Li, H. Zhao, and H. Gao, *A numerical study of electro-migration voiding by evolving level set functions on a fixed cartesian grid*, J. Comput. Phys., 152 (1999), pp. 281–304.

[21] M. Mahadevan and R. M. Bradley, *Phase field model of surface electromigration in single crystal metal thin films*, Phys. D, 126 (1999), pp. 201–213.

[22] M. Mahadevan and R. M. Bradley, *Simulations and theory of electromigration induced slit formation in unpassivated single-crystal metal lines*, Phys. Rev. B, 59 (1999), pp. 11037–11046.

[23] R. Nürnberg, *Finite Element Approximation of some Degenerate Nonlinear Parabolic Systems*, Ph.D. thesis, University of London, London, 2003.

[24] A. Schmidt and K. G. Siebert, *Albert — software for scientific computations and applications*, Acta Math. Univ. Comenian. (N.S.), 70 (2001), pp. 105–122.

[25] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge University Press, Cambridge, UK, 1999.

[26] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura. Appl. (4), 146 (1987), pp. 65–96.

[27] R. Temam, *Navier-Stokes Equations*, AMS Chelsea, Providence, RI, 2001.

[28] L. Xia, A. F. Bower, Z. Suo, and C. Shih, *A finite element analysis of the motion and evolution of voids due to strain and electromigration induced surface diffusion*, J. Mech. Phys. Solids, 45 (1997), pp. 1473–1493.

[29] L. Zhornitskaya and A. L. Bertozzi, *Positivity-preserving numerical schemes for lubrication-type equations*, SIAM J. Numer. Anal., 37 (2000), pp. 523–555.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 773–799

SURFACE DIFFUSION OF GRAPHS: VARIATIONAL FORMULATION, ERROR ANALYSIS, AND SIMULATION*

EBERHARD BÄNSCH†, PEDRO MORIN‡, AND RICARDO H. NOCHETTO§

Abstract. Surface diffusion is a (fourth-order highly nonlinear) geometric driven motion of a surface with normal velocity proportional to the surface Laplacian of mean curvature. We present a novel variational formulation for graphs and derive a priori error estimates for a time-continuous finite element discretization. We also introduce a semi-implicit time discretization and a Schur complement approach to solve the resulting fully discrete, linear systems. After computational verification of the orders of convergence for polynomial degrees 1 and 2, we show several simulations in one dimension and two dimensions with and without forcing which explore the smoothing effect of surface diffusion, as well as the onset of singularities in finite time, such as infinite slopes and cracks.

Key words. surface diffusion, fourth-order parabolic problem, finite elements, a priori error estimates, Schur complement, smoothing effect

AMS subject classifications. 35K55, 65M12, 65M15, 65M60, 65Z05

DOI. 10.1137/S0036142902419272

1. Introduction.

Controlling morphological changes in stressed epitaxial films is of paramount importance in materials science. The film may be thought of as subjected to mechanical stresses to model its misfit with the crystalline structure of the substrate. This in turn causes a plastic deformation of the free surface of the film, a morphological instability of the free surface which may eventually lead to crack formation and fracture. The simplest model couples surface diffusion of the free surface with linear elasticity in the bulk [1, 6, 13, 14, 15, 18, 19, 20]. Investigating this complicated nonlinear dynamics requires effective and reliable computational tools.

This paper studies the *geometric* motion law of *surface diffusion* with given forcing but without elasticity. The dynamics of the free surface $\Gamma(t)$ is thus governed by the (highly nonlinear) fourth-order geometric PDE

$$V = -\Delta_S(\kappa + f), \tag{1.1}$$

where V is the normal velocity of $\Gamma(t)$, κ is its mean curvature, and Δ_S is the Laplace–Beltrami operator on $\Gamma(t)$. In this reduced model f is given, whereas in the full model f corresponds to the elastic energy density of the bulk $\Omega(t)$ restricted to $\Gamma(t)$. Our goal is to present novel variational formulations and finite element methods for (1.1), which may be viewed as building blocks towards solving the fully coupled system.

*Received by the editors December 10, 2002; accepted for publication (in revised form) October 20, 2003; published electronically June 4, 2004. This research was partially supported by the international cooperation NSF-DAAD grants INT-9910086 and INT-0129243.

http://www.siam.org/journals/sinum/42-2/41927.html

†Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D-10117 Berlin, Germany and Freie Universität Berlin, Arnimallee 2-6, 14195 Berlin, Germany (baensch@wias-berlin.de).

‡Departamento de Matemática, Universidad Nacional del Litoral, Instituto de Matemática Aplicada del Litoral (IMAL), Güemes 3450, 3000 Santa Fe, Argentina (pmorin@math.unl.edu.ar). The research of this author was partially supported by Programa FOMEC de la Universidad Nacional del Litoral and CONICET of Argentina and NSF grant DMS-9971450. This work was partly developed while this author was visiting the University of Maryland.

§Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu). The research of this author was partially supported by NSF grants DMS-9971450 and DMS-0204670.

We study the *graph* case in this paper and the *parametric* case in [3]. From now on we assume that $\Omega \subseteq \mathbb{R}^d$ $(d \geq 1)$ is a fixed domain and $\Gamma(t) := \{(x, u(t,x)) \mid x \in \Omega\} \subseteq \mathbb{R}^{d+1}$ is the free surface for $0 \leq t \leq T$ described by the unknown function u. If $Q = Q(u) = \sqrt{1 + |\nabla u|^2}$ denotes the elementary surface area, then the unit normal ν to $\Gamma(t)$, its mean curvature κ, and the normal velocity V of $\Gamma(t)$ can be expressed as follows:

$$\nu = \frac{1}{Q}(-\nabla u, 1)^T, \qquad \kappa = \nabla \cdot \left(\frac{\nabla u}{Q}\right), \qquad V = \frac{\partial_t u}{Q}.$$

Therefore, (1.1) can be written as the following system of second-order nonlinear PDEs:

$$\frac{\partial_t u}{Q} = -\Delta_S(\kappa + f), \qquad \kappa = \nabla \cdot \left(\frac{\nabla u}{Q}\right), \tag{1.2}$$

for (u, κ). Once completed with initial and boundary conditions, this system constitutes our starting point. Issues about existence, uniqueness, and regularity are not yet settled, not even for the graph formulation; we refer the reader to [12] for local existence for closed surfaces, as well as global existence and exponential asymptotic behavior for solutions close to a sphere. It is known, however, that the graph property may be lost in finite time [11], an intriguing situation corroborated by simulations in section 7.3.

We introduce in section 2 a new variational formulation with several crucial stability properties. Using C^0 finite elements of any degree $k \geq 1$, we obtain a space discretization in section 3 with solutions (u_h, κ_h) and show corresponding stability properties. After deriving a number of auxiliary results for the semidiscrete scheme in section 4, we use them to prove the quasi-optimal estimate in section 5 for the errors $e_u = u - u_h$ and $e_\kappa = \kappa - \kappa_h$:

$$\sup_{t\in[0,T]} \left(\|e_u\|^2_{L^2(\Omega)} + \int_{\Gamma_h(t)} |\nabla_S e_u|^2 \right) + \int_0^T \left(\|e_\kappa\|^2_{L^2(\Omega)} + \int_{\Gamma_h(t)} |\nabla_S e_\kappa|^2 \right) \leq C\, h^{2k}. \tag{1.3}$$

Here $C > 0$ depends on the regularity of u and κ, $k \geq 1$ is the polynomial degree, and h is the mesh size. It is worth comparing our results with the existing literature. A space-time finite element method for *axially symmetric* surfaces is presented by Coleman, Falk, and Moakher in [7], along with several stability properties and very interesting dynamics, some not predicted by linearized stability. More recently, Deckelnick, Dziuk, and Elliott provided an error analysis [9] for the axially symmetric case. Our formulation, discretization, and analysis differ from those in [7, 9].

In section 6 we introduce a *semi-implicit* time discretization in the spirit of Deckelnick and Dziuk [8] and Dziuk [10]. This leads to a sequence of surfaces Γ^n and *linear* elliptic PDEs on them. We derive again several crucial stability properties and discuss a Schur complement approach for doing effective numerical linear algebra. Finally, we show a number of numerical experiments in section 7. Their purpose is twofold: first, we computationally verify the rate (1.3) for $k = 1, 2$, and, second, we explore the nonlinear regime of (1.1) via simulation. In fact, we examine the regularizing effect of surface diffusion, as well as whether (1.1) is capable of forming singularities. They manifest themselves as vertical slopes $|\nabla u| = \infty$ for $f = 0$ and cracks for $f \neq 0$ of a special form. We display results for both $d = 1, 2$ computed with the finite element toolbox ALBERT [17, 16].

2. Variational formulation. In this section we write (1.2) in weak form. We start with some notation and basic formulas.

2.1. Elementary differential geometry. Let $v, w : \Omega \to \mathbb{R}$ be (smooth) functions. Since the area element is given by Q, then

$$\int_\Gamma v = \int_\Omega v\, Q;$$

in particular, the area $A(t)$ of $\Gamma(t)$ reads $A(t) = \int_\Omega Q$ at time t. If $\tilde{v}$ is the trivial extension of v to $\mathbb{R}^{d+1}$, namely, $\tilde{v}(x_1, \dots, x_{d+1}) := v(x_1, \dots, x_d)$, then the *tangential* gradient ∇_S is given by

$$\nabla_S v = \nabla_{d+1}\tilde{v} - \nabla_{d+1}\tilde{v} \cdot \nu\, \nu,$$

where ∇_{d+1} denotes the gradient in $\mathbb{R}^{d+1}$. Since $\nabla_{d+1}\tilde{v} = (\nabla v^T, 0)^T$, we readily get

$$\nabla_S v \cdot \nabla_S w = \nabla v \cdot \nabla w - \frac{1}{Q^2} \nabla v \cdot \nabla u\ \nabla w \cdot \nabla u.$$

Note that there is also an intrinsic definition of ∇_S. If $\gamma = \partial\Gamma$ indicates the boundary of Γ, then this expression, together with integration by parts, yields

$$\begin{aligned} -\int_\Gamma \Delta_S v\, w + \int_\gamma \underline{\partial}_{\nu_\gamma} v\, w &= \int_\Gamma \nabla_S v \cdot \nabla_S w = \int_\Omega \nabla_S v \cdot \nabla_S w\, Q \\ &= \int_\Omega \left(\nabla v \cdot \nabla w\, Q - \frac{\nabla v \cdot \nabla u\, \nabla w \cdot \nabla u}{Q} \right). \end{aligned} \tag{2.1}$$

Here ν_γ denotes the intrinsic outer unit normal of Γ at γ, given by $\nu_\gamma = \nu_\Gamma \wedge \tau_\gamma$ with τ_γ the tangential unit vector of γ with the appropriate sign for $\Gamma \subseteq \mathbb{R}^3$.

2.2. Boundary conditions and function spaces. Let $L^p(\Omega)$, $1 \le p \le \infty$, be the usual space of Lebesgue measurable functions with norm $\|v\|_p := (\int_\Omega |v|^p)^{1/p}$. By $\langle \cdot, \cdot \rangle$ we denote the L^2 inner product $\langle v, w \rangle := \int_\Omega vw$ for $v, w \in L^2(\Omega)$. We indicate with $H^{m,p}(\Omega)$ the Sobolev space of functions in $L^p(\Omega)$ with mth weak derivatives also in $L^p(\Omega)$ equipped with the norm $\|v\|_{m,p} := (\sum_{|\alpha| \le m} \int_\Omega |\partial_\alpha v|^p)^{1/p}$ and $H^m := H^{m,2}$. Furthermore, $\mathring{H}^1(\Omega)^p$ is the subspace of functions in $H^{1,p}$ with vanishing boundary values in the sense of traces.

Finally, for a time interval $[0, T]$ and a function space V we define the parabolic spaces $L^p(V)$ of V-valued functions that are measurable in time with $\|v\|_{L^p(V)} := (\int_0^T \|v(t)\|_V^p dt)^{1/p} < \infty$.

To simplify the notation we will write $\|v\|_\infty = \|v\|_{L^\infty(L^\infty)}$. This ambiguity of notation will not lead to confusion.

We now discuss boundary conditions and corresponding function spaces $\mathcal{X}$.

Periodic boundary condition. Let $\Omega = \Pi_{i=1}^d (0, X_i)$ be a parallelogram. If $u(t, x + X_i e_i) = u(t, x), \kappa(t, x + X_i e_i) = \kappa(t, x)$ for all $x \in \partial\Omega$ and $1 \le i \le d$, then

$$\mathcal{X} := \{ v \in H^1(\Omega) \mid v(x + X_i e_i) = v(x) \quad \text{for } x \in \partial\Omega,\ 1 \le i \le d \}.$$

Neumann boundary condition. If $\nu_\gamma \cdot \nabla_S u(t, x) = \nu_\gamma \cdot \nabla_S \kappa(t, x) = 0$ for $x \in \partial\Omega$, then $\mathcal{X} := H^1(\Omega)$.

Dirichlet boundary condition. If $u(t, x) = \kappa(t, x) = 0$ for $x \in \partial\Omega$, then $\mathcal{X} := \mathring{H}^1(\Omega)$.

2.3. Weak form. We are now in a position to introduce two bilinear forms in (v, w) and state the variational formulation of (1.2). Let

$$\boldsymbol{a}(u; v, w) := \int_\Omega \left(\nabla v \cdot \nabla w \, Q - \frac{\nabla v \cdot \nabla u \, \nabla w \cdot \nabla u}{Q} \right), \tag{2.2}$$

$$\tilde{\boldsymbol{a}}(u; v, w) := \int_\Omega \frac{\nabla v \cdot \nabla w}{Q}. \tag{2.3}$$

LEMMA 2.1 (equivalence). *Let $u \in C^1([0,T]; C^4(\bar\Omega))$, let $\kappa \in C^0([0,T]; C^2(\bar\Omega))$, and let $\mathcal{X}$ be as defined in section 2.2. Then (u, κ) is a solution of (1.2) with initial value u_0 and boundary conditions as in section 2.2 iff $u(t)$, $\kappa(t) \in \mathcal{X}$ for all $t \in [0,T]$, $u(0,\cdot) = u_0$, and*

$$\langle \partial_t u, \psi \rangle - \boldsymbol{a}(u; \kappa, \psi) = \boldsymbol{a}(u; f, \psi) \qquad \forall \psi \in \mathcal{X}, \tag{2.4}$$

$$\langle \kappa, \varphi \rangle + \tilde{\boldsymbol{a}}(u; u, \varphi) = 0 \qquad \forall \varphi \in \mathcal{X}. \tag{2.5}$$

Proof. Multiply the first equation in (1.2) by ψ, integrate over Γ, and use formula (2.1). Observe that the boundary term vanishes because of the choice of function space $\mathcal{X}$. Equation (2.5) follows similarly from the second equation in (1.2) integrating by parts over Ω. ∎

Remark 2.2 (mean curvature flow). In contrast to the mean curvature flow, for which a divergence formulation reads [8, 10]

$$\int_\Omega \frac{\partial_t u \, v}{Q} + \tilde{\boldsymbol{a}}(u; u, v) = 0 \qquad \forall\, v \in \mathcal{X},$$

we do not have the factor $\frac{1}{Q}$ in the parabolic term.

Remark 2.3 (comparing $\boldsymbol{a}$ and $\tilde{\boldsymbol{a}}$). The forms $\boldsymbol{a}$ and $\tilde{\boldsymbol{a}}$ are symmetric and nonnegative. If $d = 1$, they coincide; i.e., $\boldsymbol{a}(u; \cdot, \cdot) = \tilde{\boldsymbol{a}}(u; \cdot, \cdot)$. If $d > 1$, instead,

$$\boldsymbol{a}(u; v, u) = \tilde{\boldsymbol{a}}(u; v, u) \qquad \forall\, v \in \mathcal{X}$$

because $Q(1 - \frac{|\nabla u|^2}{Q^2}) = \frac{1}{Q}$. Similarly,

$$\boldsymbol{a}(u; v, v) = \int_\Gamma \nabla_S v \cdot \nabla_S v = \int_\Omega \left(|\nabla v|^2 Q - \frac{|\nabla v \cdot \nabla u|^2}{Q} \right) \geq \int_\Omega \frac{|\nabla v|^2}{Q} = \tilde{\boldsymbol{a}}(u; v, v).$$

Remark 2.4 (equivalent forms of $\boldsymbol{a}$). Let $\zeta := \frac{\nabla u}{|\nabla u|}$ be a unit vector in the direction of ∇u, provided that $\nabla u \neq 0$, and be arbitrary otherwise. Let $(\chi_i)_{i=1}^{d-1}$ be a complementary orthonormal set perpendicular to ζ. A simple calculation then yields

$$\boldsymbol{a}(u; v, w) = \int_\Omega \left(\frac{\nabla v \cdot \zeta \, \nabla w \cdot \zeta}{Q} + Q \sum_{i=1}^{d-1} \nabla v \cdot \chi_i \, \nabla w \cdot \chi_i \right) \quad \forall v, w \in \mathcal{X}. \tag{2.6}$$

Another equivalent form is obtained using $\otimes$ to denote the tensor product in $\mathbb{R}^d$:

$$\boldsymbol{a}(u; v, w) = \int_\Omega \nabla v^T \left(Q\, I - \frac{\nabla u \otimes \nabla u}{Q} \right) \nabla w; \tag{2.7}$$

here I denotes the identity matrix in $\mathbb{R}^d$.

Remark 2.5 (volume conservation and area decrease). If the function $v = 1 \in \mathcal{X}$, then (2.4) yields $0 = \langle \partial_t u, 1\rangle = \frac{d}{dt}\int_\Omega u$, which is the formula for conservation of volume. On the other hand, if the forcing term $f \equiv 0$, then the area of $\Gamma(t)$ is decreasing regardless of boundary conditions (see Lemma 2.6). Both of these properties will also hold true for the semidiscrete and fully discrete formulations of section 3 and section 6.

With the help of the above variational form of the equations, we are in a position to prove a stability result for the continuous solution.

LEMMA 2.6 (continuous stability). *Let (u, κ) be a solution of* (1.2) *fulfilling the assumptions of Lemma* 2.1, *and let $A(t)$ denote the area of $\Gamma(t)$. There are two constants, $C_1 = C_1(\Omega)$ and $C_2 = C_2\big(\|\nabla f\|_\infty, A(0)\big)$, such that*

$$\sup_{t\in[0,T]} \|u(t)\|_2^2 + \int_0^T \|\kappa\|_2^2 \le \|u(0)\|_2^2 + C_1 \int_0^T \|\nabla f\|_2,$$

$$\sup_{t\in[0,T]} A(t) + \int_0^T \boldsymbol{a}(u;\kappa,\kappa) \le C_2.$$

Moreover, if $f \equiv 0$, then the function $A(t)$ is decreasing (strictly provided that $\Delta_S \kappa \not\equiv 0$).

Proof. We omit the proof because it is the same as that of Proposition 3.2. ∎

3. Space discretization. Let $(\mathcal{T}_h)_{h>0}$ be a family of (possibly graded) shape regular triangulations of Ω, with h being the largest size of elements in $\mathcal{T}_h$. We fix $k \in \mathbb{N}$ and denote by $\mathcal{X}_h \subseteq \mathcal{X}$ the subspace of continuous finite elements of polynomial degree k with appropriate boundary conditions. Let $I_h : \mathcal{X} \cap C^0(\bar\Omega) \to \mathcal{X}_h$ be an interpolation operator fulfilling

$$\|I_h v - v\|_p + h\|\nabla(I_h - v)\|_p \le C\, h^{k+1}\|v\|_{k+1,p} \tag{3.1}$$

for $1 \le p \le \infty$ and $v \in H^{k+1,p}(\Omega)$ [5]. We will not need an inverse estimate for the error analysis and thus do not require quasi uniformity of the underlying meshes.

DEFINITION 3.1 (semidiscrete solution). *A pair u_h, κ_h with $u_h \in C^1([0,T], \mathcal{X}_h)$, $\kappa_h \in C^0([0,T]; \mathcal{X}_h)$ is called a* semidiscrete solution *of* (1.2) *if $u_h(0,\cdot) = I_h u_0$ and*

$$\langle \partial_t u_h, \psi_h\rangle - \boldsymbol{a}(u_h;\kappa_h,\psi_h) = \boldsymbol{a}(u_h; f, \psi_h) \qquad \forall \psi_h \in \mathcal{X}_h, \tag{3.2}$$

$$\langle \kappa_h, \varphi_h\rangle + \tilde{\boldsymbol{a}}(u_h; u_h, \varphi_h) = 0 \qquad \forall \varphi_h \in \mathcal{X}_h. \tag{3.3}$$

From now on we consider $d \ge 2$; the analysis for $d = 1$ is just a simplified version of this case. We recall from Remark 2.4 that $\{\zeta, \chi_1, \dots, \chi_{d-1}\}$ is a set of orthonormal vectors for which (2.6) holds. If $\{\zeta_h, \chi_{h,1}, \dots, \chi_{h,d-1}\}$ denotes likewise a semidiscrete orthonormal set and $Q_h = Q(u_h)$, then

$$\boldsymbol{a}(u_h; v, w) = \int_\Omega \left(\frac{\nabla v\cdot\zeta_h\, \nabla w\cdot\zeta_h}{Q_h} + \sum_{i=1}^{d-1} \nabla v\cdot\chi_{h,i}\, \nabla w\cdot\chi_{h,i}\, Q_h \right). \tag{3.4}$$

PROPOSITION 3.2 (semidiscrete stability). *Let (u_h, κ_h) be a semidiscrete solution in the sense of Definition* 3.1, *and let $A_h(t) := \int_\Omega Q_h$ denote the area of the surface $\Gamma_h(t) := \{(x, u_h(t,x)) \,|\, x \in \Omega\}$. There are two constants, $C_1 = C_1(\Omega)$ and*

$C_2 = C_2\big(\|\nabla f\|_\infty, A_h(0)\big)$, such that

$$\sup_{t\in[0,T]} \|u_h(t)\|_2^2 + \int_0^T \|\kappa_h\|_2^2 \le \|u_h(0)\|_2^2 + C_1 \int_0^T \|\nabla f\|_2, \tag{3.5}$$

$$\sup_{t\in[0,T]} A_h(t) + \int_0^T \boldsymbol{a}(u_h;\kappa_h,\kappa_h) \le C_2. \tag{3.6}$$

Moreover, if $f \equiv 0$, the function $A_h(t)$ is decreasing (strictly if $\boldsymbol{a}(u_h;\kappa_h,\kappa_h) > 0$).

Proof. First, choose $\psi_h := u_h$, $\varphi_h := \kappa_h$ as test functions in (3.2) and (3.3), respectively. In view of Remark 2.3, we get

$$\langle \partial_t u_h, u_h\rangle + \langle \kappa_h, \kappa_h\rangle + \underbrace{\tilde{\boldsymbol{a}}(u_h;u_h,\kappa_h) - \boldsymbol{a}(u_h;\kappa_h,u_h)}_{=0} = \boldsymbol{a}(u_h;f,u_h),$$

and, since $|\nabla u_h|/Q_h \le 1$,

$$\boldsymbol{a}(u_h;f,u_h) = \tilde{\boldsymbol{a}}(u_h;f,u_h) = \int_\Omega \frac{\nabla f\cdot\nabla u_h}{Q_h} \le \|\nabla f\|_2 \left(\int_\Omega \frac{|\nabla u_h|^2}{Q_h^2}\right)^{1/2} \le C_1\|\nabla f\|_2.$$

Integrating in time gives (3.5). We next set $\psi_h := -\kappa_h$, $\varphi_h := \partial_t u_h$ to derive

$$\underbrace{-\langle \partial_t u_h, \kappa_h\rangle + \langle \kappa_h, \partial_t u_h\rangle}_{=0} + \boldsymbol{a}(u_h;\kappa_h,\kappa_h) + \tilde{\boldsymbol{a}}(u_h;u_h,\partial_t u_h) = -\boldsymbol{a}(u_h;f,\kappa_h).$$

Observing that

$$\tilde{\boldsymbol{a}}(u_h;u_h,\partial_t u_h) = \int_\Omega \frac{\nabla u_h\cdot\nabla\partial_t u_h}{Q_h} = \frac{d}{dt}\int_\Omega Q_h = \frac{d}{dt}A_h(t), \tag{3.7}$$

we get

$$\frac{d}{dt}A_h(t) + \boldsymbol{a}(u_h;\kappa_h,\kappa_h) = -\boldsymbol{a}(u_h;f,\kappa_h),$$

which implies that $A_h(t)$ is decreasing, provided that $f \equiv 0$. To prove (3.6) for $f \not\equiv 0$, we have to bound $\boldsymbol{a}(u_h;f,\kappa_h)$. Making use of (3.4), we obtain

$$\begin{aligned}
\boldsymbol{a}(u_h;f,\kappa_h) &= \int_\Omega \left(\frac{\nabla f\cdot\zeta_h\,\nabla\kappa_h\cdot\zeta_h}{Q_h} + \sum_{i=1}^{d-1} \nabla f\cdot\chi_{h,i}\,\nabla\kappa_h\cdot\chi_{h,i}\,Q_h\right)\\
&\le \|\nabla f\|_\infty \int_\Omega \left(\frac{|\nabla\kappa_h\cdot\zeta_h|}{Q_h} + \sum_{i=1}^{d-1} |\nabla\kappa_h\cdot\chi_{h,i}|\,Q_h\right)\\
&\le \|\nabla f\|_\infty \left(\frac{|\Omega|}{4\epsilon} + \epsilon\int_\Omega \left(\frac{|\nabla\kappa_h\cdot\zeta_h|^2}{Q_h} + \sum_{i=1}^{d-1} |\nabla\kappa_h\cdot\chi_{h,i}|^2 Q_h\right) + \frac{\int_\Omega Q_h}{4\epsilon}\right)\\
&= \|\nabla f\|_\infty \Big(C_\epsilon + \epsilon\,\boldsymbol{a}(u_h;\kappa_h,\kappa_h) + C_\epsilon A_h(t)\Big),
\end{aligned}$$

where we have used that $Q_h \ge 1$. Choosing ϵ sufficiently small, a Gronwall argument finally yields (3.6). ∎

COROLLARY 3.3 (global existence of semidiscrete solution). *For $h > 0$ and $T > 0$ there is a unique semidiscrete solution (u_h, κ_h) fulfilling* (3.2) *and* (3.3).

Proof. Observing that (3.2)–(3.3) is equivalent to a system of ODEs with a locally Lipschitz right-hand side, we get a local in time existence of the semidiscrete solution. Using the above stability estimate, this solution can be extended to the time interval $[0,T]$ by standard arguments. Uniqueness follows from the local Lipschitz continuity of the right-hand side. ∎

4. Auxiliary estimates. In this section we present some auxiliary lemmas and results that will be instrumental in deriving the error estimates. Since they will be used several times and might be of independent interest, we present them separately.

We start by introducing the following notation:

$$e_u := u - u_h, \qquad e_\kappa := \kappa - \kappa_h, \qquad N_h := \int_\Omega |\nu - \nu_h|^2 Q_h.$$

LEMMA 4.1 (basic geometric formulas). *Using the notation introduced above, the following inequalities hold:*

$$\left|\frac{1}{Q} - \frac{1}{Q_h}\right| \le |\nu - \nu_h|, \qquad |Q - Q_h| \le Q Q_h |\nu - \nu_h|, \tag{4.1}$$

and

$$\left|\frac{\nabla u \otimes \nabla u}{Q} - \frac{\nabla u_h \otimes \nabla u_h}{Q_h}\right| \le 3\, Q\, Q_h\, |\nu - \nu_h|. \tag{4.2}$$

Proof. Recalling that $\nu = \frac{1}{Q}(\nabla u, -1)^T$ and $\nu_h = \frac{1}{Q_h}(\nabla u_h, -1)^T$, the inequalities in (4.1) are immediate. To prove (4.2), let us introduce the notation $z := \frac{\nabla u}{Q}$ and $z_h := \frac{\nabla u_h}{Q_h}$, and observe that

$$\begin{aligned}\frac{\nabla u \otimes \nabla u}{Q} - \frac{\nabla u_h \otimes \nabla u_h}{Q_h} &= z \otimes z\, Q - z_h \otimes z_h\, Q_h \\ &= (z - z_h) \otimes z\, Q + z_h \otimes z\, (Q - Q_h) + z_h \otimes (z - z_h)\, Q_h.\end{aligned}$$

Therefore, the triangle inequality and the fact that $|z - z_h| \le |\nu - \nu_h|$ yield (4.2). ∎

The following lemma is crucial for our error analysis and provides a coercivity estimate for $\tilde{\boldsymbol{a}}$. The estimate is the same as the one that appears in the error analysis for mean curvature flow and is due to Deckelnick and Dziuk [8] and Dziuk [10]. Even though its proof can be found in [8, p. 347], we sketch it here for completeness.

LEMMA 4.2 (coercivity of $\tilde{\boldsymbol{a}}$). *The following estimate holds true:*

$$\tilde{\boldsymbol{a}}(u; u, \partial_t e_u) - \tilde{\boldsymbol{a}}(u_h; u_h, \partial_t e_u) \ge \frac{1}{2}\frac{d}{dt} N_h(t) - \|\nabla \partial_t u(t)\|_\infty N_h(t).$$

Proof. We start with two geometric relations which follow by simple calculation:

$$1 - \frac{1 + \nabla u \cdot \nabla u_h}{Q Q_h} = \frac{1}{2}\big|\nu - \nu_h\big|^2, \quad \left|\left(\frac{1}{Q} - \frac{1}{Q_h}\right)\left(\frac{\nabla u}{Q} - \frac{\nabla u_h}{Q_h}\right)\right| \le \frac{1}{2}\big|\nu - \nu_h\big|^2. \tag{4.3}$$

We now use the first equality in (4.3) to realize that

$$\begin{aligned}\frac{1}{2}\partial_t\big(|\nu-\nu_h|^2Q_h\big) &= \partial_t\left(\left(1-\frac{1+\nabla u\cdot\nabla u_h}{QQ_h}\right)Q_h\right)\\ &= \frac{\nabla u_h\cdot\nabla\partial_t u_h}{Q_h}+\frac{\nabla u\cdot\nabla\partial_t u}{Q^3}\big(1+\nabla u\cdot\nabla u_h\big)\\ &\quad -\frac{1}{Q}\big(\nabla u_h\cdot\nabla\partial_t u+\nabla u\cdot\nabla\partial_t u_h\big),\end{aligned}$$

and, upon adding and subtracting $\frac{\nabla u_h\cdot\partial_t\nabla u}{Q_h}$ and reordering terms, we find out that

$$\begin{aligned}\frac{1}{2}\partial_t\big(|\nu-\nu_h|^2Q_h\big) &= \left(\frac{\nabla u}{Q}-\frac{\nabla u_h}{Q_h}\right)\cdot\nabla\partial_t(u-u_h)\\ &\quad -\partial_t\nabla u\cdot\left(\frac{\nabla u}{Q}-\frac{\nabla u_h}{Q_h}+\frac{\nabla u_h}{Q}-\frac{1+\nabla u\cdot\nabla u_h}{Q^2}\frac{\nabla u}{Q}\right).\end{aligned}$$

We next integrate over Ω, use the definition of N_h, and add and subtract $\nabla\partial_t u\cdot\nabla u\frac{Q_h}{Q^2}$ to obtain

$$\begin{aligned}\tilde{\boldsymbol{a}}(u;u,\partial_t e_u)-\tilde{\boldsymbol{a}}(u_h;u_h,\partial_t e_u) &= \int_\Omega\left(\frac{\nabla u}{Q}-\frac{\nabla u_h}{Q_h}\right)\cdot\nabla\partial_t(u-u_h)\\ &= \frac{1}{2}\frac{d}{dt}\int_\Omega|\nu-\nu_h|^2Q_h\\ &\quad +\int_\Omega\partial_t\nabla u\cdot\left(\frac{\nabla u}{Q}-\frac{\nabla u_h}{Q_h}\right)\left(\frac{1}{Q_h}-\frac{1}{Q}\right)Q_h\\ &\quad +\int_\Omega\partial_t\nabla u\cdot\frac{\nabla u}{Q^2}\left(1-\frac{1+\nabla u\cdot\nabla u_h}{QQ_h}\right)Q_h\\ &\geq \frac{1}{2}\frac{d}{dt}N_h(t)-\|\nabla\partial_t u(t)\|_\infty N_h(t),\end{aligned}$$

where we have employed both estimates (4.3). This finally concludes the proof. ∎

The following two lemmas are consistency estimates for the bilinear forms $\boldsymbol{a}$ and $\tilde{\boldsymbol{a}}$, respectively.

LEMMA 4.3 (consistency estimate for $\boldsymbol{a}$). *For every $\epsilon>0$ there exists a constant $C=C(\epsilon,\|\nabla\kappa\|_\infty,\|Q\|_\infty,\|\nabla f\|_\infty)>0$ such that*

$$|\boldsymbol{a}(u;\kappa,w)-\boldsymbol{a}(u_h;\kappa_h,w)| \;\leq\; \epsilon\,\boldsymbol{a}(u_h;e_\kappa,e_\kappa)\;+\;C\,\|\nabla w\|_\infty^2\;+\;N_h(t)\quad\forall w\in\mathcal{X}.$$

Proof. We first add and subtract the term $\boldsymbol{a}(u_h;\kappa,w)$ to obtain

$$\boldsymbol{a}(u;\kappa,w)-\boldsymbol{a}(u_h;\kappa_h,w)=\boldsymbol{a}(u_h;\kappa-\kappa_h,w)+\big(\boldsymbol{a}(u;\kappa,w)-\boldsymbol{a}(u_h;\kappa,w)\big)=:(I)+(II)$$

and analyze (I) and (II) separately. By the Cauchy–Schwarz inequality,

$$(I)\leq\epsilon\,\boldsymbol{a}(u_h;e_\kappa,e_\kappa)+\frac{1}{4\epsilon}\boldsymbol{a}(u_h;w,w),$$

and, using the definition (2.2) of $\boldsymbol{a}(u_h;\cdot,\cdot)$, we get

$$\boldsymbol{a}(u_h;w,w)=\int_\Omega|\nabla w|^2Q_h-\frac{1}{Q_h}|\nabla w\cdot\nabla u_h|^2\leq\int_\Omega|\nabla w|^2Q_h\leq\|\nabla w\|_\infty^2A_h(t).$$

Therefore,

$$(I) \le \epsilon\, \boldsymbol{a}(u_h; e_\kappa, e_\kappa) + \frac{1}{4\epsilon}\|\nabla w\|_\infty^2 A_h(t).$$

We now turn to estimate (II). Using the equivalent form (2.7) for $\boldsymbol{a}$, we have

$$(II) = \int_\Omega \nabla\kappa^T \left((Q - Q_h)I - \left(\frac{\nabla u \otimes \nabla u}{Q} - \frac{\nabla u_h \otimes \nabla u_h}{Q_h} \right) \right) \nabla w.$$

By (4.1) and (4.2), the integrand in (II) is bounded by $4\,Q\,Q_h\,|\nabla\kappa|\,|\nabla w|\,|\nu - \nu_h|$, which by the Cauchy–Schwarz inequality gives

$$\begin{aligned}(II) &\le 4\int_\Omega Q^2 |\nabla\kappa|^2 |\nabla w|^2 Q_h + \int_\Omega |\nu - \nu_h|^2 Q_h \\ &\le 4\|Q\|_\infty^2 \|\nabla\kappa\|_\infty^2 \|\nabla w\|_\infty^2 A_h(t) + N_h(t).\end{aligned}$$

Since $A_h(t) \le C$ from (3.6), the bounds for (I) and (II) yield the assertion. ∎

LEMMA 4.4 (consistency estimate for $\tilde{\boldsymbol{a}}$). *For every $\epsilon > 0$ we have*

$$|\tilde{\boldsymbol{a}}(u; u, w) - \tilde{\boldsymbol{a}}(u_h; u_h, w)| \le \epsilon \tilde{\boldsymbol{a}}(u_h; w, w) + \frac{1}{4\epsilon} N_h(t) \quad \forall w \in \mathcal{X}.$$

Proof. Using the definition (2.3) of $\tilde{\boldsymbol{a}}$ and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}|\tilde{\boldsymbol{a}}(u; u, w) - \tilde{\boldsymbol{a}}(u_h; u_h, w)| &\le \int_\Omega \left| \frac{\nabla u}{Q} - \frac{\nabla u_h}{Q_h} \right| |\nabla w| \le \int_\Omega |\nu - \nu_h|\,|\nabla w| \\ &\le \epsilon \int_\Omega \frac{|\nabla w|^2}{Q_h} + \frac{1}{4\epsilon} N_h(t) = \epsilon\, \tilde{\boldsymbol{a}}(u_h; w, w) + \frac{1}{4\epsilon} N_h(t),\end{aligned}$$

which is the desired estimate. ∎

The following lemma establishes another consistency estimate for $\boldsymbol{a}$, this time provided that solely the *nonlinear* part of $\boldsymbol{a}$ changes.

LEMMA 4.5. *There exists a constant $C = C(\|Q\|_\infty) > 0$ such that for every $\epsilon > 0$*

$$|\boldsymbol{a}(u; v, w) - \boldsymbol{a}(u_h; v, w)| \le \epsilon\, \boldsymbol{a}(u_h; w, w) + \frac{C}{\epsilon}\|\nabla v\|_\infty^2 N_h(t) \qquad \forall v, w \in \mathcal{X}.$$

Proof. With $R := \|Q\|_\infty$, we consider the following disjoint splitting of Ω: $\Omega = \Omega^+ \cup \Omega^-$ with $\Omega^+ := \{x \in \Omega \,|\, Q_h(x) > 2R\}$ and $\Omega^- := \{x \in \Omega \,|\, Q_h(x) \le 2R\}$.

We first estimate the integrand of $\boldsymbol{a}(u; \cdot, \cdot) - \boldsymbol{a}(u_h; \cdot, \cdot)$ in the case $x \in \Omega^-$. According to (2.7), we consider this integrand written in the form

$$\nabla v^T \left((Q - Q_h)I - \left(\frac{\nabla u \otimes \nabla u}{Q} - \frac{\nabla u_h \otimes \nabla u_h}{Q_h} \right) \right) \nabla w =: (I).$$

Since $Q_h(x) \le 2R$ for $x \in \Omega^-$, in view of (4.1) and (4.2) we have

$$\begin{aligned}(I) &\le 4|\nabla v|\,|\nabla w|\,Q\,Q_h|\nu - \nu_h| \le 8R^2 |\nabla v| \frac{|\nabla w|}{\sqrt{Q_h}} |\nu - \nu_h| \sqrt{Q_h} \\ &\le \epsilon \frac{|\nabla w|^2}{Q_h} + 16 \frac{R^4}{\epsilon} \|\nabla v\|_\infty^2 |\nu - \nu_h|^2 Q_h.\end{aligned}$$

To analyze the case $x \in \Omega^+$, we choose ζ, ζ_h, χ_i, $\chi_{h,i}$ as in Remark 2.4. Since $Q(x) \le R$ and $Q_h(x) > 2R$ we have

$$|\nu - \nu_h| \ge \frac{1}{Q} - \frac{1}{Q_h} \ge \frac{1}{2R} \quad \text{and} \quad |\zeta - \zeta_h|\,, |\chi_i - \chi_{h,i}| \le 2 \le 4R|\nu - \nu_h|. \tag{4.4}$$

Consider the integrand in the form (2.6):

$$\underbrace{\left(\frac{\nabla v\cdot\zeta\,\nabla w\cdot\zeta}{Q} - \frac{\nabla v\cdot\zeta_h\,\nabla w\cdot\zeta_h}{Q_h}\right)}_{(II)} + \sum_{i=1}^{d-1} \underbrace{\nabla w^T\big(\chi_i\otimes\chi_i\,Q - \chi_{h,i}\otimes\chi_{h,i}\,Q_h\big)\nabla v}_{(III)_i}.$$

Since $R \ge 1$, we have for (II)

$$\begin{aligned}
(II) &= \nabla v^T\left(\frac{\zeta\otimes\zeta}{Q} - \frac{\zeta_h\otimes\zeta_h}{Q_h}\right)\nabla w \\
&= \nabla v^T\left[(\zeta-\zeta_h)\otimes\zeta\frac{1}{Q} + \zeta_h\otimes\zeta\left(\frac{1}{Q}-\frac{1}{Q_h}\right) + \zeta_h\otimes(\zeta-\zeta_h)\frac{1}{Q_h}\right]\nabla w \\
&\le C\,R\,|\nabla v|\,|\nabla w|\,|\nu-\nu_h| \le \frac{C^2\,R^2}{4\epsilon}\|\nabla v\|_\infty^2\,|\nu-\nu_h|^2 Q_h + \epsilon\frac{|\nabla w|^2}{Q_h}.
\end{aligned}$$

For $(III)_i$, instead, we proceed as follows with the aid of (4.1):

$$\begin{aligned}
(III)_i &= \nabla w^T\Big((\chi_i-\chi_{h,i})\otimes\chi_i\,Q + \chi_{h,i}\otimes\chi_i\,(Q-Q_h) + \chi_{h,i}\otimes(\chi_i-\chi_{h,i})Q_h\Big)\nabla v \\
&\le 4\,R^2\,|\nu-\nu_h|\,|\nabla v|\,|\nabla w| + 5R\,|\nu-\nu_h|Q_h|\chi_{h,i}\cdot\nabla w|\,|\nabla v| \\
&\le \epsilon\frac{|\nabla w|^2}{Q_h} + \epsilon|\nabla w\cdot\chi_{h,i}|^2 Q_h + \frac{CR^4}{\epsilon}\|\nabla v\|_\infty^2|\nu-\nu_h|^2 Q_h.
\end{aligned}$$

Collecting the estimates for both cases, $x \in \Omega^-$ and $x \in \Omega^+$, integrating over Ω, and recalling (2.6), we obtain the assertion after relabeling ϵ. ∎

Using Lemma 4.5 we obtain a coercivity estimate for $\boldsymbol{a}$.

COROLLARY 4.6 (coercivity of $\boldsymbol{a}$). *There exists $C = C(\|Q\|_\infty) > 0$ such that*

$$\boldsymbol{a}(u;\kappa,e_\kappa) - \boldsymbol{a}(u_h;\kappa_h,e_\kappa) \ge \frac{1}{2}\boldsymbol{a}(u_h;e_\kappa,e_\kappa) - C\|\nabla\kappa\|_\infty^2 N_h(t).$$

Proof. Adding and subtracting $\boldsymbol{a}(u_h;\kappa,e_\kappa)$, and using Lemma 4.5 with $\epsilon = 1/2$, we readily obtain the desired estimate. ∎

LEMMA 4.7 (coercivity of $N_h(t)$). *There exists $C = C(\|Q\|_\infty)$ such that*

$$\boldsymbol{a}(u_h;e_u,e_u) \le C N_h(t).$$

Proof. In light of Remark 2.4, we can write

$$\boldsymbol{a}(u_h;e_u,e_u) = \int_\Omega \underbrace{\frac{|\nabla e_u\cdot\zeta_h|^2}{Q_h}}_{(I)} + \sum_{i=1}^{d-1}\int_\Omega \underbrace{|\nabla e_u\cdot\chi_{h,i}|^2 Q_h}_{(II)_i}.$$

By virtue of (4.1), (I) satisfies

$$\begin{aligned}
(I) &\le \frac{|\nabla e_u|^2}{Q_h} = \frac{|\nabla u - \nabla u_h|^2}{Q_h} \le \frac{|\nu Q - \nu_h Q_h|^2}{Q_h} \\
&\le \frac{|\nu(Q-Q_h) + (\nu-\nu_h)Q_h|^2}{Q_h}Q_h \le 4\|Q\|_\infty^2\,Q_h|\nu-\nu_h|^2.
\end{aligned} \tag{4.5}$$

To treat the integrand $(II)_i$ we again split Ω into Ω^- and Ω^+, as in Lemma 4.5. Consider first $x \in \Omega^-$, namely, $Q_h(x) \le 2R$ with $R := \|Q(t)\|_\infty$. As in (4.5), we get

$$|\nabla e_u \cdot \chi_{h,i}|^2 Q_h \le 4R^2 \frac{|\nabla e_u|^2}{Q_h} \le 16 \|Q\|_\infty^4 Q_h |\nu - \nu_h|^2.$$

Now we consider $Q_h(x) > 2R$. Since $\nabla u_h \cdot \chi_{h,i} = 0$, it follows from (4.4) that

$$\nabla e_u \cdot \chi_{h,i} = \nabla(u - u_h) \cdot \chi_{h,i} = \nabla u \cdot \chi_{h,i} \le |\nabla u| \le 2\, R\, \|\nabla u\|_\infty |\nu - \nu_h|,$$

whence

$$|\nabla e_u \cdot \chi_{h,i}|^2 Q_h \le 4 \|Q\|_\infty^4 |\nu - \nu_h|^2 Q_h.$$

The desired estimate then follows by integration over Ω. ∎

5. A priori error analysis. In this section we prove the main theoretical result of this article, which can be stated as follows.

THEOREM 5.1. *Let (u_h, κ_h) be the semidiscrete solution of Definition* 3.1*, and let $e_u := u - u_h$, $e_\kappa := \kappa - \kappa_h$. There exists a constant C depending on $\|\nabla f\|_\infty$, $\|\partial_t u\|_{L^2(H^{k+1}(\Omega))}$, $\|\partial_t \nabla u\|_\infty$, $\|\kappa\|_{L^2(H^{k+1}(\Omega))}$, $\|\partial_t \kappa\|_{L^2(H^k(\Omega))}$, and $\|\nabla \kappa\|_\infty$ such that*

$$\sup_{t\in[0,T]} \left(\|e_u(t)\|_2^2 + \int_{\Gamma_h(t)} |\nabla_S e_u|^2 \right) + \int_0^T \left(\|e_\kappa\|_2^2 + \int_{\Gamma_h(t)} |\nabla_S e_\kappa|^2 \right) \le C\, h^{2k}.$$

The proof of Theorem 5.1 is a consequence of two estimates, the strong and the weak estimates, derived from the error equations (5.1) and (5.2) below by choosing appropriate test functions.

Remark 5.2 (H^1 estimate). The estimate for $\nabla_S e_u$ might seem surprising at first sight since direct H^1 estimates are unavailable for minimal surfaces for dimension $d > 2$. It is thus worth stressing that, instead of the usual H^1 norm, we have an integral over the discrete surface $\Gamma_h(t)$ which involves the tangential gradient ∇_S. In its derivation, we exploit parabolicity to prove first an estimate for $N_h(t)$ and then use Lemma 4.7 (see section 5.1).

Remark 5.3 (regularity). The required regularity of (u, κ) in Theorem 5.1 might appear inconsistent with that of Ω for polynomial degree $k > 1$. In fact, we have assumed that Ω can be partitioned exactly into finite elements but not that Ω is polyhedral (see section 3). This would thus entail the use of isoparametric elements for $k > 1$, but still $\Omega = \Omega_h$ might fail to hold. Accounting for the effect of $\Omega \ne \Omega_h$ is mostly a technical issue and is, therefore, omitted in the subsequent discussion, which is already rather technical.

5.1. Proof of Theorem 5.1. Subtracting (3.2) and (3.3) from (2.4) and (2.5), respectively, we get the following error equations:

$$\langle \partial_t e_u, \psi_h \rangle - (\boldsymbol{a}(u; \kappa, \psi_h) - \boldsymbol{a}(u_h; \kappa_h, \psi_h)) = \boldsymbol{a}(u; f, \psi_h) - \boldsymbol{a}(u_h; f, \psi_h), \tag{5.1}$$

$$\langle e_\kappa, \varphi_h \rangle + (\tilde{\boldsymbol{a}}(u; u, \varphi_h) - \tilde{\boldsymbol{a}}(u_h; u_h, \varphi_h)) = 0 \tag{5.2}$$

for all $\psi_h, \varphi_h \in \mathcal{X}_h$. The strong and weak estimates below are formulated in terms of the following interpolation errors, which can be bounded via (3.1):

$$\rho_u := u - I_h u, \qquad \rho_\kappa := \kappa - I_h \kappa_h. \tag{5.3}$$

The strong estimate of section 5.2 reads as follows: *For all $\epsilon > 0$ there exists a constant C_0 depending only on $\|\nabla f\|_\infty, \|\nabla\kappa\|_\infty, \|\partial_t\nabla u\|_\infty$, and ϵ such that for $t \in [0,T]$*

$$
\begin{aligned}
N_h(t) + \int_0^t \boldsymbol{a}(u_h; e_\kappa, e_\kappa) &\le N_h(0) + C_0 \int_0^t \Big(N_h + \|e_u\|_2^2 \Big) \\
&+ 2\epsilon \left(\|e_u(t)\|_2^2 + \int_0^t \|e_\kappa\|_2^2 \right) \\
&+ \frac{1}{2\epsilon} \|\rho_\kappa(t)\|_2^2 + \|e_u(0)\|_2^2 + \|\rho_\kappa(0)\|_2^2 \\
&+ C_0 \int_0^t \Big(\|\nabla\rho_\kappa\|_\infty^2 + \|\partial_t\nabla\rho_u\|_2^2 + \|\partial_t\rho_u\|_2^2 + \|\partial_t\rho_\kappa\|_2^2 \Big).
\end{aligned} \tag{5.4}
$$

It is clear that to close the argument we need separate control on the term multiplied by ϵ of the right-hand side of (5.4). This is provided by the weak estimate of section 5.3, which reads as follows: *There exist constants C_1, C_2 depending on $\|\nabla f\|_\infty$ and $\|Q\|_\infty$ such that for $t \in [0,T]$ we have*

$$
\begin{aligned}
\frac{1}{2}\|e_u(t)\|_2^2 + \int_0^t \|e_\kappa\|_2^2 &\le \frac{1}{2}\|e_u(0)\|_2^2 + \int_0^t \|e_u\|_2^2 \\
&+ C_1 \int_0^t N_h + \frac{1}{2}\int_0^t \boldsymbol{a}(u_h; e_\kappa, e_\kappa) \\
&+ \int_0^t \Big(\|\nabla\rho_\kappa\|_2^2 + \|\rho_\kappa\|_2^2 + \|\partial_t\rho_u\|_2^2 \Big) \\
&+ 2\|\rho_u(t)\|_2^2 + C_2 \int_0^t \|\nabla\rho_u\|_\infty^2.
\end{aligned} \tag{5.5}
$$

To prove Theorem 5.1 we add (5.4) and (5.5) and then choose $\epsilon = 1/8$ to eliminate $\|e_u(t)\|_2^2 + \int_0^t \|e_\kappa\|_2^2$ from the right-hand side. Employing a Gronwall argument, we can also remove $\int_0^t \big(N_h(s) + \|e_u\|_2^2\big)$ from the right-hand side at the expense of an exponential depending on C_0, C_1, and T. Finally, Lemma 4.7, in conjunction with $\boldsymbol{a}(u_h; v, v) = \int_{\Gamma_h} \nabla_S v \cdot \nabla_S v$, yields the left-hand side of the asserted estimate. Its right-hand side and underlying a priori regularity result from applying (3.1) to the terms involving ρ_κ, ρ_u defined in (3.1).

5.2. Strong estimate (5.4). To prove (5.4), we choose the discrete functions

$$
\begin{aligned}
-\psi_h &:= I_h\kappa - \kappa_h = (\kappa - \kappa_h) + (I_h\kappa - \kappa) = e_\kappa - \rho_\kappa \in \mathcal{X}_h, \\
\varphi_h &:= \partial_t(I_h u - u_h) = \partial_t e_u - \partial_t\rho_u \in \mathcal{X}_h.
\end{aligned}
$$

Adding (5.1) and (5.2), and invoking Lemma 4.2 and Corollary 4.6, we get

$$
\begin{aligned}
\frac{1}{2}\frac{d}{dt}N_h(t) &+ \frac{1}{2}\boldsymbol{a}(u_h; e_\kappa, e_\kappa) - C\, N_h(t) \\
&\le \boldsymbol{a}(u; \kappa, e_\kappa) - \boldsymbol{a}(u_h; \kappa_h, e_\kappa) + \tilde{\boldsymbol{a}}(u; u, \partial_t e_u) - \tilde{\boldsymbol{a}}(u_h; u_h, \partial_t e_u) \\
&= \big(\boldsymbol{a}(u; \kappa, \rho_\kappa) - \boldsymbol{a}(u_h; \kappa_h, \rho_\kappa)\big) - \big(\boldsymbol{a}(u; f, e_\kappa - \rho_\kappa) - \boldsymbol{a}(u_h; f, e_\kappa - \rho_\kappa)\big) \\
&\quad + \big(\tilde{\boldsymbol{a}}(u; u, \partial_t\rho_u) - \tilde{\boldsymbol{a}}(u_h; u_h, \partial_t\rho_u)\big) - \langle \partial_t e_u, \rho_\kappa\rangle + \langle e_\kappa, \partial_t\rho_u\rangle \\
&=: (I) + (II) + (III) + (IV) + (V),
\end{aligned}
$$

with C depending only on $\|\partial_t \nabla u\|_\infty$, $\|\nabla \kappa\|_\infty$, and $\|Q\|_\infty$. We now proceed to estimate each term on the right-hand side separately.

By Lemma 4.3, there is a constant $C = C(\epsilon, \|\nabla\kappa\|_\infty, \|Q\|_\infty, \|\nabla f\|_\infty)$ such that

$$\big|(I)\big| \le \epsilon\, \boldsymbol{a}(u_h; e_\kappa, e_\kappa) + C\|\nabla \rho_\kappa\|_\infty^2 + N_h(t).$$

Using Lemma 4.4 with $\epsilon = 1$, we obtain

$$\big|(III)\big| \le \tilde{\boldsymbol{a}}(u_h; \partial_t \rho_u, \partial_t \rho_u) + N_h(t) \le \|\nabla \partial_t \rho_u\|_2^2 + N_h(t).$$

For any $t \in [0,T]$ we integrate (IV) by parts on $[0,t]$, thereby obtaining

$$\begin{aligned}\int_0^t (IV) &= \langle e_u(0), \rho_\kappa(0)\rangle - \langle e_u(t), \rho_\kappa(t)\rangle + \int_0^t \langle e_u, \partial_t \rho_\kappa\rangle \\ &\le \frac{1}{2}\|e_u(0)\|_2^2 + \frac{1}{2}\|\rho_\kappa(0)\|_2^2 + \frac{\epsilon}{2}\|e_u(t)\|_2^2 + \frac{1}{2\epsilon}\|\rho_\kappa(t)\|_2^2 + \frac{1}{2}\int_0^t \Big(\|e_u\|_2^2 + \|\partial_t \rho_\kappa\|_2^2\Big).\end{aligned}$$

For (V) we readily have

$$|(V)| \le \frac{\epsilon}{2}\|e_\kappa\|_2^2 + \frac{1}{2\epsilon}\|\partial_t \rho_u\|_2^2.$$

We decompose (II) into discretization and interpolation errors as follows:

$$-(II) = \underbrace{\big(\boldsymbol{a}(u; f, e_\kappa) - \boldsymbol{a}(u_h; f, e_\kappa)\big)}_{(II)_e} - \underbrace{\big(\boldsymbol{a}(u; f, \rho_\kappa) - \boldsymbol{a}(u_h; f, \rho_\kappa)\big)}_{(II)_\rho}. \tag{5.6}$$

In light of Lemma 4.5, there is a constant $C = C(\|Q\|_\infty)$ such that

$$\big|(II)_e\big| \le \frac{1}{4}\,\boldsymbol{a}(u_h; e_\kappa, e_\kappa) + C\|\nabla f\|_\infty^2 N_h(t).$$

Using Lemma 4.1 and (3.6), we find a constant $C = C(\|\nabla f\|_\infty, \|Q\|_\infty, A_h(0))$ such that

$$\begin{aligned}\big|(II)_\rho\big| &= \left|\int_\Omega \nabla f^T \left((Q - Q_h)I - \left(\frac{\nabla u \otimes \nabla u}{Q} - \frac{\nabla u_h \otimes \nabla u_h}{Q_h}\right)\right)\nabla \rho_\kappa\right| \\ &\le 4\int_\Omega |\nabla f|\,|\nabla \rho_\kappa|\,Q\,Q_h\,|\nu - \nu_h| \\ &\le 4\|\nabla f\|_\infty^2 \|\nabla \rho_\kappa\|_\infty^2 \|Q\|_\infty^2 \int_\Omega Q_h + \int_\Omega |\nu - \nu_h|^2 Q_h \\ &\le C\|\nabla \rho_\kappa\|_\infty^2 A_h(t) + N_h(t) \le C\|\nabla \rho_\kappa\|_\infty^2 + N_h(t).\end{aligned}$$

Finally, collecting the above estimates for (I) to (V), subtracting $\frac{1}{4}\,\boldsymbol{a}(u_h; e_\kappa, e_\kappa)$, and integrating in time from 0 to $t \in [0,T]$, we arrive at (5.4).

5.3. Weak estimate (5.5). To prove (5.5), we choose the discrete functions

$$\begin{aligned}\psi_h &:= I_h u - u_h = e_u - \rho_u \in \mathcal{X}_h, \\ \varphi_h &:= I_h \kappa - \kappa_h = e_\kappa - \rho_\kappa \in \mathcal{X}_h.\end{aligned}$$

Adding the error equations (5.1) and (5.2), we obtain

(5.7)
$$\begin{aligned}\langle \partial_t e_u, e_u\rangle + \langle e_\kappa, e_\kappa\rangle \;=\; & \big(\boldsymbol{a}(u;\kappa,e_u) - \boldsymbol{a}(u_h;\kappa_h,e_u)\big) - \big(\tilde{\boldsymbol{a}}(u;u,e_\kappa) - \tilde{\boldsymbol{a}}(u_h;u_h,e_\kappa)\big)\\ & + \langle \partial_t e_u, \rho_u\rangle - \big(\boldsymbol{a}(u;\kappa,\rho_u) - \boldsymbol{a}(u_h;\kappa_h,\rho_u)\big)\\ & + \big(\boldsymbol{a}(u;f,e_u-\rho_u) - \boldsymbol{a}(u_h;f,e_u-\rho_u)\big) + \langle e_\kappa,\rho_\kappa\rangle\\ & + \big(\tilde{\boldsymbol{a}}(u;u,\rho_\kappa) - \tilde{\boldsymbol{a}}(u_h;u_h,\rho_\kappa)\big)\\ =:\; & (I)+\cdots+(VII).\end{aligned}$$

We proceed now to bound each term from (I) to (VII) separately.

Adding and subtracting $\boldsymbol{a}(u_h;\kappa,e_u)$ to (I), and employing Lemma 4.5 with $\epsilon=\frac{1}{6}$, we readily have

$$\begin{aligned}|(I)| &\le |\boldsymbol{a}(u;\kappa,e_u)-\boldsymbol{a}(u_h;\kappa,e_u)| + |\boldsymbol{a}(u_h;e_\kappa,e_u)|\\ &\le C\|\nabla\kappa\|_\infty^2 N_h(t) + \frac{5}{2}\,\boldsymbol{a}(u_h;e_u,e_u) + \frac{1}{6}\,\boldsymbol{a}(u_h;e_\kappa,e_\kappa).\end{aligned}$$

Consequently, Lemma 4.7 yields the following bound with $C=C(\|\nabla\kappa\|_\infty,\|Q\|_\infty)$:

$$\big|(I)\big| \le C N_h(t) + \frac{1}{6}\,\boldsymbol{a}(u_h;e_\kappa,e_\kappa).$$

Making use of Lemma 4.4 and Remark 2.3, we readily deduce (using $\epsilon=\frac{1}{6}$)

$$\big|(II)\big| \le \frac{1}{6}\,\tilde{\boldsymbol{a}}(u_h;e_\kappa,e_\kappa) + \frac{3}{2}N_h(t) \le \frac{1}{6}\,\boldsymbol{a}(u_h;e_\kappa,e_\kappa) + \frac{3}{2}N_h(t),$$

as well as (using $\epsilon=\frac{1}{2}$)

$$\big|(VII)\big| \le \frac{1}{2}\tilde{\boldsymbol{a}}(u_h;\rho_\kappa,\rho_\kappa) + \frac{1}{2}N_h(t) \le \frac{1}{2}\|\nabla\rho_\kappa\|_2^2 + \frac{1}{2}N_h(t).$$

Using Lemma 4.3 with $\epsilon=\frac{1}{6}$ we find a constant $C=C(\|\nabla\kappa\|_\infty,\|Q\|_\infty,\|\nabla f\|_\infty)$ such that

$$\big|(IV)\big| \le \frac{1}{6}\,\boldsymbol{a}(u_h;e_\kappa,e_\kappa) + N_h(t) + C\|\nabla\rho_u\|_\infty^2.$$

For (VI), we obviously have $\big|(VI)\big| \le \frac{1}{2}\|e_\kappa\|_2^2 + \frac{1}{2}\|\rho_\kappa\|_2^2$. For (III), instead, we integrate by parts on $[0,t]$ for any $t\in[0,T]$ to obtain

$$\begin{aligned}\int_0^t (III) &= \langle e_u(t),\rho_u(t)\rangle - \langle e_u(0),\rho_u(0)\rangle - \int_0^t \langle e_u,\partial_t\rho_u\rangle\\ &\le \frac{1}{4}\,\|e_u(t)\|_2^2 + \|\rho_u(t)\|_2^2 + \frac{1}{2}\int_0^t \Big(\|e_u\|_2^2 + \|\partial_t\rho_u\|_2^2\Big).\end{aligned}$$

It remains to bound (V), which involves the right-hand side f. Applying Lemma 4.5 (with $\epsilon=1$) and Lemma 4.7, we obtain

$$\big|\boldsymbol{a}(u;f,e_u) - \boldsymbol{a}(u_h;f,e_u)\big| \le C\|\nabla f\|_\infty^2 N_h(t) + \boldsymbol{a}(u_h;e_u,e_u) \le C N_h(t),$$

with $C=C(\|\nabla f\|_\infty,\|Q\|_\infty)$. Since $\boldsymbol{a}(u;f,\rho_u)-\boldsymbol{a}(u_h;f,\rho_u)$ is similar to $(II)_\rho$ in (5.6), we likewise deduce

$$\big|\boldsymbol{a}(u;f,\rho_u) - \boldsymbol{a}(u_h;f,\rho_u)\big| \le C\|\nabla\rho_u\|_\infty^2 A_h(t) + N_h(t) \le C\|\nabla\rho_u\|_\infty^2 + N_h(t),$$

whence, for C depending on $\|\nabla f\|_\infty$ and $\|Q\|_\infty$, we end up with

$$\big|(V)\big| \le C\|\nabla \rho_u\|_\infty^2 + CN_h(t).$$

Inserting the above bounds for (I) to (VII) back into (5.7) and integrating from 0 to t, we finally obtain the desired estimate (5.5).

6. Full discretization. In this section we introduce the fully discrete scheme actually used in simulations, along with the linear algebra approach to its solution.

6.1. Definition and properties. To discretize in time we subdivide the time interval into $t_0 = 0 < t_1 < \cdots < t_N = T$ and set $\tau_n := t_{n+1} - t_n$. We define the notion of *semi-implicit* fully discrete problem as follows: Set $u_h^0 = u_h(0)$ and for $n = 0, 1, \ldots, N-1$ determine u_h^{n+1}, $\kappa_h^{n+1} \in \mathcal{X}_h$ by

$$\langle u_h^{n+1}, \psi_h\rangle - \tau_n \boldsymbol{a}(u_h^n; \kappa_h^{n+1}, \psi_h) = \tau_n \boldsymbol{a}(u_h^n; f^n, \psi_h) + \langle u_h^n, \psi_h\rangle \qquad \forall \psi_h \in \mathcal{X}_h, \tag{6.1}$$

$$\langle \kappa_h^{n+1}, \varphi_h\rangle + \tilde{\boldsymbol{a}}(u_h^n; u_h^{n+1}, \varphi_h) = 0 \qquad \forall \varphi_h \in \mathcal{X}_h, \tag{6.2}$$

with $f^n := f(t_n)$. Existence and uniqueness of solutions u_h^n, κ_h^n follow from the considerations in section 6.2.

We now establish a stability estimate analogous to (3.6) in Lemma 2.6.

THEOREM 6.1 (fully discrete stability). *Let $(u_h^n, \kappa_h^n)_{n=0}^N$ be a solution of the fully discrete equations* (6.1) *and* (6.2)*, and let $A_h^n := \int_\Omega Q(u_h^n)$ denote the area of the surface $\Gamma_h^n := \{(x, u_h^n(x)) \mid x \in \Omega\}$. There exists $C = C(\|\nabla f\|_\infty, A_h^0)$ such that*

$$\sup_{1\le n\le N} A_h^n + \sum_{n=1}^N \tau_n \int_{\Gamma_h^{n-1}} |\nabla_S \kappa_h^n|^2 \le C. \tag{6.3}$$

Moreover, if $f \equiv 0$, A_h^n is a decreasing sequence (strictly if $\boldsymbol{a}(u_h^{n-1}; \kappa_h^n, \kappa_h^n) > 0$).

Proof. Choose as test functions $-\kappa_h^{n+1}$ and $(u_h^{n+1} - u_h^n)$ in (6.1) and (6.2), respectively, and add both equations. One readily gets

$$\tau_n \boldsymbol{a}(u_h^n; \kappa_h^{n+1}, \kappa_h^{n+1}) + \int_\Omega \frac{\nabla u_h^{n+1} \cdot \nabla (u_h^{n+1} - u_h^n)}{Q(u_h^n)} = -\tau_n \boldsymbol{a}(u_h^n; f^n, \kappa_h^{n+1}). \tag{6.4}$$

The next step consists of finding a discrete counterpart of (3.7). Observing that

$$|a| - |b| \le \frac{a \cdot (a-b)}{|b|} \qquad \forall a, b \in \mathbb{R}^{d+1}$$

and setting $a := (\nabla u_h^{n+1}, 1)^T$, $b := (\nabla u_h^n, 1)^T$, we obtain

$$Q(u_h^{n+1}) - Q(u_h^n) \le \frac{\nabla u_h^{n+1} \cdot \nabla (u_h^{n+1} - u_h^n)}{Q(u_h^n)}.$$

Inserting this into (6.4) gives $A_h^{n+1} \le A_h^n$ if $f \equiv 0$. To prove (6.3) for $f \not\equiv 0$, we have to bound the right-hand side in (6.4). This can be done similarly to (3.5), obtaining

$$|\boldsymbol{a}(u_h^n; f^n, \kappa_h^{n+1})| \le C(1 + A_h^n) \; + \; \epsilon\, \boldsymbol{a}(u_h^n; \kappa_h^{n+1}, \kappa_h^{n+1}),$$

with $C = C(\epsilon, \|\nabla f\|_\infty)$. Multiplying by τ_n, choosing ϵ sufficiently small, summing up over all n, and using a discrete Gronwall argument, the result follows. ∎

6.2. Schur complement strategy. Let $\mathcal{X}_h = \operatorname{span}\{\varphi_j\} \subseteq \mathcal{X}$ with the usual nodal basis functions φ_j and the corresponding nodal space $\underline{\mathbf{X}}$. Then, for the time instant t_{n+1}, the fully discrete system of equations can be rewritten as

$$\begin{bmatrix} \tilde{A} & M \\ M^T & -\tau_n A \end{bmatrix} \begin{bmatrix} \underline{U}^{n+1} \\ \underline{K}^{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ M^T \underline{U}^n + \tau_n \underline{F}^n \end{bmatrix}, \tag{6.5}$$

where $\underline{U}^n$, $\underline{K}^n$ denote the vector of nodal values for u_h^n, κ_h^n, respectively,

$$u_h^n = \sum_j \underline{U}_j^n \, \varphi_j, \qquad \kappa_h^n = \sum_j \underline{K}_j^n \, \varphi_j,$$

the vector $\underline{F}^n$ is given by $\underline{F}_j^n = \boldsymbol{a}(u_h^n; f^n, \varphi_j)$, and the matrices M, A, $\tilde{A}$ are given by

$$M_{i,j} = \langle \varphi_j, \varphi_i \rangle, \qquad A_{i,j} = \boldsymbol{a}(u_h^n; \varphi_j, \varphi_i), \qquad \tilde{A}_{i,j} = \tilde{\boldsymbol{a}}(u_h^n; \varphi_j, \varphi_i).$$

Notice that the matrices A and $\tilde{A}$ depend on u_h^n and thus have to be reassembled in every time step.

To derive a Schur complement formulation, we have to distinguish between the various boundary conditions (see section 2.2).

Dirichlet boundary conditions. In this case, since $\mathcal{X} := \mathring{H}^1(\Omega)$, the matrix $\tilde{A}$ is invertible, and a Schur complement for $\underline{K}^{n+1}$ is thus given by

$$\begin{aligned} \left(M^T \tilde{A}^{-1} M \; + \; \tau_n A \right) \underline{K}^{n+1} &= -M^T \underline{U}^n - \tau_n \underline{F}^n, \\ \tilde{A} \, \underline{U}^{n+1} &= -M \, \underline{K}^{n+1}. \end{aligned}$$

This system is decoupled and uniquely solvable for both $\underline{K}^{n+1}$ and $\underline{U}^{n+1}$.

Periodic and Neumann boundary conditions. This case is a bit more involved because constant functions are in $\mathcal{X}_h$, whence $\tilde{A}$ has a kernel $\ker(\tilde{A}) = \operatorname{span}\{\underline{1}\}$.

Let $\underline{\mathbf{V}}, \underline{\mathbf{W}} \subseteq \underline{\mathbf{X}}$ be the spaces of nodal values for $\underline{U}^{n+1}$ defined by

$$\underline{\mathbf{V}} := \{\underline{V} \mid \underline{1} \cdot MV = 0\}, \qquad \underline{\mathbf{W}} := \{\underline{V} \mid \underline{1} \cdot \underline{V} = 0\} = \operatorname{span}\{\underline{1}\}^{\perp}.$$

Multiplying the first equation in (6.5) by $\underline{1}$, we see that $\underline{1} \cdot M\underline{K}^{n+1} = 0$, which means that $\underline{K}^{n+1} \in \underline{\mathbf{V}}$. Let P be the orthogonal projection onto $\operatorname{span}\{\underline{1}\}$ with respect to the Euclidean inner product in $\mathbb{R}^I$, with $I = \dim \underline{\mathbf{X}}$. If $S := (\tilde{A}_{|\underline{\mathbf{W}}})^{-1}$, then

$$SM\underline{K}^{n+1} = -S\tilde{A}\underline{U}^{n+1} = -(Id - P)\underline{U}^{n+1} = -\underline{U}^{n+1} + P\underline{U}^{n+1},$$

or $\underline{U}^{n+1} = -SM\underline{K}^{n+1} + P\underline{U}^{n+1}$. Consequently, using the second equation in (6.5),

$$(M^T SM + \tau_n A) \, \underline{K}^{n+1} - M^T P \, \underline{U}^{n+1} = -M^T \underline{U}^n - \tau_n \underline{F}^n. \tag{6.6}$$

Now let $\Pi := Id - \frac{M^T \underline{1} \otimes M^T \underline{1}}{|M^T \underline{1}|^2}$ be the orthogonal projection onto $\underline{\mathbf{V}}$. Applying Π to both sides of (6.6) and using that $\Pi M P \underline{U}^{n+1} = 0$ and $\Pi \underline{K}^{n+1} = \underline{K}^{n+1}$, we arrive at

$$\Pi(M^T SM + \tau_n A)\Pi \, \underline{K}^{n+1} = -\Pi(M^T \underline{U}^n + \tau_n \underline{F}^n). \tag{6.7}$$

The matrix $\Pi(M^T SM + \tau_n A)\Pi$ is symmetric and positive definite in $\underline{\mathbf{V}}$, and thus (6.7) is uniquely solvable for $\underline{K}^{n+1}$ in $\underline{\mathbf{V}}$. Finally, $\underline{U}^{n+1}$ is uniquely determined by

$$\tilde{A} \, \underline{U}^{n+1} = -M \, \underline{K}^{n+1}, \qquad \underline{1} \cdot M^T \underline{U}^{n+1} = \underline{1} \cdot M^T \underline{U}^n. \tag{6.8}$$

Note that the last equation is the conservation of volume $\int_\Omega (U^{n+1} - U^n) = 0$ written in matrix-vector form; compare with Remark 2.5.

7. Numerical experiments. The purpose of this section is to document via several experiments the performance of the discretization scheme proposed in this article. We open this section with some comments about the implementation of the algorithm within the flexible adaptive finite element toolbox ALBERT [16, 17]. We continue with a verification of the experimental orders of convergence (EOCs) achieved by the method with different polynomial degrees and relations between time step τ and mesh size h. We next illustrate the smoothing effect of surface diffusion (case $f = 0$), and we finally present simulations driven by a forcing term which exhibit singularity formation in finite time in both one dimension and two dimensions (case $f \neq 0$).

7.1. Implementation. The matrices of section 6 were assembled using the standard assembling tools of ALBERT, and the solution to the linear systems (6.7)–(6.8) was obtained by a conjugate gradient method.

For the assembling of the linear systems, quadrature rules exact for polynomials of degree $2k$ were used, where k is the degree of the finite element. For computing the errors versus the exact solution, quadratures of order $2k + 2$ were used.

For all the experiments presented in this article, domains with periodic boundary conditions were considered. Experiments with other boundary conditions were also carried out and will be shown elsewhere. The results were similar.

7.2. EOCs. To test the performance of the discretization scheme we consider the domain $\Omega = (-1, 1) \times (-1, 1) \subset \mathbb{R}^2$ with the exact solution

$$u(x, y, t) = 1 + 0.1 \sin(\pi x) \sin(2\pi y) \cos(\pi t) \qquad \forall\, t \in [0, 1].$$

The exact curvature $\kappa = \nabla \cdot \left(\frac{\nabla u}{Q}\right)$ and right-hand side $F = \partial_t u - Q\Delta_S \kappa$ were obtained using the symbolic capabilities of Mathematica. The finite element method of section 6 is used to compute (u_h, κ_h), and a comparison with (u, κ) is presented in Tables 7.1–7.4. They display the errors

$$\mathrm{err}_\nu := \sup_{0\le t\le T} \left(\int_\Omega |\nu - \nu_h|^2 Q_h \right)^{1/2}, \qquad \mathrm{err}_u := \sup_{0\le t\le T} \boldsymbol{a}(u_h; e_u, e_u)^{1/2},$$

$$\mathrm{err}_\kappa := \left(\int_0^T \boldsymbol{a}(u_h; e_\kappa, e_\kappa) \right)^{1/2}, \quad \mathrm{err}_{u,2} := \sup_{0\le t\le T} \|e_u\|_2, \quad \mathrm{err}_{\kappa,2} := \left(\int_0^T \|e_\kappa\|_2^2 \right)^{1/2}$$

for different values of h and τ along with the EOCs. Given two meshes with mesh sizes H, h and errors err_H, err_h, respectively, the EOC is determined according to

$$\mathrm{EOC} = \frac{\log(\mathrm{err}_H/\mathrm{err}_h)}{\log(H/h)},$$

which gives the computational exponent k in the expression $\mathrm{err}_h \cong Ch^k$.

In Table 7.1 we show the results obtained using *linear* elements and a time step $\tau = h$. Even though τ seems to be large as compared to h, the convergence rate is still linear, and no instabilities arise. This is not so surprising if we recall that the fully discrete system is *unconditionally* stable (see Theorem 6.1). In order to verify the error analysis in section 5 for the *semidiscretization in space*, we also compute the EOCs for smaller values of τ, namely $\tau = 0.1h$ and $\tau = h^2$; see Tables 7.2 and 7.3. Here again, we observe that the EOCs are at least 1. Moreover, as one would expect, the errors measured in $L^2(\Omega)$ norms are approximately of second order, provided

TABLE 7.1
Linear elements and time step $\tau = h$.

h	err_ν	EOC	err_u	EOC	err_κ	EOC	$\mathrm{err}_{u,2}$	EOC	$\mathrm{err}_{\kappa,2}$	EOC
1/2	0.5601		0.6055		18.2		0.0836		2.1921	
1/4	0.2549	1.14	0.2884	1.07	7.70	1.24	0.0287	1.54	0.4366	2.33
1/8	0.1297	0.97	0.1448	0.99	4.66	0.73	0.0121	1.24	0.1773	1.30
1/16	0.0636	1.03	0.0708	1.03	2.41	0.95	0.0049	1.32	0.0630	1.49
1/32	0.0310	1.03	0.0344	1.04	1.21	0.99	0.0021	1.24	0.0262	1.26

TABLE 7.2
Linear elements and time step $\tau = 0.1h$.

h	err_ν	EOC	err_u	EOC	err_κ	EOC	$\mathrm{err}_{u,2}$	EOC	$\mathrm{err}_{\kappa,2}$	EOC
1/2	0.5594		0.6048		18.4		0.0834		2.2249	
1/4	0.2463	1.18	0.2772	1.13	7.67	1.26	0.0251	1.73	0.4071	2.45
1/8	0.1240	0.99	0.1364	1.02	4.67	0.71	0.0081	1.62	0.1484	1.46
1/16	0.0611	1.02	0.0669	1.03	2.40	0.96	0.0022	1.87	0.0397	1.90
1/32	0.0304	1.01	0.0332	1.01	1.19	1.00	0.0006	1.85	0.0102	1.97

TABLE 7.3
Linear elements and time step $\tau = h^2$.

h	err_ν	EOC	err_u	EOC	err_κ	EOC	$\mathrm{err}_{u,2}$	EOC	$\mathrm{err}_{\kappa,2}$	EOC
1/2	0.5597		0.6051		18.4		0.0835		2.2214	
1/4	0.2470	1.18	0.2782	1.12	7.67	1.26	0.0254	1.71	0.4073	2.45
1/8	0.1240	0.99	0.1365	1.03	4.61	0.73	0.0082	1.63	0.1466	1.47
1/16	0.0611	1.02	0.0669	1.03	2.38	0.96	0.0022	1.93	0.0392	1.90
1/32	0.0304	1.01	0.0332	1.01	1.19	1.00	0.0005	1.98	0.0099	1.99

TABLE 7.4
Quadratic elements and time step $\tau = h^2$.

h	err_ν	EOC	err_u	EOC	err_κ	EOC	$\mathrm{err}_{u,2}$	EOC	$\mathrm{err}_{\kappa,2}$	EOC
1/2	0.1271		0.1376		7.38		0.0101		0.3277	
1/4	0.0419	1.60	0.0487	1.50	2.47	1.58	0.0040	1.35	0.0797	2.04
1/8	0.0102	2.03	0.0122	1.99	0.71	1.80	0.0009	2.19	0.0152	2.39
1/16	0.0025	2.01	0.0030	2.00	0.17	2.07	0.0002	2.11	0.0032	2.24

that $\tau = h^2$; this is not predicted by our theory though. For $\tau = h, 0.1h$ we do not recover second-order errors because the time discretization error—expected to be of first order—dominates the space error in $L^2(\Omega)$ norms.

To further verify experimentally the error estimates of section 5, which are valid for any polynomial degree, we also compute the EOCs for *quadratic* elements. Table 7.4 displays the results obtained with quadratics and $\tau = h^2$. The EOCs are about 2 in all the error norms, as predicted by theory, including those in $L^2(\Omega)$. In fact, the latter cannot exhibit EOCs close to 3 due to the choice of the time step $\tau = h^2$.

7.3. Smoothing effect in one dimension: Case $f \equiv 0$. In this section we present experimental results in $\Omega = (-1, 1)$ concerning the behavior of the discrete solution when $f \equiv 0$ and $u_0(x) = 1 + \delta(x)$ is a perturbation of the stationary solution $u \equiv 1$.

Superposition of sines. We consider the perturbation

$$\delta(x) = 0.1 \sin(\pi x) + 0.3 \sin(16\,\pi x), \tag{7.1}$$

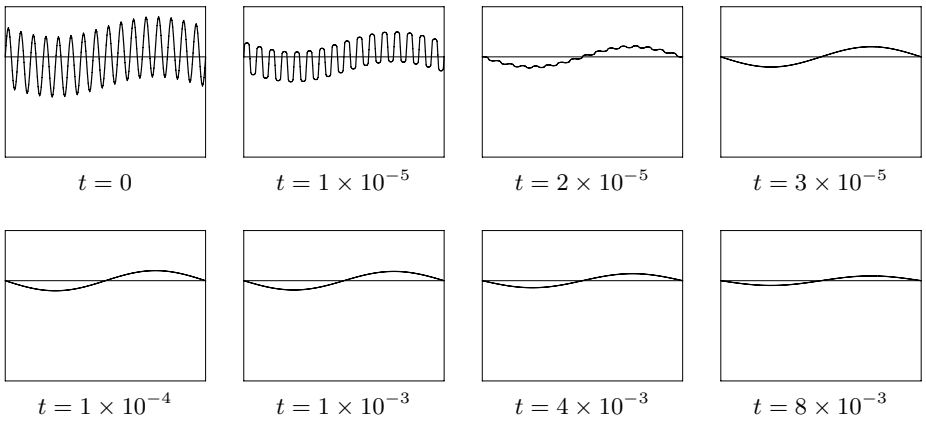


FIG. 7.1. *Solutions for* $f \equiv 0$ *and* $u_0(x) = 1 + 0.1\sin(\pi x) + 0.3\sin(16\pi x)$ *at various instants* t. *In all the plots, the* x*-axis ranges from* -1 *to* 1, *and the* y*-axis ranges from* 0 *to* 1.5.

which results from the superposition of two frequencies. We compute the approximate solution with linear elements and parameters $h = 1/128$, $\tau = 10^{-6}$. This choice of discretization parameters is necessary to reflect the intrinsic time scale for this example. Figure 7.1 depicts the solution for different time instants and shows that high frequencies are rapidly damped, whereas the amplitude of low frequency waves decays *very* slowly. To quantify the difference in the time scales it is worth noting that the time elapsed between the first and the last plot of the first row of Figure 7.1 is 3×10^{-5}, whereas that of the second row is almost 10^{-2}, a difference of three orders of magnitude. This is related to the fourth-order operator of surface diffusion.

Nonnegative perturbation. Let the perturbation be $\delta(x) = 0.3\,\delta_0(0.15x)$, with

$$\delta_0(x) = \min(1, \max(0, 2 - |x|)), \tag{7.2}$$

which is nonnegative and rather singular for this fourth-order flow because of its kinks (see Figure 7.2). We compute the approximate solution with linear elements and parameters $h = 1/128$, $\tau = 10^{-6}$. Figure 7.2 displays the solution for different time instants and confirms the strong smoothing effect of surface diffusion alluded to before. Another important feature that can be visualized in Figure 7.2 is the lack of maximum principle for this equation: we start with a function $u_0 \geq 1$ and, after the first time step, there are already points x with $u(x) < 1$. This is consistent with the fourth-order structure of the operator. It is also worth observing that the spectrum of u_0 is rather full due to the kinks and that high and low frequencies have drastically different decay rates.

Steep perturbation. This example shows that global in time existence may not be expected for a classical solution of (1.1), thereby revealing some limitations of the graph formulation. For $K = 1 + \frac{\sqrt{5}}{2}$, we take the perturbation $\delta(x) = 0.3\delta_0(0.15x)$, with

$$\delta_0(x) = \begin{cases} -K + (1+K)|x| & \text{if } |x| < 1, \\ 2 - |x| & \text{if } 1 \leq |x| < 2, \\ 0 & \text{otherwise.} \end{cases} \tag{7.3}$$

δ is steep, and its mean value vanishes (see Figure 7.3). We compute the approximate solution with linear elements and parameters $h = 1/128$, $\tau = 10^{-7}$. The most impor-

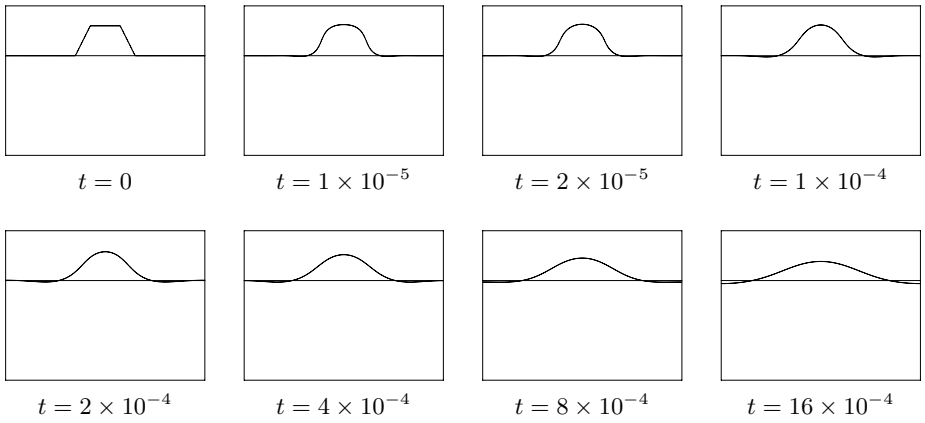


FIG. 7.2. *Solutions for $f \equiv 0$ and $u_0(x) = 1 + \delta(x)$, with $\delta(x)$ a positive perturbation at various times t. In all the plots, the x-axis ranges from -1 to 1, and the y-axis ranges from 0 to 1.5.*

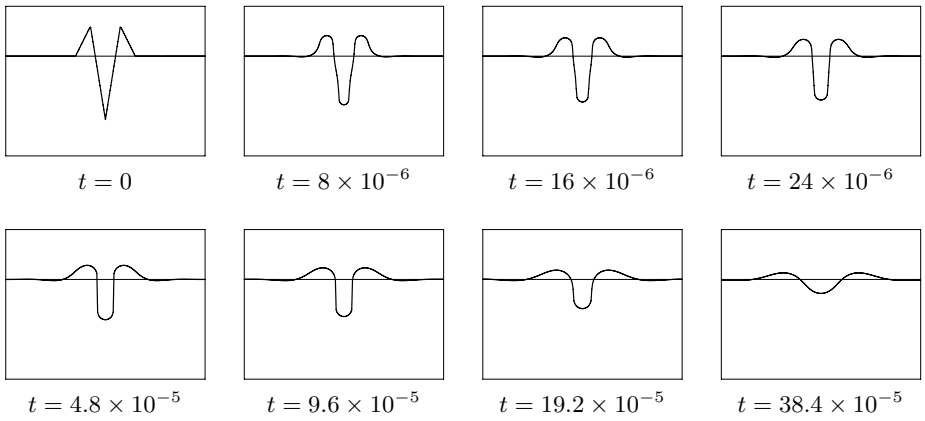


FIG. 7.3. *Solutions for $f \equiv 0$ and $u_0(x) = 1 + \delta(x)$ at various times t, with a steep perturbation $\delta(x)$. In all the plots, the x-axis ranges from -1 to 1, and the y-axis ranges from 0 to 1.5.*

tant features of δ are its steep slope, together with a big jump of the first derivative around $x = 0$. As can be seen in Figure 7.3, the slope seems to become vertical around $t = 4.8 \times 10^{-5}$, which indicates that the classical solution might cease to exist in finite time; in contrast, the *discrete* solution exists globally in time (see section 3). We stress that the lack of smoothness of u_0 plays a secondary role since starting with the (smooth) solution $u(t)$ for some small $t > 0$ would yield the same evolution.

To investigate the formation of singularities in finite time, we use the parametric formulation of [2, 3] with the same initial data; for more examples and details about the discretization for parametric surfaces, we refer the reader to [2, 3]. Since the parametric formulation works for *closed* curves and surfaces, we thus embed the graph of u_0 into a closed curve (see Figure 7.4, top left). For the time scale of Figure 7.3, the effect of this extension is negligible. Figure 7.4 displays a sequence of solutions obtained for the same eight time instants as in Figure 7.3. We see that the parametric evolution by surface diffusion tends to form a *mushroom* starting with this initial condition. Therefore, we conclude that the *continuous* solution will cease to be the graph of a function in finite time; i.e., the exact solution to the graph formulation of

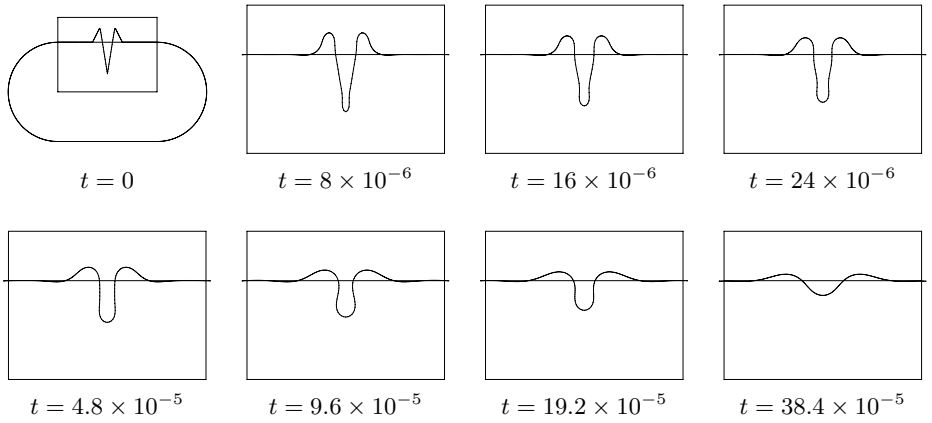


FIG. 7.4. *Solutions obtained with a discretization for parametric curves from* $[2, 3]$ *at the same times as in Figure* 7.3. *In all the plots, the rectangles in thin lines are* $[-1, 1] \times [0, 1.5]$.

surface diffusion exists only *locally* in time for certain initial conditions. To assess the range of validity of the graph formulation, namely, to be able to detect blow-up, time and space adaptivity might be relevant. It is worth noticing the striking similarity of the solutions obtained with both methods. Even though the parametric solution develops a mushroom at $t = 9.6 \times 10^{-5}$, and thus the solution to the graph formulation is questionable thereon, they still exhibit an excellent quantitative agreement for $t > 9.6 \times 10^{-5}$ (compare the last two plots of Figures 7.4 and 7.3).

7.4. Smoothing effect in two dimensions: Case $f \equiv 0$. In this section we present experimental results in $\Omega = (-1, 1) \times (-1, 1)$ concerning the behavior of the discrete solution when $f \equiv 0$ and $u_0(x) = 1 + \delta(x)$ is a perturbation of the solution $u \equiv 1$.

Positive perturbation. We consider a positive perturbation as depicted in Figure 7.5 and compute the approximate solution with linear elements and parameters $h = 1/16$, $\tau = 10^{-6}$. Figure 7.5 displays the solution for different time instants. We observe, as in the one-dimensional case, a strong smoothing effect much faster for high frequencies than for low frequencies, as well as the solution becoming less than 1 (lack of maximum principle).

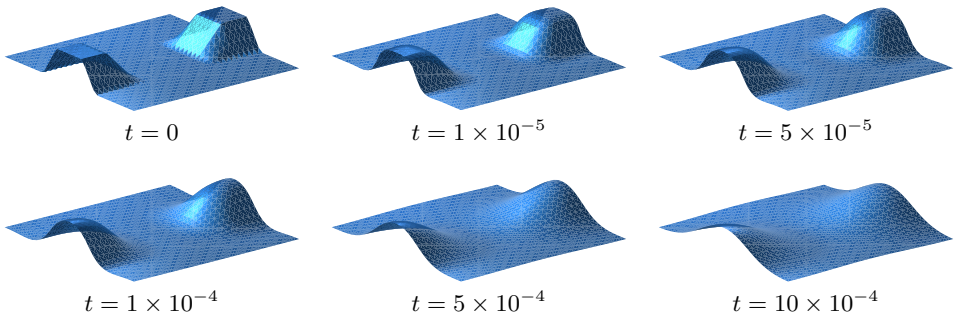


FIG. 7.5. *Solutions for* $f \equiv 0$ *and* $u_0(x) = 1 + \delta(x)$ *at various time instants, with* $\delta(x)$ *a positive perturbation touching the periodic boundary.*

7.5. Crack formation in one dimension: Case $f = -C/u$. We study here the effect of a prescribed forcing of the form $f = -C/u$, which is motivated by the following stationary situation in one dimension and corresponding linearized stability analysis. The nonlinear evolution undergoes two distinct regimes: coarsening and crack formation. We define the latter as the instance when the height u becomes zero at one or more points.

Equilibrium shape of deformable solids. Following [4], we consider a two-dimensional thin solid occupying the domain $\{(x, y) : -1 \le x \le 1\,,\ 0 \le y \le u(x)\}$ and undergoing a plastic deformation due to competition of elastic effects and surface tension with volume constraint $\int_{-1}^{1} u = 2$. The solid is to adjust its shape in order to minimize the following energy:

$$\mathcal{I}(u, v, \lambda) := \int_{-1}^{1} \sqrt{1 + |u_x|^2} + \frac{1}{2} \int_{-1}^{1} u|v_x|^2 - \lambda \left(\int_{-1}^{1} u - 2 \right), \tag{7.4}$$

where $u(x)$ describes the free surface of the film, $v(x)$ is the displacement of the solid, and λ is a Lagrange multiplier associated with the volume constraint. Hence, the first term in (7.4) corresponds to surface tension, whereas the second one is the elastic energy, provided that the displacement v solely depends on the horizontal variable x. Upon variational differentiation with respect to u, v, and λ, the Euler–Lagrange equations turn out to be

$$-\left(\frac{u_x}{\sqrt{1 + |u_x|^2}} \right)_x + \frac{1}{2}|v_x|^2 - \lambda = 0, \qquad (uv_x)_x = 0, \qquad \int_{-1}^{1} u = 2.$$

This immediately yields $v_x = \frac{C}{u}$, whence the equation for u reads

$$-\left(\frac{u_x}{\sqrt{1 + |u_x|^2}} \right)_x + \frac{C}{u^2} - \lambda = 0.$$

Linearized stability analysis. Since $u \equiv 1$ is a solution of (1.1), then a perturbation w of u evolves for a short time according to the linearized PDE around u:

$$\partial_t w = -\Delta(\Delta w + f'(u)w),$$

where $f(u) = -C/u^\gamma$ from the previous discussion, with $\gamma > 0$. Taking an ansatz $w = e^{\mu t} e^{i\pi k x}$ periodic in $(-1, 1)$, we obtain the spectral relation

$$\mu = -(\pi k)^4 + C\gamma(\pi k)^2. \tag{7.5}$$

This implies that $\mu > 0$, provided that $(\pi k)^2 < C\gamma$, whence low frequency perturbations grow and the rest decay for a short time (*linear regime*).

In the simulations below, we make the simplest choice $\gamma = 1$ and take $C = 50$. Our goal is to explore the long-time behavior of (1.1) not predicted by (7.5) (*nonlinear regime*). We discretize the nonlinear forcing term $f(u)$ explicitly, namely, $f_{n+1} = -I_h(C/u_h^n)$, and use linear finite elements with parameters $h = 1/128$, $\tau = 10^{-5}$.

Superposition of sines. We consider the sinusoidal perturbation of (7.1). Figure 7.6 displays the solution at different time instants and shows that high frequencies are rapidly damped, whereas the low frequencies slowly lead to a crack formation. This is consistent with the linearized stability analysis (7.5), according to which the frequency $k = 1$ is the only unstable mode.

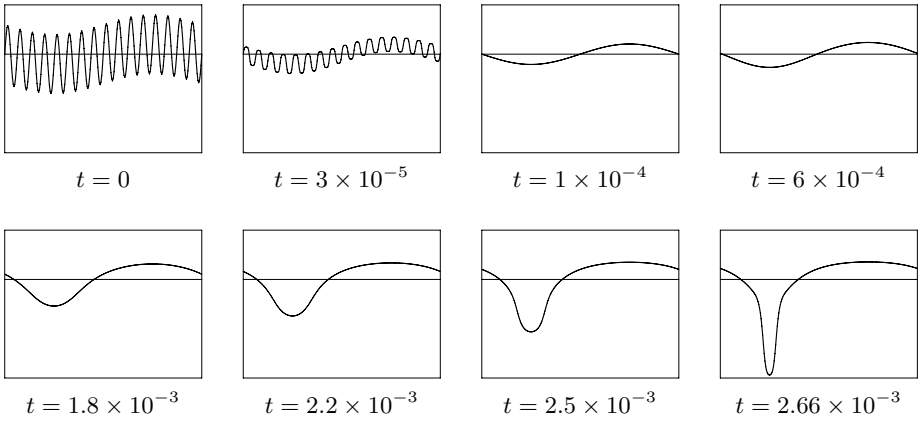


FIG. 7.6. *Solutions for* $f = -50/u$ *and* $u_0(x) = 1 + 0.1\sin(\pi x) + 0.3\sin(16\pi x)$ *at various time instants. In all the plots, the* x*-axis ranges from* -1 *to* 1*, and the* y*-axis ranges from* 0 *to* 1.5.

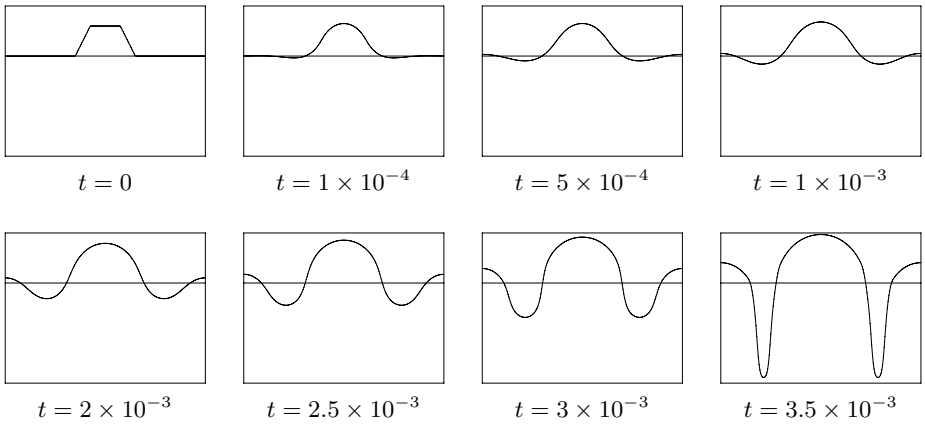


FIG. 7.7. *Solutions for* $f = -50/u$ *and* $u_0(x) = 1 + \delta(x)$ *at various time instants, with* $\delta(x)$ *the positive perturbation of* (7.2)*. In all the plots, the* x*-axis ranges from* -1 *to* 1*, and the* y*-axis ranges from* 0 *to* 1.5.

Positive perturbation. We consider the perturbation δ of (7.2) and display the results in Figure 7.7, which shows an evolution towards crack formation in finite time.

Small perturbation. We consider a perturbation $\delta(x) = 0.1\delta_0(0.02x)$, with δ_0 given in (7.3). Simulations are depicted in Figure 7.8, which shows that by $t = 2 \times 10^{-5}$ the solution is smoothed out. It seems that we have reached a constant equilibrium for a relatively long time $t \cong 7.5 \times 10^{-3}$ (*metastable state*). Then an instability grows, and a fracture starts to form. The latter develops rather fast.

In order to shed light on the actual evolution during the transition between the fast smoothing of the perturbation and the crack development, we show in Figure 7.9 the solution at some time instants between 2×10^{-5} and 7.5×10^{-3}, with the y-axis ranging between 0.998 and 1.001. Even though $u(t)$ looks constant to the eye in Figure 7.8 for t in this interval, a magnification of the y-axis shows that this is not the case: some long waves survive the smoothing effect, and at some point their amplitudes start to increase.

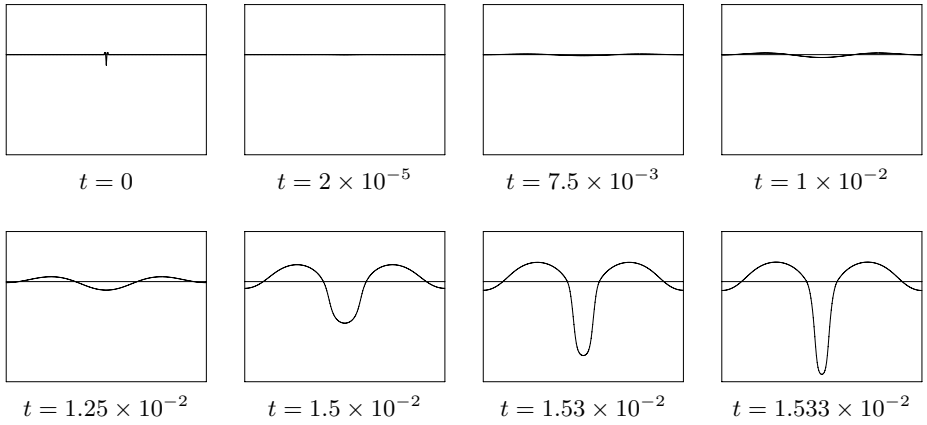


FIG. 7.8. *Solutions for $f = -50/u$ and $u_0(x) = 1 + \delta(x)$ at various time instants, with $\delta(x)$ a small Lipschitz perturbation. In all the plots, the x-axis ranges from -1 to 1, and the y-axis ranges from 0 to 1.5.*

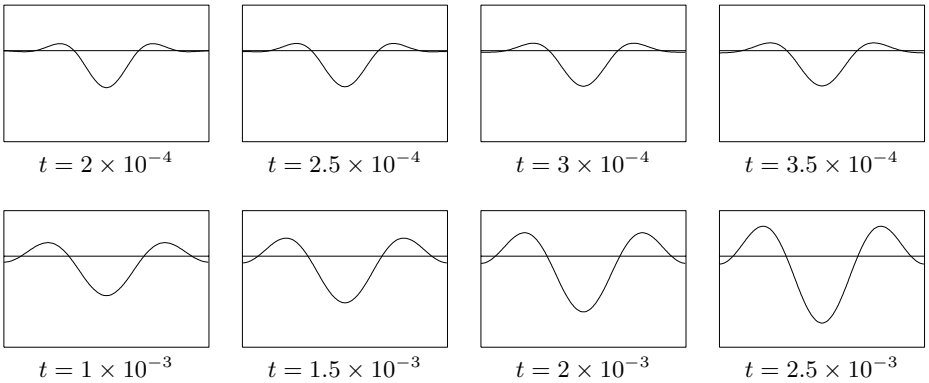


FIG. 7.9. *Solutions for $f = -50/u$ and $u_0(x) = 1 + \delta(x)$ at various time instants between $t = 2 \times 10^{-5}$ and $t = 7.5 \times 10^{-3}$, with the small perturbation of Figure 7.8. In all the plots, the x-axis ranges from -1 to 1, and the y-axis ranges from 0.998 to 1.001.*

Figure 7.10 displays the Fourier modes of $u(t)$ at times $t = 0, 10^{-5}, 10^{-2}, 3 \times 10^{-2}$. We observe that all the modes except the first two decrease immediately, whereas the first two modes increase. This is consistent with the prediction (7.5) of linearized stability because $k^2\pi^2 < 50$ implies $k \leq 2$.

Other simulations, also with forcing $f = -50/u$, do not corroborate this apparent consistency with the linearized stability analysis. We observe that, for a fixed *high* frequency, the solution either develops a crack or tends to the steady solution $u = 1$, depending on the *size* of the perturbation; for instance, if $u_0(x) = 1 + \alpha \sin(4\pi x)$, then a crack forms for $\alpha \geq 0.2375$, thus violating the prediction $k^2\pi^2 < 50$ of (7.5). On the other hand, for a low frequency, the solution always develops a crack regardless of the magnitude of perturbation; for instance, if $u_0(x) = 1 + \alpha \sin(\pi x)$, then a crack forms for all $\alpha \in [0.001, 0.5]$ tested. These simulations will be reported elsewhere. We also refer the reader to [7, 9], where simulations under the assumption of axial symmetry, but without forcing, are performed and singularities are observed as well, which do not conform to the linearized stability analysis either.

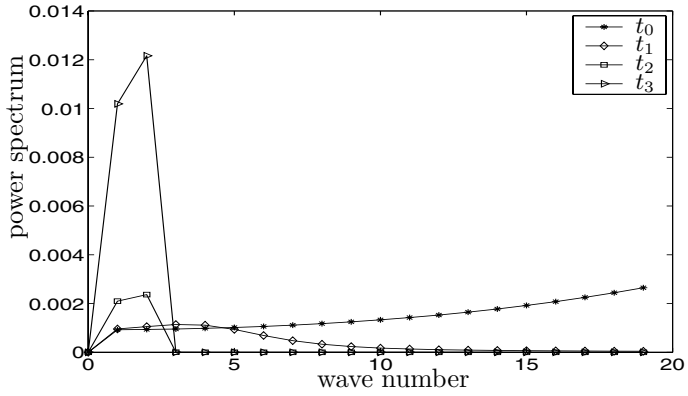


FIG. 7.10. *Power spectrum for the solutions with* $f = -50/u$ *and* $u_0(x) = 1 + \delta(x)$*, with the perturbation* δ *of Figure* 7.8*. The time instants are* $t_0 = 0$*,* $t_1 = 10^{-5}$*,* $t_2 = 10^{-2}$*, and* $t_3 = 3 \times 10^{-2}$*.*

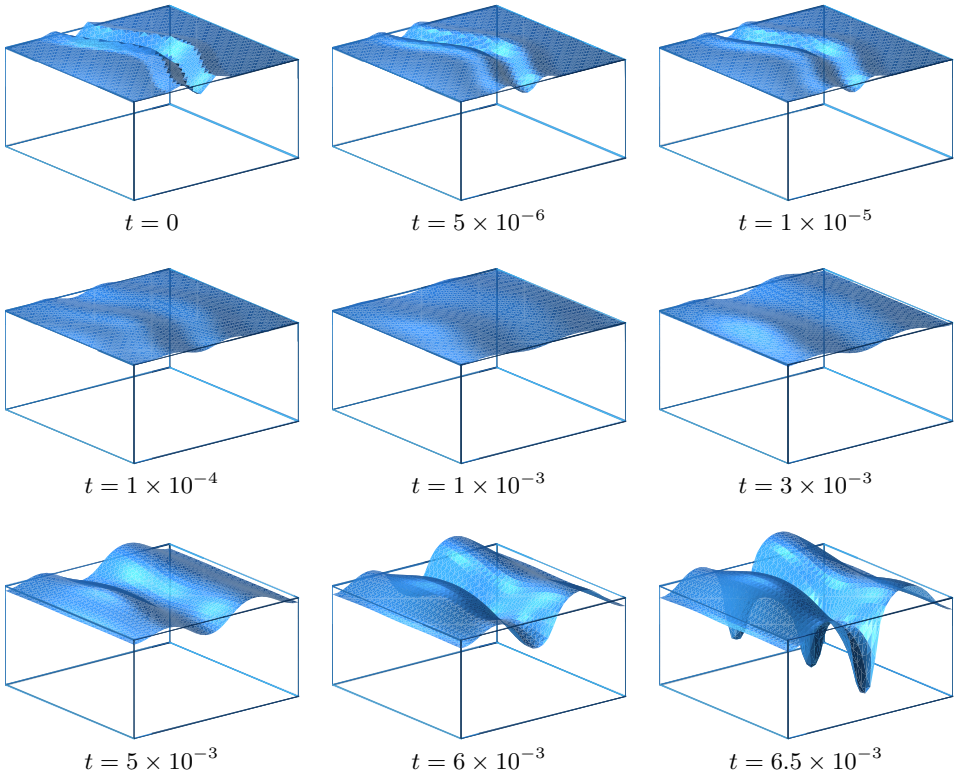


FIG. 7.11. *Solutions for* $f = -50/u$ *and* $u_0(x) = 1 + \delta(x)$ *at various time instants, with* $\delta(x)$ *a small perturbation across* $y = \cos x$*.*

7.6. Crack formation in two dimensions: Case $f = -C/u$. We conclude this section with the evolution of two-dimensional surfaces immersed in $\mathbb{R}^3$. We consider again the initial surface to be $u_0 = 1 + \delta$, where δ is a perturbation similar to that of Figure 7.3. First, we choose such δ across the periodic curve $y = \cos x$ (see Figure 7.11) and, finally, across the circle $x^2 + y^2 = 1/4$ centered at the ori-

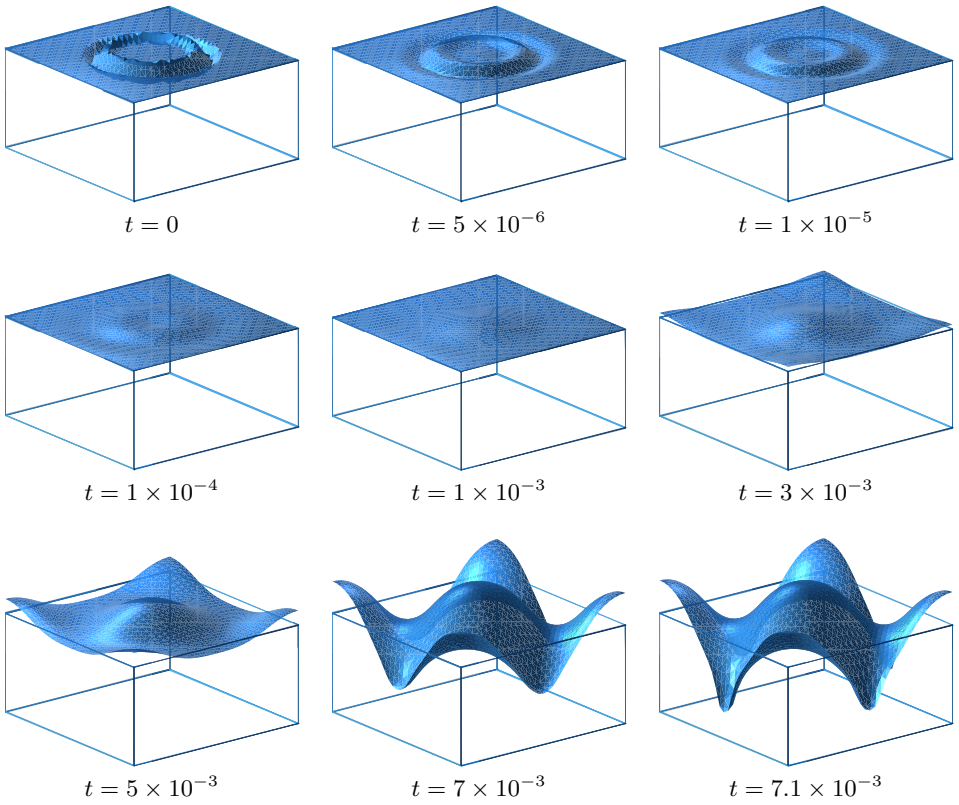


FIG. 7.12. *Solutions for* $f = -50/u$ *and* $u_0(x) = 1 + \delta(x)$ *at various time instants, with* $\delta(x)$ *a small perturbation across* $x^2 + y^2 = 1/4$.

gin (see Figure 7.12). We compute with linear elements and parameters $h = 1/16$, $\tau = 10^{-6}$.

We observe first a smoothing effect followed by crack formation. The latter seems to occur at isolated points rather than at lines, as illustrated in Figures 7.11 and 7.12. This happens even for one-dimensional profiles in two dimensions: point singularities seem to be preferred by this evolution.

Acknowledgments. We would like to thank J. Colin and M. Grinfeld, who originated our interest in surface diffusion, and F. Haußer for discovering (and fixing) a bug in our code.

REFERENCES

[1] R. J. ASARO AND W. A. TILLER, *Surface morphology development during stress corrosion cracking: Part* I: *Via surface diffusion*, Metall. Trans., 3 (1972), pp. 1789–1796.

[2] E. BÄNSCH, P. MORIN, AND R. H. NOCHETTO, *Finite element methods for surface diffusion*, in Free Boundary Problems, Internat. Ser. Numer. Math. 147, P. Colli, C. Verdi, and A. Visintin, eds., Birkhäuser, Basel, 2003, pp. 53–63.

[3] E. BÄNSCH, P. MORIN, AND R. H. NOCHETTO, *A Finite Element Method for Surface Diffusion: The Parametric Case*, preprint.

[4] E. BONNETIER, R. FALK, AND M. GRINFELD, *Analysis of a one-dimensional variational model of the equilibrium shape of a deformable crystal*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 573–591.

[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 2002.
[6] J. Cahn and J. Taylor, *Surface motion by surface diffusion*, Acta Metall. Mater., 42 (1994), pp. 1045–1063.
[7] B. Coleman, R. Falk, and M. Moakher, *Space–time finite element methods for surface diffusion with applications to the theory of the stability of cylinders*, SIAM J. Sci. Comput., 17 (1996), pp. 1434–1448.
[8] K. Deckelnick and G. Dziuk, *Error estimates for a semi-implicit fully discrete finite element scheme for the mean curvature flow of graphs,* Interfaces Free Bound., 2 (2000), pp. 341–359.
[9] K. Deckelnick, G. Dziuk, and C. M. Elliott, *Error analysis of a semidiscrete numerical scheme for diffusion in axially symmetric surfaces*, SIAM J. Numer. Anal., 41 (2003), pp. 2161–2179.
[10] G. Dziuk, *Numerical schemes for the mean curvature flow of graphs*, in Variations of Domain and Free-Boundary Problems in Solid Mechanics, Solid Mech. Appl. 66, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 63–70.
[11] C. M. Elliott and S. Maier-Paape, *Losing a graph with surface diffusion*, Hokkaido Math. J., 30 (2001), pp. 297–305.
[12] J. Escher, U. F. Mayer, and G. Simonett, *The surface diffusion flow for immersed hypersurfaces*, SIAM J. Math. Anal., 29 (1998), pp. 1419–1433.
[13] M. A. Grinfeld, *Thermodynamic Methods in the Theory of Heterogeneous Systems*, Longman, New York, 1991.
[14] C. Herring, *Surface tension as a motivation for sintering*, in The Physics of Powder Metallurgy, W. E. Kingston, ed., McGraw-Hill, New York, 1951, pp. 143–179.
[15] W. W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.
[16] A. Schmidt and K. G. Siebert, *ALBERT: An Adaptive Hierarchical Finite Element Toolbox*, Documentation, Preprint 06/2000, Universität Freiburg, Freiburg, Germany, p. 244.
[17] A. Schmidt and K. G. Siebert, *ALBERT — Software for Scientific Computations and Applications*, Acta Math. Univ. Comenian. (N.S.), 70 (2001), pp. 105–122.
[18] B. J. Spencer, S. H. Davis, and P. W. Voorhees, *Morphological instability in epitaxially-strained dislocation-free solid films: Nonlinear evolution*, Phys. Rev. B, 47 (1993), pp. 9760–9777.
[19] B. J. Spencer and D. I. Meiron, *Nonlinear evolution of the stress-driven morphological instability in a two-dimensional semi-infinite solid*, Acta Metall. Mater., 42 (1994), pp. 3629–3641.
[20] D. J. Srolovitz, *On the stability of surfaces of stressed solids*, Acta Metall., 37, (1989), pp. 621–625.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 800–825

GALERKIN FINITE ELEMENT APPROXIMATIONS OF STOCHASTIC ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS*

IVO BABUŠKA†, RAÚL TEMPONE†, AND GEORGIOS E. ZOURARIS‡

Abstract. We describe and analyze two numerical methods for a linear elliptic problem with stochastic coefficients and homogeneous Dirichlet boundary conditions. Here the aim of the computations is to approximate statistical moments of the solution, and, in particular, we give a priori error estimates for the computation of the expected value of the solution. The first method generates independent identically distributed approximations of the solution by sampling the coefficients of the equation and using a standard Galerkin finite element variational formulation. The Monte Carlo method then uses these approximations to compute corresponding sample averages. The second method is based on a finite dimensional approximation of the stochastic coefficients, turning the original stochastic problem into a deterministic parametric elliptic problem. A Galerkin finite element method, of either the h- or p-version, then approximates the corresponding deterministic solution, yielding approximations of the desired statistics. We present a priori error estimates and include a comparison of the computational work required by each numerical approximation to achieve a given accuracy. This comparison suggests intuitive conditions for an optimal selection of the numerical approximation.

Key words. stochastic elliptic equation, perturbation estimates, Karhunen–Loève expansion, finite elements, Monte Carlo method, $k \times h$-version, $p \times h$-version, expected value, error estimates

AMS subject classifications. 65N30, 65N15, 65C05, 65C20

DOI. 10.1137/S0036142902418680

1. Introduction.

Due to the great development in computational resources and scientific computing techniques, more mathematical models can be solved efficiently. Ideally, this artillery could be used to solve many classical partial differential equations, the mathematical models we shall focus on here, to high accuracy. However, in many cases, the information available to solve a given problem is far from complete and is in general very limited. This is the case when solving a partial differential equation whose coefficients depend on material properties that are known to some accuracy. The same may occur with its boundary conditions and even with the geometry of its domain; see, for example, the works [5, 4]. Naturally, since the current engineering trends are going toward more reliance on computational predictions, the need for assessing the level of accuracy in the results grows accordingly. More than ever, the goal then becomes to represent and propagate uncertainties from the available data to the desired result through our partial differential equation. By uncertainty we mean either intrinsic variability of physical quantities or simply lack of knowledge about some physical behavior; cf. [38]. If variability is interpreted as randomness,

*Received by the editors November 26, 2002; accepted for publication (in revised form) September 26, 2003; published electronically June 4, 2004.
http://www.siam.org/journals/sinum/42-2/41868.html

†Institute for Computational and Engineering Sciences (ICES), University of Texas at Austin, Austin, TX 78759 (babuska@ices.utexas.edu, rtempone@ices.utexas.edu). The first author was partially supported by the National Science Foundation (grant DMS-9802367) and the Office of Naval Research (grant N00014-99-1-0724). The second author was partially supported by the Swedish Council for Engineering Science grant 222-148, National Science Foundation grant DMS-9802367, UdelaR and UdeM in Uruguay, and the European network HYKE (contract HPRN-CT-2002-00282).

‡Department of Mathematics, University of the Aegean, GR–832 00 Karlovassi, Samos, Greece, and Institute of Applied and Computational Mathematics, FO.R.T.H., GR–711 10 Heraklion, Crete, Greece (zouraris@aegean.gr). This author was partially supported by the European network HYKE (contract HPRN-CT-2002-00282).

then naturally we can apply probability theory. To be fruitful, probability theory requires considerable empirical information about the random quantities in question, usually in the form of probability distributions or their statistical moments. On the other hand, if the only available information comes in the form of some bounds for the uncertain variables, the description and analysis of uncertainty may be based on other methods, such as convexity methods; cf. [8, 18]. This approach is closely related to the so-called worst case scenario.

This work addresses elliptic partial differential equations with stochastic coefficients, with applications to physical phenomena, e.g., random vibrations, seismic activity, oil reservoir management, and composite materials; see [2, 17, 19, 22, 27, 28, 30, 39, 43] and the references therein. Solving a stochastic partial differential equation entails finding the joint probability distribution of the solution, which is a hard problem. In practice we shall usually be satisfied with much less, namely, the computation of some moments, e.g., the expected value of the solution, or some probability related to a given event, e.g., the probability of some eventual failure; cf. [26, 34]. Although the results presented in this paper can be generalized to linear elliptic stochastic partial differential equations we now focus our study on the standard model problem, a second order linear elliptic equation with homogeneous Dirichlet boundary conditions.

Let D be a convex bounded polygonal domain in $\mathbb{R}^d$ and $(\Omega, \mathcal{F}, P)$ be a complete probability space. Here Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events, and $P : \mathcal{F} \to [0,1]$ is a probability measure. Consider the following stochastic linear elliptic boundary value problem: find a stochastic function, $u : \Omega \times \overline{D} \to \mathbb{R}$, such that P-a.e. in Ω, or, in other words, almost surely (a.s.), the following equation holds:

$$\begin{aligned} -\nabla \cdot (a(\omega,\cdot)\nabla u(\omega,\cdot)) &= f(\omega,\cdot) \quad \text{on } D, \\ u(\omega,\cdot) &= 0 \quad \text{on } \partial D. \end{aligned} \tag{1.1}$$

Here $a, f : \Omega \times D \to \mathbb{R}$ are stochastic functions with continuous and bounded covariance functions. If we denote by $B(D)$ the Borel σ-algebra generated by the open subsets of D, then a, f are assumed measurable with the σ-algebra $(\mathcal{F} \otimes B(D))$. In what follows we shall assume that a is bounded and uniformly coercive, i.e.,

$$\exists\, a_{\min}, a_{\max} \in (0, +\infty) : \quad P\big(\omega \in \Omega :\ a(\omega,x) \in [a_{\min}, a_{\max}] \quad \forall x \in \overline{D}\big) = 1. \tag{1.2}$$

To ensure regularity of the solution u we assume also that a has a uniformly bounded and continuous first derivative; i.e., there exists a real deterministic constant C such that

$$P\big(\omega \in \Omega :\ a(\omega,\cdot) \in C^1(\overline{D}) \ \text{ and } \ \max_{\overline{D}} |\nabla_x a(\omega,\cdot)| < C\big) = 1 \tag{1.3}$$

and that the right-hand side in (1.1) satisfies

$$\int_\Omega \int_D f^2(\omega,x)dx\, dP(\omega) < +\infty, \ \text{ which implies } \ \int_D f^2(\omega,x)dx < +\infty \text{ a.s.} \tag{1.4}$$

Depending on the structure of the noise that drives an elliptic partial stochastic differential equation, there are different numerical approximations. For example, when the size of the noise is relatively small, a Neumann expansion around the mean value of the elliptic operator in (1.1) is a popular approach. It requires only the solution of standard deterministic partial differential equations, the number of them being equal to the number of terms in the expansion. Equivalently, a Taylor expansion of the solution around its mean value with respect to the noise yields the same result. Similarly, the work [30] uses formal Taylor expansions up to second order of the solution but does not study their convergence properties. Recently, the work [3] proposed a perturbation method with successive approximations. It also proves that uniform coercivity of the diffusion is sufficient for the convergence of the perturbation method.

When only the load f is stochastic, it is also possible to derive deterministic equations for the statistical moments of the solution. This case was analyzed in [1, 32] and more recently in the work [40], where a new method to solve these equations with optimal complexity is presented.

On the other hand, the works by Deb [14], Deb, Babuška, and Oden [15], Ghanem and Red-Horse [21], and Ghanem and Spanos [22] address the general case where all the coefficients are stochastic. Both approaches transform the original stochastic problem into a deterministic one with a large dimensional parameter, and they differ in the choice of the approximating functional spaces. The works [14, 15] use finite elements to approximate the noise dependence of the solution, while [21, 22] use a formal expansion in terms of Hermite polynomials. The approximation error in the approach [14, 15] can then be bounded in terms of deterministic quantities, as described in this work. After finishing this paper the authors became aware of the work [9], which developed a related error analysis for elliptic stochastic differential equations. The work [9] gives approximation error estimates for functionals of the solution, while our work focuses on error estimates for the strong approximation of the statistical moments of the solution. Besides, we use the Karhunen–Loève expansion and characterize the regularity of the solution, yielding, e.g., exponential rates of convergence; cf. section 6. On the other hand, the analysis in [9] uses the regularity of the computed functional together with estimates in negative spaces for the approximation error in the solution of the stochastic partial differential equation. This negative estimate can in principle accommodate rough solutions; however, they require H^2 spatial regularity, an assumption that is not clearly fulfilled by rough solutions.

Monte Carlo methods are both general and simple to code, and they are naturally suited for parallelization. They generate a set of independent identically distributed (iid) approximations of the solution by sampling the coefficients of the equation, using a spatial discretization of the partial differential equation, e.g., by a Galerkin finite element method. Then using these approximations we can compute corresponding sample averages of the desired statistics. Monte Carlo methods have a rate of convergence that may be considered slow, but its computational work grows only like a polynomial with respect to the number of random variables present in the problem. It is worth mentioning that in particular cases their convergence can be accelerated by variance reduction techniques; see [29]. The convergence rate of the Monte Carlo method is interpreted in the probability sense, and a practical estimate of its error needs an a posteriori estimate of the variance of the sampled random variable, which in turn requires an a priori bound on higher statistical moments; cf. the Berry-Essen theorem in [16]. Besides this, if the probability density of a random variable is smooth,

the convergence rate of the Monte Carlo method for the approximation of its expected value can be improved; cf. [35, 45]. Quasi-Monte Carlo methods (see [12, 41, 42]) offer a way to get a better convergence rate than the Monte Carlo method, although this advantage seems to deteriorate in general when the number of random variables present in the problem becomes large.

Another way to provide a notion of stochastic partial differential equations is based on the Wick product and the Wiener chaos expansion; see [27] and [46]. This approach yields solutions in Kondratiev spaces of stochastic distributions that are based on a different interpretation of (1.1); the solutions proposed in [27] and [46] are not the same as those arising from (2.1). The choice between (2.1) and [27] is a modeling decision, based on the physical situation under study. For example, with the Wick product we have $E[a \diamond u] = E[a]E[u]$, regardless of the correlation between a and u, whereas this is in general not true with the usual product. A numerical approximation for Wick stochastic linear elliptic partial differential equations is studied in [44], yielding a priori convergence rates.

This work studies the case of stochastic linear elliptic partial differential equations with random diffusion and load coefficients, stating and proving conditions for existence and uniqueness of solutions; for example, to obtain a meaningful numerical solution for (1.1) its diffusion coefficient should be uniformly coercive. Besides, it compares a Monte Carlo Galerkin method with the stochastic Galerkin finite element method introduced in [14] and introduces a related p-version, developing a priori error estimates for each case. A priori error estimates are useful to characterize the convergence, and ultimately they provide information to compare the number of operations required by numerical methods. The conclusion for now is that if the noise is described by a small number of random parameters or if the accuracy requirement is sufficiently strict, then a stochastic Galerkin method is preferred; otherwise, a Monte Carlo Galerkin method still seems to be the best choice; see section 8. It is worth mentioning that the development of numerical methods for stochastic differential equations is still very much ongoing, and better numerical methods are expected to appear.

2. Theoretical aspects of the continuous problem.

2.1. Notation and function spaces. Let d be a positive integer and D be an open, connected, bounded, and convex subset of $\mathbb{R}^d$ with polygonal boundary ∂D. Denote the volume of D by $|D| \equiv \int_D 1dx$. For a nonnegative integer s and $1 \leq p \leq +\infty$, let $W^{s,q}(D)$ be the Sobolev space of functions having generalized derivatives up to order s in the space $L^q(D)$. Using the standard multi-index notation, $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a d-tuple of nonnegative integers, and the length of α is given by $|\alpha| = \sum_{i=1}^d \alpha_i$. The standard Sobolev norm of $v \in W^{s,q}(D)$ will be denoted by $\|v\|_{W^{s,q}(D)}$, $1 \leq q \leq +\infty$. Whenever $q = 2$, we shall use the notation $H^s(D)$ instead of $W^{s,2}(D)$. As usual, the function space $H_0^1(D)$ is the subspace of $H^1(D)$ consisting of functions which vanish at the boundary of D in the sense of trace, equipped with the norm $\|v\|_{H_0^1(D)} = \{\int_D |\nabla v|^2 \, dx\}^{1/2}$. Whenever $s = 0$ we shall keep the notation with $L^q(D)$ instead of $W^{0,q}(D)$. For the sake of generality, sometimes we shall let H be a Hilbert space with inner product $(\cdot, \cdot)_H$. In that case we shall also denote the dual space of H, H', that contains linear bounded functionals, $\mathcal{L} : H \to \mathbb{R}$, and is endowed with the operator norm $\|\mathcal{L}\|_{H'} = \sup_{0 \neq v \in H} \frac{\mathcal{L}(v)}{\|v\|_H}$.

Since stochastic functions intrinsically have different structure with respect to ω and with respect to x, the analysis of numerical approximations requires tensor

spaces. Let H_1, H_2 be Hilbert spaces. The tensor space $H_1 \otimes H_2$ is the completion of formal sums $u(y,x) = \sum_{i=1,\dots,n} v_i(y)w_i(x)$, $\{v_i\} \subset H_1$, $\{w_i\} \subset H_2$, with respect to the inner product $(u,\hat{u})_{H_1\otimes H_2} = \sum_{i,j}(v_i,\hat{v}_j)_{H_1}(w_i,\hat{w}_j)_{H_2}$. For example, let us consider two domains, $y \in \Gamma, x \in D$, and the tensor space $L^2(\Gamma)\otimes H^1(D)$, with tensor inner product

$$(u,\hat{u})_{L^2(\Gamma)\otimes H^1(D)} = \int_\Gamma \left(\int_D u(y,x)\hat{u}(y,x)dx \right) dy + \int_\Gamma \left(\int_D \nabla_x u(y,x) \cdot \nabla_x \hat{u}(y,x)dx \right) dy.$$

Thus, if $u \in L^2(\Gamma) \otimes H^k(D)$, then $u(y,\cdot) \in H^k(D)$ a.e. on Γ and $u(\cdot,x) \in L^2(\Gamma)$ a.e. on D. Moreover, we have the isomorphism $L^2(\Gamma) \otimes H^k(D) \simeq L^2(\Gamma; H^k(D)) \simeq H^k(D; L^2(\Gamma))$ with the definitions

$$L^2(\Gamma; H^k(D)) = \left\{ v : \Gamma \times D \to \mathbb{R} \mid v \text{ is strongly measurable and } \int_\Gamma \|v(y,\cdot)\|^2_{H^k(D)} < +\infty \right\},$$

$$H^k(D; L^2(\Gamma)) = \Bigg\{ v : \Gamma \times D \to \mathbb{R} \mid v \text{ is strongly measurable and } \forall |\alpha| \le k \ \exists\, \partial^\alpha v \in L^2(\Gamma)\otimes L^2(D) \text{ with } \int_\Gamma \int_D \partial_\alpha v(y,x)\varphi(y,x)dxdy = (-1)^{|\alpha|} \int_\Gamma \int_D v(y,x)\partial_\alpha \varphi(y,x)dxdy \quad \forall \varphi \in C_0^\infty(\Gamma \times D) \Bigg\}.$$

Similar constructions can be done for the tensor product of Banach spaces, although the norm for the tensor space used to obtain the completion of the formal sums has to be defined explicitly on each case. Here the Banach space $C(\Gamma; H)$ comprises all continuous functions $u : \Gamma \to H$ with the norm $\|u\|_{C(\Gamma;H)} \equiv \sup_{y\in\Gamma} \|u(y)\|_H$. Similar definitions apply to the spaces $C^k(\Gamma; H)$, $k = 1, \dots$; cf. [20, p. 285].

Let Y be an $\mathbb{R}^N$-valued random variable in $(\Omega, \mathcal{F}, P)$. If $Y \in L^1_P(\Omega)$, we denote its expected value by $E[Y] = \int_\Omega Y(\omega)dP(\omega) = \int_{\mathbb{R}^N} y \, d\mu_Y(y)$, where μ_Y is the distribution measure for Y, defined for the Borel sets $\tilde{b} \in B(\mathbb{R}^N)$ by $\mu_Y(\tilde{b}) \equiv P(Y^{-1}(\tilde{b}))$. If μ_Y is absolutely continuous with respect to the Lebesgue measure, then there exists a density function $\rho_Y : \mathbb{R} \to [0, +\infty)$ such that $E[Y] = \int_{\mathbb{R}^N} y \, \rho_Y(y)dy$. Analogously, whenever $Y_i \in L^2_P(\Omega)$ for $i = 1, \dots, d$, the covariance matrix of Y, $Cov[Y] \in \mathbb{R}^{d\times d}$, is defined by $Cov[Y](i,j) = Cov(Y_i, Y_j) = E[(Y_i - E[Y_i])(Y_j - E[Y_j])]$, $i, j = 1, \dots, d$. Besides this, whenever $u(\omega, x)$ is a stochastic process the positive semidefinite function $Cov[u](x_1, x_2) = Cov[u(x_1), u(x_2)] = Cov[u(x_2), u(x_1)]$ is the covariance function of the stochastic process u.

To introduce the notion of stochastic Sobolev spaces we first recall the definition of stochastic weak derivatives. Let $v \in L^2_P(\Omega) \otimes L^2(D)$; then the α stochastic weak derivative of v, $w = \partial_\alpha v \in L^2_P(\Omega) \otimes L^2(D)$, satisfies $\int_D v(\omega,x)\partial^\alpha \phi(x)dx = (-1)^{|\alpha|} \int_D w(\omega,x)\phi(x)dx \, \forall \phi \in C_0^\infty(D)$ a.s.

We shall work with stochastic Sobolev spaces $\widetilde{W}^{s,q}(D) = L^q_P(\Omega, W^{s,q}(D))$ containing stochastic functions, $v : \Omega \times D \to \mathbb{R}$, that are measurable with respect to the

product σ-algebra $\mathcal{F} \otimes B(D)$ and equipped with the averaged norms $\|v\|_{\widetilde{W}^{s,q}(D)} = (E[\|v\|^q_{W^{s,q}(D)}])^{1/q} = (E[\sum_{|\alpha|\le s} \int_D |\partial^\alpha v|^q dx])^{1/q}$, $1 \le q < +\infty$, and $\|v\|_{\widetilde{W}^{s,\infty}(D)} = \max_{|\alpha|\le s} \left(\text{ ess } \sup_{\Omega\times D} |\partial^\alpha v| \right)$. Observe that if $v \in \widetilde{W}^{s,q}(D)$, then $v(\omega,\cdot) \in W^{s,q}(D)$ a.s. and $\partial^\alpha v(\cdot,x) \in L^q_P(\Omega)$ a.e. on D for $|\alpha| \le s$. Whenever $q = 2$, the above space is a Hilbert space, i.e., $\widetilde{W}^{s,2}(D) = \widetilde{H}^s(D) \simeq L^2_P(\Omega) \otimes H^s(D)$.

2.2. Existence and uniqueness for the solution of a linear stochastic elliptic problem. Let us consider the tensor product Hilbert space $H = \widetilde{H}^1_0(D) \simeq L^2_P(\Omega; H^1_0(D))$ endowed with the inner product $(v,u)_H \equiv E[\int_D \nabla v \cdot \nabla u dx]$.

Define the bilinear form, $\mathcal{B} : H \times H \to \mathbb{R}$, by $\mathcal{B}(v,w) \equiv E[\int_D a \nabla v \cdot \nabla w dx] \, \forall v, w \in H$. The standard assumption (1.2) yields both the continuity and the coercivity of $\mathcal{B}$; i.e., $|\mathcal{B}(v,w)| \le a_{max} \|v\|_H \|w\|_H \, \forall v, w \in H$, and $a_{min}\|v\|^2_H \le \mathcal{B}(v,v) \, \forall v \in H$. A direct application of the Lax–Milgram lemma (cf. [11]) implies the existence and uniqueness for the solution to the following variational formulation: find $u \in H$ such that

$$\mathcal{B}(u,v) = \mathcal{L}(v) \quad \forall v \in H. \tag{2.1}$$

Here $\mathcal{L}(v) \equiv E[\int_D f v dx] \, \forall v \in H$ defines a bounded linear functional since the random field f satisfies (1.4). Since the domain D is convex and bounded and assumptions (1.2), (1.3) on the diffusion a hold, the theory of elliptic regularity (cf. [20]) implies that the solution of (1.1) satisfies $u(\omega,\cdot) \in H^2(D) \cap H^1_0(D)$ a.s. Moreover, standard arguments from measure theory show that the solution to (2.1) also solves (1.1). The formulation (2.1), together with assumption (2.1) on finite dimensional noise, gives the basis for the stochastic Galerkin finite element method (SGFEM) introduced in sections 5 and 6, while formulation (1.1) is the basis for the Monte Carlo Galerkin finite element method (MCGFEM), discussed in section 4.

2.3. Continuity with respect to the coefficients a and f. Since the coefficients a and f are not known exactly, in the next proposition we consider a perturbed weak formulation and estimate the size of the corresponding perturbation in the solution. The proof uses standard estimates and is included in [6].

PROPOSITION 2.1. *Let $(H, (\cdot,\cdot)_H)$ be a Hilbert space. Consider two symmetric bilinear forms $\mathcal{B}, \widehat{\mathcal{B}} : H \times H \to \mathbb{R}$ that are H-coercive and bounded; i.e., there exist real constants $0 < a_{min} \le a_{max}$ such that*

$$a_{min} \|v\|^2_H \le \min\{\mathcal{B}(v,v), \widehat{\mathcal{B}}(v,v)\} \quad \forall v \in H \tag{2.2}$$

and

$$\max\{|\mathcal{B}(v,w)|, |\widehat{\mathcal{B}}(v,w)|\} \le a_{max} \|v\|_H \|w\|_H \quad \forall v, w \in H. \tag{2.3}$$

Consider two bounded linear functionals, $\mathcal{L}, \hat{\mathcal{L}} \in H'$, and let u,$\hat{u} \in H$ be the solutions of the problems

$$\begin{aligned} \mathcal{B}(u,v) &= \mathcal{L}(v) \quad \forall v \in H, \\ \hat{\mathcal{B}}(\hat{u},v) &= \hat{\mathcal{L}}(v) \quad \forall v \in H. \end{aligned}$$

If, in addition, we know that there exist Banach spaces, V_1,V_2, and positive constants, C,γ', such that $u \in V_2 \subseteq H \subset V_1$, $\|\cdot\|_{V_1} \le C\|\cdot\|_H$, and

$$|(\hat{\mathcal{B}} - \mathcal{B})(w,v)| \le \gamma' \|w\|_{V_1} \|v\|_{V_2} \quad \forall w \in H, \ v \in V_2, \tag{2.4}$$

then

$$\|u-\hat{u}\|_H \leq \frac{1}{a_{min}}(\|\mathcal{L}-\hat{\mathcal{L}}\|_{H'} + C\gamma'\|u\|_{V_2}). \tag{2.5}$$

Next we consider a perturbation of (2.1). A direct application of Proposition 2.1 yields the following estimate.

COROLLARY 2.1. *Let* $1 < p < +\infty$ *with* $1/p+1/q = 1$. *Consider the Hilbert space* $H = \widetilde{H}_0^1(D)$ *and perturbed coefficients,* $\hat{a}, \hat{f}$, *satisfying* $0 < a_{min} \leq \hat{a} \leq a_{max} < \infty$, $(P\otimes dx)$ *a.e. on* $D\times\Omega, \hat{f} \in \widetilde{L}^2(D)$. *Let* u *and* $\hat{u}$ *solve* $E[\int_D \hat{a}\nabla\hat{u}\cdot\nabla v dx] = E[\int_D \hat{f}v dx]$ $\forall v \in H, E[\int_D a\nabla u\cdot\nabla v dx] = E[\int_D fv dx]\,\forall v \in H$. *Besides this, assume that the solution* u *belongs to the stochastic Sobolev space* $\widetilde{W}^{1,2q}(D)$. *Then*

$$\|u-\hat{u}\|_{\widetilde{H}_0^1(D)} \leq \tfrac{1}{a_{min}}(C_D\|\hat{f}-f\|_{\widetilde{L}^2(D)} + \|a-\hat{a}\|_{\widetilde{L}^{2p}(D)}\|u\|_{\widetilde{W}^{1,2q}(D)}),$$

with $C_D > 0$ *being the Poincaré constant for the domain* D; *i.e.,* $\|v\|_{L^2(D)} \leq C_D$ $\|v\|_{H_0^1(D)}\forall v \in H_0^1(D)$.

Proof. Take $V_1 = \widetilde{H}_0^1(D)$ and $V_2 = \widetilde{W}^{1,2q}(D)$. In order to apply (2.5) it is enough to bound the difference of bilinear forms

$$\begin{aligned}
&\left|E\left[\int_D (a-\hat{a})\nabla u\cdot\nabla v dx\right]\right| \\
&\leq \left(E\left[\int_D (a-\hat{a})^2|\nabla u|^2 dx\right]\right)^{1/2}\left(E\left[\int_D |\nabla v|^2 dx\right]\right)^{1/2} \\
&\leq \left(E\left[\int_D (a-\hat{a})^{2p} dx\right]\right)^{1/2p}\left(E\left[\int_D |\nabla u|^{2q} dx\right]\right)^{1/2q}\left(E\left[\int_D |\nabla v|^2 dx\right]\right)^{1/2}. \qquad \square
\end{aligned}$$

2.4. Karhunen–Loève expansions and finite dimensional noise. Here we recall the Karhunen–Loève expansion, a suitable tool for the approximation of stochastic processes. Consider a stochastic process a with continuous covariance function $Cov[a] : \overline{D}\times\overline{D} \to \mathbb{R}$. Besides this, let $\{(\lambda_n, b_n)\}_{n=1}^{\infty}$ denote the sequence of eigenpairs associated with the compact self-adjoint operator that maps

$$f \in L^2(D) \mapsto \int_D Cov[a](x,\cdot)f(x)dx \in L^2(D).$$

Its nonnegative eigenvalues, $\sqrt{\int_{D\times D}(Cov[a](x_1,x_2))^2 dx_1\, dx_2} \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, satisfy $\sum_{n=1}^{+\infty}\lambda_n = \int_D Var[a](x)dx$. The corresponding eigenfunctions are orthonormal, i.e., $\int_D b_i(x)b_j(x)dx = \delta_{ij}$. The truncated Karhunen–Loève expansion of the stochastic process a (cf. [33]) is

$$a_N(\omega,x) = E[a](x) + \sum_{n=1}^{N}\sqrt{\lambda_n}b_n(x)Y_n(\omega),$$

where the real random variables, $\{Y_n\}_{n=1}^{\infty}$, are mutually uncorrelated and have mean zero and unit variance. These random variables are uniquely determined by $Y_n(\omega) = \frac{1}{\sqrt{\lambda_n}}\int_D(a(\omega,x) - E[a](x))b_n(x)dx$ for $\lambda_n > 0$. Then, by Mercer's theorem (cf. [37, p. 245]), we have

$$\sup_{x\in D} E[(a-a_N)^2](x) = \sup_{x\in D}\sum_{n=N+1}^{+\infty}\lambda_n b_n^2(x) \to 0 \ \text{ as } \ N\to\infty.$$

If, in addition,

- the images $Y_n(\Omega)$, $n = 1, \ldots$, are uniformly bounded in $\mathbb{R}$,
- the eigenfunctions b_n are smooth, which is the case when the covariance function is smooth,
- and the eigenpairs have at least the decay $\sqrt{\lambda_n}\|b_n\|_{L^\infty(D)} = \mathcal{O}(\frac{1}{1+n^s})$ for some $s > 1$,

then $\|a - a_N\|_{\widetilde{L}^\infty(D)} \to 0$. Notice that for larger values of the decay exponent s we can also obtain the convergence of higher spatial derivatives of a_N in $\widetilde{L}^\infty(D)$. The last two conditions can be readily verified once the covariance function of a is known. However, observe that it is also necessary to verify the uniform coercivity of a_N, which depends on the probability distributions of Y_n, $n = 1, \ldots$.

In many problems the source of the randomness can be approximated using just a small number of mutually uncorrelated, sometimes mutually independent, random variables. Take, for example, the case of a truncated Karhunen–Loève expansion described previously.

Assumption 2.1 (finite dimensional noise). Whenever we apply some numerical method to solve (1.1) we assume that the coefficients used in the computations, $a, f : \Omega \times D \to \mathbb{R}$, are finite Karhunen–Loève expansions; i.e., $a(\omega, x) = E[a](x) + \sum_{n=1}^{N} \sqrt{\lambda_n} b_n(x) Y_n(\omega)$ and $f(\omega, x) = E[f](x) + \sum_{n=1}^{N} \sqrt{\hat{\lambda}_n} \hat{b}_n(x) Y_n(\omega)$, where $\{Y_n\}_{n=1}^{N}$ are real random variables with mean value zero and unit variance, are uncorrelated, and have images, $\Gamma_n \equiv Y_n(\Omega)$, that are bounded intervals in $\mathbb{R}$ for $n = 1, \ldots, N$. Moreover, we assume that each Y_n has a density function $\rho_n : \Gamma_n \to \mathbb{R}^+$ for $n = 1, \ldots, N$.

In what follows we use the notation $\rho(y)\ \forall y \in \Gamma$ for the joint probability density of $(Y_1, \ldots, Y_N)$ and $\Gamma \equiv \prod_{n=1}^{N} \Gamma_n \subset \mathbb{R}^N$ for the support of such probability density.

After making Assumption 2.1, we have by the Doob–Dynkin lemma (cf. [36]) that u, the solution corresponding to the stochastic partial differential equation (1.1), can be described by just a finite number of random variables, i.e., $u(\omega, x) = u(Y_1(\omega), \ldots, Y_N(\omega), x)$. The number N has to be large enough so that the approximation error is sufficiently small. Now the goal is to approximate the function $u(y, x)$. In addition, the stochastic variational formulation (2.1) has a deterministic equivalent in the following: find $u \in L^2_\rho(\Gamma) \otimes H^1_0(D)$ such that

$$\begin{aligned} &\int_\Gamma \rho(y) \int_D a(y,x) \nabla u(y,x) \cdot \nabla v(y,x) dx dy \\ &\quad = \int_\Gamma \rho(y) \int_D f(y,x) v(y,x) dx dy \quad \forall\ v \in L^2_\rho(\Gamma) \otimes H^1_0(D). \end{aligned} \tag{2.6}$$

In this work the gradient notation, ∇, always means differentiation with respect to $x \in D$ only, unless otherwise stated. The corresponding strong formulation for (2.6) is an elliptic partial differential equation with an N-dimensional parameter, i.e.,

$$\begin{aligned} -\nabla \cdot (a(y,x)\ \nabla u(y,x)) &= f(y,x) \quad \forall (y,x) \in \Gamma \times D, \\ u(y,x) &= 0 \qquad\quad \forall (y,x) \in \Gamma \times \partial D. \end{aligned} \tag{2.7}$$

Making Assumption 2.1 is a crucial step, turning the original stochastic elliptic equation (1.1) into a deterministic parametric elliptic one and allowing the use of finite element and finite difference techniques to approximate the solution of the resulting deterministic problem.

Truncation of the outcomes set, Γ. For the sake of efficiency, it may be useful to compute the solution of (2.7) in a subdomain with strictly positive probability, $\Gamma_0 \subset \Gamma$. Besides, we assume the probability density of Y to be strictly positive in Γ_0. In that case, we approximate the function

$$E[u(Y,\cdot)\ 1_{\{Y\in\Gamma_0\}}] = E[u(Y,\cdot)|Y\in\Gamma_0]\ \ P(Y\in\Gamma_0)$$

instead of the original $E[u]$. If $\bar{u}$ is an approximation of u in Γ_0, then we have the splitting

(2.8)
$$\begin{aligned}&\|E[u(Y,\cdot)] - E[\bar{u}(Y,\cdot)\ 1_{\{Y\in\Gamma_0\}}]\|\\ &\qquad \le \|E[u(Y,\cdot)] - E[u(Y,\cdot)\ 1_{\{Y\in\Gamma_0\}}]\| + \|E[u(Y,\cdot)-\bar{u}(Y,\cdot)|Y\in\Gamma_0]\|P(Y\in\Gamma_0).\end{aligned}$$

Property 2.1 below gives a simple estimate for the first error contribution, which is related to the truncation of Γ. The second error contribution in (2.8) is the discretization error, and it will be analyzed for each numerical approximation separately; see sections 4, 5, and 6. In those sections we shall simplify the notation by writing $\Gamma = \Gamma_0$ and work with the corresponding conditional probability space.

PROPERTY 2.1. *Let u be the solution of the problem* (2.7); *then there exists a constant C such that*

$$\big\|E[u(Y,\cdot)]-E[u(Y,\cdot)\ 1_{\{Y\in\Gamma\}}]\big\|_{H_0^1(D)} \le C\sqrt{P(Y\notin\Gamma_0)}\ \|f\|_{L^2_\rho(\Gamma\setminus\Gamma_0)\otimes L^2(D)}. \tag{2.9}$$

3. The finite element spaces. In this section, we define tensor product finite element spaces on the set $\Gamma\times D$, which we will use to construct approximations of the solution of the parametric boundary value problem (2.7), stating their approximation properties.

3.1. Finite element spaces on the spatial set $D\subset\mathbb{R}^d$: h-version. Consider a family of finite element approximation spaces, $X_h^d \subset H_0^1(D)$, consisting of piecewise linear continuous functions on conforming triangulations (of simplices), $\mathcal{T}_h^d$, of the convex polyhedral domain, $D\subset\mathbb{R}^d$, with a maximum mesh spacing parameter $h>0$. We will always assume that the triangulations are nondegenerate (sometimes also called regular); cf., [11, p. 106]. Then (cf. [11, 13]) the finite element spaces X_h^d satisfy a standard approximation estimate, namely, that for all $v\in H^2(D)\cap H_0^1(D)$

$$\min_{\chi\in X_h^d}\|v-\chi\|_{H_0^1(D)} \le C\,h\,\|v\|_{H^2(D)}, \tag{3.1}$$

where $C>0$ is a constant independent of v and h.

3.2. Tensor product finite element spaces on the outcomes set $\Gamma\subset\mathbb{R}^N$: k-version. Let $\Gamma=\prod_{n=1}^N\Gamma_n$ be as in subsection 2.4. Consider a partition of Γ consisting of a finite number of disjoint $\mathbb{R}^N$-boxes, $\gamma=\prod_{n=1}^N(a_n^\gamma,b_n^\gamma)$, with $(a_n^\gamma,b_n^\gamma)\subset\Gamma_n$ for $n=1,\dots,N$. The mesh spacing parameters, $k_n>0$, are defined by $k_n\equiv\max_\gamma|b_n^\gamma-a_n^\gamma|$ for $1\le n\le N$. For every nonnegative integer multi-index $q=(q_1,\dots,q_N)$ consider the finite element approximation space of (discontinuous) piecewise polynomials with degree at most q_n on each direction y_n, $Y_k^q\subset L^2(\Gamma)$. Thus, if $\varphi\in Y_k^q$, its restriction to each of the partition boxes satisfies $\varphi|_\gamma\in\operatorname{span}\big(\prod_{n=1}^N y_n^{\alpha_n}:\ \alpha_n\in\mathbb{N}\text{ and }\alpha_n\le q_n,\quad n=1,\dots,N\big)$. It is easy to verify

that the finite element spaces Y_k^q have the following approximation property: for all $v \in H^{q+1}(\Gamma)$,

$$\min_{\varphi \in Y_k^q} \|v - \varphi\|_{L^2(\Gamma)} \le \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{q_n+1} \frac{\|\partial_{y_n}^{q_n+1} v\|_{L^2(\Gamma)}}{(q_n+1)!}. \tag{3.2}$$

3.3. Tensor product finite element spaces on $\Gamma \times D$: $k \times h$-version. Here we will discuss some approximation properties of the following tensor product finite element spaces:

(3.3)

$$Y_k^q \otimes X_h^d \equiv \left\{\psi = \psi(y,x) \in L^2(\Gamma \times D) : \psi \in \operatorname{span}\big(\varphi(y)\,\chi(x) : \varphi \in Y_k^q,\, \chi \in X_h^d\big)\right\}$$

with X_h^d and Y_k^q as in subsections 3.1 and 3.2.

For later use we recall the definition of the standard L^2-projection operators $\Pi_k^q : L^2(\Gamma) \to Y_k^q$ by

$$(\Pi_k^q w - w, \varphi)_{L^2(\Gamma)} = 0 \quad \forall \varphi \in Y_k^q, \quad \forall w \in L^2(\Gamma) \tag{3.4}$$

and the H_0^1-projection operator $\mathcal{R}_h : H_0^1(D) \to X_h^d$ by

$$(\nabla(\mathcal{R}_h v - v), \nabla\chi)_{L^2(D)} = 0 \quad \forall \chi \in X_h^d, \quad \forall v \in H_0^1(D). \tag{3.5}$$

Estimates (3.1) and (3.2) imply

$$\begin{aligned} \|v - \mathcal{R}_h v\|_{H_0^1(D)} &\le C\,h\,\|v\|_{H^2(D)}, \\ \|w - \Pi_k^q w\|_{L^2(\Gamma)} &\le \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{q_n+1} \frac{\|\partial_{y_n}^{q_n+1} w\|_{L^2(\Gamma)}}{(q_n+1)!} \end{aligned} \tag{3.6}$$

for all $v \in H^2(D) \cap H_0^1(D)$ and $w \in H^{q+1}(\Gamma)$. We now state an approximation property for the tensor product finite element spaces defined in (3.3) which is a direct implication of the approximation properties of the spaces Y_k^q and X_h^d.

PROPOSITION 3.1. *There exists a constant $C > 0$ independent of h, N, q, and k such that*

$$\begin{aligned} &\inf_{\psi \in Y_k^q \otimes X_h^d} \|v - \psi\|_{L^2(\Gamma; H_0^1(D))} \\ &\qquad \le C \left\{ h\,\|v\|_{L^2(\Gamma; H^2(D))} + \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{q_n+1} \frac{\|\partial_{y_n}^{q_n+1} v\|_{L^2(\Gamma; H_0^1(D))}}{(q_n+1)!} \right\} \end{aligned} \tag{3.7}$$

for all $v \in C^{q+1}(\Gamma; H^2(D) \cap H_0^1(D))$.

Proof. Since $\Pi_k^q(\mathcal{R}_h v) \in Y_k^q \otimes X_h^d$, using (3.6) we obtain

$$\begin{aligned} \inf_{\psi \in Y_k^q \otimes X_h^d} \|v - \psi\|_{L^2(\Gamma; H_0^1(D))} &\le \|v - \Pi_k^q(\mathcal{R}_h v)\|_{L^2(\Gamma; H_0^1(D))} \\ &\le \|v - \mathcal{R}_h v\|_{L^2(\Gamma; H_0^1(D))} \\ &\quad + \|\mathcal{R}_h v - \Pi_k^q(\mathcal{R}_h v)\|_{L^2(\Gamma; H_0^1(D))} \\ &\le C\,h\,\|v\|_{L^2(\Gamma; H^2(D))} \\ &\quad + \|\mathcal{R}_h v - \Pi_k^q(\mathcal{R}_h v)\|_{L^2(\Gamma; H_0^1(D))}. \end{aligned} \tag{3.8}$$

Applying the estimate (3.2) and using the boundedness of $\mathcal{R}_h$ in $H_0^1(D)$ yield

$$\begin{aligned}\|\mathcal{R}_h v - \Pi_k^q(\mathcal{R}_h v)\|_{L^2(\Gamma;H_0^1(D))} &\leq \|v - \Pi_k^q v\|_{L^2(\Gamma;H_0^1(D))} \\ &\leq \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{q_n+1} \frac{\|\partial_{y_n}^{q_n+1} v\|_{L^2(\Gamma;H_0^1(D))}}{(q_n+1)!}.\end{aligned}$$

The estimate (3.7) follows, combining (3.8) with the last estimate. ∎

3.4. Tensor product finite element spaces on $\Gamma \times D$: $p \times h$-version. This approximation space is in fact a particular case of the $k \times h$-version with no k partition of Γ, i.e., $k_n = |\Gamma_n|$, $n = 1, \ldots, N$. Instead, only the polynomial degree is increased. Here the multi-index $p = (p_1, \ldots, p_N)$ plays the role of the q from section 3.3, and we use the tensor finite element space $Z^p = \bigotimes_{n=1}^N Z_n^{p_n}$, where the one dimensional global polynomial subspaces, $Z_n^{p_n}$, are defined by $Z_n^{p_n} = \{v : \Gamma_n \to \mathbb{R} : v \in \text{span}(y^s, s = 0, \ldots, p_n)\}, n = 1, \ldots, N$.

4. Monte Carlo Galerkin finite element method. In this section we describe the use of the standard Monte Carlo Galerkin finite element method (MCGFEM) to construct approximations of the expected value of the solution.

Formulation of the MCGFEM.

• Give a number of realizations, M, a piecewise linear finite element space on D, X_h^d, as defined in subsection 3.1.

• For each $j = 1, \ldots, M$, sample iid realizations of the diffusion $a(\omega_j, \cdot)$ and the load $f(\omega_j, \cdot)$, based on realizations of $\{Y_n\}_{n=1}^N$, and find a corresponding approximation $u_h(\omega_j, \cdot) \in X_h^d$ such that

$$(a(\omega_j, \cdot)\nabla u_h(\omega_j, \cdot), \nabla\chi)_{L^2(D)} = (f(\omega_j, \cdot), \chi)_{L^2(D)} \quad \forall \chi \in X_h^d. \tag{4.1}$$

• Finally, use the sample average $\frac{1}{M}\sum_{j=1}^M u_h(\omega_j; \cdot)$ to approximate $E[u]$. ∎

Here we consider only the case where X_h^d is the same for all realizations; i.e., the spatial triangulation is deterministic. The computational error naturally separates into two parts:

$$E[u] - \frac{1}{M}\sum_{j=1}^{M} u_h(\omega_j, \cdot) = \big(E[u] - E[u_h]\big) + \left(E[u_h] - \frac{1}{M}\sum_{j=1}^{M} u_h(\omega_j, \cdot)\right) \equiv \mathcal{E}_h + \mathcal{E}_S. \tag{4.2}$$

The size of the spatial triangulation controls the space discretization error $\mathcal{E}_h$, while the number of realizations, M of u_h, controls the statistical error $\mathcal{E}_S$.

To study the behavior of the statistical error, let us first consider the random variable $\|\mathcal{E}_S\|_{H_0^1(D)}$ which, due to the independence of the realizations $u_h(\omega_j, \cdot)$, $j = 1, \ldots, M$, satisfies the estimate

$$M\ E\big[\|\mathcal{E}_S\|_{H_0^1(D)}^2\big] \leq \|u_h\|_{\widetilde{H}_0^1(D)}^2 \leq \left(\frac{C_D}{a_{min}}\right)^2 \|f\|_{\widetilde{L}^2(D)}^2, \tag{4.3}$$

and a similar result also holds in $L^2(D)$. Then, thanks to (4.3) we have that, for either $H = L^2(D)$ or $H = H_0^1(D)$, the statistical error $\|\mathcal{E}_S\|_H$ tends a.s. to zero as we increase the number of realizations; i.e., we have the following.

PROPOSITION 4.1. *Suppose that there exists a constant $C > 0$ independent of M and h such that the statistical error in H norm satisfies*

$$M\ E[\|\mathcal{E}_S\|_H^2] \le C \quad \forall M, h. \tag{4.4}$$

Then, taking the number of realizations, M_k, increasingly from the set $\{2^k : k \in \mathbb{N}\}$, we have, for any $\alpha \in (0, 1/2)$ and any choice of mesh size h, $\lim_{M_k\to\infty} {M_k}^{\alpha}\|\mathcal{E}_S\|_H = 0$ a.s.

Proof. Let $\epsilon > 0$. Then (4.4) and Markov's inequality give

$$P((M_k)^{\alpha}\|\mathcal{E}_S\|_H > \epsilon) \le \frac{E[(M_k)^{2\alpha}\|\mathcal{E}_S\|_H^2]}{\epsilon^2} \le \frac{C}{\epsilon^2 (M_k)^{1-2\alpha}}.$$

Thus, for $\alpha \in (0, 1/2)$ we have

$$\sum_{k=1}^{\infty} P(M_k^{\alpha}\|\mathcal{E}_S\|_H > \epsilon) \le \frac{C}{\epsilon^2}\sum_{k=1}^{\infty}\frac{1}{M_k^{1-2\alpha}} \le \frac{C}{\epsilon^2}\sum_{k=1}^{\infty}\frac{1}{(2^{1-2\alpha})^k} < \infty,$$

which, together with the Borel–Cantelli lemma, finishes the proof. ∎

Under the same assumptions as in Proposition 4.1 we have that for any given $\epsilon > 0$ there exists a constant $C > 0$ independent of ϵ, M, and h such that

$$P\left(\|\mathcal{E}_S\|_H > \frac{\epsilon}{\sqrt{M}}\right) \le \frac{C}{\epsilon^2}. \tag{4.5}$$

Thus, within a given confidence level we have the usual convergence rate for the Monte Carlo method, which is independent of the mesh size h. Next we present error estimates for the space discretization error, namely, we have the following.

PROPOSITION 4.2 (spatial discretization error estimates). *There holds*

$$h\|u - u_h\|_{H_0^1(D)} + \|u - u_h\|_{L^2(D)} \le C\, h^2 \|f\|_{L^2(D)} \quad a.s.,$$
$$h\|E[u] - E[u_h]\|_{H_0^1(D)} + \|E[u] - E[u_h]\|_{L^2(D)} \le C h^2 E[\|f\|_{L^2(D)}^2]^{1/2}.$$

The results from Proposition 4.2 and estimate (4.5) will be used in section 8 to compare the MCGFEM with other discretizations for (1.1).

5. Stochastic Galerkin finite element method: $k \times h$-version. This section defines and analyzes the $k \times h$-version of the stochastic Galerkin finite element method ($k \times h$-SGFEM) which, via a Galerkin variational formulation, yields approximations, $u_{kh} \in Y_k^q \otimes X_h^d$, of the solution u of the parametric elliptic boundary value problem (2.7).

Formulation of the $k \times h$-SGFEM. Denote by $q = (q_1, \ldots, q_N) \in \mathbb{N}^N$ a multi-index, and let Γ be a bounded box in $\mathbb{R}^N$. The $k \times h$-SGFEM approximation is the tensor product, $u_{kh} \in Y_k^q \otimes X_h^d$, such that

$$(u_{kh}, \psi)_E \equiv \int_\Gamma \rho\, \left(a\, \nabla u_{kh}, \nabla\psi\right)_{L^2(D)}\, dy = \int_\Gamma \rho\, \left(f, \psi\right)_{L^2(D)}\, dy \quad \forall\, \psi \in Y_k^q \otimes X_h^d. \tag{5.1}$$

Recall that $\rho : \Gamma \to (0, +\infty)$ is the density function of the vector-valued random variable $Y : \Omega \to \Gamma \subset \mathbb{R}^N$. Hence, the assumption (1.2) on the random function $a(\omega, x) \equiv a(Y(\omega), x)$ reads

$$a(y, x) \in [a_{\min}, a_{\max}] \quad \forall (y, x) \in \overline{\Gamma \times D}. \tag{5.2}$$

Although the analysis can be generalized [6], we now focus on the practical case where both a and f are truncated Karhunen–Loève expansions. Later, section 7 discusses how to compute efficiently u_{kh}, the solution of (5.1), by a double orthogonal polynomials technique. By Lemma 4.1 in [31], the solution u of (2.7) satisfies $u \in C^\infty(\Gamma; H^2(D) \cap H_0^1(D))$. Use (5.2) and (2.7) to obtain

$$\|u(y,\cdot)\|_{H_0^1(D)} \leq \tfrac{C_D}{a_{\min}} \|f(y,\cdot)\|_{L^2(D)} \quad \forall y \in \Gamma, \tag{5.3}$$

where C_D is the constant of the Poincaré–Friedrichs inequality on D. Also, elliptic regularity yields

$$\|u(y,\cdot)\|_{H^2(D)} \leq C_{0,B} \|f(y,\cdot)\|_{L^2(D)} \quad \forall y \in \Gamma, \tag{5.4}$$

where $C_{0,B}$ is a constant which depends on D and $\|a\|_{L^\infty(\Gamma;W^{1,\infty}(D))}$. Finally, take derivatives with respect to y_n in (2.7), proceed as in the derivation of (5.3), and follow an inductive argument arriving at

$$\frac{\|\partial_{y_n}^{q_n+1} u(y,\cdot)\|_{H_0^1(D)}}{(q_n+1)!} \leq (r_n)^{q_n} \frac{C_D}{a_{min}} \big(\|\partial_{y_n} f(y,\cdot)\|_{L^2(D)} + r_n \|f(y,\cdot)\|_{L^2(D)}\big), \quad q_n \geq 0, \tag{5.5}$$

with $r_n \equiv \sqrt{\lambda_n}\|\frac{b_n}{a}\|_{L^\infty(\Gamma\times D)}$, and $n = 1, \ldots, N$. As a consequence of (3.7), (5.4), and (5.5) we have an a priori error estimate for the $k \times h$-SGFEM in the energy norm.

PROPOSITION 5.1. *Let u be the solution of the problem* (2.7) *and $u_{kh} \in Y_k^q \otimes X_h^d$ be the $k \times h$-SGFEM approximations of u defined in* (5.1). *If $\rho \in L^\infty(\Gamma)$ and $f \neq 0$, then*

$$\begin{aligned} &\frac{\|u - u_{kh}\|_E}{\sqrt{\|a\,\rho\|_{L^\infty(\Gamma\times D)}}\|f\|_{L^2(\Gamma;L^2(D))}} \\ &\qquad \leq Ch + \frac{C_D}{a_{min}} \sum_{n=1}^{N} \frac{k_n}{2} \left(\frac{k_n r_n}{2}\right)^{q_n} \left(\frac{\|\partial_{y_n} f\|_{L^2(\Gamma;L^2(D))}}{\|f\|_{L^2(\Gamma;L^2(D))}} + r_n\right), \end{aligned} \tag{5.6}$$

where the constant C depends on D and a and is independent of q, k, h, and u.

The next step is to use Proposition 5.1 together with a duality technique to estimate the $H_0^1(D)$ and $L^2(D)$ errors in the approximation of the expected value of $u(Y,\cdot)$.

THEOREM 5.1. *Let u be the solution of the problem* (2.7) *and $u_{kh} \in Y_k^q \otimes X_h^d$ be the $k \times h$-SGFEM approximations of u defined in* (5.1). *If $\rho \in L^\infty(\Gamma)$, then for $\ell = 0, 1$ we have*

$$\frac{\big\|E[u(Y,\cdot)] - E[u_{kh}(Y,\cdot)]\big\|_{H^\ell(D)}}{\|a\,\rho\|_{L^\infty(\Gamma\times D)}} \leq C \left(h^{2-\ell} + \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{2-\ell} \left(\frac{k_n r_n}{2}\right)^{(2-\ell)q_n}\right). \tag{5.7}$$

The constant C depends on D, f, and a, and it is independent of k, q, and h.

Proof. The case $\ell = 1$ follows directly from Proposition 5.1, so we now prove the case $\ell = 0$. Denote $e \equiv u - u_{kh}$ and use the auxiliary function $\hat{u}$ that solves

$$\begin{aligned} -\nabla_{x^\bullet}(a(y,\cdot)\,\nabla\hat{u}(y,\cdot)) &= E[e](\cdot) \ \text{ in } \ D, \\ \hat{u}(y,\cdot) &= 0 \ \text{ on } \ \partial D. \end{aligned} \tag{5.8}$$

Then (5.4) reads

$$\|\hat{u}(y,\cdot)\|_{H^2(D)} \le C_{0,B}\,\|E[e]\|_{L^2(D)} \quad \forall y \in \Gamma, \tag{5.9}$$

and, since $E[e]$ is independent of y, the estimate (5.5) for the problem (5.8) reads

$$\frac{\|\partial_{y_n}^{q_n+1}\hat{u}(y,\cdot)\|_{H_0^1(D)}}{(q_n+1)!} \le (r_n)^{q_n+1}\frac{C_D}{a_{min}}\|E[e]\|_{L^2(D)} \quad \forall y \in \Gamma. \tag{5.10}$$

Now use Galerkin orthogonality, together with (5.8), to obtain

$$\int_\Gamma \rho\,(E[e],e)_{L^2(D)}\,dy = \int_\Gamma \rho\,\big(a\nabla e, \nabla(\hat{u}-\psi)\big)_{L^2(D)}\,dy \quad \forall \psi \in Y_k^q \otimes X_h^d,$$

which yields, by the Cauchy–Schwarz inequality,

$$\|E[e]\|_{L^2(D)}^2 \le \widetilde{B}_1\,\widetilde{B}_2, \tag{5.11}$$

with

$$\widetilde{B}_1 \equiv \left(\int_\Gamma \rho\,\|\sqrt{a}\,\nabla e\|_{L^2(D)}^2\,dy\right)^{\frac{1}{2}}$$

and

$$\widetilde{B}_2 \equiv \inf_{\psi \in Y_k^q \otimes X_h^d}\left(\int_\Gamma \rho\,\|\sqrt{a}\,\nabla(\hat{u}-\psi)\|_{L^2(D)}^2\,dy\right)^{\frac{1}{2}}.$$

Next observe that $\widetilde{B}_1$ can be bounded using (5.6). Finally, use (3.7), (5.9), and (5.10) to bound $\widetilde{B}_2$ as follows:

(5.12)

$$\begin{aligned}\widetilde{B}_2 &\le C\,\|a\,\rho\|_{L^\infty(\Gamma\times D)}^{\frac{1}{2}}\left\{h\,\|\hat{u}\|_{L^2(\Gamma;H^2(D))} + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{y_n}^{q_n+1}\hat{u}\|_{L^2(\Gamma;H_0^1(D))}}{(q_n+1)!}\right\}\\ &\le C\,\|a\,\rho\|_{L^\infty(\Gamma;L^\infty(D))}^{\frac{1}{2}}\left(h + \frac{C_D}{a_{min}}\sum_{n=1}^{N}\left(\frac{r_n\,k_n}{2}\right)^{q_n+1}\right)\|E[e]\|_{L^2(D)}.\end{aligned}$$

Combining (5.11), (5.6), and (5.12), the estimate (5.7) follows. ∎

The estimates given in Proposition 5.1 and Theorem 5.1 give the optimal order of convergence with respect to k but are not optimal with respect to q. They can be improved by the analysis given in section 6, yielding exponential convergence with respect to q without the need to decrease k.

6. Stochastic Galerkin finite element method: $p \times h$-version. The goal of this section is to analyze the $p \times h$-version of the SGFEM method, which does not refine the set Γ. This method yields an exponential rate of convergence with respect to p, the degree of the polynomials used for approximation; cf. Theorem 6.2. The application of the p-version in the y-direction is motivated by the fact that u is analytic with respect to $y \in \Gamma$; cf. Lemma 6.1. The basic assumption for this section is the following.

Assumption 6.1. Let $\hat{\Gamma}_n \equiv \prod_{1\le j\le N, j\neq n} \Gamma_j$, and let $\hat{y}_n$ be an arbitrary element of $\hat{\Gamma}_n$. Then for each $\hat{y}_n \in \hat{\Gamma}_n$ let $\tilde{a}_n(\hat{y}_n) \equiv \min_{x\in D}\{E[a](x) + \sum_{1\le j\le N, j\neq n} \sqrt{\lambda_j} b_j(x) y_j\}$, and assume a slightly stronger uniform coercivity requirement; i.e., there exists a constant $\nu > 0$, independent of N, such that

$$\tilde{a}_n(\hat{y}_n) - \sqrt{\lambda_n}\|b_n\|_{L^\infty(D)} \max_{y\in\Gamma_n} |y| \ge \nu > 0 \quad \forall \hat{y}_n \in \hat{\Gamma}_n. \qquad \square$$

Observe that with the above construction we have $0 < \nu \le a_{min}$.

$p \times h$-version of the SGFEM method. The $p \times h$-version SGFEM approximation is the tensor product $u_h^p \in Z^p \otimes X_h^d$ (cf. section 3.4) that satisfies

$$(u_h^p, \chi)_E \equiv \int_\Gamma \rho \left(a \nabla u_h^p, \nabla \chi\right)_{L^2(D)} dy = \int_\Gamma \rho \left(f, \chi\right)_{L^2(D)} dy \quad \forall\, \chi \in Z^p \otimes X_h^d. \tag{6.1}$$

6.1. Error estimates. A first step in the analysis of the $p \times h$-version is to study the energy error, i.e., to consider

$$\begin{aligned}\|u - u_h^p\|_E &\le \sqrt{\|\rho a\|_{L^\infty(\Gamma\times D)}} \min_{v\in Z^p\otimes X_h^d} \|u - v\|_{L^2(\Gamma)\otimes H_0^1(D)} \\ &\le \sqrt{\|\rho a\|_{L^\infty(\Gamma\times D)}} \Big\{ \min_{v\in Z^p\otimes H_0^1(D)} \|u - v\|_{L^2(\Gamma)\otimes H_0^1(D)} \\ &\qquad + \min_{v\in L^2(\Gamma)\otimes X_h^d} \|u - v\|_{L^2(\Gamma)\otimes H_0^1(D)} \Big\}.\end{aligned}$$

This bound splits the error into an $L^2(\Gamma)$ approximation error and a standard $H_0^1(D)$ FEM approximation error. The rest of this section studies the first one, since for the second we can apply the results from Proposition 3.1 together with a density argument. The minimizer

$$\|u - u^p\|_{L^2(\Gamma)\otimes H_0^1(D)} = \min_{v\in Z^p\otimes H_0^1(D)} \|u - v\|_{L^2(\Gamma)\otimes H_0^1(D)}$$

is the projection $u^p = (\Pi_1 \dots \Pi_N)u$ with $\Pi_n : L^2(\Gamma) \otimes H_0^1(D) \to L^2(\Gamma) \otimes H_0^1(D)$ being the natural extension of the L^2 projection $\bar{\Pi}_n : L^2(\Gamma_n) \to Z_n^{p_n}$, so the difference $u - u^p$ splits into $u - u^p = (1 - \Pi_1)u + \dots + (\Pi_1 \dots \Pi_{N-1})(1 - \Pi_N)u$. In addition, the boundedness of the projections Π_n, $n = 1, \dots, N$, yields

$$\|u - u^p\|_{L^2(\Gamma)\otimes H_0^1(D)} \le \sum_{n=1}^N \|(1 - \Pi_n)u\|_{L^2(\Gamma)\otimes H_0^1(D)}. \tag{6.2}$$

Without loss of generality we now estimate the first term on the right-hand side of (6.2), since the other terms have a completely similar behavior. Moreover, since

$$\|(1 - \Pi_1)u\|^2_{L^2(\Gamma)\otimes H_0^1(D)} = \int_{\hat{\Gamma}_1} \left(\int_{\Gamma_1} \|(1 - \Pi_1)u(y_1, \hat{y}_1, \cdot)\|^2_{H_0^1(D)} dy_1 \right) d\hat{y}_1$$

it is enough to estimate

$$(E_1)^2(\hat{y}_1) \equiv \int_{\Gamma_1} \|(1 - \Pi_1)u(y_1, \hat{y}_1, \cdot)\|^2_{H_0^1(D)} dy_1, \tag{6.3}$$

and thus our analysis requires only one dimensional arguments in the y-direction. Let $\Gamma_1 = (y_{min}, y_{max})$, and consider the map $\Psi : [-1,1] \to H_0^1(D)$ defined by

$$\Psi(t) = u(y_1(t), \hat{y}_1, \cdot) \in H_0^1(D)$$

with the affine transformation, $y_1 : [-1,1] \to \Gamma_1$, $y_1(t) \equiv \left(\frac{y_{max}+y_{min}}{2}\right) + \left(\frac{y_{max}-y_{min}}{2}\right)\ t$. In the upcoming estimate of the quantities $\|d_n\|_{H_0^1(D)}$, to be proved in Lemma 6.2, we need to consider a continuation of Ψ to the complex plane, namely, the following.

LEMMA 6.1 (complex continuation). *The function $\Psi : [-1,1] \to H_0^1(D)$ can be analytically extended to the complex domain.*

Proof. Let $t_0 \in (-1,1)$. We shall prove that the real function Ψ can be represented as a power series for $|t - t_0| < r_{t_0}$ for some $r_{t_0} > 0$. Since Ψ depends linearly on f, let us assume that $f(y,x) = f(x)$ only, without loss of generality. Let $y(t) = (y_1(t), \hat{y}_1)$, and consider the formal series

$$u_F(t) \equiv \sum_{j=0}^{+\infty} \left(\frac{|\Gamma_1|(t-t_0)}{2}\right)^j u_j,$$

with $u_j \in H_0^1(D)$ satisfying

$$\int_D a(y(t_0), \cdot)\nabla u_0 \cdot \nabla v = \int_D fv \quad \forall v \in H_0^1(D) \tag{6.4}$$

and, for $j \geq 0$,

$$\int_D a(y(t_0), \cdot)\nabla u_{j+1} \cdot \nabla v = -\int_D \sqrt{\lambda_1} b_1 \nabla u_j \cdot \nabla v \quad \forall v \in H_0^1(D). \tag{6.5}$$

This construction implies $\|u_j\|_{H_0^1(D)} \leq (\sqrt{\lambda_1}\|\frac{b_1}{a(y(t_0),\cdot)}\|_{L^\infty(D)})^j \frac{C_D\|f\|_{L^2(D)}}{a_{min}}$, $j \geq 1$, and then

$$\|u_F\|_{H_0^1(D)} \leq \frac{C_D\|f\|_{L^2(D)}}{a_{min}} \frac{1}{1-q} < \infty$$

for $q \equiv \frac{|t-t_0||\Gamma_1|\sqrt{\lambda_1}}{2}\|\frac{b_1}{a(y(t_0),\cdot)}\|_{L^\infty(D)} < 1$. Thus, for any $t_0 \in (-1,1)$ and $|t - t_0| < r_{t_0} \equiv \frac{2}{|\Gamma_1|\sqrt{\lambda_1}\left\|\frac{b_1}{a(y(t_0),\cdot)}\right\|_{L^\infty(D)}}$, the function u_F can be represented as a power series in $t - t_0$. At the same time, we have the equality $u_F(t) = \Psi(t)$ for $t \in (-1,1)$ since both functions solve the linear elliptic equation

$$\begin{array}{ll} -\nabla \cdot (a(y(t),x)\ \nabla u(y(t),x)) = f(x) & \forall x \in D, \\ u(y(t),x) = 0 & \forall x \in \partial D, \end{array}$$

which has a unique solution. Then u_F is the analytic continuation of Ψ, and the proof is complete. ∎

Remark 6.1. Consider the natural extension of the real variable t to the complex variable η. Observe that $\Psi(\eta)$ from Lemma 6.1 solves

$$\begin{array}{ll} -\nabla \cdot (a(y(\eta),x)\ \nabla \Psi(\eta,x)) = f(x) & \forall x \in D, \\ \Psi(\eta,x) = 0 & \forall x \in \partial D. \end{array} \tag{6.6}$$

Following [24], we use the Legendre polynomials to prove approximation estimates for the $p \times h$-version of the SGFEM. Since the Legendre polynomials

$$p_n(t) \equiv \frac{1}{2^n n!}\frac{d^n}{dt^n}((t^2-1)^n),\ n = 0, 1, \ldots,$$

are orthogonal with respect to the $L^2(-1,1)$ inner product we have the error representation

$$(E_1)^2(\hat{y}_1) = \frac{|\Gamma_1|}{2}\sum_{n=p_1+1}^{+\infty}\frac{2}{2n+1}\|d_n\|^2_{H_0^1(D)} \tag{6.7}$$

with the corresponding Fourier coefficients

$$d_n \equiv \frac{2n+1}{2}\int_{-1}^{1}\Psi(t)p_n(t)dt \in H_0^1(D).$$

Therefore, to obtain an estimate for E_1 we shall study the convergence of the tail series in (6.7).

NOTATION 6.1. *For each $\hat{y}_1 \in \hat{\Gamma}_1$, consider the natural extension of the variable t to the complex η and introduce the real function, $\mathcal{A} : \mathbb{C} \to \mathbb{R}$:*

$$\begin{aligned}\mathcal{A}(\eta) &\equiv \min_{x\in D}\ \mathrm{Re}\{a(y_1(\eta), \hat{y}_1, x)\}\\ &= \min_{x\in D}[a(0, \hat{y}_1, x) + y_1(\mathrm{Re}\{\eta\})\sqrt{\lambda_1}b_1(x)],\end{aligned}$$

with $\mathrm{Re}\{\eta\}$ *being the real part of* $\eta \in \mathbb{C}$. *Whenever* $\mathcal{A}(\eta) \neq 0$, *the extended function* Ψ, *the solution of* (6.6), *satisfies the bound*

$$\|\Psi(\eta)\|_{H_0^1(D)} \leq C_D\frac{\|f(y_1(\eta), \hat{y}_1, \cdot)\|_{L^2(D)}}{\mathcal{A}(\eta)}, \tag{6.8}$$

with C_D *being the Poincaré constant for the domain* D. *Besides this, observe that*

$$\mathcal{A}(\eta) \geq \tilde{a}_1(\hat{y}_1) - |y_1(\mathrm{Re}\{\eta\})|\ \sqrt{\lambda_1}\|b_1\|_{L^\infty(D)}. \tag{6.9}$$

We are now ready to estimate the Fourier coefficients in (6.7).

LEMMA 6.2. *Let* $\tau \in (0,1)$. *Under Assumption* 6.1 *there exists a positive constant* $\theta_f(\hat{y}_1, \tau) > 0$ *such that*

$$\|d_n\|_{H_0^1(D)} \leq \frac{C_D\ \ \theta_f\ \ (2n+1)}{\tau\,\nu\,2^n}\int_{-1}^{1}\left(\frac{1-t^2}{t+1+\delta}\right)^n dt,$$

with $0 < \delta = \frac{2(1-\tau)\nu}{|\Gamma_1|\sqrt{\lambda_1}\|b_1\|_{L^\infty(D)}}$.

Proof. Consider

$$d_n = \frac{2n+1}{2}\int_{-1}^{1}\Psi(t)p_n(t)dt = \frac{(2n+1)(-1)^n}{n!\ \ 2^{n+1}}\int_{-1}^{1}\frac{d^n}{dt^n}\Psi(t)(1-t^2)^n dt.$$

Use the analytic continuation of the real function Ψ to the complex domain as in Lemma 6.1. The application of Cauchy's formula gives

$$\frac{d^n}{dt^n}\Psi(t) = \frac{n!\big(-1\big)^n}{2\pi i}\int_{\gamma_t}\frac{\Psi(\eta)}{(\eta-t)^{n+1}}d\eta,$$

where γ_t is a positively oriented closed circumference with the center at the real point $t \in (-1,1)$, with radius $R(t)$, and such that all singularities from Ψ are exterior to γ_t. Estimate (6.8) implies

$$\begin{aligned}\left\| \frac{d^n}{dt^n}\Psi(t) \right\|_{H_0^1(D)} &\leq \frac{C_D \ n!}{2\pi} \int_{\gamma_t} \frac{\|f(y_1(\eta), \hat{y}_1, \cdot)\|_{L^2(D)}}{\mathcal{A}(\eta)|\eta - t|^{n+1}} |d\eta| \\ &\leq \frac{C_D \ n!}{2\pi} \left(\sup_{\eta \in \gamma_t} \|f(y_1(\eta), \hat{y}_1, \cdot)\|_{L^2(D)} \right) \int_{\gamma_t} \frac{|d\eta|}{\mathcal{A}(\eta)|\eta - t|^{n+1}} \\ &\leq \frac{C_D \ n!}{(R(t))^n} \left(\sup_{\eta \in \gamma_t} \|f(y_1(\eta), \hat{y}_1, \cdot)\|_{L^2(D)} \right) \sup_{\eta \in \gamma_t} \frac{1}{\mathcal{A}(\eta)}. \end{aligned} \tag{6.10}$$

Let

$$\theta_f \equiv \sup_{t \in [-1,1]} \sup_{\eta \in \gamma_t} \|f(y_1(\eta), \hat{y}_1, \cdot)\|_{L^2(D)};$$

then estimate (6.10) implies

$$\|d_n\|_{H_0^1(D)} \leq \frac{(2n+1) C_D \theta_f}{2^{n+1}} \int_{-1}^{1} \left(\frac{1}{\inf_{\eta \in \gamma_t} \mathcal{A}(\eta)} \right) \left(\frac{1 - t^2}{R(t)} \right)^n dt. \tag{6.11}$$

Let $\tau \in (0,1)$. We want to choose $R(t)$ such that

$$\inf_{t \in [-1,1]} \inf_{\eta \in \gamma_t} \mathcal{A}(\eta) \geq \tau \nu. \tag{6.12}$$

Since Assumption 6.1 holds, then (6.12) is satisfied taking $R(t) = 1 - |t| + \delta$, with $\delta = \frac{2(1-\tau)\nu}{|\Gamma_1|\sqrt{\lambda_1}\|b_1\|_{L^\infty(D)}}$. Finally, the proof concludes by combining (6.11)–(6.12) and the definition of $R(t)$. ∎

Now we use a result from [24], namely, that we have the following.

LEMMA 6.3 (integral estimate). *Let $\xi < -1$, and define*

$$r \equiv \frac{1}{|\xi| + \sqrt{\xi^2 - 1}}, \ 0 < r < 1.$$

Then there holds

$$(-1)^n \int_{-1}^{1} \left(\frac{t^2 - 1}{t + \xi} \right)^n dt = (2r)^n 2^{n+1} \frac{n!}{(2n+1)!!} \Phi_{n,0}(r^2),$$

where $\Phi_{n,0}(r^2)$ is the Gauss hypergeometric function. Moreover, we have

$$\Phi_{n,0}(r^2) = \sqrt{1 - r^2} + \mathcal{O}\left(\frac{1}{n^{1/3}} \right)$$

uniformly with respect to $0 < r < 1$.

Finally, we can state the estimate for the size of the series in (6.7).

LEMMA 6.4. *Let $\tau \in (0,1)$. Under Assumption 6.1 there exists a positive constant $\theta_f > 0$ such that*

$$(E_1)(\hat{y}_1) \leq \frac{C_D \theta_f \sqrt{|\Gamma_1|}}{\tau \tilde{a}_1} \left(\sqrt{1 - r^2} + \mathcal{O}\left(\frac{1}{(p_1)^{1/3}} \right) \right) \sqrt{\pi} \ \ \frac{r^{p_1+1}}{\sqrt{1 - r^2}},$$

with $r \equiv \frac{1}{|\xi|+\sqrt{\xi^2-1}}, 0 < r < 1$, *and* $\xi < -(1 + \frac{2(1-\tau)\nu}{|\Gamma_1|\sqrt{\lambda_1}\|b_1\|_{L^\infty(D)}})$.

Proof. Use Lemmas 6.2 and 6.3, together with the asymptotic equivalence $\frac{(2n)!!}{(2n-1)!!} \sim \sqrt{\frac{\pi n}{2}}, n \to \infty$, yielding

$$\|d_n\|_{H_0^1(D)} \leq \frac{2C_D\theta_f}{\tau\tilde{a}_1}\sqrt{\frac{\pi n}{2}} \quad \left(\sqrt{1-r^2} + \mathcal{O}\left(\frac{1}{n^{1/3}}\right)\right) r^n.$$

Then use the result to estimate the tail of the series:

$$(E_1)^2(\hat{y}_1) = \frac{|\Gamma_1|}{2} \sum_{n=p_1+1}^{+\infty} \frac{2}{2n+1}\|d_n\|^2_{H_0^1(D)}. \qquad \square$$

The main result of this section, namely the exponential convergence with respect to the multi-index p as in [24], follows from the above lemmas; i.e., we have the following.

THEOREM 6.2. *Let* $\tau \in (0,1)$ *and* u *be the solution of* (2.6), $u \in L^2(\Gamma) \otimes H_0^1(D)$, *which is analytic with respect to* y, *onto the subspace* $Z^p \otimes H_0^1(D)$. *Under Assumption* 6.1 *there exist positive constants,* $0 < C, C_f$, *such that*

$$\begin{aligned} E_p &\equiv \min_{v \in Z^p \otimes H_0^1(D)} \|u - v\|_{L^2(\Gamma)\otimes H_0^1(D)} \\ &\leq \frac{C_D C_f}{\tau\nu}\sqrt{\pi|\Gamma|}\sum_{n=1}^{N}\left(1 + \frac{1}{\sqrt{1-r_n^2}}\mathcal{O}\left(\frac{1}{(p_n)^{1/3}}\right)\right)(r_n)^{p_n+1}, \end{aligned} \tag{6.13}$$

with $0 < r_n \equiv \frac{1}{|\xi_n|+\sqrt{\xi_n^2-1}} < 1$, *and* $\xi_n < -(1 + \frac{2(1-\tau)\nu}{|\Gamma_n|\sqrt{\lambda_n}\|b_n\|_{L^\infty(D)}})$ *for* $n = 1, \ldots, N$.

Similarly, as in the $k \times h$-version (cf. (5.7)), the p-version has a convergence result for the approximation to the expected value of the solution.

THEOREM 6.3. *With the same assumptions as in Theorem* 6.2 *and for* $\ell = 0, 1$, *we have*

$$\|E[u - u_h^p]\|_{H^\ell(D)} \leq \quad C\left(h^{2-\ell} + \frac{1}{\tau}\sum_{n=1}^{N}(r_n)^{(2-\ell)(p_n+1)}\right),$$

with $0 < r_n < 1$ *as in Theorem* 6.2 *and* $C > 0$ *independent of* h, p_n, *and* r_n.

The proof of the previous theorem uses Theorem 6.2 and is completely similar to the proof of Theorem 5.1.

Remark 6.2. Whenever the coefficients a and f are independent the constants θ_f from Lemma 6.2, C_f from Theorem 6.2, and C from Theorem 6.3 do not depend on N.

7. Double orthogonal polynomials. Here we explain the idea of double orthogonal polynomials, used by various authors in different contexts; see, e.g., [47] to compute efficiently the solution of the $k \times h$-version and the $p \times h$-version studied in sections 5 and 6, respectively. The idea is to use a special basis to decouple the system in the y-direction, yielding just a number of uncoupled systems, each one with the size and structure of one Monte Carlo realization of (4.1). The double orthogonal polynomials are able to perform the decoupling whenever the random variables in the Karhunen–Loève expansion of $a, Y_n, n = 1, \ldots, N$, are independent.

Without loss of generality we focus on the p-version, i.e., find $u_h^p \in Z^p \otimes X_h^d$ such that

$$\begin{aligned}\int_\Gamma \rho(y)(a(y,\cdot)\nabla u_h^p(y,\cdot), \nabla v(y,\cdot))_{L^2(D)}dy \\ = \int_\Gamma \rho(y)(f(y,\cdot), v(y,\cdot))_{L^2(D)}dy \quad \forall v \in Z^p \otimes X_h^d.\end{aligned} \tag{7.1}$$

Let $\{\psi_j(y)\}$ be a basis of the subspace $Z^p \subset L^2(\Gamma)$ and $\{\varphi_i(x)\}$ be a basis of the subspace $X_h^d \subset H_0^1(D)$. Write the approximate solution as

$$u_h^p(y,x) = \sum_{j,i} u_{ij}\psi_j(y)\varphi_i(x) \tag{7.2}$$

and use test functions $v(y,x) = \psi_k(y)\varphi_\ell(x)$ to find the coefficients u_{ij}. Then (7.1) gives

$$\begin{aligned}\sum_{j,i}\left(\int_\Gamma \rho(y)\psi_k(y)\psi_j(y)(a(y,\cdot)\nabla\varphi_i, \nabla\varphi_\ell)_{L^2(D)}dy\right)u_{ij} \\ = \int_\Gamma \rho(y)\psi_k(y)(f(y,\cdot),\varphi_\ell)_{L^2(D)}dy \quad \forall k,\ell,\end{aligned}$$

which can be rewritten as

$$\sum_{j,i}\left(\int_\Gamma \rho(y)\psi_k(y)\psi_j(y)K_{i,\ell}(y)dy\right)u_{ij} = \int_\Gamma \rho(y)\psi_k(y)f_\ell(y)dy \quad \forall k,\ell,$$

with $K_{i,\ell}(y) \equiv (a(y,\cdot)\nabla\varphi_i, \nabla\varphi_\ell)_{L^2(D)}$ and $f_\ell(y) \equiv (f(y,\cdot),\varphi_\ell)_{L^2(D)}$. If the diffusion coefficient, a, is a truncated Karhunen–Loève expansion, $a(y,x) = E[a](x) + \sum_{n=1}^N b_n(x)y_n$, and by the independence of the $Y_n, n = 1,\dots,N$, we have a corresponding "Karhunen–Loève" expression for the stiffness matrix $K_{i,\ell}(y) \equiv \int_D (E[a](x) + \sum_{n=1}^N b_n(x)y_n)\nabla\varphi_i(x)\cdot\nabla\varphi_\ell(x)dx = K_{i,\ell}^0 + \sum_{n=1}^N y_n K_{i,\ell}^n$ with deterministic coefficients $K_{i,\ell}^0 \equiv (E[a]\nabla\varphi_i, \nabla\varphi_\ell)_{L^2(D)}$ and $K_{i,\ell}^n \equiv (b_n\nabla\varphi_i, \nabla\varphi_\ell)_{L^2(D)}$. By the same token we have

$$\begin{aligned}\int_\Gamma \rho(y)\psi_k(y)\psi_j(y)K_{i,\ell}(y)dy = K_{i,\ell}^0\int_\Gamma \rho(y)\psi_k(y)\psi_j(y)dy \\ + \sum_{n=1}^N K_{i,\ell}^n \int_\Gamma y_n\rho(y)\psi_k(y)\psi_j(y)dy.\end{aligned}$$

Since $\psi_k \in Z^p$, with multi-index $p = (p_1,\dots,p_N)$, it is enough to take it as the product $\psi_k(y) = \prod_{r=1}^N \psi_{kr}(y_r)$, where $\psi_{kr}\colon \Gamma_r \to \mathbb{R}$ is a basis function of the subspace

$$Z^{p_r} = \operatorname{span}[1, y, \dots, y^{p_r}] = \operatorname{span}[\psi_{hr} : h = 1,\dots,p_r+1].$$

Keeping this choice of ψ_k in mind,

$$\begin{aligned}\int_\Gamma \rho(y)\psi_k(y)\psi_j(y)K_{i,\ell}(y)dy = K_{i,\ell}^0\int_\Gamma \prod_{m=1}^N \rho_m(y_m)\psi_{km}(y_m)\psi_{jm}(y_m)dy \\ + \sum_{n=1}^N K_{i,\ell}^n \int_\Gamma y_n \prod_{m=1}^N \rho_m(y_m)\psi_{km}(y_m)\psi_{jm}(y_m)dy.\end{aligned}$$

Now, for every set Γ_n, $n = 1, \ldots, N$, choose the polynomials, $\psi_j(y) = \prod_{n=1}^{N} \psi_{jn}(y_n)$, to be biorthogonal; i.e., for $n = 1, \ldots, N$ they must satisfy

$$\begin{aligned} \int_{\Gamma_n} \rho_n(z)\psi_{kn}(z)\psi_{jn}(z)dz &= \delta_{kj}, \\ \int_{\Gamma_n} z\rho_n(z)\psi_{kn}(z)\psi_{jn}(z)dz &= c_{kn}\delta_{kj}. \end{aligned} \tag{7.3}$$

To find the polynomials ψ_k we have to solve N eigenproblems, each of them with size $(1 + p_n)$. The computational work required by these eigenproblems is negligible with respect to the one required to solve for u_{ij}; cf. [23, section 8.7.2]. The orthogonality properties (7.3) for ψ_k imply the decoupling $\int_\Gamma \rho\psi_k\psi_j dy = \delta_{kj}, \int_\Gamma y_n\rho\psi_k\psi_j dy = c_{kn}\delta_{kj}$. By means of this decoupling we now conclude that

$$\begin{aligned} &\sum_{j,i} \left(\int_\Gamma \rho(y)\psi_k(y)\psi_j(y)K_{i,\ell}(y)dy \right) \\ &\quad = K^0_{i,\ell} \int_\Gamma \rho(y)\psi_k(y)\psi_j(y)dy + \sum_{n=1}^{N} K^n_{i,\ell} \int_\Gamma y_n\rho(y)\psi_k(y)\psi_j(y)dy \\ &\quad = \left(K^0_{i,\ell} + \sum_{n=1}^{N} c_{kn} \;\; K^n_{i,\ell} \right) \delta_{kj}. \end{aligned}$$

The structure of the linear system that determines u_{ij} now becomes block diagonal, with each block being coercive and with the sparsity structure identical to one deterministic FEM stiffness matrix, i.e.,

$$\begin{bmatrix} \left(K^0 + \sum\limits_{n=1}^{N} c_{1n}\, K^n \right) & 0 & \cdots & 0 \\ 0 & \left(K^0 + \sum\limits_{n=1}^{N} c_{2n}\, K^n \right) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \left(K^0 + \sum\limits_{n=1}^{N} c_{Nn}\, K^n \right) \end{bmatrix}.$$

Observe that as a consequence of the uniform coercivity assumption, each of the diagonal blocks in the system above is symmetric and strictly positive definite. The conclusion is that the computational work to find the coefficients u_{ij} in (7.2) is the same as the one needed to compute $\prod_{i=1}^{N}(1 + p_i)$ Monte Carlo realizations of u_h defined in (4.1), but the accuracies of these methods may differ. Section 8 studies this issue.

8. Asymptotical efficiency comparisons. In this section we compare the asymptotical numerical complexity for the Monte Carlo Galerkin finite element method (cf. section 4) with the stochastic Galerkin finite element method introduced in sections 5 and 6. The quantity of interest, i.e., the goal of the computation, is the expected value of the solution, $E[u]$, and its approximation is studied in both the $L^2(D)$ and the $H^1_0(D)$ sense. In all the cases, the spatial discretization is done by piecewise linear finite elements on globally quasi-uniform meshes. For the $k \times h$-SGFEM the Γ partitions are also assumed to be globally quasi-uniform. Besides this, the diffusion function a is assumed to be a truncated Karhunen–Loève expansion with independent random variables $Y_n,\ n = 1, \ldots, N$.

8.1. MCGFEM versus $k \times h$-SGFEM. Here we consider the computational work to achieve a given accuracy bounded by a positive constant TOL for both the MCGFEM and the $k \times h$-SGFEM methods. This optimal computational work indicates under which circumstances one method may be best suited. When using the MCGFEM method to approximate the solution of (1.1) in $H_0^1(D)$ the error becomes, applying Proposition 4.2 together with (4.5), that given a confidence level, $0 < c_0 < 1$, there exists a constant $C > 0$ depending only on c_0 such that

$$P\left(\left\|E[u] - \frac{1}{M}\sum_{j=1}^{M} u_h(\cdot;\omega_j)\right\|_{H_0^1(D)} \leq C\left(h + \frac{1}{\sqrt{M}}\right)\right) \geq c_0. \tag{8.1}$$

Then in the sense of (8.1) we write $E_{MCGFEM} = O(h) + O(1/\sqrt{M})$. The corresponding computational work for the MCGFEM method is $Work_{MCGFEM} = O((1/h^d)^r + 1/h^d)M$, where the parameter $1 \leq r \leq 3$ relates to the computational effort devoted to solve one linear system with n unknowns, $O(n^r)$. From now on we continue the discussion with the optimal $r = 1$ that can be achieved by means of the multigrid method; cf. [10]. Thus, choosing h and M to minimize the computational work for a given desired level of accuracy TOL > 0 yields the optimal work

$$Work^*_{MCGFEM} = O(\mathrm{TOL}^{-(2+d)}). \tag{8.2}$$

On the other hand, if we apply a $k \times h$-SGFEM with piecewise polynomials of order q in the y-direction, the computational error in $H_0^1(D)$ norm is (cf. Theorem 5.1)

$$E_{SGFEM} = O(h) + O(k^{q+1}),$$

and the corresponding computational work for the $k \times h$-version is

$$Work_{SGFEM} = O(h^{-d}(1+q)^N k^{-N}).$$

Here N is the number of terms in the truncated Karhunen–Loève expansion of the coefficients a and f and k is the discretization parameter in the y-direction. Similarly as before, we can compute the optimal work for the $k \times h$-SGFEM method, yielding

$$Work^*_{SGFEM} = O((1+q)^N \mathrm{TOL}^{-\frac{N}{q+1}} \mathrm{TOL}^{-d}).$$

Therefore, a $k \times h$-version SGFEM is likely to be preferred whenever TOL is sufficiently small and $N/2 < 1 + q$; i.e., if the number of terms in the Karhunen–Loève expansion of a is large, then the degree of approximation in the y-direction, q, has to become correspondingly large. We summarize the comparison results in Table 1, where we also include corresponding results from the $p \times h$-version, to be derived in subsection 8.2. Similarly, if we are interested in controlling the difference $\|E[u] - E[u_h]\|_{L^2(D)}$, the application of (4.2) and Proposition 4.2 for the MCGFEM method and Theorem 5.1 on the convergence of the $k \times h$-SGFEM method imply the results shown in Table 2. In this case $k \times h$-SGFEM is likely to be preferred whenever $N/4 < (q+1)$ and TOL is sufficiently small. In addition, the comparison tells us that to be able to be competitive with the Monte Carlo method when the number of relevant terms in the Karhunen–Loève expansion is not so small, an optimal method should have a high order of approximation and should avoid as much as possible the coupling between the different components of the numerical solution to preserve computational efficiency. The approach proposed by Ghanem and Spanos [22] based on

TABLE 1
Approximation of the function $E[u]$ in $H_0^1(D)$. Asymptotical numerical complexity for the MCGFEM and SGFEM methods.

	MCGFEM	$k \times h$-version SGFEM	$p \times h$-version SGFEM
$Work$	M/h^d	$\frac{(1+q)^N}{h^d k^N}$	$\frac{(1+p)^N}{h^d}$
$H_0^1(D)$ $Error$	$h + \frac{1}{\sqrt{M}}$	$h + \quad k^{q+1}$	$h + r^{(p+1)}$
$H_0^1(D)$ $Work^*$	$\mathrm{TOL}^{-(2+d)}$	$\mathrm{TOL}^{-\frac{N}{q+1}}\mathrm{TOL}^{-d}$	$(\log_r(\mathrm{TOL}))^N\mathrm{TOL}^{-d}$

TABLE 2
Approximation of the function $E[u]$ in $L^2(D)$. Asymptotical numerical complexity for the MCGFEM and SGFEM methods.

	MCGFEM	$k \times h$-version SGFEM	$p \times h$-version SGFEM
$Work$	M/h^d	$\frac{(1+q)^N}{h^d k^N}$	$\frac{(1+p)^N}{h^d}$
$L^2(D)$ $Error$	$h^2 + \frac{1}{\sqrt{M}}$	$h^2 + k^{2(q+1)}$	$h^2 + r^{2(p+1)}$
$L^2(D)$ $Work^*$	$\mathrm{TOL}^{-(2+d/2)}$	$(\mathrm{TOL})^{-\frac{N}{2(q+1)}}\ \mathrm{TOL}^{-d/2}$	$(\log_r(\mathrm{TOL}))^N\mathrm{TOL}^{-d/2}$

orthogonal polynomials has, whenever the approximate diffusion satisfies (1.2), a high order of approximation but introduces coupling between the different components of the numerical solution. The uncoupling can be achieved for linear equations using double orthogonal polynomials; see the description in section 7. With this motivation, section 6 studies the convergence of the $p \times h$-SGFEM.

8.2. MCGFEM versus $p \times h$-SGFEM. Here we consider the computational work to achieve a given accuracy for both the $p \times h$-version of SGFEM defined in (6.1) and the MCGFEM method for the approximation of $E[u]$ defined in section 4; i.e., we are interested in controlling the difference $\|E[u] - E[u_h^p]\|_{L^2(D)}$ or $\|E[u] - \frac{1}{M}\sum_{j=1}^M u_h(\cdot;\omega_j)\|_{L^2(D)}$, respectively. This computational work indicates under which circumstances one method may be better suited than the other. Besides this, let us assume that we use in our computations a piecewise linear FEM space in D. When using the MCGFEM method to approximate the expected value of the solution of (1.1), we have the optimal work required to achieve a given desired level of accuracy $\mathrm{TOL} > 0$ (cf. (8.2)):

$$Work^*_{MCGFEM} = O(1/\mathrm{TOL}^{2+\frac{d}{2}}).$$

On the other hand, if we apply a $p \times h$-version of the SGFEM, with $p_i = p,\ i = 1, \ldots, N$, the computational error is (cf. Theorem 6.3)

$$E_{SGFEM} = O(h^2) + O(r^{2(p+1)}), \quad 0 < r < 1,$$

and the corresponding computational work is (cf. section 7)

$$Work_{SGFEM} = O\left(\frac{(1+p)^N}{h^d}\right).$$

Recall that N is the number of terms in the truncated Karhunen–Loève expansion of the coefficients a and f, and k is the discretization parameter in the y-direction. As

before, we can compute the optimal work for the SGFEM method, yielding

$$Work^*_{SGFEM} \leq O((\log_r(\mathrm{TOL}))^N \mathrm{TOL}^{-\frac{d}{2}})$$

and the asymptotical comparison

$$\lim_{\mathrm{TOL}\to 0} \frac{Work^*_{SGFEM}}{Work^*_{MCGFEM}} = \lim_{\mathrm{TOL}\to 0} (\log_r(\mathrm{TOL}))^N \mathrm{TOL}^2 = 0.$$

Therefore, for sufficiently strict accuracy requirements, i.e., sufficiently small TOL, in the computation of $E[u]$, SGFEM requires less computational effort than MCGFEM. The work of Bahvalov and its subsequent extensions (cf. [7, 25, 45, 35]) generalize the standard Monte Carlo method, taking advantage of the available integrand's smoothness and yielding a faster order of convergence. The optimal work of such a method is for our case, i.e., the approximation of $E[u]$ in $L^2(D)$, $\mathcal{O}(C(N)\mathrm{TOL}^{-\frac{1}{1/2+q/N}}\mathrm{TOL}^{-d/2})$, where it is assumed that the integrand u has bounded derivatives up to order q with respect to y and the integral is done in the N-dimensional unit cube.

The result on the computational work of the $p \times h$-version of the SGFEM presented in this work is then related to the case $q = \infty$, since u is analytic with respect to y. This analyticity allows the exponential convergence with respect to p; cf. Theorem 6.3.

Notice that we discussed only the optimal asymptotical computational work required by both methods, but in practice the constants involved in the asymptotic approximations make these comparisons just indicative and not conclusive. In addition, we have studied only the case where the integrals $\int_{\Gamma_i} \rho_i y^k dy$ can be computed exactly for $k = 0, 1, \ldots$, and not considered the more general case where quadrature rules are needed to approximate such integrals. For further information we refer the reader to [6].

Acknowledgment. The authors would like to thank Prof. Anders Szepessy for many valuable discussions regarding this work.

REFERENCES

[1] I. Babuška, *On randomized solutions of Laplace's equation*, Casopis Pest. Mat., 86 (1961), pp. 269–276.

[2] I. Babuška, B. Andersson, P. J. Smith, and K. Levin, *Damage analysis of fiber composites. I. Statistical analysis on fiber scale*, Comput. Methods Appl. Mech. Engrg., 172 (1999), pp. 1–4.

[3] I. Babuška and P. Chatzipantelidis, *On solving linear elliptic stochastic partial differential equations*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4093–4122.

[4] I. Babuška and J. Chleboun, *Effects of uncertainties in the domain on the solution of Neumann boundary value problems in two spatial dimensions*, Math. Comp., 71 (2002), pp. 1339–1370.

[5] I. Babuška and J. Chleboun, *Effects of uncertainties in the domain on the solution of Dirichlet boundary value problems*, Numer. Math., 93 (2003), pp. 583–610.

[6] I. Babuška, R. Tempone, and G. Zouraris, *Galerkin Finite Element Methods for Stochastic Elliptic Partial Differential Equations*, Technical report 02-38, TICAM, Austin, TX, 2002.

[7] N. S. Bahvalov, *Approximate computation of multiple integrals*, Vestnik Moskov. Univ. Ser. Mat. Meh. Astr. Fiz. Him., 1959 (1959), pp. 3–18.

[8] Y. Ben-Haim and I. Elishakoff, *Convex Models of Uncertainty in Applied Mechanics*, Elsevier, Amsterdam, 1990.

[9] F. E. Benth and J. Gjerde, *Convergence rates for finite element approximations of stochastic partial differential equations*, Stochastics Stochastics Rep., 63 (1998), pp. 313–326.

[10] J. H. Bramble, *Multigrid Methods*, Longman Scientific and Technical, Harlow, UK, 1993.
[11] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer–Verlag, New York, 1994.
[12] R. E. Caflisch, *Monte Carlo and Quasi-Monte Carlo Methods*, Acta Numer. 7, Cambridge University Press, Cambridge, UK, 1998.
[13] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, New York, 1978.
[14] M.-K. Deb, *Solution of Stochastic Partial Differential Equations (SPDEs) Using Galerkin Method: Theory and Applications*, Ph.D. thesis, University of Texas at Austin, Austin, TX, 2000.
[15] M.-K. Deb, I. M. Babuška, and J. T. Oden, *Solution of stochastic partial differential equations using Galerkin finite element techniques*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6359–6372.
[16] R. Durrett, *Probability: Theory and Examples*, 2nd ed., Duxbury Press, Belmont, CA, 1964.
[17] I. Elishakoff, ed., *Whys and Hows in Uncertainty Modelling*, Springer–Verlag, Vienna, 1999.
[18] I. Elishakoff, Y. K. Lin, and L. P. Zhu, *Probabilistic and Convex Modelling of Acoustically Excited Structures*, Elsevier, Amsterdam, 1994.
[19] I. Elishakoff and Y. Ren, *The bird's eye view on finite element method for structures with large stochastic variations*, Comput. Methods Appl. Mech. Engrg., 168 (1999), pp. 51–61.
[20] L. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
[21] R. Ghanem and J. Red-Horse, *Propagation of probabilistic uncertainty in complex physical systems using a stochastic finite element approach*, Phys. D, 133 (1999), pp. 137–144.
[22] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer–Verlag, New York, 1991.
[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[24] W. Gui and I. Babuška, *The h, p and h-p versions of the finite element method in* 1 *dimension*. I. *The error analysis of the p-version*, Numer. Math., 49 (1986), pp. 577–612.
[25] S. Haber, *Stochastic quadrature formulas*, Math. Comp., 23 (1969), pp. 751–764.
[26] E. Haugen, *Probabilistic Mechanical Design*, Wiley, New York, 1980.
[27] H. Holden, B. Øksendal, J. Ubøe, and T. Zhang, *Stochastic Partial Differential Equations*, Birkhäuser Boston, Boston, MA, 1996.
[28] J. L. Jensen, L. W. Lake, P. W. Corbett, and D. Logan, *Statistics for Petroleum Engineers and Geoscientists*, Prentice–Hall, Upper Saddle River, NJ, 1997.
[29] E. Jouini, J. Cvitanić, and M. Musiela, eds., *Option Pricing, Interest Rates and Risk Management*, Cambridge University Press, Cambridge, UK, 2001.
[30] M. Kleiber and T.-D. Hien, *The Stochastic Finite Element Method*, Wiley, Chichester, 1992.
[31] J. E. Lagnese, *General boundary value problems for differential equations of Sobolev type*, SIAM J. Math. Anal., 3 (1972), pp. 105–119.
[32] S. Larsen, *Numerical Analysis of Elliptic Partial Differential Equations with Stochastic Input Data*, Ph.D. thesis, University of Maryland, College Park, MD, 1986.
[33] P. Lévy, *Processus Stochastiques et Mouvement Brownien*, 10th ed., Éditions Jacques Gabay, Paris, 1992.
[34] G. Maymon, *Some Engineering Applications in Random Vibrations and Random Structures*, Progress in Astronautics and Aeronautics 178, P. Zarchan, ed., American Institute of Aeronautics and Astronautics, Reston, VA, 1998.
[35] E. Novak, *Stochastic properties of quadrature formulas*, Numer. Math., 53 (1988), pp. 609–620.
[36] B. Øksendal, *Stochastic Differential Equations. An Introduction with Applications*, 5th ed., Springer–Verlag, Berlin, 1998.
[37] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Dover, New York, 1990.
[38] P. J. Roache, *Verification and Validation in Computational Science and Engineering*, Hermosa, Albuquerque, NM, 1998.
[39] J. B. Roberts and P.-T. D. Spanos, *Random Vibration and Statistical Linearization*, Wiley, Chichester, 1990.
[40] C. Schwab and R.-A. Todor, *Sparse finite elements for elliptic problems with stochastic loading*, Numer. Math., 95 (2003), pp. 707–734.
[41] I. M. Sobol′, *A Primer for the Monte Carlo Method*, CRC Press, Boca Raton, FL, 1994.
[42] I. M. Sobol′, *On quasi-Monte Carlo integrations*, Math. Comput. Simulation, 47 (1998), pp. 103–112.
[43] J. Sólnes, *Stochastic Processes and Random Vibrations*, Wiley, New York, 1997.
[44] T. G. Theting, *Solving Wick-stochastic boundary value problems using a finite element*

method, Stochastics Stochastics Rep., 70 (2000), pp. 241–270.

[45] J. F. Traub and A. G. Werschulz, *Complexity and Information*, Cambridge University Press, Cambridge, UK, 1998.

[46] G. Våge, *Variational methods for PDEs applied to stochastic partial differential equations*, Math. Scand., 82 (1998), pp. 113–137.

[47] T. Werder, K. Gerdes, D. Schötzau, and C. Schwab, *hp-discontinuous Galerkin time stepping for parabolic problems*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 49–50.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 826–842

LEAST-SQUARES METHODS FOR LINEAR ELASTICITY*

ZHIQIANG CAI† AND GERHARD STARKE‡

Abstract. This paper develops least-squares methods for the solution of linear elastic problems in both two and three dimensions. Our main approach is defined by simply applying the L^2 norm least-squares principle to a stress-displacement system: the constitutive and the equilibrium equations. It is shown that the homogeneous least-squares functional is elliptic and continuous in the $H(\mathrm{div};\Omega)^d \times H^1(\Omega)^d$ norm. This immediately implies optimal error estimates for finite element subspaces of $H(\mathrm{div};\Omega)^d \times H^1(\Omega)^d$. It admits optimal multigrid solution methods as well if Raviart–Thomas finite element spaces are used to approximate the stress tensor. Our method does not degrade when the material properties approach the incompressible limit. Least-squares methods that impose boundary conditions weakly and use an inverse norm are also considered. Numerical results for a benchmark test problem of planar elasticity are included in order to illustrate the robustness of our method in the incompressible limit.

Key words. least-squares method, mixed finite element method, linear elasticity, incompressible limit

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/S0036142902418357

1. Introduction.

The primitive physical equations for linear elastic problems are the constitutive equation, which expresses a relation between the stress and strain tensors, and the equilibrium equation. This first-order partial differential system is called the stress-displacement formulation. Substituting the stress into the equilibrium equation leads to a second-order elliptic partial differential system called the pure displacement formulation. However, the stress-displacement formulation is preferable to the pure displacement formulation for some important practical problems, e.g., modeling of nearly incompressible or incompressible materials and modeling of plastic materials where the elimination of the stress tensor is difficult. In addition, the stress is usually a physical quantity of primary interest. It can be obtained in the pure displacement method by differentiating displacement, but this degrades the order of the approximation.

A mixed finite element method is based on the weak form of the stress-displacement formulation, and it requires a stable combination of finite element spaces to approximate these variables. Unlike mixed methods for second-order scalar elliptic boundary value problems, stress-displacement finite elements are extremely difficult to construct. This is caused by the symmetry constraint of the stress tensor. Recently, Arnold and Winther in [3] constructed a family of stable conforming elements in two dimensions on a triangular tessellation. Their simplest element has 21 stress and 3 displacement degrees of freedom per triangle. The local degrees of freedom are reduced to 12 for the stress and 3 for the displacement for a stable nonconforming element in [4]. For previous work on mixed methods for linear elasticity, see [3] and

*Received by the editors November 20, 2002; accepted for publication (in revised form) October 9, 2003; published electronically June 4, 2004.
http://www.siam.org/journals/sinum/42-2/41835.html
†Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067 (zcai@math.purdue.edu). This author's work was sponsored in part by the National Science Foundation under grants INT-9910010 and INT-0139053 and by Korea Research Foundation under grant KRF-2002-015-C00014.
‡Institut für Angewandte Mathematik, Universität Hannover, Welfengarten 1, 30167 Hannover, Germany (starke@ifam.uni-hannover.de).

references therein. Like scalar elliptic problems, mixed methods lead to saddle-point problems. Many solution methods that work well for symmetric and positive definite problems cannot be applied directly. Although substantial progress in solution methods for saddle-point problems has been achieved, these problems may still be difficult and expensive to solve.

In recent years there has been increasing interest in the use of least-squares principles for numerical approximations of partial differential equations and systems (see, e.g., the survey paper [6], the monograph [18], and references therein). Their advantages over the usual mixed finite element discretizations include that the choice of finite element spaces is not subject to the stability condition (see, e.g., [9]), that the resulting algebraic equations can be solved efficiently by standard multigrid methods or preconditioned by well-known techniques, and that the value of a least-squares functional provides a free, sharp, and practical a posteriori error indicator which can be used efficiently in a local refinement process. For linear elasticity, in particular, least-squares methods have an additional edge over mixed methods in that the known stable mixed elements are very limited and they have a large number of degrees of freedom. In [11], Cai, Manteuffel, McCormick, and Parter proposed a two-stage least-squares approach that first solves for the displacement gradient and then solves for the displacement itself (if desired). Physical quantities such as the strain, the stress, and the rotation are then simple linear combinations of the displacement gradient. At the first stage, it has four (nine) variables in two (three) dimensions, compared to five (nine) variables for the stress-displacement formulation. One drawback of this approach is its requirement of sufficient smoothness on the original problem if using standard continuous finite element approximations. Another approach was proposed by engineers in [19] based on a displacement-stress-rotation formulation; it has the same drawback as that of [11]. In addition, it introduces extra variables (the rotation): one (three) variable in two (three) dimensions. For other least-squares approaches in the engineering literature in solid mechanics, see references in [19].

In contrast to these approaches, our aim is to develop a least-squares approach that does not have the above-mentioned drawbacks, and that computes the stress and the displacement directly. Thus it would be easier to extend this method to applications such as nonlinear elasticity, plasticity, etc. The stress components are physical quantities of primary interest in many practical applications including coupling of elastic deformation with fluid flow models. The method to be developed in this paper is based on the primitive physical equations of linear elasticity: the stress-displacement formulation, *without introducing any new variables or any new equations*. Applying the L^2 norm least-squares principle to this first-order system with an appropriately scaled constitutive equation, we develop a least-squares formulation for linear elasticity. It is shown that the homogeneous least-squares functional is elliptic and continuous in the $H(\mathrm{div};\Omega)$ norm for the stress and in the H^1 norm for the displacement uniformly with respect to material constants. This immediately implies optimal error estimates for finite element subspaces of $H(\mathrm{div};\,\Omega)^d \times H^1(\Omega)^d$. It also admits optimal multigrid solution methods if Raviart–Thomas finite element spaces (see, e.g., [9]) are used to approximate the stress tensor. Both discretization accuracy and multigrid convergence rate of the method do not degrade when the material properties approach the incompressible limit. As usual, the evaluation of the least-squares functional on each element is a practical and sharp a posteriori error indicator for adaptive mesh refinements. The practical performance of the resulting

adaptive strategy will be tested numerically for a common benchmark problem of linear elasticity in the final section of this paper.

The method here is closely related to our previous work in [12, 10]. The main difference is the scale in the constitutive equation. The homogeneous least-squares functionals in [12, 10] are equivalent to the $H(\mathrm{div};\Omega)$ norm for the stress and the *energy* norm for the displacement. This means that the least-squares variational problems in [12, 10] do not apply for incompressible materials and require effective discretizations and efficient solvers for the pure displacement problem when materials are nearly incompressible. These tasks remain difficult and expensive although some progress has been achieved (see, e.g., [8] and references therein).

For completeness, we study an inverse norm least-squares functional and show that its homogeneous form is equivalent to the $L^2(\Omega)^{d\times d}\times H^1(\Omega)^d$ norm for the stress and the displacement. This functional can be used to develop a discrete inverse norm least-squares method (see, e.g., [7]). For some applications, it is convenient to impose boundary conditions weakly by adding boundary functionals. Such a functional is also studied in this paper. See [21] for how to use these types of functionals to develop a computationally feasible numerical method.

An outline of the paper is as follows. The stress-displacement formulation for the linear elastic problem is introduced in section 2, along with some notation, the pure displacement formulation, and some regularity estimates. Section 3 develops the least-squares functionals based on the stress-displacement formulation and establishes their ellipticity and continuity. Section 4 discusses finite element approximations. Section 5 studies a least-squares functional with boundary terms that enforces boundary conditions weakly. Finally, numerical results for a benchmark test problem of linear elasticity are presented in section 6.

1.1. Notation. We use the standard notation and definitions for the Sobolev spaces $H^s(\Omega)^d$ and $H^s(\partial\Omega)^d$ for $s\geq 0$; the standard associated inner products are denoted by $(\cdot,\cdot)_{s,\Omega}$ and $(\cdot,\cdot)_{s,\partial\Omega}$, and their respective norms are denoted by $\|\cdot\|_{s,\Omega}$ and $\|\cdot\|_{s,\partial\Omega}$. (We suppress the superscript d because their dependence on dimension will be clear by context. We also omit the subscript Ω from the inner product and norm designation when there is no risk of confusion.) For $s=0$, $H^s(\Omega)^d$ coincides with $L^2(\Omega)^d$. In this case, the inner product and norm will be denoted by $\|\cdot\|$ and $(\cdot,\cdot)$, respectively. Set $H^1_D(\Omega):=\{q\in H^1(\Omega)\,:\,q=0 \text{ on } \Gamma_D\}$. We use $H_D^{-1}(\Omega)$ and $H^{-\frac{1}{2}}(\partial\Omega)$ to denote the dual of $H^1_D(\Omega)$ and $H^{\frac{1}{2}}(\partial\Omega)$ with norms defined by

$$\|\phi\|_{-1,\,D}=\sup_{0\neq\psi\in H^1_D(\Omega)}\frac{(\phi,\,\psi)}{\|\psi\|_1}\quad\text{and}\quad\|\phi\|_{-1/2,\,\partial\Omega}=\sup_{0\neq\psi\in H^{\frac{1}{2}}(\partial\Omega)}\frac{(\phi,\,\psi)}{\|\psi\|_{1/2,\partial\Omega}}.$$

Denote the product space $H_D^{-1}(\Omega)^d=\prod_{i=1}^d H_D^{-1}(\Omega)$ with the standard product norm. Set

$$H(\mathrm{div};\Omega)=\{\mathbf{v}\in L^2(\Omega)^2\,:\,\nabla\cdot\mathbf{v}\in L^2(\Omega)\},$$

which is a Hilbert space under the norm

$$\|\mathbf{v}\|_{H(\mathrm{div};\,\Omega)}=\left(\|\mathbf{v}\|^2+\|\nabla\cdot\mathbf{v}\|^2\right)^{\frac{1}{2}}.$$

2. Linear elasticity and preliminaries. Let Ω be a bounded, open, connected subset of $\Re^d$ ($d=2$ or 3) with a Lipschitz continuous boundary $\partial\Omega$. Denote $\mathbf{n}=(n_1,\,\ldots\,,n_d)^t$ as the outward unit vector normal to the boundary. We partition the

boundary of the domain Ω into two open subsets Γ_D and Γ_N such that $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. For simplicity, we assume that Γ_D is not empty (i.e., mes $(\Gamma_D) \neq 0$). Our approaches proposed in this paper can be easily extended to the pure traction problem ($\Gamma_D = \emptyset$) by excluding the space of infinitesimal rigid motions.

Let $\mathbf{f} = (f_1, \ldots, f_d)^t$ be a given body force defined on Ω. The linear elastic problem consists of finding a displacement field $\mathbf{u} = (u_1, \ldots, u_d)^t$ and a stress tensor $\boldsymbol{\sigma} = \big(\sigma_{ij}\big)_{d\times d}$ that satisfy the equilibrium equation

$$\sum_{j=1}^{d} \frac{\partial \sigma_{ij}}{\partial x_j} + f_i = 0 \quad \text{for } i = 1, \ldots, d \tag{2.1}$$

and boundary conditions

$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \sum_{j=1}^{d} \sigma_{ij}\, n_j = 0 \quad \text{on } \Gamma_N \quad \text{for } i = 1, \ldots, d. \tag{2.2}$$

For simplicity, here we assume that the boundary conditions are homogeneous.

Denote $\boldsymbol{\epsilon}(\mathbf{u}) = (\epsilon_{ij}(\mathbf{u}))_{d\times d}$ as the linearized strain tensor, where

$$\epsilon_{ij}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right).$$

The constitutive law expresses a relation between the stress and the strain tensors:

$$\boldsymbol{\sigma} = \mathcal{C}\,\boldsymbol{\epsilon}(\mathbf{u}) \quad \text{or} \quad \boldsymbol{\epsilon}(\mathbf{u}) = \mathcal{A}\,\boldsymbol{\sigma}, \tag{2.3}$$

where $\mathcal{C}$ and $\mathcal{A}$ are the elasticity and the compliance tensors of fourth order, respectively. Denote by tr the trace operator

$$\mathrm{tr}\big(\boldsymbol{\epsilon}(\mathbf{u})\big) = \epsilon_{11}(\mathbf{u}) + \cdots + \epsilon_{dd}(\mathbf{u}) = \nabla \cdot \mathbf{u},$$

where $\nabla\cdot$ stands for the divergence operator. For an isotropic elastic material, the elasticity tensor has the following simple expression:

$$\mathcal{C}\,\boldsymbol{\epsilon}(\mathbf{u}) = \lambda\,\mathrm{tr}\big(\boldsymbol{\epsilon}(\mathbf{u})\big)\boldsymbol{\delta} + 2\mu\,\epsilon(\mathbf{u}), \tag{2.4}$$

where $\boldsymbol{\delta} = (\delta_{ij})_{d\times d}$ is the identity tensor, and positive constants λ and μ are the Lamé constants such that $\mu \in [\mu_1, \mu_2]$ with $0 < \mu_1 < \mu_2$ and $\lambda \in (0, \infty)$. Materials are said to be nearly incompressible or incompressible when λ is very large or infinite, respectively. Note that both the stress and the strain tensors are symmetric. Such symmetry of the stress stems from the conservation of angular momentum.

For a second-order tensor $\boldsymbol{\tau} = (\tau_{ij})_{d\times d}$, define its divergence and normal by

$$\nabla \cdot \boldsymbol{\tau} = \begin{pmatrix} \partial\tau_{11}/\partial x_1 + \cdots + \partial\tau_{1d}/\partial x_d \\ \vdots \\ \partial\tau_{d1}/\partial x_1 + \cdots + \partial\tau_{dd}/\partial x_d \end{pmatrix} \quad \text{and} \quad \mathbf{n}\cdot\boldsymbol{\tau} = \begin{pmatrix} n_1\tau_{11} + \cdots + n_d\tau_{1d} \\ \vdots \\ n_1\tau_{d1} + \cdots + n_d\tau_{dd} \end{pmatrix},$$

respectively. That is, the divergence and normal operators apply to each row of the tensor. Then the stress-displacement system in (2.3), (2.1), and (2.2) may be rewritten in the compact form

$$\begin{cases} \boldsymbol{\sigma} - \mathcal{C}\,\boldsymbol{\epsilon}(\mathbf{u}) &= \mathbf{0} & \text{in} \quad \Omega, \\ \nabla\cdot\boldsymbol{\sigma} &= -\mathbf{f} & \text{in} \quad \Omega \end{cases} \tag{2.5}$$

with the boundary conditions

$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot \boldsymbol{\sigma} = \mathbf{0} \quad \text{on } \Gamma_N. \tag{2.6}$$

There are three approaches to treating this first-order partial differential system. One is to substitute the stress into the equilibrium equation to get the pure displacement formulation in (2.7). Numerical methods based on this formulation are not desirable for accurate approximations of the stress and for some important practical problems such as the modeling of nearly incompressible or incompressible or plastic materials. Another approach is to find the unique saddle point $(\boldsymbol{\sigma}, \mathbf{u}) \in H_N^S(\text{div}; \Omega)^d \times L^2(\Omega)^d$ of the Hellinger–Reissner functional

$$\mathcal{J}(\boldsymbol{\tau}, \mathbf{v}) = \frac{1}{2}(\mathcal{A}\boldsymbol{\tau}, \boldsymbol{\tau}) + (\nabla \cdot \boldsymbol{\tau} + \mathbf{f}, \mathbf{v}).$$

Here $H_N^S(\text{div}; \Omega)^d$ denotes the space of square-integrable symmetric tensors with square-integrable divergence and homogeneous normal on Γ_N. Equivalently, $(\boldsymbol{\sigma}, \mathbf{u})$ satisfies the following weak formulation:

$$(\mathcal{A}\boldsymbol{\sigma}, \boldsymbol{\tau}) + (\mathbf{u}, \nabla \cdot \boldsymbol{\tau}) + (\nabla \cdot \boldsymbol{\sigma}, \mathbf{v}) = (-\mathbf{f}, \mathbf{v}) \quad \forall\, (\boldsymbol{\tau}, \mathbf{v}) \in H_N^S(\text{div}; \Omega)^d \times L^2(\Omega)^d.$$

Numerical methods based on this formulation require a stable combination of finite element spaces to approximate the stress and the displacement. Known stable mixed elements are very limited and have a large number of degrees of freedom. In addition, the resulting indefinite algebraic system is still difficult and expensive to solve. In this paper, we study the third approach based on the least-squares principle that automatically stabilizes the stress-displacement system (see section 3).

We complete this section by deriving the pure displacement formulation and describing some regularity estimates. To this end, eliminating the stress in system (2.5)–(2.6) yields the pure displacement formulation which satisfies the following second-order elliptic partial differential system:

$$\begin{cases} 2\mu \nabla \cdot \boldsymbol{\epsilon}(\mathbf{u}) + \lambda \nabla(\nabla \cdot \mathbf{u}) &= -\mathbf{f} \quad \text{in} \quad \Omega, \\ \mathbf{u} &= \mathbf{0} \quad \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot (2\mu\, \boldsymbol{\epsilon}(\mathbf{u}) + \lambda\, (\nabla \cdot \mathbf{u})\, \boldsymbol{\delta}) &= \mathbf{0} \quad \text{on} \quad \Gamma_N, \end{cases} \tag{2.7}$$

where ∇ stands for the gradient operator. The energy norm associated with the above problem is defined as follows:

$$|||\mathbf{v}||| = \left(2\mu\, \|\boldsymbol{\epsilon}(\mathbf{v})\|^2 + \lambda\, \|\nabla \cdot \mathbf{v}\|^2\right)^{\frac{1}{2}}. \tag{2.8}$$

By using Korn's inequality (see [14]),

$$\|\mathbf{v}\|_1 \le C\, \|\boldsymbol{\epsilon}(\mathbf{v})\| \quad \forall\, \mathbf{v} \in H_D^1(\Omega)^d, \tag{2.9}$$

the energy norm is equivalent to the H^1 norm for a fixed λ. In this paper, we use C with or without subscripts to denote a generic positive constant, possibly different at different occurrences, which is independent of the Lamé constant λ and the mesh size h introduced in section 4 but may depend on the domain Ω. Note that one could scale the variables and the right-hand side accordingly so that μ is equal to one. We will frequently use the term *uniform* in reference to a relation to mean that it holds independent of λ and h.

The weak form of boundary value problem (2.7) has a unique solution $\mathbf{u} \in H_D^1(\Omega)^d$ for any $\mathbf{f} \in H_D^{-1}(\Omega)^d$ (see [14]). Moreover, the solution satisfies the following H^1-regularity estimate (see, e.g., [8, 14]):

$$\|\mathbf{u}\|_1 + \lambda \|\nabla \cdot \mathbf{u}\| \leq C \|\mathbf{f}\|_{-1,D}. \tag{2.10}$$

Furthermore, if the domain Ω is convex or its boundary is $C^{1,1}$ and if either Γ_D or Γ_N is empty, then the H^2-regularity estimate

$$\|\mathbf{u}\|_2 + \lambda \|\nabla \cdot \mathbf{u}\|_1 \leq C \|\mathbf{f}\| \tag{2.11}$$

holds. Both the regularity estimates in (2.10) and (2.11) suggest that the divergence of the displacement has a different scale from the displacement itself for large λ.

3. Least-squares variational formulation. In this section, we first discuss an appropriate stress-displacement formulation and then consider the corresponding least-squares functionals based on such a first-order system. Our primary objective here is to establish continuity and ellipticity of these least-squares functionals in the appropriate Hilbert spaces.

It is convenient to view $d \times d$-matrices as d^2-vectors and vice versa. For example, view $(\sigma_{ij})_{d\times d}$ as $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_d)^t$, where $\boldsymbol{\sigma}_j = (\sigma_{j1}, \ldots, \sigma_{jd})$ is the jth row of $(\sigma_{ij})_{d\times d}$ for $j = 1, \ldots, d$. Let

$$\mathbf{b} = \begin{cases} (1,\,0,\,0,\,1)^t & d = 2, \\ (1,\,0,\,0,\,0,\,1,\,0,\,0,\,0,\,1)^t & d = 3, \end{cases}$$

which may be viewed as the $d \times d$ identity matrix $I_{d\times d}$ or the identity tensor $\boldsymbol{\delta}$. Thus,

$$\operatorname{tr} \boldsymbol{\sigma} = \operatorname{tr} (\sigma_{ij})_{d\times d} = \sum_{i=1}^{d} \sigma_{ii} = \mathbf{b}^t \begin{pmatrix} \boldsymbol{\sigma}_1^t \\ \vdots \\ \boldsymbol{\sigma}_d^t \end{pmatrix} = \mathbf{b}^t \boldsymbol{\sigma}.$$

By viewing a $d \times d$-matrix as a d^2-vector, we can then write the fourth-order elasticity tensor as $d^2 \times d^2$-matrix

$$\mathcal{C} = \lambda\, \mathbf{b}\mathbf{b}^t + 2\mu\, I.$$

It is clear that $\mathcal{C}$ is symmetric and that $\mathcal{C}$ is positive definite for finite λ. The compliance tensor has the form

$$\mathcal{A} = \frac{1}{2\mu} \left(I - \frac{\lambda}{d\lambda + 2\mu} \mathbf{b}\mathbf{b}^t \right). \tag{3.1}$$

Note that $\mathcal{A} = \mathcal{C}^{-1}$ for finite λ. When the λ approaches ∞, the elasticity tensor blows up and the compliance tensor tends to

$$\frac{1}{2\mu} \left(I - \frac{1}{d}\, \mathbf{b}\mathbf{b}^t \right) \equiv \frac{1}{2\mu}\, \mathbf{dev}, \tag{3.2}$$

which is not invertible. For any tensor $\boldsymbol{\tau}$, $\mathbf{dev}\boldsymbol{\tau} = \boldsymbol{\tau} - \frac{1}{d}\,(\operatorname{tr}\boldsymbol{\tau})\boldsymbol{\delta}$ is the deviatoric part of $\boldsymbol{\tau}$. Hence, for nearly incompressible or incompressible materials, it is preferable to use the following strain and stress relation:

$$\boldsymbol{\epsilon}(\mathbf{u}) = \mathcal{A}\, \boldsymbol{\sigma} = \frac{1}{2\mu} \left(\boldsymbol{\sigma} - \frac{\lambda}{d\lambda + 2\mu} (\operatorname{tr} \boldsymbol{\sigma}) \boldsymbol{\delta} \right). \tag{3.3}$$

Now, the first-order system for the stress and the displacement of linear elasticity is as follows:

$$\left\{\begin{array}{rcll} \mathcal{A}\,\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u}) &=& \mathbf{0} & \text{in} \quad \Omega, \\ \nabla\cdot\boldsymbol{\sigma} + \mathbf{f} &=& \mathbf{0} & \text{in} \quad \Omega \end{array}\right. \tag{3.4}$$

with boundary conditions (2.6), where $\mathcal{A}$ is given in (3.1).

It is important to note that the stress is symmetric; i.e.,

$$\boldsymbol{\sigma} - \boldsymbol{\sigma}^t = \mathbf{0}, \tag{3.5}$$

where $\boldsymbol{\sigma}^t$ denotes the transpose of $\boldsymbol{\sigma}$ as a $d\times d$-matrix. One can impose this symmetry condition either in the solution space (in a strong sense) or in the equation (in a weak sense). In [12], we augment (3.5) with the stress-displacement system so that our least-squares methods have freedom to treat it either strongly or weakly depending on discretization and solution methods. In [10], we show that the symmetry constraint of the stress is enforced at the continuous level even without the term $\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^t\|^2$ in the least-squares functionals. This is because for any $\boldsymbol{\tau}\in L^2(\Omega)^{d\times d}$ and any $\mathbf{v}\in H^1(\Omega)^d$, we have

$$\|\boldsymbol{\tau} - \boldsymbol{\tau}^t\| \le C\,\|\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})\|. \tag{3.6}$$

Thus, $\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v}) = 0$ implies that $\boldsymbol{\tau}$ is symmetric. Therefore, we will apply the least-squares principle to first-order system (3.4) without augmenting (3.5). Inequality (3.6) follows from the symmetry of $\boldsymbol{\epsilon}(\mathbf{v})$, (3.1), and the triangle inequality that

$$\begin{aligned} \|\boldsymbol{\tau} - \boldsymbol{\tau}^t\| &= 2\mu \left\| \left(\frac{1}{2\mu}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})\right) - \left(\frac{1}{2\mu}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})\right)^t \right\| \\ &= 2\mu\,\|\,(\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})) - (\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v}))^t\,\| \le 4\mu\,\|\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})\|. \end{aligned}$$

Before defining least-squares functionals, let us first describe solution spaces. If $\Gamma_N = \emptyset$, then $\int_\Omega \nabla\cdot\mathbf{u}\,dx = \int_{\partial\Omega}\mathbf{n}\cdot\mathbf{u}\,ds = 0$, which implies

$$\int_\Omega \operatorname{tr}\boldsymbol{\sigma}\,dx = 0.$$

Therefore, we are at liberty to impose such a condition for the stress $\boldsymbol{\sigma}$. Let

$$\mathbf{X} = \left\{\begin{array}{ll} H(\mathrm{div};\,\Omega)^d & \text{if } \Gamma_N \neq \emptyset, \\ \{\boldsymbol{\tau}\in H(\mathrm{div};\,\Omega)^d \mid \int_\Omega \operatorname{tr}\boldsymbol{\sigma}\,dx = 0\} & \text{otherwise,} \end{array}\right.$$

and denote its subspace by

$$\mathbf{X}_N = \{\boldsymbol{\tau}\in\mathbf{X} \,:\, \mathbf{n}\cdot\boldsymbol{\tau} = \mathbf{0} \ \text{ on } \ \Gamma_N\}.$$

Let

$$\mathcal{V}_B = \mathbf{X}_N \times H^1_D(\Omega)^d.$$

For $\mathbf{f}\in L^2(\Omega)^d$, we define the following least-squares functionals:

$$G_{-1}(\boldsymbol{\sigma},\,\mathbf{u}\,;\,\mathbf{f}) = \|\mathcal{A}\,\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla\cdot\boldsymbol{\sigma} + \mathbf{f}\|^2_{-1,D} \tag{3.7}$$

and

$$G(\boldsymbol{\sigma}, \mathbf{u}; \mathbf{f}) = \|\mathcal{A}\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla \cdot \boldsymbol{\sigma} + \mathbf{f}\|^2 \tag{3.8}$$

for $(\boldsymbol{\sigma}, \mathbf{u}) \in \mathcal{V}_B$. Least-squares variational problems for the stress-displacement of linear elasticity are then to minimize the above least-squares functionals over $\mathcal{V}_B$. In this paper, we concentrate on the least-squares problem based on the L^2 norm functional in (3.8): find $(\boldsymbol{\sigma}, \mathbf{u}) \in \mathcal{V}_B$ such that

$$G(\boldsymbol{\sigma}, \mathbf{u}; \mathbf{f}) = \inf_{(\boldsymbol{\tau}, \mathbf{v}) \in \mathcal{V}_B} G(\boldsymbol{\tau}, \mathbf{v}; \mathbf{f}). \tag{3.9}$$

Note that the inverse norm functional in (3.7) can be used to develop a discrete inverse norm least-squares method (see [7]) as well.

REMARK 3.1. *Since the minimum of the quadratic functional $G(\boldsymbol{\sigma}, \mathbf{u}; \mathbf{f})$ is zero, by* (3.6) *the symmetry of the stress tensor is guaranteed by the first term of the functional, i.e., the constitutive equation.*

REMARK 3.2. *The least-squares functionals defined in* (3.7) *and* (3.8) *differ from those in* [12, 10] *mainly in the first term with an extra weight of $\mathcal{C}^{-\frac{1}{2}}$. More precisely, the first term of the functionals in* [12, 10] *is $\|\mathcal{C}^{-\frac{1}{2}}\boldsymbol{\sigma} - \mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{u})\|^2$. Note that $\|\mathcal{C}^{\frac{1}{2}}\boldsymbol{\epsilon}(\mathbf{u})\|^2 = \lambda \|\nabla \mathbf{u}\|^2 + 2\mu \|\boldsymbol{\epsilon}(\mathbf{u})\|^2$. This means that the least-squares variational problems in* [12, 10] *do not apply for incompressible materials and require effective discretizations and efficient solvers for the pure displacement problem when materials are nearly incompressible.*

Below we establish uniform continuity and ellipticity (i.e., equivalence) of the homogeneous functionals $G_{-1}(\boldsymbol{\tau}, \mathbf{v}; \mathbf{0})$ and $G(\boldsymbol{\tau}, \mathbf{v}; \mathbf{0})$ in terms of the respective functionals $M_{-1}(\boldsymbol{\tau}, \mathbf{v})$ and $M(\boldsymbol{\tau}, \mathbf{v})$ defined on $\mathcal{V}_B$ by

$$M_{-1}(\boldsymbol{\tau}, \mathbf{v}) = \|\boldsymbol{\epsilon}(\mathbf{v})\|^2 + \|\boldsymbol{\tau}\|^2$$

and

$$M(\boldsymbol{\tau}, \mathbf{v}) = \|\boldsymbol{\epsilon}(\mathbf{v})\|^2 + \|\boldsymbol{\tau}\|^2 + \|\nabla \cdot \boldsymbol{\tau}\|^2.$$

To do so, we need the following fundamental inequality on the trace of $\mathbf{X}_N$:

$$\|\operatorname{tr} \boldsymbol{\tau}\| \le C \left(\sqrt{(\mathcal{A}\boldsymbol{\tau}, \boldsymbol{\tau})} + \|\nabla \cdot \boldsymbol{\tau}\|_{-1,D} \right) \quad \forall\, \boldsymbol{\tau} \in \mathbf{X}_N, \tag{3.10}$$

where C is a positive constant independent of λ. This inequality should be a classic result. But the only references that we know for its proof are [1] for two dimensions and Dirichlet boundary conditions (i.e., $d = 2$ and $\Gamma_N = \emptyset$) and [12] for both two and three dimensions and general boundary conditions. Note that

$$\begin{aligned} (\mathcal{A}\boldsymbol{\tau}, \boldsymbol{\tau}) &= \frac{1}{2\mu} \left(\|\boldsymbol{\tau}\|^2 - \frac{\lambda}{d\lambda + 2\mu} \|\operatorname{tr} \boldsymbol{\tau}\|^2 \right) \\ &= \frac{1}{2\mu} \|\mathbf{dev}\, \boldsymbol{\tau}\|^2 + \frac{1}{d(d\lambda + 2\mu)} \|\operatorname{tr} \boldsymbol{\tau}\|^2, \end{aligned} \tag{3.11}$$

where $\mathbf{dev}\, \boldsymbol{\tau}$ and $\operatorname{tr} \boldsymbol{\tau}$ are the respective deviatoric and volumetric parts of $\boldsymbol{\tau}$. It is then obvious that the divergence term in (3.10) is necessary to bound the L^2 norm of the trace. From the definition of the inverse norm and the Cauchy–Schwarz inequality, we have that

$$\|\nabla \cdot \boldsymbol{\tau}\|_{-1,\,D} \le \|\boldsymbol{\tau}\|. \tag{3.12}$$

By (3.10), (3.11), and (3.12) it is easy to see that

$$\|\boldsymbol{\tau}\|_a \equiv \left((\mathcal{A}\,\boldsymbol{\tau},\,\boldsymbol{\tau}) + \|\nabla\cdot\boldsymbol{\tau}\|^2_{-1,\,D}\right)^{\frac{1}{2}}$$

is equivalent to the L^2 norm; i.e., there exists a positive constant C independent of λ such that

$$\frac{1}{C}\,\|\boldsymbol{\tau}\|^2 \le \|\boldsymbol{\tau}\|_a^2 \le C\,\|\boldsymbol{\tau}\|^2 \quad \forall\ \boldsymbol{\tau}\in\mathbf{X}_N. \tag{3.13}$$

THEOREM 3.1. *The homogeneous functionals $G_{-1}(\boldsymbol{\tau},\,\mathbf{v};\,\mathbf{0})$ and $G(\boldsymbol{\tau},\,\mathbf{v};\,\mathbf{0})$ are uniformly equivalent to the functionals $M_{-1}(\boldsymbol{\tau},\,\mathbf{v})$ and $M(\boldsymbol{\tau},\,\mathbf{v})$, respectively; i.e., there exist positive constants C_1 and C_2, independent of λ, such that*

$$\frac{1}{C_1}\,M_{-1}(\boldsymbol{\tau},\,\mathbf{v}) \le G_{-1}(\boldsymbol{\tau},\,\mathbf{v}\,;\,\mathbf{0}) \le C_1\,M_{-1}(\boldsymbol{\tau},\,\mathbf{v}) \tag{3.14}$$

and

$$\frac{1}{C_2}\,M(\boldsymbol{\tau},\,\mathbf{v}) \le G(\boldsymbol{\tau},\,\mathbf{v}\,;\,\mathbf{0}) \le C_2\,M(\boldsymbol{\tau},\,\mathbf{v}) \tag{3.15}$$

hold for all $(\boldsymbol{\tau},\,\mathbf{v})\in\mathcal{V}_B$.

Proof. It follows from (3.1) that

$$\begin{aligned}\|\mathcal{A}\,\boldsymbol{\tau}\|^2 &= \left(\frac{1}{2\mu}\right)^2\left(\|\boldsymbol{\tau}\|^2 - \frac{2\lambda}{d\lambda+2\mu}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2 + d\left(\frac{\lambda}{d\lambda+2\mu}\right)^2\|\mathrm{tr}\,\boldsymbol{\tau}\|^2\right)\\ &= \left(\frac{1}{2\mu}\right)^2\left(\|\boldsymbol{\tau}\|^2 - \frac{\lambda\,(d\lambda+4\mu)}{(d\lambda+2\mu)^2}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2\right) \le \left(\frac{1}{2\mu}\right)^2\|\boldsymbol{\tau}\|^2.\end{aligned}$$

Thus, $\mathcal{A}\,\boldsymbol{\tau}$ is bounded above by $\boldsymbol{\tau}$ in the L^2 norm:

$$\|\mathcal{A}\,\boldsymbol{\tau}\| \le \frac{1}{2\mu}\,\|\boldsymbol{\tau}\|. \tag{3.16}$$

The upper bounds in both (3.14) and (3.15) follow easily from the triangle inequality, (3.16), and (3.12). To show the validity of the lower bound in (3.14), we first prove that $\boldsymbol{\tau}$ in the L^2 norm is bounded above by the homogeneous functional:

$$\|\boldsymbol{\tau}\|^2 \le C\,G_{-1}(\boldsymbol{\tau},\,\mathbf{v}\,;\mathbf{0}) \quad \forall\ (\boldsymbol{\tau},\,\mathbf{v})\in\mathcal{V}_B. \tag{3.17}$$

To this end, by the triangle inequality and (3.16) we have

$$\|\boldsymbol{\epsilon}(\mathbf{v})\| \le \|\boldsymbol{\epsilon}(\mathbf{v}) - \mathcal{A}\,\boldsymbol{\tau}\| + \|\mathcal{A}\,\boldsymbol{\tau}\| \le \|\boldsymbol{\epsilon}(\mathbf{v}) - \mathcal{A}\,\boldsymbol{\tau}\| + \frac{1}{2\mu}\,\|\boldsymbol{\tau}\|. \tag{3.18}$$

Since $\boldsymbol{\epsilon}(\mathbf{v}) = \frac{1}{2}\,(\nabla\mathbf{v} + (\nabla\mathbf{v})^t)$ is symmetric, then integration by parts; the triangle, Cauchy–Schwarz, and Korn inequalities; and (3.6) give

$$\begin{aligned}|(\boldsymbol{\tau},\,\boldsymbol{\epsilon}(\mathbf{v}))| &= \left|(\boldsymbol{\tau},\,\nabla\mathbf{v}) - \left(\frac{\boldsymbol{\tau}-\boldsymbol{\tau}^t}{2},\,\nabla\mathbf{v}\right)\right| = \left|(\nabla\cdot\boldsymbol{\tau},\,\mathbf{v}) + \left(\frac{\boldsymbol{\tau}-\boldsymbol{\tau}^t}{2},\,\nabla\mathbf{v}\right)\right|\\ &\le \left(\|\nabla\cdot\boldsymbol{\tau}\|_{-1,\,D} + \left\|\frac{\boldsymbol{\tau}-\boldsymbol{\tau}^t}{2}\right\|\right)\,\|\mathbf{v}\|_1\\ &\le C\,\left(\|\nabla\cdot\boldsymbol{\tau}\|_{-1,\,D} + \|\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})\|\right)\,\|\boldsymbol{\epsilon}(\mathbf{v})\|,\end{aligned}$$

which, together with (3.18), implies that

$$|(\boldsymbol{\tau},\, \boldsymbol{\epsilon}(\mathbf{v}))| \le C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0}) + C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0})^{\frac{1}{2}}\, \|\boldsymbol{\tau}\|, \tag{3.19}$$

where $G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0})^{\frac{1}{2}}$ denotes the square root of $G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0})$. Now, it follows from the Cauchy–Schwarz inequality, (3.19), and (3.13) that

$$\begin{aligned}(\mathcal{A}\,\boldsymbol{\tau},\, \boldsymbol{\tau}) &= (\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v}),\, \boldsymbol{\tau}) + (\boldsymbol{\epsilon}(\mathbf{v}),\, \boldsymbol{\tau}) \\ &\le \|\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})\|\, \|\boldsymbol{\tau}\| + C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0}) + C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0})^{\frac{1}{2}}\, \|\boldsymbol{\tau}\| \\ &\le C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0}) + C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0})^{\frac{1}{2}}\, \left((\mathcal{A}\,\boldsymbol{\tau},\, \boldsymbol{\tau}) + \|\nabla\cdot\boldsymbol{\tau}\|^2_{-1,\,D}\right)^{\frac{1}{2}} \\ &\le C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0}) + C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0})^{\frac{1}{2}}\, (\mathcal{A}\,\boldsymbol{\tau},\, \boldsymbol{\tau})^{\frac{1}{2}}.\end{aligned}$$

Hence,

$$(\mathcal{A}\,\boldsymbol{\tau},\, \boldsymbol{\tau}) \le C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0}),$$

which, together with (3.13), implies the validity of (3.17). With (3.18) and (3.17), it is then easy to see that $\|\boldsymbol{\epsilon}(\mathbf{v})\|^2$ is also bounded above by the homogeneous functional $G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\mathbf{0})$. This completes the proof of the lower bound in (3.14). Since

$$G_{-1}(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{0}) \le G(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{0}) \quad \text{and} \quad \|\nabla\cdot\boldsymbol{\tau}\|^2 \le G(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{0}),$$

then the lower bound in (3.15) follows from (3.14). The proof of the theorem is therefore completed. ∎

4. Finite element approximation. We approximate the minimum of the least-squares functional $G(\boldsymbol{\sigma}, \mathbf{u}; \mathbf{f})$ in (3.9) using a Rayleigh–Ritz type finite element method. For convenience, we use two-dimensional terminology ($d = 2$). Assuming that the domain Ω is a polygon, let $\mathcal{T}_h$ be a triangulation of Ω with triangular elements of size $\mathcal{O}(h)$ that is regular (see [13]). We restrict ourselves to triangular elements for convenience because extension to either rectangular or a combination of triangular and rectangular elements is straightforward.

Since the homogeneous functional $G(\boldsymbol{\sigma},\, \mathbf{u}\,;\mathbf{0})$ is equivalent to the $H(\mathrm{div};\, \Omega)$ norm for the stress and the H^1 norm for the displacement by Theorem 3.1 and Korn's inequality (2.9), it is then natural to approximate the stress (each row) by the standard $H(\mathrm{div};\, \Omega)$ conforming Raviart–Thomas space of order k (see [20]) and the standard (conforming) continuous piecewise polynomials of degree $k+1$ for the displacement:

$$\Sigma_h^k = \{\boldsymbol{\tau} \in \mathbf{X}_N \,:\, \boldsymbol{\tau}|_K \in RT_k(K)^2 \;\forall\, K \in \mathcal{T}_h\} \subset \mathbf{X}_N, \tag{4.1}$$

$$V_h^k = \{\mathbf{v} \in C^0(\Omega)^2 \,:\, \mathbf{v}|_K \in P_k(K)^2 \;\forall\, K \in \mathcal{T}_h,\; \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D\} \subset H^1_D(\Omega)^2, \tag{4.2}$$

where $RT_k(K)$ is local Raviart–Thomas space of order k defined by

$$RT_k(K) = P_k(K)^2 + \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} P_k(K),$$

and $P_k(K)$ is the space of polynomials of degree k on triangle K. These spaces have the following approximation properties: if $k \ge 0$ is an integer and $l \in (0,\, k+1]$, then

$$\inf_{\boldsymbol{\tau}\in\Sigma_h^k} \|\boldsymbol{\sigma} - \boldsymbol{\tau}\|_{H(\mathrm{div};\,\Omega)} \le C\, h^l\, (\|\boldsymbol{\sigma}\|_l + \|\nabla\cdot\boldsymbol{\sigma}\|_l) \tag{4.3}$$

for $\boldsymbol{\sigma} \in H^l(\Omega)^{2\times 2} \cap \mathbf{X}_N$ with $\nabla \cdot \boldsymbol{\sigma} \in H^l(\Omega)^2$, and

$$\inf_{\mathbf{u}\in V_h^{k+1}} \|\mathbf{u}-\mathbf{v}\|_1 \le C\, h^l\, \|\mathbf{u}\|_{l+1} \tag{4.4}$$

for $\mathbf{u} \in H^{l+1}(\Omega)^2 \cap H_D^1(\Omega)^2$.

The finite element approximation for minimizing $G(\boldsymbol{\sigma}, \mathbf{u}; \mathbf{f})$ in (3.9) on $\mathcal{V}_B$ becomes: find $(\boldsymbol{\sigma}_h, \mathbf{u}_h) \in \Sigma_h^k \times V_h^{k+1}$ such that

$$G(\boldsymbol{\sigma}_h, \mathbf{u}_h; \mathbf{f}) = \min_{(\boldsymbol{\tau}, \mathbf{v})\in \Sigma_h^k \times V_h^{k+1}} G(\boldsymbol{\tau}, \mathbf{v}; \mathbf{f}). \tag{4.5}$$

By Theorem 3.1, (2.9), and the fact that $\Sigma_h^k \times V_h^{k+1}$ is a subspace of $\mathcal{V}_B$, we conclude that (4.5) has a unique solution and is equivalent to the weak form: find $(\boldsymbol{\sigma}_h, \mathbf{u}_h) \in \Sigma_h^k \times V_h^{k+1}$ such that

$$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h; \boldsymbol{\tau}, \mathbf{v}) = (-\mathbf{f}, \nabla \cdot \boldsymbol{\tau}) \quad \forall\ (\boldsymbol{\tau}, \mathbf{v}) \in \Sigma_h^k \times V_h^{k+1}, \tag{4.6}$$

where the bilinear form $\mathcal{F}(\cdot\,;\,\cdot)$ has the form of

$$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h; \boldsymbol{\tau}, \mathbf{v}) = (\mathcal{A}\,\boldsymbol{\sigma}_h - \boldsymbol{\epsilon}(\mathbf{u}_h), \mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})) + (\nabla \cdot \boldsymbol{\sigma}_h, \nabla \cdot \boldsymbol{\tau}).$$

Moreover, the error $(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h)$ satisfies the orthogonality property

$$\mathcal{F}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h; \boldsymbol{\tau}, \mathbf{v}) = 0 \quad \forall\ (\boldsymbol{\tau}, \mathbf{v}) \in \Sigma_h^k \times V_h^{k+1}. \tag{4.7}$$

THEOREM 4.1. *Assume that the solution,* $(\boldsymbol{\sigma}, \mathbf{u})$*, of* (3.9) *is in* $H^l(\Omega)^{2\times 2} \times H^{l+1}(\Omega)^2$ *and that the divergence of the stress,* $\nabla \cdot \boldsymbol{\sigma}$*, is in* $H^l(\Omega)^2$*. Let* $k+1$ *be the smallest integer greater than or equal to* l*. Then with* $(\boldsymbol{\sigma}_h, \mathbf{u}_h) \in \Sigma_h^k \times V_h^{k+1}$*, the following error estimate holds:*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{H(\mathrm{div};\,\Omega)} + \|\mathbf{u} - \mathbf{u}_h\|_1 \le C\, h^l(\|\boldsymbol{\sigma}\|_l + \|\nabla \cdot \boldsymbol{\sigma}\|_l + \|\mathbf{u}\|_{l+1}). \tag{4.8}$$

Proof. The proof is a simple consequence of the orthogonality property (4.7) and the approximation properties (4.3) and (4.4) of the finite element spaces $\Sigma_h^k \times V_h^{k+1}$. ∎

Theorem 3.1 indicates that the bilinear form $\mathcal{F}(\cdot\,;\,\cdot)$ is elliptic and continuous with respect to the $H(\mathrm{div};\,\Omega)$ norm for the stress and the H^1 norm for the displacement. It is then well known that multigrid methods applied to the resulting discrete system (4.6) are optimally convergent (see, e.g., [16, 2, 17, 23]).

It is obvious that the finite element approximation in (4.5) does not preserve the symmetry of the stress. But the finite element approximation of the stress is approximately symmetric. Moreover, one can obtain symmetric stress approximation with the same accuracy as $\boldsymbol{\sigma}_h$ by simply computing

$$\tilde{\boldsymbol{\sigma}}_h = \frac{1}{2}\left(\boldsymbol{\sigma}_h + \boldsymbol{\sigma}_h^t\right). \tag{4.9}$$

It should also be noted that many mixed finite element approaches commonly used produce stress approximations which do not satisfy symmetry exactly (cf. [9, sect. VII.2]).

COROLLARY 4.2. *Under the assumptions of Theorem* 4.1*, we have that*

$$\|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^t\| \le C\, h^l\, (\|\boldsymbol{\sigma}\|_l + \|\nabla \cdot \boldsymbol{\sigma}\|_l + \|\mathbf{u}\|_{l+1}) \tag{4.10}$$

and that

$$\|\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h\| \le C\, h^l \left(\|\boldsymbol{\sigma}\|_l + \|\nabla \cdot \boldsymbol{\sigma}\|_l + \|\mathbf{u}\|_{l+1}\right). \tag{4.11}$$

Proof. Since the stress $\boldsymbol{\sigma}$ is symmetric, by the triangle inequality we have that

$$\|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^t\| = \|(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h) - (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)^t\| \le 2\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|$$

and that

$$\|\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h\| = \left\|\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h) + \frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)^t\right\| \le \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|.$$

Now, (4.10) and (4.11) follow from the error bound in (4.8). ∎

Nevertheless, there may be a possible reluctance in the engineering community to accept nonsymmetric stress approximation since the symmetry is due to the conservation of angular momentum. In order to directly preserve the symmetry of finite element approximations to the stress tensor, one may enforce the symmetry constraint in the finite element approximation space. To this end, let $\mathbf{X}_N^s$ denote the symmetric stress space,

$$\mathbf{X}_N^s = \{\boldsymbol{\tau} \in \mathbf{X}_N \,|\, \boldsymbol{\tau}^t = \boldsymbol{\tau} \text{ in } \Omega\}.$$

A simple and obvious choice is to use continuous piecewise polynomials of degree k for each component of the symmetric stress:

$$\Sigma_h^{k,s} = \{\boldsymbol{\tau} \in C^0(\Omega)^{2\times 2} \cap \mathbf{X}_N^s \,:\, \boldsymbol{\tau}|_K \in P_k(K)^{2\times 2}\ \forall\, K \in \mathcal{T}_h\} \subset \mathbf{X}_N^s. \tag{4.12}$$

This space has the following approximation property: if $k \ge 1$ is an integer and $l \in (0,\, k]$, then

$$\inf_{\boldsymbol{\tau} \in \Sigma_h^{k,s}} \|\boldsymbol{\sigma} - \boldsymbol{\tau}\|_{H(\mathrm{div};\,\Omega)} \le C\, h^l\, \|\boldsymbol{\sigma}\|_{l+1} \tag{4.13}$$

for $\boldsymbol{\sigma} \in H^{l+1}(\Omega)^2 \cap \mathbf{X}_N$. Now, the least-squares finite element approximation problem is to minimize G over $\Sigma_h^{k,s} \times V_h^k$: find $(\boldsymbol{\sigma}_h,\, \mathbf{u}_h) \in \Sigma_h^{k,s} \times V_h^k$ such that

$$G(\boldsymbol{\sigma}_h,\, \mathbf{u}_h\,;\, \mathbf{f}) = \inf_{(\boldsymbol{\tau},\, \mathbf{v}) \in \Sigma_h^{k,s} \times V_h^k} G(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{f}). \tag{4.14}$$

It is easy to see that (4.14) has a unique solution $(\boldsymbol{\sigma}_h,\, \mathbf{u}_h)$, that $\boldsymbol{\sigma}_h$ is symmetric, and that $\boldsymbol{\sigma}_h$ has the following error bound:

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{H(\mathrm{div};\,\Omega)} + \|\mathbf{u} - \mathbf{u}_h\|_1 \le C\, h^l \left(\|\boldsymbol{\sigma}\|_{l+1} + \|\mathbf{u}\|_{l+1}\right) \tag{4.15}$$

if the solution, $(\boldsymbol{\sigma},\, \mathbf{u})$, of (3.9) is in $H^{l+1}(\Omega)^{2\times 2} \times H^{l+1}(\Omega)^2$. Note that this estimate is not optimal in the regularity of the displacement. Nevertheless, the nodal elements $\Sigma_h^{k,s}$ have many fewer local average degrees of freedom than the Raviart–Thomas elements Σ_h^k. Developing a better finite element space of the symmetric stress will be a topic of our further study.

5. Weakly imposed boundary conditions. In previous sections, boundary conditions are imposed in the solution space. This leads to least-squares finite element approximations that are much more accurate on the boundary than in the interior of the domain. In the context of the least-squares method, it is natural to treat boundary conditions weakly through boundary functionals. For many applications, this is also convenient. In this section, we study a least-squares functional with boundary terms minimized over a solution space free of boundary conditions. We focus on establishing continuity and ellipticity of this functional here. See [21] for the development of computable finite element approximations and the corresponding iterative solvers based on this functional.

Assume the following nonhomogeneous boundary conditions:

$$\mathbf{u} = \mathbf{g} \ \text{ on } \ \Gamma_D \quad \text{and} \quad \mathbf{n}\cdot\boldsymbol{\sigma} = \mathbf{h} \ \text{ on } \ \Gamma_N. \tag{5.1}$$

Let

$$\mathcal{V} = \mathbf{X} \times H^1(\Omega)^d,$$

and for $\mathbf{g} \in H^{1/2}(\Gamma_D)$ and $\mathbf{h} \in H^{-1/2}(\Gamma_N)$ define the least-squares functional as follows:

$$\begin{aligned}\tilde{G}(\boldsymbol{\sigma},\, \mathbf{u}; \mathbf{f}, \mathbf{g}, \mathbf{h}) = {} & \|\mathcal{A}\,\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla\cdot\boldsymbol{\sigma} + \mathbf{f}\|^2 \\ & + \|\mathbf{u} - \mathbf{g}\|^2_{\frac{1}{2},\,\Gamma_D} + \|\mathbf{n}\cdot\boldsymbol{\sigma} - \mathbf{h}\|^2_{-\frac{1}{2},\,\Gamma_N}\end{aligned} \tag{5.2}$$

for $(\boldsymbol{\sigma},\, \mathbf{u}) \in \mathcal{V}$. The least-squares variational problem is then to minimize $\tilde{G}$ over $\mathcal{V}$: find $(\boldsymbol{\sigma},\, \mathbf{u}) \in \mathcal{V}$ such that

$$\tilde{G}(\boldsymbol{\sigma},\, \mathbf{u}\,;\, \mathbf{f},\, \mathbf{g},\, \mathbf{h}) = \inf_{(\boldsymbol{\tau},\,\mathbf{v})\in\mathcal{V}} \tilde{G}(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{f},\, \mathbf{g},\, \mathbf{h}). \tag{5.3}$$

To establish the continuity and ellipticity of the homogeneous least-squares functional $\tilde{G}(\mathbf{u},\, \boldsymbol{\sigma}\,;\, \mathbf{0},\, \mathbf{0},\, \mathbf{0})$ in $\mathcal{V}$, we need the trace inequalities (see [15])

$$\begin{aligned}&\|u\|_{\frac{1}{2},\,\partial\Omega} \le \|u\|_1 && \forall\ u \in H^1(\Omega),\\ &\|\mathbf{n}\cdot\mathbf{v}\|_{-\frac{1}{2},\,\partial\Omega} \le \|\mathbf{v}\|_{H(\mathrm{div};\,\Omega)} && \forall\ \mathbf{v} \in H(\mathrm{div};\,\Omega)\end{aligned}$$

and the generalized Korn inequality

$$\|\mathbf{v}\|_1 \le C\,(\|\boldsymbol{\epsilon}(\mathbf{v})\| + \|\mathbf{v}\|_{0,\,\Gamma_D}) \quad \forall\ \mathbf{v} \in H^1(\Omega)^d. \tag{5.4}$$

THEOREM 5.1. *The homogeneous functional $\tilde{G}(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{0},\, \mathbf{0},\, \mathbf{0})$ is uniformly equivalent to the functional $M(\mathbf{v},\, \boldsymbol{\tau})$; i.e., there exists a positive constant C independent of λ such that*

$$\frac{1}{C}\, M(\boldsymbol{\tau},\, \mathbf{v}) \le \tilde{G}(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{0},\mathbf{0},\mathbf{0}) \le C\, M(\boldsymbol{\tau},\, \mathbf{v}) \tag{5.5}$$

holds for all $(\boldsymbol{\tau},\, \mathbf{v}) \in \mathcal{V}$.

Proof. The upper bound in (5.5) follows easily from the triangle inequality, (3.16), and trace inequalities. The proof of the lower bound in (5.5) is the same as that for Theorem 3.1 except the proof on the upper bound of $|(\boldsymbol{\tau},\, \boldsymbol{\epsilon}(\mathbf{v}))|$. This is because $\mathbf{v}$

and $\mathbf{n} \cdot \boldsymbol{\tau}$ do not satisfy any boundary conditions and our new functional has extra boundary terms. Therefore, it suffices to show that

$$|(\boldsymbol{\tau},\, \boldsymbol{\epsilon}(\mathbf{v}))| \le C\, \tilde{G}(\boldsymbol{\tau},\, \mathbf{v}\,; \mathbf{0}, \mathbf{0}, \mathbf{0}) + C\, \tilde{G}(\boldsymbol{\tau},\, \mathbf{v}\,; \mathbf{0}, \mathbf{0}, \mathbf{0})^{\frac{1}{2}}\, \|\boldsymbol{\tau}\|. \tag{5.6}$$

To this end, the triangle, Cauchy–Schwarz, and generalized Korn inequalities give

$$\begin{aligned}
\left| \int_{\partial\Omega} \mathbf{v} \cdot \big(\mathbf{n} \cdot \boldsymbol{\tau}\big)\, ds \right| &= \left| \int_{\Gamma_D} \mathbf{v} \cdot \big(\mathbf{n} \cdot \boldsymbol{\tau}\big)\, ds + \int_{\Gamma_N} \mathbf{v} \cdot \big(\mathbf{n} \cdot \boldsymbol{\tau}\big)\, ds \right| \\
&\le \|\mathbf{v}\|_{\frac{1}{2},\,\Gamma_D}\, \|\mathbf{n} \cdot \boldsymbol{\tau}\|_{-\frac{1}{2},\,\Gamma_D} + \|\mathbf{v}\|_{\frac{1}{2},\,\Gamma_N}\, \|\mathbf{n} \cdot \boldsymbol{\tau}\|_{-\frac{1}{2},\,\Gamma_N} \\
&\le \|\mathbf{v}\|_{\frac{1}{2},\,\Gamma_D}\, \|\boldsymbol{\tau}\|_{H(\mathrm{div};\,\Omega)} + \|\mathbf{v}\|_1\, \|\mathbf{n} \cdot \boldsymbol{\tau}\|_{-\frac{1}{2},\,\Gamma_N}.
\end{aligned}$$

Now, it follows from the symmetry of $\boldsymbol{\epsilon}(\mathbf{v})$; integration by parts; the triangle, Cauchy–Schwarz, and generalized Korn inequalities; and (3.6) that

$$\begin{aligned}
|(\boldsymbol{\tau},\, \boldsymbol{\epsilon}(\mathbf{v}))| &= \left| (\boldsymbol{\tau},\, \nabla\mathbf{v}) - \left(\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2},\, \nabla\mathbf{v} \right) \right| \\
&= \left| (\nabla \cdot \boldsymbol{\tau},\, \mathbf{v}) - \int_{\partial\Omega} \mathbf{v} \cdot \big(\mathbf{n} \cdot \boldsymbol{\tau}\big)\, ds + \left(\frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2},\, \nabla\mathbf{v} \right) \right| \\
&\le \left(\|\nabla \cdot \boldsymbol{\tau}\| + \left\| \frac{\boldsymbol{\tau} - \boldsymbol{\tau}^t}{2} \right\| \right) \|\mathbf{v}\|_1 + \|\mathbf{v}\|_{\frac{1}{2},\,\Gamma_D}\, \|\boldsymbol{\tau}\|_{H(\mathrm{div};\,\Omega)} + \|\mathbf{v}\|_1\, \|\mathbf{n} \cdot \boldsymbol{\tau}\|_{-\frac{1}{2},\,\Gamma_N} \\
&\le C \left(\|\nabla \cdot \boldsymbol{\tau}\| + \|\mathcal{A}\,\boldsymbol{\tau} - \boldsymbol{\epsilon}(\mathbf{v})\| + \|\mathbf{n} \cdot \boldsymbol{\tau}\|_{-\frac{1}{2},\,\Gamma_N} \right) \left(\|\boldsymbol{\epsilon}(\mathbf{v})\| + \|\boldsymbol{\tau}\| \right) + \|\mathbf{v}\|_{\frac{1}{2},\,\Gamma_D}\, \|\nabla \cdot \boldsymbol{\tau}\|,
\end{aligned}$$

which, together with (3.18), implies (5.6) and, hence, the theorem. ∎

6. Numerical results. In this section, numerical results for a benchmark problem of linear elasticity taken from [22] are presented. The problem to be considered is given by a quadratic membrane of elastic isotropic material with a circular hole in the center. Traction forces act on the upper and lower edges of the strip. Because of the symmetry of the domain, it suffices to discretize only a fourth of the total geometry. The computational domain is then given by

$$\Omega = \{\mathbf{x} \in \Re^2 : 0 < x_1 < 10, 0 < x_2 < 10, x_1^2 + x_2^2 > 1\}$$

(see Figure 6.1). The boundary conditions on the top edge of the computational domain ($x_2 = 10,\ 0 < x_1 < 10$) are set to $\boldsymbol{\sigma} \cdot \mathbf{n} = 4.5$, the boundary conditions on the bottom ($x_2 = 0,\ 1 < x_1 < 10$) are set to $(\sigma_{11}, \sigma_{12}) \cdot \mathbf{n} = 0,\ u_2 = 0$ (symmetry condition), and, finally, the boundary conditions on the left ($x_1 = 0,\ 1 < x_2 < 10$) are given by $u_1 = 0$, $(\sigma_{21}, \sigma_{22}) \cdot \mathbf{n} = 0$ (symmetry condition). The material parameters are $E = 206900$ for Young's modulus and $\nu = 0.29$ for Poisson's ratio, and their relation with the Lamé constants is given by

$$\lambda = \frac{E\,\nu}{(1+\nu)(1-2\nu)} \quad \text{and} \quad \mu = \frac{E}{2(1+\nu)}.$$

Obviously, the definition of the functional in (3.8) implies

$$G(\boldsymbol{\sigma}_h, \mathbf{u}_h; \mathbf{f}) = \sum_{K \in \mathcal{T}_h} \left(\|\mathcal{A}\,\boldsymbol{\sigma}_h - \boldsymbol{\epsilon}(\mathbf{u}_h)\|_{0,K}^2 + \|\nabla \cdot \boldsymbol{\sigma}_h + \mathbf{f}\|_{0,K}^2 \right) =: \sum_{K \in \mathcal{T}_h} G_K(\boldsymbol{\sigma}_h, \mathbf{u}_h; \mathbf{f})\,.$$

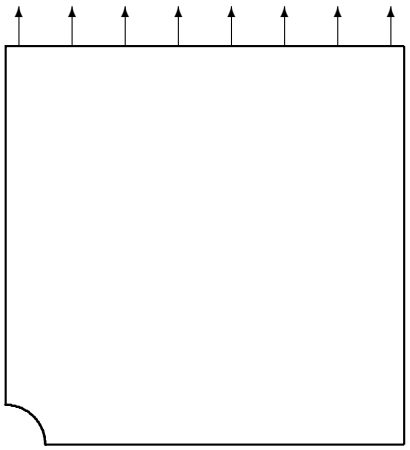


FIG. 6.1. *Computational domain and boundary conditions.*

TABLE 6.1
Adaptive finite element approximation ($k = 1$, $\nu = 0.29$).

	# elements	dim Σ_h^1	dim V_h^2	Functional	$(\sigma_h)_{22}(1,0)$
$l = 0$	52	504	224	2.56e-1	9.8830
$l = 1$	115	1130	480	3.78e-2	12.5226
$l = 2$	243	2400	1002	7.61e-2	13.4090
$l = 3$	511	5058	2096	1.62e-3	13.7213
$l = 4$	1069	10600	4366	4.82e-3	13.8259
$l = 5$	2164	21468	8828	1.21e-4	13.8630
$l = 6$	4384	43532	17844	3.51e-5	13.8771
$l = 7$	8678	86190	35302	9.20e-6	13.8912
$l = 8$	17152	170398	69730	2.59e-6	13.8884

Due to the equivalence (3.15), the local contributions $G_K(\boldsymbol{\sigma}_h, \mathbf{u}_h; \mathbf{f})$ to the least-squares functional constitute an a posteriori error estimator to be used for adaptive refinement (cf. [5]). The results in Table 6.1 are computed on a sequence of adaptively refined meshes based on this error estimator. In each refinement step those triangles with the largest values of $G_K(\boldsymbol{\sigma}_h, \mathbf{u}_h; \mathbf{f})$ (roughly 25 percent) were refined regularly (by dividing each into four congruent subtriangles). The Raviart–Thomas spaces of order one for the stress approximation are coupled with standard quadratic conforming elements for the displacement ($\Sigma_h^1 \times V_h^2$ in the terminology of section 4).

Table 6.1 provides a strong indication that the minimum of the functional is inversely proportional to the square of the number of degrees of freedom:

$$\mathcal{F}_h(\boldsymbol{\sigma}_h, \mathbf{u}_h) \sim \frac{1}{(\dim\, \Sigma_h^1 + \dim\, V_h^2)^2}\,.$$

This is the optimal asymptotic convergence rate achievable with piecewise quadratic finite elements. Of particular interest in this example is the stress component σ_{22} at the point $(1, 0)$. The size of this stress component is responsible for failure of the

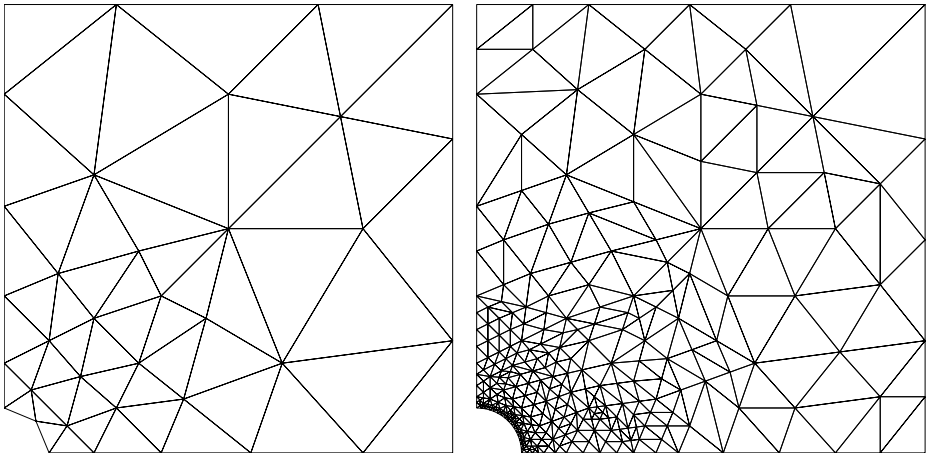


FIG. 6.2. *Initial triangulation and result after three adaptive refinement steps.*

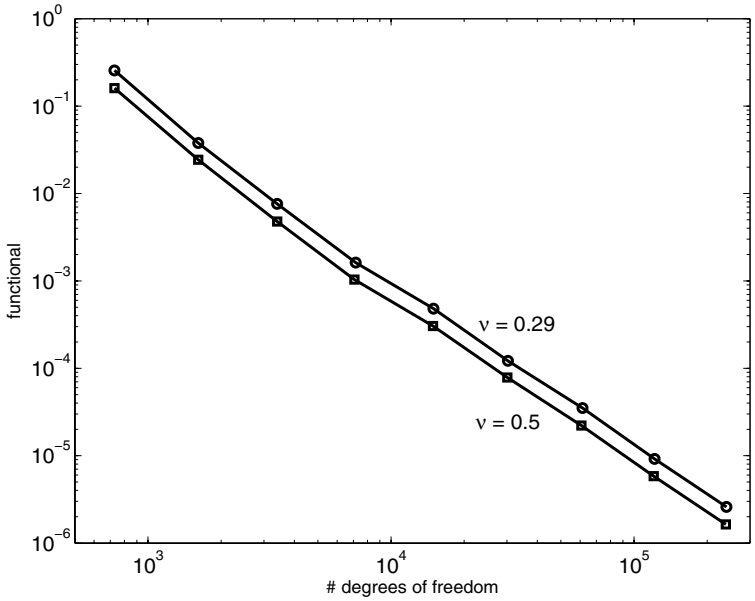


FIG. 6.3. *Adaptive finite element approximation for* $k = 1$ *(*$\nu = 0.29, 0.5$*).*

material at this point. For $\nu = 0.29$ the value of $\sigma_{22}(1,0) = 13.8873$ is given in [22] for a reference solution computed by a polynomial approximation of high degree. The corresponding column in Table 6.1 shows the convergence of the solutions obtained with our least-squares approach to that reference value as the mesh is refined. The initial triangulation and the result of three adaptive refinement steps are shown in Figure 6.2.

The robustness with respect to the incompressible limit can be seen in the doubly logarithmic convergence graphs in Figure 6.3. In addition to the numbers of Table 6.1, the results for the incompressible limit ($\nu = 0.5$) are shown in Figure 6.3.

REFERENCES

[1] D. N. Arnold, J. Douglas, and C. P. Gupta, *A family of higher order mixed finite element methods for plane elasticity*, Numer. Math., 45 (1984), pp. 1–22.

[2] D. N. Arnold, R. S. Falk, and R. Winther, *Multigrid in H(div) and H(curl)*, Numer. Math., 85 (2000), pp. 197–217.

[3] D. N. Arnold and R. Winther, *Mixed finite elements for elasticity*, Numer. Math., 42 (2002), pp. 401–419.

[4] D. N. Arnold and R. Winther, *Nonconforming mixed elements for elasticity*, Math. Models Methods Appl. Sci., 13 (2003), pp. 295–307.

[5] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Local error estimates and adaptive refinement for first-order system least squares*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43.

[6] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[7] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

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

[10] Z. Cai, J. Korsawe, and G. Starke, *A least squares mixed finite element method for the stress-displacement formulation of linear elasticity: Error estimation and adaptive refinement*, Numer. Methods Partial Differential Equations, to appear.

[11] Z. Cai, T. A. Manteuffel, S. F. McCormick, and S. V. Parter, *First-order system least squares (FOSLS) for planar linear elasticity: Pure traction problem*, SIAM J. Numer. Anal., 35 (1998), pp. 320–335.

[12] Z. Cai and G. Starke, *First-order system least squares for the stress-displacement formulation: Linear elasticity*, SIAM J. Numer. Anal., 41 (2003), pp. 715–730.

[13] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[14] P. G. Ciarlet, *Mathematical Elasticity Volume* I: *Three-Dimensional Elasticity*, North–Holland, Amsterdam, 1988.

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

[16] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer-Verlag, Berlin, 1985.

[17] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.

[18] B. Jiang, *The Least-Squares Finite Element Method*, Springer-Verlag, Berlin, 1998.

[19] B. Jiang and J. Wu, *The least-squares finite element method in elasticity—Part* I: *Plane stress or strain with drilling degrees of freedom*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 621–636.

[20] P.-A. Raviart and J.-M. Thomas, *A mixed finite element method for* 2*-nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, New York, 1977, pp. 292–315.

[21] G. Starke, *Multilevel boundary functionals for least-squares mixed finite element methods*, SIAM J. Numer. Anal., 36 (1999), pp. 1065–1077.

[22] E. Stein, P. Wriggers, A. Rieger, and M. Schmidt, *Benchmarks*, in Error-Controlled Adaptive Finite Elements in Solid Mechanics, E. Stein, ed., John Wiley and Sons, New York, 2002, pp. 385–404.

[23] P. S. Vassilevski and J. Wang, *Multilevel iterative methods for mixed finite element discretizations of elliptic problems*, Numer. Math., 63 (1992), pp. 503–520.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 843–859

LEAST-SQUARES METHODS FOR INCOMPRESSIBLE NEWTONIAN FLUID FLOW: LINEAR STATIONARY PROBLEMS*

ZHIQIANG CAI†, BARRY LEE‡, AND PING WANG§

Abstract. This paper develops and analyzes two least-squares methods for the numerical solution of linear, stationary incompressible Newtonian fluid flow in two and three dimensions. Both approaches use the L^2 norm to define least-squares functionals. One is based on the stress-velocity formulation (see section 3.2), and it applies to general boundary conditions. The other is based on an equivalent formulation for the *pseudostress* and velocity (see section 4.2), and it applies to pure velocity Dirichlet boundary conditions. The velocity gradient and vorticity can be obtained algebraically from this new tensor variable. It is shown that the homogeneous least-squares functionals are elliptic and continuous in the $H(\mathrm{div};\Omega)^d \times H^1(\Omega)^d$ norm. This immediately implies optimal error estimates for conforming finite element approximations. As well, it admits optimal multigrid solution methods if Raviart–Thomas finite element spaces are used to approximate the stress or the pseudostress tensor.

Key words. least-squares method, mixed finite element method, Navier–Stokes, Stokes, incompressible Newtonian flow

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/S0036142903422673

1. Introduction.

For incompressible Newtonian fluid flow with homogeneous density, the primitive physical equations are the conservation of momentum and the constitutive law. The constitutive law relates the stress tensor to the deformation rate tensor and pressure, and it states the incompressibility condition. It is a first-order partial differential system for the physical variables stress, velocity, and pressure. By differentiating and eliminating the stress, one obtains the well-known second-order incompressible Navier–Stokes equations in the velocity-pressure formulation. A tremendous amount of computational research has been done on this second-order partial differential system (see, e.g, mathematical books [17, 18]), but these equations may still be one of the most challenging problems in computational fluid mechanics and computational mathematics.

In recent years there has been substantial interest in the use of least-squares principles for the numerical approximation of Newtonian fluid flow problems (see, e.g., the survey paper [5], the monograph [21], and references therein). In particular, there are many research articles in both the mathematics and engineering communities on least-squares methods for the stationary Stokes equations (see [5]). Specifically, least-squares methods based on five first-order partial differential systems have been proposed, analyzed, implemented, and tested. These five first-order systems are formulations for variables (i) velocity, vorticity, and pressure [5, 21], (ii) velocity, pres-

*Received by the editors February 9, 2003; accepted for publication (in revised form) October 27, 2003; published electronically June 4, 2004.
http://www.siam.org/journals/sinum/42-2/42267.html

†Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067 (zcai@math.purdue.edu). The research of this author was sponsored in part by the National Science Foundation under grant INT-0139053 and by the Korea Research Foundation under grant KRF-2002-015-C00014.

‡Center for Applied Scientific Computing, Lawrence Livermore National Lab, Livermore, CA 94551 (lee123@llnl.gov).

§DNT Computing Applications Division, Lawrence Livermore National Lab, Livermore, CA 94551 (wang32@llnl.gov).

sure, and "stress" [4], (iii) velocity, velocity gradient, and pressure [11], (iv) velocity, velocity gradient, and pressure with additional constraints [11], and (v) constrained velocity gradient and pressure [15]. The new "stress" variable in (ii) is actually the deformation rate tensor and not the physical stress. Least-squares methods based on the first three formulations employ either discrete inverse norms (see [10, 7]) or mesh-weighted L^2 norms (see [3]) in order to achieve optimal finite element approximations. The inverse norm approach is very expensive due to its discrete inverse norm evaluations, and fast multigrid solvers are still a missing ingredient for the mesh-weighted L^2 norm approaches. If the original problem is sufficiently smooth, methods based on the last two formulations are equivalent to the H^1 norm. Such equivalence implies optimal finite element approximations and optimal convergence of multigrid solvers. But the smoothness requirement is restrictive.

A common feature of all these formulations is that they do not involve the primitive physical equations. Based on the velocity-pressure formulation of the Stokes equations, they are derived by introducing new variables such as vorticity in (i), "stress" in (ii), velocity gradient in (iii) and (iv), and constrained velocity gradient in (v). Some of these new variables have physical meanings, but they are not original physical quantities of interest.

The first objective of this paper is to develop a new least-squares method that does not have the above mentioned drawbacks and that computes the original physical quantities directly. For linear, stationary problems of incompressible Newtonian fluid flow, our least-squares method is based directly on the primitive first-order partial differential system: the stress-velocity-pressure formulation, without introducing any new variables nor any new equations. We define the least-squares functional by applying a L^2 norm least-squares principle to this first-order system. It is shown that the homogeneous least-squares functional is elliptic and continuous in the $H(\mathrm{div};\,\Omega)^d$ norm for the stress, the $H^1(\Omega)^d$ norm for the velocity, and the L^2 norm for the pressure. This immediately implies optimal error estimates for conforming finite element approximations in $H(\mathrm{div};\,\Omega)^d \times H^1(\Omega)^d \times L^2(\Omega)$. It also admits optimal multigrid solution methods if Raviart–Thomas finite element spaces are used to approximate the stress tensor. Both discretization accuracy and multigrid convergence rates are uniform in the viscosity parameter.

Since the pressure can be represented in terms of the normal stress and since the stress is an independent variable in the first-order system, the pressure can be eliminated from the first-order system. By replacing the pressure with the normal stress, we derive the stress-velocity formulation for incompressible Newtonian fluid flow. We can then define the corresponding least-squares method and show identical numerical properties to those of the stress-velocity-pressure formulation, since the stress-velocity formulation is a special case of the stress-velocity-pressure formulation. It is important to note that, mathematically, the stress-velocity formulation for linear, stationary problems of incompressible Newtonian fluid flow is the limiting case of the stress-displacement formulation for elastic problems when $2\mu = \nu$. This indicates that this paper, together with [14], develops a unified least-squares approach for both elastic solids and incompressible Newtonian fluids with respect to spatial discretization and fast solution solvers, even though the variables and materials have different physical meanings. Hence, our method can be extended to problems coupling elastic deformation with fluid flow.

Many applications in incompressible Newtonian fluid flow do not have traction boundary conditions. It is then not necessary to use the stress as an independent

variable. This is especially true because the stress does not contain any information on the vorticity that is a physical quantity of great interest in fluid mechanics. Thus, for pure velocity Dirichlet boundary conditions, we define a new independent variable, *pseudostress*, in terms of the velocity gradient and pressure, and then derive an equivalent first-order system containing the pseudostress and velocity. The pressure, the velocity gradient, and, hence, the vorticity are expressed in terms of the pseudostress. The L^2 norm least-squares functional based on this first-order system is again shown to be elliptic and continuous in the $H(\mathrm{div};\,\Omega)^d \times H^1(\Omega)^d$ norm. Hence, Raviart–Thomas finite elements for the pseudostress and standard continuous piecewise polynomials for the velocity yield optimal approximation, and the resulting algebraic equations can be solved with optimal multigrid methods.

For completeness, we also study inverse norm least-squares functionals and show that their homogeneous forms are elliptic and continuous in appropriate Hilbert spaces. These functionals can be used to develop discrete inverse norm least-squares methods (see, e.g., [6]). Also, for many applications, it is convenient to impose boundary conditions weakly through boundary functionals. Such functionals are also studied in this paper (see section 4.5). (See [23] for the computational feasibility of methods based on these types of functionals.)

Least-squares methods developed in this paper for linear, stationary problems can be easily extended to nonlinear incompressible Newtonian flows, at least in principle. One can simply include an appropriate form of the nonlinear convection term in the residual of the momentum equations. Possible choices for this form can (1) involve only the velocity or (2) involve the (pseudo-) stress which replaces the velocity gradient. Mathematical analysis for least-squares methods applied to nonlinear problems is much more difficult, but it still can be established using the abstract theory of [9]. Formulations of our methods can be easily extended to incompressible non-Newtonian flows as well: only a simple modification is needed in the constitutive equation.

An outline of the paper is as follows. In section 2, the stress-velocity-pressure formulation for incompressible Newtonian fluid flow problems and the corresponding linear, stationary problems are introduced, as well as some notation and the Stokes equations. In section 3, least-squares functionals based on the stress-velocity-pressure and stress-velocity formulations are developed, their ellipticity and continuity are established, and finite element approximations and multigrid solvers are discussed. In section 4, least-squares methods for pure Dirichlet boundary conditions are developed.

1.1. Notation. We use the standard notation and definitions for the Sobolev spaces $H^s(\Omega)^d$ and $H^s(\partial\Omega)^d$ for $s \geq 0$. The standard associated inner products are denoted by $(\cdot,\,\cdot)_{s,\Omega}$ and $(\cdot,\,\cdot)_{s,\partial\Omega}$, and their respective norms are denoted by $\|\cdot\|_{s,\Omega}$ and $\|\cdot\|_{s,\partial\Omega}$. (We suppress the superscript d because the dependence on dimension will be clear by context. We also omit the subscript Ω from the inner product and norm designation when there is no risk of confusion.) For $s = 0$, $H^s(\Omega)^d$ coincides with $L^2(\Omega)^d$. In this case, the inner product and norm will be denoted by $\|\cdot\|$ and $(\cdot,\,\cdot)$, respectively. Set $H^1_D(\Omega) := \{q \in H^1(\Omega) \,:\, q = 0 \text{ on } \Gamma_D\}$. We denote the duals of $H^1_D(\Omega)$ and $H^{\frac{1}{2}}(\partial\Omega)$ by $H_D^{-1}(\Omega)$ and $H^{-\frac{1}{2}}(\partial\Omega)$ with norms defined by

$$\|\phi\|_{-1,\,D} = \sup_{0\neq\psi\in H^1_D(\Omega)} \frac{(\phi,\,\psi)}{\|\psi\|_1} \quad \text{and} \quad \|\phi\|_{-1/2,\,\partial\Omega} = \sup_{0\neq\psi\in H^{\frac{1}{2}}(\partial\Omega)} \frac{(\phi,\,\psi)}{\|\psi\|_{1/2,\partial\Omega}}.$$

When $D = \partial\Omega$, we denote the dual of $H^1_0(\Omega) = H^1_D(\Omega)$ and its norm by $H_0^{-1}(\Omega)$ and $\|\cdot\|_{-1,\,0}$, respectively. When D is empty, the dual of $H^1(\Omega)$ and its norm are

denoted by the respective $H^{-1}(\Omega)$ and $\|\cdot\|_{-1}$. Also, we denote the product space $\prod_{i=1}^{d} H_D^{-1}(\Omega)$ with the standard product norm by $H_D^{-1}(\Omega)^d$. Finally, set

$$H(\mathrm{div};\Omega) = \{\mathbf{v} \in L^2(\Omega)^2 \,:\, \nabla\cdot\mathbf{v} \in L^2(\Omega)\},$$

which is a Hilbert space under the norm

$$\|\mathbf{v}\|_{H(\mathrm{div};\,\Omega)} = \left(\|\mathbf{v}\|^2 + \|\nabla\cdot\mathbf{v}\|^2\right)^{\frac{1}{2}},$$

and define the subspace

$$H_N(\mathrm{div};\Omega) = \{\mathbf{v} \in H(\mathrm{div};\Omega) \,:\, \mathbf{n}\cdot\mathbf{v} = 0\}.$$

2. Mathematical equations for incompressible Newtonian fluid flow. Let Ω be a bounded, open, connected subset of $\Re^d$ ($d = 2$ or 3) with a Lipschitz continuous boundary $\partial\Omega$. Denote the outward unit vector normal to the boundary by $\mathbf{n} = (n_1, \ldots, n_d)^t$. We partition the boundary of Ω into two open subsets Γ_D and Γ_N such that $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. For simplicity, we will assume that Γ_D is not empty (i.e., $\mathrm{mes}\,(\Gamma_D) \neq 0$).

For a second-order tensor $\boldsymbol{\tau} = (\tau_{ij})_{d\times d}$, define its divergence and normal by

$$\nabla\cdot\boldsymbol{\tau} = \begin{pmatrix} \partial\tau_{11}/\partial x_1 + \cdots + \partial\tau_{1d}/\partial x_d \\ \vdots \\ \partial\tau_{d1}/\partial x_1 + \cdots + \partial\tau_{dd}/\partial x_d \end{pmatrix} \quad \text{and} \quad \mathbf{n}\cdot\boldsymbol{\tau} = \begin{pmatrix} n_1\tau_{11} + \cdots + n_d\tau_{1d} \\ \vdots \\ n_1\tau_{d1} + \cdots + n_d\tau_{dd} \end{pmatrix},$$

respectively. That is, the divergence and normal operators apply to each row of the tensor. Also denote the matrix trace operator by tr:

$$\mathrm{tr}\,\boldsymbol{\tau} = \tau_{11} + \cdots + \tau_{dd}.$$

Let $\mathbf{f} = (f_1, \ldots, f_d)^t$ be a given external body force defined in Ω and $\mathbf{g} = (g_1, \ldots, g_d)^t$ be a given external surface traction applied on Γ_N. Let $\mathbf{u}(\mathbf{x}, t) = (u_1, \ldots, u_d)^t$ be the velocity vector field of a particle of fluid that is moving through $\mathbf{x}$ at time t, and let $\boldsymbol{\sigma} = (\sigma_{ij})_{d\times d}$ be the stress tensor field. Without loss of generality, we assume that the homogeneous density is one. Then conservation of momentum implies both symmetry of the stress tensor and the local relation

$$\left\{ \begin{array}{rcll} \frac{D\mathbf{u}}{Dt} - \nabla\cdot\boldsymbol{\sigma} &=& \mathbf{f} & \text{in} \quad \Omega, \\ \mathbf{n}\cdot\boldsymbol{\sigma} &=& \mathbf{g} & \text{on} \quad \Gamma_N, \end{array} \right. \tag{2.1}$$

where $\frac{D}{Dt}$ is the material derivative

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla = \frac{\partial}{\partial t} + \sum_{i=1}^{d} u_i \frac{\partial}{\partial x_i}.$$

In this paper, we restrict ourselves to linear, stationary problems, i.e., problems where the momentum equation in (2.1) is of the form

$$-\nabla\cdot\boldsymbol{\sigma} = \mathbf{f}. \tag{2.2}$$

Let ν be the viscosity parameter, p the pressure, and

$$\boldsymbol{\epsilon}(\mathbf{u}) = \frac{1}{2}\left(\nabla\,\mathbf{u} + (\nabla\,\mathbf{u})^t\right)$$

the deformation rate tensor, where $\nabla \mathbf{u}$ is the velocity gradient tensor with entries $(\nabla \mathbf{u})_{ij} = \partial u_i / \partial x_j$. Then the constitutive law for incompressible Newtonian fluids is

$$\begin{cases} \boldsymbol{\sigma} &= \nu\, \boldsymbol{\epsilon}(\mathbf{u}) - p\, I & \text{in} \quad \Omega, \\ \nabla \cdot \mathbf{u} &= 0 & \text{in} \quad \Omega. \end{cases} \tag{2.3}$$

The second equation in (2.3) is the incompressibility condition. Without loss of generality, we assume that $\nu = 1$, since otherwise $\mathbf{u}$ can be rescaled to $\nu\, \mathbf{u}$. Now, combining (2.2) and (2.3), we have the stress-velocity-pressure formulation for incompressible Newtonian fluid flow:

$$\begin{cases} -\nabla \cdot \boldsymbol{\sigma} &= \mathbf{f} & \text{in} \quad \Omega, \\ \boldsymbol{\sigma} + p\, I - \boldsymbol{\epsilon}(\mathbf{u}) &= \mathbf{0} & \text{in} \quad \Omega, \\ \nabla \cdot \mathbf{u} &= 0 & \text{in} \quad \Omega. \end{cases} \tag{2.4}$$

Differentiating and eliminating the stress in the above system leads to the well-known incompressible Stokes equations:

$$\begin{cases} -\nabla \cdot \boldsymbol{\epsilon}(\mathbf{u}) + \nabla p &= \mathbf{f} & \text{in} \quad \Omega, \\ \nabla \cdot \mathbf{u} &= 0 & \text{in} \quad \Omega. \end{cases} \tag{2.5}$$

3. General boundary conditions. For simplicity, we assume that the boundary conditions are homogeneous:

$$\mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{n} \cdot \boldsymbol{\sigma} = \mathbf{0} \quad \text{on } \Gamma_N. \tag{3.1}$$

When Γ_N is nonempty, because of the traction boundary condition, it is natural and necessary to have the stress be the independent variable. Hence, we study least-squares functionals based on formulations for stress-velocity-pressure (section 3.1) and for stress-velocity (section 3.2). Our primary goal in this section is to establish continuity and ellipticity for these least-squares functionals in appropriate Hilbert spaces. The least-squares finite element method based on the stress-velocity formulation is described in section 3.3.

3.1. Least-squares functionals based on the stress-velocity-pressure formulation. The first-order system (2.4), together with boundary conditions (3.1), is the stress-velocity-pressure formulation for linear, stationary incompressible Newtonian flow. Taking the trace of the second equation in (2.4) and using the fact that

$$\operatorname{tr} \boldsymbol{\epsilon}(\mathbf{u}) = \nabla \cdot \mathbf{u} = 0,$$

we have the following important relation between the pressure and normal stress:

$$\operatorname{tr} \boldsymbol{\sigma} + d\, p = 0. \tag{3.2}$$

Before defining least-squares functionals, let us first describe solution spaces. When $\Gamma_D = \partial\Omega$, Stokes system (2.5) and (3.1) have a unique solution, provided that

$$\int_\Omega p\, dx = 0. \tag{3.3}$$

Together with (3.2), this implies

$$\int_\Omega \mathrm{tr}\boldsymbol{\sigma}\, dx = 0.$$

Therefore, we are at liberty to impose these conditions on the stress and pressure. Thus, define the spaces

$$\mathbf{X}_N = \begin{cases} H_N(\mathrm{div};\,\Omega)^d & \text{if } \Gamma_N \neq \emptyset, \\ \mathbf{X}_0 \equiv \left\{ \boldsymbol{\tau} \in H(\mathrm{div};\,\Omega)^d \mid \displaystyle\int_\Omega \mathrm{tr}\,\boldsymbol{\tau}\, dx = 0 \right\} & \text{otherwise} \end{cases}$$

and

$$L^2_N(\Omega) = \begin{cases} L^2(\Omega) & \text{if } \Gamma_N \neq \emptyset, \\ L^2_0(\Omega) = \left\{ q \in L^2(\Omega) \mid \displaystyle\int_\Omega q\, dx = 0 \right\} & \text{otherwise.} \end{cases}$$

Then for $\mathbf{f} \in L^2(\Omega)^d$ we define the following least-squares functionals:

$$G_{-1}(\boldsymbol{\sigma},\, \mathbf{u},\, p\,;\, \mathbf{f}) = \|\nabla\cdot\boldsymbol{\sigma} + \mathbf{f}\|^2_{-1,D} + \|\boldsymbol{\sigma} + p\,I - \boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla\cdot\mathbf{u}\|^2 \tag{3.4}$$

and

$$G(\boldsymbol{\sigma},\, \mathbf{u},\, p\,;\, \mathbf{f}) = \|\nabla\cdot\boldsymbol{\sigma} + \mathbf{f}\|^2 + \|\boldsymbol{\sigma} + p\,I - \boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla\cdot\mathbf{u}\|^2 \tag{3.5}$$

for $(\boldsymbol{\sigma},\, \mathbf{u},\, p) \in \mathcal{V} \equiv \mathbf{X}_N \times H^1_D(\Omega)^d \times L^2_N(\Omega)$. We will first establish uniform boundedness and ellipticity (i.e., equivalence) of the homogeneous functionals $G_{-1}(\boldsymbol{\tau},\, \mathbf{v},\, q\,;\, \mathbf{0})$ and $G(\boldsymbol{\tau},\, \mathbf{v},\, q\,;\, \mathbf{0})$ in terms of the respective functionals $M_{-1}(\boldsymbol{\tau},\, \mathbf{v},\, q)$ and $M(\boldsymbol{\tau},\, \mathbf{v},\, q)$ defined on $\mathcal{V}$ by

$$M_{-1}(\boldsymbol{\tau},\, \mathbf{v},\, q) = \|\mathbf{v}\|^2_1 + \|q\|^2 + \|\boldsymbol{\tau}\|^2$$

and

$$M(\boldsymbol{\tau},\, \mathbf{v},\, q) = \|\mathbf{v}\|^2_1 + \|q\|^2 + \|\boldsymbol{\tau}\|^2 + \|\nabla\cdot\boldsymbol{\tau}\|^2.$$

To accomplish this, let $\mathcal{A}_\lambda : R^{d\times d} \longrightarrow R^{d\times d}$ be a linear map defined by

$$\mathcal{A}_\lambda\,\boldsymbol{\tau} = \boldsymbol{\tau} - \frac{\lambda}{d\lambda + 2\mu}(\mathrm{tr}\,\boldsymbol{\tau})\,I \quad \forall\ \boldsymbol{\tau} \in R^{d\times d}.$$

The $\mathcal{A}_\lambda$ is the compliance tensor of fourth order, a terminology from elasticity. Parameters λ and μ are material constants for both solids and fluids. We will use the following fundamental inequality for the trace of $\mathbf{X}_N$:

$$\|\mathrm{tr}\,\boldsymbol{\tau}\| \le C\left(\sqrt{(\mathcal{A}_\lambda\,\boldsymbol{\tau},\, \boldsymbol{\tau})} + \|\nabla\cdot\boldsymbol{\tau}\|_{-1,D}\right) \quad \forall\ \boldsymbol{\tau} \in \mathbf{X}_N, \tag{3.6}$$

where C is a positive constant independent of λ. This inequality was proved in [1] for two dimensions and Dirichlet boundary conditions (i.e., $d = 2$ and $\Gamma_N = \emptyset$) and in [13] for both two and three dimensions and general boundary conditions. When λ approaches ∞, the limit of the linear map $\mathcal{A}_\lambda$ is

$$\mathcal{A}_\infty\,\boldsymbol{\tau} = \boldsymbol{\tau} - \frac{1}{d}\,(\mathrm{tr}\,\boldsymbol{\tau})\,I \ :\ R^{d\times d} \longrightarrow R^{d\times d}.$$

Note that $\mathcal{A}_\infty$ is not an invertible map. A simple calculation gives

$$\begin{cases} (\mathcal{A}_\infty\,\boldsymbol{\tau},\,\boldsymbol{\tau}) = \|\boldsymbol{\tau}\|^2 - \frac{1}{d}\,\|\mathrm{tr}\,\boldsymbol{\tau}\|^2 = \|\mathcal{A}_\infty\,\boldsymbol{\tau}\|^2, \\ (\mathcal{A}_\lambda\,\boldsymbol{\tau},\,\boldsymbol{\tau}) = \|\boldsymbol{\tau}\|^2 - \frac{\lambda}{d\lambda+2\mu}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2 = \|\mathcal{A}_\infty\,\boldsymbol{\tau}\|^2 + \frac{2\mu}{d(d\lambda+2\mu)}\|\mathrm{tr}\,\boldsymbol{\tau}\|^2. \end{cases} \tag{3.7}$$

Since the constant in (3.6) is independent of λ, taking the limit of (3.6) as $\lambda \to \infty$ and using the first equation in (3.7) we obtain

$$\|\mathrm{tr}\,\boldsymbol{\tau}\| \le C\,(\|\mathcal{A}_\infty\,\boldsymbol{\tau}\| + \|\nabla\cdot\boldsymbol{\tau}\|_{-1,D}) \quad \forall\ \boldsymbol{\tau} \in \mathbf{X}_N. \tag{3.8}$$

Let $\|\boldsymbol{\tau}\|_a \equiv \left(\|\mathcal{A}_\infty\,\boldsymbol{\tau}\|^2 + \|\nabla\cdot\boldsymbol{\tau}\|^2_{-1,\,D}\right)^{\frac{1}{2}}$; then $\|\boldsymbol{\tau}\|_a$ is equivalent to the L^2 norm.

LEMMA 3.1. *There exists a positive constant C such that*

$$\frac{1}{C}\|\boldsymbol{\tau}\|^2 \le \|\boldsymbol{\tau}\|_a^2 \le C\,\|\boldsymbol{\tau}\|^2 \quad \forall\ \boldsymbol{\tau} \in \mathbf{X}_N. \tag{3.9}$$

Proof. From the definition of the inverse norm and the Cauchy–Schwarz inequality, we have that

$$\|\nabla\cdot\boldsymbol{\tau}\|_{-1,\,D} \le \|\boldsymbol{\tau}\|. \tag{3.10}$$

Equation (3.9) follows easily from (3.7), (3.8), and (3.10). ∎

Now we are ready to establish equivalence between functionals G_{-1} and M_{-1} and equivalence between functionals G and M.

THEOREM 3.2. *The homogeneous functionals $G_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\,\mathbf{0})$ and $G(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\,\mathbf{0})$ are uniformly equivalent to the functionals $M_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q)$ and $M(\boldsymbol{\tau},\,\mathbf{v},\,q)$, respectively; i.e., there exist positive constants C_1 and C_2 such that*

$$\frac{1}{C_1}M_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q) \le G_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\,\mathbf{0}) \le C_1 M_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q) \tag{3.11}$$

and

$$\frac{1}{C_2}M(\boldsymbol{\tau},\,\mathbf{v},\,q) \le G(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\,\mathbf{0}) \le C_2 M(\boldsymbol{\tau},\,\mathbf{v},\,q) \tag{3.12}$$

hold for all $(\boldsymbol{\tau},\,\mathbf{v},\,q) \in \mathcal{V}$.

Proof. The upper bounds in both (3.11) and (3.12) follow easily from the triangle inequality and (3.10).

To show the validity of the lower bound in (3.11), we first note that

$$\begin{aligned} \|\boldsymbol{\tau} - \boldsymbol{\tau}^t\| &= \|\,(\boldsymbol{\tau} + q\,I - \boldsymbol{\epsilon}(\mathbf{v})) - (\boldsymbol{\tau} + q\,I - \boldsymbol{\epsilon}(\mathbf{v}))^t\,\| \\ &\le 2\,\|\boldsymbol{\tau} + q\,I - \boldsymbol{\epsilon}(\mathbf{v})\|. \end{aligned}$$

We used symmetry of I and $\boldsymbol{\epsilon}(\mathbf{v})$ and the triangle inequality above. Now integration by parts and the Cauchy–Schwarz and Korn inequalities lead to

$$\begin{aligned} |(\boldsymbol{\tau},\,\boldsymbol{\epsilon}(\mathbf{v}))| &= \left|\left(\frac{\boldsymbol{\tau}+\boldsymbol{\tau}^t}{2},\,\boldsymbol{\epsilon}(\mathbf{v})\right)\right| = \left|\left(\frac{\boldsymbol{\tau}+\boldsymbol{\tau}^t}{2},\,\nabla\,\mathbf{v}\right)\right| \\ &= \left|(\boldsymbol{\tau},\,\nabla\mathbf{v}) - \left(\frac{\boldsymbol{\tau}-\boldsymbol{\tau}^t}{2},\,\nabla\mathbf{v}\right)\right| = \left|(-\nabla\cdot\boldsymbol{\tau},\,\mathbf{v}) - \left(\frac{\boldsymbol{\tau}-\boldsymbol{\tau}^t}{2},\,\nabla\mathbf{v}\right)\right| \\ &\le \|\nabla\cdot\boldsymbol{\tau}\|_{-1,\,D}\,\|\mathbf{v}\|_1 + \|\boldsymbol{\tau} + q\,I - \boldsymbol{\epsilon}(\mathbf{v})\|\,\|\nabla\,\mathbf{v}\| \\ &\le C\,(\|\nabla\cdot\boldsymbol{\tau}\|_{-1,\,D} + \|\boldsymbol{\tau} + q\,I - \boldsymbol{\epsilon}(\mathbf{v})\|)\,\|\boldsymbol{\epsilon}(\mathbf{v})\| \\ &\le C\,G_{-1}^{\frac{1}{2}}(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\mathbf{0})\,\|\boldsymbol{\epsilon}(\mathbf{v})\|, \end{aligned} \tag{3.13}$$

where $G_{-1}^{\frac{1}{2}}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0})$ denotes the square root of $G_{-1}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0})$. To bound the deformation rate tensor $\boldsymbol{\epsilon}(\mathbf{v})$, it follows from the fact that

$$(q\, I, \boldsymbol{\epsilon}(\mathbf{v})) = (q, \nabla \cdot \mathbf{v}),$$

the Cauchy–Schwarz inequality, and (3.13) that

$$\begin{aligned}
\|\boldsymbol{\epsilon}(\mathbf{v})\|^2 &= (\boldsymbol{\epsilon}(\mathbf{v}) - \boldsymbol{\tau} - q\, I, \boldsymbol{\epsilon}(\mathbf{v})) + (\boldsymbol{\tau}, \boldsymbol{\epsilon}(\mathbf{v})) + (q, \nabla \cdot \mathbf{v}) \\
&\le \|\boldsymbol{\epsilon}(\mathbf{v}) - \boldsymbol{\tau} - q\, I\|\, \|\boldsymbol{\epsilon}(\mathbf{v})\| + C\, G_{-1}^{\frac{1}{2}}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0})\, \|\boldsymbol{\epsilon}(\mathbf{v})\| + \|q\|\, \|\nabla \cdot \mathbf{v}\| \\
&\le \|\boldsymbol{\epsilon}(\mathbf{v}) - \boldsymbol{\tau} - q\, I\|^2 + C\, G_{-1}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + \frac{1}{2}\, \|\boldsymbol{\epsilon}(\mathbf{v})\|^2 + \|q\|\, \|\nabla \cdot \mathbf{v}\|.
\end{aligned}$$

This implies that

$$\|\boldsymbol{\epsilon}(\mathbf{v})\|^2 \le C\, G_{-1}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + 2\, \|q\|\, \|\nabla \cdot \mathbf{v}\|. \tag{3.14}$$

Now to bound $\|q\|$ in (3.14), since $\operatorname{tr}\, (\boldsymbol{\tau} + q\, I - \boldsymbol{\epsilon}(\mathbf{v})) = \operatorname{tr} \boldsymbol{\tau} + d\, q - \nabla \cdot \mathbf{v}$, we have

$$\|\operatorname{tr} \boldsymbol{\tau} + d\, q - \nabla \cdot \mathbf{v}\| \le d\, \|\boldsymbol{\tau} + q\, I - \boldsymbol{\epsilon}(\mathbf{v})\|.$$

It then follows from the triangle inequality that

$$\begin{aligned}
\|q\| &\le \frac{1}{d}\, (\|\operatorname{tr} \boldsymbol{\tau} + d\, q - \nabla \cdot \mathbf{v}\| + \|\operatorname{tr} \boldsymbol{\tau}\| + \|\nabla \cdot \mathbf{v}\|) \\
&\le d\, G_{-1}^{\frac{1}{2}}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + \|\operatorname{tr} \boldsymbol{\tau}\|. \qquad (3.15)
\end{aligned}$$

Next, we bound $\|\operatorname{tr} \boldsymbol{\tau}\|$ above by the homogeneous functional and the L^2 norm of the deformation rate tensor. To do so, we first establish a similar upper bound for $\|\mathcal{A}_\infty \boldsymbol{\tau}\|$. Note that $\mathcal{A}_\infty^2 = \mathcal{A}_\infty$ and that $(q\, I, \mathcal{A}_\infty \boldsymbol{\tau}) = 0$. These identities and the Cauchy–Schwarz inequality lead to

$$\begin{aligned}
\|\mathcal{A}_\infty \boldsymbol{\tau}\|^2 &= (\boldsymbol{\tau}, \mathcal{A}_\infty \boldsymbol{\tau}) = (\boldsymbol{\tau} + q\, I - \boldsymbol{\epsilon}(\mathbf{v}), \mathcal{A}_\infty \boldsymbol{\tau}) + (\boldsymbol{\epsilon}(\mathbf{v}), \mathcal{A}_\infty \boldsymbol{\tau}) \\
&\le (\|\boldsymbol{\tau} + q\, I - \boldsymbol{\epsilon}(\mathbf{v})\| + \|\boldsymbol{\epsilon}(\mathbf{v})\|)\, \|\mathcal{A}_\infty \boldsymbol{\tau}\|,
\end{aligned}$$

which implies that

$$\|\mathcal{A}_\infty \boldsymbol{\tau}\| \le \|\boldsymbol{\tau} + q\, I - \boldsymbol{\epsilon}(\mathbf{v})\| + \|\boldsymbol{\epsilon}(\mathbf{v})\|. \tag{3.16}$$

Together with (3.9), inequality (3.16) yields

$$\begin{aligned}
\|\operatorname{tr} \boldsymbol{\tau}\| \le \|\boldsymbol{\tau}\| &\le C\, (\|\mathcal{A}_\infty \boldsymbol{\tau}\| + \|\nabla \cdot \boldsymbol{\tau}\|_{-1,D}) \\
&\le C\, (\|\boldsymbol{\tau} + q\, I - \boldsymbol{\epsilon}(\mathbf{v})\| + \|\boldsymbol{\epsilon}(\mathbf{v})\| + \|\nabla \cdot \boldsymbol{\tau}\|_{-1,D}) \\
&\le C\, \left(G_{-1}^{\frac{1}{2}}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + \|\boldsymbol{\epsilon}(\mathbf{v})\|\right). \qquad (3.17)
\end{aligned}$$

Now, combining upper bounds in (3.14), (3.15), and (3.17) and using the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\|\boldsymbol{\epsilon}(\mathbf{v})\|^2 &\le C\, G_{-1}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + \left(d\, G_{-1}^{\frac{1}{2}}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + \|\operatorname{tr} \boldsymbol{\tau}\|\right)\, \|\nabla \cdot \mathbf{v}\| \\
&\le C\, G_{-1}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + C\, \left(G_{-1}^{\frac{1}{2}}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + \|\boldsymbol{\epsilon}(\mathbf{v})\|\right)\, \|\nabla \cdot \mathbf{v}\| \\
&\le C\, G_{-1}(\boldsymbol{\tau}, \mathbf{v}, q; \mathbf{0}) + C\, \|\boldsymbol{\epsilon}(\mathbf{v})\|\, \|\nabla \cdot \mathbf{v}\|.
\end{aligned}$$

Hence,

$$\|\boldsymbol{\epsilon}(\mathbf{v})\|^2 \le C\, G_{-1}(\boldsymbol{\tau},\, \mathbf{v},\, q\,;\mathbf{0}),$$

which, together with (3.17), (3.15), and (3.9), implies that both $\|\boldsymbol{\tau}\|^2$ and $\|q\|^2$ are also bounded above by the homogeneous functional $G_{-1}(\boldsymbol{\tau},\, \mathbf{v},\, q\,;\mathbf{0})$. This completes the proof of the lower bound in (3.11). Since

$$G_{-1}(\boldsymbol{\tau},\, \mathbf{v},\, q\,;\, \mathbf{0}) \le G(\boldsymbol{\tau},\, \mathbf{v},\, q\,;\, \mathbf{0}) \quad \text{and} \quad \|\nabla\cdot\boldsymbol{\tau}\|^2 \le G(\boldsymbol{\tau},\, \mathbf{v},\, q\,;\, \mathbf{0}),$$

then the lower bound in (3.12) follows from (3.11). The proof of the theorem is therefore completed. ∎

3.2. Least-squares functionals based on the stress-velocity formulation. In this section, we derive the stress-velocity formulation by using relation (3.2) to eliminate the pressure. We then define least-squares functionals based on this formulation and establish their ellipticity and continuity.

Assume that the first equation in (2.3) holds. Then it is easy to see that (3.2) is equivalent to the incompressible condition, the second equation in (2.3). Relation (3.2) says that the pressure is the negative of the arithmetic average of the normal stress. Since the stress is a variable in our first-order system, using (3.2) we eliminate the pressure in the first equation of (2.3) to obtain the following constitutive equation:

$$\mathcal{A}_\infty\,\boldsymbol{\sigma} = \boldsymbol{\sigma} - \frac{1}{d}(\operatorname{tr}\boldsymbol{\sigma})\, I = \boldsymbol{\epsilon}(\mathbf{u}) \quad \text{in} \quad \Omega. \tag{3.18}$$

Note that taking the trace of this equation yields the incompressible condition. This and the momentum equation define the stress-velocity formulation for incompressible Newtonian fluid flow problems. In particular, for linear stationary problems, we have

$$\begin{cases} \mathcal{A}_\infty\,\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u}) &=& \mathbf{0} & \text{in} \quad \Omega, \\ \nabla\cdot\boldsymbol{\sigma} + \mathbf{f} &=& \mathbf{0} & \text{in} \quad \Omega, \end{cases} \tag{3.19}$$

with boundary conditions (3.1). Let

$$\tilde{\mathcal{V}} = \mathbf{X}_N \times H^1_D(\Omega)^d.$$

For $\mathbf{f} \in L^2(\Omega)^d$, we define the following least-squares functionals:

$$\tilde{G}_{-1}(\boldsymbol{\sigma},\, \mathbf{u}\,;\, \mathbf{f}) = \|\mathcal{A}_\infty\,\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla\cdot\boldsymbol{\sigma} + \mathbf{f}\|^2_{-1,D} \tag{3.20}$$

and

$$\tilde{G}(\boldsymbol{\sigma},\, \mathbf{u}\,;\, \mathbf{f}) = \|\mathcal{A}_\infty\,\boldsymbol{\sigma} - \boldsymbol{\epsilon}(\mathbf{u})\|^2 + \|\nabla\cdot\boldsymbol{\sigma} + \mathbf{f}\|^2 \tag{3.21}$$

for $(\boldsymbol{\sigma},\, \mathbf{u}) \in \tilde{\mathcal{V}}$. We also define the norm functionals

$$\tilde{M}_{-1}(\boldsymbol{\tau},\, \mathbf{v}) = \|\mathbf{v}\|_1^2 + \|\boldsymbol{\tau}\|^2$$

and

$$\tilde{M}(\boldsymbol{\tau},\, \mathbf{v}) = \|\mathbf{v}\|_1^2 + \|\boldsymbol{\tau}\|^2 + \|\nabla\cdot\boldsymbol{\tau}\|^2.$$

THEOREM 3.3. *The homogeneous functionals* $\tilde{G}_{-1}(\boldsymbol{\tau}, \mathbf{v}; \mathbf{0})$ *and* $\tilde{G}(\boldsymbol{\tau}, \mathbf{v}; \mathbf{0})$ *are uniformly equivalent to the functionals* $\tilde{M}_{-1}(\boldsymbol{\tau}, \mathbf{v})$ *and* $\tilde{M}(\boldsymbol{\tau}, \mathbf{v})$*, respectively; i.e., there exist positive constants* C_1 *and* C_2 *such that*

$$\frac{1}{C_1}\,\tilde{M}_{-1}(\boldsymbol{\tau}, \mathbf{v}) \le \tilde{G}_{-1}(\boldsymbol{\tau}, \mathbf{v}\,; \mathbf{0}) \le C_1\,\tilde{M}_{-1}(\boldsymbol{\tau}, \mathbf{v}) \tag{3.22}$$

and

$$\frac{1}{C_2}\,\tilde{M}(\boldsymbol{\tau}, \mathbf{v}) \le \tilde{G}(\boldsymbol{\tau}, \mathbf{v}\,; \mathbf{0}) \le C_2\,\tilde{M}(\boldsymbol{\tau}, \mathbf{v}) \tag{3.23}$$

hold for all $(\boldsymbol{\tau}, \mathbf{v}) \in \tilde{\mathcal{V}}$.

Proof. Since $\|\mathrm{tr}\,\boldsymbol{\tau}\| \le d\,\|\boldsymbol{\tau}\|$, Theorem 3.2 with the choice of $q = -\mathrm{tr}\,\boldsymbol{\tau}/d$ yields the upper bounds in both (3.22) and (3.23) and the following lower bounds:

$$\tilde{M}_{-1}(\boldsymbol{\tau}, \mathbf{v}) \le C\,\left(\tilde{G}_{-1}(\boldsymbol{\tau}, \mathbf{v}\,; \mathbf{0}) + \|\nabla \cdot \mathbf{v}\|^2\right)$$

and

$$\tilde{M}(\boldsymbol{\tau}, \mathbf{v}) \le C\,\left(\tilde{G}(\boldsymbol{\tau}, \mathbf{v}\,; \mathbf{0}) + \|\nabla \cdot \mathbf{v}\|^2\right).$$

Now the lower bounds in both (3.22) and (3.23) are a direct consequence of the bound

$$\|\nabla \cdot \mathbf{v}\| = \|\mathrm{tr}\,(\boldsymbol{\epsilon}(\mathbf{v}) - \mathcal{A}_\infty\,\boldsymbol{\tau})\| \le d\,\|\boldsymbol{\epsilon}(\mathbf{v}) - \mathcal{A}_\infty\,\boldsymbol{\tau}\|. \qquad \square$$

3.3. Least-squares finite element methods. In this section, we restrict our attention to the least-squares method based on the L^2 norm least-squares functional $\tilde{G}$ for the stress-velocity formulation, although the method developed in this section can be developed in the same manner for the stress-velocity-pressure formulation, and discrete inverse norm methods can be developed for the inverse norm functionals (see [6]). In fact, it seems that the least-squares method based on the stress-velocity formulation may be preferable since it does not involve the pressure and, more importantly, since it has mathematical structure similar to that of linear elasticity. Consequently, we develop a unified numerical approach for both linear elasticity and linear, stationary incompressible Newtonian flows. The pressure, if desired, can be recovered using (3.2).

The variational problem corresponding to the L^2 norm least-squares functional for the stress-velocity formulation is to minimize functional (3.21) over $\tilde{\mathcal{V}}$, that is, to find $(\boldsymbol{\sigma}, \mathbf{u}) \in \tilde{\mathcal{V}}$ such that

$$\tilde{G}(\boldsymbol{\sigma}, \mathbf{u}\,; \mathbf{f}) = \inf_{(\boldsymbol{\tau}, \mathbf{v}) \in \tilde{\mathcal{V}}} \tilde{G}(\boldsymbol{\tau}, \mathbf{v}\,; \mathbf{f}). \tag{3.24}$$

By Theorem 3.3, we can conclude that (3.24) has a unique solution.

Now (3.24) very much resembles the variational problem for the least-squares formulation of linear elasticity developed in [14]. In particular, the elasticity least-squares problem for limiting case $\lambda \to \infty$ is precisely (3.24). In [14], optimal accuracy for the least-squares finite element approximations and optimal multigrid convergence rates for solving the resulting algebraic equations are established to be uniform in λ. This indicates that using the finite elements in [14] to discretize the least-squares problem in (3.24) will give optimal accuracy, and multigrid methods with optimal

complexity can be used to solve the resulting algebraic equations. For completeness, we describe these finite elements and their approximation properties and comment on multigrid methods for solving the resulting algebraic systems. For simplicity, we take the two-dimensional case ($d = 2$).

Assuming that the domain Ω is polygonal, let $\mathcal{T}_h$ be a regular triangulation of Ω (see [16]) with triangular elements of size $\mathcal{O}(h)$. Let $P_k(K)$ be the space of polynomials of degree k on triangle K, and denote the local Raviart–Thomas space of order k on K:

$$RT_k(K) = P_k(K)^2 + \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} P_k(K).$$

Then the standard $H(\mathrm{div};\,\Omega)$ conforming Raviart–Thomas space of order k [22] and the standard (conforming) continuous piecewise polynomials of degree $k+1$ are defined, respectively, by

$$\Sigma_h^k = \{\boldsymbol{\tau} \in \mathbf{X}_N \,:\, \boldsymbol{\tau}|_K \in RT_k(K)^2 \;\forall\, K \in \mathcal{T}_h\} \subset \mathbf{X}_N, \tag{3.25}$$

$$V_h^{k+1} = \{\mathbf{v} \in C^0(\Omega)^2 \,:\, \mathbf{v}|_K \in P_{k+1}(K)^2 \;\forall\, K \in \mathcal{T}_h,\ \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D\} \subset H_D^1(\Omega)^2. \tag{3.26}$$

Space Σ_h^k is used to approximate the stress, and space V_h^{k+1} is used to approximate the velocity. These spaces have the following approximation properties: let $k \geq 0$ be an integer, and let $l \in (0,\, k+1]$:

$$\inf_{\boldsymbol{\tau} \in \Sigma_h^k} \|\boldsymbol{\sigma} - \boldsymbol{\tau}\|_{H(\mathrm{div};\,\Omega)} \leq C\, h^l \left(\|\boldsymbol{\sigma}\|_l + \|\nabla \cdot \boldsymbol{\sigma}\|_l\right) \tag{3.27}$$

for $\boldsymbol{\sigma} \in H^l(\Omega)^{2\times 2} \cap \mathbf{X}_N$ with $\nabla \cdot \boldsymbol{\sigma} \in H^l(\Omega)^2$ and

$$\inf_{\mathbf{u} \in V_h^{k+1}} \|\mathbf{u} - \mathbf{v}\|_1 \leq C\, h^l\, \|\mathbf{u}\|_{l+1} \tag{3.28}$$

for $\mathbf{u} \in H^{l+1}(\Omega)^2 \cap H_D^1(\Omega)^2$. Based on the smoothness of $\boldsymbol{\sigma}$ and $\mathbf{u}$, we will choose $k+1$ to be the smallest integer greater than or equal to l.

The finite element discretization of our stress-velocity least-squares variational problem is as follows: find $(\boldsymbol{\sigma}^h,\, \mathbf{u}^h) \in \Sigma_h^k \times V_h^{k+1}$ such that

$$\tilde{G}(\boldsymbol{\sigma}^h,\, \mathbf{u}^h;\, \mathbf{f}) = \min_{(\boldsymbol{\tau},\, \mathbf{v}) \in \Sigma_h^k \times V_h^{k+1}} \tilde{G}(\boldsymbol{\tau},\, \mathbf{v};\, \mathbf{f}). \tag{3.29}$$

By Theorem 3.3 and the fact that $\Sigma_h^k \times V_h^{k+1}$ is a subspace of $\tilde{\mathcal{V}}$, (3.29) has a unique solution. As proved in [14], we have the following error estimations.

THEOREM 3.4. *Assume that the solution* $(\boldsymbol{\sigma},\, \mathbf{u})$ *of* (3.24) *is in* $H^l(\Omega)^{2\times 2} \times H^{l+1}(\Omega)^2$ *and that the divergence of the stress* $\nabla \cdot \boldsymbol{\sigma}$ *is in* $H^l(\Omega)^2$. *Let* $k+1$ *be the smallest integer greater than or equal to* l. *Then with* $(\boldsymbol{\sigma}^h,\, \mathbf{u}^h) \in \Sigma_h^k \times V_h^{k+1}$ *denoting the solution to* (3.29), *the following error estimate holds:*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\|_{H(\mathrm{div};\,\Omega)} + \|\mathbf{u} - \mathbf{u}^h\|_1 \leq C\, h^l \left(\|\boldsymbol{\sigma}\|_l + \|\nabla \cdot \boldsymbol{\sigma}\|_l + \|\mathbf{u}\|_{l+1}\right). \tag{3.30}$$

As for the pressure, it can be recovered algebraically using (3.2):

$$p^h = -\frac{1}{d} \operatorname{tr} \boldsymbol{\sigma}^h. \tag{3.31}$$

It follows from (3.2), (3.31), the triangle inequality, and Theorem 3.4 that

$$\|p - p^h\| = \frac{1}{d}\|\mathrm{tr}\,(\boldsymbol{\sigma} - \boldsymbol{\sigma}^h)\| \le \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\| \le C\,h^l. \tag{3.32}$$

Remark. Theorem 3.3 states that the homogeneous functional $\tilde{G}(\boldsymbol{\tau}, \mathbf{v}; \mathbf{0})$ is equivalent to the $H(\mathrm{div}; \Omega)$ norm for the tensor variable and the H^1 norm for the vector variable. It is then well known that multigrid methods applied to discrete linear system (3.29) have optimal convergence properties (see, e.g., [19, 2, 12, 20, 24]).

4. Pure Dirichlet boundary conditions. Many applications in incompressible Newtonian fluid flow are not posed under traction boundary conditions. It is then not necessary to use the stress as the independent variable. In fact, the stress and the deformation rate tensor may not be the variables of choice, especially if the vorticity is needed. This is because the vorticity is the skew-symmetric part of the velocity gradient, and thus the stress and deformation rate tensor do not contain information on the vorticity. For this reason, in this section we develop a least-squares method involving variables that can recover the velocity gradient and vorticity without differentiation. This least-squares method will use the finite element spaces described in section 3.3.

For simplicity, we assume the homogeneous Dirichlet boundary condition

$$\mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega. \tag{4.1}$$

4.1. First-order systems. Whereas the vorticity is the skew-symmetric part of the velocity gradient, the deformation rate tensor $\boldsymbol{\epsilon}(\mathbf{u})$ is the symmetric part of the velocity gradient. From the second equation of first-order system (2.4), it is then not possible to algebraically obtain the vorticity from the stress tensor. To accomplish this, a new variable must be introduced in place of the stress. This new variable should be chosen such that the resulting least-squares functionals have properties similar to G_{-1} and G ($\tilde{G}_{-1}$ and $\tilde{G}$) and such that both the stress and vorticity can be algebraically obtained from this variable. Insight into designing this new variable can be obtained by noting that for incompressible fluids the divergence of $(\nabla\mathbf{u})^t$ vanishes:

$$\nabla\cdot(\nabla\mathbf{u})^t = \nabla\cdot\begin{pmatrix} \partial_1 u_1 & \cdots & \partial_1 u_d \\ \vdots & \vdots & \vdots \\ \partial_d u_1 & \cdots & \partial_d u_d \end{pmatrix} = \nabla\,(\nabla\cdot\mathbf{u}) = \mathbf{0}. \tag{4.2}$$

Specifically, defining the new independent tensor variable, the pseudostress, to be

$$\tilde{\boldsymbol{\sigma}} = \frac{1}{2}\,\nabla\,\mathbf{u} - p\,I, \tag{4.3}$$

then

$$\boldsymbol{\sigma} = \tilde{\boldsymbol{\sigma}} + \frac{1}{2}\,(\nabla\,\mathbf{u})^t, \tag{4.4}$$

and so by (4.2) we have

$$\nabla\cdot\tilde{\boldsymbol{\sigma}} = \nabla\cdot\boldsymbol{\sigma}. \tag{4.5}$$

Moreover, using the incompressibility of $\mathbf{u}$, we have

$$\mathrm{tr}\,\tilde{\boldsymbol{\sigma}} = \mathrm{tr}\,\boldsymbol{\sigma} = -d\,p. \tag{4.6}$$

The pseudostress is not symmetric and probably not a primitive physical quantity. However, the resulting first-order system is

$$\left\{\begin{array}{rcll} -\nabla\cdot\tilde{\boldsymbol{\sigma}} &=& \mathbf{f} & \text{in } \Omega,\\ \tilde{\boldsymbol{\sigma}} + p\,I - \frac{1}{2}\nabla\mathbf{u} &=& \mathbf{0} & \text{in } \Omega,\\ \nabla\cdot\mathbf{u} &=& 0 & \text{in } \Omega, \end{array}\right. \tag{4.7}$$

which is essentially equivalent to (2.4). Differentiating and eliminating $\tilde{\boldsymbol{\sigma}}$ in (4.7) leads to the incompressible Stokes equations:

$$\left\{\begin{array}{rcll} -\frac{1}{2}\Delta\mathbf{u} + \nabla p &=& \mathbf{f} & \text{in } \Omega,\\ \nabla\cdot\mathbf{u} &=& 0 & \text{in } \Omega. \end{array}\right. \tag{4.8}$$

4.2. Least-squares functionals. For $\mathbf{f}\in L^2(\Omega)^d$, we define the following least-squares functionals based on first-order system (4.7):

$$F_{-1}(\tilde{\boldsymbol{\sigma}},\,\mathbf{u},\,p\,;\,\mathbf{f}) = \|\nabla\cdot\tilde{\boldsymbol{\sigma}}+\mathbf{f}\|^2_{-1,0} + \left\|\tilde{\boldsymbol{\sigma}} + p\,I - \frac{1}{2}\nabla\mathbf{u}\right\|^2 + \|\nabla\cdot\mathbf{u}\|^2 \tag{4.9}$$

and

$$F(\tilde{\boldsymbol{\sigma}},\,\mathbf{u},\,p\,;\,\mathbf{f}) = \|\nabla\cdot\tilde{\boldsymbol{\sigma}}+\mathbf{f}\|^2 + \left\|\tilde{\boldsymbol{\sigma}} + p\,I - \frac{1}{2}\nabla\mathbf{u}\right\|^2 + \|\nabla\cdot\mathbf{u}\|^2 \tag{4.10}$$

for $(\tilde{\boldsymbol{\sigma}},\,\mathbf{u},\,p)\in\mathcal{V}_0\equiv\mathbf{X}_0\times H^1_0(\Omega)^d\times L^2_0(\Omega)$.

THEOREM 4.1. *The homogeneous functionals $F_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\,\mathbf{0})$ and $F(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\,\mathbf{0})$ are uniformly equivalent to the functionals $M_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q)$ and $M(\boldsymbol{\tau},\,\mathbf{v},\,q)$, respectively; i.e., there exist positive constants C_1 and C_2 such that*

$$\frac{1}{C_1}M_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q) \le F_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\,\mathbf{0}) \le C_1 M_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q) \tag{4.11}$$

and

$$\frac{1}{C_2}M(\boldsymbol{\tau},\,\mathbf{v},\,q) \le F(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\,\mathbf{0}) \le C_2 M(\boldsymbol{\tau},\,\mathbf{v},\,q) \tag{4.12}$$

hold for all $(\boldsymbol{\tau},\,\mathbf{v},\,q)\in\mathcal{V}_0$.

Proof. The theorem can be proved in a similar manner as in Theorem 3.2. Actually, the key inequality

$$\|\nabla\mathbf{v}\|^2 \le C\,F_{-1}(\boldsymbol{\tau},\,\mathbf{v},\,q\,;\mathbf{0}) + C\,\|q\|\,\|\nabla\cdot\mathbf{v}\|, \tag{4.13}$$

which is similar to inequality (3.14), can be established easily: integration by parts and the Cauchy–Schwarz inequality lead to

$$\begin{aligned}\frac{1}{2}\|\nabla\mathbf{v}\|^2 &= \left(\frac{1}{2}\nabla\mathbf{v} - \boldsymbol{\tau} - q\,I,\,\nabla\mathbf{v}\right) - (\nabla\cdot\boldsymbol{\tau},\,\mathbf{v}) + (q,\,\nabla\cdot\mathbf{v})\\ &\le \left\|\frac{1}{2}\nabla\mathbf{v} - \boldsymbol{\tau} - q\,I\right\|\,\|\nabla\mathbf{v}\| + \|\nabla\cdot\boldsymbol{\tau}\|_{-1,0}\,\|\mathbf{v}\|_1 + \|q\|\,\|\nabla\cdot\mathbf{v}\|.\end{aligned}$$

Now (4.13) follows from the Poincaré and ϵ inequalities. ∎

As in section 3.2, we can derive the following first-order system without the pressure:

$$\left\{\begin{array}{rcll} \mathcal{A}_\infty\,\tilde{\boldsymbol{\sigma}} - \frac{1}{2}\,\nabla\,\mathbf{u} &=& \mathbf{0} & \text{in} \quad \Omega, \\ \nabla\cdot\,\tilde{\boldsymbol{\sigma}} + \mathbf{f} &=& \mathbf{0} & \text{in} \quad \Omega. \end{array}\right. \tag{4.14}$$

The corresponding least-squares functionals are

$$\tilde{F}_{-1}(\tilde{\boldsymbol{\sigma}},\,\mathbf{u}\,;\,\mathbf{f}) = \left\|\mathcal{A}_\infty\,\tilde{\boldsymbol{\sigma}} - \frac{1}{2}\,\nabla\,\mathbf{u}\right\|^2 + \|\nabla\cdot\,\tilde{\boldsymbol{\sigma}} + \mathbf{f}\|^2_{-1,0} \tag{4.15}$$

and

$$\tilde{F}(\tilde{\boldsymbol{\sigma}},\,\mathbf{u}\,;\,\mathbf{f}) = \left\|\mathcal{A}_\infty\,\tilde{\boldsymbol{\sigma}} - \frac{1}{2}\,\nabla\,\mathbf{u}\right\|^2 + \|\nabla\cdot\,\tilde{\boldsymbol{\sigma}} + \mathbf{f}\|^2 \tag{4.16}$$

for $(\tilde{\boldsymbol{\sigma}},\,\mathbf{u}) \in \tilde{\mathcal{V}}_0 \equiv \mathbf{X}_0 \times H^1_0(\Omega)^d$.

THEOREM 4.2. *The homogeneous functionals $\tilde{F}_{-1}(\boldsymbol{\tau},\,\mathbf{v};\,\mathbf{0})$ and $\tilde{F}(\boldsymbol{\tau},\,\mathbf{v};\,\mathbf{0})$ are uniformly equivalent to the functionals $\tilde{M}_{-1}(\boldsymbol{\tau},\,\mathbf{v})$ and $\tilde{M}(\boldsymbol{\tau},\,\mathbf{v})$, respectively; i.e., there exist positive constants C_1 and C_2 such that*

$$\frac{1}{C_1}\,\tilde{M}_{-1}(\boldsymbol{\tau},\,\mathbf{v}) \le \tilde{F}_{-1}(\boldsymbol{\tau},\,\mathbf{v}\,;\,\mathbf{0}) \le C_1\,\tilde{M}_{-1}(\boldsymbol{\tau},\,\mathbf{v}) \tag{4.17}$$

and

$$\frac{1}{C_2}\,\tilde{M}(\boldsymbol{\tau},\,\mathbf{v}) \le \tilde{F}(\boldsymbol{\tau},\,\mathbf{v}\,;\,\mathbf{0}) \le C_2\,\tilde{M}(\boldsymbol{\tau},\,\mathbf{v}) \tag{4.18}$$

hold for all $(\boldsymbol{\tau},\,\mathbf{v}) \in \tilde{\mathcal{V}}_0$.

Proof. The theorem can be shown in a similar fashion as in Theorem 3.3. ∎

Remark. The mixed variational problem based on (4.14) is to find $(\tilde{\boldsymbol{\sigma}},\,\mathbf{u}) \in \mathbf{X}_0 \times L^2_0(\Omega)^d$ such that

$$\left\{\begin{array}{lcll} (\mathcal{A}_\infty\,\tilde{\boldsymbol{\sigma}},\,\boldsymbol{\tau}) + \frac{1}{2}(\mathbf{u},\,\nabla\cdot\,\boldsymbol{\tau}) &=& \mathbf{0} & \forall\ \boldsymbol{\tau} \in \mathbf{X}_0, \\ (\nabla\cdot\,\tilde{\boldsymbol{\sigma}},\,\mathbf{v}) &=& -(\mathbf{f},\,\mathbf{v}) & \forall\ \mathbf{v} \in L^2(\Omega)^d. \end{array}\right. \tag{4.19}$$

It is easy to see that (4.19) is essentially a vector version of the mixed formulation for the second-order elliptic problems. Therefore, any stable pair of finite elements for the second-order elliptic problems (see [8]) is also a stable approximation for (4.19). This will be studied in a forthcoming paper.

4.3. Least-squares finite element methods. The variational problem for the L^2 norm least-squares formulation of (4.14) is to minimize least-squares functional (4.16) over $\tilde{\mathcal{V}}_0$, that is, to find $(\tilde{\boldsymbol{\sigma}},\,\mathbf{u}) \in \tilde{\mathcal{V}}_0$ such that

$$\tilde{F}(\tilde{\boldsymbol{\sigma}},\,\mathbf{u}\,;\,\mathbf{f}) = \inf_{(\boldsymbol{\tau},\,\mathbf{v})\in\tilde{\mathcal{V}}_0} \tilde{F}(\boldsymbol{\tau},\,\mathbf{v}\,;\,\mathbf{f}). \tag{4.20}$$

By Theorem 4.2, (4.20) has a unique solution. The discrete finite element problem is to find $(\tilde{\boldsymbol{\sigma}}^h,\,\mathbf{u}^h) \in \Sigma^k_h \times V^{k+1}_h$ such that

$$\tilde{F}(\tilde{\boldsymbol{\sigma}}^h,\,\mathbf{u}^h;\,\mathbf{f}) = \min_{(\boldsymbol{\tau},\,\mathbf{v})\in\Sigma^k_h\times V^{k+1}_h} \tilde{F}(\boldsymbol{\tau},\,\mathbf{v};\,\mathbf{f}). \tag{4.21}$$

Since $\Sigma_h^k \times V_h^{k+1}$ is a subspace of $\tilde{\mathcal{V}}_0$ ($\Gamma_D = \partial\Omega$ and $\Gamma_N = \emptyset$), by Theorem 4.2, (4.21) has a unique solution. We have the following error estimate for the finite element approximation.

THEOREM 4.3. *Assume that the solution* $(\tilde{\boldsymbol{\sigma}}, \mathbf{u})$ *of* (4.20) *is in* $H^l(\Omega)^{2\times 2} \times H^{l+1}(\Omega)^2$ *and that* $\nabla \cdot \tilde{\boldsymbol{\sigma}}$ *is in* $H^l(\Omega)^2$. *Let* $k+1$ *be the smallest integer greater than or equal to* l. *Then with* $(\tilde{\boldsymbol{\sigma}}^h, \mathbf{u}^h) \in \Sigma_h^k \times V_h^{k+1}$ *denoting the solution to* (4.21), *the following error estimate holds:*

$$\|\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}^h\|_{H(\mathrm{div};\,\Omega)} + \|\mathbf{u} - \mathbf{u}^h\|_1 \le C\, h^l \left(\|\tilde{\boldsymbol{\sigma}}\|_l + \|\nabla \cdot \tilde{\boldsymbol{\sigma}}\|_l + \|\mathbf{u}\|_{l+1}\right). \tag{4.22}$$

4.4. Computation of pressure, stress, and vorticity. Physical quantities such as pressure, stress, and vorticity can be approximated in terms of $\tilde{\boldsymbol{\sigma}}^h$. For the pressure, (4.6) gives

$$p = -\frac{1}{d} \operatorname{tr} \tilde{\boldsymbol{\sigma}}. \tag{4.23}$$

For the stress, note that the first equation in (4.14) gives

$$\nabla \mathbf{u} = 2\, \mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}}. \tag{4.24}$$

This, together with (4.4), implies

$$\boldsymbol{\sigma} = \tilde{\boldsymbol{\sigma}} + \frac{1}{2} (\nabla \mathbf{u})^t = \tilde{\boldsymbol{\sigma}} + (\mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}})^t = \mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}} + \tilde{\boldsymbol{\sigma}}^t. \tag{4.25}$$

For the vorticity $\boldsymbol{\omega} = \nabla \times \mathbf{u}$, it can be expressed in terms of the entries of the skew-symmetric part of the velocity gradient and, hence, the pseudostress $\tilde{\boldsymbol{\sigma}}$. More precisely, letting $\mathbf{s} = \frac{1}{2} (\nabla \mathbf{u} - (\nabla \mathbf{u})^t)$, the definition of the curl operator gives

$$\boldsymbol{\omega} = \begin{cases} 2 s_{21}(\mathbf{u}) & \text{if } d = 2, \\ 2(s_{32}(\mathbf{u}),\, s_{13}(\mathbf{u}),\, s_{21}(\mathbf{u}))^t & \text{if } d = 3. \end{cases}$$

Then by (4.24) we have

$$\mathbf{s}(\mathbf{u}) = \mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}} - (\mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}})^t = \tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}^t$$

and, hence,

$$\boldsymbol{\omega} = 2 \begin{cases} \tilde{\sigma}_{21} - \tilde{\sigma}_{12} & \text{if } d = 2, \\ (\tilde{\sigma}_{32} - \tilde{\sigma}_{23},\, \tilde{\sigma}_{13} - \tilde{\sigma}_{31},\, \tilde{\sigma}_{21} - \tilde{\sigma}_{12})^t & \text{if } d = 3. \end{cases} \tag{4.26}$$

Equations (4.23), (4.25), and (4.26) suggest that we can approximate the pressure, stress, and vorticity as

$$p^h = -\frac{1}{d} \operatorname{tr} \tilde{\boldsymbol{\sigma}}^h, \qquad \boldsymbol{\sigma}^h = \mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}}^h + (\tilde{\boldsymbol{\sigma}}^h)^t,$$

$$\boldsymbol{\omega}_h = 2 \begin{cases} \tilde{\sigma}_{21}^h - \tilde{\sigma}_{12}^h & \text{if } d = 2, \\ (\tilde{\sigma}_{32}^h - \tilde{\sigma}_{23}^h,\, \tilde{\sigma}_{13}^h - \tilde{\sigma}_{31}^h,\, \tilde{\sigma}_{21}^h - \tilde{\sigma}_{12}^h)^t & \text{if } d = 3. \end{cases}$$

From (4.23), (4.25), (4.26), the triangle inequality, and Theorem 4.3, we have the following error estimates:

$$\|p - p^h\| = \frac{1}{d} \|\mathrm{tr}(\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}^h)\| \le C\, h^l,$$

$$\|\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}^h\| = \|\mathcal{A}_\infty(\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}^h) + (\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}^h)^t\| \le C\, h^l,$$

$$\|\boldsymbol{\omega} - \boldsymbol{\omega}^h\| \le C\, \|\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}^h\| \le C\, h^l.$$

4.5. Weakly imposed boundary conditions. In the previous sections, boundary conditions were imposed on the solution spaces. This leads to least-squares finite element approximations that are more accurate on the boundary than in the interior of the domain. In the context of least-squares methods, it is natural to treat boundary conditions weakly through boundary functionals. This is also convenient for many applications.

As an example of least-squares boundary functionals, we describe a least-squares functional with boundary terms for first-order system (4.14):

$$\tilde{F}_b(\tilde{\boldsymbol{\sigma}},\, \mathbf{u}\,;\, \mathbf{f}) = \left\| \mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}} - \frac{1}{2}\, \nabla\, \mathbf{u} \right\|^2 + \|\nabla \cdot \tilde{\boldsymbol{\sigma}} + \mathbf{f}\|^2 + \|\mathbf{u}\|^2_{\frac{1}{2},\partial\Omega}. \tag{4.27}$$

The least-squares variational problem is to minimize this functional over a solution space free of imposed boundary conditions: find $(\tilde{\boldsymbol{\sigma}},\, \mathbf{u}) \in \tilde{\mathcal{V}}_b \equiv \mathbf{X}_0 \times H^1(\Omega)^d$ such that

$$\tilde{F}_b(\tilde{\boldsymbol{\sigma}},\, \mathbf{u}\,;\, \mathbf{f}) = \inf_{(\boldsymbol{\tau},\, \mathbf{v}) \in \tilde{\mathcal{V}}_b} \tilde{F}_b(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{f}). \tag{4.28}$$

Using techniques in this paper and in the proof of Theorem 5.1 in [14], we can show that there exists a positive constant C such that

$$\frac{1}{C}\, \tilde{M}(\boldsymbol{\tau},\, \mathbf{v}) \le \tilde{F}_b(\boldsymbol{\tau},\, \mathbf{v}\,;\, \mathbf{0}) \le C\, \tilde{M}(\boldsymbol{\tau},\, \mathbf{v}) \tag{4.29}$$

for all $(\boldsymbol{\tau},\, \mathbf{v}) \in \tilde{\mathcal{V}}_b$. To develop computable finite element methods and the corresponding iterative solvers based on this functional, see [23].

4.6. Relation to existing least-squares methods. There are many existing least-squares methods for the Stokes equations. Since the pseudostress $\tilde{\boldsymbol{\sigma}}$ involves the velocity gradient and the pressure, our approach has some similarities with the methods in [11, 15]. In [11], the velocity gradient is introduced as an independent variable; two additional (consistent) constraints (vanishing trace and curl of the velocity gradient) are added to the original system; the variables of the least-squares method are the velocity, velocity gradient, and pressure; and the homogeneous L^2 norm least-squares functional is elliptic and continuous in $(H(\mathrm{div};\, \Omega)^d \cap H(\mathbf{curl}\,;\, \Omega)^d) \times H^1(\Omega)^d \times H^1(\Omega)$, where $H(\mathbf{curl}\,;\, \Omega)$ is the Hilbert space consisting of square-integrable vectors whose curls are also square-integrable. In [15], a constrained velocity gradient (the velocity gradient satisfying the incompressibility condition) is introduced as an independent variable; the least-squares method is based on the div-curl system of the constraint velocity gradient and the pressure; and the homogeneous functional is elliptic and continuous in $(H(\mathrm{div};\, \Omega)^d \cap H(\mathbf{curl}\,;\, \Omega)^d) \times H^1(\Omega)$. Both methods require sufficient smoothness for the original problem, and, hence, their applicability is very limited.

As a side remark, we comment that the div-curl least-squares method can be developed for our formulations. To see this, applying the curl operator to the first equation of (4.14) leads to the following div-curl system:

$$\begin{cases} \nabla \times (\mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}}) &=& \mathbf{0} & \text{in} & \Omega, \\ \nabla \cdot\, \tilde{\boldsymbol{\sigma}} + \mathbf{f} &=& \mathbf{0} & \text{in} & \Omega, \end{cases} \tag{4.30}$$

with boundary conditions $\mathbf{n} \times (\mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}}) = \mathbf{n} \times \nabla\, \mathbf{u} = \mathbf{0}$ on $\partial\Omega$. The corresponding least-squares functionals are defined as

$$\bar{F}_{-1}(\tilde{\boldsymbol{\sigma}};\, \mathbf{f}) = \|\nabla \times (\mathcal{A}_\infty\, \tilde{\boldsymbol{\sigma}})\,\|^2_{-1,0} + \|\nabla \cdot \tilde{\boldsymbol{\sigma}} + \mathbf{f}\|^2_{-1} \tag{4.31}$$

and

$$\bar{F}(\tilde{\boldsymbol{\sigma}}; \mathbf{f}) = \|\nabla \times (\mathcal{A}_\infty \tilde{\boldsymbol{\sigma}})\|^2 + \|\nabla \cdot \tilde{\boldsymbol{\sigma}} + \mathbf{f}\|^2. \tag{4.32}$$

These div-curl approaches will be studied in a forthcoming paper.

REFERENCES

[1] D. N. Arnold, J. Douglas, and C. P. Gupta, *A family of higher order mixed finite element methods for plane elasticity*, Numer. Math., 45 (1984), pp. 1–22.

[2] D. N. Arnold, R. S. Falk, and R. Winther, *Multigrid in H(div) and H(curl)*, Numer. Math., 85 (2000), pp. 197–218.

[3] P. B. Bochev and M. D. Gunzburger, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., 63 (1994), pp. 479–506.

[4] P. B. Bochev and M. D. Gunzburger, *Least-squares for the velocity-pressure-stress formulation of the Stokes equations*, Comput. Methods Appl. Mech. Engrg., 126 (1995), pp. 267–287.

[5] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[6] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order system*, Math. Comp., 66 (1997), pp. 935–955.

[7] J. H. Bramble and J. E. Pasciak, *Least-squares method for Stokes equations based on a discrete minus one inner product*, J. Comput. Appl. Math., 74 (1996), pp. 155–173.

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

[9] F. Brezzi, J. Rappaz, and P. A. Raviart, *Finite-dimensional approximation of nonlinear problems, Part* 1: *Branches of nonsingular solutions*, Numer. Math., 36 (1980), pp. 1–25.

[10] Z. Cai, T. Manteuffel, and S. McCormick, *First-order system least squares for velocity-vorticity- pressure form of the Stokes equations, with application to linear elasticity*, Electron. Trans. Numer. Anal., 3 (1995), pp. 150–159.

[11] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for the Stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[12] Z. Cai, R. R. Parashkevov, T. F. Russell, J. D. Wilson, and X. Ye, *Domain decomposition for a mixed finite element method in three dimensions*, SIAM J. Numer. Anal., 41 (2003), pp. 181–194.

[13] Z. Cai and G. Starke, *First-order system least squares for the stress-displacement formulation: Linear elasticity*, SIAM J. Numer. Anal., 41 (2003), pp. 715–730.

[14] Z. Cai and G. Starke, *Least-squares methods for linear elasticity*, SIAM J. Numer. Anal., 42 (2004), pp. 826–842.

[15] C. Chang, *A mixed finite element method for the Stokes problem: An acceleration pressure formulation*, Appl. Math. Comput., 36 (1990), pp. 135–146.

[16] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[17] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer, New York, 1986.

[18] M. D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, Boston, 1989.

[19] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer, New York, 1985.

[20] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.

[21] B. Jiang, *The Least-Squares Finite Element Method: Theory and Applications in Computational Fluid Dynamics and Electromagnetics*, Springer, Berlin, 1998.

[22] P. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer, New York, 1977, pp. 292–315.

[23] G. Starke, *Multilevel boundary functionals for least-squares mixed finite element methods*, SIAM J. Numer. Anal., 36 (1999), pp. 1065–1077.

[24] P. S. Vassilevski and J. Wang, *Multilevel iterative methods for mixed finite element discretizations of elliptic problems*, Numer. Math., 63 (1992), pp. 503–520.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 860–877

APPROXIMATIONS OF VERY WEAK SOLUTIONS TO BOUNDARY-VALUE PROBLEMS*

MARTIN BERGGREN†

Abstract. Standard weak solutions to the Poisson problem on a bounded domain have square-integrable derivatives, which limits the admissible regularity of inhomogeneous data. The concept of solution may be further weakened in order to define solutions when data is rough, such as for inhomogeneous Dirichlet data that is only square-integrable over the boundary. Such very weak solutions satisfy a nonstandard variational form $(u, v) = G(v)$. A Galerkin approximation combined with an approximation of the right-hand side G defines a finite-element approximation of the very weak solution. Applying conforming linear elements leads to a discrete solution equivalent to the text-book finite-element solution to the Poisson problem in which the boundary data is approximated by L^2-projections. The L^2 convergence rate of the discrete solution is $O(h^s)$ for some $s \in (0, 1/2)$ that depends on the shape of the domain, assuming a polygonal (two-dimensional) or polyhedral (three-dimensional) domain without slits and (only) square-integrable boundary data.

Key words. finite-element methods, very weak solution, transposition, rough data

AMS subject classifications. 35J05, 35J20, 65N30

DOI. 10.1137/S0036142903382048

1. Introduction.

Applications such as optimal control, inverse problems, and shape optimization sometimes call for boundary data that are rougher than the theory for elliptic or parabolic boundary-value problems routinely assumes. This note addresses numerical issues in the presence of rough boundary data.

We restrict the discussion to a simple case, the Poisson equation with inhomogeneous Dirichlet conditions,

$$-\Delta u = f \qquad \text{in } \Omega, \tag{1a}$$

$$u = g \qquad \text{on } \Gamma, \tag{1b}$$

where Ω is a bounded domain in $\mathbb{R}^2$ or $\mathbb{R}^3$, Γ the domain boundary, and f and g are given data. Integration by parts yields that smooth solutions to (1) satisfy

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \tag{2}$$

for each smooth v vanishing on Γ. The "standard" weak solution to (1), the basis for finite-element discretizations, satisfies variational expression (2) with u and v being elements in certain subspaces of $H^1(\Omega)$, the Sobolev space of order one.

Which type of boundary data g makes sense to specify in (1)? The answer depends on which type of functions we accept as being solutions. For standard weak solutions, it is necessary that g can be extended continuously into a function in $H^1(\Omega)$. Such extensions are not always possible—there are even continuous functions g on the

*Received by the editors March 26, 2003; acceped for publication (in revised form) October 6, 2003; published electronically June 4, 2004. This research was supported in part by the Computer Science Research Institute at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockhead Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

http://www.siam.org/journals/sinum/42-2/38204.html

†Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden (Martin.Berggren@it.uu.se).

boundary Γ that do not extend continuously. The situation is even worse for nonsmooth data. For instance, no function with jump discontinuities on the boundary can be extended continuously into a function in $H^1(\Omega)$.

However, even weaker solutions to system (1) than the standard weak solution relax the requirements on g. Integrating (2) by parts once more reveals that solutions to (1) satisfy

$$-\int_\Omega u\,\Delta v\,dx = -\int_\Gamma g\frac{\partial v}{\partial n}\,d\Gamma + \int_\Omega f v\,dx \tag{3}$$

for each smooth v vanishing on Γ. The boundary integral on the right-hand side of expression (3) now makes sense for g being merely square-integrable, as long as the normal derivative of v is also square-integrable on the boundary. Variational expression (3)—in a version made precise in section 4—is the basis for defining very weak solutions to the Poisson equation with boundary data being no more than square-integrable. Compared to the variational form (2), the nonstandard variational form (3) relaxes the regularity requirement on u and g at the price of higher regularity requirements on v, a particular example of the method of *transposition*, treated in great generality in the classic three-volume treatise of Lions and Magenes [15].

This note considers a numerical approximation based on the *Lions-type* variational expression (3) and proves optimal-order convergence rates for linear, conforming elements. To the best of my knowledge, this approach to discretizing problems with rough boundary data has not previously been reported in the literature. Previous analysis of problems with rough data has concerned other discretization approaches.

Babuška [1] defines weak solutions using a generalized Lax–Milgram lemma, a version that would apply, for instance, to the form (3) in the case of full elliptic regularity and a smooth boundary. (Because of lack of regularity, we cannot directly use his method here.) He considers finite-element approximations of the Poisson problem based on the standard variational form (2) with homogeneous boundary data and proves error bounds that cover also very weak solutions such as Green's functions.

French and King [8] analyze a parabolic initial–boundary-value problem for convex domains in $\mathbb{R}^2$. Using temporal averaging combined with spatial L^2-projections for the boundary data, they prove error estimates for a standard finite-element approximation in space, combined with the backward Euler scheme in time, of a very weak solution to the parabolic problem.

In the context of an optimal-control problem for a second-order elliptic equation on convex domains in $\mathbb{R}^2$, French and King [7] introduce a standard finite-element approximation and show that it converges to the very weak solution defined by transposition. Another contribution in this direction is by Bramble and King [2]. They consider elliptic problems on smooth, curved domains in $\mathbb{R}^2$, and their error estimates hold also for rough boundary data.

In all articles cited above, $L^2(\Gamma)$-projections approximate the rough boundary data, which allows the use of the standard variational form (2). In contrast, this article uses the Lions-type variational form (3) as the basis for discretization. Thus, the variational form in the discretization is identical to the one used to define the (very) weak solution, and projection of the data is not needed. Nonetheless, a perhaps surprising result of this article (Theorem 5.2) is that the discrete solution obtained with the current approach is *equivalent* to the standard finite-element approximation combined with $L^2(\Gamma)$-projections of the inhomogeneous data. This observation removes

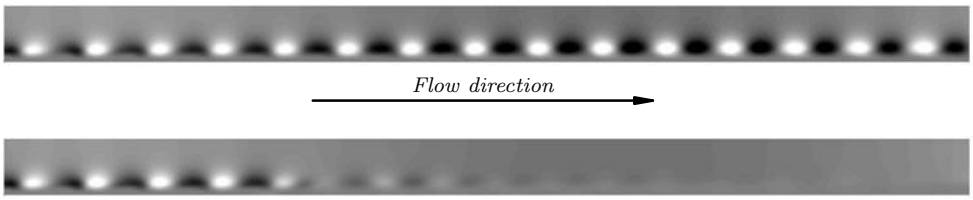


FIG. 1. *Wall-normal velocity disturbance levels depicting Tollmien–Schlichting waves in a boundary layer over a flat plate. The lower picture shows the result after applying boundary control at a portion of the wall.*

some of the arbitrariness of the standard method with data projections, showing its equivalence with a systematic scheme based on the Lions-type variational form.

Another difference from previous work is that the analysis below covers both two and three space dimensions with polygonal or polyhedral boundaries, without assuming convexity of the domain. The analysis in the articles cited above is restricted to two dimensions and assumes convex, polygonal domains.

Solutions as weak as the ones considered here have interest beyond mathematical curiosity. Applications in which very weak solutions appear naturally are *boundary control* and *inverse* problems [4, 9, 14]. The rest of the presentation starts with a short outline of this background in section 2. Section 3 reviews some fundamentals, notation, and the approximation properties that are needed, setting the stage for a more precise description in section 4 of the Poisson-equation solution based on transposition. Section 5 introduces a finite-element approximation of the very weak solution, proving convergence rates and equivalence with the "standard" approximation with projection of the data.

2. Background. Scientific and technical reasons prompt the need for *controlling* the behavior of solutions to partial differential equations, for instance through boundary action. In fluid-dynamics applications, the object is typically to manage the evolution of disturbances. Figure 1 depicts a boundary layer with evolving *Tollmien–Schlichting* waves, the most unstable disturbance according to linear stability theory. The pictures are snapshots from numerical solutions of the unsteady, incompressible Navier–Stokes equations in three space dimensions. A suitable blowing and suction at a portion of the lower boundary dramatically damps the disturbances, as shown in the lower picture of Figure 1. Chevalier et al. [4] report details of this and several similar computations. The blowing and suction is the numerical solution to a nonlinear optimization problem that minimizes the disturbances in the domain over a space of admissible controls. The admissible controls are boundary conditions with no more regularity a priori than being square-integrable functions on a portion of the lower boundary and during a finite time interval. This is a weaker regularity requirement on the boundary condition than needed for weak solutions of the Navier–Stokes equations.

The same regularity concern is an issue also for simpler problems that are easier to analyze. Consider, for instance, the following *inverse problem* for the Poisson equation (1). Given a function z defined in a subdomain $\omega \subset \Omega$, find the boundary condition g that yields $u = z$ in ω. Thinking of (1) as a model for steady heat conduction in a homogeneous, isotropic solid, the inverse problem consists of estimating the temperature g on the boundary given measurements in the interior (in ω).

The inverse problem above is only solvable for a very restricted class of *targets* z. Perhaps the easiest way of getting around this restriction is to solve a linear least-squares problem and minimize the *objective function*

$$J(g) = \frac{\epsilon}{2}\int_\Gamma g^2\,d\Gamma + \frac{1}{2}\int_\omega (u-z)^2\,dx,$$

where $\epsilon > 0$ is a (Tikhonov) regularization parameter, included to prevent g from becoming unbounded. The classical exposition of optimal control problems of this sort is the book by Lions [14]. A newer review by Glowinski and Lions [9] covers also numerical aspects.

To minimize J among all $g \in L^2(\Gamma)$, u needs to be well defined and square-integrable for each $g \in L^2(\Gamma)$. However, the standard variational form (2) of the Poisson equation requires also the derivatives of u to be elements in $L^2(\Omega)$, a property that does not hold for all $g \in L^2(\Omega)$. Including also derivatives (possibly fractional) of g along the boundary in the regularization term of J fixes this problem. However, the derivatives of g complicate a numerical solution of the control problem and introduce an extra *smoothing*, which may be unwanted, of the g that minimizes J.

There are also other reasons to prefer L^2 norms. For instance, for studies of stability and transition in fluid mechanics, the customary measure of the "size" of velocity quantities is expressed as L^2-like norms, because of the connection to the kinetic energy of the fluid (Schmid and Henningson [16]).

3. Preliminaries.

3.1. Notation, function spaces. Consider an open, bounded, and connected domain Ω in $\mathbb{R}^2$ or $\mathbb{R}^3$ with a Lipschitz boundary Γ; that is, the boundary is locally the graph of a Lipschitz function (for details see Definition 1.2.1.1 in Grisvard [11], for instance). We denote by $H^s(\Omega)$ the Sobolev space of order s on Ω. When s is a nonnegative integer, $H^s(\Omega)$ is the space in which each function and all its (weak) partial derivatives up to order s are square-integrable over Ω. We use the convention $H^0(\Omega) = L^2(\Omega)$ and $H^0(\Gamma) = L^2(\Gamma)$. Introducing a norm containing integrals over the domain, as in Definition 1.2.1 in Grisvard [12], generalizes the definition of $H^s(\Omega)$ to any real positive s. An alternative generalization uses interpolation of Hilbert spaces, as in section 2 of Chapter 1 in Lions and Magenes [15, Volume 1]. Brenner and Scott [3, Theorem 12.2.7] provide a proof that the spaces generated in these two ways are equivalent when the boundary is Lipschitz.

The *trace* γv of a function $v \in H^s(\Omega)$ generalizes to Sobolev spaces the *restriction* $v|_\Gamma$ of a smooth function v to the boundary. Unfortunately, the presence of "edges" and "corners" on a nonsmooth boundary complicates the trace concept compared to the case when the boundary is smooth. Nevertheless, it follows from Theorem 1.5.1.2 in Grisvard [11] that for $s \in (1/2, 1]$, each function v in $H^s(\Omega)$ has a well-defined trace γv in the Sobolev space $H^{s-1/2}(\Gamma)$, and that there exists a $C > 0$ such that

$$\|\gamma v\|_{s-1/2,\Gamma} \le C\|v\|_s \qquad \forall v \in H^s(\Omega). \tag{4}$$

Expression (4) uses the notation $\|\cdot\|_{s,\Gamma}$ for norms on $H^s(\Gamma)$. Analogously to $H^s(\Omega)$, integrals over the Lipschitz boundary Γ define a norm on $H^s(\Gamma)$ as long as $s \in [0, 1]$ (section 1.3.3 in Grisvard [11]). Restricting the domain of integration, we may also define norms $\|\cdot\|_{s,\Gamma_i}$ on open subsets Γ_i of Γ.

The closure of $C_0^\infty(\Omega)$, the infinitely differentiable functions with compact support in Ω, with respect to the norm in $H^s(\Omega)$ forms a subspace denoted $H_0^s(\Omega)$. In

particular it holds that

$$H_0^1(\Omega) = \left\{ v \mid v \in H^1(\Omega), \gamma v = 0 \right\};$$

that is, $H_0^1(\Omega)$ is the subspace of functions in $H^1(\Omega)$ with zero trace.

Negative norms are defined by

$$\|v\|_{-s} = \sup_{w \in H_0^s(\Omega)\setminus\{0\}} \frac{1}{\|w\|_s} \int_\Omega vw \, dx, \qquad s > 0. \tag{5}$$

This norm can be used to define $H^{-s}(\Omega)$, a space of distributions on Ω strictly larger than $L^2(\Omega)$ (for instance, the space $H^{-1}(\Omega)$ can then be identified with the dual space of $H_0^1(\Omega)$). However, we will need the norm (5) only for estimates of functions $v \in L^2(\Omega)$.

The "dual" to definition (5),

$$\|w\|_s = \sup_{v \in L^2(\Omega))\setminus\{0\}} \frac{1}{\|v\|_{-s}} \int_\Omega vw \, dx, \tag{6}$$

holds for any $w \in H_0^s(\Omega)$. Yosida [18, Chapter III, section 10] provides a detailed proof for $s = 1$, but the arguments are unchanged for any $s > 0$. From definition (6) one immediately obtains the Cauchy–Schwarz-like inequality

$$\int_\Omega vw \, dx \leq \|v\|_{-s} \|w\|_s. \tag{7}$$

Similarly,

$$\|g\|_{-s,\Gamma_i} = \sup_{h \in H_0^s(\Gamma_i)\setminus\{0\}} \frac{1}{\|h\|_{s,\Gamma_i}} \int_{\Gamma_i} gh \, d\Gamma \tag{8}$$

defines negative norms on open subsets Γ_i of Γ. Again, we will apply this norm only on functions $g \in L^2(\Gamma_i)$. Let $g \in L^2(\Gamma)$ and $h \in C_0^\infty(\Gamma_i) \setminus \{0\}$ be given. Extending h by zero on $\Gamma \setminus \Gamma_i$, we see that

$$\frac{1}{\|h\|_{s,\Gamma_i}} \int_{\Gamma_i} gh \, d\Gamma = \frac{1}{\|h\|_{s,\Gamma}} \int_\Gamma gh \, d\Gamma.$$

Noting that the extended h is in $H_0^s(\Gamma_i)$ as well as in $H_0^s(\Gamma)$, and taking supremum, it follows that

$$\|g\|_{-s,\Gamma_i} \leq \|g\|_{-s,\Gamma}. \tag{9}$$

Throughout the following, C denotes a positive constant, independent of the choice of functions and, later, of the mesh parameter h. However, adhering to a customary abuse of notation, the actual value of C may change, even within the same chain of inequalities.

3.2. Regularity of the Poisson problem with homogeneous boundary conditions. Let Ω be a polyhedral domain in $\mathbb{R}^3$ or a polygonal domain in $\mathbb{R}^2$. We require the domain to be Lipschitz, which excludes domains with slits. Given $v \in L^2(\Omega)$, there is a unique $z \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla z \cdot \nabla w \, dx = \int_\Omega vw \, dx \qquad \forall w \in H_0^1(\Omega). \tag{10}$$

The regularity properties of solutions to (10) are crucial in the development below. In fact, the possibility of defining the very weak solutions is a *consequence* of the fact that the regularity is better than merely $z \in H_0^1(\Omega)$.

Indeed, if the boundary is smooth, the additional regularity $z \in H^2(\Omega)$ holds. This is still true for polygonal or polyhedral boundaries if the domain is convex. The regularity is reduced, however, in the vicinity of nonconvex portions of polygonal or polyhedral boundaries. Grisvard [12] proves precise regularity results (Theorem 2.4.3 for the two-dimensional case and Corollary 2.6.7 for the three-dimensional case), stating that there exists an $\epsilon \in (0, 1/2]$, which depends on the shape of the domain, such that solutions to (10) are actually in $H^{3/2+\epsilon}(\Omega)$. The following estimate will be needed.

THEOREM 3.1. *There exist an* $\epsilon \in (0, 1/2]$ *and a* $C > 0$ *such that the solution* $z \in H_0^1(\Omega)$ *to* (10) *satisfies*

$$\|z\|_{3/2+\epsilon-s} \le C\|v\|_{-s} \qquad \forall v \in L^2(\Omega) \tag{11}$$

for each $s \in [0, 1]$.

Proof. By the regularity result quoted above, the closed-graph theorem yields that there exists an $\epsilon \in (0, 1/2]$ such that

$$\|z\|_{3/2+\epsilon} \le C\|v\|_0 \qquad \forall v \in L^2(\Omega). \tag{12}$$

Using the notation

$$\| \, |\nabla w| \, \|_0^2 = \int_\Omega |\nabla w|^2 \, dx,$$

(10) implies that

$$\frac{1}{\| \, |\nabla w| \, \|_0} \int_\Omega \nabla z \cdot \nabla w \, dx = \frac{1}{\| \, |\nabla w| \, \|_0} \int_\Omega vw \, dx$$

for each nonzero $w \in H_0^1(\Omega)$. Taking the supremum yields that

$$\begin{aligned} \|z\|_{1/2+\epsilon} &\le C\|z\|_1 \le C\| \, |\nabla z| \, \|_0 \\ &= C \sup_{w \in H_0^1(\Omega)\setminus\{0\}} \frac{1}{\| \, |\nabla w| \, \|_0} \int_\Omega \nabla z \cdot \nabla w \, dx \\ &= C \sup_{w \in H_0^1(\Omega)\setminus\{0\}} \frac{1}{\| \, |\nabla w| \, \|_0} \int_\Omega vw \, dx \le C\|v\|_{-1} \qquad \forall v \in L^2(\Omega), \end{aligned} \tag{13}$$

where the second and the last inequality follow from the fact that the seminorm $\| \, |\nabla z| \, \|_0$ is equivalent to $\|z\|_1$ for $z \in H_0^1(\Omega)$. Estimates (12) and (13) imply that the linear mapping $v \mapsto z$ is bounded from $L^2(\Omega)$ into $H^{3/2+\epsilon}$ as well as from H^{-1} into $H^{1/2+\epsilon}(\Omega)$. Estimate (11) then follows by operator interpolation of the mapping $v \mapsto z$. ∎

We will also need expressions for the *boundary flux* associated with the solution z to (10) and the regularity properties of the boundary flux. Let us first assume full elliptic regularity, that is, $z \in H^2(\Omega) \cap H_0^1(\Omega)$. Integration by parts of the product $-\phi\Delta z$ then yields that the boundary flux $\partial z/\partial n$ satisfies

$$-\int_\Gamma \frac{\partial z}{\partial n} \phi \, d\Gamma = \int_\Omega v\phi \, dx - \int_\Omega \nabla z \cdot \nabla \phi \, dx \qquad \forall \phi \in H^1(\Omega). \tag{14}$$

If, moreover, the boundary is smooth, the boundary flux is an element in $H^{1/2}(\Gamma)$. On polyhedral boundaries, however, the discontinuity of boundary normals complicates the definition of boundary-flux spaces. Nevertheless, the integration-by-parts property (14) holds also for Lipschitz domains in $\mathbb{R}^n$ as long as $z \in H^2(\Omega) \cap H_0^1(\Omega)$; it follows from Proposition 5.1.6 in Brenner and Scott [3], for instance. The following theorem proves a similar expression for less regular z.

THEOREM 3.2. *There are an $\epsilon \in (0, 1/2]$ and a $C > 0$ such that, associated with any $v \in L^2(\Omega)$ and corresponding solution $z \in H_0^1(\Omega)$ to (10), there exists a unique $\lambda \in L^2(\Gamma)$ satisfying*

$$-\int_\Gamma \lambda\phi \, d\Gamma = \int_\Omega v\phi \, dx - \int_\Omega \nabla z \cdot \nabla\phi \, dx \qquad \forall \phi \in H^1(\Omega) \tag{15}$$

and the estimates

$$\begin{aligned} \|\lambda\|_{0,\Gamma} &\le C\|v\|_0, \\ \|\lambda\|_{\epsilon-s,\Gamma_i} &\le C\|v\|_{-s} \qquad \text{for } i = 1, \dots, I \text{ and } \forall s \in [0, \epsilon). \end{aligned} \tag{16}$$

Proof. By assumption, the boundary of Ω can be written $\Gamma = \cup_{i=1}^I \overline{\Gamma}_i$, where each (open and bounded) Γ_i is a planar polygon or a line segment embedded in $\mathbb{R}^3$ and $\mathbb{R}^2$, respectively. Let n_i denote the (constant) outward-directed unit normal associated with each polygonal surface (or line segment) Γ_i. By Theorem 3.1, we know that there is an $\epsilon \in (0, 1/2]$ such that, for each $v \in L^2(\Omega)$ supplied to (10), the solution satisfies $z \in H^{3/2+\epsilon-s}$ $\forall s \in [0, 1]$. Differentiation is a continuous operator from $H^\alpha(\Omega)$ into $H^{\alpha-1}(\Omega)$ as long as $\alpha \neq 1/2$ [11, Theorem 1.4.4.6]. Thus, for each $i = 1, \dots, I$,

$$\frac{\partial z}{\partial n_i} = n_i \cdot \nabla z$$

resides in $H^{1/2+\epsilon-s}(\Omega)$ since $3/2 + \epsilon - s \neq 1/2$ for any $\epsilon \in (0, 1/2]$ and $s \in [0, 1]$. By further restricting s, we can apply the trace theorem for Lipschitz boundaries (Grisvard [11, Theorem 1.5.1.2]) to obtain the bounds, for $s \in [0, \epsilon)$, $i = 1, \dots, I$,

$$\left\| \frac{\partial z}{\partial n_i} \right\|_{\epsilon-s,\Gamma} \le C \left\| \frac{\partial z}{\partial n_i} \right\|_{1/2+\epsilon-s} \le C \, \|z\|_{3/2+\epsilon-s} \,, \tag{17}$$

where the second inequality follows from the above-mentioned continuity of differentiation.

Now, for $i = 1, \dots, I$, we define $\lambda_i \in L^2(\Gamma)$ as

$$\lambda_i = \begin{cases} \dfrac{\partial z}{\partial n_i} & \text{on } \Gamma_i, \\ 0 & \text{on } \Gamma \setminus \Gamma_i, \end{cases}$$

and $\lambda \in L^2(\Gamma)$ as

$$\lambda = \sum_{i=1}^I \lambda_i. \tag{18}$$

The definition of λ, together with inequality (17) and Theorem 3.1, yields estimate (16).

Noting that λ in effect is a weak representation for $\partial z/\partial n$, we will prove the integration-by-parts formula (15) by approximating z by smooth functions. Indeed, since $C_0^\infty(\Omega)$ is dense in $H^{3/2+\epsilon}(\Omega)\cap H_0^1(\Omega)$, there exists a sequence $\{\zeta_n\}_{n=1}^\infty \subset C_0^\infty(\Omega)$ such that

$$\zeta_n \to z \qquad \text{in } H^{3/2+\epsilon}(\Omega) \tag{19}$$

as $n \to \infty$. It follows from inequality (17) that

$$\frac{\partial \zeta_n}{\partial n_i} \to \lambda_i \qquad \text{in } L^2(\Gamma_i) \tag{20}$$

as $n \to \infty$.

Forming $-\Delta\zeta_n$, multiplying by $w \in H_0^1(\Omega)$, integrating by parts, and noting that strong convergence in $H^{3/2+\epsilon}(\Omega) \cap H_0^1(\Omega)$ implies convergence in $H_0^1(\Omega)$, it follows from (19) and (10) that

$$\begin{aligned} \int_\Omega w(-\Delta\zeta_n)\,dx &= \int_\Omega \nabla w \cdot \nabla\zeta_n\,dx \to \int_\Omega \nabla w \cdot \nabla z\,dx \\ &= \int_\Omega wv\,dx \qquad \forall w \in H_0^1(\Omega) \end{aligned} \tag{21}$$

as $n \to \infty$. Since $H_0^1(\Omega)$ is dense in $L^2(\Omega)$, expression (21) yields that

$$-\Delta\zeta_n \to v \text{ weakly in } L^2(\Omega). \tag{22}$$

Let $\phi \in H^1(\Omega)$. Integration by parts yields

$$\begin{aligned} \int_\Omega \phi(-\Delta\zeta_n)\,dx &= -\int_\Gamma \frac{\partial\zeta_n}{\partial n}\phi\,d\Gamma + \int_\Omega \nabla\phi\cdot\nabla\zeta_n\,dx \\ &= -\sum_{i=1}^I \int_{\Gamma_i} \frac{\partial\zeta_n}{\partial n_i}\phi\,d\Gamma + \int_\Omega \nabla\phi\cdot\nabla\zeta_n\,dx. \end{aligned} \tag{23}$$

Letting $n \to \infty$, it follows from (18), (20), (21), and (22) that expression (23) converges to

$$\int_\Omega \phi v\,dx = -\int_\Gamma \lambda\phi\,d\Gamma + \int_\Omega \nabla\phi\cdot\nabla z\,dx,$$

which proves that the λ defined in expression (18) satisfies expression (15). Finally, since λ depends linearly on $v \in L^2(\Omega)$, estimate (16) also provides uniqueness of λ for each given $v \in L^2(\Omega)$. ∎

3.3. Approximation properties. Let us now triangulate the polygonal or polyhedral domain Ω and introduce a mesh parameter $h > 0$ that characterizes the triangulation. We assume nondegenerate meshes [3, Def. 4.4.13]; that is, there is a limit to how "thin" the tetrahedral may become as the mesh is refined. Denote by V^h the space of continuous functions that are linear on each triangle or tetrahedron in the mesh, and denote by V_0^h the subspace of functions in V^h vanishing on Γ. We have $V^h \subset H^1(\Omega)$ and $V_0^h \subset H_0^1(\Omega)$. The restriction to Γ of functions in V^h is denoted γV^h. We also define M^h as the space of all functions $v^h \in V^h$ that vanish at each mesh point in the strict interior of the domain. We have that $V^h = M^h \oplus V_0^h$; that

is, each function in V^h is the sum of unique functions in M^h and V_0^h. Also note that each function $g_h \in \gamma V^h$ uniquely extends to a function $\widehat{g}_h \in M^h$ that equals g_h at the boundary but vanishes at all nodes in Ω. This extension property is useful when solving inhomogeneous boundary-value problems: Given an approximation $g_h \in \gamma V^h$ of the boundary data g, extend g_h to $\widehat{g}_h \in M^h$, write the solution $u_h \in V_h$ as $u_h = u_{h,0} + \widehat{g}_h$, where $u_{h,0} \in V_0^h$, and solve for $u_{h,0}$.

An interpolation operator Π_h from $H^s(\Omega)$ into V^h characterizes the approximation properties of V^h. If $s > d/2$, where d is the space dimension, it follows from the Sobolev embeddings that Π_h can simply be chosen as the linear interpolator of function values at the nodes of the triangulation. However, we need to consider small values of s, so pointwise values may not be well defined. It is therefore appropriate to choose the Scott and Zhang interpolator [17], which uses a local averaging to generate nodal values. This interpolator yields optimal-order estimates, and the averaging is constructed to preserve piecewise-polynomial boundary conditions, a property that Lemma 5.3 exploits.

The following approximation properties hold for V^h: There exists a constant $C > 0$ such that, for all $h > 0$,

$$\|v - \Pi_h v\|_1 \leq Ch^s \|v\|_{1+s} \qquad \forall v \in H^{1+s}(\Omega),\ s \in [0,1], \tag{24a}$$

$$\|v - \Pi_h v\|_0 \leq Ch^s \|v\|_s \qquad \forall v \in H^s(\Omega),\ s \in [0,2]. \tag{24b}$$

Standard textbooks, such as Ciarlet [5], prove these properties for integral values of s. Scott and Zhang [17] supply a proof for the particular case of the above-mentioned Π_h. Operator-interpolation arguments, discussed by Brenner and Scott [3, Chapter 12], for instance, extend the estimates to intermediate real numbers s.

We also need to approximate functions defined on the boundary. Recall that the domain is polyhedral or polygonal, so $\Gamma = \cup_{i=1}^I \overline{\Gamma}_i$, where each Γ_i is an open planar polygon or an open line segment that does not overlap any other Γ_i. The space $\gamma_i V^h$ of traces of function in V^h on Γ_i is a space of continuous, piecewise-linear functions on the triangles (or intervals) of Γ_i. We may thus define an interpolation operator $\Pi_h^{\gamma_i}$ from $H^s(\Gamma_i)$ into $\gamma_i V^h$ with properties analogous to Π_h,

$$\|g - \Pi_h^{\gamma_i} g\|_{0,\Gamma_i} \leq Ch^s \|g\|_{s,\Gamma_i} \qquad \forall g \in H^s(\Gamma_i),\ s \in [0,2]. \tag{25}$$

Another type of approximation in V^h, γV^h, and $\gamma_i V^h$ are the *L^2-projections*, that is, the functions $P_h v \in V^h$, $P_h^\gamma g \in \gamma V^h$, and $P_h^{\gamma_i} g \in \gamma_i V^h$ satisfying

$$\int_\Omega P_h v\, w_h \, dx = \int_\Omega v w_h \, dx \qquad \forall w_h \in V^h, \tag{26a}$$

$$\int_\Gamma P_h^\gamma g\, \varphi_h \, d\Gamma = \int_\Gamma g \varphi_h \, d\Gamma \qquad \forall \varphi_h \in \gamma V^h, \tag{26b}$$

$$\int_{\Gamma_i} P_h^{\gamma_i} g\, \varphi_h \, d\Gamma = \int_{\Gamma_i} g \varphi_h \, d\Gamma \qquad \forall \varphi_h \in \gamma_i V^h,\ i = 1, \ldots, I, \tag{26c}$$

which are well defined for each $v \in L^2(\Omega)$ and $g \in L^2(\Gamma)$. The L^2-projections produce the discrete functions that minimize the L^2 error.

For any $g \in L^2(\Gamma_i)$ and $\psi \in H_0^s(\Gamma_i)$, we have

$$\begin{aligned}(27)\quad \int_{\Gamma_i} (g - P_h^{\gamma_i} g)\psi \, d\Gamma &= \int_{\Gamma_i} (g - P_h^{\gamma_i} g)(\psi - P_h^{\gamma_i}\psi)\, d\Gamma \\ &\le \|g - P_h^{\gamma_i} g\|_{0,\Gamma_i} \|\psi - P_h^{\gamma_i}\psi\|_{0,\Gamma_i} \le \|g - \Pi_h^{\gamma_i} g\|_{0,\Gamma_i} \|\psi - \Pi_h^{\gamma_i}\psi\|_{0,\Gamma_i} \\ &\le C\|g\|_{0,\Gamma_i} h^s \|\psi\|_{s,\Gamma_i} \quad \forall s \in [0,2],\end{aligned}$$

where the first equality follows from definition (26c), the second inequality from the fact that the $L^2(\Gamma_i)$-projection is optimal, and the third from estimate (25). Dividing expression (27) by $\|\psi\|_{s,\Gamma_i}$ and taking the supremum over all $\psi \in H_0^s(\Gamma_i) \setminus \{0\}$ yields, by definition (5),

$$\|g - P_h^{\gamma_i} g\|_{-s,\Gamma_i} \le C h^s \|g\|_{0,\Gamma_i} \qquad \forall s \in [0,2]. \tag{28}$$

We will also need the inverse estimate

$$\|v_h\|_{s,\Gamma} \le C h^{-s} \|v_h\|_{0,\Gamma}, \qquad s \in [0,1]. \tag{29}$$

In contrast to an approximation estimate like (25), the inverse estimate requires quasi-uniform mesh refinements [3, Definition 4.4.13]. That is, the quotient between the largest and smallest diameter of any triangle or line segment should stay uniformly bounded as the mesh is refined. Brenner and Scott [3, Theorem 4.5.11], for instance, prove inverse estimates for integral s and domains in $\mathbb{R}^n$. Local bi-Lipschitz change of variables, partition of unity, and operator interpolation extend these estimates to estimate (29).

4. The variational form. We will make precise the idea of a solution to the Poisson equation based on the Lions-type variational expression (3). Let us define a linear form $G : L^2(\Omega) \to \mathbb{R}$ by the following procedure.

1. Given an element $v \in L^2(\Omega)$, find $z \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla z \cdot \nabla w \, dx = \int_\Omega v w \, dx \qquad \forall w \in H_0^1(\Omega). \tag{30}$$

2. From v and z, find $\lambda \in L^2(\Gamma)$ such that

$$-\int_\Gamma \lambda \phi \, d\Gamma = \int_\Omega v\phi \, dx - \int_\Omega \nabla z \cdot \nabla \phi \, dx \qquad \forall \phi \in H^1(\Omega). \tag{31}$$

3. Set, for given $g \in L^2(\Gamma)$ and $f \in L^2(\Omega)$, uniquely associated with each G,

$$G(v) = -\int_\Gamma g\lambda \, d\Gamma + \int_\Omega f z \, dx. \tag{32}$$

THEOREM 4.1. *The form G is a bounded linear functional on $L^2(\Omega)$.*

Proof. From Theorem 3.2 it follows that λ and the boundary integral involved in the definition of G are well defined for each $g \in L^2(\Gamma)$. The form G is linear in v since λ and z are linear in v. From (30) follows the estimate

$$\|z\|_1 \le C\|v\|_{-1}. \tag{33}$$

Thus,

$$\begin{aligned}(34)\quad |G(v)| &\le \|g\|_{0,\Gamma} \|\lambda\|_{0,\Gamma} + \|f\|_{-1}\|z\|_1 \le C\left(\|g\|_{0,\Gamma}\|\lambda\|_{0,\Gamma} + \|f\|_{-1}\|v\|_{-1}\right) \\ &\le C(\|g\|_{0,\Gamma} + \|f\|_0)\|v\|_0,\end{aligned}$$

where estimates (7), (33), and (16) are used in the first, second, and third inequality, respectively. ∎

By Theorem 4.1 and Riesz representation, the following problem thus has a unique solution:

$$\begin{aligned} &\text{Find } u \in L^2(\Omega) \text{ such that} \\ &\int_\Omega uv\,dx = G(v) \qquad \forall v \in L^2(\Omega). \end{aligned} \tag{35}$$

Problem (35) defines a weak solution to the Poisson problem (1), in which the boundary data g needs only to be square-integrable. The price to pay for the reduced regularity requirement on g is that $u \notin H^1(\Omega)$ in general and that the meaning of boundary condition $u = g$ will be weak; it will be satisfied only in a distributional sense acute a la Theorem 6.5 in Chapter 2 of Lions and Magenes [15, Volume 1].

Similar to solutions to (10), solutions to problem (35) have higher regularity than asked for.

THEOREM 4.2. *For each $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ associated with definition* (32) *of G, there exists an $\epsilon \in (0, 1/2]$ so that the solution to problem* (35) *resides in $H^s(\Omega)$ for each $s \in [0, \epsilon)$.*

Proof. First, note that $H_0^s(\Omega) = H^s(\Omega)$ for $s \in (0, 1/2]$, which follows from the fact that $C_0^\infty(\Omega)$ is dense in $H^s(\Omega)$ for $s \in (0, 1/2]$ [11, Theorem 1.4.2.4]. This equivalence allows the use of expression (6) to estimate $\|u\|_s$.

To estimate the boundary-integral term in the definition (32) of G, write the $L^2(\Gamma)$-norm of λ as a sum over contributions from each polygonal surface and utilize estimate (16) in Theorem 3.2. It then follows that there exists an $\epsilon \in (0, 1/2]$ such that, for each $s \in [0, \epsilon)$,

$$\begin{aligned} \left(\int_\Gamma g\lambda\,d\Gamma\right)^2 &\le \|\lambda\|_{0,\Gamma}^2 \|g\|_{0,\Gamma}^2 = \sum_{i=1}^I \|\lambda\|_{0,\Gamma_i}^2 \|g\|_{0,\Gamma}^2 \\ &\le C \sum_{i=1}^I \|\lambda\|_{\epsilon-s,\Gamma_i}^2 \|g\|_{0,\Gamma}^2 \le C\|v\|_{-s}^2 \|g\|_{0,\Gamma}^2 \qquad \forall v \in L^2(\Omega). \end{aligned} \tag{36}$$

We also estimate the second integral in the definition (32) of G as

$$\int_\Omega fz\,dx \le \|f\|_{-1}\|z\|_1 \le \|f\|_{-1}\|z\|_{3/2+\epsilon} \le C\|f\|_{-1}\|v\|_{-s} \quad \forall v \in L^2(\Omega), \tag{37}$$

where the last inequality follows from Theorem 3.1. Equation (35), the definition (32) of G, and estimates (36) and (37) yield that, for some $\epsilon \in (0, 1/2]$,

$$\int_\Omega uv\,dx \le C\|v\|_{-s}(\|g\|_{0,\Gamma} + \|f\|_{-1}) \qquad \forall v \in L^2(\Omega),\ \forall s \in [0, \epsilon). \tag{38}$$

Dividing expression (38) by $\|v\|_{-s}$, taking the supremum over all $v \in L^2(\Omega) \setminus \{0\}$, and using property (6) yields the conclusion. ∎

5. Numerical approximations. Recall the standard Galerkin approximation of the Poisson problem with inhomogeneous boundary data: If $g_h \in \gamma V_h$ approximates the boundary data, we solve the problem:

$$\begin{aligned} &\text{Find } u_h \in V^h \text{ such that } u_h|_\Gamma = g_h \text{ and} \\ &\int_\Omega \nabla u_h \cdot \nabla v_h\,dx = \int_\Omega fv_h\,dx \qquad \forall v_h \in V_0^h. \end{aligned} \tag{39}$$

The error in this approximation depends on how the boundary data is approximated. For *homogeneous* boundary data, $g_h = 0$, standard error estimates and the regularity according to Theorem 3.1 yield that there exists an $\epsilon \in (0, 1/2]$ such that

$$\|u_h - u\|_1 \leq Ch^{1/2+\epsilon}\|f\|_0, \tag{40}$$

where the estimate holds for $\epsilon = 1/2$ when the domain is convex.

Since $V^h \subset L^2(\Omega)$, we can apply a Galerkin approximation to problem (35):

$$\begin{aligned} &\text{Find } \widetilde{u}_h \in V^h \text{ such that} \\ &\int_\Omega \widetilde{u}_h v_h \, dx = G(v_h) \qquad \forall v_h \in V^h. \end{aligned} \tag{41}$$

Subtracting (35) with $v = v_h$ from (41), we obtain

$$\int_\Omega (\widetilde{u}_h - u) v_h \, dx = 0 \qquad \forall v_h \in V^h,$$

implying that the Galerkin approximation is optimal in $L^2(\Omega)$,

$$\|\widetilde{u}_h - u\|_0 = \inf_{v_h \in V^h} \|u - v_h\|_0 \leq \|u - \Pi_h u\|. \tag{42}$$

Approximation (41) is useless as a numerical method, however, since to compute $G(v_h)$, we need the exact solutions z and λ to problems (30) and (31) for each $v_h \in V^h$. A natural alternative is to use numerical approximations z_h and λ_h instead, which pertains to a modification of G—a so-called variational crime. For this, define the linear form $G_h : L^2(\Omega) \to \mathbb{R}$ as follows:

1. Given $v \in L^2(\Omega)$, find $z_h \in V_0^h$ such that

$$\int_\Omega \nabla z_h \cdot \nabla w_h \, dx = \int_\Omega v w_h \, dx \qquad \forall w_h \in V_0^h. \tag{43}$$

2. From v and z_h, compute $\lambda_h \in \gamma V^h$ such that

$$-\int_\Gamma \lambda_h \phi_h \, d\Gamma = \int_\Omega v \phi_h \, dx - \int_\Omega \nabla z_h \cdot \nabla \phi_h \, dx \qquad \forall \phi_h \in M^h. \tag{44}$$

3. Set, given $g \in L^2(\Gamma)$ and $f \in L^2(\Omega)$,

$$G_h(v) = -\int_\Gamma g \lambda_h \, d\Gamma + \int_\Omega f z_h \, dx. \tag{45}$$

A second approximation to (35) is as follows:

$$\begin{aligned} &\text{Find } u_h \in V^h \text{ such that} \\ &\int_\Omega u_h v_h \, dx = G_h(v_h) \qquad \forall v_h \in V^h. \end{aligned} \tag{46}$$

At first glance, approximation (46) appears unreasonably costly to implement, since the computation of each component of the vector $G_h(v_h)$ requires the solution of (43) and (44)! However, a remarkable property of approximation (46), shown in Theorem 5.2, is its equivalence to the standard Galerkin approximation (39), provided that the $L^2(\Gamma)$ projection is used to approximate the inhomogeneous boundary

conditions. Thus, in practical computations, approximation (46) can be implemented as a standard Galerkin approximation combined with L^2-projections of the boundary data. The equivalence of Theorem 5.2 is a consequence of the properties of the mapping $v_h \mapsto (z_h, \lambda_h)$ involved in the definition of G_h.

LEMMA 5.1. *The mapping $v_h \mapsto (z_h, \lambda_h)$, defined by solving* (43) *and* (44)*, is bijective as a mapping $V^h \to V_0^h \times \gamma V^h$. Moreover, the functions v_h, z_h, and λ_h satisfy*

$$\int_\Omega v_h \psi_h \, dx = -\int_\Gamma \lambda_h \psi_h \, d\Gamma + \int_\Omega \nabla z_h \cdot \nabla \psi_h \, dx \qquad \forall \psi_h \in V^h. \tag{47}$$

Proof. Let $v_h \in V^h$ be given, and let $z_h \in V_0^h$ and $\lambda_h \in \gamma V^h$ be the unique solutions to (43) and (44) for $v = v_h$.

Expression (43) with $v = v_h$ can be written

$$0 = \int_\Omega v_h w_h \, dx - \int_\Omega \nabla z_h \cdot \nabla w_h \, dx \qquad \forall w_h \in V_0^h. \tag{48}$$

Adding (48) to (44) with $v = v_h$ yields

$$-\int_\Gamma \lambda_h \phi_h \, d\Gamma = \int_\Omega v_h(\phi_h + w_h) \, dx - \int_\Omega \nabla z_h \cdot \nabla(\phi_h + w_h) \, dx \tag{49}$$

for each $\phi_h \in M^h$ and each $w_h \in V_0^h$. Since $V^h = V_0^h \oplus M^h$, and since functions in V_0^h vanish on Γ, it follows that v_h, z_h, and λ_h are related through expression (47).

Conversely, let $z_h \in V_0^h$ and $\lambda_h \in \gamma V^h$ be given. Expression (47) defines an equation for v_h corresponding to a square linear system with a positive-definite matrix. Equation (47) thus has a unique solution $v_h \in V^h$. The mapping $v_h \mapsto (z_h, \lambda_h)$ is thus bijective, since the mapping itself as well as its inverse are one-to-one. ∎

With the aid of the mapping of Lemma 5.1, we can transfer between the "new" approximation (46) and the traditional (39), as follows.

THEOREM 5.2. *The function $u_h \in V^h$ is a solution to problem* (46) *if and only if*

$$u_h \in V^h \text{ such that}$$

$$\int_\Omega \nabla u_h \cdot \nabla z_h \, dx = \int_\Omega f z_h \, dx \qquad \forall z_h \in V_0^h, \tag{50a}$$

$$u_h = P_h^\gamma g \qquad \text{on } \Gamma, \tag{50b}$$

where $P_h^\gamma g$ is the $L^2(\Gamma)$-projection of g on γV^h, that is,

$$P_h^\gamma g \in \gamma V^h \text{ such that}$$

$$\int_\Gamma P_h^\gamma g \, r_h \, d\Gamma = \int_\Gamma g \, r_h \, d\Gamma \qquad \forall r_h \in \gamma V^h. \tag{51}$$

Proof. (i) Let u_h be the solution to problem (50). Let $v_h \in V^h$ be given, and compute $z_h \in V_0^h$ and $\lambda_h \in \gamma V^h$ by solving (43) and (44) with $v = v_h$. By Lemma 5.1, v_h, z_h, and λ_h are related through the expression

$$-\int_\Gamma \lambda_h \psi_h \, d\Gamma + \int_\Omega \nabla z_h \cdot \nabla \psi_h \, dx = \int_\Omega v_h \psi_h \, dx \qquad \forall \psi_h \in V^h.$$

Choosing $\psi_h = u_h$, it follows that

$$-\int_\Gamma \lambda_h u_h \, d\Gamma + \int_\Omega \nabla z_h \cdot \nabla u_h \, dx = \int_\Omega v_h u_h \, dx. \tag{52}$$

Using (50a) to replace the second term in expression (52), we obtain

$$\begin{aligned} \int_\Omega u_h v_h \, dx &= -\int_\Gamma \lambda_h u_h \, d\Gamma + \int_\Omega f z_h \, dx = -\int_\Gamma \lambda_h \, P_h^\gamma g \, d\Gamma + \int_\Omega f z_h \, dx \\ &= -\int_\Gamma \lambda_h g \, d\Gamma + \int_\Omega f z_h \, dx = G_h(v_h), \end{aligned} \tag{53}$$

where we have used (50b) in the second equality, definition (51) of $P_h^\gamma g$ in the third, and definition (45) of G_h in the fourth equality. Since $v_h \in V^h$ was arbitrary, we have shown that if u_h solves (50), it also solves (46).

(ii) Conversely, let u_h be the solution of problem (46), and let $\lambda_h \in \gamma V^h$ and $z_h \in V_0^h$ be given. By Lemma 5.1, there is a unique $v_h \in V^h$ satisfying

$$\int_\Omega v_h \psi_h \, dx = -\int_\Gamma \lambda_h \psi_h \, d\Gamma + \int_\Omega \nabla z_h \cdot \nabla \psi_h \, dx \qquad \forall \psi_h \in V^h. \tag{54}$$

Choosing $\psi_h = u_h$, we find

$$-\int_\Gamma \lambda_h u_h \, d\Gamma + \int_\Omega \nabla z_h \cdot \nabla u_h \, dx = \int_\Omega v_h u_h \, dx. \tag{55}$$

Since u_h is a solution of problem (46), the right-hand side of expression (55) satisfies

$$\int_\Omega v_h u_h \, dx = -\int_\Gamma g \widehat{\lambda}_h \, d\Gamma + \int_\Omega f \widehat{z}_h \, dx, \tag{56}$$

where $\widehat{z}_h$ and $\widehat{\lambda}_h$ are the solutions to (43) and (44) with $v = v_h$. However, since by Lemma 5.1, $\widehat{z}_h$ and $\widehat{\lambda}_h$ are uniquely defined by v_h, we have $\widehat{z}_h = z_h$ and $\widehat{\lambda}_h = \lambda_h$. Substituting expression (56) into expression (55), we obtain

$$\begin{aligned} \int_\Omega \nabla u_h \cdot \nabla z_h \, dx &= \int_\Gamma \lambda_h (u_h - g) \, d\Gamma + \int_\Omega f z_h \, dx \\ &= \int_\Gamma \lambda_h (u_h - P_h^\gamma g) \, d\Gamma + \int_\Omega f z_h \, dx, \end{aligned}$$

where in the last equality we introduce $P_h^\gamma g \in \gamma V^h$ as the solution of (51). Since the choices of z_h and λ_h were arbitrary, it follows that both (50a) and the boundary condition (50b) must be satisfied. □

The difference between problems (41) and (46) lies in the use of an approximated linear form G_h in problem (46). It is therefore crucial to analyze the error that the use of G_h introduces. Lemma 5.4, which estimates $(G - G_h)$, needs the following discrete extension result.

LEMMA 5.3. *There exists a $C > 0$, independent of $h > 0$, such that for each $g_h \in \gamma V^h$, a $u_h \in V^h$ exists satisfying $u_h|_\Gamma = g_h$ and*

$$\|u_h\|_1 \le C \|g_h\|_{1/2,\Gamma}.$$

Proof. Let $g_h \in \gamma V^h$ be given. The functions in γV^h are continuous and piecewise linear on the boundary, so $\gamma V^h \subset H^{1/2}(\Gamma)$ (in fact, $\gamma V^h \subset H^1(\Gamma)$). By Theorem 1.5.1.3 of Grisvard [11], the trace map $\gamma : H^1(\Omega) \to H^{1/2}(\Omega)$ has a right continuous inverse E. The Scott and Zhang interpolator Π_h, discussed in section 3.3, continuously maps functions in $H^1(\Omega)$ into V^h. Composing Π_h and E, we define $u_h \in V^h$ such that $u_h = \Pi_h E g_h$. Note that $\gamma u_h = g_h$ since the Scott and Zhang interpolator preserves piecewise-polynomial boundary conditions. Moreover, since both E and Π_h are continuous, we find that

$$\|u_h\|_1 \leq C\|g_h\|_{1/2,\Gamma}. \qquad \square$$

Remark 1. Similar results are reported in the estimate (5.5) of Scott and Zhang [17] and in Lemma 11 of Gunzburger and Hou [13].

LEMMA 5.4. *Assume a quasi-uniform triangulation, characterized by the mesh parameter h, of the polyhedral (or polygonal) domain Ω having a Lipschitz boundary. There are an $\epsilon \in (0, 1/2]$ and a $C > 0$ such that, given $g \in L^2(\Gamma)$ and $f \in L^2(\Omega)$, the linear forms G and G_h, defined in expressions* (32) *and* (45)*, satisfy, for each $h > 0$,*

$$(G - G_h)(v) \leq C\left(h^\epsilon \|g\|_{0,\Gamma} + h^{1/2+\epsilon}\|f\|_{-1}\right)\|v\|_0 \qquad \forall v \in L^2(\Omega).$$

Proof. Let $v \in L^2(\Omega)$ be given, and let z and λ be the solutions to (30) and (31) associated with the given v. Likewise, let z_h and λ_h be the solutions to (43) and (44) associated with v. By definitions (32) and (45), we find that

$$\begin{aligned}(G - G_h)(v) &= -\int_\Gamma g(\lambda - \lambda_h)\,d\Gamma + \int_\Omega f(z - z_h)\,dx \\ &= -\int_\Gamma (\lambda - \lambda_h)P_h^\gamma g\,d\Gamma - \int_\Gamma \lambda(g - P_h^\gamma g)\,d\Gamma + \int_\Omega f(z - z_h)\,dx,\end{aligned} \tag{57}$$

introducing $P_h^\gamma g$, the $L^2(\Gamma)$ projection of g on γV^h, defined as in (26b).

We will estimate each of the terms in the right-hand side of expression (57), starting with the first. Let $\varphi_h \in \gamma V^h$ be given. Choose $\phi_h \in V_h$ so that $\phi_h|_\Gamma = \varphi_h$ and so that the estimate of Lemma 5.3 is satisfied. From (31) and (44) it follows that

$$\begin{aligned}-\int_\Gamma (\lambda - \lambda_h)\,\varphi_h\,d\Gamma &= -\int_\Omega \nabla(z - z_h)\cdot\nabla\phi_h\,dx \\ &\leq \|z_h - z\|_1\|\phi_h\|_1 \leq C\|z_h - z\|_1\|\varphi_h\|_{1/2,\Gamma} \\ &\leq C\|z_h - z\|_1\, h^{-1/2}\|\varphi_h\|_{0,\Gamma} \leq Ch^{1/2+\epsilon}\|v\|_0\, h^{-1/2}\|\varphi_h\|_{0,\Gamma} \\ &= Ch^\epsilon\|v\|_0\|\varphi_h\|_{0,\Gamma},\end{aligned} \tag{58}$$

where the second inequality follows from Lemma 5.3 and the third from inverse estimates (29) (the inverse estimate needs the assumption of quasi-uniform mesh refinements); expression (40) yields the existence of an $\epsilon \in (0, 1/2]$ such that the fourth inequality holds.

Next we estimate the second term in the right-hand side of expression (57) as

follows:

$$\begin{aligned}\int_\Gamma \lambda(g - P_h^\gamma g)\,d\Gamma &= \sum_{i=1}^I \int_{\Gamma_i} \lambda(g - P_h^\gamma g)\,d\Gamma \le \sum_{i=1}^I \|\lambda\|_{\epsilon,\Gamma_i}\|g - P_h^\gamma g\|_{-\epsilon,\Gamma_i} \\ &\le C\|v\|_0 \sum_{i=1}^I \|g - P_h^\gamma g\|_{-\epsilon,\Gamma_i} \le C\|v\|_0\|g - P_h^\gamma g\|_{-\epsilon,\Gamma} \\ &\le Ch^\epsilon \|v\|_0\|g\|_{0,\Gamma},\end{aligned} \tag{59}$$

where the second, third, and fourth inequalities use estimates (16), (9), and (28), respectively. The third term in the right-hand side of expression (57) is estimated by expression (40),

$$\int_\Omega f(z - z_h)\,dx \le \|f\|_{-1}\|z - z_h\|_1 \le Ch^{1/2+\epsilon}\|f\|_{-1}\|v\|_0. \tag{60}$$

Substituting estimates (58) (with $\varphi_h = P_h^\gamma g$), (59), and (60) into expression (57) yields the required estimate

$$\begin{aligned}(G - G_h)(v) &\le C\left(h^\epsilon \|P_h^\gamma g\|_{0,\Gamma} + h^\epsilon\|g\|_{0,\Gamma} + h^{1/2+\epsilon}\|f\|_{-1}\right)\|v\|_0 \\ &\le C\left(h^\epsilon\|g\|_{0,\Gamma} + h^{1/2+\epsilon}\|f\|_{-1}\right)\|v\|_0,\end{aligned}$$

where in the second inequality we have used the bound $\|P_h^\gamma g\|_{0,\Gamma} \le \|g\|_{0,\Gamma}$ that holds for an L^2-projection. ∎

The final result of this article is that the solution to (46) converges to the solution of (35) at a rate that depends, through ϵ, on the shape of the domain, where $\epsilon = 1/2$ corresponds to a convex domain.

THEOREM 5.5. *Assume a quasi-uniform triangulation, characterized by the mesh parameter h, of the polyhedral (or polygonal) domain Ω having a Lipschitz boundary. There are an $\epsilon \in (0, 1/2]$ and a $C > 0$ such that, given $g \in L^2(\Gamma)$ and $f \in L^2(\Omega)$, the solutions u and u_h to problems* (35) *and* (46) *satisfy, for each $h > 0$,*

$$\|u - u_h\|_0 \le C\left(h^s\|u\|_s + h^\epsilon\|g\|_{0,\Gamma} + h^{1/2+\epsilon}\|f\|_{-1}\right)$$

for each $s \in [0, \epsilon)$.

Proof. The solution error may be decomposed as $u - u_h = (u - \widetilde{u}_h) + (\widetilde{u}_h - u_h)$, where $\widetilde{u}$ is the solution to problem (41). Thus,

$$\|u - u_h\|_0 \le \|u - \widetilde{u}_h\| + \|\widetilde{u}_h - u_h\|. \tag{61}$$

By estimate (42), approximation property (24b), and Theorem (4.2), there exists an $\epsilon \in (0, 1/2]$ so that

$$\|u - \widetilde{u}_h\|_0 \le Ch^s\|u\|_s \qquad \forall s \in [0, \epsilon). \tag{62}$$

Equations (41) and (46) yield that

$$\int_\Omega (\widetilde{u}_h - u_h)v_h\,dx = G(v_h) - G_h(v_h) \qquad \forall v_h \in V^h,$$

so choosing $v_h = \widetilde{u}_h - u_h$ means that

$$\|\widetilde{u}_h - u_h\|_0^2 = (G - G_h)(\widetilde{u}_h - u_h). \tag{63}$$

Using Lemma 5.4 in expression (63) implies that

$$\|\widetilde{u}_h - u_h\|_0 \leq C \left(h^{\epsilon} \|g\|_{0,\Gamma} + h^{1/2+\epsilon} \|f\|_{-1} \right). \tag{64}$$

Substituting estimates (62) and (64) into expression (61) provides the required estimate. ∎

Remark 2. Theorem 5.5 only provides convergence rates for boundary data in $L^2(\Gamma)$. By the equivalence proven in Theorem 5.2, smoother data will improve the convergence rate, since error estimates for the standard approach then apply. Fix, Gunzburger, and Peterson [6], French and King [7, 8], and Bramble and King [2] prove various error estimates that apply for smoother data.

Acknowledgments. I am grateful to my former advisor Roland Glowinski, who once told his surprised student that the ordinary finite-element approximation converges, although in a weaker sense, also for rough data. This comment sparked an interest, eventually leading to the investigations reported here. For careful readings of the manuscript, which uncovered many issues in need of improvement, I owe much to Ridgway Scott. I also thank Max Gunzburger, Steven Hou, and Christer Kiselman for helpful discussions.

REFERENCES

[1] I. Babuška, *Error-bounds for finite element method*, Numer. Math., 16 (1971), pp. 322–333.
[2] J. H. Bramble and J. T. King, *A robust finite element method for nonhomogeneous Dirichlet problems in domains with curved boundaries*, Math. Comp., 63 (1994), pp. 1–17.
[3] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[4] M. Chevalier, M. Högberg, M. Berggren, and D. S. Henningson, *Linear and nonlinear optimal control in spatial boundary layers*, in Proceedings of the AIAA 3rd Theoretical Fluid Mechanics Meeting, St. Louis, MO, AIAA Paper 2002-2755, 2002.
[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[6] G. J. Fix, M. D. Gunzburger, and J. S. Peterson, *On finite element approximations of problems having inhomogeneous essential boundary conditions*, Comput. Math. Appl., 9 (1983), pp. 687–700.
[7] D. A. French and J. T. King, *Approximation of an elliptic control problem by the finite element method*, Numer. Funct. Anal. Optim., 12 (1991), pp. 299–314.
[8] D. A. French and J. T. King, *Analysis of a robust finite element approximation for a parabolic equation with rough boundary data*, Math. Comp., 60 (1993), pp. 79–104.
[9] R. Glowinski and J. L. Lions, *Exact and approximate controllability for distributed parameter systems*, Acta Numer., 1994, pp. 269–378.
[10] R. Glowinski and J. L. Lions, *Exact and approximate controllability for distribtuted parameter systems*, Acta Numer., 1995, p. 159–333.
[11] P. Grisvard, *Elliptic Problems in Non-Smooth Domains*, Pitman, London, 1985.
[12] P. Grisvard, *Singularities in Boundary Value Problems*, Masson, Paris, and Springer-Verlag, Berlin, 1992.
[13] M. D. Gunzburger and S. L. Hou, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424.
[14] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[15] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972.
[16] P. J. Schmid and D. S. Henningson, *Stability and Transition in Shear Flows*, Appl. Math. Sci. 142, Springer-Verlag, New York, 2000.
[17] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[18] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, 1980.

SIAM J. NUMER. ANAL.
Vol. 42, No. 2, pp. 878–904

INTEGRATION OF PERTURBED INITIAL VALUE PROBLEMS THROUGH REDUCTION THEORY*

JESÚS PALACIÁN†

Abstract. We propose a method to integrate an initial value problem formed by the differential equations of a perturbed dynamical system plus the initial condition. The approach consists of several steps. First of all, if the original system does not enjoy a continuous symmetry, it is transformed up to a certain order of approximation M into an equivalent but equivariant one after truncation at order M. Second, this new symmetry allows us to define a reduction map which produces in its whole phase space a splitting of the transformed equations. Specifically, the latter system is decoupled into two subsystems, also initial value problems, one of them defined in the orbit space associated with the reduction and the other linear in its coordinates. As a third step the reduced subsystems are analytically resolved or numerically integrated using standard methods. Finally, the solution of the original system is recovered by inverting back the reduction map and thereafter the analytic transformations. We apply our procedure to three examples.

Key words. initial value problems, Lie transformations, generalized normal forms, reduced equations, invariant theory, orbit spaces, Lie groups, semianalytic integration, numerical simulation

AMS subject classifications. 34C14, 34C20, 34C27, 37C80, 37M05, 65L05

DOI. 10.1137/S0036142903420996

1. Introduction.

The goal of this paper is the analysis of initial value problems composed by ordinary differential equations (ODEs) of the type

$$D\boldsymbol{x}(t) = \sum_{i=0}^{L} \frac{\varepsilon^i}{i!} \boldsymbol{F}_i(\boldsymbol{x}(t)), \tag{1.1}$$

where $t \in I \subseteq \boldsymbol{R}$ stands for the independent variable, $\boldsymbol{x} \in \boldsymbol{R}^m$, ε is a small parameter, and, for each $0 \leq i \leq L$, $\boldsymbol{F}_i$ represents a smooth vector field with m components defined in an open set $\Omega \subseteq \boldsymbol{R}^m$. Symbol D represents a differential operator, and thus $D\boldsymbol{\eta}(s) = d\boldsymbol{\eta}/ds$; see, for instance, [20]. The initial condition we add is written as $\boldsymbol{x}(t_0) = \boldsymbol{x}_0 \in \boldsymbol{R}^m$, with $t_0 \in I$. Since L can be understood as the degree reached by the Taylor series of a certain smooth vector field with respect to ε, it could be infinity, at least formally speaking.

In the context of perturbation theory, systems modeled by an ODE such as (1.1) are called perturbed dynamical systems; see, for example, [34]. Indeed, these equations are formed by the sum of a principal part $\boldsymbol{F}_0$ plus a small and regular perturbation. Our purpose is to give a methodology for the analytical or numerical approximation of system (1.1) with adequate initial values, taking advantage of previous manipulations of the equations.

First of all, the initial value problem is converted into a simpler one, called a generalized normal form, by means of formal changes of variables: "generalized" meaning that we extend the standard approach for smooth vector fields based on simplifications through normal forms in a sense that will be specified later on and "simpler" indicating

*Received by the editors January 9, 2003; accepted for publication (in revised form) October 7, 2003; published electronically July 2, 2004. This work was partially supported by Project ESP99-1074-C02-01 of Ministerio de Educación y Ciencia, Spain.
http://www.siam.org/journals/sinum/42-2/42099.html
†Departamento de Matemática e Informática, Universidad Pública de Navarra, 31006 Pamplona, Navarra, Spain (palacian@unavarra.es).

that in the process of the transformation we introduce a formal (continuous) symmetry in the equations; in other words, we make the transformed system equivariant.

Afterwards, we define a reduction map and make use of a decomposition lemma (or splitting lemma) so that the transformed vector field can be split into two subsystems defined in two different invariant spaces. One of the subsystems, the so-called reduced system, contains the fundamental dynamics of the original system and is defined in a quotient space (called an orbit space; see [1, 35, 21, 13, 7]), whereas the other is a system of linear differential equations with time-dependent coefficients and is defined in a Lie group. Actually, the reduction can be performed thanks to the formal symmetry introduced through the normal form approach.

As the initial vector $\boldsymbol{x}(t_0)$ is also transformed using the change of coordinates, we build two initial value problems, one defined in the orbit space and the other on the Lie group. More precisely, the initial value problem defined in the orbit space is a differential system of equations with constraints among the unknowns. These restrictions appear in the reduction process as a result of constructing the orbit space.

Once we write the subsystems in the appropriate coordinates we have to solve both differential equations with analytical—if possible—or numerical methods. Note that in some cases the reduction is so drastic that we can arrive at two initial value problems that can be solved straightforwardly by means of quadratures, as we shall see in the first and second examples of section 4.

The solution of the original initial value problem is recovered through two steps. First, we shall pass from the vector-valued solutions obtained in the orbit space and in the Lie group to the extended normal form system solving a system of smooth equations (in many cases algebraic equations, as we shall discuss later). Second, the solution of the departure system will be determined by using the inverse change of coordinates.

Our implementation uses Lie transformations for differential equations. The method was introduced by Kamel [27] and Henrard [25]; see also the basic paper by Deprit [15] for Hamiltonian systems. The advantage of working with Lie transformations is that the computations are done in an ascendent way (starting at first order, e.g., with terms factored by ε), arriving at the desired order, say M, where the changes of coordinates—direct and inverse changes—can be obtained explicitly. This feature is in contrast to other techniques based on near-identity transformations such as the classical Poincaré method [43] or the Krylov, Bogoliubov, and Mitropolsky theory of averaging [29, 5], as these procedures provide the formula in mixed coordinates (the original and the transformed ones), and so an inversion of the resulting series is needed. Note, however, that the Lie transformations of Kamel, Henrard, and Deprit and the theory of averaging are based on and inspired by the work of Poincaré.

Here we use the setting given by Meyer [34] through his general perturbation theorem. At this point we emphasize that our procedure is global in the sense that we do not use local expansions around equilibrium points. Nevertheless, the convergence of the transformations is not discussed through the paper, though it is well known that generally transformations based on normal form techniques diverge. Basically, in the setting of smooth vector fields a convergent transformation can be guaranteed if there is a nontrivial local one-parameter group of symmetries; see [10, 52] and the recent book by Cicogna and Gaeta [9]. For polynomial vector fields and the Poincaré–Dulac normal form approach, strong hypotheses of the eigenvalues of the linear part are used to obtain convergence of normal forms. In this framework it is worth mentioning the works by Poincaré [42], Siegel [47], Markhashov [31], and Bryuno [6].

The connection of the general perturbation theorem with the reduction of a dynamical system through the introduction of symmetries has been given for polynomial vector fields in [40]; see also a previous paper by Cicogna and Gaeta [8] dealing with a simpler version. Here we enlarge those studies, considering smooth vector fields and reduction techniques for the case that the Lie group related to the formal symmetry introduced is finitely generated, as we shall detail in section 3. The extension to nonpolynomial vector fields is justified by the use of reduction techniques from the point of view of global analysis of dynamical systems. As examples of global analysis of dynamical sytems based on normal forms of nonpolynomial vector fields we mention the case of perturbed Keplerian systems; see Deprit [16], Cushman [12], Barrio and Palacián [4] and the references therein. Note that in these cases the authors deal with normal forms avoiding collision trajectories, and therefore the perturbations of the two-body problem are smooth functions (but are, in general, nonpolynomial).

Possible fields of application of our approach for perturbed differential equations are classic and quantum mechanics and astrodynamics. The different versions of the n-body problem (general setting or restricted n-body problems; see, for instance, [48]), the motion of artificial satellites around a planet taking into account the gravity potential of the planet (which depends on its shape) [11], or the orbital motion of the Moon [17] are examples of very complex problems to be treated numerically. In classical and quantum mechanics we cite the examples of motion of electrons under the influence of electrical or magnetic fields [18] and the case of ion traps modeled by resonant Hamiltonians with n degrees of freedom [22]. However, generically in these problems, poor numerical approximations are obtained. There are various reasons for this; we mention two: (i) usually very long time spans are required in the simulations; (ii) an initial value problem can depend on external parameters and the corresponding solutions when fixing different values of the parameters and can be very sensitive with respect to them. In this case, a numerical method often exhibits a blind behavior and is not useful for dealing with the solution of the problem. Other sources of problems when dealing with numerical methods are related to the involved expressions defining the original vector field; on other occasions the system of differential equations has a high dimension, and it leads to using very expensive methods from a computational point of view. Finally, if the equations to be resolved are expected to present a chaotic behavior, a numerical treatment can yield unsatisfactory results.

A common idea present in many perturbed ODEs of mechanics [29, 5, 44] is that these problems can be dramatically simplified due to the fact that they are formulated in such a way that some of the variables are angles which oscillate very rapidly in comparison with other angles. Terms of the equation related to the "fast" angles are called short-period terms, whereas terms related to the "slow" ones are called long-period terms. However, short- and long-period terms appear coupled in most applications. If one is interested in the solution of a problem for a long time—for instance to extract information about the numerical stability of some trajectories—a typical approach consists in converting the original equation into an equivalent one but "eliminating" the fast angles up to a certain order (with an appropriate change of coordinates). Thus, one arrives at a simplified problem which can be studied using long steps of time. This is the theory (or the method) of averaging in its modern version formulated by Krylov, Bogoliubov, and Mitropolsky. Using these techniques, Laskar [30] has analyzed the stability of the solar system for a time interval of 10^9 years; see also the review paper by Marmi [32]. A rigorous analysis of artificial satellites orbiting at low altitude has to take into account the oblateness of the Earth

and the atmospheric friction. In [28] a numerical ODE-solver is proposed based on previous averaging of the initial equations. A semianalytical theory [3, 4] has been applied to simplify the original problem, transforming it into another one whose terms are only of long-period type. Hence, the simplified system is numerically solved using a Runge–Kutta method of order eight due to Dormand and Prince and equipped with variable stepsize and continuous output [23].

Nevertheless, to our knowledge, the approach described above has not been followed by the community of applied mathematicians. Besides, the simplification of a system through formal changes of variables can be generalized with the aim of considering not only the case of angular variables but other types of coordinates. This is, consequently, better formulated by the use of reduction to the orbit space (see the seminal paper by Michel [35] and the recent book by Chossat and Lauterbach [7]) but adding the equation in the Lie group. In section 3 we shall show that the averaging method is a particular case of the reduction method we use. Indeed, the variables "eliminated" are the coordinates of the Lie group; therefore, to recover all the departure variables, one needs to solve the linear system of equations in the Lie group.

In this respect the splitting lemma (that is, our splitting decomposition of the transformed equations), together with the relation of the transformed system with the original one, is the essential tool to achieve the solution of the initial ODE via normal forms. In [37] the standard treatment of building formal symmetries of smooth vector fields is extended with the aim of constructing some invariant manifolds of the original ODE using Lie transformations, invariant theory, and reduction techniques. However, the integration of a certain initial value problem such as (1.1) using splitting decompositions is also new.

It can happen that the integration—numerical or analytical—of the corresponding systems in the orbit space and in the Lie group becomes a difficult task. This feature can be caused because the constraints defining the orbit space are involved expressions or due to the fact that the differential system in the orbit space is of high dimension; note that the number of variables in the orbit space can be bigger than the number of original coordinates. Under those circumstances our approach is not recommended, and one must resort to numerical integrations. However, on many interesting occasions, the use of Lie transformations and the subsequent reduction simplify the problem enormously, as we will see in section 4.

The paper has five sections. Section 2 contains the required setting for the generalized normal form formalism. Section 3 contains the central part of the paper and is devoted to the solution of the reduced equations and to describe how one can obtain the solution of the original system. We start this section concretely by describing the geometrical aspects of the reduction after the application of generalized normal forms, dealing with the invariants related to the symmetry introduced by the Lie transformations. We continue showing how the reduced phase spaces are constructed and how the original differential equation is decomposed into the corresponding systems. After that we relate the reduction techniques used with the integration of initial value problems related to such equations. In section 4 we illustrate the technique with three examples. Finally, in section 5 the conclusions of our work are outlined.

2. Analytic reduction.

2.1. Lie transformations for vector fields. Meyer's approach to the calculation of formal symmetries employs Lie transformations and is based on previous work by Kamel [27] and Henrard [25]. In [34], Meyer presents a Lie transformation

treatment in the context of tensor fields. We start by recalling the method applied to smooth vector fields.

Let us consider the differential system

$$D\boldsymbol{x}(t) = \boldsymbol{F}_0(\boldsymbol{x}(t)) + \sum_{i=1}^{\infty} \frac{\varepsilon^i}{i!} \boldsymbol{F}_i(\boldsymbol{x}(t)), \tag{2.1}$$

where t represents the time variable, $\boldsymbol{x} \in \boldsymbol{R}^m$, and ε stands for a dimensionless small parameter. The vector field $\boldsymbol{F}_i$, for $i \geq 0$, has m components, which are analytic functions in $\boldsymbol{x}$, and we truncate the series (2.1) at order L; then $\boldsymbol{F}_i \equiv 0$ for $i > L$. We define by $[\cdot, \cdot]$ the Lie bracket of two vector fields $\boldsymbol{g}_1(\boldsymbol{x})$ and $\boldsymbol{g}_2(\boldsymbol{x})$ in $\boldsymbol{R}^m$, that is, $[\boldsymbol{g}_1, \boldsymbol{g}_2] = D\boldsymbol{g}_1(\boldsymbol{x})\boldsymbol{g}_2 - D\boldsymbol{g}_2(\boldsymbol{x})\boldsymbol{g}_1$.

Let us describe the typical algorithm of Lie transformations for ODEs. A smooth vector field (2.1) depending on a small parameter ε is transformed into another vector field

$$D\boldsymbol{y}(t) = \boldsymbol{G}_0(\boldsymbol{y}(t)) + \sum_{i=1}^{\infty} \frac{\varepsilon^i}{i!} \boldsymbol{G}_i(\boldsymbol{y}(t)), \tag{2.2}$$

where $\boldsymbol{G}_0(\boldsymbol{y}(t)) \equiv \boldsymbol{F}_0(\boldsymbol{x}(t))$, through a generating function

$$\boldsymbol{W}(\boldsymbol{x}; \varepsilon) = \sum_{i=0}^{\infty} \frac{\varepsilon^i}{i!} \boldsymbol{W}_{i+1}(\boldsymbol{x}),$$

following the recursive formula

$$\boldsymbol{F}_i^{(j)} = \boldsymbol{F}_{i+1}^{(j-1)} + \sum_{k=0}^{i} \binom{i}{k} \big[\boldsymbol{F}_{i-k}^{(j-1)}, \boldsymbol{W}_{k+1}\big], \tag{2.3}$$

with $i \geq 0$, $j \geq 1$. Besides, $\boldsymbol{F}_i^{(0)} \equiv \boldsymbol{F}_i$ and $\boldsymbol{F}_0^{(i)} \equiv \boldsymbol{G}_i$ for all $i \geq 0$.

Note that $\boldsymbol{W}(\boldsymbol{x}; \varepsilon)$ is conserved under the transformation, and thus it can also be expressed as $\boldsymbol{W}(\boldsymbol{y}; \varepsilon)$, that is, $\boldsymbol{W}(\boldsymbol{x}; \varepsilon) \equiv \boldsymbol{W}(\boldsymbol{y}; \varepsilon)$. Hence, (2.3) yields the partial differential identity

$$\mathcal{L}_{\boldsymbol{F}_0}(\boldsymbol{W}_i) + \boldsymbol{G}_i = \widetilde{\boldsymbol{F}}_i, \tag{2.4}$$

where $\widetilde{\boldsymbol{F}}_i$ collects all the terms known from the previous orders plus $\boldsymbol{F}_i$. In this identity, called the homology equation, $\boldsymbol{W}_i$ and $\boldsymbol{G}_i$ are determined according to the specific requirements of the Lie transformation one performs. Besides, $\mathcal{L}_{\boldsymbol{F}_0}$ denotes the Lie operator associated with the Lie bracket of two vector functions, i.e., given two vector fields $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$: $\mathcal{L}_{\boldsymbol{g}_1}(\boldsymbol{g}_2) = [\boldsymbol{g}_2, \boldsymbol{g}_1]$.

The transformation $\boldsymbol{x} = \boldsymbol{X}(\boldsymbol{y}; \varepsilon)$ relates the "old" variables $\boldsymbol{x}$ with the "new" ones $\boldsymbol{y}$ and is a near-identity change of coordinates. The direct change is given by

$$\boldsymbol{x} = \boldsymbol{y} + \sum_{i=1}^{\infty} \frac{\varepsilon^i}{i!} \boldsymbol{y}_0^{(i)}. \tag{2.5}$$

Vectors $\boldsymbol{y}_0^{(i)}$, $i \geq 1$, are calculated recursively with the aid of

$$\boldsymbol{y}_i^{(j)} = \boldsymbol{y}_{i+1}^{(j-1)} + \sum_{k=0}^{i} \binom{i}{k} \big(\boldsymbol{y}_k^{(j-1)}, \boldsymbol{W}_{i+1-k}\big), \tag{2.6}$$

with $i \geq 0$, $j \geq 1$, and $\boldsymbol{y}_i^{(0)} \equiv \mathbf{0}$ for $i \geq 1$, $\boldsymbol{y}_0^{(0)} \equiv \boldsymbol{y}$, and each $\boldsymbol{W}_{i+1-k}$ is written in terms of $\boldsymbol{y}$. Besides, given two vector fields $\boldsymbol{g}_1(\boldsymbol{y})$ and $\boldsymbol{g}_2(\boldsymbol{y})$, the operator $(\boldsymbol{g}_1, \boldsymbol{g}_2)$ is computed as $D\boldsymbol{g}_1(\boldsymbol{y})\boldsymbol{g}_2$. Consequently, (2.5) gives the set of coordinates $\boldsymbol{x}$ in terms of $\boldsymbol{y}$ with the use of the generating function $\boldsymbol{W}$.

Similar formulae can be used to obtain the inverse transformation $\boldsymbol{y} = \boldsymbol{Y}(\boldsymbol{x}; \varepsilon)$, which explicitly reads as

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{i=1}^{\infty} \frac{\varepsilon^i}{i!} \boldsymbol{x}_i^{(0)}. \tag{2.7}$$

Now $\boldsymbol{x}_0^{(0)} \equiv \boldsymbol{x}$, and for $i \geq 1$ vectors $\boldsymbol{x}_i^{(0)}$ are calculated recursively by means of

$$\boldsymbol{x}_i^{(j)} = \boldsymbol{x}_{i-1}^{(j+1)} + \sum_{k=0}^{i-1} \binom{i-1}{k} (\boldsymbol{x}_{i-k-1}^{(j)}, \boldsymbol{W}_{k+1}), \tag{2.8}$$

with $i \geq 1$, $j \geq 0$. This time $\boldsymbol{x}_0^{(i)} \equiv \mathbf{0}$ for $i \geq 1$, the Jacobians appearing in the operators of (2.8) are computed with respect to $\boldsymbol{x}$, and $\boldsymbol{W}_{k+1}$ is also written in $\boldsymbol{x}$.

Formulae (2.5) and (2.6) can be used to transform any function expressed in the old coordinates $\boldsymbol{x}$, say $\boldsymbol{P}(\boldsymbol{x}; \varepsilon) = \sum_{i=0}^{\ell} (\varepsilon^i / i!) \boldsymbol{P}_i(\boldsymbol{x})$, as a function of the new variables $\boldsymbol{y}$, after identifying each $\boldsymbol{P}_i(\boldsymbol{x})$ with $\boldsymbol{y}_i^{(0)}$ (and replacing in the identification $\boldsymbol{x}$ by $\boldsymbol{y}$). The result is the function $\boldsymbol{P}$ in terms of $\boldsymbol{y}$. In a similar fashion, (2.7) and (2.8) should be used to transform any function written in terms of $\boldsymbol{y}$ as a function of the coordinates $\boldsymbol{x}$.

2.2. Generalized normal forms. The above method is formal in the sense that the convergence of the various series is not discussed. Moreover, the series diverge in many applications. However, the first orders of the transformed system can give interesting information, and the process can be stopped at a certain order M. It means that these terms of the series are useful to construct both the transformed vector field and the generating function since they are unaffected by the divergent character of the whole process. Now we are ready to build formal symmetries for vector fields using Lie transformations.

THEOREM 2.1 (generalized normal forms). *Let $M \geq 1$ be given, let $\{\mathcal{P}_i\}_{i=0}^{M}$, $\{\mathcal{Q}_i\}_{i=1}^{M}$, and $\{\mathcal{R}_i\}_{i=1}^{M}$ be sequences of vector spaces of smooth vector fields in $\boldsymbol{x} \in \boldsymbol{R}^m$ defined in a common domain Ω in $\boldsymbol{R}^m$, and let $\boldsymbol{T} \equiv \boldsymbol{T}(\boldsymbol{x})$ be a vector field in some $\{\mathcal{P}_i\}_{i=0}^{M}$ with the following properties:*

(i) $\mathcal{Q}_i \subseteq \mathcal{P}_i$, $i = 1, \ldots, M$;
(ii) $\boldsymbol{F}_i \in \mathcal{P}_i$, $i = 0, 1, \ldots, M$;
(iii) *for any $\boldsymbol{D} \in \mathcal{P}_i$ and any $\boldsymbol{K} \in \mathcal{R}_j$ one has $[\boldsymbol{D}, \boldsymbol{K}] \subseteq \mathcal{P}_{i+j}$ for all $i + j = 1, \ldots, M$;*
(iv) *for any $\boldsymbol{D} \in \mathcal{P}_i$, $i = 1, \ldots, M$, one can find $\boldsymbol{E} \in \mathcal{Q}_i$ and $\boldsymbol{K} \in \mathcal{R}_i$ such that*

$$\boldsymbol{E} = \boldsymbol{D} + [\boldsymbol{F}_0, \boldsymbol{K}] \qquad \text{and} \qquad [\boldsymbol{E}, \boldsymbol{T}] = \mathbf{0}.$$

Then there is a smooth vector field $\boldsymbol{W}$,

$$\boldsymbol{W}(\boldsymbol{x}; \varepsilon) = \sum_{i=0}^{M-1} \frac{\varepsilon^i}{i!} \boldsymbol{W}_{i+1}(\boldsymbol{x}),$$

with $\boldsymbol{W}_i \in \mathcal{R}_i$, $i = 1, \ldots, M$, *such that the change of variables* $\boldsymbol{x} = \boldsymbol{X}(\boldsymbol{y}; \varepsilon)$ *is the general solution of the initial value problem*

$$\begin{aligned} D\boldsymbol{x}(\varepsilon) &= D_1 \boldsymbol{W}(\boldsymbol{x}; \varepsilon), \\ \boldsymbol{x}(0) &= \boldsymbol{y}, \end{aligned}$$

and transforms the convergent vector field

$$\boldsymbol{F}(\boldsymbol{x}; \varepsilon) = \sum_{i=0}^{\infty} \frac{\varepsilon^i}{i!} \boldsymbol{F}_i(\boldsymbol{x})$$

to the convergent vector field

$$\boldsymbol{G}(\boldsymbol{y}; \varepsilon) = \sum_{i=0}^{M} \frac{\varepsilon^i}{i!} \boldsymbol{G}_i(\boldsymbol{y}) + \mathcal{O}(\varepsilon^{M+1}),$$

with $\boldsymbol{G}_i \in \mathcal{Q}_i$ *and* $[\boldsymbol{G}_i, \boldsymbol{T}] = \boldsymbol{0}$, $i = 1, \ldots, M$. *Besides, if* $[\boldsymbol{F}_0, \boldsymbol{T}] = \boldsymbol{0}$, *then* $\boldsymbol{T} \equiv \boldsymbol{T}(\boldsymbol{y})$ *is a formal symmetry of* $\boldsymbol{G}$.

Proof. It appears in [40]. Note that we are requiring that functions $\boldsymbol{E} \in \mathcal{Q}_i$ satisfy $[\boldsymbol{E}, \boldsymbol{T}] = \boldsymbol{0}$. In the above $D_1 \boldsymbol{W}(\boldsymbol{x}; \varepsilon)$ stands for the derivatives of $\boldsymbol{W}$ with respect to $\boldsymbol{x}$, that is, the Jacobian. In general, $D_i \boldsymbol{\eta}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell)$ refers to $d\boldsymbol{\eta}/d\boldsymbol{u}_i$, if $i \in \{1, \ldots, \ell\}$, whereas $D_j \boldsymbol{\omega}(\boldsymbol{s}) = d\boldsymbol{\omega}/ds_j$, using Bourbaki notation [20]. □

If we drop the remainder of $\boldsymbol{G}$ and make $M \to +\infty$, the resulting formal expansion does not converge, in general, as further and deep conditions on the vector field $\boldsymbol{F}$ and on the change of coordinates of the Lie transformation should be added; see, for example, [10, 52].

The number of possible normal forms of $\boldsymbol{F}$ one calculates depends on the different Lie transformations one can execute or, in other words, on the independent symmetries $\boldsymbol{T}$ of $\boldsymbol{F}_0$ selected in order to apply Theorem 2.1. This is in contrast to the usual approach of normal forms in which one chooses $\boldsymbol{T}$ to be $\boldsymbol{F}_0$ or, in the polynomial setting, the semisimple part of $\boldsymbol{F}_0$. However, we allow $\boldsymbol{T}$ to be any symmetry of $\boldsymbol{F}_0$. Hence, the vector field $\boldsymbol{G}$ is called a generalized normal form of $\boldsymbol{F}$; see also [38, 40].

Next, we can determine a formal continuous symmetry of the original system by going back to the departure system. Specifically, we compute $\boldsymbol{T}^*(\boldsymbol{x}; \varepsilon)$ as

$$\boldsymbol{T}^*(\boldsymbol{x}; \varepsilon) = \boldsymbol{T}(\boldsymbol{x}) + \sum_{i=1}^{M} \frac{\varepsilon^i}{i!} \boldsymbol{T}(\boldsymbol{x})_i^{(0)}, \tag{2.9}$$

where $\boldsymbol{T}(\boldsymbol{x})_i^{(0)}$ are calculated using

$$\boldsymbol{T}(\boldsymbol{x})_i^{(j)} = \boldsymbol{T}(\boldsymbol{x})_{i-1}^{(j+1)} + \sum_{k=0}^{i-1} \binom{i-1}{k} (\boldsymbol{T}(\boldsymbol{x})_{i-k-1}^{(j)}, \boldsymbol{W}_{k+1}), \tag{2.10}$$

with $i \geq 1$ and $j \geq 0$. Now $\boldsymbol{T}(\boldsymbol{x})_0^{(0)} \equiv \boldsymbol{T}(\boldsymbol{x})$ and, for $i \geq 1$, $\boldsymbol{T}(\boldsymbol{x})_0^{(i)} \equiv \boldsymbol{0}$. Then $\boldsymbol{T}^*(\boldsymbol{x}; \varepsilon)$ is an asymptotic symmetry of $\boldsymbol{F}$, i.e., $[\boldsymbol{F}, \boldsymbol{T}^*] = \mathcal{O}(\varepsilon^{M+1})$.

We have to note that given a vector field $\boldsymbol{T}$ with $[\boldsymbol{F}_0, \boldsymbol{T}] = \boldsymbol{0}$ it is not always possible in practice to solve the homology equation (2.4) due to the difficulties in finding out the pair $(\boldsymbol{G}_i, \boldsymbol{W}_i)$ satisfying it. Therefore, on some occasions we will stop the computation of a normal form at the order we had reached without trouble.

Estimates of the error committed by the application of Theorem 2.1 are obtained from the theory developed for the method of averaging. Taking into consideration that $\boldsymbol{y} = \boldsymbol{Y}(\boldsymbol{x};\varepsilon)$ and $\boldsymbol{x} = \boldsymbol{X}(\boldsymbol{y};\varepsilon)$, for a given vector field $\boldsymbol{S}(\boldsymbol{x};\varepsilon)$ we call $\boldsymbol{S}^*(\boldsymbol{x};\varepsilon) = \boldsymbol{S}(\boldsymbol{X}(\boldsymbol{Y}(\boldsymbol{x};\varepsilon);\varepsilon);\varepsilon)$; then one can conclude that by using an adequate norm $\|\boldsymbol{S}^*(\boldsymbol{x};\varepsilon) - \boldsymbol{S}(\boldsymbol{x};\varepsilon)\| = \mathcal{O}(\varepsilon^{M+1})$ on a time scale $1/\varepsilon$; see [44] for details. This remark gives the key to know how accurate we get when computing approximate solutions in (2.1). Using these considerations we shall bound the errors in the first and second examples of section 4.

3. Solution of the original equations.

3.1. Decoupling of the transformed equation. From a geometrical standpoint the consequence of introducing a symmetry by making use of Theorem 2.1 is that the dimension of the phase space where the transformed system is defined—the so-called reduced phase space—is reduced from m to s (s denoting the number of functionally independent first integrals associated with $\boldsymbol{T}(\boldsymbol{y})$). Let us see how this is achieved with some details.

System (2.1) is defined over an open subset of $\boldsymbol{R}^m$. This is the phase space of the dynamical system determined by (2.1). Given $\boldsymbol{T}$, an m-dimensional vector field such that $[\boldsymbol{F}_0, \boldsymbol{T}] = 0$, the application of Theorem 2.1 after truncating at order M leads to a smooth vector field that we denote by $\boldsymbol{H}(\boldsymbol{y};\varepsilon)$:

$$D\boldsymbol{y}(t) = \boldsymbol{H}(\boldsymbol{y};\varepsilon) = \sum_{i=0}^{M} \frac{\varepsilon^i}{i!} \boldsymbol{G}_i(\boldsymbol{y}), \tag{3.1}$$

where $\boldsymbol{H}_0 \equiv \boldsymbol{F}_0$, and each $\boldsymbol{G}_i$ is constructed so that $[\boldsymbol{G}_i, \boldsymbol{T}] = \boldsymbol{0}$ for $1 \leq i \leq M$.

We need to show how the transformation performed in section 2 is effective in the sense that we really simplify the departure system. We took inspiration from the result obtained by Gaeta in [21], adapting it to our requirements. Associated with the one-parameter group of symmetries introduced through the Lie transformation there is an $(m-s)$-dimensional Lie group $G_{\boldsymbol{T}}$ such that $\boldsymbol{H}$ is $G_{\boldsymbol{T}}$-equivariant, that is, fixed $\varepsilon > 0$, for any $\boldsymbol{y} \in \boldsymbol{R}^m$ and any $g \in G_{\boldsymbol{T}}$, and the identity $\boldsymbol{H}(\boldsymbol{y};\varepsilon) = \boldsymbol{H}(g\boldsymbol{y};\varepsilon)$ holds.

In [46] Schwarz generalized a result obtained by Hilbert for polynomial first integrals. Specifically, Schwarz showed that given a compact Lie group $G_{\boldsymbol{T}}$ and any $G_{\boldsymbol{T}}$-equivariant vector field, there is a set of smooth functions defined in a domain $\Omega \subseteq \boldsymbol{R}^m$ (in other words, $\mathcal{C}^\infty(\Omega)$-functions) such that any $G_{\boldsymbol{T}}$-equivariant smooth function defined over Ω can be written as a $\mathcal{C}^\infty(\Omega)$-function of those functions. These functions, designated by $\varphi_i(\boldsymbol{y})$, $i = 1, \ldots, r$ (with $\boldsymbol{y} \in \Omega$), correspond to the r linearly independent first integrals of the system $D\boldsymbol{y}(t) = \boldsymbol{T}(\boldsymbol{y}(t))$, from which s (with $1 \leq s \leq r$) are functionally independent. The requisite of $G_{\boldsymbol{T}}$ to be compact cannot be avoided in order to ensure the existence of a finite number of first integrals. However, we can relax somewhat this hypothesis in three important cases:

(i) When $\boldsymbol{T}$ is a homogeneous linear vector field, we do not care about the compactness of the Lie group. Indeed, a classic theorem by Weitzenböck [53] guarantees that the polynomial invariant algebra of the symmetry $\boldsymbol{T}$ (i.e., the set of all polynomial first integrals associated with $\boldsymbol{T}$), denoted by $I(\boldsymbol{T})$, is finitely generated. The semisimple case was treated by Walcher [50, 51], whereas the extension to nonsemisimple vector fields is treated, for instance, in [14]. In summary, if $\boldsymbol{T}$ is a polynomial vector we can always find a finite number of polynomial first integrals, regardless of the compactness of $G_{\boldsymbol{T}}$. Note that even for a nonpolynomial ODE one could obtain a

nonpolynomial normal form using a polynomial vector field $\boldsymbol{T}$ as the symmetry to be extended in the process. Hence in this case the algebra $I(\boldsymbol{T})$ is also finitely generated.

(ii) If $G_{\boldsymbol{T}}$ is a linearly reductive Lie group, Hilbert proved that $G_{\boldsymbol{T}}$ has a finite number of generators, and therefore $I(\boldsymbol{T})$ is finitely generated. A constructive algorithm to compute the required first integrals of $I(\boldsymbol{T})$ is given by Derksen [19].

(iii) If $G_{\boldsymbol{T}}$ is not compact and (i) and (ii) do not hold we can still apply the Frobenius theorem in the neighborhood of any point $\boldsymbol{y} \in \Omega$, finding out a finite set of linearly independent generators from which s are functionally independent. See Theorem 2.17 of the book by Olver [36] for a derivation of these invariants.

Suppose now that we are in the situation that $G_{\boldsymbol{T}}$ is finitely generated and that we have already determined its generators. The set $\{\varphi_1, \ldots, \varphi_r\}$ has the structure of a ring of scalar fields with the standard product and addition of $\mathcal{C}^\infty$-functions. Denote by ${\mathcal{L}^*}_{\boldsymbol{T}}(z)$ the Lie derivative of a function $z : \Omega \to \boldsymbol{R}$ related to $\boldsymbol{T}$, e.g., ${\mathcal{L}^*}_{\boldsymbol{T}}(z(\boldsymbol{y})) = \langle Dz(\boldsymbol{y}), \boldsymbol{T}\rangle$. So ${\mathcal{L}^*}_{\boldsymbol{T}}(\varphi_i(\boldsymbol{y})) = 0$ for $i \in \{1, \ldots, r\}$. Hence, the φ_i are the linearly independent solutions of the linear partial differential equation ${\mathcal{L}^*}_{\boldsymbol{T}}(\varphi_i(\boldsymbol{x})) = 0$. Note that $s \leq m$, but r can be any value.

We follow now the construction of Walcher [50] but in the context of smooth functions. First, we make $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_r)$ and $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_{m-s})$. Then we consider the map $\varrho_{\boldsymbol{T}}$ over $\boldsymbol{R}^m$ as follows:

$$\begin{aligned} \varrho_{\boldsymbol{T}} : \Omega \subseteq \boldsymbol{R}^m &\longrightarrow \boldsymbol{R}^r, \\ \boldsymbol{y} &\mapsto \boldsymbol{\varphi}(\boldsymbol{y}). \end{aligned}$$

Let $S(\boldsymbol{R}^r)$ be the associative and graded algebra of all $\mathcal{C}^\infty$-maps from $\boldsymbol{R}^r$ to $\boldsymbol{R}$, and let $J = \{\gamma \in S(\boldsymbol{R}^r) : \gamma(\boldsymbol{\varphi}(\boldsymbol{y})) = 0 \text{ for all } \boldsymbol{y} \in \Omega\}$; then J is a prime ideal in $S(\boldsymbol{R}^r)$ because for $\gamma_1, \gamma_2 \in S(\boldsymbol{R}^r)$ and $\gamma_1 \cdot \gamma_2 \in J$ one has $\gamma_1(\boldsymbol{\varphi}(\boldsymbol{y})) \cdot \gamma_2(\boldsymbol{\varphi}(\boldsymbol{y})) = 0$ for all $\boldsymbol{y} \in \Omega$. Let $Y = \{\boldsymbol{y} \in \Omega : \gamma(\boldsymbol{y}) = 0 \text{ for all } \gamma \in J\}$ be an irreducible variety. It is clear that $\varrho_{\boldsymbol{T}}(\Omega) \subseteq Y$ and $\dim Y = s$ (that is, the maximum number of functionally independent $\mathcal{C}^\infty(\Omega)$-functions on Y), and the rank of $D\varrho_{\boldsymbol{T}}(\boldsymbol{y})$ is equal to s at most points of Ω.

Next, we define the reduction map as the surjective map

$$\begin{aligned} \pi_{\boldsymbol{T}} : \Omega \subseteq \boldsymbol{R}^m &\longrightarrow Y, \\ \boldsymbol{y} &\mapsto \boldsymbol{\varphi}(\boldsymbol{y}). \end{aligned}$$

By construction, the map $\pi_{\boldsymbol{T}}$ is a solution-preserving map from the equivariant vector field $\boldsymbol{H}(\boldsymbol{y}; \varepsilon)$ to a differential equation defined in the reduced space (or reduced phase space) Y and reflects the actual reduction process; see the details in [50]. In addition, Y has the structure of a variety (semialgebraic whenever $\boldsymbol{T}$ is of polynomial character) and is defined as a set of (smooth) equalities and inequalities that we call reduced phase space. However, the passage to the space Y must be combined with an additional differential equation in the Lie group $G_{\boldsymbol{T}}$ to make the splitting explicit. We have the following result.

THEOREM 3.1 (splitting lemma). *Given the generalized normal form system* (3.1) *with* $\boldsymbol{H}$ *a smooth function of* $\boldsymbol{y}$ *and* ε *(fixed) defined in* $\Omega \subseteq \boldsymbol{R}^m$ *and with the necessary assumptions on the smooth vector field* $\boldsymbol{T}$ *to make* $G_{\boldsymbol{T}}$ *finitely generated,* $\boldsymbol{H}$

can be transformed into a triangular system as

$$
\begin{aligned}
D\boldsymbol{\varphi}(t) &= \boldsymbol{\alpha}(\boldsymbol{\varphi}(t);\varepsilon) = \sum_{i=0}^{M} \frac{\varepsilon^i}{i!}\boldsymbol{\alpha}_i(\boldsymbol{\varphi}(t)),\\
D\boldsymbol{\vartheta}(t) &= \boldsymbol{\beta}(\boldsymbol{\vartheta}(t),\boldsymbol{\varphi}(t);\varepsilon) = \sum_{i=0}^{M} \frac{\varepsilon^i}{i!}\boldsymbol{\beta}_i(\boldsymbol{\vartheta}(t),\boldsymbol{\varphi}(t)),
\end{aligned}
\tag{3.2}
$$

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ being smooth functions obtained constructively from $\boldsymbol{H}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\vartheta}$ and having dimensions r and $m-s$, respectively.

Proof. A similar result is proven in [21], but we propose a different version. (See also [1] for details and the seminal paper by Michel [35].) The reason for the appearance of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ comes from the fact that they are constructed order by order in powers of ε. The first equation of (3.2) depends exclusively on the φ_i, is called the reduced system, and is defined over $\boldsymbol{R}^r$ (or, using the reduction map $\pi_{\boldsymbol{T}}$, over Y), whereas the second equation of (3.2) is defined in the Lie group $G_{\boldsymbol{T}}$. We use that $D\boldsymbol{\varphi}(t) = D\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{H}(\boldsymbol{y};\varepsilon)$ and take into account that the right-hand member of this equation can be expressed completely in terms of $\boldsymbol{\varphi}$, as $\boldsymbol{\alpha}$ is $G_{\boldsymbol{T}}$-invariant. Next, we make the identification

$$
\boldsymbol{\alpha}(\boldsymbol{\varphi};\varepsilon) = D\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{H}(\boldsymbol{y};\varepsilon), \quad \text{that is,} \quad \boldsymbol{\alpha}_i(\boldsymbol{\varphi}) = D\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{H}_i(\boldsymbol{y}).
$$

For each i, the construction of $\boldsymbol{\beta}_i$ is done with the aid of $\boldsymbol{\alpha}_i$ and $\boldsymbol{H}_i$. It must be performed once the coordinates $\boldsymbol{\vartheta}$ have been calculated as functions of $\boldsymbol{y}$. Now there is no systematic way to encounter the components of $\boldsymbol{\vartheta}$, as it depends on each particular case. The dimensions of the vector fields $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ follow, respectively, from the dimensions of $\boldsymbol{\varphi}$ and $\boldsymbol{\vartheta}$. □

Theorem 3.1 is also called the splitting decomposition. Note that as there is not a unique set of coordinates, there is not a unique function $\boldsymbol{\beta}$. From a qualitative standpoint the relevant part of the normal form is given by the equation on Y. However, the equation on the Lie group must be taken into account if an integration (either analytical or numerical) of (3.1) is needed. In this case, if the solution of the equation involving the φ_i is known, then the solution of the remaining equation on $G_{\boldsymbol{T}}$ can be obtained. In particular for polynomial normal forms, this equation is linear in $\boldsymbol{\vartheta}$. As there are $r-s$ functionally independent relations among the $\varphi_i(\boldsymbol{y})$, these relations are indeed the constraints determining the phase space where the normal form system in Y is defined. Besides, the basic properties of system (3.1) are also reflected in Y. For instance, asymptotic expressions at a certain order M of the analytic integrals of the initial system must be found from the analysis of the normal form in the reduced space. The invariance of some subsets of $\boldsymbol{R}^m$ is formally preserved when passing to the reduced space; see a proof in [50].

Whenever $G_{\boldsymbol{T}}$ is compact the set Y can be identified with the orbit space $\Omega/G_{\boldsymbol{T}}$, and our reduction process coincides with the reduction to the orbit space because in this case the invariants separate orbits; see, for instance, [7]. However, for noncompact Lie groups the variety Y is topologically better behaved than $\Omega/G_{\boldsymbol{T}}$; see [50].

The relation of the procedure described through Theorem 3.1 and the theory of averaging is rather clear; see, for instance, [5, 44]. Indeed, it is easy to see that the passage from the original equation to the system defined in the reduced phase space can be interpreted as an average of the equation over all "angular" variables ϑ_i since the coordinates of the Lie group are absent in Y. (The latter has sense, provided that no small divisor appears in the generating functions; otherwise a partial averaging

is needed; that is, one needs to choose adequately those angles out of the resonance domain; therefore, a different selection of the symmetry of the principal part is used.) However, the way we have followed seems to be more transparent and general, as the reduction process does not depend on the variables we use, and the coordinates of $G_{\boldsymbol{T}}$ do not need to be actual angles; see some examples in [41, 56].

For Hamiltonian systems we can perform symplectic transformations. In this context there is a scalar function associated with the initial vector field $\boldsymbol{F}$. This function is called a Hamiltonian or Hamilton function and is usually denoted by $\mathcal{H}$. If $\mathcal{J}$ designates the skew-symmetric matrix of order $2n$,

$$\mathcal{J} = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix},$$

one has $\boldsymbol{F} = \mathcal{J} D\mathcal{H}(\boldsymbol{x})$. Moreover, we need to select $\boldsymbol{T}$ (a symmetry of $\boldsymbol{F}_0$) such that it has a Hamiltonian nature. That is, $\boldsymbol{T}$ has to be related to a constant of motion (an integral) that we denote by $\mathcal{T}$ through $\boldsymbol{T} = \mathcal{J} D\mathcal{T}(\boldsymbol{x})$. As $\mathcal{T}$ is constant over the solutions of $\boldsymbol{F}$ one can fix a real value for it, i.e., $\mathcal{T} \equiv c \in I \subseteq \boldsymbol{R}$.

The Lie bracket of two vector fields is replaced by the Poisson bracket of two scalar fields $\mathcal{P}$ and $\mathcal{Q}$; specifically, if $\boldsymbol{x} = (x_1, \ldots, x_{2n})$, the Poisson bracket is defined over an open domain of $\boldsymbol{R}^{2n}$ as the quantity

$$\{\mathcal{P}, \mathcal{Q}\}(\boldsymbol{x}) = \sum_{i=1}^{n} D_i \mathcal{P}(\boldsymbol{x}) D_{i+n} \mathcal{Q}(\boldsymbol{x}) - D_{i+n} \mathcal{P}(\boldsymbol{x}) D_i \mathcal{Q}(\boldsymbol{x}).$$

Note that $(x_1, \ldots, x_n)$ stand for the coordinates, whereas $(x_{n+1}, \ldots, x_{2n})$ refer to their conjugate momenta. Moreover, all computations are carried out in the scalar framework, and the generating function $\boldsymbol{W}$ is related to a scalar $\mathcal{W}$ through $\boldsymbol{W} = \mathcal{J} D\mathcal{W}(\boldsymbol{x})$.

As a consequence, the normal form is by construction a function depending exclusively on $\boldsymbol{q}$ [38]. Thus, the second equation of system (3.2) gets reduced to $D\boldsymbol{p}(t) = \boldsymbol{\beta}(\boldsymbol{q}(t); \varepsilon)$. Observe that the variables $\boldsymbol{q}(t)$ are not canonical; however, the vector field $\boldsymbol{\beta}$ admits a symplectic structure because the generating function is also a Hamiltonian. This time the reduction is done adding an extra step. First of all, Theorem 3.1 is applied, and $\boldsymbol{q}$ and $\boldsymbol{\alpha}$ are calculated. Consequently, if a departure Hamiltonian defines a dynamical system on a $2n$-dimensional phase space, that is, a system of n degrees of freedom, after a symplectic reduction, the resulting Hamiltonian lies on a phase space of dimension s if s is even or of dimension $s-1$ if s is odd. Strictly speaking, there is an infinite number of reduced phase spaces, one for each value of $c \in I \subseteq \boldsymbol{R}$. See [37] for an application to the Hénon and Heiles family of Hamiltonians with two degrees of freedom.

Several reductions of a departure system can be performed successively. Indeed, if $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_k$ correspond to k functionally independent vector fields commuting with the principal part of a dynamical system such as (2.1), it is possible (at least theoretically) to apply up to k different reductions (conversions to normal forms followed by the passage to the corresponding invariants and the splitting decompositions). Thus an originally m-dimensional system ($m \geq 3$) could be reduced to a system of dimension one. We shall see an example of this in section 4.3. However, in practice it is quite unlikely to execute more than one transformation due to the difficulty in solving the homology equation in the Lie transformation.

The coordinates of the orbit space receive also the name of generators of the reduced phase space and are indeed the r linearly independent first integrals related

to $\boldsymbol{T}$. Since the dimension of the quotient space $\Omega/G_{\boldsymbol{T}}$ is s, there are s functionally independent invariants. However, the number r of algebraically independent invariants cannot be obtained in a systematic manner, and it depends on each reduction; that is, it is determined by the choice of the vector field $\boldsymbol{T}$, but $r \geq s$ is always satisfied. There are at least $r - s$ relations involving the φ_i. These relations are used to define the reduced phase space.

The reduced phase space can have singular points. In particular, the reduction is called regular whenever $\Omega/G_{\boldsymbol{T}}$ is a regular manifold [33], whereas if there is a vector $\boldsymbol{x} \in \boldsymbol{R}^m$ such that its isotropy subgroup is nontrivial, the reduced phase space is a manifold with singularities. That reduction is called singular [2]. See details in [37].

If the reduction is symplectic, there is another possibility of introducing singularities in the reduced phase space. After determining the corresponding invariants and computing the reduced (Hamiltonian) equations up to the desired order, the value of $\mathcal{T}$ has to be fixed to a constant $c \in \boldsymbol{R}$. This constant appears as a parameter in the constraints which define the reduced phase spaces. In other words, one has a parametric family of reduced phase spaces with at least one parameter.

Thus, these reduced phase spaces have different numbers of singularities according to the values the parameter c takes. We stress that this situation cannot be detected by analyzing the corresponding isotropy subgroups. A straightforward way of calculating the singularities consists in parametrizing the reduced phase space and computing thereafter its gradient vector. The new singularities are those points where the gradient vanishes.

3.2. Integration of the reduced equations. System (3.2) is solved analytically if the reduction procedure has simplified the original system enough; otherwise one resorts to numerical schemes.

First of all, we need to add to (3.2) the adequate initial conditions. Notice that $\boldsymbol{x}_0$ must be converted into a vector $\boldsymbol{y}_0$. It is done by constructing first the inverse change $(\boldsymbol{x} \to \boldsymbol{y})$ using formula (2.7) and truncating it at order M. Then $\boldsymbol{x}$ is replaced by $\boldsymbol{x}_0$, and one obtains an m-dimensional vector $\boldsymbol{y}_0$. Then, by means of $\pi_{\boldsymbol{T}}$, the vector $\boldsymbol{q}_0$ is calculated, and, finally, $\boldsymbol{p}_0$ is determined from the explicit expressions of the coordinates ϑ_i.

Associated with the system

$$\begin{aligned} D\boldsymbol{q}(t) &= \boldsymbol{\alpha}(\boldsymbol{q}(t);\varepsilon), \\ \boldsymbol{q}(t_0) &= \boldsymbol{q}_0, \end{aligned} \tag{3.3}$$

one has some algebraic constraints among the φ_i, which are responsible for defining the orbit space. The number of relations to take into account is $r - s$ (it is zero in some cases). This makes (3.3) a differential equation with constraints, which must either be solved analytically or integrated using adequate numerical methods (see, for example, [24] for generic systems and [26] for a specific algorithm in the Hamiltonian context). Note that the method chosen to execute the numerical computation must have a global error of a size similar to $\mathcal{O}(\varepsilon^{M+1})$ in order not to lose accuracy obtained with the analytic transformation.

Once the vector $\boldsymbol{q}(t)$ has been determined (either explicitly by analytic means or approximately with numerical tools) we are ready to solve the linear problem

$$\begin{aligned} D\boldsymbol{p}(t) &= \boldsymbol{\beta}(\boldsymbol{p}(t), \boldsymbol{q}(t);\varepsilon), \\ \boldsymbol{p}(t_0) &= \boldsymbol{p}_0. \end{aligned} \tag{3.4}$$

We remark that at this step $\boldsymbol{q}(t)$ is a known function of time. Sometimes system (3.4) is readily solved analytically, for instance in some one-dimensional systems (then one has $s = m - 1$). On other occasions we apply an explicit Runge–Kutta method maintaining the global error such as $\mathcal{O}(\varepsilon^{M+1})$. Thus, we obtain $\boldsymbol{p}(t)$ either analytically or numerically.

Finally, if the reduction process is in the Hamiltonian context one can make use of numerical methods based upon symplectic integrators (see, for instance, [45, 57]) to approximate the reduced equations.

3.3. Going back to the original equations. The recovering of the original ODE is a crucial step that makes our approach different from others based on averaging techniques, since our goal is to provide the (approximate) solution of a certain initial value problem. Thus, we need to pass to the initial set of variables through the inversion of the reduction map and the truncated change of coordinates, as we detail now.

We obtain $\boldsymbol{y}(t)$ from $\boldsymbol{q}(t)$ and $\boldsymbol{p}(t)$. This is achieved by solving a system of equations usually of nonlinear nature. In general, in many applications, the coordinates φ_i and ϑ_i are easily expressible in terms of $\boldsymbol{y}$, and this step is immediate (see, for instance, the examples of [39]). However, it could occur that a Newton (or quasi-Newton) method would be needed to obtain each $y_i(t)$. This would make the whole method rather complicated, and one would probably resort to applying numerical integrations from the very beginning.

From $\boldsymbol{y}(t)$ we make use of the direct change of coordinates given by (2.5) to determine $\boldsymbol{x}(t)$. This is the last step of the entire process.

4. Applications. All the calculations we present in the three examples have been done with the numerical and analytical routines of Mathematica, Version 4.2 [55] on a PowerMac G4. The formulae obtained have been translated to LaTeX using the Mathematica function TeXForm[formulae], so as to avoid mistakes in the transcription process.

4.1. Stiff perturbed ODE in two dimensions. Let the following planar nonlinear perturbed initial value problem be given:

$$\begin{aligned} Dx_1(t) &= -2x_1 + x_2 + \varepsilon(-2x_1^2 + x_2^2),\\ Dx_2(t) &= 998x_1 - 999x_2 + \varepsilon(x_1^2 + x_1x_2 + x_2^2), \end{aligned} \tag{4.1}$$

with $x_1(0) = 1/2$, $x_2(0) = 3/5$, and $\varepsilon = 10^{-4}$. As the eigenvalues of the linear part of (4.1) are $\lambda_1 = -1000$ and $\lambda_2 = -1$, the equation is clearly stiff.

We start by using an analytic method to approximate the solution. First, we make a linear transformation of the original equation so as to diagonalize the linear part of it. This is achieved by defining $y_1 = (-x_1 + x_2)/999$ and $y_2 = (998x_1 + x_2)/999$. The new system, in short $D\boldsymbol{y}(t) = \boldsymbol{f}(\boldsymbol{y};\varepsilon)$, reads as

$$\begin{aligned} Dy_1(t) &= -1000y_1 + \varepsilon(996002y_1^2 + 2000y_1y_2 - y_2^2),\\ Dy_2(t) &= -y_2 + 3\varepsilon(331669y_1^2 + 997y_1y_2 + y_2^2), \end{aligned} \tag{4.2}$$

with initial conditions $y_1(0) = \frac{1}{9990}$ and $y_2(0) = \frac{2498}{4995}$.

Now the linear part of (4.2) can be written as $A\boldsymbol{y}$ with $A = \operatorname{diag}\{-1000, -1\}$ and $\boldsymbol{y} = (y_1, y_2)^t$. We choose now the vector field $\boldsymbol{T}(\boldsymbol{y}) = A\boldsymbol{y}$. Clearly, $\boldsymbol{T}$ represents a symmetry of the linear part because $[A\boldsymbol{y}, \boldsymbol{T}(\boldsymbol{y})] = \boldsymbol{0}$. Our purpose is to extend

this symmetry to higher orders by applying an adequate Lie transformation, passing from $\boldsymbol{y}$ to the new coordinates $\boldsymbol{z} = (z_1, z_2)^t$. Let us denote by $\boldsymbol{w}(\boldsymbol{y}) \equiv \boldsymbol{w}(\boldsymbol{z}) = (w_1(\boldsymbol{z}), w_2(\boldsymbol{z}))^t$ the generating function, by $\boldsymbol{g}(\boldsymbol{z}) = (g_1(\boldsymbol{z}), g_2(\boldsymbol{z}))^t$ the nonlinear part of the resulting equation, and by $\widetilde{\boldsymbol{f}}(\boldsymbol{z}) = (\widetilde{f}_1(\boldsymbol{z}), \widetilde{f}_2(\boldsymbol{z}))^t$ the nonlinear part of (4.2) (without ε and replacing $\boldsymbol{y}$ by $\boldsymbol{z}$). Now the homology equation (2.4) which has to be solved in the unknowns $\boldsymbol{w}$ and $\boldsymbol{g}$ takes the following form:

$$\begin{aligned} 1000w_1(\boldsymbol{z}) - 1000z_1 D_1 w_1(\boldsymbol{z}) - z_2 D_2 w_1(\boldsymbol{z}) &= \widetilde{f}_1(\boldsymbol{z}) - g_1(\boldsymbol{z}), \\ w_2(\boldsymbol{z}) - 1000z_1 D_1 w_2(\boldsymbol{z}) - z_2 D_2 w_2(\boldsymbol{z}) &= \widetilde{f}_2(\boldsymbol{z}) - g_2(\boldsymbol{z}). \end{aligned} \tag{4.3}$$

As $\boldsymbol{g}$ is chosen so that $[\boldsymbol{g}, \boldsymbol{T}(\boldsymbol{z})] = \boldsymbol{0}$, we take $\boldsymbol{g}(\boldsymbol{z}) = (|z_1|^{999/1000} z_2, 0)^t$, and hence the transformed equation after truncation is $D\boldsymbol{z}(t) = \boldsymbol{h}(\boldsymbol{z}; \varepsilon) = A\boldsymbol{z} + \varepsilon \boldsymbol{g}(\boldsymbol{z})$. Observe that due to the trivial form of $\boldsymbol{g}$ we can get the solution of this latter differential equation and deduce from it the solution of $\boldsymbol{y}$ in terms of t. Then we would avoid the passage to the reduced equations and the subsequent study. However, we prefer to continue with the standard treatment in order to show all the steps. The expression for the generating function reads as

$$\begin{aligned} w_1(\boldsymbol{z}) &= -\frac{498001}{500} z_1^2 - 2000 z_1 z_2 - \frac{1}{998} z_2^2 + \frac{1}{1000} |z_1|^{999/1000} z_2 \log |z_1|, \\ w_2(\boldsymbol{z}) &= -\frac{995007}{1999} z_1^2 - \frac{2991}{1000} z_1 z_2 - 3 z_2^2. \end{aligned}$$

This completes the normal form procedure at first order. Note that since the limit of w_1 when z_1 tends to zero is finite, we extend the domain of definition of the generating function and consider also the axis $z_1 = 0$.

The vector spaces $\mathcal{P}_i$ and $\mathcal{R}_i$ are chosen as the spaces of continuous vector fields in $\boldsymbol{R}^2$ whose coefficients are in $\boldsymbol{R}$. For each i, the space $\mathcal{Q}_i$ corresponds to the vector subspace of $\mathcal{P}_i$ for which the Lie bracket $[\boldsymbol{q}, T\boldsymbol{y}] = \boldsymbol{0}$ for any continuous two-dimensional vector field $\boldsymbol{q}$. Of course, if we remove the axis $y_1 = 0$ from our analysis, we can consider all vector spaces smooth functions.

Now we calculate the first integrals and the coordinates of the Lie group arriving at the expressions $\varphi = z_1 z_2^{-1000}$ and $\vartheta = z_2$. Thus $r = s = 1$, and $m - s = 1$. We need to transform the initial conditions $y_1(0)$ and $y_2(0)$ in order to obtain $z_1(0)$ and $z_2(0)$. This is done by means of the inverse Lie transformation (2.7), yielding $\boldsymbol{z} = \boldsymbol{Z}(\boldsymbol{y}; \varepsilon)$, where

$$\begin{aligned} z_1 &= y_1 + \varepsilon \left(\frac{498001}{500} y_1^2 + 2000 y_1 y_2 + \frac{1}{998} y_2^2 - \frac{999}{1000} |y_1|^{999/1000} y_2 \log |y_1| \right), \\ z_2 &= y_2 + \varepsilon \left(\frac{995007}{1999} y_1^2 + \frac{2991}{1000} y_1 y_2 + 3 y_2^2 \right). \end{aligned}$$

Replacing in the latter equation $y_1(0)$ and $y_2(0)$ by their specific values we get $z_1(0) = 1.101382 \times 10^{-4}$ and $z_2(0) = 0.500175$. Now we can obtain the reduced equations together with their initial conditions φ_0 and ϑ_0:

$$D\varphi(t) = \varepsilon |\varphi|^{999/1000}, \qquad \varphi_0 = 8.31439 \times 10^{296} \tag{4.4}$$

and

$$D\vartheta(t) = -\vartheta, \qquad \vartheta_0 = 0.500175. \tag{4.5}$$

Note that (4.4) is defined in $\boldsymbol{R}$, which is indeed the reduced phase space (as $r = s$ there is no constraint for the reduced phase space). The Lie group is also isomorphic to $\boldsymbol{R}$. We are lucky this time and can calculate explicit solutions of φ and ϑ as functions of time. We have that

$$\varphi(t) = \left(|\varphi_0|^{1/1000} + \frac{\varepsilon t}{1000}\right)^{1000}, \qquad \vartheta(t) = \vartheta_0 \exp(-t).$$

Now z_1 and z_2 are directly written as functions of time: $z_1(t) = \varphi(t)\vartheta(t)^{1000}$ and $z_2(t) = \vartheta(t)$. The next step consists in calculating the vector $\boldsymbol{y}$ in terms of $\boldsymbol{z}$, i.e., the direct change $\boldsymbol{y} = \boldsymbol{Y}(\boldsymbol{z};\varepsilon)$. This is done by means of (2.5) arriving at

$$y_1 = z_1 - \varepsilon\left(\frac{498001}{500}z_1^2 + 2000 z_1 z_2 + \frac{1}{998}z_2^2 - \frac{999}{1000}|z_1|^{999/1000} z_2 \log|z_1|\right),$$

$$y_2 = z_2 - \varepsilon\left(\frac{995007}{1999}z_1^2 + \frac{2991}{1000}z_1 z_2 + 3z_2^2\right).$$

Finally, we put the initial variables $\boldsymbol{x}$ in terms of $\boldsymbol{y}$ and arrive at an approximate solution of the stiff problem $\boldsymbol{x}(t)$ up to order $\mathcal{O}(\varepsilon^2)$.

A bound of the error made in the process of the Lie transformation is obtained after computing $\|\boldsymbol{h}(\boldsymbol{z};\varepsilon) - \boldsymbol{f}(\boldsymbol{Y}(\boldsymbol{z};\varepsilon);\varepsilon)\| = \varepsilon^2\|\boldsymbol{e}(\boldsymbol{z})\| + \mathcal{O}(\varepsilon^3)$, where $\|\boldsymbol{e}(\boldsymbol{z})\|^2$ is given by

$$\begin{aligned}\frac{1}{|z_1|^{1/500}}\big\{&z_1^2 z_2^2[0.995505 z_1 + 0.002991 z_2\\ &+(-0.000995505 z_1 - 2.991\times 10^{-6} z_2)\log|z_1|]^2\\ &+[-0.497752 z_1^3 + 0.995003 z_1^2 z_2 - 10^{-6}|z_1|^{999/1000} z_2^2 + 0.001 z_1 z_2^2 - 10^{-6} z_2^3\\ &+(0.000497752 z_1^3 - 0.000994007 z_1^2 z_2 + 10^{-6} z_1 z_2^2 + 1.001\times 10^{-9} z_2^3)\log|z_1|]^2\big\}.\end{aligned}$$

Bounding $\boldsymbol{z}$ such that $\max\{|z_1|, |z_2|\} = 1$ and fixing $\varepsilon = 10^{-4}$ we conclude that $\|\boldsymbol{e}(\boldsymbol{z})\| \leq 1.79177\times 10^{-8}$ is an upper bound of the global error. Finally, using the changes $\boldsymbol{y} = \boldsymbol{Y}(\boldsymbol{z};\varepsilon)$, together with the linear transformation relating $\boldsymbol{y}$ and $\boldsymbol{x}$, we obtain that whenever the variable $\boldsymbol{x}$ is controlled such that $|x_1| \leq 2$ and $|x_2| \leq 1000$, the total error of the computation never exceeds 1.79177×10^{-8} in an approximate time interval $[0, 10^4]$.

Notice that we could have chosen another vector field $\boldsymbol{T}(\boldsymbol{y})$ to make the transformation and construct the reduced equations, for instance $\boldsymbol{T}(\boldsymbol{y}) = (y_1, -y_2)^t$. Then the process would not be the same, and it would lead to a different asymptotic approximation of the original equation.

If we perform a numerical integration of (4.1) in a time interval $[0, 10^4]$ and use a Runge–Kutta–Felhberg embedded scheme of orders 4–5 equipped with dense output, the iteration stops at $t = 367.75$, as the process does not converge. However, using the backward-difference formula [23, 24] with variable order (i.e., a stiff-ODE-solver) there is convergence after 258 steps, obtaining a result which is exact up to 14 or 15 digits, and thus in many cases it is better than our approach. However, varying the initial conditions in the domain defined by $|x_1| \leq 2$ and $|x_2| \leq 1000$, the numerical integration by means of the backward-difference integrator does not converge where the analytical procedure gives satisfactory results.

4.2. Analytical integration of the Lorenz equations. We apply the theory of sections 2 and 3 to the Lorenz system given by

(4.6)
$$Dx_1(t) = 10(x_2 - x_1), \quad Dx_2(t) = 28x_1 - x_2 - x_1x_3, \quad Dx_3(t) = x_1x_2 - \frac{8}{3}x_3,$$

where t represents the time variable. To the latter system we add some initial conditions $\boldsymbol{x}(0) = (1/15, -1/11, 1/10)^t$. Equation (4.6) has been widely analyzed through the literature (see, for instance, the book by Verhulst [49] and the references therein), mainly with regard to its chaotic behavior and the existence of a strange attractor. Our aim now is to apply the method described in section 2 to system (4.6) with the goal of analyzing the Lorenz equations in a vicinity of the origin.

The linear part of (4.6) is given by $A\boldsymbol{x}$, where $\boldsymbol{x} = (x_1, x_2, x_3)^t$ and

$$A = \begin{pmatrix} -10 & 10 & 0 \\ 28 & -1 & 0 \\ 0 & 0 & -8/3 \end{pmatrix},$$

and its eigenvalues are $\lambda_{1,2} = \frac{1}{2}(-11 \pm \sqrt{1201})$ and $\lambda_3 = -8/3$. Now we make the corresponding linear change of variables $\boldsymbol{x} \to \boldsymbol{y}'$ so that A becomes diagonal (we call it A_J) with the eigenvalues in its diagonal. We stretch the variables, say $\boldsymbol{y}' \to \varepsilon\boldsymbol{y}$, so as to introduce a dimensionless small parameter $\varepsilon > 0$. Then the resulting differential equation $D\boldsymbol{y}(t) = \boldsymbol{f}(\boldsymbol{y}; \varepsilon)$ can be considered as a perturbed dynamical system. We look for a symmetry of the linear part $A_J\boldsymbol{y}$. It is achieved by taking $T = \operatorname{diag}\{1, \sqrt{2}, 0\}$, and thus $A_JT = TA_J$ and $[A_J\boldsymbol{y}, T\boldsymbol{y}] = \boldsymbol{0}$. This time $s = r = 1$, $m - s = 2$.

We perform the Lie transformation with the aim of constructing a new set of coordinates $\boldsymbol{z}$ in terms of $\boldsymbol{y}$. Let $\boldsymbol{w}_i(\boldsymbol{y}) \equiv \boldsymbol{w}_i(\boldsymbol{z}) = (w_{1i}(\boldsymbol{z}), w_{2i}(\boldsymbol{z}), w_{3i}(\boldsymbol{z}))^t$ be the generating function at order i, $\boldsymbol{g}_i(\boldsymbol{z}) = (g_{1i}(\boldsymbol{z}), g_{2i}(\boldsymbol{z}), g_{3i}(\boldsymbol{z}))^t$ be the nonlinear part of the transformed equation, and $\widetilde{\boldsymbol{f}}_i(\boldsymbol{z}) = (\widetilde{f}_{1i}(\boldsymbol{z}), \widetilde{f}_{2i}(\boldsymbol{z}), \widetilde{f}_{3i}(\boldsymbol{z}))^t$ be the known part at each order; hence the homology equation (2.4) reads as

$$\frac{8}{3}w_{1i}(\boldsymbol{z}) + n_1z_3D_3w_{1i}(\boldsymbol{z}) - n_2z_2D_2w_{1i}(\boldsymbol{z}) - \frac{8}{3}z_1D_1w_{1i}(\boldsymbol{z}) = \widetilde{f}_{1i}(\boldsymbol{z}) - g_{1i}(\boldsymbol{z}),$$

$$n_2w_{2i}(\boldsymbol{z}) + n_1z_3D_3w_{2i}(\boldsymbol{z}) - n_2z_2D_2w_{2i}(\boldsymbol{z}) - \frac{8}{3}z_1D_1w_{2i}(\boldsymbol{z}) = \widetilde{f}_{2i}(\boldsymbol{z}) - g_{2i}(\boldsymbol{z}),$$

$$-n_1w_{3i}(\boldsymbol{z}) + n_1z_3D_3w_{3i}(\boldsymbol{z}) - n_2z_2D_2w_{3i}(\boldsymbol{z}) - \frac{8}{3}z_1D_1w_{3i}(\boldsymbol{z}) = \widetilde{f}_{3i}(\boldsymbol{z}) - g_{3i}(\boldsymbol{z}),$$

where $n_1 = (-11 + \sqrt{1201})/2$ and $n_2 = (11 + \sqrt{1201})/2$.

We denote by $D\boldsymbol{z}(t) = \boldsymbol{h}(\boldsymbol{z}; \varepsilon)$ the normal form up to order three, and it is given by

(4.7)
$$Dz_1(t) = -\frac{8}{3}z_1 + \varepsilon\frac{(-9+\sqrt{1201})}{56}z_3^2 - \varepsilon^3\frac{238268911748107 + 3427671328157\sqrt{1201}}{3389702400(32622739 + 543621\sqrt{1201})}z_3^4,$$

$$Dz_2(t) = -\frac{11+\sqrt{1201}}{2}z_2 + \varepsilon^2\frac{3(1893759619 - 24165531\sqrt{1201})}{10117320080(25+3\sqrt{1201})}z_3^3,$$

$$Dz_3(t) = \frac{-11+\sqrt{1201}}{2}z_3 + \varepsilon^2\frac{15(1201 - 1689\sqrt{1201})}{134512(25+3\sqrt{1201})}z_3^3.$$

Terms in (4.7) factored by ε form the vector $\boldsymbol{h}_1$, whereas those factored by ε^2 are part of $\boldsymbol{h}_2$, and those factored by ε^3 define $\boldsymbol{h}_3$. Besides, $[\boldsymbol{h}_i, T\boldsymbol{y}] = \mathbf{0}$ for $1 \le i \le 3$. The expressions of $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, and $\boldsymbol{w}_3$ have been calculated using rational arithmetic, but we display an approximation with 10 digits. Thus we obtain the generating function $\boldsymbol{w} = \boldsymbol{w}_1 + \varepsilon \boldsymbol{w}_2 + (\varepsilon^2/2)\boldsymbol{w}_3$:

$$
\begin{aligned}
\boldsymbol{w}(\boldsymbol{z};\varepsilon) = (&0.0181340700 z_2^2 + 0.0385714285 z_2 z_3,\\
&- 0.1082080981 z_1 z_2 - 0.0053011728 z_1 z_3,\\
&- 0.0131559168 z_1 z_2 + 0.1082080981 z_1 z_3)^t\\
&+ \varepsilon(-0.0031438583 z_1 z_2^2 + 0.0008811978 z_1 z_2 z_3 + 0.0087311650 z_1 z_3^2,\\
&\quad - 0.0009063498 z_1^2 z_2 + 0.0017330286 z_2^3 - 0.0013559321 z_1^2 z_3\\
&\quad + 0.0028053572 z_2^2 z_3 + 0.0038427758 z_2 z_3^2,\\
&\quad 0.0024674265 z_1^2 z_2 + 0.0000168333 z_2^3 + 0.0009063498 z_1^2 z_3\\
&\quad - 0.0020552276 z_2^2 z_3 - 0.0032459561 z_2 z_3^2)^t\\
&+ \varepsilon^2(-0.0000795658 z_1^2 z_2^2 + 0.0001163808 z_2^4 + 0.0004142502 z_1^2 z_2 z_3\\
&\quad + 0.0001652742 z_2^3 z_3 - 0.0003707416 z_1^2 z_3^2\\
&\quad - 0.0000956330 z_2^2 z_3^2 - 0.0002239279 z_2 z_3^3,\\
&\quad - 0.0000618375 z_1^3 z_2 + 0.0001945584 z_1 z_2^3 - 0.0000817780 z_1^3 z_3\\
&\quad + 0.0004607450 z_1 z_2^2 z_3 - 0.0004185807 z_1 z_2 z_3^2 + 0.0000749699 z_1 z_3^3,\\
&\quad 0.0001084983 z_1^3 z_2 - 0.0000998823 z_1 z_2^3 + 0.0000618375 z_1^3 z_3\\
&\quad - 0.0004245437 z_1 z_2^2 z_3 - 0.0005441518 z_1 z_2 z_3^2\\
&\quad - 0.0014274482 z_1 z_3^3)^t + \mathcal{O}(\varepsilon^3).
\end{aligned}
$$

The vector spaces $\mathcal{P}_i$ and $\mathcal{R}_i$ coincide and are identified with the homogeneous polynomial spaces of dimension 3 and of degree i with coefficients in $\boldsymbol{R}$. Given an i, the space $\mathcal{Q}_i$ corresponds to the vector subspace of $\mathcal{P}_i$ for which the Lie bracket $[\boldsymbol{q}, T\boldsymbol{y}] = \mathbf{0}$ for any homogeneous polynomial vector $\boldsymbol{q}$ of degree i.

The first integrals associated with $T\boldsymbol{z}$ are the invariant polynomials determined from the partial differential equation $\langle D\varphi(\boldsymbol{z}), T\boldsymbol{z}\rangle = 0$, obtaining $\varphi = z_3$; therefore the coordinates of the Lie group are $\vartheta_1 = z_1$ and $\vartheta_2 = z_2$. The key point to obtain only one invariant—and, consequently, transform the initial system to a simpler one—is the nonexistence of resonant conditions among the entries of T. After computing the normal form up to order three, we pass to the reduced equations, yielding that

$$
D\varphi(t) = \frac{-11 + \sqrt{1201}}{2}\varphi + \frac{15(1201 - 1689\sqrt{1201})}{134512(25 + 3\sqrt{1201})}\varepsilon^2\varphi^3, \tag{4.8}
$$

and

$$
\begin{aligned}
D\vartheta_1(t) &= \frac{-9 + \sqrt{1201}}{56}\varepsilon\varphi^2 - \frac{238268911748107 + 3427671328157\sqrt{1201}}{3389702400(32622739 + 543621\sqrt{1201})}\varepsilon^3\varphi^4 - \frac{8}{3}\vartheta_1,\\
D\vartheta_2(t) &= \frac{3(1893759619 - 24165531\sqrt{1201})}{10117320080(25 + 3\sqrt{1201})}\varepsilon^2\varphi^3 - \frac{11 + \sqrt{1201}}{2}\vartheta_2.
\end{aligned} \tag{4.9}
$$

Observe that if we were interested in calculating some invariant sets of (4.6) through the normal forms, it could be enough to analyze the equilibria of the scalar equation (4.8) defined in $\mathbf{R}$. These critical points are in correspondence with two-dimensional

tori of the original equation which would be approximated using the direct Lie transformation. See [41, 37] for more details.

Solutions of (4.8) and (4.9) are obtained straightforwardly, as φ is first written explicitly as a function of time. Specifically, if $\varphi_0 = \varphi(0)$, we have

(4.10)
$$\varphi(t) = \frac{8\sqrt{8407(416-\sqrt{1201})}\exp\left(\frac{-11+\sqrt{1201}}{2}t\right)\varphi_0}{\sqrt{538048(416-\sqrt{1201})+15(1201-1689\sqrt{1201})\{1-\exp[(-11+\sqrt{1201})t]\}\varepsilon^2\varphi_0^2}}.$$

This expression is inserted in the linear equations (4.9). We rewrite system (4.9) as

$$D\vartheta_1(t) = a(t;\varepsilon) - \frac{8}{3}\vartheta_1, \qquad D\vartheta_2(t) = b(t;\varepsilon) - \frac{11+\sqrt{1201}}{2}\vartheta_2, \tag{4.11}$$

where $a(t;\varepsilon)$ and $b(t;\varepsilon)$ come from (4.9) after substituting φ by its expression given in (4.10). Hence we get

$$\begin{aligned}
\vartheta_1(t) &= \exp\left(-\frac{8}{3}t\right)\left((\vartheta_1)_0 + \int_0^t \exp\left(\frac{8}{3}s\right)a(s;\varepsilon)\,ds\right),\\
\vartheta_2(t) &= \exp\left(-\frac{11+\sqrt{1201}}{2}t\right)\left((\vartheta_2)_0 + \int_0^t \exp\left(\frac{11+\sqrt{1201}}{2}s\right)b(s;\varepsilon)\,ds\right),
\end{aligned} \tag{4.12}$$

the constants $(\vartheta_1)_0$ and $(\vartheta_2)_0$ being, respectively, $\vartheta_1(0)$ and $\vartheta_2(0)$. Note that $z_3(t) = \varphi(t)$, $z_1(t) = \vartheta_1(t)$, and $z_2(t) = \vartheta_2(t)$. Moreover, $\varphi_0 = z_3(0)$, $(\vartheta_1)_0 = z_1(0)$, and $(\vartheta_2)_0 = z_2(0)$. Thus, $\boldsymbol{z}(0)$ is obtained from $\boldsymbol{y}(0)$ using the expression $\boldsymbol{z} = \boldsymbol{Z}(\boldsymbol{y};\varepsilon)$, i.e., the inverse Lie transformation. As the expressions with rational arithmetic are quite involved, we display the corresponding numerical approximations with 10 digits. We have $\boldsymbol{z} = \boldsymbol{Z}(\boldsymbol{y};\varepsilon)$ with

$$\begin{aligned}
z_1 &= y_1 + \varepsilon(-0.0181340700y_2^2 - 0.0385714y_2y_3)\\
&\quad + \varepsilon^2(0.00314386y_1y_2^2 - 0.00873117y_1y_3^2 - 0.000881198y_1y_2y_3)\\
&\quad + \varepsilon^3(-0.000263461y_1^2y_2^2 + 2.06654\times10^{-6}y_2^4 - 0.000133513y_1^2y_2y_3\\
&\qquad + 0.0000146633y_2^3y_3 - 0.000783226y_1^2y_3^2 + 0.0000128441y_2^2y_3^2\\
&\qquad + 0.0000176879y_2y_3^3),\\
z_2 &= y_2 + \varepsilon(0.108208y_1y_2 + 0.00530117y_1y_3)\\
&\quad + \varepsilon^2(0.00090635y_1^2y_2 - 0.00173303y_2^3 + 0.00135593y_1^2y_3\\
&\qquad - 0.00280536y_2^2y_3 - 0.00384278y_2y_3^2)\\
&\quad + \varepsilon^3(0.0000154594y_1^3y_2 + 0.000295189y_1y_2^3 + 0.000182807y_1^3y_3\\
&\qquad - 0.000140385y_1y_2^2y_3 - 0.000605481y_1y_2y_3^2 - 0.0000504421y_1y_3^3),\\
z_3 &= y_3 + \varepsilon(0.0131559y_1y_2 - 0.108208y_1y_3)\\
&\quad + \varepsilon^2(-0.00246743y_1^2y_2 - 0.0000168333y_2^3 - 0.00090635y_1^2y_3\\
&\qquad + 0.00205523y_2^2y_3 + 0.00324596y_2y_3^2)\\
&\quad + \varepsilon^3(0.00022467y_1^3y_2 + 4.6008581025\times10^{-6}y_1y_2^3 - 0.0000154594y_1^3y_3\\
&\qquad - 0.000358644y_1y_2^2y_3 + 0.000191187y_1y_2y_3^2 + 0.00117751y_1y_3^3).
\end{aligned}$$

Now $\boldsymbol{y}$ must be expressed in terms of $\boldsymbol{z}$ so as to obtain the direct change $\boldsymbol{y} = \boldsymbol{Y}(\boldsymbol{z};\varepsilon)$ that we do not write down for reasons of space.

Fixing a value for t we evaluate $\boldsymbol{z}(t)$ with the aid of (4.10) and (4.12) (the initial conditions φ_0, $(\vartheta_1)_0 = z_1(0)$, and $(\vartheta_2)_0 = z_2(0)$ are supposed to be calculated previously). Then we determine $\boldsymbol{y}(t)$ by replacing $\boldsymbol{z}(t)$ in the latter system. Hence, we undo the scaling $\boldsymbol{y}(t) = \boldsymbol{y}'(t)/\varepsilon$, and in the end we get $\boldsymbol{x}(t)$ inverting back the Jordan linear change of variables.

Concerning an estimation of the error we know that

$$\|\boldsymbol{f}(\boldsymbol{y};\varepsilon) - \boldsymbol{h}(\boldsymbol{Z}(\boldsymbol{y};\varepsilon);\varepsilon)\| = \varepsilon^4\|\boldsymbol{e}(\boldsymbol{y})\| + \mathcal{O}(\varepsilon^5),$$

where $\|\boldsymbol{e}(\boldsymbol{y})\|^2$ stands for the quantity

$$\begin{aligned}
&1.75\times 10^{-10}y_1^8(y_2^2+y_3^2)\\
&\quad +10^{-9}y_1^6y_3(-1.806y_2^3-1.744y_2^2y_3+1.277y_2y_3^2-2.705y_3^3)\\
&\quad +10^{-10}y_2^2y_3^2(1.42y_2^6+2.84y_2^5y_3+3.47y_2^4y_3^2+1.95y_2^3y_3^3+2.80y_2y_3^5+3.31y_3^6)\\
&\quad +10^{-8}y_1^4(0.0763y_2^6+0.2359y_2^5y_3+0.4409y_2^4y_3^2+0.5054y_2^3y_3^3+0.04450y_2^2y_3^4\\
&\qquad\qquad -1.0003y_2y_3^5+1.5684y_3^6)\\
&\quad +10^{-9}y_1^2(-0.122y_2^8-0.688y_2^7y_3-1.725y_2^6y_3^2-2.249y_2^5y_3^3-1.083y_2^4y_3^4\\
&\qquad\qquad -1.396y_2^3y_3^5-0.545y_2^2y_3^6+3.645y_2y_3^7+0.389y_3^8).
\end{aligned}$$

We restrict ourselves to a vicinity of the origin of the original system. So we bound $\boldsymbol{y}$ by taking $\max\{|y_1|,|y_2|,|y_3|\} = 10$, and choosing $\varepsilon = 10^{-2}$ we have that the global error of our approach is upper bounded by $\varepsilon^4\|\boldsymbol{e}(\boldsymbol{y})\| \leq 1.33969\times 10^{-7}$ on a time scale of order $t \approx 100$. Then, going back to the variable $\boldsymbol{x}$, the solution of the Lorenz equation is calculated with an error less than 1.33969×10^{-7}, provided that we consider $|x_1| \leq 0.26$, $|x_2| \leq 2.28$, and $|x_3| \leq 1.18$.

Note that the solution of (4.6) is calculated almost explicitly. The only drawback is the determination of $\vartheta_1(t)$ and $\vartheta_2(t)$ since it is impossible to put the integrals appearing in (4.12) in terms of known functions of time. However, it is not hard to obtain an approximation of such integrals using polynomial interpolation or even to evaluate them explicitly for a given value of t by means of a standard algorithm capable of approximating numerically definite integrals maintaining the global error smaller than 10^{-7}.

4.3. Reduction and numerical integration of free particles in three degrees of freedom. Suppose we have the case of a free particle in three degrees of freedom (point mass) whose position is given by the vector $\boldsymbol{x} = (x,y,z)^t$, whereas its corresponding velocity is the vector $\boldsymbol{X} = (X,Y,Z)^t$. If the particle is subject to a weak perturbation, the dynamical system describing this effect is given by the Hamilton function

$$\mathcal{H}(\boldsymbol{x},\boldsymbol{X};\boldsymbol{c}) = \frac{1}{2}(X^2+Y^2+Z^2) + \mathcal{P}(\boldsymbol{x},\boldsymbol{X};\boldsymbol{c}), \tag{4.13}$$

where $\mathcal{P}$ refers to the perturbation and is a generic polynomial in $\boldsymbol{x}$ and $\boldsymbol{X}$ of degree three or higher and $\boldsymbol{c}$ contains some parameter(s) of the problem.

First, we define the symplectic change

$$x = X^*, \qquad y = Y^*, \qquad z = Z^*, \qquad X = x^*, \qquad Y = y^*, \qquad Z = z^*. \tag{4.14}$$

Then, after dropping the asterisks to avoid tedious notation, Hamiltonian (4.13) becomes

$$\mathcal{H}(\boldsymbol{x};\boldsymbol{X};\boldsymbol{c}) = \mathcal{H}_0(\boldsymbol{x}) + \mathcal{P}(\boldsymbol{x},\boldsymbol{X};\boldsymbol{c}) = \frac{1}{2}(x^2+y^2+z^2) + \mathcal{H}_1(\boldsymbol{x},\boldsymbol{X};\boldsymbol{c}), \tag{4.15}$$

where we have identified $\mathcal{P}$ with $\mathcal{H}_1$. By means of the linear change (4.14) the unperturbed part of $\mathcal{H}_0$ is proportional to the square of the distance of the particle to the origin, and we can use the whole $\mathcal{H}_0$ as a sole coordinate. Indeed, we introduce a set of orbital variables $(r, \vartheta, \nu, R, \Theta, N)$—called polar-nodal variables—where r stands for the radial distance from the origin of reference to the particle, ϑ represents the argument of latitude, and ν is the right ascension of the node, whereas R, Θ, and N are the conjugate momenta of r, ϑ, and ν, respectively. Besides, $rR = \langle \boldsymbol{x}, \boldsymbol{X} \rangle$, the action Θ designates the modulus of the angular momentum vector, i.e., $\Theta = \|\boldsymbol{x} \times \boldsymbol{X}\|$, and $N = xY - yX$ stands for the third component of the angular momentum; see more details in [11]. Whittaker [54] demonstrated that the transformation

$$\varrho : (r, \vartheta, \nu, R, \Theta, N) \longrightarrow (x, y, z, X, Y, Z),$$

defined by

$$\begin{array}{ll} x = \bar{x}\cos\nu - \bar{y}\cos I \sin\nu, & X = \bar{X}\cos\nu - \bar{Y}\cos I \sin\nu, \\ y = \bar{x}\sin\nu + \bar{y}\cos I \cos\nu, & Y = \bar{X}\sin\nu + \bar{Y}\cos I \cos\nu, \\ z = \bar{y}\sin I, & Z = \bar{Y}\sin I, \end{array} \tag{4.16}$$

where $\cos I = N/\Theta$ and $\bar{x}$, $\bar{y}$, $\bar{X}$, and $\bar{Y}$ are given through

$$\begin{array}{ll} \bar{x} = r\cos\vartheta, & \bar{X} = R\cos\vartheta - \dfrac{\Theta}{r}\sin\vartheta, \\ \bar{y} = r\sin\vartheta, & \bar{Y} = R\sin\vartheta + \dfrac{\Theta}{r}\cos\vartheta, \end{array} \tag{4.17}$$

is symplectic. We have to take into account that the transformation ϱ is singular for $\Theta = 0$, and $\Theta = |N|$ as I is an angle defined in $(0, \pi)$. Therefore, the domain of validity of the change given by ϱ is the subset of $\boldsymbol{R}^6$ defined by

$$\Delta = [0, +\infty) \times [0, 2\pi) \times [0, 2\pi) \times \boldsymbol{R} \times (0, \infty) \times (-\Theta, \Theta).$$

Let us denote by $\boldsymbol{P_N}$ the set of polar-nodal coordinates. In these variables we have $\mathcal{H}_0 = r^2/2$ and $\mathcal{H}_1(\boldsymbol{x}, \boldsymbol{X}; \boldsymbol{c}) \equiv \mathcal{H}_1(\boldsymbol{P_N}; \boldsymbol{c})$. Every term of $\mathcal{H}_1$ contains powers of r, R, Θ, N, or $(1 - N^2/\Theta^2)^{1/2}$, whereas ϑ and ν appear through sines and cosines of multiples of them. Now the Lie operator associated with $\mathcal{H}_0$ is the nilpotent[1] linear operator

$$\mathcal{L}_{\mathcal{H}_0} = rD_4(\cdot(\boldsymbol{P_N})),$$

and the transformation can be executed readily. In fact, if $\mathcal{W}_i$ designates the scalar generating function at order i, the homology equation which has to be solved at each order is

$$rD_4\mathcal{W}_i(\boldsymbol{P_N}) + \mathcal{K}_i = \widetilde{\mathcal{H}_i}. \tag{4.18}$$

Our goal is to construct a formal change of coordinates $\boldsymbol{P_N} \to \boldsymbol{P_N}'$, that is,

$$(r, \vartheta, \nu, R, \Theta, N) \longrightarrow (r', \vartheta', \nu', R', \Theta', N'),$$

so that Θ' and N' become formal integrals of the transformed Hamiltonian. We drop the primes for the polar-nodal variables to avoid tedious notation. Indeed, since both

[1] Its nilpotent character is analyzed in [39].

Θ and N are integrals of $\mathcal{H}_0$, it could be possible to build a unique transformation such that the new Hamiltonian $\mathcal{K}$ has them as integrals. (Note that the corresponding vector fields $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are the constant six-dimensional vectors (respectively, (0,0,0,0,−1,0) and (0,0,0,0,0,−1)).)

Next, one has to bring $\mathcal{H}$ to polar-nodal coordinates by means of the transformations (4.16) and (4.17). Then the homology equation has to be solved, and $\widetilde{\mathcal{H}_i}$ is split as

$$\widetilde{\mathcal{H}_i}(r,\vartheta,\nu,R,\Theta,N)=\widetilde{\mathcal{H}_i}^{*}(r,-,-,R,\Theta,N)+\widetilde{\mathcal{H}_i}^{\#}(r,\vartheta,\nu,R,\Theta,N),$$

where $\widetilde{\mathcal{H}_i}^{*}$ groups all the terms of $\widetilde{\mathcal{H}_i}$ independent of ϑ and of ν. So a term belongs to $\ker(\mathcal{L}_\Theta)\cap\ker(\mathcal{L}_N)$ if and only if it is independent of ϑ and of ν. Hence, $\mathcal{K}_i$ is taken to be $\widetilde{\mathcal{H}_i}^{*}$, and $\mathcal{W}_i$ is obtained by solving the quadrature (4.18). The process can be iterated to any order. After M steps, the transformed Hamiltonian $\mathcal{K}$ is given by

$$\mathcal{K}(r,\vartheta,\nu,R,\Theta,N)=\frac{1}{2}r^2+\varepsilon\mathcal{K}_1(r,R;\Theta,N)+\cdots+\frac{\varepsilon^M}{M!}\mathcal{K}_M(r,R;\Theta,N)+\mathcal{O}(\varepsilon^{M+1}).$$

Dropping the tail to $\mathcal{K}$ and giving the same name to the resulting Hamiltonian, we observe that it is independent of ϑ and of ν. In fact, our Lie transformation can be understood as an average over the angles ϑ and ν. Then Θ and N are integrals of $\mathcal{K}$. In this situation $\mathcal{K}$ defines a Hamiltonian vector field with one degree of freedom in the coordinate r and its momentum R.

Next, as we have introduced the integrals Θ and N, we build the invariants of the group of symmetry associated with them. The Poisson brackets of the terms $\varphi_1=x^2+y^2+z^2$, $\varphi_2=X^2+Y^2+Z^2$, $\varphi_3=xX+yY+zZ$, $\varphi_4=xY-yX$, and $\varphi_5=(xY-yX)^2+(xZ-zX)^2+(yZ-zY)^2$ with respect to Θ and N vanish, and they are related through $\varphi_1\varphi_2=\varphi_3^2+\varphi_5$. Thus, we have $r=5$, $s=4$, and $m-s=2$. However, two of the five differential equations associated with the φ_i are trivial, as Θ and N become constants of motion. Note that the cases excluded in polar-nodal variables, i.e., $\Theta=0$ and $|N|=\Theta$, are recovered with the use of the invariants, as they are polynomials, and thus they define perfectly the reduced phase space. Nevertheless, it is relevant only if one is interested in analyzing the behavior of the whole original system through its normal form.

Now we fix values for Θ and N, that is, $\varphi_4=c_1$, $\varphi_5=c_2\geq 0$, with $|c_1|\leq\sqrt{c_2}$. So $\mathcal{K}$ is transformed into the following Hamilton function in nonsymplectic variables:

$$\mathcal{T}(\varphi_1,\varphi_2,\varphi_3;c_1,c_2)=\frac{1}{2}\varphi_1+\varepsilon\mathcal{T}_1(\varphi_1,\varphi_2,\varphi_3;c_1,c_2)+\cdots+\frac{\varepsilon^M}{M!}\mathcal{T}_M(\varphi_1,\varphi_2,\varphi_3;c_1,c_2).$$

Note that the generators of the phase space satisfy

$$\varphi_1\varphi_2=\varphi_3^2+c_2, \tag{4.19}$$

defining, therefore, a two-dimensional surface with respect to the axes φ_1, φ_2, and φ_3, that is, a hyperbolic paraboloid. In the case $c_2=0$ the surface has a singularity at the origin of the frame defined by φ_1, φ_2, and φ_3. However, it is still possible to analyze this situation in the context of singular reduction. See the images depicted in Figure 4.1.

There still remains the extreme situation $\varphi_1=0$. It implies that $x=y=z\equiv 0$, which forces us to consider $\varphi_3=\varphi_4=\varphi_5\equiv 0$, whereas $\varphi_2\geq 0$. In this case the reduced phase space gets reduced to the straight line.

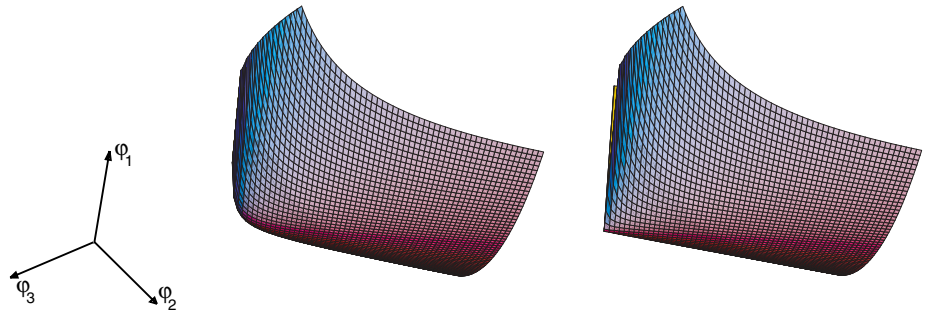


FIG. 4.1. *Reduced phase spaces (orbit spaces) for a Hamiltonian like* $\mathcal{T}(\varphi_1, \varphi_2, \varphi_3)$ *satisfying* (4.19). *The surface on the left is regular and corresponds to the cases* $c_2 \neq 0$. *The surface on the right has a singularity at the origin and is related to* $c_2 = 0$.

As an application of the above paragraphs we take in Hamiltonian (4.15) the perturbation $\mathcal{H}_1 = c\,xyz$, with the constant $c > 0$ being sufficiently small to consider $\mathcal{P}$ a small perturbation of $\mathcal{H}_0$ with small parameter c. Note that with this choice function $\mathcal{H}$ represents a three-degrees-of-freedom Hamilton function in the six-dimensional space spanned by $\boldsymbol{x}$ and $\boldsymbol{X}$.

To simplify the problem by means of coordinate formal transformations, we first need to write xyz in terms of polar-nodal variables. We do it using formulae (4.16) and (4.17). Then we apply the Lie transformation up to order two, i.e., up to an error of size $\mathcal{O}(c^3)$, calculating $\mathcal{K}_1(=0)$, $\mathcal{K}_2$ together with $\mathcal{W}_1$ and $\mathcal{W}_2$.

Hamiltonian $\mathcal{K}$, after truncation, reads as

$$\begin{aligned}\mathcal{K}(r, R; \Theta, N) = \frac{1}{2}r^2 + \frac{c^2}{1536r^8}[&18\Theta^6 \sin^2 I(8 - 12\sin^2 I + 5\sin^4 I)\\ &+ (-8 - 24\sin^2 I + 21\sin^4 I)r^6R^6\\ &+ \Theta^2(-8 - 120\sin^2 I + 105\sin^4 I)r^4R^4\\ &- 6\Theta^4(8 + 16\sin^2 I - 17\sin^4 I + 3\sin^6 I)r^2R^2].\end{aligned}$$

The generating function $\mathcal{W}$ is a finite Fourier series in ϑ and ν whose coefficients are powers of r, R, Θ, and $(1 - N^2/\Theta^2)^{1/2}$. Specifically, it is written as the sum $\mathcal{W}_1 + c\,\mathcal{W}_2$, where $\mathcal{W}_1$ contains 48 terms and $\mathcal{W}_2$ is composed of 200 terms.

The vector spaces $\mathcal{P}_i$ and $\mathcal{R}_i$ are spaces of functions whose terms are rational in r, R, Θ, and $(1 - N^2/\Theta^2)^{1/2}$, whereas the dependence on ϑ and on ν is through sines and cosines. The functions characterizing these spaces are often called Poisson series [17, 16]. However, the spaces $\mathcal{Q}_i$ are purely rational and are indeed the vector subspaces of $\mathcal{P}_i$ which do not contain the trigonometric part of $\mathcal{P}_i$.

The direct and inverse changes of coordinates are also large formulae. More precisely, the passage from the "old" r to the new one is a trigonometric sum in ϑ and ν with 312 terms. The equations of the direct changes for ϑ, ν, R, Θ, and N are also Fourier sums with 579, 454, 336, 340, and 236 terms, respectively. The inverse change contains 338 terms for r, 579 terms for ϑ, 452 for ν, 333 for R, 340 for Θ, and 236 terms for N.

The passage from $\mathcal{K}$ to $\mathcal{T}$ is constructive. Using the identities $rR = \varphi_3$ and

$r^2=\varphi_1$ and fixing $\Theta=\sqrt{c_2}$ and $N=c_1$ we conclude that

$$\begin{aligned}\mathcal{T}(\varphi_1,\varphi_3;\Theta,N)=\frac{1}{2}\varphi_1+\frac{c^2}{1536\varphi_1^4}[&-18c_1^6c_2(5c_2-\varphi_3^2)\\&+18c_1^2c_2(c_2+\varphi_3^2)(c_2^2-4c_2\varphi_3^2-\varphi_3^4)\\&+c_2^2(18c_2^3-60c_2^2\varphi_3^2-23c_2\varphi_3^4-11\varphi_3^6)\\&+3c_1^4(18c_2^3+16c_2^2\varphi_3^2+35c_2\varphi_3^4+7\varphi_3^6)].\end{aligned}$$

Observe that as $\mathcal{T}$ is independent of φ_2—a feature of the perturbation we have chosen—the differential equation in the φ_i contains only two terms, although we can determine φ_2 with the aid of (4.19). Besides, we have to work with trajectories not very close to the origin, as φ_1 appears in the denominator of $\mathcal{T}$. The first equation of (3.2) is determined now taking into account that $Dr(t)=D_4\mathcal{K}(\boldsymbol{P_N})$ and $DR(t)=-D_1\mathcal{K}(\boldsymbol{P_N})$ and then $D\varphi_1(t)=2rD_4\mathcal{K}(\boldsymbol{P_N})$ and $D\varphi_3(t)=RD_4\mathcal{K}(\boldsymbol{P_N})-rD_1\mathcal{K}(\boldsymbol{P_N})$. We obtain

(4.20)
$$\begin{aligned}D\varphi_1(t)&=\frac{c^2\varphi_3}{384c_2^2\varphi_1^3}[18c_1^6c_2-18c_1^2c_2(3c_2+\varphi_3^2)(c_2+3\varphi_3^2)\\&\qquad-c_2^2(60c_2^2+46c_2\varphi_3^2+33\varphi_3^4)+c_1^4(48c_2^2+210c_2\varphi_3^2+63\varphi_3^4)],\\D\varphi_3(t)&=-\varphi_1+\frac{c^2}{192c_2^2\varphi_1^4}[-18c_1^6c_2(5c_2-\varphi_3^2)\\&\qquad+18c_1^2c_2(c_2+\varphi_3^2)(c_2^2-4c_2\varphi_3^2-\varphi_3^4)\\&\qquad+c_2^2(18c_2^3-60c_2^2\varphi_3^2-23c_2\varphi_3^4-11\varphi_3^6)\\&\qquad+3c_1^4(18c_2^3+16c_2^2\varphi_3^2+35c_2\varphi_3^4+7\varphi_3^6)].\end{aligned}$$

The relevant dynamics of (4.13) is reflected in (4.20). However, we are interested in a numerical approximation of the original equation to which we add some initial conditions. We need to find out an numerical solution of (4.20) and to pose the corresponding system in the Lie group.

Now as candidates for coordinates of the Lie group we take ϑ and ν, since they are the angles which have "disappeared" in the analytic transformation. From $D\vartheta(t)=D_5\mathcal{K}(\boldsymbol{P_N})$, $D\nu(t)=D_6\mathcal{K}(\boldsymbol{P_N})$ we arrive at the second equation of (3.2) which is, as expected, independent of ϑ and of ν:

(4.21)
$$\begin{aligned}D\vartheta(t)&=-\frac{c^2}{768c_2^{5/2}\varphi_1^4}[18c_1^6c_2\varphi_3^2-18c_1^2c_2(c_2-\varphi_3^2)^2(2c_2+\varphi_3^2)\\&\qquad+c_2^3(-54c_2^2+120c_2\varphi_3^2+23\varphi_3^4)\\&\qquad+c_1^4(-54c_2^3+105c_2\varphi_3^4+42\varphi_3^6)],\\D\nu(t)&=\frac{c^2c_1}{128c_2^2\varphi_1^4}[9c_1^4c_2(-5c_2+\varphi_3^2)+3c_2(c_2+\varphi_3^2)(c_2^2-4c_2\varphi_3^2-\varphi_3^4)\\&\qquad+c_1^2(18c_2^3+16c_2^2\varphi_3^2+35c_2\varphi_3^4+7\varphi_3^6)].\end{aligned}$$

Observe that once $\varphi_1(t)$ and $\varphi_3(t)$ will be numerically calculated from (4.20), the angles $\vartheta(t)$ and $\nu(t)$ shall be readily determined.

We choose initial conditions for Hamiltonian (4.15), concretely up to 10 digits we put positions $\boldsymbol{x}_0=(-0.0639671104,0,0.0581813272)^t$, and for velocities we write $\boldsymbol{X}_0=(-0.0369885382,-0.1719633719,0.0336429492)^t$. Using (4.16) and (4.17), we

have that $r_0 = 0.086468827$, $\vartheta_0 = 1.5707963267$, $\nu_0 = 1.5707963267$, $R_0 = 0.05$, $\Theta_0 = 0.014869471$, and $N_0 = 0.011$. Fixing the value $c = 10^{-4}$, we obtain $\mathcal{H}_0(\boldsymbol{x}_0) = 0.003738429$, whereas $\mathcal{H}_1(\boldsymbol{x}_0, \boldsymbol{X}_0; c) = 2.139918249 \times 10^{-8}$, and therefore $\mathcal{H}_1$ can be considered as a small perturbation of $\mathcal{H}_0$.

We switch now to polar-nodal variables. First, we need to know the value of the transformed variables, obtained using the inverse Lie change. The transformed initial conditions are given by the six-tuple $(r'_0, \vartheta'_0, \nu'_0, R'_0, \Theta'_0, N'_0) = (0.086469074, 1.570796662, 1.570796248, 0.0500000954, 0.0148695127, 0.0110000280)$. (Now we have recovered the primes for the transformed variables.) Then the initial conditions for system (4.20) are determined, yielding $\varphi_{1_0} = 0.007476900$ and $\varphi_{3_0} = 0.004323461$. Note that the initial conditions for (4.21) are given by ϑ'_0 and ν'_0. Besides, the values for the constants are $c_1 = 0.0110000280$ and $c_2 = 0.0002211024$. Note that $\Theta'(t) = \Theta'_0 \equiv \sqrt{c_2}$ and $N'(t) = N'_0 \equiv c_1$.

System (4.20) has been numerically solved using an Adams–Moulton method and 76 steps in the interval $[0, 15]$. With the approximations $\varphi_1(t)$ and $\varphi_3(t)$ we determine $r'(t) = \sqrt{\phi_1(t)}$ and $R'(t) = \varphi_3(t)/\sqrt{\varphi_1(t)}$. Equation (4.21) is solved obtaining $\vartheta'(t)$ and $\nu'(t)$ for $t \in [0, 15]$, also with the same Adams–Moulton formula and using 108 steps. Next, the passage from the prime polar-nodal variables to the original ones is done after using the direct change of coordinates. Finally, the transformation from the polar-nodal variables to the Cartesian ones is performed with the aid of (4.16) and (4.17). This completes our seminumerical treatment.

Instead of estimating the size of the global error committed with our approach, we have calculated the solution of the initial value problem given by Hamiltonian (4.15) with initial conditions $\boldsymbol{x}_0$ and $\boldsymbol{X}_0$ given above. We have used an Adams–Moulton method, obtaining a solution after 196 steps in the interval $[0, 15]$. Comparison between the fully numerical and seminumerical approaches gives a coincidence of seven digits in the six coordinates for all t in $[0, 15]$. As the value of the Hamiltonian is an invariant of the whole process (as it represents the total energy of the problem), we have compared the value of $\mathcal{H}(\boldsymbol{x}_0, \boldsymbol{X}_0; c)$ with $\mathcal{H}(\boldsymbol{x}(t), \boldsymbol{X}(t); c)$ for $t \in [0, 15]$, obtaining that the difference coincides in at least eight digits. Note that within this approximation our approach behaves better, as the integration of (4.20) and (4.21) is much simpler than the integration of the equation defined through $\mathcal{H}$.

If we enlarge the interval $[0, 15]$, the seminumerical solution begins to become worse. This is explainable from the point of view of the type of solution related to the original unperturbed problem (e.g., (4.13) or (4.15) with $c \equiv 0$). Indeed, if $c = 0$, excepting the radial velocity R, the rest of the polar-nodal variables are constants. Moreover, $R(t) = -r_0 t + R_0$, which means that as t becomes big, the term $|R(t)|$ also increases. Adding now the perturbation $c \neq 0$ it implies that the solution is not bounded while t becomes big. This is an unavoidable feature of the present problem, in contrast to other systems which are formed by a principal equation with bounded solutions to which we attach a small perturbation, as it occurs, for quasi-periodic equations and the elimination of some angular variables (or the time) using averaging procedures.

5. Concluding remarks. This paper outlines a new method for building approximate solutions of initial value problems in the cases where the differential equation is formed by a principal part plus a small regular perturbation. The main features of the technique we present are the following:

- The method is based on the analytic reduction of the original problem by means of the introduction of a formal symmetry through appropriate Lie transformations. After truncating, the reduction process carries out the decoupling of the

transformed original system into two subsystems which are either solved analytically or approximated numerically. We stress that the symmetry properties of the generalized normal form are responsible for allowing the symmetry reduction.

• Compared with the standard numerical routines our approach has the advantage of modifying analytically the initial equation, arriving therefore at a simplified version of it. These transformed systems can sometimes be solved analytically. This leads to a full analytic approximation of the problem with the typical error committed after truncating the normal formal transformation.

• However, if the transformed equations cannot be solved directly, one can still use a numerical algorithm for differential equations with constraints, which involves less computational effort than the solution of the original equations. This is really useful when the dimensions of the two subsystems are small.

• On some occasions the choice of adequate coordinates is crucial to perform the computations, especially those concerning the Lie transformation treatment. Besides, as the splitting of the transformed equation into the two subsystems is usually not apparent, one has to put much care in the determination of the coordinates of the Lie group. Unfortunately, there is no systematic manner for the selection of the variables in which one carries out the Lie transformation, though the coordinates of the orbit space can be obtained constructively.

Acknowledgments. I thank Patricia Yanguas (Universidad Pública de Navarra) for fruitful discussions concerning Lie transformations for vector fields and reduction procedures we shared during the elaboration of the paper. The remarks and suggestions of the anonymous referees have contributed significantly to improve a previous version of the paper.

REFERENCES

[1] M. Abud and G. Sartori, *The geometry of spontaneous symmetry breaking*, Ann. Physics, 150 (1983), pp. 307–372.

[2] J. M. Arms, R. H. Cushman, and M. J. Gotay, *A universal reduction procedure for Hamiltonian group actions*, in The Geometry of Hamiltonian Systems, T. Ratiu, ed., Springer–Verlag, Berlin, New York, 1991, pp. 33–51.

[3] R. Barrio and J. Palacián, *Semianalytical methods for high–eccentric orbits*: *Zonal harmonics and air drag terms*, Adv. Astron. Sci., 95 (1997), pp. 331–339.

[4] R. Barrio and J. Palacián, *High–order averaging of eccentric artificial satellites perturbed by the Earth's potential and air–drag terms*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 1517–1534.

[5] N. N. Bogoliubov and Y. A. Mitropolski, *Asymptotic Methods in the Theory of Nonlinear Oscillators*, Gordon and Breach, New York, 1961.

[6] A. D. Bryuno, *Local Methods in Nonlinear Differential Equations. Part* I. *The Local Method of Nonlinear Analysis of Differential Equations. Part* II. *The Sets of Analyticity of a Normalizing Transformation*, Springer Series in Soviet Mathematics, Springer–Verlag, Berlin, New York, 1989.

[7] P. Chossat and R. Lauterbach, *Methods in Equivariant Bifurcations and Dynamical Systems*, Adv. Ser. Nonlinear Dynam. 15, World Scientific, Singapore, 2000.

[8] G. Cicogna and G. Gaeta, *Normal forms and nonlinear symmetries*, J. Phys. A, 27 (1994), pp. 7115–7124.

[9] G. Cicogna and G. Gaeta, *Symmetry and Perturbation Theory in Nonlinear Dynamics*, Lecture Notes in Phys. New Ser. M Monogr. 57, Springer–Verlag, Berlin, New York, 1999.

[10] G. Cicogna and S. Walcher, *Convergence of normal form transformations*: *The role of symmetries*, Acta Appl. Math., 70 (2002), pp. 95–111.

[11] S. L. Coffey and A. Deprit, *Third order solution to the main problem in satellite theory*, J. Guidance Control Dynam., 5 (1982), pp. 363–371.

[12] R. H. Cushman, *Reduction, Brouwer's Hamiltonian, and the critical inclination*, Celestial Mech., 31 (1983), pp. 401–429.

[13] R. H. Cushman and L. M. Bates, *Global Aspects of Classical Integrable Systems*, Birkhäuser, Basel, 1997.
[14] R. H. Cushman, J. A. Sanders, and N. White, *Normal form for the* $(2;n)$*-nilpotent vector field, using invariant theory*, Phys. D, 30 (1988), pp. 399–412.
[15] A. Deprit, *Canonical transformations depending on a small parameter*, Celestial Mech., 1 (1969), pp. 12–30.
[16] A. Deprit, *Delaunay normalisations*, Celestial Mech., 26 (1982), pp. 9–21.
[17] A. Deprit, J. Henrard, and A. Rom, *Analytical lunar ephemeris*: *Delaunay's theory*, Astronom. J., 76 (1971), pp. 269–272.
[18] A. Deprit, V. Lanchares, M. Iñarrea, J. P. Salas, and J. D. Sierra, *Teardrop bifurcation for Rydberg atoms in parallel electric and magnetic fields*, Phys. Rev. A, 54 (1996), pp. 3885–3893.
[19] H. Derksen, *Computation of invariants for reductive groups*, Adv. Math., 141 (1999), pp. 366–384.
[20] J. Dieudonné, *Foundations of Modern Analysis*, Pure Appl. Math. 10, Academic Press, New York, 1969.
[21] G. Gaeta, *A splitting lemma for equivariant dynamics*, Lett. Math. Phys., 33 (1995), pp. 313–320.
[22] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics*, Interdiscip. Appl. Math. 1, Springer–Verlag, New York, 1990.
[23] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I. *Nonstiff Problems*, 2nd ed., Springer–Verlag, Berlin, New York, 1993.
[24] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential–Algebraic Problems*, Springer–Verlag, Berlin, New York, 1996.
[25] J. Henrard, *On a perturbation theory using Lie transforms*, Celestial Mech., 1 (1970), pp. 107–120.
[26] L. Jay, *Symplectic partitioned Runge–Kutta methods for constrained Hamiltonian systems*, SIAM J. Numer. Anal., 33 (1996), pp. 368–387.
[27] A. A. Kamel, *Expansion formulae in canonical transformations depending on a small parameter*, Celestial Mech., 1 (1969), pp. 190–199.
[28] U. Kirchgraber, *An ODE–solver based on the method of averaging*, Numer. Math., 53 (1988), pp. 621–652.
[29] N. Krylov and N. Bogoliubov, *Introduction to Nonlinear Mechanics*, Academic Press, New York, 1947.
[30] J. Laskar, *Large–scale chaos in the solar system*, Astron. Astrophys. Lett., 287 (1994), pp. 9–12.
[31] L. M. Markhashov, *On the reduction of differential equations to the normal form by an analytic transformation*, J. Appl. Math. Mech., 38 (1974), pp. 738–740.
[32] S. Marmi, *Chaotic behaviour in the solar system (following J. Laskar)*, Astérisque, 266 (2000), pp. 113–136.
[33] K. R. Meyer, *Symmetries and integrals in mechanics*, in Dynamical Systems, M. M. Peixoto, ed., Academic Press, New York, London, 1973, pp. 259–272.
[34] K. R. Meyer, *A Lie transform tutorial* II, in Computer Aided Proofs in Analysis, K. Meyer and D. S. Schmidt, eds., IMA Vol. Math. Appl. 28, Springer–Verlag, Berlin, New York, 1991, pp. 190–210.
[35] L. Michel, *Points critiques des fonctions invariantes sur une* G*–varieté*, C. R. Acad. Sci. Paris Sér. A–B, 272 (1971), pp. 433–436.
[36] P. J. Olver, *Applications of Lie Groups to Differential Equations*, Grad. Texts in Math. 107, Springer–Verlag, Berlin, New York, 1986.
[37] J. Palacián, *Invariant manifolds of an autonomous ordinary differential equation from its generalised normal forms*, Chaos, 13 (2003), pp. 1188–1204.
[38] J. Palacián and P. Yanguas, *Reduction of polynomial Hamiltonians by the construction of formal integrals*, Nonlinearity, 13 (2000), pp. 1021–1055.
[39] J. Palacián and P. Yanguas, *Reduction of polynomial planar Hamiltonians with quadratic unperturbed part*, SIAM Rev., 42 (2000), pp. 671–691.
[40] J. Palacián and P. Yanguas, *Generalized normal forms for polynomial vector fields*, J. Math. Pures Appl. (9), 80 (2001), pp. 445–469.
[41] J. Palacián and P. Yanguas, *Periodic orbits of the Lorenz system through perturbation theory*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 11 (2001), pp. 2559–2566.
[42] H. Poincaré, *Mémoire sur les courbes définies par une équation différentielle*, J. Math. Pures Appl., 1 (1885), pp. 167–244.
[43] H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste*, Vol. II, Gauthier-Villars, Paris,

1893.
[44] J. A. Sanders and F. Verhulst, *Averaging Methods in Nonlinear Dynamical Systems*, Appl. Math. Sci. 59, Springer–Verlag, Berlin, New York, 1985.
[45] J. M. Sanz–Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Appl. Math. Math. Comput. 7, Chapman and Hall, London, 1994.
[46] G. Schwarz, *Smooth functions invariant under the action of a compact Lie group*, Topology, 14 (1975), pp. 63–68.
[47] C. L. Siegel, *Über die Existenz einer Normalform analytischer Hamiltonscher Differentialgleichungen in der Nähe einer Gleichgewichtlösung*, Math. Ann., 128 (1954), pp. 144–170.
[48] V. Szebehely, *Theory of Orbits*, Academic Press, New York, 1967.
[49] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Springer–Verlag, Berlin, 1996.
[50] S. Walcher, *On differential equations in normal form*, Math. Ann., 291 (1991), pp. 293–314.
[51] S. Walcher, *On transformations into normal form*, J. Math. Anal. Appl., 180 (1993), pp. 617–632.
[52] S. Walcher, *On convergent normal form transformations in presence of symmetries*, J. Math. Anal. Appl., 244 (2000), pp. 17–26.
[53] R. Weitzenböck, *Uber die Invarianten von linearen Gruppen*, Acta Math., 58 (1932), pp. 231–293.
[54] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, Cambridge University Press, Cambridge, UK, 1927.
[55] S. Wolfram, *The Mathematica Book*, 4th ed., Wolfram Media/Cambridge University Press, Cambridge, UK, 1999.
[56] P. Yanguas, *Lowering the dimension of polynomial vector fields in* $\mathbf{R}^2$ *and* $\mathbf{R}^3$, Chaos, 11 (2001), pp. 306–318.
[57] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A, 150 (1990), pp. 262–268.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 905–933

EFFICIENT SOLVERS FOR SADDLE-POINT PROBLEMS ARISING FROM DOMAIN DECOMPOSITIONS WITH LAGRANGE MULTIPLIERS*

QIYA HU†, ZHONGCI SHI†, AND DEHAO YU†

Abstract. In this paper, we are concerned with the domain decomposition method with Lagrange multipliers for solving three-dimensional elliptic problems with variable coefficients. We shall first introduce a weighted saddle-point problem resulting from this domain decomposition, which can be solved by existing iterative methods. Then we will construct two simple preconditioners, one for the system associated with the displacement variable and the other for the Schur complement system associated with the multiplier variable, that are applicable to various discretization schemes. The new preconditioners possess such local properties that they can be implemented cheaply. We will show that the condition number of the global preconditioned system grows only as the logarithm of the dimension of the local problem associated with an individual substructure and is independent of the large variations of the coefficients across the local interfaces.

Key words. domain decomposition, nonmatching grids, mortar element, Lagrange multiplier, weighted saddle-point problem, preconditioner, condition number

AMS subject classifications. 65F10, 65N30, 65N55

DOI. 10.1137/S0036142902404108

1. Introduction.

In recent years, there has been a fast growing interest in the domain decomposition methods (DDMs) with Lagrange multipliers, which were studied early in [11], [12], [13], and [26]. This kind of DDM has many advantages over the traditional DDMs in applications (cf. [1], [19], [25], [34], and [35]).

It is known that the DDM with Lagrange multipliers results in a saddle-point problem of the displacement variable and the multiplier variable. There exist two different approaches for solving such a saddle-point problem: (1) eliminate the displacement to build an interface equation of the multiplier and solve the interface equation by some PCG method (see [1], [14], [15], [19], [29], and [34]); (2) solve the saddle-point problem itself by some preconditioned iterative method (see [24], [25], and [35]). Each of the approaches has its individual merits: the PCG iteration of the first approach has the optimal convergence, while the interface equation need not be built in the second approach, and so inexact solvers can be applied to both the primal and the Schur complement system (this is important for solving nonlinear problems and problems with variable coefficients). As we know, the displacement corresponds to a singular (positive semidefinite) problem on a floating subdomain. There exist many techniques to deal with such singularity, for example FETI-type methods (see [13], [15], [14], and [24]). The interface equation in a FETI method is always derived in a subspace by using a projection, and the interface equation can be solved by the *projected* PCG method with the Dirichlet preconditioner. The projection is realized by

*Received by the editors March 15, 2002; accepted for publication (in revised form) October 29, 2003; published electronically July 14, 2004. This work was supported by Special Funds for Major State Basic Research Projects of China G1999032804 and Natural Science Foundation of China grant G10371129.

http://www.siam.org/journals/sinum/42-3/40410.html

†Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, China (hqy@lsec.cc.ac.cn, szc@lsec.cc.ac.cn, ydh@lsec.cc.ac.cn).

solving a coarse problem, and each local solver in the preconditioner, which is defined on the boundary of a subdomain, can be implemented by solving a local Dirichlet problem (see [13] and [15] for details).

In the present paper, we consider the second approach mentioned above for solving saddle-point problems and propose a new strategy to handle the singularity. We discuss a general DDM with Lagrange multipliers for three-dimensional elliptic problems with variable coefficients, which allows finite element discretization with nonmatching grids and coupling discretization by the finite element and the spectral element. We will first transform the resulting saddle-point system into another equivalent saddle-point system in which the system corresponding to the displacement is positive definite. Then we construct two new preconditioners for the primal problem corresponding to the displacement and the Schur complement system corresponding to the multiplier, respectively. Each local solver in the preconditioner for the Schur complement is defined on the common face of two neighboring subdomains and can be implemented in a more efficient manner. The new method successfully avoids the action of the projections used in [24] and reduces the cost of computation significantly. It is shown that the global preconditioned system has a nearly optimal condition number, which is independent of the large variations of the coefficient across the local interfaces.

Also, the method can be extended to nonlinear problems with some elliptic properties.

The outline of the remainder of the paper is as follows. We introduce an augmented saddle-point problem in section 2. In section 3, we construct the preconditioners for the saddle-point system and give general convergence results. The main results of the paper will be shown in section 4. In section 5, we discuss various choices of the multiplier space and introduce one class of cheap local solver. Finally, we report some numerical results in section 6.

2. Domain decomposition and the saddle-point problem. This section is devoted to the introduction of the augmented saddle-point problem.

Consider the model problem

$$\begin{cases} -\mathrm{div}(a\nabla u) = f \ \text{ in } \ \Omega, \\ \qquad u = 0 \quad \text{on } \ \partial\Omega, \end{cases} \tag{2.1}$$

where Ω is a bounded, connected Lipschitz domain in $\mathbb{R}^3$ and $a \in L^\infty(\Omega)$ is a positive function.

Let $H_0^1(\Omega)$ denote the standard Sobolev space, and define the bilinear form

$$\mathcal{A}(v,w) = \int_\Omega a\nabla v \cdot \nabla w dx, \quad v, w \in H_0^1(\Omega).$$

Let $(\cdot,\cdot)$ denote the $L^2(\Omega)$ inner product. The weak formulation of (2.1) in $H_0^1(\Omega)$ is then given by the following: find $u \in H_0^1(\Omega)$ such that

$$\mathcal{A}(u,v) = (f,v) \quad \forall v \in H_0^1(\Omega). \tag{2.2}$$

In the following, we define a discrete problem of (2.2) based on nonoverlapping domain decomposition.

Let the domain Ω be decomposed into $\bar{\Omega} = \bigcup_{k=1}^N \bar{\Omega}_k$, which satisfy $\Omega_i \cap \Omega_j = \emptyset$ when $i \neq j$. For simplicity of exposition, we consider only the case of geometrically conforming partitionings of the region into subdomains:

(i) If $\bar{\Omega}_i \cap \bar{\Omega}_j \neq \emptyset$ for some $i \neq j$, then $\Gamma_{ij} = \bar{\Omega}_i \cap \bar{\Omega}_j$ is a face of both Ω_i and Ω_j. Define $\Gamma = \cup \Gamma_{ij}$.

(ii) Each subdomain Ω_k has the same "size" d in the usual way (see [8] and [38]).

As usual, we assume that each Ω_k is a polyhedron. Let $V(\Omega_k)$ be a finite-dimensional space on Ω_k, for example a finite element space or a spectral element space. We assume that $V(\Omega_k) \subset H^1(\Omega_k)$ for each subdomain Ω_k. If $\partial\Omega_k \cap \partial\Omega \neq \emptyset$, we require that all functions in $V(\Omega_k)$ vanish on $\partial\Omega_k \cap \partial\Omega$. Define $V(\Omega) = \prod_{k=1}^N V(\Omega_k)$ and $V(\partial\Omega_k) = V(\Omega_k)|_{\partial\Omega_k}$. In the next section, we will use several additional spaces. For $\Gamma_{ij} \subset \Gamma$, define the *local* trace spaces $V_i(\Gamma_{ij}) = V(\partial\Omega_i)|_{\Gamma_{ij}}$ and $V_j(\Gamma_{ij}) = V(\partial\Omega_j)|_{\Gamma_{ij}}$. Besides, define

$$V_i^0(\Gamma_{ij}) = V_i(\Gamma_{ij}) \cap H_0^1(\Gamma_{ij}) \quad \text{and} \quad V_j^0(\Gamma_{ij}) = V_j(\Gamma_{ij}) \cap H_0^1(\Gamma_{ij}).$$

It is possible that $V_i(\Gamma_{ij}) \neq V_j(\Gamma_{ij})$ and $V_i^0(\Gamma_{ij}) \neq V_j^0(\Gamma_{ij})$, for example for finite element discretization with nonmatching grids or coupling discretization of the finite element and the spectral element.

Let $W(\Gamma_{ij})$ be a given finite-dimensional space on the face Γ_{ij}, which will be chosen as the *local* multiplier space. We assume that the space $W(\Gamma_{ij})$ contains the constant function. There are many ways to define the *local* multiplier space $W(\Gamma_{ij})$, for example the well-known mortar element method (see [6], [4], [22], and section 5 for details). Define $W(\Gamma) = \prod_{\Gamma_{ij} \subset \Gamma} W(\Gamma_{ij})$.

Let $P_{ij} : L^2(\Gamma_{ij}) \to W(\Gamma_{ij})$ be the orthogonal projection with respect to the $L^2(\Gamma_{ij})$ inner product. For $v \in V(\Omega)$, set $v|_{\Omega_k} = v_k$. Define

$$\tilde{V}(\Omega) = \{v \in V(\Omega) : \ P_{ij}(v_i|_{\Gamma_{ij}} - v_j|_{\Gamma_{ij}}) = 0 \text{ for each } \Gamma_{ij} \subset \Gamma\}.$$

Note that we do not require that $\tilde{V}(\Omega) \subset H_0^1(\Omega)$.

Define the *local* bilinear form

$$\mathcal{A}_k(v, w) = \int_{\Omega_k} a \nabla v \cdot \nabla w dx, \quad v, w \in H^1(\Omega_k).$$

The discrete problem of (2.2) is the following: find $\bar{u} \in \tilde{V}(\Omega)$ such that

$$\sum_{k=1}^N \mathcal{A}_k(\bar{u}_k, v_k) = (f, v) \quad \forall v \in \tilde{V}(\Omega). \tag{2.3}$$

We assume that the system (2.3) has a unique solution.

Let $A_k : V(\Omega_k) \to V(\Omega_k)$ be the *local* operator defined by

$$(A_k v, w)_{\Omega_k} = \mathcal{A}_k(v, w), \quad v \in V(\Omega_k) \quad \forall w \in V(\Omega_k).$$

By introducing the *sign* function

$$\sigma_{ij} = \begin{cases} 1, & i < j, \\ -1, & i > j, \end{cases}$$

we define the operator $B_k : V(\Omega_k) \to W(\Gamma)$ as follows:

$$(B_k u)|_{\Gamma_{ij}} = \begin{cases} \sigma_{ij} P_{ij}(u_k|_{\Gamma_{ij}}), & \Gamma_{ij} \subset \partial\Omega_k, \\ 0, & \Gamma_{ij} \not\subset \partial\Omega_k. \end{cases}$$

Define the operators $A : V(\Omega) \to V(\Omega)$ and $B : V(\Omega) \to W(\Gamma)$ by

$$A = \mathrm{diag}(A_1\ A_2 \cdots A_N)$$

and

$$Bv = \sum_{k=1}^{N} B_k v_k, \quad v \in V(\Omega),$$

respectively. Let $\langle \cdot, \cdot \rangle$ denote the $L^2(\Gamma)$ inner product and $B^t : W(\Gamma) \to V(\Omega)$ denote the adjoint of B, which satisfies

$$(B^t \mu, v) = \langle \mu, Bv \rangle \quad \forall \mu \in W(\Gamma),\ v \in V(\Omega).$$

It is easy to see that the space $\tilde{V}(\Omega)$ can be written as

$$\tilde{V}(\Omega) = \{v \in V(\Omega) : Bv = 0\}.$$

Then (2.3) is equivalent to the following saddle-point problem: find $(\bar{u}, \lambda) \in V(\Omega) \times W(\Gamma)$ such that

$$\begin{cases} A\bar{u} + B^t\lambda = f, \\ B\bar{u} = 0. \end{cases} \tag{2.4}$$

Here the unknown λ is called the Lagrange multiplier for the constraint $B\bar{u} = 0$.

Although the operator A is block diagonal, the system (2.4) cannot be solved in the standard way (refer to [26] and [19]). The main difficulty is that each *local* operator A_k corresponding to some internal subdomain Ω_k is singular on $V(\Omega_k)$ (so the *global* operator A is also singular on $V(\Omega)$). To handle this singularity, many existing methods have to consider a subspace of λ and project λ repeatedly into this subspace, which increase the cost of calculation. For example, the paper [24] discussed the preconditioned conjugate residual method for solving the system (2.4) based on this technique.

In the following we propose a new way to deal with such singularity. It is easy to see that

$$\ker(A) \cap \ker(B) = \{0\}, \tag{2.5}$$

and so the operator $A + rB^tB$ is positive definite for any positive number r. Because of this, we may consider the augmented Lagrange multiplier framework

$$\begin{cases} (A + d^{-1}B^tB)\bar{u} + B^t\lambda = f, \\ B\bar{u} = 0, \end{cases} \tag{2.6}$$

which has the same solution as (2.4). A similar method has been discussed in [35], but an additional interface unknown φ was introduced.

The coefficient $a(x)$ has possible large variation from one subdomain to another. To avoid the influence of these large variations, we will consider another augmented Lagrange multiplier formulation instead of (2.6). Let α_k and β_k be the positive constants defined by

$$\alpha_k \le a(x) \le \beta_k \quad \forall x \in \Omega_k,\ (k = 1, \ldots, N). \tag{2.7}$$

Then

$$\alpha_k |v|^2_{1,\Omega_k} \le (A_k v, v)_{\Omega_k} \le \beta_k |v|^2_{1,\Omega_k} \quad \forall \in H^1(\Omega_k). \tag{2.8}$$

For a face $\Gamma_{ij} \subset \Gamma$, define $\alpha_{ij} = \min\{\alpha_i, \alpha_j\}$.

Define the operator $\bar{B}_k : V(\Omega_k) \to W(\Gamma)$,

$$(\bar{B}_k u)|_{\Gamma_{ij}} = \begin{cases} \sigma_{ij} \alpha_{ij}^{\frac{1}{2}} P_{ij}(u_k|_{\Gamma_{ij}}), & \Gamma_{ij} \subset \partial\Omega_k, \\ 0, & \Gamma_{ij} \not\subset \partial\Omega_k, \end{cases}$$

and define $\bar{B} : V(\Omega) \to W(\Gamma)$ by

$$\bar{B}v = \sum_{k=1}^{N} \bar{B}_k v_k, \quad v \in V(\Omega).$$

It is easy to see that $\bar{B}v = 0$ if and only if $Bv = 0$. Thus, the system (2.4) has the same solution with the *weighted* saddle-point problem

$$\begin{cases} (A + d^{-1}\bar{B}^t\bar{B})\bar{u} + B^t\lambda = f, \\ B\bar{u} = 0. \end{cases} \tag{2.9}$$

Note that the operator $A^* = A + d^{-1}\bar{B}^t\bar{B}$ is also symmetric and positive definite on $V(\Omega)$ and generates a sparse stiffness matrix such as A.

Since the operator A^* is not yet block diagonal, it is not practical to eliminate directly the variable $\bar{u}$ in (2.9). Fortunately, many iterative methods have been developed for solving saddle-point problems such as (2.9), for example the inexact Uzawa-type methods (see [2], [9], [20], and [21]), the preconditioned MINRES (i.e., conjugate residual) method (see [30]), and the PCG method based on a positive definite reformulation (see [10]). For all the iterative methods, the projections P and P_R used in [24] need not be introduced (since A^* is positive definite).

In order to illustrate an advantage of the new method, we consider the preconditioned MINRES method for solving the system (2.9). Define

$$M = \begin{pmatrix} A^* & B^t \\ B & 0 \end{pmatrix}, \quad U = \begin{pmatrix} \bar{u} \\ \lambda \end{pmatrix}, \quad F = \begin{pmatrix} f \\ 0 \end{pmatrix}.$$

The system (2.9) can be written as

$$MU = F. \tag{2.10}$$

Let $\bar{A} : V(\Omega) \to V(\Omega)$ be a preconditioner for A^*, and let $\bar{S} : W(\Gamma) \to W(\Gamma)$ be a preconditioner for the Schur complement $\tilde{S} = B\bar{A}^{-1}B^t$. Define the block-diagonal preconditioner $\bar{M}$ for M by

$$\bar{M}^{-1} = \begin{pmatrix} \bar{A}^{-1} & 0 \\ 0 & \bar{S}^{-1} \end{pmatrix}.$$

Then the preconditioned MINRES method for (2.10) is the MINRES method (see [30]) applied to the preconditioned system

$$\bar{M}^{-1}MU = \bar{M}^{-1}F. \tag{2.11}$$

The convergence speed of this method is determined by the condition number $cond(\bar{M}^{-1}M)$, which depends on $cond(\bar{A}^{-1}A^*)$ and $cond(\bar{S}^{-1}\tilde{S})$.

In the next section, we consider the construction of the preconditioners $\bar{A}$ and $\bar{S}$. As we will see, $\bar{A}^{-1}$ has the form $\bar{A}^{-1} = A_0^{-1} + \hat{A}^{-1}$, where A_0 is a coarse solver, and $\hat{A}$ is a block-diagonal operator. Moreover, $\bar{S}$ is also a block-diagonal operator. Thus, the coarse solver needs to be implemented only once at each MINRES iteration for (2.11). For the method proposed in [24], two similar coarse solvers (contained, respectively, in the projections P and P_R) need to be implemented at least twice at each MINRES iteration. Besides, the local solvers in $\bar{S}$ are also cheaper than the existing local solvers.

3. General results. In this section we present the main resuts of this paper. Here we do not involve concrete structures of the spaces $V(\Omega)$ and $W(\Gamma)$.

For convenience, following [38], the symbols $\lesssim$, $\gtrsim$, and $\eqsim$ will be used in the rest of this paper. $x_1 \lesssim y_1,\ x_2 \gtrsim y_2$, and $x_3 \eqsim y_3$ mean that $x_1 \le C_1 y_1,\ x_2 \ge c_2 y_2$, and $c_3 x_3 \le y_3 \le C_3 x_3$ for some constants $C_1,\ c_2,\ c_3$, and C_3 that are independent of the dimensions of the approximate spaces $V(\Omega)$ and $W(\Gamma)$. Define $S = B(A^*)^{-1}B^t$.

We first construct a preconditioner for A^*. To handle the "nonlocal" operator $\bar{B}^t\bar{B}$, we need a *coarse* solver.

Define $V_0 = \ker(A)$, which consists of piecewise constant functions. The natural choice of the coarse solver is the restriction of A^* on V_0. But, for numerical consideration, we can also define the coarse solver $A_0 : V_0 \to V_0$ by

$$(A_0 v_0, w_0) = d \sum_{\Gamma_{ij}} \alpha_{ij}(a_i - a_j)(b_i - b_j), \quad v_0 \in V_0 \quad \forall w_0 \in V_0,$$

where $a_k = v_0|_{\Omega_k}$ and $b_k = w_0|_{\Omega_k}$ (const.). It is easy to see that the *coarse* solver A_0 is positive definite on V_0. The action of A_0^{-1} is very cheap to implement, which is similar to that of the coarse solver in the substructuring preconditioner for *two-dimensional* problems (refer to [7]). Since the space $W(\Gamma_{ij})$ contains the piecewise constant function, we have $P_{ij}(v_{0i}|_{\Gamma_{ij}} - v_{0j}|_{\Gamma_{ij}}) = a_i - a_j$ (const.) for $v_0 \in V_0$. Thus, the operator A_0 is spectrally equivalent to the restriction operator of A^* on V_0. Namely,

$$(A_0 v_0, v_0) \eqsim d^{-1}\langle \bar{B}v_0, \bar{B}v_0\rangle = (A^* v_0, v_0) \quad \forall v_0 \in V_0. \tag{3.1}$$

Note that the coarse solver A_0 is much simpler and cheaper than the one in the classical substructuring method for three-dimensional problems (compare [8]).

Define the (local) solver $\bar{A}_k : V(\Omega_k) \to V(\Omega_k)$ by

$$(\bar{A}_k v, w)_{\Omega_k} = \alpha_k[(\nabla v, \nabla w)_{\Omega_k} + d^{-2}(v, w)_{\Omega_k}], \quad v \in V(\Omega_k) \quad \forall w \in V(\Omega_k).$$

In some applications, we would like to consider an inexact solver for $\bar{A}_k$. Let $\hat{A}_k$ be a symmetric and positive definite operator on $V(\Omega_k)$. Assume that there is a positive number γ_k, which may depend slightly on the dimension of $V(\Omega_k)$, such that

$$(\bar{A}_k v, v)_{\Omega_k} \lesssim (\hat{A}_k v, v)_{\Omega_k} \lesssim \gamma_k (\bar{A}_k v, v)_{\Omega_k} \quad \forall v \in V(\Omega_k). \tag{3.2}$$

Define $\hat{A} = \mathrm{diag}(\hat{A}_1,\ \hat{A}_2, \ldots, \hat{A}_N)$. Let $Q_0 : V(\Omega) \to V_0$ denote the L^2 orthogonal projection. Define $\gamma = \max_{1\le k\le N}(\beta_k/\alpha_k)$ and $\hat{\gamma} = \max_{1\le k\le N}\gamma_k$.

THEOREM 3.1. *Define* $\bar{A}^{-1} = \hat{A}^{-1} + A_0^{-1}Q_0$. *Then*

$$cond(\bar{A}^{-1}A^*) \lesssim \gamma \cdot \hat{\gamma}, \tag{3.3}$$

which is independent of the large variants of $a(x)$ *across the local interfaces* Γ_{ij}.

Remark 3.1. When Ω_k is a regular domain, the action of $\bar{A}_k^{-1}$ can be implemented by FFT, which is much cheaper than that of the exact solver corresponding to A_k. For such Ω_k, we can define $\hat{A}_k = \bar{A}_k$, and so $\gamma_k = 1$. In general, the operator $\hat{A}_k$ can be chosen as the (algebraic) multigrid preconditioner, the hierarchical basis preconditioner, or the ILU preconditioner for $\bar{A}_k$. For the case of the multigrid preconditioner, we have $\gamma_k \lesssim 1$.

We next construct a preconditioner for the Schur complement $\tilde{S}$ or S.

To explain our motivation, we first consider a direct choice of the preconditioner.

Let $B_k^t : W(\Gamma) \to V(\Omega_k)$ be the adjoint of B_k, which is defined by

$$(B_k^t\mu, v)_{\Omega_k} = \langle \mu, B_k v\rangle = \langle \mu, B_k v\rangle_{\partial\Omega_k} \quad \forall \mu \in W(\Gamma),\ v \in V(\Omega_k),$$

and define the operator $R_k : W(\Gamma) \to V(\Omega_k)$ by $R_k = \bar{A}_k^{-1}B_k^t$. Let $I_{ij} : W(\Gamma) \to W(\Gamma_{ij})$ denote the natural restriction operator and $I_{ij}^t : W(\Gamma_{ij}) \to W(\Gamma)$ denote the zero extention operator, which is just the adjoint of I_{ij}.

For a face Γ_{ij}, let $\langle\cdot,\cdot\rangle_{\Gamma_{ij}}$ denote the L^2 inner product on the local interface Γ_{ij}. Define the operator $S_{ij} : W(\Gamma_{ij}) \to W(\Gamma_{ij})$ by

$$\begin{aligned}\langle S_{ij}\lambda_{ij}, \mu_{ij}\rangle_{\Gamma_{ij}} &= (\bar{A}_i R_i I_{ij}^t \lambda_{ij}, R_i I_{ij}^t \mu_{ij})_{\Omega_i} \\ &+ (\bar{A}_j R_j I_{ij}^t \lambda_{ij}, R_j I_{ij}^t \mu_{ij})_{\Omega_j} \quad \forall \mu_{ij} \in W(\Gamma_{ij}).\end{aligned}$$

It can be verified that S_{ij} is spectrally equivalent to the restriction of the operator $\tilde{S}$ (or S) on $W(\Gamma_{ij})$.

By the general framework given in [17], we can define a preconditioner by (refer to [19])

$$\hat{S}^{-1} = \sum_{\Gamma_{ij}} I_{ij}^t S_{ij}^{-1} I_{ij}. \tag{3.4}$$

Since S_{ij} results in a dense stiffness matrix, the action of S_{ij}^{-1} is expensive to implement (when $W(\Gamma_{ij})$ has high dimensions). Because of this, we will construct another preconditioner by replacing S_{ij} with a cheaper local solver.

Throughout this paper, we assume that the local multiplier space $W(\Gamma_{ij})$ is *associated* with the local trace space $V_i(\Gamma_{ij})$ (to distinguish it from $V_j(\Gamma_{ij})$) in the sense that $\dim(W(\Gamma_{ij})) = \dim(V_i^0(\Gamma_{ij}))$ and $W(\Gamma_{ij})$ has a vector space structure similar to that of $V_i^0(\Gamma_{ij})$ (refer to the mortar element method in [6] and [4]). Here we do not require $W(\Gamma_{ij}) \subset V_i(\Gamma_{ij})$. It is also possible that $W(\Gamma_{ij})$ is independent of $V_i(\Gamma_{ij})$ and $V_j(\Gamma_{ij})$ (refer to [19] and [26]).

When the coefficient $a(x)$ has a large variation across the local interface Γ_{ij}, we need a particular choice of the index i. Throughout this paper, we always assume that

H_1: the index i is chosen such that $\alpha_i \le \alpha_j$ (so $\alpha_{ij} = \alpha_i$).

Note that the particular choice of the index i will not influence applications of our method.

Remark 3.2. When $V_i^0(\Gamma_{ij}) = V_j^0(\Gamma_{ij})$ (namely, the meshes on Γ_{ij} are matching), we need not choose a particular index i. If the two local trace spaces on Γ_{ij} have the

relation that either a space is a subspace of another one or a space has much smaller dimension than another one the index i can also be chosen such that $V_i^0(\Gamma_{ij}) \subset V_j^0(\Gamma_{ij})$ or $\dim(V_i^0(\Gamma_{ij})) \ll \dim(V_j^0(\Gamma_{ij}))$. But, for simplicity of exposition, we will not discuss this choice in this paper (refer to [16]).

Let $\|\cdot\|_{1/2,\Gamma_{ij}^0}$ denote the norm on the space $H_{00}^{1/2}(\Gamma_{ij})$ (namely, the norm $\|\cdot\|_{H_{00}^{1/2}(\Gamma_{ij})}$ in [28]). Define the *discrete* dual seminorm $\|\cdot\|_{-*,\Gamma_{ij}}$ by

$$\|\mu\|_{-*,\Gamma_{ij}} = \sup_{v\in V_i^0(\Gamma_{ij})} \frac{|\langle \mu, v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}^0}}, \quad \mu \in W(\Gamma_{ij}).$$

For a face Γ_{ij}, let $n_0(i,j)$ denote the dimension of the space $W(\Gamma_{ij})$. Assume that

H_2: for each face Γ_{ij} there is a number $G(n_0(i,j)) > 1$ such that the inverse estimate holds:

$$\|\mu\|_{0,\Gamma_{ij}} \le G(n_0(i,j))\|\mu\|_{-*,\Gamma_{ij}} \quad \forall \mu \in W(\Gamma). \tag{3.5}$$

Hypothesis H_2 implies that $\|\cdot\|_{-*,\Gamma_{ij}}$ is a norm on $W(\Gamma_{ij})$.

For each face Γ_{ij}, let Λ_{ij} be a symmetric and positive definite operator on $W(\Gamma_{ij})$. As the desired local solver, we hope that Λ_{ij} is spectrally equivalent to S_{ij}. But, for convenience, we assume that

H_3: norm $(\langle \Lambda_{ij}\cdot,\cdot\rangle_{\Gamma_{ij}})^{1/2}$ is spectrally equivalent to dual norm $\alpha_i^{-1/2}\|\cdot\|_{-*,\Gamma_{ij}}$.

Now we define the preconditioner $\bar{S}$ by

$$\bar{S}^{-1} = \sum_{\Gamma_{ij}} I_{ij}^t \Lambda_{ij}^{-1} I_{ij}.$$

To estimate the condition number of the preconditioned system $\bar{S}^{-1}S$, we further assume that

H_4: when $W(\Gamma_{ij}) \not\subset V_i(\Gamma_{ij})$, the inequality holds:

$$G(n_0(i,j)) \cdot \inf_{v_\delta \in V_i(\Gamma_{ij})} \|v - v_\delta\|_{0,\Gamma_{ij}} \lesssim \|v\|_{\frac{1}{2},\Gamma_{ij}} \quad \forall v \in V_j(\Gamma_{ij}). \tag{3.6}$$

Set $G(n) = \max_{\Gamma_{ij}\subset\Gamma} G(n_0(i,j))$.

THEOREM 3.2. *Let conditions* H_1–H_4 *be satisfied. Then we have*

$$cond(\bar{S}^{-1}\tilde{S}) \le C\hat{\gamma}[1 + \log(dG^2(n))]^2 \tag{3.7}$$

and

$$cond(\bar{S}^{-1}S) \le C[1 + \log(dG^2(n))]^2. \tag{3.8}$$

Moreover, the constant C *in* (3.7) *and* (3.8) *can be bounded by* γ^2*, which is independent of the large variations of the coefficient* $a(x)$ *across the local interfaces* Γ_{ij}.

Theorems 3.1 and 3.2 will be proved in the next section. From their proofs, we will see that

$$\lambda_{\min}(\bar{A}^{-1}A^*) \gtrsim \hat{\gamma}^{-1} \quad \text{and} \quad \lambda_{\min}(\bar{S}^{-1}\tilde{S}) \gtrsim \hat{\gamma}^{-1}\gamma^{-1}.$$

Thus, using Lemma 2.1 of [30], together with (3.3) and (3.7), we deduce the following.

COROLLARY. *Let $\bar{M}$ and M be defined as in the last section. Then*

$$cond(\bar{M}^{-1}M \le C\hat{\gamma}[1+\log(dG^2(n))]^2. \tag{3.9}$$

Moreover, the constant C in (3.9) *can be bounded by γ^2, which is independent of the large variations of the coefficient $a(x)$ across the local interfaces Γ_{ij}.*

In section 5, we will estimate the positive number $G(n_0(i,j))$ for different choices of $W(\Gamma_{ij})$. For example, for the finite element discretization we have $G(n_0(i,j)) \lesssim h_i^{-1/2}$, where h_i denotes the diameter of the triangulation on Ω_i. A kind of cheap *local* solver Λ_{ij}^{-1} will also be derived in section 5, which needs only $O(n_0^2(i,j))$ arithmetic operations without any particular requirement to $W(\Gamma_{ij})$.

4. Analyses. This section is devoted to prove the results given in the last section.

4.1. On Theorem 3.1. To prove Theorem 3.1, we need a lemma.

LEMMA 4.1. *The following inequality holds:*

$$d^{-1}(\bar{B}^t\bar{B}v,v) \lesssim (\hat{A}v,v) \quad \forall v \in V(\Omega). \tag{4.1}$$

Proof. As in section 2, we set $v|_{\Omega_k} = v_k$. For any index k, the number of indices l such that $\langle \bar{B}_k v_k, \bar{B}_l v_l\rangle \neq 0$ is independent of N (which equals six at most). Thus,

$$(\bar{B}^t\bar{B}v,v) = \langle \bar{B}v, \bar{B}v\rangle = \left\| \sum_{k=1}^N \bar{B}_k v_k \right\|_{0,\Gamma}^2 \lesssim \sum_{k=1}^N \|\bar{B}_k v_k\|_{0,\Gamma}^2. \tag{4.2}$$

By the definition of the operator $\bar{B}_k$, we have

$$\|\bar{B}_k v_k\|_{0,\Gamma}^2 = \sum_{\Gamma_{kj}\subset\partial\Omega_k} \alpha_{kj}\|P_{kj}(v_k|_{\Gamma_{kj}})\|_{0,\Gamma_{kj}}^2 \le \sum_{\Gamma_{kj}\subset\partial\Omega_k} \alpha_{kj}\|v_k\|_{0,\Gamma_{kj}}^2,$$

which implies that

$$\begin{aligned}\sum_{k=1}^N \|\bar{B}_k v_k\|_{0,\Gamma}^2 &\le \sum_{\Gamma_{ij}} \alpha_{ij}(\|v_i\|_{0,\Gamma_{ij}}^2 + \|v_j\|_{0,\Gamma_{ij}}^2)\\ &\le \sum_{\Gamma_{ij}} (\alpha_i\|v_i\|_{0,\Gamma_{ij}}^2 + \alpha_j\|v_j\|_{0,\Gamma_{ij}}^2)\\ &= \sum_{k=1}^N \alpha_k\|v_k\|_{0,\partial\Omega_k}^2.\end{aligned}$$

Substituting the above inequality into (4.2) and using the trace theorem lead to

$$\begin{aligned}d^{-1}(\bar{B}^t\bar{B}v,v) &\lesssim \sum_{k=1}^N \alpha_k(d^{-1}\|v_k\|_{0,\partial\Omega_k}^2)\\ &\lesssim \sum_{k=1}^N \alpha_k(|v_k|_{1,\Omega_k}^2 + d^{-2}\|v_k\|_{0,\Omega_k}^2) = \sum_{k=1}^N (\bar{A}v_k, v_k)_{\Omega_k},\end{aligned}$$

which, together with (3.2), yields (4.1). ∎

Proof of Theorem 3.1. The inequality (3.3) can be derived by

$$\hat{\gamma}^{-1}(v, A^*v) \lesssim ((\hat{A}^{-1} + A_0^{-1}Q_0)A^*v, A^*v) \lesssim \gamma(v, A^*v) \quad \forall v \in V(\Omega).$$

Consider the space decomposition $V(\Omega) = V_0 + \bar{V}$, with $\bar{V} \subset V(\Omega)$. By the abstract Schwarz theory, we need only to prove that

(a) for any $\varphi_0 \in V_0$ and $\bar{\varphi} \in \bar{V}$, we have

$$(A^*(\varphi_0 + \bar{\varphi}), \varphi_0 + \bar{\varphi}) \lesssim \gamma[(A_0\varphi_0, \varphi_0) + (\hat{A}\bar{\varphi}, \bar{\varphi})]; \tag{4.3}$$

(b) for any $v \in V(\Omega)$, there is a decomposition $v = v_0 + \bar{v}$ with $v_0 \in V_0$ and $\bar{v} \in \bar{V}$ such that

$$(A_0v_0, v_0) + (\hat{A}\bar{v}, \bar{v}) \lesssim \hat{\gamma}(A^*v, v). \tag{4.4}$$

We first prove (a). By the triangle inequality and (3.1), we have

$$\begin{aligned}(A^*(\varphi_0 + \bar{\varphi}), \varphi_0 + \bar{\varphi}) &\le 2[(A^*\varphi_0, \varphi_0) + (A^*\bar{\varphi}, \bar{\varphi})] \\ &\lesssim (A_0\varphi_0, \varphi_0) + (A\bar{\varphi}, \bar{\varphi}) + d^{-1}(\bar{B}^t\bar{B}\bar{\varphi}, \bar{\varphi}).\end{aligned} \tag{4.5}$$

But (2.8) and (3.2) imply that

$$\begin{aligned}(A\bar{\varphi}, \bar{\varphi}) &\le \sum_{k=1}^{N} \beta_k |\bar{\varphi}|^2_{1,\Omega_k} \le \sum_{k=1}^{N} (\beta_k/\alpha_k)(\bar{A}_k(\bar{\varphi}|_{\Omega_k}), \bar{\varphi}|_{\Omega_k})_{\Omega_k} \\ &\lesssim \sum_{k=1}^{N} (\beta_k/\alpha_k)(\hat{A}_k(\bar{\varphi}|_{\Omega_k}), \bar{\varphi}|_{\Omega_k})_{\Omega_k} \le \gamma(\hat{A}\bar{\varphi}, \bar{\varphi}).\end{aligned}$$

Substituting this inequality into (4.5) and using (4.1) yield (4.3).

Now we prove (b). Let $\gamma_{\Omega_k}(v_k)$ denote the average value of v_k on Ω_k. Define $v_0 \in V_0$ as follows: for the internal subdomains Ω_k, define $v_0|_{\Omega_k} = \gamma_{\Omega_k}(v_k)$; otherwise, $v_0|_{\Omega_k} = 0$. Moreover, we define $\bar{v} = v - v_0$. Thus, by Friedrich's inequality, we obtain

$$d^{-2}\alpha_k(\bar{v}_k, \bar{v}_k)_{\Omega_k} = d^{-2}\alpha_k \|v_k - (v_0|_{\Omega_k})\|^2_{0,\Omega_k} \lesssim \alpha_k |v_k|^2_{1,\Omega_k}. \tag{4.6}$$

Using (3.2) and (4.6), we deduce

$$\begin{aligned}(\hat{A}\bar{v}, \bar{v}) &= \sum_{k=1}^{N} (\hat{A}_k\bar{v}_k, \bar{v}_k)_{\Omega_k} \le \sum_{k=1}^{N} \gamma_k(\bar{A}_k\bar{v}_k, \bar{v}_k)_{\Omega_k} \\ &\lesssim \sum_{k=1}^{N} \gamma_k\alpha_k(|\bar{v}_k|^2_{1,\Omega_k} + |v_k|^2_{1,\Omega_k}) = 2\sum_{k=1}^{N} \gamma_k\alpha_k |v_k|^2_{1,\Omega_k}.\end{aligned}$$

This, together with (2.8), leads to

$$(\hat{A}\bar{v}, \bar{v}) \lesssim \hat{\gamma}(A^*v, v). \tag{4.7}$$

On the other hand, we have

$$\begin{aligned}(A_0v_0, v_0) &= (A^*v_0, v_0) = (A^*(v - \bar{v}), v - \bar{v}) \\ &\le 2[(A^*v, v) + (A^*\bar{v}, \bar{v})].\end{aligned} \tag{4.8}$$

It follows by (4.1) and (4.7) that

$$(A^*\bar{v},\bar{v}) \lesssim (\hat{A}\bar{v},\bar{v}) \lesssim \hat{\gamma}(A^*v,v).$$

Substituting the above inequality into (4.8) yields

$$(A_0v_0,v_0) \lesssim \hat{\gamma}(A^*v,v),$$

which, together with (4.7), gives (4.4). ∎

4.2. On Theorem 3.2. In this subsection we prove Theorem 3.2. We consider only the inequality (3.7). The inequality (3.8) can be proved in the same way. To prove (3.7), we need some auxiliary results.

Consider the natural space decomposition $W(\Gamma)=\sum_{\Gamma_{ij}} I_{ij}^t W(\Gamma_{ij})$. The following result can be derived by Theorem 2.1 of [17] (refer to [33]). This result can be regarded as a variant of the abstract Schwarz theory (see [32] and [37]).

LEMMA 4.2. *Assume that the following conditions are satisfied:*

(i) *for each* $\mu \in W(\Gamma)$, *we have*

$$\sum_{\Gamma_{ij}} \langle \Lambda_{ij} I_{ij}\mu, I_{ij}\mu\rangle_{\Gamma_{ij}} \lesssim C_1\hat{\gamma}\langle \tilde{S}\mu,\mu\rangle; \tag{4.9}$$

(ii) *for any* $\phi_{ij} \in W(\Gamma_{ij})$, *we have*

$$\left\langle \tilde{S}\left(\sum_{\Gamma_{ij}} I_{ij}^t\phi_{ij}\right), \sum_{\Gamma_{ij}} I_{ij}^t\phi_{ij}\right\rangle \lesssim C_2[1+\log(dG^2(n))]^2 \sum_{\Gamma_{ij}} \langle \Lambda_{ij}\phi_{ij},\phi_{ij}\rangle_{\Gamma_{ij}}. \tag{4.10}$$

Then the inequality (3.7) *holds with* $C \lesssim C_1C_2$.

The above lemma gives a convenient way to estimate the condition number of the preconditioned Schur complements. To estimate the constants C_1 and C_2 in Lemma 4.2, we need to study carefully a discrete dual norm on the local boundary $\partial\Omega_k$ $(k=1,\dots,N)$.

Define $V(\partial\Omega_k)=V(\Omega_k)|_{\partial\Omega_k}$ and $W(\partial\Omega_k)=W(\Gamma)|_{\partial\Omega_k}$. Let $\|\cdot\|_{-*,\partial\Omega_k}$ be the *discrete* dual norm defined by

$$\|\mu\|_{-*,\partial\Omega_k} = \sup_{v\in V(\partial\Omega_k)} \frac{|\langle \mu,v\rangle_{\partial\Omega_k}|}{\|v\|_{\frac{1}{2},\partial\Omega_k}}, \quad \mu \in W(\partial\Omega_k),$$

with

$$\|v\|_{\frac{1}{2},\partial\Omega_k} = (|v|^2_{\frac{1}{2},\partial\Omega_k} + d^{-1}\|v\|^2_{0,\partial\Omega_k})^{\frac{1}{2}}.$$

For a function $\mu \in W(\Gamma)$, define $\pm\mu \in W(\Gamma)$ by

$$(\pm\mu)|_{\Gamma_{ij}} = \sigma_{ij}(\mu|_{\Gamma_{ij}}) \quad \text{for each } \Gamma_{ij}\subset\Gamma.$$

LEMMA 4.3. *For any index* k, *we have*

$$\beta_k^{-1}\|\pm\mu\|^2_{-*,\partial\Omega_k} \lesssim (\bar{A}_kR_k\mu,R_k\mu)_{\Omega_k} \lesssim \alpha_k^{-1}\|\pm\mu\|^2_{-*,\partial\Omega_k} \quad \forall\mu\in W(\Gamma). \tag{4.11}$$

Proof. We first prove that for any extension $\tilde{v}$ of $v\in V(\partial\Omega_k)$ which satisfies $\tilde{v}\in V(\Omega_k)$ and $\tilde{v}|_{\partial\Omega_k}=v$ we have

$$\langle \mu, B_k\tilde{v}\rangle = \langle \pm\mu, v\rangle_{\partial\Omega_k}. \tag{4.12}$$

In fact, by the definitions of the operator B_k, we deduce

$$\begin{aligned}\langle \mu, B_k\tilde{v}\rangle &= \sum_{\Gamma_{kj}\subset\partial\Omega_k} \langle \mu, \sigma_{kj}P_{kj}(v|_{\Gamma_{kj}})\rangle_{\Gamma_{kj}}\\ &= \sum_{\Gamma_{kj}\subset\partial\Omega_k} \langle \sigma_{kj}(\mu|_{\Gamma_{kj}}), P_{kj}(v|_{\Gamma_{kj}})\rangle_{\Gamma_{kj}}\\ &= \sum_{\Gamma_{kj}\subset\partial\Omega_k} \langle \sigma_{kj}(\mu|_{\Gamma_{kj}}), v\rangle_{\Gamma_{kj}}.\end{aligned}$$

This implies (4.12). Here we have used the fact that $\sigma_{kj}(\mu|_{\Gamma_{kj}}) \in W(\Gamma_{kj})$.

We then prove the second inequality of (4.11).

Since $R_k = \bar{A}_k^{-1}B_k^t$ on $W(\Gamma)$, it follows by (4.12) that

$$\begin{aligned}(\bar{A}_kR_k\mu, R_k\mu)_{\Omega_k} &= (B_k^t\mu, R_k\mu)_{\Omega_k}\\ &= \langle \mu, B_kR_k\mu\rangle = \langle \pm\mu, R_k\mu\rangle_{\partial\Omega_k}.\end{aligned} \tag{4.13}$$

By the Cauchy inequality and the trace theorem, we obtain

$$\begin{aligned}\langle \pm\mu, R_k\mu\rangle_{\partial\Omega_k} &\le \|\pm\mu\|_{-*,\partial\Omega_k}\cdot\|R_k\mu\|_{\frac{1}{2},\partial\Omega_k}\\ &\lesssim \|\pm\mu\|_{-*,\partial\Omega_k}\cdot\|R_k\mu\|_{1,\Omega_k}.\end{aligned} \tag{4.14}$$

Using (4.14) and the equality

$$\alpha_k\|R_k\mu\|_{1,\Omega_k}^2 = (\bar{A}_kR_k\mu, R_k\mu)_{\Omega_k}$$

leads to

$$\langle \pm\mu, R_k\mu\rangle_{\partial\Omega_k} \lesssim \alpha_k^{-\frac{1}{2}}\|\pm\mu\|_{-*,\partial\Omega_k}\cdot((\bar{A}_kR_k\mu, R_k\mu)_{\Omega_k})^{\frac{1}{2}}.$$

Substituting the above inequality into (4.13) yields

$$(\bar{A}_kR_k\mu, R_k\mu)_{\Omega_k} \lesssim \alpha_k^{-1}\|\pm\mu\|_{-*,\partial\Omega_k}^2. \tag{4.15}$$

Now we prove the first inequality of (4.11).

For any $v \in V(\partial\Omega_k)$ there is an extension $\tilde{v} \in V(\Omega_k)$ such that $\tilde{v}|_{\partial\Omega_k} = v$ and

$$\|\tilde{v}\|_{1,\Omega_k} \lesssim \|v\|_{\frac{1}{2},\partial\Omega_k}. \tag{4.16}$$

Using the definition of the operator R_k and the Cauchy inequality yields

$$\begin{aligned}|\langle \mu, B_k\tilde{v}\rangle_{\partial\Omega_k}| &= |(B_k^t\mu, \tilde{v})_{\Omega_k}| = |(\bar{A}_kR_k\mu, \tilde{v})_{\Omega_k}|\\ &\le ((\bar{A}_kR_k\mu, R_k\mu)_{\Omega_k})^{\frac{1}{2}}\cdot((\bar{A}_k\tilde{v}, \tilde{v})_{\Omega_k})^{\frac{1}{2}}.\end{aligned}$$

This, together with (4.16) and the inequality

$$(\bar{A}_k\tilde{v}, \tilde{v})_{\Omega_k} \le \beta_k\|\tilde{v}\|_{1,\Omega_k}^2,$$

leads to

$$|\langle \mu, B_k\tilde{v}\rangle_{\partial\Omega_k}| \lesssim ((\bar{A}_kR_k\mu, R_k\mu)_{\Omega_k})^{\frac{1}{2}}\cdot\beta_k^{\frac{1}{2}}\|v\|_{\frac{1}{2},\partial\Omega_k}.$$

Thus, it follows by (4.12) that

$$\begin{aligned}\|\pm\mu\|_{-*,\partial\Omega_k} &= \sup_{v\in V(\partial\Omega_k)} frac|\langle\mu, B_k\tilde{v}\rangle_{\partial\Omega_k}|\|v\|_{\frac{1}{2},\partial\Omega_k}\\ &\lesssim \beta_k^{\frac{1}{2}}((\bar{A}_k R_k\mu, R_k\mu)_{\Omega_k})^{\frac{1}{2}},\end{aligned}$$

which, together with (4.15), gives the desired result. □

When estimating the constant C_2, some extension results of the discrete dual norm $\|\cdot\|_{-*,\partial\Omega_k}$ will play a key role (refer to [19]). Before proving these results, we give two restriction operators and their properties.

For an open face $\Gamma_{ij}\subset\Gamma$, make a regular and quasi-uniform triangulation with the diameter $h_{ij} = 1/G^2(n_0(i,j))$. Let $\mathcal{N}_{ij}$ and $Z(\Gamma_{ij})$ denote, respectively, the set of the corresponding nodes and the space which consists of continuous piecewise linear polynomials associated with this triangulation. For $v_{ij}\in Z(\Gamma_{ij})$, we define $I^0_{\partial\Gamma_{ij}}v_{ij},\ I^0_{\Gamma_{ij}}v_{ij}\in Z(\Gamma_{ij})$ by

$$I^0_{\partial\Gamma_{ij}}v_{ij}(s) = \begin{cases} v_{ij}(s) & \text{if } s\in\mathcal{N}_{ij}\cap\partial\Gamma_{ij},\\ 0 & \text{if } s\in\mathcal{N}_{ij}\cap\Gamma_{ij}\end{cases}$$

and

$$I^0_{\Gamma_{ij}}v_{ij}(s) = \begin{cases} v_{ij}(s) & \text{if } s\in\mathcal{N}_{ij}\cap\Gamma_{ij},\\ 0 & \text{if } s\in\mathcal{N}_{ij}\cap\partial\Gamma_{ij}.\end{cases}$$

It is easy to see that $v_{ij} = I^0_{\partial\Gamma_{ij}}v_{ij} + I^0_{\Gamma_{ij}}v_{ij}$ on $\bar{\Gamma}_{ij}$.

The following two inequalities can be found in [12] and [38].

LEMMA 4.4. *Let $v_{ij}\in Z(\Gamma_{ij})$. Then*

$$\|I^0_{\Gamma_{ij}}v_{ij}\|_{\frac{1}{2},\Gamma_{ij}} \lesssim [1+\log(d/h_{ij})]\|v_{ij}\|_{\frac{1}{2},\Gamma_{ij}} \tag{4.17}$$

and

$$\|I^0_{\partial\Gamma_{ij}}v_{ij}\|_{0,\partial\Gamma_{ij}} \lesssim [1+\log(d/h_{ij})]^{\frac{1}{2}}\|v_{ij}\|_{\frac{1}{2},\Gamma_{ij}}. \tag{4.18}$$

The following result can be viewed as a variant of the $H^{-\frac{1}{2}}$-extension proved in [18] (see also [19]).

LEMMA 4.5. *Let hypothesis* H_2 *hold. Then, for each $\Gamma_{ij}\subset\Gamma$, we have*

$$\sup_{v\in V_i(\Gamma_{ij})}\frac{|\langle\mu_{ij}, v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} \lesssim [1+\log(dG^2(n_0(i,j)))]\|\mu_{ij}\|_{-*,\Gamma_{ij}} \quad \forall\mu_{ij}\in W(\Gamma_{ij}). \tag{4.19}$$

Proof. For any $v\in V_i(\Gamma_{ij})$, let v_{ij} be the L^2 projection of v on $Z(\Gamma_{ij})$. Then

$$\begin{aligned}|\langle\mu_{ij}, v\rangle_{\Gamma_{ij}}| &\le |\langle\mu_{ij}, v-v_{ij}\rangle_{\Gamma_{ij}}|\\ &\quad + |\langle\mu_{ij}, v_{ij}\rangle_{\Gamma_{ij}}|.\end{aligned} \tag{4.20}$$

It is well known that

$$\|v_{ij}-v\|_{0,\Gamma_{ij}} \le Ch_{ij}^{\frac{1}{2}}\|v\|_{\frac{1}{2},\Gamma_{ij}}.$$

This, together with (3.5), leads to

$$\begin{aligned}|\langle \mu_{ij}, v - v_{ij}\rangle_{\Gamma_{ij}}| &\le \|\mu_{ij}\|_{0,\Gamma_{ij}} \cdot \|v - v_{ij}\|_{0,\Gamma_{ij}} \\ &\lesssim \|\mu_{ij}\|_{0,\Gamma_{ij}} \cdot h_{ij}^{\frac{1}{2}} \|v\|_{\frac{1}{2},\Gamma_{ij}} \\ &\lesssim \|\mu_{ij}\|_{-*,\Gamma_{ij}} \cdot \|v\|_{\frac{1}{2},\Gamma_{ij}}. \end{aligned} \tag{4.21}$$

Here we have used the relation $h_{ij}^{\frac{1}{2}} = 1/G(n_0(i,j))$. Since

$$v_{ij} = I^0_{\Gamma_{ij}} v_{ij} + I^0_{\partial\Gamma_{ij}} v_{ij},$$

we have

$$\begin{aligned}|\langle \mu_{ij}, v_{ij}\rangle_{\Gamma_{ij}}| &\le |\langle \mu_{ij}, I^0_{\Gamma_{ij}} v_{ij}\rangle_{\Gamma_{ij}}| + |\langle \mu_{ij}, I^0_{\partial\Gamma_{ij}} v_{ij}\rangle_{\Gamma_{ij}}| \\ &\le |\langle \mu_{ij}, I^0_{\Gamma_{ij}} v_{ij}\rangle_{\Gamma_{ij}}| + \|\mu_{ij}\|_{0,\Gamma_{ij}} \cdot \|I^0_{\partial\Gamma_{ij}} v_{ij}\|_{0,\Gamma_{ij}} \\ &\le \|\mu_{ij}\|_{-*,\Gamma_{ij}} \cdot \|I^0_{\Gamma_{ij}} v_{ij}\|_{\frac{1}{2},\Gamma_{ij}} \\ &\quad + \|\mu_{ij}\|_{0,\Gamma_{ij}} \cdot h_{ij}^{\frac{1}{2}} \|I^0_{\partial\Gamma_{ij}} v_{ij}\|_{0,\partial\Gamma_{ij}}, \end{aligned} \tag{4.22}$$

where a direct computation is used to bound the term $\|I^0_{\Gamma_{ij}} v_{ij}\|_{\frac{1}{2},\Gamma_{ij}}$ by $h_{ij}^{\frac{1}{2}} \|I^0_{\partial\Gamma_{ij}} v_{ij}\|_{0,\partial\Gamma_{ij}}$ using the discrete L^2 norm. Substituting (4.17), (4.18), and (3.5) into (4.22), we obtain

$$|\langle \mu_{ij}, v_{ij}\rangle_{\Gamma_{ij}}| \lesssim [1 + \log(d/h_{ij})] \|\mu_{ij}\|_{-*,\Gamma_{ij}} \cdot \|v_{ij}\|_{\frac{1}{2},\Gamma_{ij}},$$

which, together with (4.20) and (4.21), yields

$$|\langle \mu_{ij}, v\rangle_{\Gamma_{ij}}| \lesssim [1 + \log(d/h_{ij})] \|\mu_{ij}\|_{-*,\Gamma_{ij}} \cdot \|v\|_{\frac{1}{2},\Gamma_{ij}}.$$

Here we have used the fact

$$\|v_{ij}\|_{\frac{1}{2},\Gamma_{ij}} \lesssim \|v\|_{\frac{1}{2},\Gamma_{ij}}.$$

Therefore, we deduce (4.19) (note that $h_{ij} = 1/G^2(n_0(i,j))$). ∎

LEMMA 4.6. *Let conditions* H_2 *and* H_4 *be satisfied. Then, for each* $\Gamma_{ij} \subset \Gamma$*, we have*

$$\|I^t_{ij} \mu_{ij}\|_{-*,\partial\Omega_i} \lesssim [1 + \log(dG^2(n_0(i,j)))] \|\mu_{ij}\|_{-*,\Gamma_{ij}} \quad \forall \mu_{ij} \in W(\Gamma_{ij}) \tag{4.23}$$

and

$$\|I^t_{ij} \mu_{ij}\|_{-*,\partial\Omega_j} \lesssim [1 + \log(dG^2(n_0(i,j)))] \|\mu_{ij}\|_{-*,\Gamma_{ij}} \quad \forall \mu_{ij} \in W(\Gamma_{ij}). \tag{4.24}$$

Proof. Since

$$\|v\|_{\frac{1}{2},\partial\Omega_i} \ge \|v\|_{\frac{1}{2},\Gamma_{ij}} \quad \forall v \in V_i(\partial\Omega_i),$$

the inequality (4.23) is a direct consequence of (4.19).

Now we consider (4.24). By (4.19), it suffices to prove that

$$\sup_{v \in V_j(\Gamma_{ij})} \frac{|\langle \mu_{ij}, v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} \lesssim \sup_{v \in V_i(\Gamma_{ij})} \frac{|\langle \mu_{ij}, v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} + \|\mu_{ij}\|_{-*,\Gamma_{ij}} \quad \forall \mu_{ij} \in W(\Gamma_{ij}). \tag{4.25}$$

Let $P_{ij}^i : L^2(\Gamma_{ij}) \to V_i(\Gamma_{ij})$ denote the L^2 projection. Then, for any $v \in V_j(\Gamma_{ij})$, we have

$$\frac{|\langle \mu_{ij}, v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} \le \begin{cases} \frac{|\langle \mu_{ij}, P_{ij}^i v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} & \text{if } W(\Gamma_{ij}) \subset V_i(\Gamma_{ij}), \\ \frac{|\langle \mu_{ij}, P_{ij}^i v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} + \frac{|\langle \mu_{ij}, v - P_{ij}^i v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} & \text{if } W(\Gamma_{ij}) \not\subset V_i(\Gamma_{ij}). \end{cases} \tag{4.26}$$

Using the Cauchy inequality and (3.5) yields

$$\begin{aligned} |\langle \mu_{ij}, v - P_{ij}^i v\rangle_{\Gamma_{ij}}| &\le \|\mu_{ij}\|_{0,\Gamma_{ij}} \cdot \|v - P_{ij}^i v\|_{0,\Gamma_{ij}} \\ &\lesssim G(n_0(i,j))\|\mu_{ij}\|_{-*,\Gamma_{ij}} \cdot \inf_{v_\delta \in V_i(\Gamma_{ij})} \|v - v_\delta\|_{0,\Gamma_{ij}}. \end{aligned}$$

Thus, it follows by condition H_4 that

$$\frac{|\langle \mu_{ij}, v - P_{ij}^i v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} \lesssim \|\mu_{ij}\|_{-*,\Gamma_{ij}} \quad (W(\Gamma_{ij}) \not\subset V_i(\Gamma_{ij})). \tag{4.27}$$

Note that $\|v\|_{\frac{1}{2},\Gamma_{ij}} \gtrsim \|P_{ij}^i v\|_{\frac{1}{2},\Gamma_{ij}}$, and using (4.26), together with (4.27), we obtain

$$\sup_{v \in V_j(\Gamma_{ij})} \frac{|\langle \mu_{ij}, v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}}} \lesssim \sup_{v \in V_j(\Gamma_{ij})} \frac{|\langle \mu_{ij}, P_{ij}^i v\rangle_{\Gamma_{ij}}|}{\|P_{ij}^i v\|_{\frac{1}{2},\Gamma_{ij}}} + \|\mu_{ij}\|_{-*,\Gamma_{ij}}.$$

Since $P_{ij}^i v \in V_i(\Gamma_{ij})$, this leads to (4.25). ∎

The following lemma is a direct consequence of the following relations:

$$I_{ij}^t(V_i^0(\Gamma_{ij})) \subset V(\partial\Omega_i) \text{ and } I_{ij}^t(V_j^0(\Gamma_{ij})) \subset V(\partial\Omega_j).$$

LEMMA 4.7. *For each face $\Gamma_{ij} \subset \Gamma$ we have*

$$\|\mu\|_{-*,\Gamma_{ij}} \lesssim \| \pm \mu\|_{-*,\partial\Omega_i} \quad \forall \mu \in W(\Gamma) \tag{4.28}$$

and

$$\sup_{v \in V_j^0(\Gamma_{ij})} \frac{|\langle \mu, v\rangle_{\Gamma_{ij}}|}{\|v\|_{\frac{1}{2},\Gamma_{ij}^0}} \lesssim \| \pm \mu\|_{-*,\partial\Omega_j} \quad \forall \mu \in W(\Gamma). \tag{4.29}$$

For ease of notation, define $S_0 = BA_0^{-1}Q_0B^t$.

LEMMA 4.8. *For $\mu \in W(\Gamma)$, define $\mu_{ij} = I_{ij}\mu \in W(\Gamma_{ij})$. Let assumptions* H_1 *and* H_3 *hold. Then*

$$\langle S_0\mu, \mu\rangle \lesssim [1 + \log(dG^2(n))]^2 \sum_{\Gamma_{ij}} \langle \Lambda_{ij}\mu_{ij}, \mu_{ij}\rangle_{\Gamma_{ij}}. \tag{4.30}$$

Proof. Define $u_0 = A_0^{-1}Q_0B^t\mu$ $(\in V_0)$. Then

$$\begin{aligned} \langle S_0\mu, \mu\rangle &= \langle Bu_0, \mu\rangle = (Q_0B^t\mu, u_0) = (Au_0, u_0) \\ &\eqsim d^{-1}(\bar{B}^t\bar{B}u_0, u_0) = d^{-1}\|\bar{B}u_0\|_{0,\Gamma}^2. \end{aligned} \tag{4.31}$$

In the following we estimate $d^{-1}\|\bar{B}u_0\|_{0,\Gamma}^2$ by using the relation

$$d^{-1}\|\bar{B}u_0\|_{0,\Gamma}^2 \eqsim \langle \mu, Bu_0\rangle. \tag{4.32}$$

From the definitions of V_0, B, and $\bar{B}$, we know that

$$(\bar{B}u_0)|_{\Gamma_{ij}} = \alpha_{ij}^{\frac{1}{2}}(Bu_0)|_{\Gamma_{ij}}$$

is a constant for each Γ_{ij}, which is denoted by a_{ij}^0. Using (4.32) and the Cauchy–Schwarz inequality yields

$$\begin{aligned} d^{-1}\|\bar{B}u_0\|_{0,\Gamma}^2 &\eqsim \sum_{\Gamma_{ij}} a_{ij}^0 \alpha_{ij}^{-\frac{1}{2}} \langle \mu, 1\rangle_{\Gamma_{ij}} \\ &\le \left(\sum_{\Gamma_{ij}} (a_{ij}^0)^2\right)^{\frac{1}{2}} \cdot \left(\sum_{\Gamma_{ij}} \alpha_{ij}^{-1}\langle \mu, 1\rangle_{\Gamma_{ij}}^2\right)^{\frac{1}{2}} \\ &\eqsim d^{-1}\|\bar{B}u_0\|_{0,\Gamma} \cdot \left(\sum_{\Gamma_{ij}} \alpha_{ij}^{-1}\langle \mu, 1\rangle_{\Gamma_{ij}}^2\right)^{\frac{1}{2}}. \end{aligned}$$

Thus,

$$\|\bar{B}u_0\|_{0,\Gamma}^2 \lesssim \sum_{\Gamma_{ij}} \alpha_{ij}^{-1} |\langle \mu, 1\rangle_{\Gamma_{ij}}|^2. \tag{4.33}$$

Since $V(\partial\Omega_i)$ contains the constant function, we have

$$\begin{aligned} |\langle \mu, 1\rangle_{\Gamma_{ij}}| &= |\langle I_{ij}^t \mu_{ij}, 1\rangle_{\partial\Omega_i}| \\ &\le \|I_{ij}^t \mu_{ij}\|_{-*,\partial\Omega_i} \cdot \|1\|_{\frac{1}{2},\partial\Omega_i} \\ &\eqsim d^{\frac{1}{2}} \|I_{ij}^t \mu_{ij}\|_{-*,\partial\Omega_i}. \end{aligned}$$

Substituting the above inequality into (4.33) and using (4.23) lead to

$$d^{-1}\|\bar{B}u_0\|_{0,\Gamma}^2 \lesssim [1 + \log(dG(n))]^2 \sum_{\Gamma_{ij}} \alpha_{ij}^{-1} \|\mu_{ij}\|_{-*,\Gamma_{ij}}^2.$$

Hence, from (4.31) we have

$$\langle S_0\mu, \mu\rangle \lesssim [1 + \log(dG(n))]^2 \sum_{\Gamma_{ij}} \alpha_{ij}^{-1} \|\mu_{ij}\|_{-*,\Gamma_{ij}}^2. \tag{4.34}$$

Note that $\alpha_{ij}^{-1} = \alpha_i^{-1}$ (see assumption H_1), and by (4.34) and the condition H_3 we obtain (4.30). Now we can prove Theorem 3.2 easily.

Proof of Theorem 3.2. By Lemma 4.2, we need only to estimate the constants C_1 and C_2 in (4.9) and (4.10).

For simplicity of exposition, we assume that $\gamma_k \eqsim 1$. It is easy to see that ($\mu \in W(\Gamma)$)

$$\begin{aligned} \langle \tilde{S}\mu, \mu\rangle &= \sum_{k=1}^N \langle B_k \hat{A}_k^{-1} B_k^t \mu, \mu\rangle_{\partial\Omega_k} + \langle BA_0^{-1}Q_0 B^t\mu, \mu\rangle \\ &\eqsim \sum_{k=1}^N (\bar{A}_k^{-1} B_k^t \mu, B_k^t \mu)_{\Omega_k} + \langle S_0\mu, \mu\rangle \\ &= \sum_{k=1}^N (\bar{A}_k R_k \mu, R_k \mu)_{\Omega_k} + \langle S_0\mu, \mu\rangle. \end{aligned} \tag{4.35}$$

We first estimate the constant C_1.

By condition H_3, we have

$$\langle \Lambda_{ij} I_{ij}\mu, I_{ij}\mu\rangle_{\Gamma_{ij}} \lesssim \alpha_i^{-1} \|I_{ij}\mu\|^2_{-*,\Gamma_{ij}}.$$

This, together with Lemmas 4.7 and 4.3, leads to

$$\langle \Lambda_{ij} I_{ij}\mu, I_{ij}\mu\rangle_{\Gamma_{ij}} \lesssim \frac{\beta_i}{\alpha_i}(\bar{A}_i R_i\mu, R_i\mu)_{\Omega_i}.$$

Summing over Γ_{ij} to the above inequality and using (4.35) yield (4.9) with $C_1 \le \max_{1\le k\le N}(\beta_k/\alpha_k) = \gamma$.

We next estimate the constant C_2.

For ease of notation, define $\phi = \sum_{\Gamma_{ij}} I_{ij}^t \phi_{ij}$. Then, from (4.35), we have

$$\begin{aligned}\left\langle \tilde{S}\left(\sum_{\Gamma_{ij}} I_{ij}^t\phi_{ij}\right), \sum_{\Gamma_{ij}} I_{ij}^t\phi_{ij}\right\rangle &= \langle \tilde{S}\phi, \phi\rangle \\ &\eqsim \sum_{k=1}^{N} (\bar{A}_k R_k\phi, R_k\phi)_{\Omega_k} + \langle S_0\phi, \phi\rangle. \end{aligned} \tag{4.36}$$

It follows by Lemma 4.3 that

$$\begin{aligned}(\bar{A}_k R_k\phi, R_k\phi)_{\Omega_k} &\lesssim \alpha_k^{-1} \| \pm \phi\|^2_{-*,\partial\Omega_k} \\ &\lesssim \alpha_k^{-1} \sum_{\Gamma_{kj}\subset\partial\Omega_k} \| \pm I_{kj}^t(\phi|_{\Gamma_{kj}})\|^2_{-*,\partial\Omega_k}.\end{aligned}$$

Substituting the above inequality into (4.36) and noting that $\phi|_{\Gamma_{ij}} = \phi_{ij}$ yield

$$\begin{aligned}\left\langle \tilde{S}\left(\sum_{\Gamma_{ij}} I_{ij}^t\phi_{ij}\right), \sum_{\Gamma_{ij}} I_{ij}^t\phi_{ij}\right\rangle &\lesssim \langle S_0\phi, \phi\rangle \\ + \sum_{\Gamma_{ij}} [\alpha_i^{-1} \|I_{ij}^t\phi_{ij}\|^2_{-*,\partial\Omega_i} &+ \alpha_j^{-1} \|I_{ij}^t\phi_{ij}\|^2_{-*,\partial\Omega_j}].\end{aligned}$$

This, together with Lemma 4.6, leads to

$$\left\langle \tilde{S}\left(\sum_{\Gamma_{ij}} I_{ij}^t\phi_{ij}\right), \sum_{\Gamma_{ij}} I_{ij}^t\phi_{ij}\right\rangle \lesssim \langle S_0\phi, \phi\rangle + [1+\log(dG^2(n))]^2 \sum_{\Gamma_{ij}} \alpha_i^{-1} \|\phi_{ij}\|^2_{-*,\Gamma_{ij}}. \tag{4.37}$$

Thus, Lemma 4.8 and condition H_3 yield (4.10) with $C_2 \le \gamma$. ∎

5. Detailed discussions. In this section, we discuss some details on the preconditioner $\bar{S}$. Theorem 3.2 indicates that the condition number of the preconditioned Schur complements is determined mainly by the positive numbers $G(n_0(i,j))$ in the inverse estimate (3.5). However, the verification of such an inverse estimate depends on the underlying approximate spaces.

5.1. Approximate spaces. As usual, we assume that each Ω_k is a polyhedron. We consider both the finite element space and the spectral element space.

(I) pure finite element discretizations.

With each subdomain Ω_k we associate a regular and quasi-uniform triangulation $\mathcal{T}_k$ made of elements that are either hexahedra or tetrahedra. We denote by h_k the mesh size of $\mathcal{T}_k$; i.e., h_k denotes the maximum diameter of any tetrahedra in the mesh $\mathcal{T}_k$. The triangulations in the subdomains are independent of one other and generally do not match at the interfaces between subdomains. Hence, each interface Γ_{ij} is provided with two different (two-dimensional) meshes $\mathcal{T}_{ij}$ and $\mathcal{T}_{ji}$, which are associated with $\mathcal{T}_i$ and $\mathcal{T}_j$, respectively. Define $V(\Omega_k)$ as the space consisting of continuous piecewise linear functions associated with $\mathcal{T}_k$. For each $\Gamma_{ij} \subset \Gamma$, we choose $\mathcal{T}_{ij}$ or $\mathcal{T}_{ji}$ to define the *local* multiplier space $W(\Gamma_{ij})$; for example, choose $\mathcal{T}_{ij}$. There are many ways to define the local Lagrange multiplier space $W(\Gamma_{ij})$.

For the case of triangulation made of parallelepipeds, the multiplier space $W(\Gamma_{ij})$ can be defined as the tensor product of two one-dimensional multiplier spaces. Thus, we consider only the case where the face Γ_{ij} is meshed with triangular elements.

Case (i). $W(\Gamma_{ij})$ is the *mortar* space (refer to [6], [3], and [4]).

We denote by x_m, $1 \le m \le n(i,j)$, all the nodes on Γ_{ij} associated with the triangulation $\mathcal{T}_{ij}$ and distinguish the internal nodes in Γ_{ij} (numbered from 1 to $n_0(i,j)$) from the boundary notes which belong to $\partial\Gamma_{ij}$ (numbered from $n_0(i,j)+1$ to $n(i,j)$). All these nodes are associated with the shape functions ϕ_m so that each function φ in $W(\Gamma_{ij})$ can be written as

$$\varphi = \sum_{m=1}^{n(i,j)} \varphi(x_m)\phi_m,$$

and each function φ in $V_i^0(\Gamma_{ij})$ can be written as

$$\varphi = \sum_{m=1}^{n_0(i,j)} \varphi(x_m)\phi_m.$$

For a node x_m on $\partial\Gamma_{ij}$ $(n_0(i,j)+1 \le m \le n(i,j))$, only the following two cases are possible: (a) there are $Q(m)$ internal nodes $x_m^l \in \Gamma_{ij}$ such that each x_m^l and x_m are just two end points of an edge of some triangle on Γ_{ij}; (b) all three vertices of the triangle containing x_m are on the boundary of Γ_{ij} (such x_m is a vertex of Γ_{ij}). For case (b), let x_m^l $(l = 1, 2)$ denote the other two vertices of the triangle containing x_m, and set $Q(m) = 2$. Now we choose $Q(m)$ positive numbers a_m^l satisfying

$$\sum_{l=1}^{Q(m)} a_m^l = 1$$

and define

$$\varphi(x_m) = \sum_{l=1}^{Q(m)} a_m^l \varphi(x_m^l).$$

The definition of the space $W(\Gamma_{ij})$ is then

$$W(\Gamma_{ij}) = \left\{ \varphi \in V_i(\Gamma_{ij}) : \forall m,\ n_0(i,j)+1 \le m \le n(i,j),\ \varphi(x_m) = \sum_{l=1}^{Q(m)} a_m^l \varphi(x_m^l) \right\}.$$

It is easy to see that the above space can also be written as

$$W(\Gamma_{ij}) = \left\{ \varphi \in V_i(\Gamma_{ij}) : \varphi = \sum_{m=1}^{n_0(i,j)} \varphi(x_m)\phi_m + \sum_{m=n_0(i,j)+1}^{n(i,j)} \left[\sum_{l=1}^{Q(m)} a_m^l \varphi(x_m^l) \right] \phi_m \right\}.$$

Case (ii). $W(\Gamma_{ij})$ is the *dual basis* space (see [23] and [36]).

We will define a map F_{ij} which takes $V_i^0(\Gamma_{ij})$ to the space of discontinuous functions which are linear when restricted to the triangles of $\mathcal{T}_{ij}$. Let τ be a triangle with vertices $\{y_l,\ l = 1, 2, 3\}$ and v_l denote the value of a function $\varphi \in V_i^0(\Gamma_{ij})$ at y_l. We define F_{ij} by the following rules:

1. If all three vertices of τ are in Γ_{ij}, then we set $F_{ij}\varphi = w$, where w is the linear function with values $w_1 = 3v_1 - v_2 - v_3$, $w_2 = 3v_2 - v_1 - v_3$, and $w_3 = 3v_3 - v_1 - v_2$.

2. If exactly one vertex (say y_1) of τ is on $\partial\Gamma_{ij}$, then we set $w_1 = (v_2 + v_3)/2$, $w_2 = (5v_2 - 3v_3)/2$, and $w_3 = (5v_3 - 3v_2)/2$.

3. If exactly one vertex (say y_1) of τ is in Γ_{ij}, then we set $w_1 = w_2 = w_3 = v_1$.

4. If none of the vertices of τ are in Γ_{ij} then we set $w_1 = w_2 = w_3 = v_l$, where v_l is value of φ at the interior vertex which is closest to the triangle.

Let $\{x_l : l = 1, \ldots, n_0(i,j)\}$ be the nodes in Γ_{ij}. We get a dual basis by defining $\psi_l = F_{ij}\varphi_l$, for $l = 1, \ldots, n_0(i,j)$. In fact, it easily follows from the above definitions that $\{\psi_l : l = 1, \ldots, n_0(i,j)\}$ is linearly independent and satisfies $(\varphi_l, \psi_m) = 0$ whenever $l \neq m$. We define $W(\Gamma_{ij})$ to be the span of $\{\psi_l : l = 1, \ldots, n_0(i,j)\}$. Note that $W(\Gamma_{ij}) \not\subset V_i(\Gamma_{ij})$.

The existence and uniqueness of the solution $\bar{u}$ of (2.3) and the error estimates of $\bar{u}$ have been shown, respectively, in [4] (for Case (i)) and in [23] (for Case (ii)).

(II) coupling spectral and finite element discretization (refer to [5]).

We use spectral discretization on some subdomains, but we use finite element discretization on the other subdomains. Consider an interface Γ_{ij}. Without loss of generality, we assume that $V(\Omega_i)$ is the spectral space of all polynomial functions with degree $\leq n_i$, but $V(\Omega_j)$ is the usual linear finite element space (see (I)).

When $\alpha_i \leq \alpha_j$, we choose the multiplier space $W(\Gamma_{ij})$ as the spectral space of all polynomial functions with degree $\leq n_i - 2$. Otherwise, we choose $W(\Gamma_{ij})$ as a suitable linear finite element space associated with the triangulation $\mathcal{T}_{ji}$ (see [5] for details).

The existence and uniqueness of the solution $\bar{u}$ of (2.3) and the error estimates of $\bar{u}$ have been shown in [5] under suitable assumptions.

Note that for all the cases considered above, we have $\dim(W(\Gamma_{ij})) = \dim(V_i^0(\Gamma_{ij}))$.

5.2. Estimate of the positive numbers $G(n_0(i,j))$. The following result can be derived directly by the definition of $\|\cdot\|_{-*,\Gamma_{ij}}$.

THEOREM 5.1. *Assume that there exists a positive number $E(n_0(i,j))$ such that for any $\mu_{ij} \in W(\Gamma_{ij})$ there is a function $\varphi \in V_i^0(\Gamma_{ij})$ satisfying*

$$|\langle \mu_{ij}, \varphi \rangle_{\Gamma_{ij}}| \geq E(n_0(i,j)) \|\mu_{ij}\|_{0,\Gamma_{ij}} \cdot \|\varphi\|_{0,\Gamma_{ij}}. \tag{5.1}$$

Then

$$G(n_0(i,j)) \leq E_1(n_0(i,j))/E(n_0(i,j)), \tag{5.2}$$

where $E_1(n_0(i,j))$ is defined by the inverse estimate

$$\|v\|_{\frac{1}{2},\Gamma_{ij}^0} \leq E_1(n_0(i,j)) \|v\|_{0,\Gamma_{ij}} \quad \forall v \in V_i^0(\Gamma_{ij}). \tag{5.3}$$

Using Theorem 5.1, we can estimate $G(n_0(i,j))$ for various cases.

PROPOSITION 5.1. *For the two cases in discretization* (I), *we have* $G(n_0(i,j)) \lesssim h_i^{-1/2}$.

Proof. From the standard inverse estimate of the finite element functions, we know that the positive number $E_1(n_0(i,j))$ in (5.3) can be estimated by $E_1(n_0(i,j)) \lesssim h_i^{-1/2}$. We need only to prove the positive number $E(n_0(i,j))$ in (5.1) satisfies $E(n_0(i,j)) \gtrsim 1$.

For Case (ii), this inequality has been shown in [23].

Now we consider Case (i). It has been shown in [4] that the *mortar* projection $\pi_{ij} : L^2(\Gamma_{ij}) \to V_i^0(\Gamma_{ij})$ defined by

$$\int_{\Gamma_{ij}} (\varphi - \pi_{ij}\varphi)\mu ds = 0, \quad \varphi \in L^2(\Gamma_{ij}) \quad \forall \mu \in W(\Gamma_{ij})$$

satisfies the L^2 stability

$$\|\pi_{ij}\varphi\|_{0,\Gamma_{ij}} \lesssim \|\varphi\|_{0,\Gamma_{ij}} \quad \forall \varphi \in L^2(\Gamma_{ij}).$$

Thus,

$$\begin{aligned} &\langle \mu_{ij}, \pi_{ij}\mu_{ij}\rangle_{\Gamma_{ij}} = \langle \mu_{ij}, \mu_{ij}\rangle_{\Gamma_{ij}} \\ &\gtrsim \|\mu_{ij}\|_{0,\Gamma_{ij}} \cdot \|\pi_{ij}\mu_{ij}\|_{0,\Gamma_{ij}}, \end{aligned}$$

which means that the function $\varphi = \pi_{ij}\mu_{ij} \in V_i^0(\Gamma_{ij})$ satisfies (5.1) with $E(n_0(i,j)) \gtrsim 1$. ∎

PROPOSITION 5.2. *For discretization* (II), *we assume that the coupling interface* Γ_{ij} *is affinely equivalent to the reference square. Then* $G(n_0(i,j)) \lesssim (n_i^3/d)^{1/2}$ *(or* $G(n_0(i,j)) \lesssim h_j^{-1/2}$*).*

Proof. From the standard inverse estimate of the spectral element, we know that the positive number $E_1(n_0(i,j))$ in (5.3) can be estimated by $E_1(n_0(i,j)) \lesssim n_i d^{-\frac{1}{2}}$. We need only to prove the positive number $E(n_0(i,j))$ in (5.1) satisfies $E(n_0(i,j)) \gtrsim n_i^{-1/2}$. This inequality can be proved as in Case (i) of Proposition 5.1 by using the inequality (3.2) of [31], which is also true for the current situation. ∎

The following result can be derived by Proposition 5.1 and the approximate property of the finite element space $V_i(\Gamma_{ij})$.

PROPOSITION 5.3. *For Case* (ii) *in discretization* (I), *assumption* H_4 *holds.*

The following proposition is a direct consequence of Proposition 5.2 and the approximate property of the finite element space $V_j^0(\Gamma_{ij})$.

PROPOSITION 5.4. *For discretization* (II), *let the coupling interface* Γ_{ij} *be affinely equivalent to the reference square. We assume that* $h_j n_i^3/d$ *is sufficiently small. Then assumption* H_4 *holds.*

Remark 5.1. From the above discussion, we know that the positive number $G(n)$ in (3.7) and (3.8) can be estimated by

$$G^2(n) \lesssim \begin{cases} h^{-1} & \text{for discretization (I)}, \\ \max\{h^{-1}, \hat{n}^3/d\} & \text{for discretization (II)}. \end{cases}$$

Here, $h = \min h_i$ and $\hat{n} = \max n_i$.

Remark 5.2. We believe that when n_i^3 and $\hat{n}^3$ are replaced by n_i^2 and $\hat{n}^2$, respectively, the results in this section still hold. However, the proof seems complicated.

5.3. On the local solvers. It is clear that the preconditioner $\bar{S}$ is determined by the local solvers Λ_{ij} ($\Gamma_{ij} \subset \Gamma$). When the local multiplier space $W(\Gamma_{ij})$ has low dimension $n_0(i,j)$, the action of the inverse S_{ij}^{-1} can be implemented exactly (which needs $O(n_0^3(i,j))$ arithmetic operations), and Λ_{ij} can be chosen as S_{ij} itself (refer to [18] and [19]). Otherwise, we have to develop a cheaper solver Λ_{ij}, which would be spectrally equivalent to S_{ij} (namely, satisfies assumption H_3). For this purpose, we need an auxiliary result.

Let $Q_{ij} : W(\Gamma_{ij}) \to V_i^0(\Gamma_{ij})$ denote the L^2 projection on $V_i^0(\Gamma_{ij})$, and let $Q_{ij}^t : V_i^0(\Gamma_{ij}) \to W(\Gamma_{ij})$ denote its adjoint with respect to the $L^2(\Gamma_{ij})$ inner product.

THEOREM 5.2. *Let $\hat{\Lambda}_{ij} : V_i^0(\Gamma_{ij}) \to V_i^0(\Gamma_{ij})$ be a symmetric and positive definite operator satisfying $\langle \hat{\Lambda}_{ij}\cdot,\cdot\rangle_{\Gamma_{ij}} \eqsim \|\cdot\|^2_{\frac{1}{2},\Gamma_{ij}^0}$. Define $\Lambda_{ij} = \alpha_i^{-1} Q_{ij}^t \hat{\Lambda}_{ij}^{-1} Q_{ij}$. Then the operator Λ_{ij} satisfies hypothesis* H_3.

Proof. Since there is a positive number $E(n_0(i,j))$ satisfying (5.1) in all cases, it can be verified that the operator $Q_{ij} : W(\Gamma_{ij}) \to V_i^0(\Gamma_{ij})$ is nonsingular. Hence, the operator Λ_{ij} is symmetric and positive definite on $W(\Gamma_{ij})$. It suffices to prove that

$$\|\mu_{ij}\|_{-*,\Gamma_{ij}} \eqsim \langle (\hat{\Lambda}_{ij}^{-1} Q_{ij}\mu_{ij}, Q_{ij}\mu_{ij}\rangle_{\Gamma_{ij}}^{\frac{1}{2}} \quad \forall \mu_{ij} \in W(\Gamma_{ij}). \tag{5.4}$$

Define $v_{ij} = \hat{\Lambda}_{ij}^{-1} Q_{ij}\mu_{ij}$. Since $v_{ij} \in V_i^0(\Gamma_{ij})$, we have from the assumption to $\hat{\Lambda}_{ij}$

$$\begin{aligned}
\|\mu_{ij}\|_{-*,\Gamma_{ij}} &\geq \frac{|\langle \mu_{ij}, v_{ij}\rangle_{\Gamma_{ij}}|}{\|v_{ij}\|_{\frac{1}{2},\Gamma_{ij}^0}} = \frac{|\langle Q_{ij}\mu_{ij}, v_{ij}\rangle_{\Gamma_{ij}}|}{\|v_{ij}\|_{\frac{1}{2},\Gamma_{ij}^0}} \\
&= \langle Q_{ij}\mu_{ij}, \hat{\Lambda}_{ij}^{-1} Q_{ij}\mu_{ij}\rangle_{\Gamma_{ij}}^{\frac{1}{2}} \cdot \frac{\langle \hat{\Lambda}_{ij} v_{ij}, v_{ij}\rangle_{\Gamma_{ij}}^{\frac{1}{2}}}{\|v_{ij}\|_{\frac{1}{2},\Gamma_{ij}^0}} \\
&\gtrsim \langle Q_{ij}\mu_{ij}, \hat{\Lambda}_{ij}^{-1} Q_{ij}\mu_{ij}\rangle_{\Gamma_{ij}}^{\frac{1}{2}} \quad \forall \mu_{ij} \in W(\Gamma_{ij}).
\end{aligned} \tag{5.5}$$

On the other hand, we have for any $v \in V_i^0(\Gamma_{ij})$

$$\begin{aligned}
|\langle \mu_{ij}, v\rangle_{\Gamma_{ij}}| &= |\langle Q_{ij}\mu_{ij}, v\rangle_{\Gamma_{ij}}| \\
&= |\langle \hat{\Lambda}_{ij}^{-\frac{1}{2}} Q_{ij}\mu_{ij}, \hat{\Lambda}_{ij}^{\frac{1}{2}} v\rangle_{\Gamma_{ij}}|.
\end{aligned}$$

Thus, using the Cauchy inequality and the assumption, we deduce

$$\begin{aligned}
|\langle \mu_{ij}, v\rangle_{\Gamma_{ij}}| &\leq \langle \hat{\Lambda}_{ij}^{-1} Q_{ij}\mu_{ij}, Q_{ij}\mu_{ij}\rangle_{\Gamma_{ij}}^{\frac{1}{2}} \cdot \langle \hat{\Lambda}_{ij} v, v\rangle_{\Gamma_{ij}}^{\frac{1}{2}} \\
&\lesssim \langle \hat{\Lambda}_{ij}^{-1} Q_{ij}\mu_{ij}, Q_{ij}\mu_{ij}\rangle_{\Gamma_{ij}}^{\frac{1}{2}} \cdot \|v\|_{\frac{1}{2},\Gamma_{ij}^0}.
\end{aligned} \tag{5.6}$$

By the definition of $\|\cdot\|_{-*,\Gamma_{ij}}$, the inequality (5.6) leads to

$$\|\mu_{ij}\|_{-*,\Gamma_{ij}} \lesssim \langle \hat{\Lambda}_{ij}^{-1} Q_{ij}\mu_{ij}, Q_{ij}\mu_{ij}\rangle_{\Gamma_{ij}}^{\frac{1}{2}} \quad \forall \mu_{ij} \in W(\Gamma_{ij}).$$

This, together with (5.5), yields (5.4). ∎

By Theorem 5.2, we need only to define a suitable operator $\hat{\Lambda}_{ij}$ so that the action of the operator Λ_{ij}^{-1} is cheap to implement.

Remark 5.3. Of course, the operator $\hat{\Lambda}_{ij}$ can be chosen as the positive square root of the discrete Laplace operator (refer to [8] and [38]). But this choice of $\hat{\Lambda}_{ij}$ is not practical for a triangular face Γ_{ij} or spectral element discretization.

We first study the action of $(Q_{ij}^t\hat{\Lambda}_{ij}^{-1}Q_{ij})^{-1}$. For $\mu \in W(\Gamma_{ij})$, set $(Q_{ij}^t\hat{\Lambda}_{ij}^{-1}Q_{ij})^{-1}\mu = \lambda \in W(\Gamma_{ij})$. We need only to explain how to get this λ from $\mu \in W(\Gamma_{ij})$. Since the operator Q_{ij} is nonsingular, we have $(Q_{ij}^t\hat{\Lambda}_{ij}^{-1}Q_{ij})^{-1} = Q_{ij}^{-1}\hat{\Lambda}_{ij}(Q_{ij}^t)^{-1}$. Thus, $Q_{ij}\lambda = \hat{\Lambda}_{ij}(Q_{ij}^t)^{-1}\mu$, which implies that

$$\langle Q_{ij}\lambda, \varphi\rangle_{\Gamma_{ij}} = \langle \hat{\Lambda}_{ij}(Q_{ij}^t)^{-1}\mu, \varphi\rangle_{\Gamma_{ij}} \quad \forall \varphi \in V_i^0(\Gamma_{ij}).$$

Namely,

$$\langle \lambda, \varphi\rangle_{\Gamma_{ij}} = \langle \tilde{\mu}, \hat{\Lambda}_{ij}\varphi\rangle_{\Gamma_{ij}} \quad \forall \varphi \in V_i^0(\Gamma_{ij}). \tag{5.7}$$

Here the function $\tilde{\mu}$ satisfies $Q_{ij}^t\tilde{\mu} = \mu$.

As in section 2, set $\dim(W(\Gamma_{ij})) = n_0(i,j)$. Let $\{\varphi_{ij}^m\}$ and $\{\psi_{ij}^m\}$ $(m = 1, \ldots, n_0(i,j))$ denote a basis of the spaces $V_i^0(\Gamma_{ij})$ and $W(\Gamma_{ij})$, respectively. Then the functions μ and λ can be written as

$$\mu = \sum_m a_m\psi_{ij}^m \quad \text{and} \quad \lambda = \sum_m z_m\psi_{ij}^m,$$

respectively. Define the vectors

$$b_{ij} = (a_1, a_2, \ldots, a_{n_0(i,j)})^t, \quad \chi_{ij} = (z_1, z_2, \ldots, z_{n_0(i,j)})^t.$$

Let K_{ij} and M_{ij} denote the coupling matrix and the mass matrix with the entries $\langle\varphi_{ij}^k, \psi_{ij}^l\rangle_{\Gamma_{ij}}$ and $\langle\psi_{ij}^k, \psi_{ij}^l\rangle_{\Gamma_{ij}}$ $(k,l = 1, \ldots, n_0(i,j))$, respectively. Besides, let N_{ij} denote the stiffness matrix of the operator $\hat{\Lambda}_{ij}$ associated with the basis $\{\varphi_{ij}^m\}$. It can be verified by (5.7) that the vector χ_{ij} is determined by

$$K_{ij}\chi_{ij} = N_{ij}\tilde{b}_{ij}, \tag{5.8}$$

with $\tilde{b}_{ij}$ satisfying

$$K_{ij}^t\tilde{b}_{ij} = M_{ij}b_{ij}. \tag{5.9}$$

The above two equations show the action of the local solver $\Lambda_{ij}^{-1} = \alpha_i(Q_{ij}^t\hat{\Lambda}_{ij}^{-1}Q_{ij})^{-1}$. It follows by (5.8) and (5.9) that the matrix form of Λ_{ij}^{-1} is $\alpha_i K_{ij}^{-1}N_{ij}(K_{ij}^{-1})^t$.

From the viewpoint of computation, we need only to know the matrix N_{ij} instead of the operator $\hat{\Lambda}_{ij}$. In the following, we define an appropriate matrix N_{ij} such that the corresponding operator $\hat{\Lambda}_{ij}$ (for a given basis) satisfies the assumption in Theorem 5.2. Since the linear system (5.8) does not involve the action of N_{ij}^{-1}, we can define N_{ij} in a more direct way. This is the main advantage of our preconditioner over the existing preconditioners.

We first consider discretization (I).

For Case (i), both $\{\varphi_{ij}^m\}$ and $\{\psi_{ij}^n\}$ are chosen as the nodal basis functions, but, for Case (ii), $\{\psi_{ij}^n\}$ is chosen as the dual basis of $\{\varphi_{ij}^m\}$. It is clear that both M_{ij} and K_{ij} are sparse and band matrices. In particular, K_{ij} is a diagonal matrix for Case (ii).

Let $\{x_{ij}^k\}$ denote the nodes associated with the triangulation $\mathcal{T}_{ij}$. Define the operator $\hat{\Lambda}_{ij} : V_i^0(\Gamma_{ij}) \to V_i^0(\Gamma_{ij})$ by

$$\langle \hat{\Lambda}_{ij}\varphi,\psi\rangle_{\Gamma_{ij}} = h_i^4 \sum_{\substack{x_{ij}^m,x_{ij}^n\in\mathcal{T}_{ij}\\ x_{ij}^m\neq x_{ij}^n}} \frac{(\varphi(x_{ij}^m)-\varphi(x_{ij}^n))(\psi(x_{ij}^m)-\psi(x_{ij}^n))}{|x_{ij}^m-x_{ij}^n|^3}$$

$$+ h_i^2 \sum_{x_{ij}^m\in\mathcal{T}_{ij}} \frac{\varphi(x_{ij}^m)\cdot\psi(x_{ij}^m)}{dist(x_{ij}^m,\partial\Gamma_{ij})}, \quad \varphi\in V_i^0(\Gamma_{ij}) \quad \forall\psi\in V_i^0(\Gamma_{ij}). \tag{5.10}$$

In particular, for the nodal basis $\{\varphi_{ij}^k\}$ of $V_i^0(\Gamma_{ij})$, we have

$$\langle \hat{\Lambda}_{ij}\varphi_{ij}^m,\varphi_{ij}^n\rangle_{\Gamma_{ij}} = \begin{cases} \sum_{\substack{x_{ij}^l\in\mathcal{T}_{ij}\\ x_{ij}^l\neq x_{ij}^m}} \frac{h_i^4}{|x_{ij}^l-x_{ij}^m|^3} + \frac{h_i^2}{dist(x_{ij}^m,\partial\Gamma_{ij})} & \text{if } n=m,\\ \frac{-h_i^4}{|x_{ij}^m-x_{ij}^n|^3} & \text{if } n\neq m. \end{cases}$$

This equality gives the entries of the matrix N_{ij}.

If choosing $\psi=\varphi$ in (5.10), then the right-hand side of (5.10) is just the discrete form of the norm $\|\varphi\|^2_{\frac{1}{2},\Gamma_{ij}^0}$ (refer to [38]). Therefore, the operator $\hat{\Lambda}_{ij}$ satisfies the assumption in Theorem 5.2.

Since (refer to [38])

$$\|\varphi\|_{\frac{1}{2},\Gamma_{ij}^0} \eqsim [1+\log(d/h_i)]^{\frac{1}{2}}\|\varphi\|_{\frac{1}{2},\Gamma_{ij}}, \quad \varphi\in V_i^0(\Gamma_{ij}),$$

the matrix N_{ij} can be replaced by $[1+\log(d/h_i)]N_{ij}^0$, where the entries of the matrix N_{ij}^0 are defined by

$$c_{mn}^{ij} = \begin{cases} \sum_{\substack{x_{ij}^l\in\mathcal{T}_{ij}\\ x_{ij}^l\neq x_{ij}^m}} \frac{h_i^4}{|x_{ij}^l-x_{ij}^m|^3} + \frac{h_i^2}{d} & \text{if } n=m,\\ \frac{-h_i^4}{|x_{ij}^m-x_{ij}^n|^3} & \text{if } n\neq m. \end{cases}$$

Remark 5.4. The calculation of the matrix N_{ij} is needed only once before implementing a certain iterative algorithm, so its complexity is much smaller than that of the global iterations. We would like to compare the arithmetic complexity of the local solver Λ_{ij}^{-1} with that of the existing local solvers. Since the matrix K_{ij} is sparse and band, solving χ by (5.8) and (5.9) needs only $O(n_0^2(i,j))$ arithmetic operations. This local solver seems cheaper than all the existing local solvers, including the one in the classical substructuring method (if there is no particular assumption to the triangulation). For example, the action of the local solver in the FETI-type methods is implemented by solving a local Dirichlet problem with $O(n_0^{3/2}(i,j))$ unknown, and so it needs $O(n_0^3(i,j))$ arithmetic operations. We point out that we need not make a particular requirement to the meshes on Γ_{ij} here. This is a very important merit in DDMs for three-dimensional problems. If the meshes on Γ_{ij} have a particular structure, we can decrease the arithmetic operations by defining special operator $\hat{\Lambda}_{ij}$.

Now we consider discretization (II) (on parallelepiped). Without loss of generality, we assume that $\Gamma_{ij}=J^2$ with $J=[-1,1]$. Let $L_k(x)$ denote the Legendre polynomial of degree k ($k=0,\ldots,n_i-1$) defined on J. Define

$$\phi_k(x) = \sqrt{k+\frac{1}{2}}\int_{-1}^{x} L_k(t)dt, \quad k=1,\ldots,n_i-1.$$

It is clear that $\phi_k(-1) = \phi_k(1) = 0$. Let the basis functions of $W(\Gamma_{ij})$ and $V_i^0(\Gamma_{ij})$ be defined by

$$\psi_{mn}(x,y) = L_m(x)L_n(y), \quad m,n = 0,\ldots,n_i-2$$

and

$$\varphi_{mn}(x,y) = \phi_m(x)\phi_n(y), \quad m,n = 1,\ldots,n_i-1,$$

respectively. Let K_0 and M_0 denote the coupling matrix and the mass matrix with the entries $\langle \phi_k, L_l\rangle_J$ $(l = 0,\ldots,n_i-2)$ and $\langle \phi_k,\phi_l\rangle_J$ $(l = 1,\ldots,n_i-1)$, respectively $(k = 1,\ldots,n_i-1)$. It is clear that

$$K_{ij} = K_0 \otimes K_0 \quad \text{and} \quad M_{ij} = M_0 \otimes M_0. \tag{5.11}$$

From the orthogonality of the Legendre polynomial, we infer that K_0 and M_0 are sparse and band matrices with bandwidth ≤ 5. Thus, the matrices K_{ij} and M_{ij} are also sparse and band (with constant bandwidth) by (5.11).

To compute the matrix N_{ij}, we need to investigate the norm $\|\cdot\|^2_{1/2,\Gamma_{ij}^0}$. It can be verified directly that (see Theorem 13.1 of [28])

$$\|v\|^2_{\frac{1}{2},\Gamma_{ij}^0} \eqsim \|\varphi\|^2_{H_{00}^{\frac{1}{2}}(J)}\cdot\|\psi\|^2_{0,J} + \|\varphi\|^2_{0,J}\cdot\|\psi\|^2_{H_{00}^{\frac{1}{2}}(J)}, \quad v(x,y) = \varphi(x)\psi(y) \in V_i^0(\Gamma_{ij}). \tag{5.12}$$

Define

$$V^0(J) = span\{\phi_1,\phi_2,\ldots,\phi_{n_i-1}\}.$$

Let $\Lambda_0 : V^0(J) \to V^0(J)$ be a symmetric and positive definite operator, which satisfies

$$\langle \Lambda_0\phi,\phi\rangle_J \eqsim \|\phi\|^2_{H_{00}^{\frac{1}{2}}(J)} \quad \forall \phi \in V^0(J). \tag{5.13}$$

Define

$$\begin{aligned} &\langle \hat{\Lambda}_{ij} v, w\rangle_{\Gamma_{ij}} = \langle \Lambda_0\varphi^1,\psi^1\rangle_J \cdot \langle \varphi^2,\psi^2\rangle_J + \langle \varphi^1,\psi^1\rangle_J \cdot \langle \Lambda_0\varphi^2,\psi^2\rangle_J, \\ &v(x,y) = \varphi^1(x)\varphi^2(y) \in V^0(J) \quad \forall w(x,y) = \psi^1(x)\psi^2(y) \in V^0(J). \end{aligned} \tag{5.14}$$

It follows by (5.13) and (5.14) that the operator $\hat{\Lambda}_{ij}$ satisfies the assumption in Theorem 5.2. We define the operator Λ_0 by

$$\begin{aligned} \langle \Lambda_0\varphi,\psi\rangle_J &= \int_J\int_J \frac{(\varphi(x)-\varphi(y))(\psi(x)-\psi(y))}{(x-y)^2}dxdy \\ &+ \int_J \frac{\varphi(x)\psi(x)}{dist(x,\partial J)}dx, \quad \varphi \in V^0(J) \quad \forall \psi \in V_0(J), \end{aligned}$$

such that the condition (5.13) is satisfied. In particular, we have

$$\begin{aligned} \langle \Lambda_0\phi_m,\phi_n\rangle_J &= \int_J\int_J \frac{(\phi_m(x)-\phi_m(y))(\phi_n(x)-\phi_n(y))}{(x-y)^2}dxdy \\ &+ \int_J \frac{\phi_m(x)\phi_n(x)}{dist(x,\partial J)}dx, \quad m,n = 1,\ldots,n_i-1. \end{aligned} \tag{5.15}$$

Let N_0 denote the stiffness matrix of the operator Λ_0 associated with the basis $\{\phi_k\}$. From (5.14), we have

$$N_{ij} = N_0 \otimes M_0 + M_0 \otimes N_0. \tag{5.16}$$

The matrix M_0 can be obtained easily, which has only $O(n_i)$ nonzero entries (so the matrix N_{ij} has $O(n_i^3)$ nonzero entries at most).

To calculate the matrix N_0, we need to derive a formula to evaluate the inner product $\langle \Lambda_0 \phi_m, \phi_n \rangle_J$. Define

$$I_{m,n}(k,l) = \int_J \int_J \frac{\int_y^x t^k L_m(t)dt \cdot \int_y^x t^l L_n(t)dt}{(x-y)^2} dxdy,$$
$$k = 0, \ldots, n_i - 1 - m;\ l = 0, \ldots, n_i - 1 - n;\ m, n = 1, \ldots, n_i - 1.$$

Then

$$\int_J \int_J \frac{(\phi_m(x) - \phi_m(y))(\phi_n(x) - \phi_n(y))}{(x-y)^2} dxdy = \sqrt{m + \frac{1}{2}}\sqrt{n + \frac{1}{2}} I_{m,n}(0,0).$$

It is well known that

$$L_{k+1}(x) = \frac{2k+1}{k+1} x L_k(x) - \frac{k}{k+1} L_{k-1}(x), \quad k = 1, \ldots, n_i - 1.$$

Thus, we can obtain the following recurrence formulas ($m, n \geq 2$):

$$I_{1,n}(k,l) = \frac{2n+1}{n+1} I_{1,n-1}(k, l+1) - \frac{n}{n+1} I_{1,n-2}(k,l),$$

$$I_{m,1}(k,l) = \frac{2m+1}{m+1} I_{m-1,1}(k+1, l) - \frac{m}{m+1} I_{m-2,1}(k,l),$$

and

$$I_{m,n}(k,l) = \frac{2m+1}{m+1} \cdot \frac{2n+1}{n+1} I_{m-1,n-1}(k+1, l+1) + \frac{m}{m+1} \cdot \frac{n}{n+1} I_{m-2,n-2}(k,l)$$
$$- \frac{2m+1}{m+1} \cdot \frac{n}{n+1} I_{m-1,n-2}(k+1, l) - \frac{m}{m+1} \cdot \frac{2n+1}{n+1} I_{m-2,n-1}(k, l+1).$$

For any $k, l = 0, \ldots, n_i - 2$, the integration $I_{1,1}(k,l)$ can be evaluated easily. Thus, the integration

$$\int_J \int_J \frac{(\phi_m(x) - \phi_m(y))(\phi_n(x) - \phi_n(y))}{(x-y)^2} dxdy$$

can be evaluated by the above recurrence formulas. In an analogous way, we can evaluate the integration $\int_J \frac{\phi_m(x)\phi_n(x)}{dist(x, \partial J)} dx$, and so we can get the value $\langle \Lambda_0 \phi_m, \phi_n \rangle_J$ by (5.15).

Remark 5.5. For the case of the spectral element, the dimension of the spaces $W(\Gamma_{ij})$ and $V_i^0(\Gamma_{ij})$ are $n_0(i,j) = (n_i - 1)^2$. It is easy to see that the computation of the matrix N_0 needs $O(n_i^4) = O(n_0^2(i,j))$ arithmetic operations. This calculation is needed only once in the preprocessing step and thus is minor to global iterations. From the above discussion, we know that solving χ by (5.8) needs only $O(n_i^3) = O(n_0^{\frac{3}{2}}(i,j))$ arithmetic operations. Note that the number of unknowns corresponding to the subdomain Ω_i is $(n_i + 1)^3$. To our knowledge, there is no literature to discuss the FETI method for spectral element discretization.

6. Numerical experiments. In this section, we apply our newly proposed DDM to solve two elliptic problems: the first one with large jumps in the coefficient across interfaces and the second one with a singular solution.

Consider the elliptic equation of the second order

$$\begin{cases} -\mathrm{div}(a\nabla u) = f & \text{in } \Omega, \\ u = g & \text{on } \partial\Omega, \end{cases} \tag{6.1}$$

where Ω is the unit cube: $\Omega = [0,1]^3$.

Let Ω be decomposed into N equal cubes of size d, with the cubes numbered as Ω_1, $\Omega_2, \ldots, \Omega_N$ in the usual way. The coefficient a is defined by $a(x,y,z) = a_k(1+xyz)$ for $(x,y,z) \in \Omega_k$, where a_k is a constant and will be given below. The functions f and g will be determined by the exact solution of (6.1) with continuous a (i.e., $a_k = 1$). Note that the exact solution with jump a is difficult to construct.

The finite element or the spectral element discretization will be used on each subdomain Ω_k. As demonstrated in section 2, the DDM with Lagrange multipliers results in the augmented saddle-point problem (2.9). The corresponding *algebraic* system will be solved by the Uzawa-type method described by Algorithm 3.1 in [21]. This method needs to compute an approximation $\Psi(\cdot)$ in its inner iteration. For our current examples, the approximation $\Psi(\cdot)$ is generated by a one-step multiplicative Schwarz iteration with solvers $\hat{A}_k$ and a coarse solver A_0, where the optimal relaxation factor is added in the standard way. In the case of finite element discretization, $\hat{A}_k$ is chosen to be the multigrid preconditioner corresponding to $\bar{A}_k$; in the case of spectral element discretization, $\hat{A}_k$ is chosen to be $\bar{A}_k$ itself (since the stiffness matrix of $\bar{A}_k$ has a particular structure). For the outer iteration, we use the preconditioner introduced in subsection 5.3.

To investigate the convergence for jump a, we have to choose a termination criterion carefully, which would not be influenced by possible large jumps. Let $|||\cdot|||$ denote the weighted Euclidean norm with the weights 1 (for the primal vector) and $\{\alpha_{ij}^2\}$ (for the multiplier vectors; refer to section 2).

The initial guess is chosen as the zero vector, and the termination criterion ε is defined to be the (relative) residual norm

$$\varepsilon = |||\mathcal{F} - \mathcal{M}\mathcal{U}_n|||/|||\mathcal{F}|||.$$

Here $\mathcal{M}$, $\mathcal{F}$, and $\mathcal{U}$ denote the algebraic form of M, F, and U, respectively. The iteration terminates when $\varepsilon \le 10^{-5}$.

Example 6.1. In (6.1), the constant a_k is defined by

$$a_k = \begin{cases} 1 & \text{if } k \text{ is odd}, \\ 10^{-5} & \text{if } k \text{ is even}. \end{cases}$$

We take the analytic solution to be $u = e^{xyz}$.

Case (i). Each subdomain Ω_k is divided into small cubes of size h, and the standard Q_1 element is used on Ω_k. The multiplier space on a local interface is spanned by the (two-dimensional) mortar multiplier basis, which is the tensor product of two one-dimensional mortar multiplier bases given in [6].

The numerical results are summarized in Table 1.

This table indicates that the iteration counts depend slightly on the ratio d/h and are almost independent of N or h itself.

Case (ii). We use the same discretization on the subdomains near the boundary $\partial\Omega$ as the one used in Case (i), but with the spectral element discretization on the

TABLE 1
Iteration counts for the case with matching grids.

d/h	$d = 1/4$ $(N = 64)$	$d = 1/5$ $(N = 125)$	$d = 1/6$ $(N = 216)$
8	40	42	43
16	45	46	47

TABLE 2
Iteration counts for the coupling discretization.

d/h	n	$d = 1/4$ $(N = 64)$	$d = 1/5$ $(N = 125)$
8	3	40	41
8	5	45	46
16	3	44	45
16	5	45	46

inner subdomains. The basis in the spectral space is defined as the tensor product of three one-dimensional Legendre polynomial bases with degree n. The multiplier space is defined as in subsection 5.1.

The iteration counts are given in Table 2.

This table indicates that the iteration counts are indeed determined by $\max\{d/h, n^2\}$ for the coupling discretization.

Example 6.2. In (6.1), we take the analytic solution to be the singular solution $u(x, y, z) = (x + y + z)^{0.2}$.

Since u is singular at the original point, we use the finite element discretization with nonmatching grids. Set $d = 1/4$ $(N = 64)$ or $d = 1/5$ $(N = 125)$, and divide each cube into small cubes with the size h_1, h_2, or h_3 $(h_1 < h_2 < h_3)$. Here the size h_1 is used in the subdomains $\Omega_k \subset [0, d]^3$, the size h_2 in the subdomains $\Omega_k \subset [0, 2d]^3 \backslash [0, d]^3$, and the size h_3 in the subdomains $\Omega_k \subset \Omega \backslash [0, 2d]^3$. Again the standard Q_1 element is used on each cube. The multiplier is chosen as the tensor product of two one-dimensional dual basis multiplier associated with the coarse triangulation on the interface.

We shall compare the performances for the cases with same a_k and different a_k:

Case 1. The constant $a_k \equiv 1$ for all subdomains Ω_k.

Case 2. The constant a_k is defined by

$$a_k = \begin{cases} 10^6 & \text{if } \Omega_k \subset [0, d]^3, \\ 10^3 & \text{if } \Omega_k \subset [0, 2d]^3 \backslash [0, d]^3, \\ 1 & \text{if } \Omega_k \subset \Omega \backslash [0, 2d]^3. \end{cases}$$

The iteration counts are reported in Tables 3 and 4, respectively.

All the numerical results reported in this section clearly demonstrate the efficiency of our method, and this confirms the theoretical results presented in this paper. In particular, we have seen that the iteration counts are not affected by nonmatching grids or jump coefficients.

7. Conclusions. In this paper we have proposed a new DDM with Lagrange multipliers for solving three-dimensional elliptic problems with variable coefficients. For this method, the singularity on floating domains is handled by combining the augmented method and the preconditioning technique. To construct a preconditioner for the Schur complement, we have developed a class of inexact local solver for three important multiplier spaces, which are less expensive than the existing local solvers. Both the theoretical results and the numerical experiments show that our method is

TABLE 3
Iteration counts for nonmatching grids (Case 1).

d/h_1	d/h_2	d/h_3	$d = 1/4$ $(N = 64)$	$d = 1/5$ $(N = 125)$
16	8	4	39	40
32	16	8	44	45

TABLE 4
Iteration counts for nonmatching grids (Case 2).

d/h_1	d/h_2	d/h_3	$d = 1/4$ $(N = 64)$	$d = 1/5$ $(N = 125)$
16	8	4	41	42
32	16	8	45	46

efficient, even if the grids are nonmatching and the coefficient has large variations across local interfaces.

Acknowledgment. The authors wish to thank two anonymous referees for many constructive comments which led to a great improvement of the results and the presentation of the paper.

REFERENCES

[1] Y. ACHDOU, Y. KUZNETNOV, AND O. PIRONNEAU, *Substructuring preconditioners for the Q_1 mortar element method*, Numer. Math., 71 (1995), pp. 419–449.
[2] R. BANK, B. WELFERT, AND H. YSERENTANT, *A class of iterative methods for solving saddle point problems*, Numer. Math., 56 (1990), pp. 645–666.
[3] F. BELGACEM, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.
[4] F. BELGACEM AND Y. MADAY, *The mortar element method for three dimensional finite elements*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 289–302.
[5] F. BEN BELGACEM AND Y. MADAY, *Coupling spectral and finite elements for second order elliptic three-dimensional equations*, SIAM J. Numer. Anal., 36 (1999), pp. 1234–1263.
[6] C. BERNARDI, Y. MADAY, AND A. T. PATERA, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications. College de France Seminar, Vol. XI (Paris, 1989–1991), Longman Scientific and Technical, Harlow, 1994, pp. 13–51.
[7] J. BRAMBLE, J. PASCIAK, AND A. SCHATZ, *The construction of preconditioners for elliptic problems by substructuring*, I. Math. Comp., 47 (1986), pp. 107–134.
[8] J. BRAMBLE, J. PASCIAK, AND A. SCHATZ, *The construction of preconditioners for elliptic problems by substructuring*, IV. Math. Comp., 53 (1989), pp. 1–24.
[9] J. H. BRAMBLE, J. E. PASCIAK, AND A. T. VASSILEV, *Analysis of the inexact Uzawa algorithm for saddle point problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1072–1092.
[10] J. BRAMBLE AND J. PASCIAK, *A preconditioning technique for indefinite systems resulting from mixied approximations of elliptic problems*, Math. Comp., 50 (1988), pp. 1–18.
[11] Q. DINH, R. GLOWINSKI, AND J. PERIAUX, *Solving elliptic problems by domain decomposition methods with applications*, in Elliptic Problem Solver II, Academic Press, New York, 1982.
[12] M. R. DORR, *On the discretization of interdomain coupling in elliptic boundary-value problems*, in Domain Decomposition Methods, T. F. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1989, pp. 17–37.
[13] C. FARHAT AND F.-X. ROUX, *A method of finite element tearing and interconnecting and its parallel solution algorithm*, Internat. J. Numer. Methods Engrg., 32 (1991), pp. 1205–1227.
[14] C. FARHAT, M. LESOINNE, AND K. PIERSON, *A scalable dual-primal domain decomposition method*, Numer. Linear Algebra Appl., 7 (2000), pp. 687–714.
[15] C. FARHAT, J. MANDEL, AND F.-X. ROUX, *Optimal convergence properties of the FETI domain decomposition method*, Comput. Methods. Appl. Mech. Engrg., 115 (1994), pp. 365–388.
[16] Q. HU, *A new kind of preconditioner for interface equations of mortar multipliers on subspaces*, in Recent Progress in Computational and Applied PDEs, T. F. Chan, Y. Huang, T. Tang, J. Xu, and L. Ying, eds., Kluwer/Plenum, New York, 2002, pp. 205–216.

[17] Q. Hu and G. Liang, *A general framework to construct interface preconditioners*, Chinese J. Numer. Math. Appl., 21 (1999), pp. 83–95.
[18] Q. Hu, *Study of Domain Decomposition Methods with Non-matching Grids*, Ph.D. thesis, Institute of Mathematics, Chinese Academy of Science, Beijing, 1998.
[19] Q. Hu, G. Liang, and J. Lui, *The construction of preconditioner for domain decomposition methods with polynomial multipliers*, J. Comp. Math., 19 (2001), pp. 213–224.
[20] Q. Hu and J. Zou, *An iterative method with variable relaxation parameters for saddle-point problems*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 317–338.
[21] Q. Hu and J. Zou, *Two new variants of nonlinear inexact Uzawa algorithms for saddle-point problems*, Numer. Math., 93 (2002), pp. 333–359.
[22] J. Huang and J. Zou, *A mortar element method for elliptic problems with discontinuous coefficients*, IMA J. Numer. Anal., 22 (2002), pp. 549–576.
[23] C. Kim, R. D. Lazarov, J. E. Pasciak, and P. S. Vassilevski, *Multiplier spaces for the mortar finite element method in three dimensions*, SIAM J. Numer. Anal., 39 (2001), pp. 519–538.
[24] A. Klawonn and O. B. Widlund, *A domain decomposition method with Lagrange multipliers and inexact solvers for linear elasticity*, SIAM J. Sci. Comput., 22 (2000), pp. 1199–1219.
[25] Y. Kuznetsov, *Efficient iterative solvers for elliptic finite element problems on non-matching grids*, Russian J. Numer. Anal. Math. Modelling, 10 (1995), pp. 187–211.
[26] G. Liang and J. He, *The non-conforming domain decomposition method for elliptic problems with Lagrangian multipliers*, Chinese J. Numer. Math. Appl., 15 (1993), pp. 8–19.
[27] G. Liang and P. Liang, *Non-conforming domain decomposition with the hybrid finite element method*, Math. Numer. Sinica, 11 (1989), pp. 323–332.
[28] J. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
[29] J. Mandel and R. Tezaur, *Convergence of a substructuring methods with Lagrangian multipliers*, Numer. Math., 73 (1996), pp. 473–487.
[30] T. Rusten and R. Winther, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.
[31] P. Seshaiyer and M. Suri, *Uniform hp convergence results for the mortar finite element method*, Math. Comp., 69 (2000), pp. 521–546.
[32] B. Smith, P. Bjorstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.
[33] P. Le Tallec, *Domain decomposition methods in computational mechanics*, Comput. Mech. Adv., 1 (1994), pp. 121–220.
[34] P. Le Tallec, T. Sassi, and M. Vidrascu, *Three-Dimensional Domain Decomposition Methods with Nonmatching Grids and Unstructured Coarse Solvers*, Contemp. Math. 180, AMS, Providence, RI, 1994, pp. 61–74.
[35] P. Le Tallec and T. Sassi, *Domain decomposition with nonmatching grids: Augmented Lagrangian approach*, Math. Comp., 64 (1995), pp. 1367–1396.
[36] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.
[37] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[38] J. Xu and J. Zou, *Some nonoverlapping domain decomposition methods*, SIAM Rev., 40 (1998), pp. 857–914.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 934–952

A RELAXATION SCHEME FOR THE NONLINEAR SCHRÖDINGER EQUATION*

CHRISTOPHE BESSE†

Abstract. In this paper, we present a new numerical scheme for the nonlinear Schrödinger equation. This is a relaxation-type scheme that avoids solving for nonlinear systems and preserves density and energy. We give convergence results for the semidiscretized version of the scheme and perform several numerical experiments.

Key words. nonlinear Schrödinger equation, relaxation method

AMS subject classifications. 65M12, 35Q55

DOI. 10.1137/S0036142901396521

1. Introduction and statement of the results. We present in this article a semidiscretized numerical scheme for the nonlinear Schrödinger equation defined for $x \in \mathbb{R}^d$ by

$$\begin{cases} iu_t + \Delta u = \lambda |u|^{2\sigma} u, & t > 0, \\ u(x,0) = u_0(x), \end{cases} \tag{1.1}$$

where we set $\lambda \in \mathbb{R}^*$, $\sigma \in \mathbb{N}^*$, and Δ the classical Laplacian operator.

Both density N and energy E are conserved quantities:

$$N(t) = \int_{\mathbb{R}^d} |u|^2(t,x)dx,$$
$$E(t) = \int_{\mathbb{R}^d} \frac{1}{2}|\nabla u|^2 + \frac{\lambda}{2\sigma+2}|u|^{2\sigma+2}dx.$$

Using these conserved quantities, it is well known that the system (1.1) is globally well posed in $H^1(\mathbb{R}^d)$ if $\lambda < 0$, and blow-up may occur if $\lambda > 0$ (see, for example, [11], [7], [12]).

A large number of articles are devoted to the numerical study of this equation using many different time discretizations. The standard cases use schemes of Crank–Nicolson type [9], Runge–Kutta type [3], [1], [2], symplectic type (see, for example, [13], [14]), or splitting type [15], [6]. Recently, Zouraris [16] proved the convergence of scheme developed by [10]. In [4], Bao, Jin, and Markowich use a time-splitting spectral scheme for the more general nonlinear Schrödinger equation (1.2):

$$i\varepsilon u_t + \frac{\varepsilon^2}{2}\Delta u - \mathcal{V}(x)u - f(|u|^2)u = 0, \quad t > 0, \quad x \in \mathbb{R}^d, \tag{1.2}$$

where $\mathcal{V}$ is a given real-valued electrostatic potential and f is a real-valued smooth function. Typically, for $f(\rho) = \delta\rho$ and $\mathcal{V}(x) = |x|^2$, this equation is called the Gross–Pitaevskii equation. In [4], the authors study the behavior of the scheme with respect to the oscillations due to ε. They show that the splitting scheme gives better results than the classical Crank–Nicolson one with respect to ε.

*Received by the editors October 16, 2001; accepted for publication (in revised form) September 26, 2003; published electronically July 14, 2004.
http://www.siam.org/journals/sinum/42-3/39652.html

†Laboratoire de Mathématiques pour l'Industrie et la Physique, UMR 5640, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 4, France (besse@mip.ups-tlse.fr).

However, only Crank–Nicolson-type schemes succeed in preserving both density and energy. The aim of the present paper is to derive a convergent explicit numerical approximation of (1.1) preserving both conserved quantities.

The scheme is constructed as follows. We rewrite (1.1) as the system of two equations:

$$\left\{ \begin{array}{ll} \phi = |u|^{2\sigma}, \quad t > 0, & \text{(a)} \\ iu_t + \Delta u = \lambda \phi u, \quad t > 0, & \text{(b)} \end{array} \right. \tag{1.3}$$

with $u(x,0) = u_0(x)$.

Let T^* be the existence time of the solution and $T_{\delta t} < T^*$ the computation time. We use N points for the time discretization, thus defining a time step $\delta t = T_{\delta t}/N$. Equations (1.3a) and (1.3b) are discretized at times $t_n = n\delta t$ and $t_{n+\frac{1}{2}} = (n+\frac{1}{2})\delta t$, $n = 1, \ldots, N$, respectively. We define the variables $\phi^{n+\frac{1}{2}}$ and u^{n+1}, which represent the approximations of $|u|^{2\sigma}$ at time $t_{n+\frac{1}{2}}$ and u at time t_{n+1}, respectively. So, we get the semidiscrete-in-time relaxation scheme which reads

$$\left\{ \begin{array}{ll} \dfrac{\phi^{n+\frac{1}{2}} + \phi^{n-\frac{1}{2}}}{2} = |u^n|^{2\sigma}, & \text{(a)} \\ i\dfrac{u^{n+1} - u^n}{\delta t} + \Delta\left(\dfrac{u^{n+1} + u^n}{2}\right) = \lambda\left(\dfrac{u^{n+1} + u^n}{2}\right)\phi^{n+\frac{1}{2}}, & \text{(b)} \end{array} \right. \tag{1.4}$$

with the initial data $u^0(x) = u_0(x)$ and $\phi^{-\frac{1}{2}}(x) = |u^0(x)|^{2\sigma}$. Since only time discretization is involved, the space discretization may be of either finite difference type or finite element type.

Compared to the Crank–Nicolson scheme, which is also based in a time-centering method, this scheme allows us to avoid a costly numerical treatment of the nonlinearity and to preserve the flexibility of spatial discretization choice. On the other hand, we show its capability of conserving the density and energy. There are indications that the order of convergence may be equal to two. We compute in the last section the numerical order p_{num} for different meshes.

The main results of this article are the proof of local existence and uniqueness of a solution in $H^s(\mathbb{R}^d)$, $s > d/2 + 2$, for the scheme (1.4) and its convergence to the solution of (1.1). Moreover, we prove for the case $\sigma = 1$ that density and energy are conserved quantities.

Let us define $u_{\delta t}(t,x) = \sum_{n=0}^{N-1} u^n(x)\mathbf{I}_{[t_n, t_{n+1}[}$ and $\phi_{\delta t}(t,x) = \sum_{n=0}^{N-1} \phi^{n+\frac{1}{2}}(x)\mathbf{I}_{[t_n, t_{n+1}[}$ for $t \in [0, T_{\delta t}]$ and $x \in \mathbb{R}^d$. Then, the results are as follows in Theorem 1.1.

THEOREM 1.1. *Let u_0 belong to $H^s(\mathbb{R}^d)$, $s > d/2 + 2$, and let u be the maximal solution to (1.1) defined in $C^2([0, T^*[; H^s(\mathbb{R}^d))$. Then, there exists a unique maximal solution $(u_{\delta t}, \phi_{\delta t})$ of (1.4) in $L^\infty([0, T_{\delta t}]; (H^s(\mathbb{R}^d))^2)$ which verifies*

$$\sup_{t \in [0, T_{\delta t}]} (\|u_{\delta t}\|_{H^s} + \|\phi_{\delta t}\|_{H^s}) \leq C(T^*, \|u_0\|_{H^s}).$$

If $s > d/2 + 4$, then $\liminf_{\delta t \to 0} T_{\delta t} \geq T^*$ and $\forall T < T^*$, the solution $(u_{\delta t}, \phi_{\delta t})$ to (1.4) converges to $(u, |u|^{2\sigma})$ in $L^\infty([0,T]; (H^s(\mathbb{R}^d))^2)$ as $\delta t \to 0$.

REMARK 1. *In the previous theorem, the solution u is maximal in the sense that if $T^* < \infty$, then $\|u\|_{H^s} \to \infty$ as $t \to T^*$.*

In sections 2 and 3, we prove the local existence of a solution and the convergence of the scheme. We emphasize the main difficulties in proving the theorem and the way to solve them. The scheme is proved to be conservative in section 4. In the last

section, we compute solutions of (1.1) with the relaxation scheme and compare to other schemes and give an estimate of the order of convergence.

An abridged version of this paper as well as the application of this scheme to the Davey–Stewartson system can be found in [5].

2. Local existence and uniqueness of strong solutions. We prove in this section the following proposition.

PROPOSITION 2.1. *Let u_0 belong to $H^s(\mathbb{R}^d)$, $s > d/2+2$, and let u be the maximal solution to* (1.1) *defined in $C^2([0,T^*[;H^s(\mathbb{R}^d))$. Then, there exists a unique maximal solution $(u_{\delta t},\phi_{\delta t})$ of* (1.4) *in $L^\infty([0,T_{\delta t}];(H^s(\mathbb{R}^d))^2)$ which verifies*

$$\sup_{t\in[0,T_{\delta t}]}(\|u_{\delta t}\|_{H^s}+\|\phi_{\delta t}\|_{H^s})\leq C(T^*,\|u_0\|_{H^s}).$$

2.1. Main problems and ideas of the proof. As we stated in the introduction, the relaxation scheme first consists of changing system (1.1) into the system (1.3). Then, both equations in (1.3) are discretized on two staggered time grids.

There are many ways in which we might prove local existence of solutions. Here we detail some solutions.

1. The first idea for proving existence of solutions is to express $\phi^{n+\frac12}$ as a function of $|u^n|^2$ and to plug it into $i\frac{u^{n+1}-u^n}{\delta t}+\Delta(\frac{u^{n+1}+u^n}{2})=\lambda(\frac{u^{n+1}+u^n}{2})\phi^{n+\frac12}$. Unfortunately, the computation of the quantity $\phi^{2p+\frac12}$ for $n=2p$ gives $\phi^{2p+\frac12}=|u_0|^{2\sigma}+\sum_{k=1}^{p}2\delta t\frac{|u^{2k}|^{2\sigma}-|u^{2k-1}|^{2\sigma}}{\delta t}$ (a similar form is obtained for $n=2p+1$). Hence, the term $\phi^{\bullet+\frac12}$ is nothing but the discretization of

$$\phi(t)=|u|^{2\sigma}(0)+\int_0^t\partial_\zeta|u|^{2\sigma}(\zeta)d\zeta. \tag{2.1}$$

Contrary to the continuous evolution equation, where $\phi=|u|^2$, we have a loss of uniformity with respect to time. Hence, we cannot preserve this idea.

2. Another possibility is to consider these equations as a system in which one simultaneously studies the evolution of the two variables ϕ and u. From this, it is possible to use usual fixed point techniques in order to prove the existence and uniqueness of solutions. Thanks to (1.4.a), the discrete derivative of $\phi^{n+1/2}$ is given by $\frac{\phi^{n+\frac12}-\phi^{n-3/2}}{2\delta t}=\frac{|u^n|^{2\sigma}-|u^{n-1}|^{2\sigma}}{\delta t}$. In a sense, the control of the discrete derivative of ϕ depends on the control of the term $\frac{|u^n|^{2\sigma}-|u^{n-1}|^{2\sigma}}{\delta t}$. As $\sigma\in\mathbb{N}^*$ and with the help of (1.4b), we have $\frac{|u^n|^{2\sigma}-|u^{n-1}|^{2\sigma}}{\delta t}=\mathrm{Im}(\Delta(\frac{u^{n+1}+u^n}{2})f(u^{n+1},u^n))$, where f is a function with parameter σ. Therefore, the operator Δ causes a loss of regularity for the time derivative of u, and we have to give up this solution.

3. The two previous demonstrations lead us to control the discrete derivative of u^n. Then, the new idea is to take $v^{n+\frac12}=\frac{u^{n+1}-u^n}{\delta t}$ as a new independent variable. In the continuous framework, this last manipulation amounts to writing

$$\begin{cases}\phi_t=|u|_t^{2\sigma},\\ iu_t+\Delta u=\lambda\phi u,\end{cases} \tag{2.2}$$

and setting $v=u_t$, which is the solution to $iv_t+\Delta v=\lambda(\phi_t u+\phi v)$. Hence, the system

(2.2) reads

$$\text{(2.3)}\qquad \begin{cases} \phi_t = 2\mathrm{Re}(\overline{u}v)(\sigma(|u|^2)^{\sigma-1}) \equiv \Phi_\sigma, \\ iu_t + \Delta u = \lambda\phi u \equiv \lambda U, \\ iv_t + \Delta v = \lambda(2\mathrm{Re}(\overline{u}v)(\sigma(|u|^2)^{\sigma-1})u + \phi v) \equiv \lambda V_\sigma. \end{cases}$$

This system is semilinear and now can be treated easily by a classical fixed point procedure in usual Sobolev spaces. Nevertheless, the discrete case again is slightly more difficult. Indeed, the term $v^{n+\frac{1}{2}}$, which is the discrete equivalent of v, does not propagate through the linear part of the Crank–Nicolson scheme $i\frac{u^{n+1}-u^n}{\delta t} + \Delta(\frac{u^{n+1}+u^n}{2})$ but through $i\frac{v^{n+\frac{3}{2}}-v^{n-\frac{1}{2}}}{2\delta t} + \Delta(\frac{v^{n+\frac{3}{2}}+2v^{n+\frac{1}{2}}+v^{n-\frac{1}{2}}}{4})$, which generates two unitary groups and makes the proof more difficult to handle than in the continuous case.

2.2. Notation. We first denote by $S(t)$ the group associated to the linear Schrödinger equation $iu_t + \Delta u = 0$, $u(0,x) = u_0(x)$. Therefore, we can represent the solution of (1.1) by $u(t,x) = S(t)u_0 - i\int_0^t S(t-\zeta)\lambda|u|^{2\sigma}u(\zeta)d\zeta$.

We introduce also the various operators $A = (1 - i\delta t\frac{\Delta}{2})^{-1}(1 + i\delta t\frac{\Delta}{2})$, $B = (1 - i\delta t\frac{\Delta}{2})^{-1}$, $\Xi^k = (1+A)^{-1}(A^k - (-1)^k)$, and $\mathcal{S}_{\delta t}(t) = \sum_{n=0}^{N-1} A^n(x)\mathbf{I}_{[t_n,t_{n+1}[}$. Then, we have the following lemma.

LEMMA 2.2.

1. *A is a unitary operator on H^s $\forall s$.*
2. *B is a bounded operator on H^s and $\|B\|_s \le 1$.*
3. *$B\Xi^k$ is a bounded operator on H^s and $\|B\Xi^k\|_s \le 1$.*
4. *$\lim_{\delta t\to 0}\mathcal{S}_{\delta t}(t) = \mathcal{S}(t)$ for the strong topology of operators.*
5. *$\lim_{\delta t\to 0} B = 1$.*

Since $\sigma \in \mathbb{N}^*$, $|u^{n+1}|^{2\sigma} - |u^n|^{2\sigma} = (|u^{n+1}|^2 - |u^n|^2)f_\sigma(|u^{n+1}|^2, |u^n|^2)$, where f_σ is a C^1 function. For example, we have $f_1(a,b) \equiv 1$ and $f_2(a,b) \equiv a + b$.

We also set $v^{n+\frac{1}{2}} = \frac{u^{n+1}-u^n}{\delta t}$. Then, we derive from the scheme (1.4) the discrete equivalents of the nonlinearities U, Φ_σ, and V_σ, namely, $U^{n+\frac{1}{2}}$, $\Phi_\sigma^{n+\frac{1}{2}}$, and $V_\sigma^{n+\frac{1}{2}}$:

$$U^{n+\frac{1}{2}} = \phi^{n+\frac{1}{2}}\left(\frac{u^{n+1}+u^n}{2}\right),$$

$$\Phi_\sigma^{n+\frac{1}{2}} = 2\mathrm{Re}\left(v^{n+\frac{1}{2}}\left(\overline{\frac{u^{n+1}+u^n}{2}}\right)\right) f_\sigma(|u^{n+1}|^2, |u^n|^2),$$

$$V_\sigma^{n+\frac{1}{2}} = \left(\frac{\phi^{n+\frac{3}{2}}+\phi^{n-\frac{1}{2}}}{2}\right)\left(\frac{v^{n+\frac{3}{2}}+2v^{n+\frac{1}{2}}+v^{n-\frac{1}{2}}}{4}\right)$$
$$+ 2\mathrm{Re}\left(v^{n+\frac{1}{2}}\left(\overline{\frac{u^{n+1}+u^n}{2}}\right)\right) f_\sigma(|u^{n+1}|^2, |u^n|^2)\left(\frac{u^{n+2}+u^{n+1}+u^n+u^{n-1}}{4}\right).$$

2.3. Proof of Proposition 2.1. The proof is made up of three steps. As stressed above, we need to modify system (1.4) in order to exhibit the equation for $v^{n+\frac{1}{2}}$. The second step consists of solving this new system by a standard fixed point procedure. And, in the last step, we prove that the solution thus obtained is indeed the solution to (1.4).

Step 1: transformation of system (1.4). To begin, let us write (1.4b) at time $t^{n-\frac{3}{2}}$, $n \ge 2$.

We have $i\frac{u^{n-1}-u^{n-2}}{\delta t}+\Delta(\frac{u^{n-1}+u^{n-2}}{2})=\lambda(\frac{u^{n-1}+u^{n-2}}{2})\phi^{n-\frac{3}{2}}$. After subtraction from (1.4b) and division by $2\delta t$, we get

$$
\begin{aligned}
&i\frac{v^{n+\frac{1}{2}}-v^{n-\frac{3}{2}}}{2\delta t}+\Delta\left(\frac{v^{n+\frac{1}{2}}+2v^{n-\frac{1}{2}}+v^{n-\frac{3}{2}}}{4}\right)\\
&\quad=\lambda\left(\frac{\phi^{n+\frac{1}{2}}+\phi^{n-\frac{3}{2}}}{2}\right)\left(\frac{v^{n+\frac{1}{2}}+2v^{n-\frac{1}{2}}+v^{n-\frac{3}{2}}}{4}\right)\\
&\qquad+\lambda\left(\frac{\phi^{n+\frac{1}{2}}-\phi^{n-\frac{3}{2}}}{2\delta t}\right)\left(\frac{u^{n+1}+u^{n}+u^{n-1}+u^{n-2}}{4}\right).
\end{aligned}
$$

Moreover, from (1.4a), the term $\frac{\phi^{n+\frac{1}{2}}-\phi^{n-\frac{3}{2}}}{2\delta t}$ equals $\frac{|u^n|^{2\sigma}-|u^{n-1}|^{2\sigma}}{\delta t}$. This equality together with the definition of f_σ yields

$$
\frac{\phi^{n+\frac{1}{2}}-\phi^{n-\frac{3}{2}}}{2\delta t}=\left(\frac{|u^n|^2-|u^{n-1}|^2}{\delta t}\right)f_\sigma(|u^n|^2,|u^{n-1}|^2).
$$

On the other hand, a simple computation gives $2\mathrm{Re}(v^{n-\frac{1}{2}}(\overline{\frac{u^n+u^{n-1}}{2}}))=\frac{|u^n|^2-|u^{n-1}|^2}{\delta t}$.

Collecting all the previous information, we get the system of equations

$$
\left\{
\begin{aligned}
&\frac{\phi^{n+\frac{3}{2}}-\phi^{n-\frac{1}{2}}}{2\delta t}=\Phi_\sigma^{n+\frac{1}{2}}, && \text{(a)}\\
&i\frac{u^{n+2}-u^{n+1}}{\delta t}+\Delta\left(\frac{u^{n+2}+u^{n+1}}{2}\right)=\lambda U^{n+\frac{3}{2}}, && \text{(b)}\\
&i\frac{v^{n+\frac{3}{2}}-v^{n-\frac{1}{2}}}{2\delta t}+\Delta\left(\frac{v^{n+\frac{3}{2}}+2v^{n+\frac{1}{2}}+v^{n-\frac{1}{2}}}{4}\right)=\lambda V_\sigma^{n+\frac{1}{2}}. && \text{(c)}
\end{aligned}
\right. \tag{2.4}
$$

Step 2: existence and uniqueness of scheme (2.4). This system (2.4) is clearly the discrete equivalent of (2.3). Expressed in this way, the system is semilinear and can be solved by a fixed point procedure in the integral formulation for the vector $(\phi,\, u,\, v)^t$.

Before going any further, we have to define the initial data. We set $u^0(x)=u_0(x)$, $\phi^{-\frac{1}{2}}=\phi^{+\frac{1}{2}}=|u^0|^{2\sigma}$, $v^{-\frac{1}{2}}=\frac{u^0-u^{-1}}{\delta t}$, and $v^{\frac{1}{2}}=\frac{u^1-u^0}{\delta t}$. Also, we define the quantity u^{-1} by

$$
Au^{-1}=(A+1)u^0-u^1.
$$

The choice of u^{-1} is not arbitrary; indeed, it is the only choice that ensures that, for the initial data $v^{\frac{1}{2}}$ and $v^{-\frac{1}{2}}$ defined as above, the solution to $i\frac{v^{n+\frac{3}{2}}-v^{n-\frac{1}{2}}}{2\delta t}+\Delta\frac{v^{n+\frac{3}{2}}+2v^{n+\frac{1}{2}}+v^{n-\frac{1}{2}}}{4}=f^{n+\frac{1}{2}}$ is bounded in $l^\infty(0,N;H^s)$ as soon as $(f^{n+\frac{1}{2}})_n$ is bounded in $l^1(0,N;H^s)$. Obviously, the definitions of $v^{\frac{1}{2}}$ and $v^{-\frac{1}{2}}$ involve the term u^1. However, since $u^1=Au^0-i\delta t\lambda B(\frac{u^1+u^0}{2})\phi^{\frac{1}{2}}$, u^1 exists in $H^s(\mathbb{R}^d)$ (at least for small δt) and is unique.

Now, we transform system (2.4) in order to obtain Duhamel's formula. First, (2.4b) becomes $u^{n+2}=Au^{n+1}-i\lambda\delta tBU^{n+\frac{3}{2}}$, which leads immediately to $u^{n+2}=A^{n+2}u^0-i\lambda\sum_{k=0}^{n+1}\delta tBA^{n+1-k}U^{k+\frac{1}{2}}$. We apply the same transformation to (2.4a). However, for $n=0$, it leads to $\phi^{\frac{3}{2}}=|u_0|^{2\sigma}+2\delta t\Phi^{\frac{1}{2}}$, and for $n=1$, we have

$\phi^{\frac{5}{2}} = |u_0|^{2\sigma} + 2\delta t \Phi^{\frac{3}{2}}$. Therefore, the equation for ϕ will be different according to the evenness of n. Indeed, for $n \geq 0$, we find

$$\begin{aligned}\phi^{n+\frac{3}{2}} &= |u_0|^{2\sigma} + \sum_{k=1}^{[n/2]} 2\delta t \Phi_\sigma^{2k-\frac{1}{2}} \text{ if } n \text{ is odd},\\ \phi^{n+\frac{3}{2}} &= |u_0|^{2\sigma} + \sum_{k=0}^{n/2} 2\delta t \Phi_\sigma^{2k+\frac{1}{2}} \text{ if } n \text{ is even}.\end{aligned}$$

Then, the last component, $v^{n+\frac{3}{2}}$, is governed by

$$v^{n+\frac{3}{2}} - (A-1)v^{n+\frac{1}{2}} - Av^{n-\frac{1}{2}} = -2i\lambda\delta t B V^{n+\frac{1}{2}}.$$

Using the roots of the characteristic equation $r^2 - (A-1)r - A = 0$, the component $v^{n+\frac{3}{2}}$ is

$$\begin{aligned}v^{n+\frac{3}{2}} &= (A+1)^{-1}\Big\{(A^{n+2} - (-1)^{n+2})v^{\frac{1}{2}} + A(A^{n+1} - (-1)^{n+1})v^{-\frac{1}{2}}\\ &\qquad - i\lambda \sum_{k=1}^{n+1} 2\delta t B (A^{n+2-k} - (-1)^{n+2-k}) V_\sigma^{k-\frac{1}{2}}\Big\}\\ &= \Xi^{n+1}(\Xi v^{\frac{1}{2}} + Av^{-\frac{1}{2}}) - i\lambda \sum_{k=0}^{n} 2\delta t\, B\, \Xi^{n+1-k}\, V_\sigma^{k+\frac{1}{2}}.\end{aligned}$$

Collecting all of the above calculations, the system (2.4) reads for $n \geq 0$ as follows:

$$\text{(2.5)}\qquad \begin{cases} \phi^{n+\frac{3}{2}} = \begin{cases} |u_0|^{2\sigma} + \displaystyle\sum_{k=1}^{[n/2]} 2\delta t \Phi_\sigma^{2k-\frac{1}{2}} \text{ if } n \text{ is odd},\\ |u_0|^{2\sigma} + \displaystyle\sum_{k=0}^{n/2} 2\delta t \Phi_\sigma^{2k+\frac{1}{2}} \text{ if } n \text{ is even}, \end{cases}\\ u^{n+2} = A^{n+2}u_0 - i\lambda \displaystyle\sum_{k=0}^{n+1} \delta t\, B\, A^{n+1-k}\, U^{k+\frac{1}{2}},\\ v^{n+\frac{3}{2}} = \Xi^{n+1}(\Xi v^{\frac{1}{2}} + Av^{-\frac{1}{2}}) - i\lambda \displaystyle\sum_{k=0}^{n} 2\delta t\, B\, \Xi^{n+1-k}\, V_\sigma^{k+\frac{1}{2}}. \end{cases}$$

Let us define the sequences

$$\begin{aligned}\mathbf{U}^N &= (u^2, \ldots, u^{N+1}),\\ \Phi^N &= (\phi^{\frac{3}{2}}, \ldots, \phi^{N+\frac{1}{2}}),\\ \mathbf{V}^N &= (v^{\frac{3}{2}}, \ldots, v^{N+\frac{1}{2}}),\\ \widetilde{\mathbf{U}}^N &= (\tilde{u}^2, \ldots, \tilde{u}^{N+1}),\\ \widetilde{\Phi}^N &= (\tilde{\phi}^{\frac{3}{2}}, \ldots, \tilde{\phi}^{N+\frac{1}{2}}),\\ \widetilde{\mathbf{V}}^N &= (\tilde{v}^{\frac{3}{2}}, \ldots, \tilde{v}^{N+\frac{1}{2}}),\end{aligned}$$

and the space $X_N = l^\infty(0, N; H^s(\mathbb{R}^d))$ endowed with the usual norm $\|.\|_{X_N} = \sup_{n\in[0,N]} \|(.)_n\|_s$, where $\|\cdot\|_s$ indicates the norm of $H^s(\mathbb{R}^d)$.

The proof of existence and uniqueness for the previous system consists of showing the existence of a unique fixed point for the map $\mathcal{T}$ defined by

$$\begin{array}{ccc} X_N^3 & \stackrel{\mathcal{T}}{\longrightarrow} & X_N^3, \\ (\widetilde{\mathbf{U}}^N, \widetilde{\Phi}^N, \widetilde{\mathbf{V}}^N) & \longmapsto & (\mathbf{U}^N, \Phi^N, \mathbf{V}^N), \end{array}$$

where, for $n \in [0, N-1]$,

$$\phi^{n+\frac{3}{2}} = \left| \begin{array}{l} |u_0|^{2\sigma} + \displaystyle\sum_{k=1}^{[n/2]} 2\delta t 2\mathrm{Re}\left(\tilde{v}^{2k-\frac{1}{2}} \left(\overline{\frac{\tilde{u}^{2k} + \tilde{u}^{2k-1}}{2}} \right) \right) f_\sigma(|\tilde{u}^{2k}|^2, |\tilde{u}^{2k-1}|^2) \text{ if } n \text{ is odd}, \\ |u_0|^\sigma 2 + \displaystyle\sum_{k=0}^{n/2} 2\delta t 2\mathrm{Re}\left(\tilde{v}^{2k+\frac{1}{2}} \left(\overline{\frac{\tilde{u}^{2k+1} + \tilde{u}^{2k}}{2}} \right) \right) f_\sigma(|\tilde{u}^{2k+1}|^2, |\tilde{u}^{2k}|^2) \text{ if } n \text{ is even}, \end{array} \right.$$

$$u^{n+2} = A^{n+2} u_0 - i\lambda \sum_{k=0}^{n+1} \delta t\, B\, A^{n+1-k} \left(\tilde{\phi}^{k+\frac{1}{2}} \left(\overline{\frac{\tilde{u}^{k+1} + \tilde{u}^k}{2}} \right) \right),$$

$$v^{n+\frac{3}{2}} = \Xi^{n+1} \left(\Xi v^{\frac{1}{2}} + A v^{-\frac{1}{2}} \right)$$

$$-i\lambda \sum_{k=0}^{n} 2\delta t B\, \Xi^{n+1-k} \left(\begin{array}{l} \left(\frac{\tilde{\phi}^{k+\frac{3}{2}} + \tilde{\phi}^{k-\frac{1}{2}}}{2} \right) \left(\frac{\tilde{v}^{k+\frac{3}{2}} + 2\tilde{v}^{k+\frac{1}{2}} + \tilde{v}^{k-\frac{1}{2}}}{4} \right) \\ + 2\mathrm{Re}\left[\tilde{v}^{k+\frac{1}{2}} \left(\frac{\tilde{u}^{k+1} + \tilde{u}^k}{2} \right) \right] f_\sigma(|\tilde{u}^{2k+1}|^2, |\tilde{u}^{2k}|^2) \left(\frac{\tilde{u}^{k+2} + \tilde{u}^{k+1} + \tilde{u}^k + \tilde{u}^{k-1}}{4} \right) \end{array} \right).$$

So that all terms are well defined, we set $\tilde{u}^{-1} = u^{-1}$, $\tilde{u}^0 = u^0$, $\tilde{u}^1 = u^1$, $\tilde{\phi}^{-\frac{1}{2}} = \tilde{\phi}^{\frac{1}{2}} = |u^0|^{2\sigma}$, $\tilde{v}^{-\frac{1}{2}} = v^{-\frac{1}{2}}$, and $\tilde{v}^{\frac{1}{2}} = v^{\frac{1}{2}}$.

Let us define B_R, the ball of X_N^3 centered on 0 with radius R endowed with the usual norm on X_N^3, by

$$B_R = \{(\mathbf{U}^N, \Phi^N, \mathbf{V}^N) \in X_N^3 / \|\mathbf{U}^N\|_{X_N} + \|\Phi^N\|_{X_N} + \|\mathbf{V}^N\|_{X_N} \leq R\}.$$

Then, we show that the application $\mathcal{T}$ sends B_R into B_R with some a priori estimates. Let us make the assumption that $\widetilde{\Phi}^N$, $\widetilde{\mathbf{U}}^N$, and $\widetilde{\mathbf{V}}^N$ belong to X_N.

In the continuous case of nonlinear Schrödinger equation (1.1), $f_\sigma = \sigma(|u|^2)^{\sigma-1}$. In our discrete case, f_σ does not involve the more powerful power of $|u|^2$. As we choose the Sobolev space $H^s(\mathbb{R}^d)$ for $s > d/2 + 2$ and $\sigma \in \mathbb{N}^*$, f_σ belongs to H^s if $u \in H^s$.

Estimate on Φ^N: $\forall n$, $\|\phi^{n+3/2}\|_s \leq \|u^0\|_s^{2\sigma} + c_1 T_{\delta t} \|\widetilde{\mathbf{U}}^N\|_{X_N}^{2\sigma-1} \|\widetilde{\mathbf{V}}^N\|_{X_N}$, with c_1 independent of n and δt.

Estimate on $\mathbf{U}^N$: Lemma 2.2 leads to $\|u^n\|_s \leq \|u^0\|_s + c_2 T_{\delta t} \|\widetilde{\Phi}^N\|_{X_N} \|\widetilde{\mathbf{U}}^N\|_{X_N}$, with c_2 independent of n and δt.

Estimate on $\mathbf{V}^N$: the term $\Xi^{n+1}(\Xi v^{\frac{1}{2}} + A v^{-\frac{1}{2}})$ is not a priori bounded. However, thanks to the definition of u^{-1}, $\Xi^{n+1}(\Xi v^{\frac{1}{2}} + A v^{-\frac{1}{2}}) = A^{n+2} v^{-\frac{1}{2}}$. This term is therefore bounded. Lemma 2.2 yields

$$\|v^{n+\frac{1}{2}}\|_s \leq \|v^{-\frac{1}{2}}\|_s + c_3 T_{\delta t} \left(\|\widetilde{\Phi}^N\|_{X_N} \|\widetilde{\Phi}^N\|_{X_N} + \|\widetilde{\mathbf{V}}^N\|_{X_N} \|\widetilde{\mathbf{U}}^N\|_{X_N}^{2\sigma-1} \right),$$

with c_3 independent of n and δt.

Collecting these three estimates, we finally have

$$\|\mathbf{U}^N\|_{X_N} + \|\Phi^N\|_{X_N} + \|\mathbf{V}^N\|_{X_N} \leq (\|u^0\|_s + \|u^0\|_s^{2\sigma} + \|v^{-\frac{1}{2}}\|_s) + c_4(R)T_{\delta t}.$$

Now, we define R by $\|u^0\|_s + \|u^0\|_s^{2\sigma} + \|v^{-\frac{1}{2}}\|_s = \frac{R}{2}$ and define $T_{\delta t}$ such that $c_6(R)T_{\delta t} \leq \frac{R}{2}$ to ensure that $\mathcal{T}$ maps B_R into itself.

By similar arguments, it is easy to show that $\mathcal{T}$ is a contraction in B_R if $T_{\delta t}$ is small enough. Therefore, $\mathcal{T}$ has a unique fixed point. This last argument shows the existence and uniqueness of the solution to (2.4).

Step 3: equivalence between systems (2.4) and (1.4). Let $(u^{n+1}, v^{n+\frac{1}{2}}, \phi^{n+\frac{1}{2}})$ be the solution of system (2.4). We already know that the couple $(u^{n+1}, \phi^{n+\frac{1}{2}})$ solves (2.4b) at time $t^{n+\frac{1}{2}}$. We have to show that $v^{n+\frac{1}{2}} = \frac{u^{n+1}-u^n}{\delta t}$ and $\frac{\phi^{n+\frac{1}{2}}+\phi^{n-\frac{1}{2}}}{2} = |u^n|^{2\sigma}$. If we rewrite (2.4b) at time $t^{n-\frac{3}{2}}$, we have $i\frac{u^{n-1}-u^{n-2}}{\delta t} + \Delta(\frac{u^{n-1}+u^{n-2}}{2}) = \lambda(\frac{u^{n-1}+u^{n-2}}{2})\phi^{n-\frac{3}{2}}$. We subtract this last equality from (2.4b) at time $t^{n+\frac{1}{2}}$ and divide the result by $2\delta t$. Then, we have

$$\begin{aligned} i\frac{\frac{u^{n+1}-u^n}{\delta t} - \frac{u^{n-1}-u^{n-2}}{\delta t}}{2\delta t} + \Delta\left(\frac{\frac{u^{n+1}-u^n}{\delta t} + 2\frac{u^n-u^{n-1}}{\delta t} + \frac{u^{n-1}-u^{n-2}}{\delta t}}{4}\right) \\ = \lambda\left(\frac{\phi^{n+\frac{1}{2}} - \phi^{n-\frac{3}{2}}}{2\delta t}\right)\left(\frac{u^{n+1}+u^n+u^{n-1}+u^{n-2}}{4}\right) \\ + \lambda\left(\frac{\phi^{n+\frac{1}{2}} + \phi^{n-\frac{3}{2}}}{2}\right)\left(\frac{\frac{u^{n+1}-u^n}{\delta t} + 2\frac{u^n-u^{n-1}}{\delta t} + \frac{u^{n-1}-u^{n-2}}{\delta t}}{4}\right). \end{aligned}$$

This identity is a discrete equivalent of the continuous equation $iu_{tt} + \Delta u_t = \lambda\phi_t u + \lambda\phi u_t$. We know that $\frac{\phi^{n+\frac{1}{2}}-\phi^{n-\frac{3}{2}}}{2\delta t} = 2\mathrm{Re}(v^{n-\frac{1}{2}}(\overline{\frac{u^n+u^{n-1}}{2}}))f_\sigma(|u^n|^2, |u^{n-1}|^2)$. Finally, we get

$$\begin{aligned} i\frac{\frac{u^{n+1}-u^n}{\delta t} - \frac{u^{n-1}-u^{n-2}}{\delta t}}{2\delta t} + \Delta\left(\frac{\frac{u^{n+1}-u^n}{\delta t} + 2\frac{u^n-u^{n-1}}{\delta t} + \frac{u^{n-1}-u^{n-2}}{\delta t}}{4}\right) \\ = \lambda 2\mathrm{Re}\left(v^{n-\frac{1}{2}}\left(\overline{\frac{u^n+u^{n-1}}{2}}\right)\right) f_\sigma(|u^n|^2, |u^{n-1}|^2)\left(\frac{u^{n+1}+u^n+u^{n-1}+u^{n-2}}{4}\right) \\ + \lambda\left(\frac{\phi^{n+\frac{1}{2}} + \phi^{n-\frac{3}{2}}}{2}\right)\left(\frac{\frac{u^{n+1}-u^n}{\delta t} + 2\frac{u^n-u^{n-1}}{\delta t} + \frac{u^{n-1}-u^{n-2}}{\delta t}}{4}\right). \end{aligned}$$

As before, the continuous case reads $iu_{tt} + \Delta u_t = \lambda 2\mathrm{Re}(v\overline{u})u + \lambda\phi u_t$. We set $w^{n+\frac{1}{2}} = v^{n+\frac{1}{2}} - (\frac{u^{n+1}-u^n}{\delta t})$. Then, the subtraction of this last equation from (2.4c) taken at time $t^{n-\frac{1}{2}}$ leads to

$$\begin{aligned} i\frac{w^{n+\frac{1}{2}} - w^{n-\frac{3}{2}}}{2\delta t} + \Delta\left(\frac{w^{n+\frac{1}{2}} + 2w^{n-\frac{1}{2}} + w^{n-\frac{3}{2}}}{4}\right) \\ = \lambda\left(\frac{\phi^{n+\frac{1}{2}} + \phi^{n-\frac{1}{2}}}{2}\right)\left(\frac{w^{n+\frac{1}{2}} + 2w^{n-\frac{1}{2}} + w^{n-\frac{3}{2}}}{4}\right), \end{aligned}$$

with initial data $w^{\frac{1}{2}} = w^{-\frac{1}{2}} = 0$. This is a linear Schrödinger equation with regular potential. So, the only solution is zero. Thus, we have $v^{n+\frac{1}{2}} = \frac{u^{n+1}-u^n}{\delta t}$. Thanks to this equality, we get $\frac{\phi^{k+\frac{1}{2}}-\phi^{k-\frac{3}{2}}}{2\delta t} = \frac{|u^n|^2-|u^{n-1}|^2}{\delta t}$. Summing this last relation for $k \in [0,n]$ yields $\frac{\phi^{n+\frac{1}{2}}+\phi^{n-\frac{1}{2}}}{2} = |u^n|^2$.

The systems (1.4) and (2.4) are equivalent, thus proving Proposition 2.1.

3. Convergence of the relaxation scheme for strong solutions. The result of this section reads as follows.

PROPOSITION 3.1. *Let u be the maximal solution to equation* (1.1) *defined in* $C^2([0,T^*[;H^s(\mathbb{R}^d))$, $s > d/2+4$. *Then, if* $T_{\delta t} = N\delta t$, $\liminf_{\delta t\to 0} T_{\delta t} \geq T^*$ *and* $\forall T < T^*$ *the solution* $(u_{\delta t}, \phi_{\delta t})$ *to* (1.4) *given by Proposition* (2.1) *converges to* $(u, |u|^{2\sigma})$ *in* $L^\infty([0,T];(H^s(\mathbb{R}^d))^2)$ *as* $\delta t \to 0$.

3.1. Ideas and difficulties of the proof. The proof of convergence is done comparing the integral formulation of solutions $(u_{\delta t}, \phi_{\delta t})$ and $(u, |u|^{2\sigma})$. However, for the same reasons as those developed for the proof of Proposition (2.1), we cannot directly show the result of convergence on system (1.4). A priori, we have to work on system (2.4). Since we wish to compare integral formulations of the solutions, we particularly focus on system (2.5).

Duhamel's formula of system (2.3) is given for $x \in \mathbb{R}^d$ and $t > 0$ by

$$\left\{\begin{array}{ll} \phi(x,t) = |u_0(x)|^{2\sigma} + \displaystyle\int_0^t \Phi_\sigma(x,\zeta)d\zeta, & \text{(a)} \\ u(x,t) = S(t)u_0(x) - i\lambda \displaystyle\int_0^t S(t-\zeta)U(x,\zeta)d\zeta, & \text{(b)} \\ v(x,t) = S(t)v(x,t=0) - i\lambda \displaystyle\int_0^t S(t-\zeta)V_\sigma(x,\zeta)d\zeta. & \text{(c)} \end{array}\right. \tag{3.1}$$

Thanks to the consistency of $BA^{n+1-k}U^{k+\frac{1}{2}}$ and $S(t-\zeta)U(x,\zeta)$ (respectively, $\Phi_\sigma^{2k+\frac{1}{2}}$ and Φ_σ), it is quite easy to show convergence of u^{n+2} to $u(x,t)$ (respectively, $\phi^{n+\frac{3}{2}}$ and $\phi(x,t)$).

Therefore, the only difficulty arises from the comparison between $v^{n+\frac{3}{2}}$ and $v(x,t)$. Thanks to the definition of u^{-1}, we have already seen that $\Xi^{n+1}(\Xi v^{\frac{1}{2}} + Av^{-\frac{1}{2}}) = A^{n+2}v^{-\frac{1}{2}}$. Thus, the contribution of the initial datum is easily handled. However, the second part of Duhamel's formula involves the bounded operator $B\Xi^{n+1-k}$ which generates two semigroups according to the evenness of $n+1-k$. Actually,

$$\lim_{\delta t\to 0} \left(B\Xi^{n+1-k}\right)_{n+1-k=2p} = \frac{\exp\left(i\Delta t_{n+1-k}/2\right)-1}{2},$$
$$\lim_{\delta t\to 0} \left(B\Xi^{n+1-k}\right)_{n+1-k=2p+1} = \frac{\exp\left(i\Delta t_{n+1-k}/2\right)+1}{2}.$$

Separately, these two semigroups are different from S. As a consequence, with the current forms of $v^{n+\frac{3}{2}}$ and $v(x,t)$, we cannot prove the convergence.

3.2. Proof of Proposition 3.1. Actually, we have to find another form for the term $\sum_{k=0}^n B\Xi^{n+1-k}V^{k+\frac{1}{2}}$. First, we remark that $\Xi^k = \sum_{l=1}^k (-1)^{l-1}A^{k-l}$. This leads us to write

$$\sum_{k=0}^n 2\delta t B\Xi^{n+1-k}V^{k+\frac{1}{2}} = \sum_{k=0}^n 2\delta t B \sum_{l=1}^{n+1-k} (-1)^{l-1}A^{n+1-k-l}V^{k+\frac{1}{2}}.$$

Finally, by rearranging the previous sums, we get

$$\sum_{k=0}^{n} B\Xi^{n+1-k}V^{k+\frac{1}{2}} = \begin{cases} \displaystyle\sum_{l=0}^{p-1} 2\delta t BA^{2l+1}\left(\sum_{j=0}^{p-l-1}\delta t\frac{V^{2j+\frac{3}{2}}-V^{2j+\frac{1}{2}}}{\delta t}\right) \\ \quad \displaystyle+\sum_{l=0}^{p} 2\delta t BA^{2l}\left(V^{\frac{1}{2}}+\sum_{j=0}^{p-l-1}\delta t\frac{V^{2j+\frac{5}{2}}-V^{2j+\frac{3}{2}}}{\delta t}\right) & \text{if } n=2p, \\ \displaystyle\sum_{l=0}^{p} 2\delta t BA^{2l+1}\left(V^{\frac{1}{2}}+\sum_{j=0}^{p-l-1}\delta t\frac{V^{2j+\frac{5}{2}}-V^{2l+\frac{3}{2}}}{\delta t}\right) \\ \quad \displaystyle+\sum_{l=0}^{p} 2\delta t BA^{2l}\left(\sum_{j=0}^{p-l}\delta t\frac{V^{2j+\frac{3}{2}}-V^{2j+\frac{1}{2}}}{\delta t}\right) & \text{if } n=2p+1. \end{cases}$$

Once more, we observe discrete forms of the time derivative. The continuous equation for $v(x,t)$ may therefore be interpreted as

$$v(x,t) = S(t)v(x,t=0) - i\lambda\int_0^t S(t-s)\int_0^s \partial_\tau V(x,\tau)d\tau ds. \tag{3.2}$$

Unfortunately, $\partial_\tau V$ involves a time derivative of v (denoted by w) and ϕ (denoted by ψ) simultaneously. Therefore, we have to find a discrete system which is equivalent to (1.4) and allows us to control the time derivatives of $v^{n+\frac{3}{2}}$ (defined by $w^{n+1} = \frac{v^{n+\frac{3}{2}}-v^{n+\frac{1}{2}}}{\delta t}$) and $\phi^{n+\frac{3}{2}}$ (defined by $\psi^{n+1} = \frac{\phi^{n+\frac{3}{2}}-\phi^{n+\frac{1}{2}}}{\delta t}$) in order to compare the new formulations of $v^{n+\frac{3}{2}}$ and $v(x,t)$. This new system is proved to be well posed, and Duhamel's formulations for ψ^{n+1} and w^{n+1} read, respectively,

$$\begin{aligned} \psi^n &= \psi^{-1} + \sum_{k=0}^{p} 2\delta t\Psi^{2k}, \quad n=2p+1, \\ \psi^n &= \psi^0 + \sum_{k=1}^{p} 2\delta t\Psi^{2k-1}, \quad n=2p, \end{aligned}$$

and with the convention $D=(1-i\delta t\Delta)^{-1}(1+i\delta t\Delta)$ and $E=(1-i\delta t\Delta)^{-1}$,

$$\begin{aligned} w^n &= D^p w^1 - i\lambda\sum_{k=0}^{p-1} 2\delta t E D^{p-k-1}W^{2(k+1)}, \quad n=2p+1, \\ w^n &= D^p w^0 - i\lambda\sum_{k=0}^{p-1} 2\delta t E D^{p-k-1}W^{2k+1}, \quad n=2p. \end{aligned}$$

The terms Ψ and W denote the nonlinearities involved in Duhamel's formula (see the appendix for the proof and a complete set of notation).

As before, the formulations are different according to the parity of n. However, the proof of convergence is not really influenced by the choice of n. So from now on,

we choose n to be odd. We define for $t \in [t_n, t_{n+1}[$ and $x \in \mathbb{R}^d$

$$\begin{array}{ll} u_{\delta t}(x,t) = u^n(x), & U_{\delta t}(x,t) = U^{n+\frac{1}{2}}(x), \\ \phi_{\delta t}(x,t) = \phi^{n+\frac{1}{2}}(x), & \Phi_{\delta t}(x,t) = \Phi^{n+\frac{1}{2}}(x), \\ v_{\delta t}(x,t) = v^{n+\frac{1}{2}}(x), & V_{\delta t}(x,t) = V^{n+\frac{1}{2}}(x), \\ \psi_{\delta t}(x,t) = \psi^n(x), & \Psi_{\delta t}(x,t) = \Psi^n(x), \\ w_{\delta t}(x,t) = w^n(x), & W_{\delta t}(x,t) = W^n(x). \end{array}$$

Taking $u \in C^2(0,T;H^s(\mathbb{R}^d))$ with $s > d/2+4$, ϕ, $v = u_t$, $\psi = \phi_t$, and $w = u_{tt}$ always belong to $H^s(\mathbb{R}^d)$, and we define the applications $\mathcal{U}_{\delta t}$ and $\mathcal{U}$ as follows:

$$\begin{array}{ccc} (H^s(\mathbb{R}^d))^5 & \xrightarrow{\mathcal{U}_{\delta t}} & (L^\infty(0,T;H^s(\mathbb{R}^d))^5, \\ (u_{\delta t}, \phi_{\delta t}, v_{\delta t}, \psi_{\delta t}, w_{\delta t})(0) & \longmapsto & (u_{\delta t}, \phi_{\delta t}, v_{\delta t}, \psi_{\delta t}, w_{\delta t})(t), \end{array}$$

$$\begin{array}{ccc} (H^s(\mathbb{R}^d))^5 & \xrightarrow{\mathcal{U}} & (L^\infty(0,T;H^s(\mathbb{R}^d))^5, \\ (u, \phi, u_t, \phi_t, u_{tt})(0) & \longmapsto & (u, \phi, u_t, \phi_t, u_{tt})(t). \end{array}$$

In the same way, let us define $\mathcal{F}_{\delta t}$ and $\mathcal{F}$ as the two applications

$$\begin{array}{ccc} (L^\infty(0,T;H^s(\mathbb{R}^d))^5 & \xrightarrow{\mathcal{F}_{\delta t}} & (L^\infty(0,T;H^s(\mathbb{R}^d))^5, \\ (u_{\delta t}, \phi_{\delta t}, v_{\delta t}, \psi_{\delta t}, w_{\delta t}) & \longmapsto & (U_{\delta t}, -i\Phi_{\delta t}, V_{\delta t}, -i\Psi_{\delta t}, W_{\delta t}), \end{array}$$

and

$$\begin{array}{ccc} (L^\infty(0,T;H^s(\mathbb{R}^d))^5, & \xrightarrow{\mathcal{F}} & (L^\infty(0,T;H^s(\mathbb{R}^d))^5, \\ (u, \phi, v, \psi, w) & \longmapsto & (U, -i\Phi, V, -i\Psi, W). \end{array}$$

Finally, we set the operators $\mathcal{B} = (B, I, B, I, E)$, $\mathcal{S}(t) = (S(t), 1, S(t), 1, S(t))$, and for $t \in [t_n, t_{n+1}[$, $\mathcal{S}_{\delta t}(t) = (A^n, 1, A^n, 1, D^{\lfloor n/2 \rfloor})$.

We proceed as in [8] for the proof of convergence. We have to evaluate $\|\mathcal{U}(t) - \mathcal{U}_{\delta t}(t)\|_{(H^s)^5}$ for $t = t_{n+1}$ (we do not replace t by t_{n+1}, in order to simplify the notation). This leads to

(3.3)

$$\begin{aligned} &\left\| \mathcal{S}(t)\mathcal{U}(0) - \mathcal{S}_{\delta t}(t)\mathcal{U}_{\delta t}(0) - i\int_0^t (\mathcal{S}(t-\tau)\mathcal{F}(\mathcal{U})(\tau) - \mathcal{B}\mathcal{S}_{\delta t}(t-\tau)\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\tau))d\tau \right\|_{(H^s)^5} \\ &\quad \le \|(\mathcal{S}(t) - \mathcal{S}_{\delta t}(t))\mathcal{U}(0)\|_{(H^s)^5} + \|\mathcal{S}_{\delta t}(t)(\mathcal{U}(0) - \mathcal{U}_{\delta t}(0))\|_{(H^s)^5} \\ &\qquad + \sum_{k=0}^n \left\| \int_{t_k}^{t_{k+1}} \mathcal{B}\mathcal{S}_{\delta t}(t-t_k)\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\kappa\delta t) - \mathcal{S}(t-\tau)\mathcal{F}(\mathcal{U})(\tau)d\tau \right\|_{(H^s)^5}. \end{aligned}$$

We introduce the variable $\kappa\delta t$ in this inequality in order to insist on the fact that the term $\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\kappa\delta t)$ does not depend on the variable τ. For example, the first component of $\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\kappa\delta t)$ is

$$\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\kappa\delta t) = \phi_{\delta t}(t_{k+\frac{1}{2}}) \left(\frac{u_{\delta t}(t_{k+1}) + u_{\delta t}(t_k)}{2} \right).$$

In order to control the third term in the right-hand side of (3.3), we write each integral as

$$\begin{aligned}\left\|\int_{t_k}^{t_{k+1}} (\mathcal{B}\mathcal{S}_{\delta t}(t-t_k)\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\kappa\delta t) - \mathcal{S}(t-\tau)\mathcal{F}(\mathcal{U})(\tau))d\tau\right\|_{(H^s)^5} \\ \leq \left\|\int_{t_k}^{t_{k+1}} [\mathcal{B}\mathcal{S}_{\delta t}(t-t_k) - \mathcal{S}(t-t_k)]\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\kappa\delta t)d\tau\right\|_{(H^s)^5} \\ + \left\|\int_{t_k}^{t_{k+1}} [\mathcal{S}(t-\tau) - \mathcal{S}(t-t_k)]\mathcal{F}(\mathcal{U})(\tau)d\tau\right\|_{(H^s)^5} \\ + \left\|\int_{t_k}^{t_{k+1}} \mathcal{S}(t-t_k)[\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\kappa\delta t) - \mathcal{F}(\mathcal{U})(\tau)]d\tau\right\|_{(H^s)^5} \\ \equiv I_1^{k,n} + I_2^{k,n} + I_3^{k,n}.\end{aligned}$$

Let $T_1 = N\delta t$. Then, the inequality (3.3) means

$$\begin{aligned}\|\mathcal{U} - \mathcal{U}_{\delta t}\|_{L^\infty([0,T_1],H^s)^5} \leq \|(\mathcal{S} - \mathcal{S}_{\delta t})\mathcal{U}(0)\|_{L^\infty([0,T_1],H^s)^5} \\ + \|\mathcal{S}_{\delta t}(\mathcal{U}(0) - \mathcal{U}_{\delta t}(0))\|_{L^\infty([0,T_1],H^s)^5} \\ + \sup_{n\in[0,N]} \sum_{k=0}^{n} (I_1^{k,n} + I_2^{k,n} + I_3^{k,n}).\end{aligned} \tag{3.4}$$

Thanks to Lemma 2.2, $\mathcal{S}_{\delta t} \longrightarrow \mathcal{S}$ for the strong topology of operators in $(H^s)^5$, and $\mathcal{B} \longrightarrow 1$ in $(H^s)^5$, we have

$$\begin{aligned}\|(\mathcal{S} - \mathcal{S}_{\delta t})\mathcal{U}(0)\|_{L^\infty([0,T_1],H^s)^5} \to 0 \quad \text{as } \delta t \to 0, \\ \sup_{n\in[0,N]} \sum_{k=0}^{n} I_1^{k,n} \to 0 \quad \text{as } \delta t \to 0, \\ \sup_{n\in[0,N]} \sum_{k=0}^{n} I_2^{k,n} \to 0 \quad \text{as } \delta t \to 0.\end{aligned}$$

Since $u \in C^2(0,T;H^s)$, we also have

$$\|\mathcal{S}_{\delta t}(\mathcal{U}(0) - \mathcal{U}_{\delta t}(0))\|_{L^\infty([0,T_1],H^s)^5} \to 0 \quad \text{as } \delta t \to 0.$$

It remains to deal with the term $I_3^{k,n}$. For this, we write, as in [8], $\mathcal{F}(\mathcal{U})(\tau) = \mathcal{F}_{\delta t}(\mathcal{U})(\kappa\delta t) + R_\kappa(\tau)$. Thus,

$$\begin{aligned}I_3^{k,n} \leq \int_{t_k}^{t_{k+1}} \|\mathcal{F}_{\delta t}(\mathcal{U}_{\delta t})(\kappa\delta t) - \mathcal{F}_{\delta t}(\mathcal{U})(\kappa\delta t)\|_{(H^s)^5} + \delta t \sup_{\tau\in[t_k,t_{k+1}]} \|R_\kappa(\tau)\|_{(H^s)^5} \\ \equiv J_k + \delta t G_k(\delta t).\end{aligned}$$

Since $u \in C^2(0,T;H^s)$, we have $\sum_{k=0}^{N} \delta t G_k(\delta t) \to 0$ as $\delta t \to 0$. Thanks to the form of $\mathcal{F}_{\delta t}$, we get

$$J_k \leq \int_{t_k}^{t_{k+1}} C_1 f(\|\mathcal{U}(\kappa\delta t)\|_{(H^s)^5}, \|\mathcal{U}_{\delta t}(\kappa\delta t)\|_{(H^s)^5}) \|\mathcal{U}(\kappa\delta t) - \mathcal{U}_{\delta t}(\kappa\delta t)\|_{(H^s)^5} d\tau,$$

where f is a C^∞ function and C_1 is a constant which is independent of n and δt. Now, referring to uniform estimates of $\mathcal{U}$ and $\mathcal{U}_{\delta t}$ in $[0,T_1]$, we have $J_k \leq$

$C_2\delta t\,\|\mathcal{U}-\mathcal{U}_{\delta t}\|_{L^\infty([0,T_1],H^s)^5}$, with C_2 a constant independent of n and δt. Finally, the inequality (3.4) leads to

$$\|\mathcal{U}-\mathcal{U}_{\delta t}\|_{L^\infty([0,T_1],H^s)^5} \le o(1) + CT_1\,\|\mathcal{U}-\mathcal{U}_{\delta t}\|_{L^\infty([0,T_1],H^s)^5}.$$

Taking $CT_1 < 1$, we get

$$\|\mathcal{U}-\mathcal{U}_{\delta t}\|_{L^\infty([0,T_1],H^s)^5} \le o(1).$$

Applying once more the proof with the initial datum $\mathcal{U}(T_1)$, we get the lower semi-continuity of the existence time $T_{\delta t}$ as $\delta t \to 0$ and the convergence on $[0,T]$ $\forall T < T^*$.

Hence, we have proved Theorem 1.1.

4. Conservation laws. As we recalled in the introduction, (1.1) has two conserved quantities which are the density defined by $\int_{\mathbb{R}^d}|u|^2dx$ and the energy $\int_{\mathbb{R}^d}\frac{1}{2}|\nabla u|^2+\frac{\lambda}{2\sigma+2}|u|^{2\sigma+2}dx$. The minimum requirement for the numerical scheme is to conserve these two quantities. For example, for $\sigma=1$, the Crank–Nicolson scheme (see [9]) has this property, contrary to the split-step scheme (see [15]) which conserves only the density. We show in this section that the relaxation scheme conserves the two quantities, but only for $\sigma=1$.

PROPOSITION 4.1. *Let $(u^n)_n$ and $(\phi^{n-\frac{1}{2}})_n$ be solutions of (1.4) with $\sigma=1$ belonging to $l^\infty([0,N];H^s(\mathbb{R}^d)^2)$, $s>d/2+2$. Then, $\forall n\le N$, we have*

$$\int_{\mathbb{R}^d}|u^n|^2dx = \int_{\mathbb{R}^d}|u^0|^2dx, \tag{4.1}$$

$$\int_{\mathbb{R}^d}\left(|\nabla u^n|^2+\lambda|u^n|^2\phi^{n-\frac{1}{2}}-\lambda\frac{(\phi^{n-\frac{1}{2}})^2}{2}\right)dx = \int_{\mathbb{R}^d}\left(|\nabla u^0|^2+\lambda|u^0|^2\phi^{-\frac{1}{2}}-\lambda\frac{(\phi^{-\frac{1}{2}})^2}{2}\right)dx. \tag{4.2}$$

Moreover, if $(\phi^{n-\frac{1}{2}})_n \in l^\infty([0,N];L^\infty(\mathbb{R}^d))$, then

$$\int_{\mathbb{R}^d}|u^n|^2dx = \int_{\mathbb{R}^d}\phi^{n+\frac{1}{2}}dx. \tag{4.3}$$

REMARK 2.

1. *We remark that $\int_{\mathbb{R}^d}|u^n|^2\phi^{n-\frac{1}{2}}dx = \int_{\mathbb{R}^d}\frac{\phi^{n+\frac{1}{2}}\phi^{n-\frac{1}{2}}}{2}dx+\int_{\mathbb{R}^d}\frac{(\phi^{n-\frac{1}{2}})^2}{2}dx$. Therefore, (4.2) reads*

$$E_0 = \int_{\mathbb{R}^d}\left(|\nabla u^n|^2+\frac{\lambda}{2}\phi^{n+\frac{1}{2}}\phi^{n-\frac{1}{2}}\right)dx = \int_{\mathbb{R}^d}\left(|\nabla u^0|^2+\frac{\lambda}{2}\phi^{\frac{1}{2}}\phi^{-\frac{1}{2}}\right)dx.$$

   Since $\phi^{n+\frac{1}{2}}$ represents $|u^n|^2$, this is the notation that we choose in order to show the conservation of the energy.
2. *Since $\phi^{\frac{1}{2}}=\phi^{-\frac{1}{2}}=|u^0|^2$, $E_0=\int_{\mathbb{R}^d}|\nabla u^0(x)|^2+\frac{\lambda}{2}|u^0(x)|^4dx$.*
3. *In the continuous case, we know that we have to find $\phi=|u|^2$. We do not exactly recover this property in the discrete case; however, (4.3) tells us that this property can be recovered in a weaker formulation.*

Proof.

- Multiplying (1.4b) by $\overline{u^{n+1}+u^n}$, integrating in space, taking the imaginary part, and summing in n, we get (4.1).
- Multiplying (1.4b) by $\overline{\frac{u^{n+1}-u^n}{\delta t}}$, integrating in space, and taking the real part, we get $-\int_{\mathbb{R}^d}\frac{|\nabla u^{n+1}|^2-|\nabla u^n|^2}{2\delta t}dx = \lambda\int_{\mathbb{R}^d}\frac{|u^{n+1}|^2-|u^n|^2}{2\delta t}\phi^{n+\frac{1}{2}}dx$. Summing these

equalities for $n \in [0, N]$ yields

$$-\int_{\mathbb{R}^d} \frac{|\nabla u^{N+1}|^2}{2\delta t} - \frac{|\nabla u^0|^2}{2\delta t} dx = \lambda \sum_{n=0}^{N} \int_{\mathbb{R}^d} \frac{|u^{n+1}|^2 - |u^n|^2}{2\delta t} \phi^{n+\frac{1}{2}} dx. \tag{4.4}$$

Multiplying (1.4a) by $\frac{\phi^{n+\frac{1}{2}} - \phi^{n-\frac{1}{2}}}{\delta t}$, we have $\frac{(\phi^{n+\frac{1}{2}})^2 - (\phi^{n-\frac{1}{2}})^2}{2\delta t} = \frac{|u^n|^2 \phi^{n+\frac{1}{2}} - |u^n|^2 \phi^{n-\frac{1}{2}}}{\delta t}$. Summing again from 0 to N gives

$$\begin{aligned}\frac{(\phi^{N+\frac{1}{2}})^2 - (\phi^{-\frac{1}{2}})^2}{2\delta t} &= \sum_{k=0}^{N} \frac{|u^k|^2 \phi^{k+\frac{1}{2}}}{\delta t} - \sum_{k=0}^{N} \frac{|u^k|^2 \phi^{k-\frac{1}{2}}}{\delta t} \\ &= \sum_{k=0}^{N} \frac{(|u^k|^2 - |u^{k+1}|^2)\phi^{k+\frac{1}{2}}}{\delta t} \\ &\quad + \frac{|u^{N+1}|^2 \phi^{N+\frac{1}{2}}}{\delta t} - \frac{|u^0|^2 \phi^{-\frac{1}{2}}}{\delta t}.\end{aligned}$$

Plugging this last equality into (4.4) yields (4.2).

- Finally, $\int_{\mathbb{R}^d} |u^n|^2 dx = \int_{\mathbb{R}^d} |u^{n-1}|^2 dx$; thus $\int_{\mathbb{R}^d} \frac{\phi^{n+\frac{1}{2}} + \phi^{n-\frac{1}{2}}}{2} dx = \int_{\mathbb{R}^d} \frac{\phi^{n-\frac{1}{2}} + \phi^{n-\frac{3}{2}}}{2} dx$. We therefore obtain $\int_{\mathbb{R}^d} \phi^{n+\frac{1}{2}} dx = \int_{\mathbb{R}^d} \phi^{n-\frac{3}{2}} dx$. As $\int_{\mathbb{R}^d} \phi^{\frac{1}{2}} dx = \int_{\mathbb{R}^d} \phi^{-\frac{1}{2}} dx$, we have $\int_{\mathbb{R}^d} \phi^{n+\frac{1}{2}} dx = \int_{\mathbb{R}^d} \phi^{n-\frac{1}{2}} dx$, which implies (4.3) because $\phi^{-\frac{1}{2}} = |u^0|^2$. □

5. Numerical experiments. We test the relaxation scheme on the nonlinear Schrödinger equation

$$i\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} = -2|u|^2 u,$$

with the initial datum $u(x, 0) = \frac{i \exp(2ik_0 x)}{\cosh(x+2)}$, $x \in \mathbb{R}$. We know the exact solution for this initial datum which is the soliton solution $u_{ex}(x, t) = \frac{i \exp i(2k_0 x + (1-4k_0^2)t)}{\cosh(x+2-4k_0 t)}$.

We want to compute in a first experiment the numerical order p_{num}. Let us define Ω as the computational domain. The numerical order p_{num} is computed by

$$p_{\text{num}} = \frac{1}{\ln n} \ln \left(\sup_t \frac{\|u_{n\delta t} - u_{ex}\|_{L^2(\Omega)}(t)}{\|u_{\delta t} - u_{ex}\|_{L^2(\Omega)}(t)} \right).$$

In order to avoid any numerical reflections due to the boundaries, we take $\Omega = [-50, 50]$ and $k_0 = 0$. Then, the Dirichlet boundary condition is almost verified. We use finite difference approximation for the spatial differential operator $\frac{\partial^2 u}{\partial x^2}$. As we want to identify only the time numerical order, we take a fine mesh with $\delta x = 5.10^{-4}$. The number of points is therefore $N = 200000$.

The full discrete scheme reads

$$\begin{aligned}&\phi_j^{n+\frac{1}{2}} = 2|u_j^n|^2 - \phi^{n-\frac{1}{2}}, \\ &i\frac{u_j^{n+1} - u_j^n}{\delta t} + \frac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n}{2\delta x^2} = \frac{u_j^{n+1} + u_j^n}{2} \phi_j^{n+\frac{1}{2}}, \\ &u_0^{n+1} = 0,\ u_N^{n+1} = 0.\end{aligned}$$

TABLE 5.1
Computation of p_{num}.

	$n = 5$	$n = 10$	$n = 50$
p_{num}	2.000713	2.003127	2.002260

TABLE 5.2
Cpu time for relaxation and Crank–Nicolson schemes.

	$N = 1000$	$N = 2000$
Relaxation scheme	7.1s	15.12s
Crank–Nicolson scheme	23.90s	53.32s

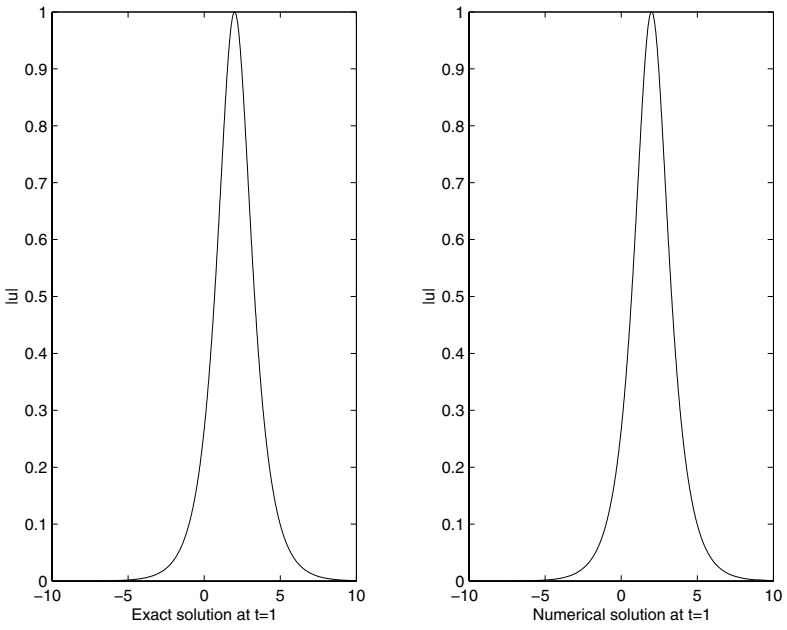


FIG. 5.1. *Comparison between exact and numerical solutions at time* $t = 1$.

The results displayed in Table 5.1 are computed for $\delta t = 10^{-3}$. Then, the order of the relaxation scheme should be two.

In our next experiment, we take $\Omega = [-10, 10]$, $N = 1024$, $\delta t = 10^{-3}$, and the final time $t = 1$. We plot in Figure 5.1 the modulus of the exact and the numerical solutions at $t = 1$. The velocity of the soliton is well preserved as its amplitude. Next, we plot the conserved quantities, namely density (here, equal to 2) and energy (equal to $\frac{11}{3}$), in Figure 5.2. On the one hand, the density is conserved with a precision of 10^{-6} and on the other hand, the energy, which is $\int_{\mathbb{R}} \frac{1}{2}|\nabla u|^2 - \frac{1}{2}|u|^4 dx$, is conserved with a precision of 10^{-2} after 1000 iterations. We plot in Figure 5.3 the numerical error $\|u_{ex} - u_{num}\|_{L^2}$. The result is very accurate. The Crank–Nicolson scheme gives the same results. However, due to the nonlinear step in this scheme, the used cpu time is more important, as we can see in Table 5.2 for different mesh size.

Finally, we experiment with the two-dimensional case. We consider the nonlinear Schrödinger equation $iu_t + \Delta u = -2|u|^2 u$. The computational domain is now $\Omega = [-10, 10] \times [-10, 10]$ and we discretize it with 200×200 points. We take $\delta t = 10^{-2}$,

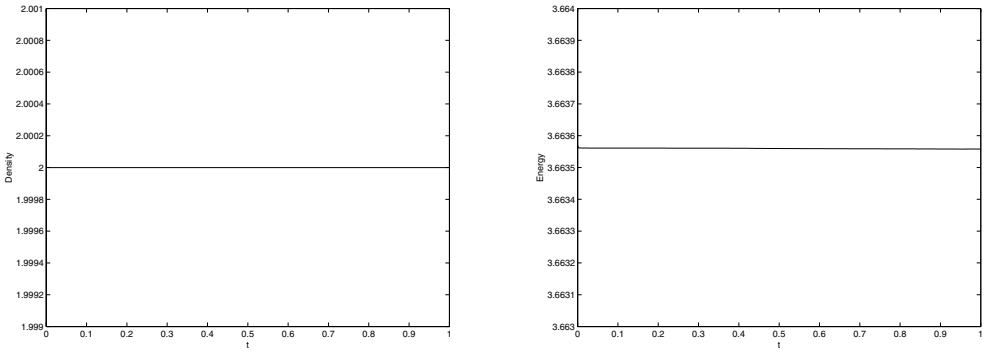


FIG. 5.2. *Density and energy.*

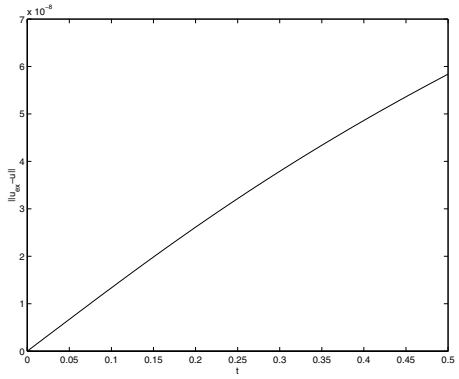


FIG. 5.3. *Error estimate.*

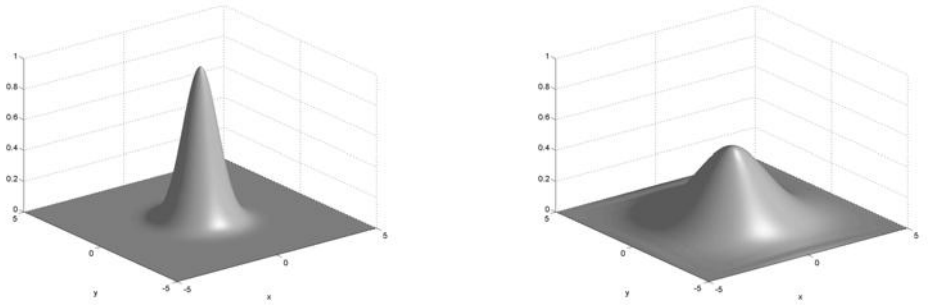


FIG. 5.4. *Initial datum and computed solution at time* $t = 0.5$.

and the final time t is 0.5. The initial datum is $\exp(-x^2 - y^2)$.

The computational time is 166s for the relaxation scheme and 336s for the Crank–Nicolson one. Actually, the nonlinear step involved in the Crank–Nicolson scheme is very costly. At each time step, two or three iterations are necessary in order for the Newton scheme to be convergent. We plot in Figure 5.4 the amplitude of the solution at time $t = 0$ and $t = 0.5$.

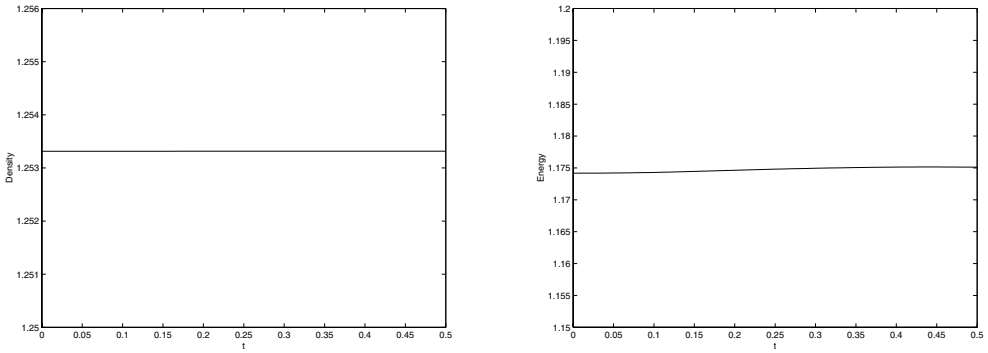


FIG. 5.5. *Density and energy in the two-dimensional case.*

The density and energy are well preserved (see Figure 5.5).

6. Conclusion. We define a new numerical scheme adapted to the nonlinear Schrödinger equation which avoids costly nonlinear computations. We prove local existence of solutions and the convergence of the scheme. Moreover, this scheme conserves the density and the energy quantities. It can be easily adapted to other systems of equations such as the Davey–Stewartson systems, for example.

Appendix. Equivalent system of (2.5). After algebraic calculations on system (2.5), this new system reads

$$\begin{cases} \frac{u^{n+2}-u^{n+1}}{\delta t} - i\Delta\left(\frac{u^{n+2}+u^{n+1}}{2}\right) = -i\lambda U^{n+\frac{3}{2}}, \\ \frac{\phi^{n+\frac{3}{2}}-\phi^{n-\frac{1}{2}}}{2\delta t} = \Phi^{n+\frac{1}{2}}, \\ \frac{v^{n+\frac{3}{2}}-v^{n-\frac{1}{2}}}{2\delta t} - i\Delta\left(\frac{v^{n+\frac{3}{2}}+2v^{n+\frac{1}{2}}+v^{n-\frac{1}{2}}}{4}\right) = -i\lambda V^{n+\frac{1}{2}}, \\ \frac{\psi^{n+1}-\psi^{n-1}}{2\delta t} = \Psi^{n}, \\ \frac{w^{n+1}-w^{n-1}}{2\delta t} - i\Delta\frac{w^{n+1}+w^{n-1}}{2} = -i\lambda W^{n}, \end{cases}$$

with $\psi^{n+1} = \frac{\phi^{n+\frac{3}{2}}-\phi^{n+\frac{1}{2}}}{\delta t}$, $w^{n+1} = \frac{v^{n+\frac{3}{2}}-v^{n+\frac{1}{2}}}{\delta t}$,

$$\begin{aligned} \Psi^n &= 2\mathrm{Re}\left(w^n\,\overline{\tfrac{u^{n+1}+2u^n+u^{n-1}}{4}}\right) + 2\left|\tfrac{v^{n+\frac{1}{2}}+v^{n-\frac{1}{2}}}{2}\right|^2, \\ W^n &= \left(\frac{\psi^{n+1}+\psi^{n-1}}{2}\right)\left(\frac{v^{n+\frac{3}{2}}+3v^{n+\frac{1}{2}}+3v^{n-\frac{1}{2}}+v^{n-\frac{3}{2}}}{8}\right) \\ &\quad + \left(\tfrac{\phi^{n+\frac{3}{2}}+\phi^{n+\frac{1}{2}}+\phi^{n-\frac{1}{2}}+\phi^{n-\frac{3}{2}}}{4}\right)\left(\tfrac{w^{n+1}+2w^n+w^{n-1}}{4}\right) \\ &\quad + 2\left[\mathrm{Re}\left(w^n\left(\overline{\tfrac{u^{n+1}+2u^n+u^{n-1}}{4}}\right)\right)+\left|\tfrac{v^{n+\frac{1}{2}}+v^{n-\frac{1}{2}}}{2}\right|^2\right]\left(\tfrac{u^{n+2}+2u^{n+1}+2u^n+2u^{n-1}+u^{n-2}}{8}\right) \\ &\quad + 2\mathrm{Re}\left(\frac{v^{n+\frac{1}{2}}\,\overline{\frac{u^{n+1}+u^n}{2}}+v^{n-\frac{1}{2}}\,\overline{\frac{u^n+u^{n-1}}{2}}}{2}\right)\left(\tfrac{v^{n+\frac{3}{2}}+v^{n+\frac{1}{2}}+v^{n-\frac{1}{2}}+v^{n-\frac{3}{2}}}{4}\right). \end{aligned}$$

The proof of existence and uniqueness of this new system is similar to that of system (2.4), but is now in $H^s(\mathbb{R}^d)$ with $s > d/2 + 4$ in order to ensure the regularity of w^n. Obviously, we can prove that system (1.4) and this new system are equivalent.

The equivalent continuous version of this system is

$$\begin{cases} u_t - i\Delta u = -i\lambda\phi u, \\ \phi_t = 2\mathrm{Re}(\overline{u}v), \\ v_t - i\Delta v = -i\lambda(2\mathrm{Re}(\overline{u}v)u + \phi v), \\ \psi_t = 2\mathrm{Re}(w\overline{u}) + 2|v|^2, \\ w_t - i\Delta w = -i\lambda(\psi v + \phi w + 2(\mathrm{Re}(w\overline{u} + |v|^2))u + 2\mathrm{Re}(v\overline{u})v), \end{cases}$$

and we can define $\Psi(x,t) = 2\mathrm{Re}(w\overline{u}) + 2|v|^2$ and $W(x,t) = \psi v + \phi w + 2(\mathrm{Re}(w\overline{u} + |v|^2))u + 2\mathrm{Re}(v\overline{u})v$.

Duhamel's formulations for ψ^n and w^n read, respectively,

$$\psi^n = \psi^{-1} + \sum_{k=0}^{p} 2\delta t \Psi^{2k}, \quad n = 2p+1,$$

$$\psi^n = \psi^0 + \sum_{k=1}^{p} 2\delta t \Psi^{2k-1}, \quad n = 2p,$$

and with the convention $D = (1 - i\delta t\Delta)^{-1}(1 + i\delta t\Delta)$ and $E = (1 - i\delta t\Delta)^{-1}$,

$$w^n = D^p w^1 - i\lambda \sum_{k=0}^{p-1} 2\delta t E D^{p-k-1} W^{2(k+1)}, \quad n = 2p+1,$$

$$w^n = D^p w^0 - i\lambda \sum_{k=0}^{p-1} 2\delta t E D^{p-k-1} W^{2k+1}, \quad n = 2p.$$

Acknowledgments. The author gratefully thanks Professor Th. Colin for his help, valuable discussions, and fruitful suggestions. He also would like to thank Professor C.-H. Bruneau for his encouragment and his help.

REFERENCES

[1] G. D. Akrivis, *Finite difference discretization of the cubic Schrödinger equation*, IMA J. Numer. Anal., 13 (1993), pp. 115–124.

[2] O. Karakashian, G. D. Akrivis, and V. A. Dougalis, *On optimal order error estimates for the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 30 (1993), pp. 377–400.

[3] G. D. Akrivis, V. A. Dougalis, and O. Karakashian, *Solving the systems of equations arising in the discretization of some nonlinear PDE's by implicit Runge-Kutta methods*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 251–287.

[4] W. Bao, S. Jin, and P. A. Markowich, *On time-splitting spectral approximations for the Schrödinger equation in the semiclassical regime*, J. Comput. Phys., 175 (2002), pp. 487–524.

[5] C. Besse, *Schéma de relaxation pour l'équation de Schrödinger non linéaire et les systèmes de Davey et Stewartson*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1427–1432.

[6] C. Besse, B. Bidégaray, and S. Descombes, *Order estimates in time of splitting methods for the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 40 (2002), pp. 26–40.

[7] Th. Cazenave, *An Introduction to Nonlinear Schrödinger Equations*, Institudo de Matemática, Rio de Janeiro, Brazil, 1996.

[8] T. Colin and P. Fabrie, *Semidiscretization in time for nonlinear Schrödinger-waves equations*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 671–690.

[9] M. Delfour, M. Fortin, and G. Payre, *Finite-difference solutions of a nonlinear Schrödinger equation*, J. Comput. Phys., 44 (1981), pp. 277–288.

[10] Z. Fei, V. Pérez-Garcia, and L. Vázquez, *Numerical simulation of nonlinear Schrödinger systems: A new conservative scheme*, Appl. Math. Comput., 71 (1995), pp. 165–177.
[11] J.-M. Ghidaglia and J.-C. Saut, *Nonelliptic Schrödinger equations*, J. Nonlinear Sci., 3 (1993), pp. 169–195.
[12] T. Kato, *On nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 46 (1987), pp. 113–129.
[13] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.
[14] J. M. Sanz-Serna and J. G. Verwer, *Conservative and nonconservative schemes for the solution of the nonlinear Schrödinger equation*, IMA J. Numer. Anal., 6 (1986), pp. 25–42.
[15] J. A. C. Weideman and B. M. Herbst, *Split-step methods for the solution of the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 23 (1986), pp. 485–507.
[16] G. Zouraris, *On the convergence of a linear two-step finite element method for the nonlinear Schrödinger equation*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 389–405.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 953–967

ON GEOMETRIC INTERPOLATION BY POLYNOMIAL CURVES*

JERNEJ KOZAK† AND EMIL ŽAGAR†

Abstract. In this paper, geometric interpolation by parametric polynomial curves is considered. Discussion is focused on the case where the number of interpolated points is equal to $r+2$, and $n=r$ denotes the degree of the interpolating polynomial curve. The interpolation takes place in $\mathbb{R}^d$ with $d=n$. Even though the problem is nonlinear, simple necessary and sufficient conditions for existence of the solution are stated. These conditions are entirely geometric and do not depend on the asymptotic analysis. Furthermore, they provide an efficient and stable way to the numeric solution of the problem.

Key words. polynomial curve, geometric interpolation, existence, uniqueness, approximation order

AMS subject classifications. 65D05, 65D07

DOI. 10.1137/S0036142903422077

1. Introduction.

Let a sequence of data points

$$\boldsymbol{T}_0, \boldsymbol{T}_1, \ldots, \boldsymbol{T}_{r+1} \in \mathbb{R}^d, \quad \boldsymbol{T}_i \neq \boldsymbol{T}_{i+1}, \tag{1.1}$$

be given. A parametric curve interpolates these points in the geometric sense if the parameter values at which it passes through the points $\boldsymbol{T}_i$ are not prescribed in advance. In the limiting case of the geometric interpolation, if two consecutive points coincide, this scheme leads to the interpolation of a point and a tangent direction at the same parameter value. Further, threefold interpolation at a point also requires the curvature to be known there, etc. The threefold interpolation by cubics in the plane can be traced back to [3], the paper that initiated the study of geometric interpolation. In order to make the proofs of the results simple, only distinct points $\boldsymbol{T}_i$ will be considered in this paper. The extension to the osculatory case will appear elsewhere.

The disadvantage of the geometric approach is obvious. Namely, the problem of finding the interpolatory curve is nonlinear, so the questions of existence, uniqueness, and computation of the solution arise.

However, there are important advantages, too. Free parameter values at which the points $\boldsymbol{T}_i$ are interpolated may raise the approximation order. This fact has been observed in [3] and in many of the subsequent papers. As a bound for the polynomial geometric interpolation, it has been conjectured in [5] that a parametric polynomial curve of degree n in d-dimensional Euclidean space can, in general, interpolate

$$r+2 = n+1+\left\lfloor \frac{n-1}{d-1} \right\rfloor$$

points in $\mathbb{R}^d$ and reach the same approximation order. The conjecture has been proved only for a few particular cases. But perhaps the most important bonus of all is that the geometric approach provides the basis for the G^m continuous spline schemes where

*Received by the editors January 24, 2003; accepted for publication (in revised form) August 29, 2003; published electronically July 14, 2004.
http://www.siam.org/journals/sinum/42-3/42207.html
†Department of Mathematics and IMFM, University of Ljubljana, Jadranska 19, SI-1000 Ljubljana, Slovenia (Jernej.Kozak@Fmf.Uni-Lj.Si, Emil.Zagar@Fmf.Uni-Lj.Si).

the interpolants do not depend on the local parametrization. This is an important and often-required property in the CAGD applications, and some of the results can be found in [4, 10, 12].

Suppose now that the interpolatory curve is a parametric polynomial curve

$$\boldsymbol{B}_n : [a, b] \to \mathbb{R}^d$$

of degree n. Since linear reparametrization does not change the degree of the polynomial, assumptions $a := 0$ and $b := 1$ can be made. Thus the construction of $\boldsymbol{B}_n$ requires determining t_i,

$$t_0 := 0 < t_1 < t_2 < \cdots < t_r < t_{r+1} := 1, \tag{1.2}$$

such that

$$\boldsymbol{B}_n(t_i) = \boldsymbol{T}_i, \quad i = 0, 1, \ldots, r+1. \tag{1.3}$$

This is the nonlinear part of the problem. Once t_i are known, it is straightforward to obtain the curve $\boldsymbol{B}_n$ in any of the well-known forms, such as Bézier, Newton, or Lagrange.

In order to keep the number of free parameters equal to the number of the unknowns, a Diophantine equation has to be satisfied [4], i.e.,

$$d\,n - (d-1)\,r = d.$$

The case

$$n = r = d \tag{1.4}$$

turns out to be the simplest to handle [6]. Nevertheless, few results can be found in the literature without the assumption that the data points are sufficiently dense and taken from some smooth underlying curve. In the plane case ($d = 2$), some results are included in [10, 8, 9], and in the space case ($d = 3$) they are included in [4].

As it will be shown below, the case $n = r = d$ can be worked out without any asymptotic assumptions. However, if $n > d$, the interpolation problems become much harder to tackle, and the asymptotic approach seems to be necessary [3, 11].

Let us assume (1.4) throughout the paper and simplify the notation of $\boldsymbol{B}_n$ to

$$\boldsymbol{B} := \boldsymbol{B}_n = \boldsymbol{B}_d.$$

The equations that determine the unknowns t_i in this particular case will be worked out in the next section.

The key role in the paper is played by the matrix of data differences,

$$\Delta T := \left(\Delta \boldsymbol{T}_i\right)_{i=0}^{d}, \quad \Delta \boldsymbol{T}_i := \boldsymbol{T}_{i+1} - \boldsymbol{T}_i, \tag{1.5}$$

and by the signs of its minors

$$D_i := \det \left(\Delta \boldsymbol{T}_j\right)_{\substack{j=0 \\ j\neq i}}^{d}, \tag{1.6}$$

i.e., the signs of the volumes of the d-simplexes, based upon the points

$$\boldsymbol{T}_0,\ \boldsymbol{T}_1,\ \ldots,\ \boldsymbol{T}_{i-1},\ \boldsymbol{T}_{i+1} - \Delta \boldsymbol{T}_i,\ \boldsymbol{T}_{i+2} - \Delta \boldsymbol{T}_i,\ \ldots,\ \boldsymbol{T}_{d+1} - \Delta \boldsymbol{T}_i.$$

If the vectors $\Delta \boldsymbol{T}_i$ do not belong to a proper subspace of $\mathbb{R}^d$, the matrix ΔT is of full rank, and the following conclusion can be made.

THEOREM 1.1. *Suppose* rank $\Delta T = d$. *Then the interpolating curve* $\boldsymbol{B}$ *exists if and only if the minors* D_i *are all of the same sign. If* $\boldsymbol{B}$ *exists, it is regular, and the parameter values* $\boldsymbol{t} := (t_i)_{i=1}^d$ *are determined uniquely.*

In the plane case, the signs of the D_i can be identified by certain angles, as has already been observed in [10, 8]. More generally, if the data are convex in the discrete sense, one has

$$\operatorname{sign}(D_0) \;=\; \operatorname{sign}(D_d)\,.$$

The additional requirements of Theorem 1.1 simply guarantee that the data stay convex under the translations

$$\boldsymbol{T}_j \to \boldsymbol{T}_j - \Delta \boldsymbol{T}_i, \quad j = i+1, i+2, \ldots, d, \quad i = 1, 2, \ldots, d-1;$$

i.e., they are not too twisted. Here the sequence of vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$, $\boldsymbol{v}_i \in \mathbb{R}^d$, $m \geq d$, is considered to be convex in the discrete sense if all the volumes of the d-simplexes, i.e.,

$$\det(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}, \ldots, \boldsymbol{v}_{i+d-1}), \quad i = 1, 2, \ldots, m-d+1,$$

are of the same sign.

Let $\mathrm{S}^- : \mathbb{R}^{d+1} \to \{0, 1, \ldots, d\}$ denote the number of strong sign changes in $\boldsymbol{x} := (x_i)_{i=0}^d \in \mathbb{R}^{d+1}$, i.e., the number of actual sign changes in the sequence $x_0, x_1, \ldots, x_d$ with zero terms discarded. Then Theorem 1.1 actually requires that the kernel of ΔT is spanned by

$$\boldsymbol{x} = \left((-1)^i D_i\right)_{i=0}^d,$$

with

$$\mathrm{S}^-(\boldsymbol{x}) = d.$$

This observation can be extended to the case of deficient rank $\Delta T < d$, but then the uniqueness or the regularity of a solution cannot be expected. Still, the following fact can be established.

THEOREM 1.2. *Let* rank $\Delta T < d$. *An interpolating curve* $\boldsymbol{B}$ *of degree* $\leq d$ *can be found if and only if there exists* $\boldsymbol{x} \in \ker \Delta T$ *such that* $\mathrm{S}^-(\boldsymbol{x}) = d$.

In the setup of Theorem 1.2, all regular $\boldsymbol{B}$ will return the same interpolatory curve, considered as a set of points. But the speed of moving along the curve will be different. The additional free parameters should be used to decrease the degree of the interpolating curve if possible. If the obtained lower-degree curve is unique, the proof of Theorem 1.1 can be repeated, and the conclusion that it is regular can also be made. Reduction of the degree is not always possible. As an example, take a cubic curve that interpolates five points in $\mathbb{R}^3$. If the data are lying on a plane, a cubic is still needed, as a quadratic curve can interpolate four planar points in general.

Although the problem of determining the unknowns t_i is nonlinear, there is an efficient and stable way to the numerical solution, given as the following result.

THEOREM 1.3. *Suppose that the requirements of Theorem* 1.1 *are satisfied. The continuation method* [1] *will always compute the numerical solution.*

Practical evidence shows that the best way is to start the continuation method as a one-step method. This step has to be reduced only if in the solution some of the t_is are close together.

2. The equations. Under the assumption (1.4) the system (1.3) can be rewritten as

$$\boldsymbol{B}(t_i) = \boldsymbol{T}_i, \quad i = 0, 1, \ldots, d+1,$$

where the unknowns are (vector) coefficients of the polynomial curve $\boldsymbol{B}$, and scalars $(t_i)_{i=1}^{d}$ have to satisfy (1.2). But the divided difference on arbitrary $d+2$ points maps a polynomial of degree $\leq d$ to zero, so

$$[t_0, t_1, \ldots, t_{d+1}]\boldsymbol{B} = \boldsymbol{0},$$

and $[t_0, t_1, \ldots, t_{d+1}]$ should map the data $\boldsymbol{T}_i$ to zero, too. Since t_i are required to be different, this fact can be written as

$$\sum_{i=0}^{d+1} \frac{1}{\dot{\omega}(t_i)} \boldsymbol{T}_i = \boldsymbol{0}, \quad \omega(t) := \prod_{i=0}^{d+1} (t - t_i), \quad \dot{\omega}(t) := \frac{d\omega}{dt}(t), \tag{2.1}$$

i.e., d scalar equations for d scalar unknowns $t_1, t_2, \ldots, t_d$. The equations (2.1) are the only nonlinear part on the way to the interpolatory curve $\boldsymbol{B}$, and one can efficiently solve them by the continuation method as stated in Theorem 1.3. The final construction of $\boldsymbol{B}$ then follows the function case and is straightforward.

3. The proofs. The assertions in the introduction seem quite simple, but the proofs will take several steps. Here is a brief outline:

1. The system (2.1) will be transformed in a form more suitable for the analysis of the existence and the uniqueness.
2. It will be shown that the existence of a unique solution of the system (2.1) implies that D_i should all be of the same sign (with Lemma 3.1 as part of the proof).
3. Lemma 3.2 will establish the fact that any solution of (2.1) that satisfies (1.2) should be simple, and Lemma 3.3 will ensure that any such solution cannot have t_is arbitrary close.
4. A proof that the system (2.1) has a unique solution for particular data will be outlined in Lemma 3.4.
5. The convex homotopy will help to carry over the conclusions from the particular to the general case in order to complete the proof of Theorem 1.1.
6. The proofs of Theorems 1.2 and 1.3 will complete the section.

Step 1. Let us recall that $[t_0, t_1, \ldots, t_{d+1}]1 = 0$. So the system (2.1) can be rewritten as

$$\sum_{i=0}^{d+1} \frac{1}{\dot{\omega}(t_i)} (\boldsymbol{T}_i - \boldsymbol{T}_0) \;=\; \sum_{i=1}^{d+1} \frac{1}{\dot{\omega}(t_i)} (\boldsymbol{T}_i - \boldsymbol{T}_0) \;=\; \boldsymbol{0} \tag{3.1}$$

or

$$(\boldsymbol{T}_i - \boldsymbol{T}_0)_{i=1}^{d+1} \boldsymbol{\omega} = \boldsymbol{0}, \tag{3.2}$$

where

$$\boldsymbol{\omega} := \left(\frac{1}{\dot{\omega}(t_i)} \right)_{i=1}^{d+1}. \tag{3.3}$$

By inserting $I = U^{-1}U$ between the two factors in (3.2), where

$$U := \begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{d+1,d+1}, \qquad U^{-1} = \begin{pmatrix} 1 & -1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

the equations (3.1) become

$$\Delta T\, \boldsymbol{\omega_\Sigma} = \mathbf{0}, \tag{3.4}$$

with

$$\Delta T = (\boldsymbol{T}_i - \boldsymbol{T}_0)_{i=1}^{d+1}\, U^{-1} \in \mathbb{R}^{d,d+1}$$

defined by (1.5), and

$$\boldsymbol{\omega_\Sigma} := U\boldsymbol{\omega} = \left(\sum_{j=i}^{d+1} \frac{1}{\dot\omega(t_j)} \right)_{i=1}^{d+1}. \tag{3.5}$$

If at least one of the determinants D_i defined in (1.6) is different from zero, then ΔT is of full rank d, and the kernel of ΔT is spanned by the vector

$$\big((-1)^{d+1-i}\, D_i\big)_{i=0}^{d}.$$

Since $\boldsymbol{\omega_\Sigma}$ should be proportional to it, the nonlinear system (3.4) becomes

$$\alpha \sum_{j=i}^{d+1} \frac{1}{\dot\omega(t_j)} = (-1)^{d+1-i} D_{i-1}, \quad i = 1, 2, \dots, d+1, \tag{3.6}$$

i.e., $d+1$ scalar equations for $d+1$ unknowns $\alpha, t_1, t_2, \dots, t_d$.

Step 2. The form of the system (3.6) is suitable to proceed with the next part of the proofs. Let $t_0 := 0 < t_1 < \cdots < t_d < t_{d+1} := 1$, and let $\alpha \neq 0$ be a unique solution of the system (3.6). Then

$$D_d = \alpha\, \frac{1}{\dot\omega(t_{d+1})} \neq 0,$$

and $\operatorname{sign}(D_d) = \operatorname{sign}(\alpha)$. Thus

$$\operatorname{sign}(D_{i-1}) \;=\; \operatorname{sign}(\alpha), \quad i = 1, 2, \dots, d,$$

if and only if

$$\mathrm{S}^-(\boldsymbol{\omega_\Sigma}) \;=\; d.$$

This equality will be established with the help of the following lemma.

LEMMA 3.1. *Let p_i, $1 \le i \le d$, be the interpolating polynomial of degree $\le d+1$ that interpolates the data*

$$p_i(t_j) = \begin{cases} 0, & j = 0, 1, \dots, i-1, \\ 1, & j = i, \quad i+1, \dots, d+1, \end{cases}$$

at $d+2$ distinct points $t_0 < t_1 < \cdots < t_{d+1}$. Then p_i is of degree $d+1$, and the sign of its leading coefficient is equal to $(-1)^{d+1-i}$.

Proof. The interpolating conditions imply $p_i \neq \text{const}$; thus $p_i' \neq 0$. By Rolle's theorem, p_i' has at least $i-1$ zeros on (t_0, t_{i-1}) and at least $d-i+1$ zeros on (t_i, t_{d+1}), i.e., at least d zeros on (t_0, t_{d+1}). Since p_i' does not vanish identically, the degree of p_i' is d, and the degree of p_i is $d+1$. Note that p_i' must be increasing on (t_{i-1}, t_i), and $\operatorname{sign}(p_i'(t_i)) = 1$. But then $\operatorname{sign}(p_i'(t_{d+1})) = (-1)^{d+1-i}$. Since the leading coefficient of p_i has to be of the same sign as $p_i'(t_{d+1})$, the lemma has been proved. ∎

Let p_i be the polynomial studied in Lemma 3.1. Its leading coefficient is equal to the divided difference

$$[t_0, t_1, \ldots, t_{d+1}]p_i = \sum_{j=0}^{d+1} \frac{p_i(t_j)}{\dot{\omega}(t_j)} = \sum_{j=i}^{d+1} \frac{1}{\dot{\omega}(t_j)},$$

and the fact

$$\operatorname{sign}\left(\sum_{j=i}^{d+1} \frac{1}{\dot{\omega}(t_j)}\right) = (-1)^{d+1-i}$$

is confirmed by the conclusion of Lemma 3.1. The first part of the proof of Theorem 1.1 is complete.

Step 3. Let us continue with the next step of the proofs. If two consecutive equations in (3.6) are subtracted, the system reads as

$$\frac{\alpha}{\dot{\omega}(t_i)} = (-1)^{d+1-i}(D_{i-1} + D_i), \quad i = 1, 2, \ldots, d+1, \quad D_{d+1} := 0. \tag{3.7}$$

The system (3.7) will have a simple solution if the Jacobian J at that point is non-singular. A straightforward computation gives J as

$$J := J(\boldsymbol{t}, \alpha)\operatorname{diag}\left(\frac{1}{\dot{\omega}(t_i)}\right)_{i=1}^{d+1} A, \tag{3.8}$$

with $A := (a_{ij})_{i,j=1}^{d+1}$, and

$$a_{ij} = \begin{cases} \dfrac{\alpha}{t_i - t_j}, & i \neq j, \quad j < d+1, \\ \displaystyle\sum_{\substack{k=0 \\ k \neq i}}^{d+1} \frac{\alpha}{t_k - t_i}, & i = j, \quad j < d+1, \\ 1, & j = d+1. \end{cases}$$

The suggestions in [7] will help us to prove the following lemma.

LEMMA 3.2. *The determinant of the matrix A is given as*

$$\det A = d!\,\alpha^d\,(t_0 - t_{d+1})\frac{1}{\dot{\omega}(t_0)}.$$

Proof. By definition, $\det A$ is a sum of terms of the form

$$\text{const} \prod_{i \neq j} \frac{1}{t_i - t_j}, \tag{3.9}$$

where the total degree of the denominator, viewed as a polynomial in the variables

$$t_\ell, \quad \ell = 0, 1, \ldots, d+1,$$

is d, but for some terms const could be zero. The terms involving

$$\frac{1}{t_i - t_j} \quad \text{or} \quad \frac{1}{(t_i - t_j)^2}, \quad i, j = 1, 2, \ldots, d+1, \quad i \neq j, \tag{3.10}$$

could not take part in (3.9). To see this, observe that for fixed $i \neq j$, $0 \leq\ i, j \leq d$, only the elements

$$\begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix} \alpha \begin{pmatrix} \dfrac{1}{t_j - t_i} & \dfrac{1}{t_i - t_j} \\ \dfrac{1}{t_j - t_i} & \dfrac{1}{t_i - t_j} \end{pmatrix} + \text{other terms}$$

in the matrix A are involved. So the contribution of (3.10) to $\det A$ is computed as the determinant of the matrix A where all the other elements in rows i and j and in columns i and j are set to zero. But then all the 2×2 minors obtained from the rows i and j vanish identically, and the Laplace expansion shows that this determinant is equal to zero. A similar argument works for $i = d+1$, $j = 0$, too. But then only the d possible divisors $t_0 - t_i$, $i = 1, 2, \ldots, d$, are left, and $\det A$ has to be of the form

$$\det A = \alpha^d \frac{c}{\prod_{i=1}^{d}(t_0 - t_i)} \alpha^d (t_0 - t_{d+1}) \frac{c}{\dot{\omega}(t_0)},$$

where c is a constant independent of t_i. Since

$$c = \frac{1}{\alpha^d (t_0 - t_{d+1})} \det \big(\operatorname{diag}(t_0 - t_i)_{i=1}^{d+1} A\big),$$

the sequence of limits $t_1 \to t_0, t_2 \to t_0, \ldots, t_d \to t_0$ simplifies c to

$$c = \frac{1}{t_0 - t_{d+1}} \det \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & d & 0 \\ -1 & -1 & \cdots & -1 & t_0 - t_{d+1} \end{pmatrix} = d!,$$

and the lemma is proved. □

It is convenient now to describe all the admissible solutions of (3.7) as a set $\mathcal{D}$,

$$\mathcal{D} := \{\boldsymbol{t} = (t_i)_{i=1}^{d} \,|\, 0 := t_0 < t_1 < \cdots < t_d < t_{d+1} := 1\} \times \{\alpha \,|\, \alpha \neq 0\}.$$

The restriction of $\mathcal{D}$ to the case $\alpha > 0$ will be denoted by $\mathcal{D}_+$ and the restriction to the case $\alpha < 0$ by $\mathcal{D}_-$.

LEMMA 3.3. *Let D_i be all positive (negative). The system* (3.7) *cannot have a solution arbitrary close to the boundary $\partial\mathcal{D}_+$ ($\partial\mathcal{D}_-$).*

Proof. Without loss of generality, one can assume that $\operatorname{sign}(\alpha) = \operatorname{sign}(D_i) > 0$. The last equation in (3.7) reads

$$\frac{\alpha}{\dot{\omega}(t_{d+1})} = \frac{\alpha}{(t_{d+1} - t_0)(t_{d+1} - t_1) \cdots (t_{d+1} - t_d)} = D_d. \tag{3.11}$$

Note that

$$t_{d+1} - t_d \le t_{d+1} - t_i \le 1, \quad i = 0, 1, \dots, d,$$

so the inequality

$$(t_{d+1} - t_d)^{d+1} \le \dot{\omega}(t_{d+1}) \le\ t_{d+1} - t_d$$

gives the bounds on α in (3.11) as

$$(t_{d+1} - t_d)^{d+1} \le \frac{\alpha}{D_d} \le\ t_{d+1} - t_d. \tag{3.12}$$

Since $[t_0, t_1, \dots, t_{d+1}]1 = 0$, summing all equations in (3.7) yields

$$\frac{\alpha}{\dot{\omega}(t_0)} = (-1)^{d+1} D_0,$$

and, further, as in (3.12),

$$(t_1 - t_0)^{d+1} \le \frac{\alpha}{D_0} \le\ t_1 - t_0. \tag{3.13}$$

The inequalities (3.12) and (3.13) show that $t_1 - t_0$ and $t_{d+1} - t_d$ should go to zero with α, i.e.,

$$t_1 - t_0 = o(1),\ t_{d+1} - t_d = o(1) \text{ as } \alpha \to 0. \tag{3.14}$$

More generally, with

$$(t_{i+1} - t_i)^{d+1-i} \prod_{j=0}^{i-1} (t_i - t_j) \le (-1)^{d+1-i}\, \dot{\omega}(t_i) \le (t_{i+1} - t_i) \prod_{j=0}^{i-1} (t_i - t_j)$$

one obtains the bounds from the ith equation in (3.7) as

$$(t_{i+1} - t_i)^{d+1-i} \le \frac{\alpha}{(D_{i-1} + D_i) \prod_{j=0}^{i-1} (t_i - t_j)} \le t_{i+1} - t_i, \quad i = 1, 2, \dots, d. \tag{3.15}$$

Let $\alpha \to 0$. The product $\prod_{j=0}^{i-1}(t_i - t_j)$ cannot go faster to zero as α since the right-hand side in (3.15) is bounded by one. So one has either

$$\frac{\alpha}{\prod_{j=0}^{i-1}(t_i - t_j)} = o(1) \text{ as } \alpha \to 0 \tag{3.16}$$

or

$$\frac{\alpha}{\prod_{j=0}^{i-1}(t_i - t_j)} = \text{const as } \alpha \to 0. \tag{3.17}$$

The possibility (3.16) cannot hold for all i since this would, together with (3.14), imply $t_{d+1} \to t_0$, but $t_{d+1} = 1,\ t_0 = 0$, a contradiction. So at least for one i the equation (3.17) holds, and (3.15) further implies that $t_{i+1} - t_i \ge \text{const}$. Suppose that

ℓ, $1 \le \ell \le d$, is the smallest index such that t_ℓ, $t_{\ell+1}$ are separated, i.e., $t_\ell - t_0 = o(1)$ as $\alpha \to 0$, but $t_{\ell+1} - t_\ell = \text{const} > 0$. Then

$$\frac{1}{t_i - t_j} = \frac{1}{t_0 - t_j}\,(1 + \mathcal{O}(t_i - t_0)), \quad i \le \ell < j,$$

and

$$\frac{1}{\dot\omega(t_i)} = \frac{1}{\prod_{\substack{j=0\\ j\neq i}}^{\ell}(t_i - t_j)}\ \frac{1}{\prod_{j=\ell+1}^{d+1}(t_0 - t_j)}\ \left(1 + \sum_{i=0}^{\ell}\mathcal{O}(t_i - t_0)\right), \quad i \le \ell. \tag{3.18}$$

Let

$$w := \prod_{j=\ell+1}^{d+1}(t_j - t_0)\ \ge (t_{\ell+1} - t_\ell)^{d+1-\ell} = \text{const}^{d+1-\ell} > 0.$$

By inserting (3.18) into (3.7), multiplied by w, one computes

$$\frac{\alpha}{\prod_{\substack{j=0\\ j\neq i}}^{\ell}(t_i - t_j)} = (-1)^{\ell-i} w(D_{i-1} + D_i) + \text{ higher-order terms}, \quad i = 0, 1, \ldots, \ell,$$

and the summing of these equations yields

$$\sum_{i=0}^{\ell} \frac{\alpha}{\prod_{\substack{j=0\\ j\neq i}}^{\ell}(t_i - t_j)}[t_0, t_1, \ldots, t_\ell]\alpha = wD_\ell + \text{ higher-order terms}. \tag{3.19}$$

The left-hand side of (3.19) vanishes, since $\ell \ge 1$ and, consequently, $[t_0, t_1, \ldots, t_\ell]\alpha = 0$. On the other hand, the right-hand side is positive if higher-order terms are small enough. This implies that not all of the first $\ell + 1$ equations of (3.7) can be satisfied if $\alpha \ll 1$. So α has to be bounded away from 0, and the equations (3.7) imply that $t_{i+1} - t_i \ge \text{const} > 0$ for all i. The lemma is confirmed. ∎

Step 4. The fourth step of the proofs considers the system (2.1) with a particular set of data points.

LEMMA 3.4. *Let us suppose that the data points* (1.1) *are taken on the polynomial curve* $\boldsymbol{f}(t) := (t^k)_{k=1}^d$ *as*

$$T_i^* := \boldsymbol{f}(\eta_i), \quad i = 0, 1, \ldots, d+1, \tag{3.20}$$

where

$$\eta_0 := 0 < \eta_1 < \cdots < \eta_d < \eta_{d+1} := 1. \tag{3.21}$$

Then the system of nonlinear equations (2.1) *has a unique solution.*

Proof. First, note that the determinants which correspond to the particular data (3.20) are

$$D_i^* := \det\,(\Delta T_i^*)_{\substack{j=0\\ j\neq i}}^{d} \tag{3.22}$$

and can be computed as

$$D_i^* = d! \int_{\eta_0}^{\eta_1} dx_1 \int_{\eta_1}^{\eta_2} dx_2 \ldots \int_{\eta_{i-1}}^{\eta_i} dx_i \int_{\eta_{i+1}}^{\eta_{i+2}} dx_{i+1} \ldots \int_{\eta_d}^{\eta_{d+1}} V(x_1, x_2, \ldots, x_d) dx_d,$$

where

$$V(x_1, x_2, \ldots, x_d) = \prod_{j>i}(x_j - x_i)$$

is the Vandermonde determinant. This implies that $D_i^* > 0$, and rank $\Delta T^* = d$, since η_i are ordered by (3.21). The necessary conditions of Theorem 1.1 are met, and one of the solutions of (2.1) for the particular data is obviously

$$t_i = \eta_i, \quad i = 1, \ldots, d.$$

In order to complete the proof of Theorem 1.1 for these data, it must be shown that this is the only solution that satisfies (1.2). The system in its basic form (2.1) is

$$\sum_{i=0}^{d+1} \frac{\eta_i^\ell}{\dot\omega(t_i)} = 0, \quad \ell = 1, 2, \ldots, d, \tag{3.23}$$

and the identity $[t_0, t_1, \ldots, t_{d+1}]1 = 0$ can always be added. But then (3.23) is reduced to the fact that the vector

$$\left(\frac{1}{\dot\omega(t_i)}\right)_{i=0}^{d+1}$$

should span the kernel of the matrix $M_{\boldsymbol{\eta}}$, where

$$M_{\boldsymbol{z}} := \left(z_i^\ell\right)_{\ell=0;i=0}^{d;d+1}, \quad \boldsymbol{z} = (z_i)_{i=0}^{d+1},$$

since by assumption (3.21) the matrix $M_{\boldsymbol{\eta}}$ is obviously of full rank, i.e., $d+1$. But

$$M_{\boldsymbol{\eta}} \left(\frac{1}{\dot\omega(\eta_i)}\right)_{i=0}^{d+1} \left([\eta_0, \eta_1, \ldots, \eta_{d+1}]_\eta\, \eta^\ell\right)_{\ell=0}^{d} = \mathbf{0}.$$

Thus $(\frac{1}{\dot\omega(\eta_i)})_{i=0}^{d+1}$ spans the kernel of $M_{\boldsymbol{\eta}}$, too, and

$$\begin{aligned}\mathbf{0} = M_{\boldsymbol{\eta}} \left(\frac{1}{\dot\omega(t_i)}\right)_{i=0}^{d+1} &= \operatorname{const} M_{\boldsymbol{\eta}} \left(\frac{1}{\dot\omega(\eta_i)}\right)_{i=0}^{d+1} \\ &= \operatorname{const} M_{\boldsymbol{t}} \left(\frac{1}{\dot\omega(t_i)}\right)_{i=0}^{d+1} = \operatorname{const}_1 M_{\boldsymbol{t}} \left(\frac{1}{\dot\omega(\eta_i)}\right)_{i=0}^{d+1}.\end{aligned}$$

So t_i and η_i are equivalent, and one can simplify further discussion by exchanging the role of the unknowns and the parameters. Thus suppose t_i to be known and η_i to be determined.

The equations (3.23) imply that the values η_i must be equal to the values $p(t_i)$ of some polynomial p of degree $\le d$, and

$$[t_0, t_1, \ldots, t_{d+1}]p^\ell = 0, \quad \ell = 1, 2, \ldots, d. \tag{3.24}$$

It is easy to see that (3.24) does not, in general, determine the polynomial p uniquely, even for small d. Take $d = 3$, and equidistant partition $t_i = \frac{i}{4}$. Then the divided

difference $[t_0, t_1, t_2, t_3, t_4]$ obviously maps to zero the powers $t^\ell,\ \ell = 1, 2, 3$, but also p^ℓ, where

$$p(t) := \frac{1}{3}t(16 - 45t + 32t^2).$$

However, this p does not produce $\eta_i = p(t_i)$ in the order as required in (3.21) since it is not monotone on $[0, 1]$.

Let us proceed to show that for a particular choice of t_i the solution of (3.24) that satisfies (3.21) is unique. Let $0 < h \ll 1$, and

$$t_i = \frac{i}{d}h, \quad i = 1, 2, \ldots, d.$$

Note that $p(0) = 0,\ p(1) = 1$. Thus p can be expressed as follows:

$$p(t) = \sum_{i=1}^{d} c_i t^i, \quad c_d := 1 - \sum_{i=1}^{d-1} c_i,$$

and the first equation of (3.24) is satisfied automatically. Let us recall that the divided difference can also be written as

$$\oint_{\partial\Omega} \frac{f(z)}{\omega(z)}dz = \sum_{i=0}^{d+1} \operatorname{Res}\left(\frac{f}{\omega}; t_i\right) = [t_0, t_1, \ldots, t_{d+1}]f, \quad t_i \in I,$$

if f is analytical on the set $\Omega \subset \mathbb{C},\ I \subset \Omega$. Here, $\operatorname{Res}(g; z)$ denotes the residuum of g at z. Thus (3.24) can be written as

$$\sum_{i=0}^{d+1} \operatorname{Res}\left(\frac{p^\ell}{\omega}; t_i\right) = 0, \quad \ell = 2, 3, \ldots, d. \tag{3.25}$$

The fraction $\frac{p^\ell}{\omega}$ has only isolated singularities in $\mathbb{C}^*$; therefore

$$\sum_{i=0}^{d+1} \operatorname{Res}\left(\frac{p^\ell}{\omega}; t_i\right) + \operatorname{Res}\left(\frac{p^\ell}{\omega}; \infty\right) = 0, \quad \ell = 2, 3, \ldots, d,$$

and the system (3.25) is simplified to

$$\operatorname{Res}\left(\frac{p^\ell}{\omega}; \infty\right) = 0, \quad \ell = 2, 3, \ldots, d. \tag{3.26}$$

The rational function $\frac{1}{\omega}$ expands at ∞ as

$$\frac{1}{\omega(z)} = \frac{1}{z^{d+2}} + \sum_{i=d+3}^{\infty} \frac{1}{z^i}\left(\frac{d+1}{2}h + \mathcal{O}(h^2)\right).$$

Also,

$$p^\ell(z) = \sum_{k=\ell}^{\ell\, d} z^k \sum_{i_1+i_2+\cdots+i_\ell=k} c_{i_1} c_{i_2} \ldots c_{i_\ell}. \tag{3.27}$$

In (3.27), only the terms with $k \geq d+1$ will contribute to the residue. Since $d+1 > \ell$, the system (3.26) reads

$$\sum_{k=d+1}^{\ell\, d} \sum_{i_1+i_2+\cdots+i_\ell=k} c_{i_1} c_{i_2} \dots c_{i_\ell} + \mathcal{O}(h) = 0, \quad \ell = 2, 3, \dots, d. \tag{3.28}$$

But $p^\ell(1) = 1$, and (3.27) simplifies (3.28) to

$$1 - \sum_{k=l}^{d} \sum_{i_1+i_2+\cdots+i_\ell=k} c_{i_1} c_{i_2} \dots c_{i_\ell} + \mathcal{O}(h) = 0, \quad \ell = d, d-1, \dots, 2. \tag{3.29}$$

First, let us consider (3.29) when $h \to 0$. Then the first two equations read as

$$1 - c_1^d = 0 \tag{3.30}$$

and

$$1 - c_1^{d-1} - d c_1^{d-2} c_2 = 0, \tag{3.31}$$

and the rest read as

$$1 - c_1^\ell - \ell c_1^{\ell-1} c_{d-\ell+1} + g_\ell(c_1, c_2, \dots, c_{d-\ell}) = 0, \quad \ell = d-2, d-3, \dots, 2. \tag{3.32}$$

Equation (3.30) implies that $c_1 = 1$ is the only real solution. This is true also for even d, because $c_1 = -1$ implies that p is not monotone. But then (3.31) implies that $c_2 = 0$, and (3.32) implies that $c_i = 0,\ i = 3, 4, \dots, d-1$. A brief look at (3.29) reveals that $g_\ell(c_1, c_2, \dots, c_{d-\ell})$ involves products that include at least two c_i, with $2 \leq i \leq d - \ell$. So the lower triangular nonlinear system (3.30), (3.31), and (3.32) has a nonsingular Jacobian at the limit point $h = 0$, and the limit solution is

$$(c_1, c_2, \dots, c_{d-1}) = (1, 0, \dots, 0).$$

Thus, by the implicit function theorem, there exists $h_0 > 0$ such that for all $h,\ 0 \leq h \leq h_0$, there is a unique monotone solution p of the system (3.24), i.e., $p(t) = t$, independently of h. Consequently, the system (3.6) has a unique solution (3.21) which implies the unique solution of the system (2.1), too. ∎

Step 5. Consider now the general case. Without loss of generality, one may assume that D_i are all positive. Let us join the particular data D_i^*, defined in (3.22), and the general data D_i with a convex homotopy,

$$D_i(\lambda) := (1-\lambda) D_i^* + \lambda D_i > 0, \quad \lambda \in [0, 1].$$

Let

$$\boldsymbol{H}(\boldsymbol{t}, \alpha; \lambda) := \left(\frac{\alpha}{\dot{\omega}(t_i)} \right)_{i=1}^{d+1} - \Big((-1)^{d+1-i} \big(D_{i-1}(\lambda) + D_i(\lambda) \big) \Big)_{i=1}^{d+1}, \quad \lambda \in [0, 1],$$

so that the system (3.7) is simplified to

$$\boldsymbol{H}(\boldsymbol{t}, \alpha; \lambda) = \boldsymbol{0}. \tag{3.33}$$

Note that $D_i(\lambda) = (1-\lambda)\, D_i^* + \lambda\, D_i \geq \min\{D_i^*, D_i\} > 0$. So the system (3.33) satisfies the requirements of Lemma 3.3 for any $\lambda \in [0,1]$. As a consequence, the zeros of $\boldsymbol{H}$ are apart from the boundary $\partial\mathcal{D}_+$ for all $\lambda \in [0,1]$. Let

$$\mathcal{S} := \{(\boldsymbol{t},\alpha) \in \mathcal{D}_+ \,|\, \boldsymbol{H}(\boldsymbol{t},\alpha;\lambda) = 0,\ \lambda \in [0,1]\}$$

be the set of solutions of (3.33) for all $\lambda \in [0,1]$, and let $Co(\mathcal{S}) \supset \mathcal{S}$ be its convex hull. The set $\mathcal{D}_+$ is convex. Thus Lemma 3.3 implies that $Co(\mathcal{S}) \subset \mathcal{D}_+$. But $\mathcal{D}_+$ is an open set, and the compact set $Co(\mathcal{S})$ can be enlarged a little to a compact set $\widetilde{\mathcal{D}}$ with a smooth boundary so that the following relations are satisfied:

$$Co(\mathcal{S}) \subset \widetilde{D} \subset \mathcal{D}_+, \qquad Co(\mathcal{S}) \cap \partial\widetilde{D} = \emptyset.$$

The map $\boldsymbol{H}$ is clearly differentiable on $\widetilde{\mathcal{D}}$ and does not vanish on the boundary $\partial\widetilde{\mathcal{D}}$. But then Brouwer's degree [2, pp. 52–53] of $\boldsymbol{H}$ is invariant for $\lambda \in [0,1]$ on $\widetilde{\mathcal{D}}$. In $\boldsymbol{H}$, only the data depend on λ, and a brief look at the homotopy reveals that its Jacobian is simply $J(\boldsymbol{t},\alpha)$, as given in (3.8). This simplifies Brouwer's degree to

$$\sum_{(\boldsymbol{t},\alpha)\in\widetilde{\mathcal{D}},\, \boldsymbol{H}(\boldsymbol{t},\alpha;\lambda)=0} \operatorname{sign}\big(\det J(\boldsymbol{t},\alpha)\big).$$

By Lemma 3.2, $\det J$ vanishes nowhere in $\mathcal{D}_+ \supset \widetilde{\mathcal{D}}$, and Brouwer's degree is further simplified to

$$\pm\, \#\,\{(\boldsymbol{t},\alpha)\,|\,(\boldsymbol{t},\alpha) \in \widetilde{\mathcal{D}},\ \boldsymbol{H}(\boldsymbol{t},\alpha;\lambda) = 0\},$$

so it provides the exact count of zeros in $\widetilde{\mathcal{D}}$. But by Lemma 3.4 the particular problem $\boldsymbol{H}(\boldsymbol{t},\alpha;0) = 0$ has a unique solution in $\widetilde{\mathcal{D}}$, as do all $\boldsymbol{H}(\boldsymbol{t},\alpha;\lambda) = 0$.

In order to complete the proof of Theorem 1.1, it remains to show that $\boldsymbol{B}$, based upon $\boldsymbol{t}$ that we have just determined, is a regular curve.

Note that $\boldsymbol{B}$ can also be written as

$$\boldsymbol{B} = \sum_{j=0}^{d+1} \boldsymbol{T}_j \ell_j, \qquad \ell_j(t) := \frac{\omega(t)}{(t-t_j)\dot{\omega}(t_j)}.$$

If $\boldsymbol{B}$ is not regular, then

$$\dot{\boldsymbol{B}}(\tilde{t}) = 0 = \Delta T U(\dot{\ell}_i(\tilde{t}))_{i=1}^{d+1}$$

for some $\tilde{t} \in [0,1]$. Since $\ker \Delta T$ is spanned by $\boldsymbol{\omega}_\Sigma = U\boldsymbol{\omega}$, given in (3.5) and (3.3), the vector $\big(\ell_i(\tilde{t})\big)_{i=1}^{d+1}$ should be proportional to $\boldsymbol{\omega}$. But then

$$\dot{\omega}(t_i)\dot{\ell}_i(\tilde{t}) = \left(\frac{\dot{\omega}(\tilde{t})}{\tilde{t}-t_i} - \frac{\omega(\tilde{t})}{(\tilde{t}-t_i)^2}\right) = \text{const}$$

for all $t_i \neq \tilde{t}$, which implies that at least two of t_i are equal, a contradiction that confirms the regularity of the interpolating curve. The proof of Theorem 1.1 is complete.

Step 6. The proof of Theorem 1.3 will precede the proof of Theorem 1.2.

Proof of Theorem 1.3. The conclusion of the theorem follows from Lemma 3.2. Since the Jacobian of $\boldsymbol{H}$ is globally nonsingular it is also nonsingular on the set $\boldsymbol{H}^{-1}(\boldsymbol{0})$; thus $\boldsymbol{0}$ is a regular value of the map $\boldsymbol{H}$. This implies [1, p. 38] that the

convergence of the Euler–Newton method as a continuation method is ensured for a suitable small step-length when the solution curve from the particular system ($\lambda = 0$) to our general one ($\lambda = 1$) is traced. □

Proof of Theorem 1.2. If the interpolating polynomial $\boldsymbol{B}$ exists, then the corresponding $\boldsymbol{\omega_\Sigma} \in \ker \Delta T$, as defined in (3.5), clearly satisfies $\mathrm{S}^-(\boldsymbol{\omega_\Sigma}) = d$. On the other hand, if $\boldsymbol{x} = (x_i)_{i=0}^d \in \ker \Delta T$ can be found such that $\mathrm{S}^-(\boldsymbol{x}) = d$, then x_i may replace the right-hand side $(-1)^{d+1-i} D_i$ in (3.6). The existence part of Theorem 1.1 still carries through, and Theorem 1.2 is proved. □

Let us illustrate the last proof by a simple example. Let data be given on a line in a plane,

$$\boldsymbol{T}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{T}_1 = \frac{1}{3}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{T}_2 = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{T}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{3.34}$$

Then

$$\Delta T = \frac{1}{6}\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 3 \end{pmatrix},$$

and rank $\Delta T = 1$. Furthermore, the vector $\boldsymbol{x} \in \ker \Delta T$ such that $\mathrm{S}^-(\boldsymbol{x}) = d = 2$ is given as a parametric family:

$$\boldsymbol{x} = \boldsymbol{x}(\mu) := (\mu, -3 - 2\mu, 1), \qquad \mu > 0.$$

For such an $\boldsymbol{x}$, the system (3.6) has the solution

$$t_1 = t_1(\mu) := \frac{1}{1-\mu}\left(1 - \sqrt{\frac{2\mu(\mu+2)}{3(\mu+1)}}\right), \qquad t_2 = t_2(\mu) := t_1(\mu) + \sqrt{\frac{\mu}{6(\mu+1)(\mu+2)}},$$

and the data (3.34) are interpolated by a quadratically parametrized line

$$\boldsymbol{B} = \phi\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \phi(t) := \frac{(1-2t_2^2)t - (1-2t_2)t^2}{2t_2(1-t_2)}.$$

Furthermore, the curve $\boldsymbol{B}$ is regular if and only if

$$1 - \frac{1}{\sqrt{2}} \le t_2 \le \frac{1}{\sqrt{2}}, \quad \left(\sqrt{3}-1\right)\left(\sqrt{2}-1\right) \le \mu \le \left(\sqrt{2}+1\right)\left(2+\sqrt{6}\right).$$

There is only one free parameter to decrease the degree of $\boldsymbol{B}$, and $t_2 = t_2(2) = \frac{1}{2}$ reduces the parametrization to the simplest case $\phi(t) = t$. This parametrization is regular since it is a unique solution of degree one of the interpolation problem. This concludes the proofs.

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Continuation Methods*, Springer-Verlag, Berlin, 1990.

[2] M. S. Berger, *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.

[3] C. de Boor, K. Höllig, and M. Sabin, *High accuracy geometric Hermite interpolation*, Comput. Aided Geom. Design, 4 (1987), pp. 269–278.

[4] Y. Y. Feng and J. Kozak, *On spline interpolation of space data*, in Mathematical Methods for Curves and Surfaces II, M. Dæhlen, T. Lyche, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, pp. 167–174.

[5] K. Höllig and J. Koch, *Geometric Hermite interpolation with maximal order and smoothness*, Comput. Aided Geom. Design, 13 (1996), pp. 681–695.

[6] J. Kozak and E. Žagar, *On curve interpolation in* $\mathbb{R}^d$, in Curve and Surface Fitting: Saint Malo 1999, A. Cohen, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 2000, pp. 263–273.

[7] C. Krattenthaler, *Advanced determinant calculus*, Sém. Lothar. Combin., 42 (1999), Article B42q, 67 pp.

[8] K. Mørken, *Parametric interpolation by quadratic polynomials in the plane*, in Mathematical Methods for Curves and Surfaces, M. Dæhlen, T. Lyche, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1995, pp. 385–402.

[9] K. Mørken and K. Scherer, *A general framework for high-accuracy parametric interpolation*, Math. Comp., 66 (1997), pp. 237–260.

[10] R. Schaback, *Interpolation with piecewise quadratic visually* C^2 *Bézier polynomials*, Comput. Aided Geom. Design, 6 (1989), pp. 219–233.

[11] K. Scherer, *Parametric polynomial curves of local approximation of order* 8, in Curve and Surface Fitting: Saint Malo 1999, A. Cohen, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 2000, pp. 375–384.

[12] E. Žagar, *On* G^2 *continuous spline interpolation of curves in* $\mathbb{R}^d$, BIT, 42 (2002), pp. 670–688.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 968–973

SATURATION OF REGULARIZATION METHODS FOR LINEAR ILL-POSED PROBLEMS IN HILBERT SPACES*

PETER MATHÉ†

Abstract. We prove the saturation of methods for solving linear ill-posed problems in Hilbert spaces for a wide class of regularization methods. It turns out that, under a certain convexity assumption, saturation must necessarily occur. We provide easy to verify assumptions, which allow us to calculate the rate at which saturation occurs.

Key words. saturation, regularization, ill-posed problem

AMS subject classifications. 47A52, 65F22

DOI. 10.1137/S0036142903420947

1. Introduction and problem formulation. We study the numerical solution of operator equations $Ax = y$ under the presence of noise, which means we are given

$$y_\delta = Ax + \delta\xi,$$

where the operator A acts between Hilbert spaces X and Y. Moreover, the noise ξ is assumed to be bounded $\|\xi\| \le 1$. If A acts injectively and has dense range, then the problem of recovering the unknown x from noisy observations y_δ is known to be ill-posed. In this case we are interested in *regularization methods* given by some operator function $\alpha \to g_\alpha(A^*A)$, $0 < \alpha \le a$; i.e., the approximation to $x \in X$ is given by choosing some $\alpha = \alpha(\delta)$ and letting

$$x_{\alpha,\delta} := g_\alpha(A^*A)A^*y_\delta.$$

The most famous method is *Tikhonov–Phillips regularization*, where

$$x_{\alpha,\delta} := (\alpha \mathrm{I} + A^*A)^{-1}A^*y_\delta.$$

The error of any regularization g_α for approximating x, based on observations y_δ, is given by $e(x, g_\alpha, \delta) := \sup_{\|\xi\|\le 1} \|x - g_\alpha(A^*A)A^*y_\delta\|$. For the performance of a given regularization it is interesting to know how sensitive it is with respect to a priori smoothness assumptions. This is captured by the notion of the *qualification* of the regularization: The higher its qualification, the more it is capable of reacting to smoothness assumptions. In this respect, Tikhonov–Phillips regularization is known to have qualification 1; more generally, it can provide the optimal order of accuracy for all concave source conditions within the framework of variable Hilbert scales; see, e.g., [3, 6]. We will not turn to this subject in this paper.

Here we ask for *saturation*, i.e., the maximal smoothness, where a given regularization can provide the best possible order of approximation. This is made precise in section 2. The main result establishes fairly general conditions under which saturation must occur; the core of them is a certain *convexity property*. The saturation phenomenon has been studied several times. We do not recall the history but refer

*Received by the editors January 8, 2003; accepted for publication (in revised form) November 12, 2003; published electronically July 14, 2004.
http://www.siam.org/journals/sinum/42-3/42094.html
†Weierstraß Institute for Applied Analysis and Stochastics, Mohrenstraße 39, D–10117 Berlin, Germany (mathe@wias-berlin.de).

the reader to [2], summarizing also the recent [7, 8]. Our attitude is different. Most importantly, the analysis is based on a *geometric concept*, which indicates why saturation occurs. As a side effect, this approach is so general that it covers fairly general regularization methods.

We conclude our study with several examples to exhibit that these conditions are easy to verify in specific situations.

2. Main result. As indicated in section 1 we discuss regularization in the form of

$$x_{\alpha,\delta} := g_\alpha(A^*A)A^*y_\delta.$$

More formally, given any bounded operator $A: X \to Y$, the operator A^*A acts in Y, and we assume it has a norm bound $\|A^*A\| \le a$. By spectral calculus, each bounded (Borel-measurable) function defined on $(0, a]$ taking real values can be assigned a respective function taking nonnegative operators to self-adjoint ones. Therefore we may and do identify g_α with its real-valued function. In terms of real functions g_α the requirements can be expressed as follows.

DEFINITION 2.1 (see [9, 6]). *A family g_α, $0 < \alpha \le a$, of bounded (Borel-measurable) functions is called regularization if there are constants γ_* and γ for which*

$$\sup_{0<\lambda\le a} |1 - \lambda g_\alpha(\lambda)| \le \gamma, \quad 0 < \alpha \le a,$$

and

$$\sup_{0<\lambda\le a} \sqrt{\lambda}\, |g_\alpha(\lambda)| \le \frac{\gamma_*}{\sqrt{\alpha}}, \quad 0 < \alpha \le a. \tag{2.1}$$

The regularization g_α is said to have qualification *ρ, for an increasing function $\rho : (0, a] \to \mathbb{R}_+$, if $\sup_{0<\lambda\le a} |1 - \lambda g_\alpha(\lambda)|\, \rho(\lambda) \le \gamma\rho(\alpha), 0 < \alpha \le a$.*

Remark 1. The above notion of qualification extends the usual one, where the functions ρ are restricted to monomials $\rho(t) := t^\mu$ for $0 < \mu < \infty$. The analysis of regularization methods in this general framework was carried out in [6]. The results of that study are not a prerequisite for the present discussion, but the ideas also rely on geometric concepts.

The main objective in this study is *saturation*, which is made precise as follows.

DEFINITION 2.2. *A family g_α, $0 < \alpha \le a$, is said to have* saturation *ψ, for a certain function $\psi: [0, a] \to \mathbb{R}_+$, if for all $0 \ne x \in X$ there are $\delta(x) > 0$ and a constant $c := c(x) > 0$ such that $\inf_{0<\alpha\le a} \frac{e(x,g_\alpha,\delta)}{\psi(\delta)} \ge c$ for $0 < \delta \le \delta(x)$.*

Remark 2. The above notion is consistent with and formalizes the usual approach. For instance, Theorem 3.1 in [8] may be rephrased as follows: The rate of Tikhonov regularization cannot exceed $\delta^{2/3}$ as $\delta \to 0$, regardless of the choice of regularization parameter and regardless of the actual smoothness of $x \ne 0$.

We shall reprove this result within a general framework.

The main result in this study is as follows.

THEOREM 2.3. *Suppose regularization $\alpha \to g_\alpha$ is chosen such that*

(i) *for some $c > 0$ the following lower bound is valid:*

$$\sup_{0<\lambda\le a} \sqrt{\lambda}\, |g_\alpha(\lambda)| \ge \frac{c}{\sqrt{\alpha}}, \quad 0 < \alpha \le a; \tag{2.2}$$

(ii) *for some increasing function ρ, the regularization has* maximal qualification *ρ; i.e., for all $0 < \lambda \le a$ there is $c := c(\lambda) > 0$ for which*

$$\inf_{0<\alpha\le a} \frac{|1-\lambda g_\alpha(\lambda)|}{\rho(\alpha)} \ge c; \tag{2.3}$$

(iii) *for all $0 < \alpha \le a$ the functions*

$$\lambda \longrightarrow |1-\lambda g_\alpha(\lambda)|^2, \quad 0 < \lambda \le a, \tag{2.4}$$

are convex.

Let $\Theta(t) := \sqrt{t}\rho(t), t > 0$. Then g_α has saturation $\rho\circ\Theta^{-1}$; i.e., there is $c > 0$ for which $\inf_{0<\alpha\le a} \frac{e(x,g_\alpha,\delta)}{\rho(\Theta^{-1}(\delta))} \ge c\|x\|$.

Let us comment on the assumptions. The bound in (2.2) serves as a normalization for the choice of the regularization parameter α and is related to (2.1).

The second assumption (2.3) is related to the qualification. Assumption (2.4) in [7] is a similar condition for the case of saturation of power type.

However, to prove saturation in [7], this assumption has to be accompanied by an additional set ((2.5)–(2.9)) of technical ones. In this paper we clearly indicate a geometric assumption (2.4), which seems to be responsible for saturation.

Last, the function Θ from above actually controls the best possible accuracy under the a priori smoothness assumption related to ρ. So, if the regularization has qualification ρ and the actual smoothness is just or not covered by ρ, then $\delta \to \rho(\Theta^{-1}(\delta))$ yields the best possible accuracy; see [6] for details.

Finally, as seen from the proof, the constant c in the saturation bound depends on x only through the quotient $\|Ax\|/\|x\|$.

The proof is based on the following two lemmas.

LEMMA 2.4. *Let $\rho\colon (0,a] \to \mathbb{R}_+$ be any increasing function. For all $0 < \alpha, \delta \le a$ we have*

$$\max\left\{\rho(\alpha), \delta/\sqrt{\alpha}\right\} \ge \rho(\Theta^{-1}(\delta)). \tag{2.5}$$

Proof. By assumption, the function Θ is increasing, and $\lim_{t\to 0}\Theta(t) = 0$. Therefore, we may let $\beta := \Theta^{-1}(\delta)$. Then (2.5) can be rewritten as

$$\max\left\{\rho(\alpha), \sqrt{\beta/\alpha}\rho(\beta)\right\} \ge \rho(\beta).$$

The latter inequality can easily be verified by distinguishing the two cases $\alpha \le \beta$ and $\alpha > \beta$. ∎

LEMMA 2.5 (Peierls–Bogolyubov [1, Problem IX.8.14]). *Suppose $\varphi : [0,a] \to \mathbb{R}_+$ is bounded measurable such that φ^2 is convex. Then $\|x\|\varphi(\frac{\|Ax\|^2}{\|x\|^2}) \le \|\varphi(A^*A)x\|$.*

Proof. The operator A^*A admits a spectral family $(E_\lambda)_{\lambda\in(0,a]}$ for which

$$\|Ax\|^2 = \langle A^*Ax, x\rangle = \int \lambda d\langle E_\lambda x, x\rangle, \quad x \in X.$$

Because $\int d\langle E_\lambda x, x\rangle = \|x\|^2$ we can use Jensen's inequality to conclude that

$$\begin{aligned}\varphi^2\left(\frac{\|Ax\|^2}{\|x\|^2}\right) &= \varphi^2\left(\frac{1}{\|x\|^2}\int \lambda d\langle E_\lambda x, x\rangle\right)\\ &\le \frac{1}{\|x\|^2}\int \varphi^2(\lambda) d\langle E_\lambda x, x\rangle = \frac{\|\varphi(A^*A)x\|^2}{\|x\|^2}.\end{aligned}$$

The proof is complete. ∎

Remark 3. We note that the above inequality is a special case of an interpolation type inequality, which was established in [4]. For more details we refer the reader to [5].

We are now ready to prove Theorem 2.3.

Proof. For each fixed α and δ we consider the function $\bar{e}\colon X \times Y \to \mathbb{R}_+$, given by

$$(x,\xi) \to \bar{e}(x, g_\alpha, \delta\xi) := \|(\mathrm{I} - g_\alpha(A^*A)A^*A)x - \delta g_\alpha(A^*A)A^*\xi\|. \tag{2.6}$$

Furthermore, we shall exploit the following fact: For any $x, y \in X$ it holds that $\max\{\|x+y\|, \|x-y\|\} \geq \|x\|$. By symmetry, for fixed x we have

$$e(x, g_\alpha, \delta) \geq \max\{\bar{e}(x, g_\alpha, \delta\xi), \bar{e}(x, g_\alpha, -\delta\xi)\} \geq \|(\mathrm{I} - g_\alpha(A^*A)A^*A)x\|.$$

Also, by similar reasoning, for fixed ξ

$$\max\{\bar{e}(x, g_\alpha, \delta\xi), \bar{e}(-x, g_\alpha, \delta\xi)\} \geq \delta\|g_\alpha(A^*A)A^*\xi\|.$$

Since

$$e(x, g_\alpha, \delta) \geq \max_{\|\xi\|\leq 1} \max\{\bar{e}(x, g_\alpha, \delta\xi), \bar{e}(-x, g_\alpha, \delta\xi)\},$$

we arrive at $e(x, g_\alpha, \delta) \geq \max\{\|(\mathrm{I} - g_\alpha(A^*A)A^*A)x\|, \delta\|g_\alpha(A^*A)A^*\colon Y \to X\|\}$. Using the convexity assumption for (2.4) and Lemma 2.5 we deduce with $\bar{\lambda} := \|Ax\|/\|x\|$ that

$$\|(\mathrm{I} - g_\alpha(A^*A)A^*A)x\| \geq \|x\| \left|1 - \bar{\lambda}^2 g_\alpha(\bar{\lambda}^2)\right|.$$

By assumption (2.3), we can find $c > 0$ for which

$$\|(\mathrm{I} - g_\alpha(A^*A)A^*A)x\| \geq c\|x\|\rho(\alpha).$$

Finally, assumption (2.2) allows us to determine $c'(x)$ for which

$$e(x, g_\alpha, \delta) \geq c'(x) \max\left\{\rho(\alpha), \delta/\sqrt{\alpha}\right\}.$$

Applying Lemma 2.4 allows us to complete the proof. ∎

It may not be clear from the very beginning whether ρ from (2.3) exists. This can be clarified under an additional monotonicity assumption. For this purpose it is convenient to introduce the function $r\colon (0,a] \times (0,a] \to \mathbb{R}$ as $r(\alpha,\lambda) := r_\alpha(\lambda) := 1 - \lambda g_\alpha(\lambda)$.

COROLLARY 2.6. *Let regularization g_α be chosen such that* (2.2) *is satisfied. If*

1. *for all α the function $\lambda \to |r_\alpha(\lambda)|^2$ is convex and*
2. *for all λ the function $\alpha \to |r_\alpha(\lambda)|$ is nondecreasing,*

then the function $\alpha \to \rho(\alpha) := \inf_{0<\lambda\leq a} |r_\alpha(\lambda)|$ is nondecreasing and obeys (2.3)*, with constant $c = 1$.*

Thus g_α has saturation at $\rho \circ \Theta^{-1}$, precisely, $e(x, g_\alpha, \delta) \geq \|x\|\rho(\Theta^{-1}(c\delta/\|x\|))$, with c from (2.2).

Proof. We need only to show that ρ is nondecreasing and obeys (2.3) with constant 1. Plainly, for $\alpha < \beta$ we have

$$\begin{aligned}\rho^2(\alpha) \leq \sup_{\xi\leq\beta} \rho^2(\xi) &= \sup_{\xi\leq\beta} \inf_{0<\lambda\leq a} |r_\xi(\lambda)|^2 \\ &\leq \inf_{0<\lambda\leq a} \sup_{\xi\leq\beta} |r_\xi(\lambda)|^2 = \inf_{0<\lambda\leq a} |r_\beta(\lambda)|^2 = \rho^2(\beta).\end{aligned}$$

Moreover, by construction, for all α, λ it holds that $|r_\alpha(\lambda)|/\rho(\alpha) \geq 1$, which allows us to complete the proof. ∎

3. Examples. We close our investigation with several examples. It is easy to see that Examples 1–3 below also obey the assumptions made in Corollary 2.6.

Example 1. Tikhonov regularization $g_\alpha(\lambda) := 1/(\alpha+\lambda)$ obeys the assumptions of the theorem with $\rho(\alpha) = \alpha$. Therefore, for each $x \neq 0$ it holds that $e(x, g_\alpha, \delta) \geq c\delta^{2/3}$, as $\delta \to 0$.

Example 2. n-fold iterated Tikhonov regularization, corresponding to

$$g_\alpha^n(\lambda) := 1/\lambda \left(1 - \left(\frac{\alpha}{\alpha+\lambda}\right)^n\right),$$

allows us to apply the theorem with $\rho(\alpha) = \alpha^n$, since for each λ there is $c'(\lambda)$, where

$$\left(\frac{\alpha}{\alpha+\lambda^2}\right)^n \geq c'\alpha^n.$$

Thus it has maximal qualification $\rho(\alpha) = \alpha^n$, and hence saturation at $\psi(\delta) := \delta^{2n/(2n+1)}$.

It is interesting to see that regularization with infinite qualification in the classical sense may have saturation. We exhibit this in the following example.

Example 3. Landweber iteration $g_\alpha(\lambda) := 1/\lambda \left(1 - (1-\mu\lambda)^{1/\alpha}\right)$, where $\mu < 1/a$ and $1/\alpha \geq 1$, corresponding to the number of iterations greater than or equal to 1, has maximal qualification $\rho_\kappa(\alpha) := \exp(-\kappa/\alpha)$ for some positive κ. Therefore, it can be seen that it has saturation at $\psi(\delta) := \delta(\log(1/\delta))^{1/2\kappa}$.

This example shows that even powerful regularization still may not allow us to regularize very mildly ill-posed problems properly.

If there is no maximal qualification, then the saturation phenomenon need not occur.

Example 4. The spectral cut-off, which is defined as

$$g_\alpha(\lambda) := \begin{cases} 1/\lambda, & \alpha \leq \lambda \leq a, \\ 0, & 0 < \lambda < \alpha, \end{cases}$$

has arbitrary qualification. More specifically, $|1 - \lambda g_\alpha(\lambda)| = 0$, provided that $\alpha < \lambda$. Therefore, there is no maximal qualification in the sense of (2.3). Also, the convexity assumption for (2.4) is violated. Thus, if, for example, the element x has a finite expansion in $u_1, u_2, \ldots$, say of length n, then the spectral cut-off has error of the order δ, which clearly is best possible, if only $\delta \leq s_n$ and $\alpha = a$. This can be seen using the representation of the error as presented in (2.6). We refer the reader to [6] for more details on the behavior of the spectral cut-off.

Acknowledgments. I thank Sergei V. Pereverzev, Kiev, for stimulating discussions on this subject and for reading the manuscript.

Furthermore, I thank an anonymous referee for pointing out the paper [7].

REFERENCES

[1] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.
[2] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.
[3] M. Hegland, *An optimal order regularization method which does not use additional smoothness assumptions*, SIAM J. Numer. Anal., 29 (1992), pp. 1446–1461.
[4] M. Hegland, *Variable Hilbert scales and their interpolation inequalities with applications to Tikhonov regularization*, Appl. Anal., 59 (1995), pp. 207–223.

[5] P. Mathé and S. V. Pereverzev, *Discretization strategy for linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 1263–1277.
[6] P. Mathé and S. V. Pereverzev, *Geometry of linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789–803.
[7] A. Neubauer, *On converse and saturation results for regularization methods*, in Beiträge zur angewandten Analysis und Informatik, Shaker, Aachen, Germany, 1994, pp. 262–270.
[8] A. Neubauer, *On converse and saturation results for Tikhonov regularization of linear ill-posed problems*, SIAM J. Numer. Anal., 34 (1997), pp. 517–527.
[9] G. M. Vaĭnikko and A. Yu. Veretennikov, *Iteratsionnye protsedury v nekorrektnykh zadachakh*, Nauka, Moscow, 1986.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 974–996

HIGH-ORDER STRONG-STABILITY-PRESERVING RUNGE–KUTTA METHODS WITH DOWNWIND-BIASED SPATIAL DISCRETIZATIONS*

STEVEN J. RUUTH† AND RAYMOND J. SPITERI‡

Abstract. Strong-stability-preserving Runge–Kutta (SSPRK) methods are a specific type of time discretization method that have been widely used for the time evolution of hyperbolic partial differential equations (PDEs). Under a suitable stepsize restriction, these methods share a desirable nonlinear stability property with the underlying PDE, e.g., stability with respect to total variation, the maximum norm, or other convex functionals. This is of particular interest when the solution exhibits shock-like or other nonsmooth behavior. Many results are known for SSPRK methods with nonnegative coefficients. However, it has recently been shown that such methods cannot exist with order greater than 4. In this paper, we give a systematic treatment of explicit SSPRK methods with general (i.e., possibly negative) coefficients up to order 5. In particular, we show how to optimally treat negative coefficients (corresponding to a change in the upwind direction of the spatial discretization) in the context of *effective CFL coefficient* maximization and provide proofs of optimality of some explicit SSPRK methods of orders 1 to 4. We also give the first known explicit fifth-order SSPRK schemes and show their effectiveness in practice versus more well-known fifth-order explicit Runge–Kutta schemes.

Key words. downwinding, strong stability preserving, total variation diminishing, Runge–Kutta methods, high-order accuracy, time discretization

AMS subject classifications. 65L06, 65M20

DOI. 10.1137/S0036142902419284

1. Introduction.

Solutions to hyperbolic partial differential equations (PDEs) are commonly approximated by sequentially discretizing the spatial and temporal derivatives. For example, in the method of lines, a discretization of the spatial derivatives of the PDE is carried out to produce a large set of coupled time-dependent ordinary differential equations (ODEs). These ODEs can then be treated by suitable time-stepping techniques such as linear multistep or Runge–Kutta methods.

In the numerical solution of hyperbolic PDEs, difficulties may arise due to the presence of shock waves or other discontinuous behavior. In particular, the numerical solution to such problems often suffers from spurious oscillations or overshoots. This usually represents unphysical behavior, and it is almost always desirable to use a numerical method that suppresses it. One of the first families of such schemes were called *total variation diminishing* (TVD); see [22, 21]. Following more recent work [4], we refer to them as *strong stability preserving* (SSP).

In particular, we are interested in the development, analysis, and optimization of SSP Runge–Kutta (SSPRK) time-stepping methods for the hyperbolic conservation law

$$u_t + f(u)_x = 0 \tag{1.1}$$

*Received by the editors December 10, 2002; accepted for publication (in revised form) October 2, 2003; published electronically July 14, 2004.
http://www.siam.org/journals/sinum/42-3/41928.html

†Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6 Canada (sruuth@sfu.ca). The work of this author was partially supported by a grant from NSERC Canada.

‡Department of Computer Science, Dalhousie University, Halifax, Nova Scotia, B3H 1W5 Canada (spiteri@cs.dal.ca). The work of this author was partially supported by a grant from NSERC Canada.

subject to appropriate initial conditions. When a SSPRK method with nonnegative coefficients is used it is convenient to consider a semidiscretization of (1.1) in space to yield a large coupled set of ODEs:

$$\dot{U} = F(U). \tag{1.2}$$

More generally, following [22, 21, 3, 4], upwind-biased ($F(U)$) and downwind-biased ($\tilde{F}(U)$) spatial discretizations may be applied in some combination to achieve favorable nonlinear stability properties for a given time-stepping scheme. For simplicity we refer to upwind-biased and downwind-biased spatial discretizations as upwind and downwind spatial discretizations, respectively.

Optimal explicit SSPRK schemes with nonnegative coefficients and where the number of stages s is equal to the order p for $s = p = 1, 2$, and 3 have been known for some time. Gottlieb and Shu [3] showed that no such method exists with nonnegative coefficients when $s = p = 4$. In [25], Spiteri and Ruuth proposed a new class of explicit SSPRK methods with nonnegative coefficients with $s > p$. They gave optimal explicit SSPRK schemes with s stages and orders 1 and 2 (see also [21, 3]), as well as specific schemes for $p = 3$, $s = 4, 5$ and $p = 4$, $s = 5$. The advantage afforded by these high-stage schemes is that the increase in the CFL coefficient allows for a large enough increase in the stable time step to more than offset the increase in computational cost per step. However, in [20] they showed that it was impossible to have an explicit SSPRK method with order greater than 4 with nonnegative coefficients. In this paper, we give a unified treatment of all explicit SSPRK schemes with positive and/or negative coefficients of up to order 5 in terms of the effective CFL coefficient. We find that many of the optimal explicit SSPRK methods under the constraint of nonnegative coefficients are also optimal in terms of the effective CFL coefficient when negative coefficients are allowed. We also present the first fifth-order explicit SSPRK methods.

We remark that explicit fifth-order SSP *multistep* schemes have been successfully constructed [21, 13, 4]. The most efficient scheme of this type that explicitly appears in the literature [21] is

$$\begin{aligned} U^{n+1} = {} & \frac{7}{20}U^n + \frac{3}{10}U^{n-1} + \frac{4}{15}U^{n-2} + \frac{7}{120}U^{n-4} + \frac{1}{40}U^{n-5} + \frac{291201}{108000}F(U^n) \\ & - \frac{198401}{86400}\tilde{F}(U^{n-1}) + \frac{88063}{43200}F(U^{n-2}) - \frac{17969}{43200}\tilde{F}(U^{n-4}) + \frac{73061}{432000}F(U^{n-5}). \end{aligned} \tag{1.3}$$

This six-step scheme involves evaluations of both upwind and downwind operators and has an effective CFL coefficient of 0.065. In this paper we construct explicit SSPRK methods with up to a 325% improvement in the effective CFL coefficient over this scheme. Comparable gains are also shown to arise in practice.

We further note that in this paper we deal with explicit Runge–Kutta methods where the number of stages s can be substantially larger than the order p. These methods are optimized with respect to the effective CFL coefficient, which is a theoretical measure of the stepsize restriction required for nonlinear stability. Although perhaps similar at first glance, this is not in general related to maximizing the area of the (linear) stability region of a Runge–Kutta method; see [3] for a counterexample. For work on the optimization of the linear stability regions of explicit Runge–Kutta methods, we refer the reader to [15] and the references therein.

The remainder of the paper is organized as follows. In section 2 we review some relevant results on SSP schemes as well as define important concepts such as the effective CFL coefficient. In sections 3 and 4 we use analytical as well as numerical techniques to find explicit SSPRK methods up to order 5 with optimal effective CFL coefficients. In section 5 we show the efficiency of the new optimized fifth-order explicit SSPRK methods versus the optimal fifth-order multistep method (1.3) and a commonly used fifth-order explicit Runge–Kutta method. Finally, in section 6 we conclude by summarizing the main findings of the paper.

2. Background on SSP schemes. In this section we give some theoretical background on SSPRK schemes. We begin by recalling the definition of strong stability.

DEFINITION 2.1. *A sequence* $\{U^n\}$ *is said to be* strongly stable *in a given seminorm* $||\cdot||$ *if* $||U^{n+1}|| \le ||U^n||$ *for all* $n \ge 0$.

Strong stability turns out to have an interesting relationship to the more classical concept of *contractivity* (see, e.g., [23, 7, 8]). In this case for equations (1.2) satisfying a one-sided Lipschitz condition, we have that the distance between all exact solutions starting from different initial conditions is nonincreasing in time. It is reasonable to then require the same property of the numerical solution; i.e., $||\tilde{U}^{n+1} - U^{n+1}|| \le ||\tilde{U}^n - U^n||$ for all $n \ge 1$. In classical stability analysis, $\tilde{U}^n$ is usually assumed to be a perturbation of U^n. It is interesting that many of the optimal SSP schemes found in [25] agree with optimal contractive schemes in [8]. In fact, recent work by Ferracina and Spijker [2] for schemes with positive coefficients shows that the stepsize coefficient C (see below) for strong stability is equivalent to the related quantity $R(A, b)$ [8] arising in contractivity studies.

To begin our analysis, assume that upwind spatial discretizations are appropriate, and consider an s-stage, explicit Runge–Kutta method written in the form

$$U^{(0)} = U^n, \tag{2.1a}$$

$$U^{(i)} = \sum_{k=0}^{i-1} (\alpha_{ik} U^{(k)} + \Delta t \beta_{ik} F(U^{(k)})), \qquad i = 1, 2, \ldots, s, \tag{2.1b}$$

$$U^{n+1} = U^{(s)}, \tag{2.1c}$$

where all the $\alpha_{ik} \ge 0$ and $\alpha_{ik} = 0$ only if $\beta_{ik} = 0$ [21].

For consistency, we must have that $\sum_{k=0}^{i-1} \alpha_{ik} = 1$, $i = 1, 2, \ldots, s$. Hence, if both sets of coefficients α_{ik}, β_{ik} are nonnegative, then (2.1) is a convex combination of forward Euler steps with various step sizes $\frac{\beta_{ik}}{\alpha_{ik}} \Delta t$. The strong stability property follows easily from this observation.

The Runge–Kutta scheme (2.1) is not written in standard Butcher array form; however, the representation (2.1) maps uniquely to a Butcher array. On the other hand, written in this form, it is particularly convenient to make use of the following result [22, 4].

THEOREM 2.2. *If the forward Euler method is strongly stable under the CFL restriction* $\Delta t \le \Delta t_{FE}$, *then the Runge–Kutta method* (2.1) *with* $\beta_{ik} \ge 0$ *is SSP, provided*

$$\Delta t \le C \Delta t_{FE},$$

where C *is the CFL coefficient*

$$C \equiv \min_{i,k} \frac{\alpha_{ik}}{\beta_{ik}}.$$

SSPRK schemes with negative coefficients β_{ik} are also possible with the appropriate interpretation. Following the procedure first suggested in [21], whenever $\beta_{ik} < 0$, the operator $\tilde{F}(\cdot)$ is used instead of $F(\cdot)$, where $\tilde{F}(\cdot)$ approximates the same derivatives as $F(\cdot)$ but is assumed to be strongly stable for Euler's method solved *backwards* in time under a suitable time-step restriction. In practice, this corresponds to a change in upwinding direction or, in other words, *downwinding*. This allows the following generalization of Theorem 2.2.

THEOREM 2.3. *Let Euler's method applied forward in time combined with the spatial discretization $F(\cdot)$ be strongly stable under the CFL restriction $\Delta t \le \Delta t_{FE}$. Let Euler's method applied backwards in time combined with the spatial discretization $\tilde{F}(\cdot)$ also be strongly stable under the same CFL restriction $\Delta t \le \Delta t_{FE}$. Then the Runge–Kutta method* (2.1) *is SSP, provided*

$$\Delta t \le C \Delta t_{FE},$$

where C is the CFL coefficient

$$C \equiv \min_{i,k} \frac{\alpha_{ik}}{|\beta_{ik}|}, \tag{2.2}$$

where $\beta_{ik} F(\cdot)$ is replaced by $\beta_{ik} \tilde{F}(\cdot)$ whenever β_{ik} is negative.

We note that the assumptions on strong stability of Euler's method applied forward and backwards in time restrict the theoretical advantages of this result to nondissipative equations such as (1.1).

Irreducible explicit Runge–Kutta methods have one (new) function evaluation per stage. We note that if every coefficient β_{ik} is positive, then the number of stages is trivially equal to the number of function evaluations. However, if both $F(U^{(k)})$ and $\tilde{F}(U^{(k)})$ are required for some k, the Runge–Kutta method (2.1) has more function evaluations[1] than stages. So the first step in creating a fair comparison of the computational cost of a given Runge–Kutta method and in deriving optimal schemes is to consider general methods that allow only one (new) function evaluation per stage. A necessary and sufficient condition for this is that the nonzero coefficients β_{ik} for a given k are all of the same sign. To see this, let $\mathcal{K}_-$ be the set of levels k such that all $\beta_{ik} \le 0$, and we consider

$$\begin{aligned} U^{(0)} &= U^n, \\ U^{(i)} &= \sum_{k=0}^{i-1} \alpha_{ik} U^{(k)} + \Delta t \begin{cases} \beta_{ik} \tilde{F}(U^{(k)})), & k \in \mathcal{K}_-, \\ \beta_{ik} F(U^{(k)})) & \text{otherwise}, \end{cases} \qquad i = 1, 2, \ldots, s-1, \\ U^{n+1} &= U^{(s)}. \end{aligned} \tag{2.3}$$

For the remainder of the paper, we will tacitly assume that the schemes under consideration are of this form. Naturally, schemes that are written combining positive and negative coefficients β_{ik} within a given level k can be augmented with additional stages to be of this form. Thus, without loss of generality, we have that the total

[1]The only difference between $\tilde{F}(\cdot)$ and $F(\cdot)$ is a change in the upwind direction; so $\tilde{F}(\cdot)$ can clearly be computed with the same cost as $F(\cdot)$ [4]. Indeed, recent studies make the assumption that if both $\tilde{F}(U^{(k)})$ and $F(U^{(k)})$ must be computed for some k, the cost, as well as the storage requirements for that k, doubles [3, 4, 25, 17]; i.e., each is given equal weight.

number of evaluations of $F(\cdot)$ and $\tilde{F}(\cdot)$ is identically equal to the number of stages of the method.

We note that this formulation allows one to search for the optimal scheme for a given order and a *given number of stages* (function evaluations). This is a more appropriate description of what should be optimized than has been considered in the literature thus far. For example, searching for the scheme with the largest CFL coefficient (or even effective CFL coefficient, see below) *for a given order* results in the number of stages tending to infinity.

Another advantage to this formulation is that schemes can be represented and implemented in Butcher array form using (3.6) since differences of the form $F(U^{(i)}) - \tilde{F}(U^{(i)})$ do not arise; i.e., the method can be implemented as

$$K_i = \begin{cases} F\left(U^n + \Delta t \sum_{j=1}^{i-1} a_{ij} K_j\right) & \text{if } b_i \geq 0, \\ \tilde{F}\left(U^n + \Delta t \sum_{j=1}^{i-1} a_{ij} K_j\right) & \text{otherwise,} \end{cases} \qquad i = 1, 2, \ldots, s,$$

$$U^{n+1} = U^n + \Delta t \sum_{i=1}^{s} b_i K_i.$$

This form is often desirable for implementing fifth-order schemes because the storage requirements can be reduced. We further remark that the differences $F(U^{(i)}) - \tilde{F}(U^{(i)})$ contribute to artificial dissipation and smearing. For example, this difference is proportional to the discrete Laplacian when first-order upwinding is applied to the linear advection equation. A natural consequence of our formulation is that during optimization these dissipative differences do not arise, leading to schemes with smaller errors and less smearing than would otherwise occur.

In section 5, we compare the computational efficiencies of various Runge–Kutta methods. In order to make a fair comparison of the relative efficiencies of these methods and to derive optimal schemes we make the following definition.

DEFINITION 2.4. *The* effective CFL coefficient C_{eff} *of an SSPRK method is* C/s*, where* C *is the CFL coefficient of the method and* s *is the number of stages (function evaluations) required for one step of the method.*

As conjectured in Shu and Osher [22] and subsequently proven in Gottlieb and Shu [3], the optimal two-stage, order-two explicit SSPRK scheme with nonnegative coefficients is the modified Euler scheme:

$$\begin{aligned} U^{(1)} &= U^n + \Delta t F(U^n), \\ U^{n+1} &= \frac{1}{2} U^n + \frac{1}{2} U^{(1)} + \frac{1}{2} \Delta t F(U^{(1)}). \end{aligned}$$

It has a CFL restriction $\Delta t \leq \Delta t_{FE}$, which implies a CFL coefficient of 1. Henceforth, we will refer to this scheme as SSP(2,2). In general, we adopt the convention of referring to the best (in terms of effective CFL coefficient) known s-stage, order-p explicit SSPRK scheme as SSP(s,p), where s is equal to the total number of function evaluations of $F(\cdot)$ and $\tilde{F}(\cdot)$. In [25] a class of s-stage, order-two explicit SSPRK schemes was given and proved to be optimal with a CFL coefficient of $s - 1$.

Shu and Osher [22] also conjectured that the optimal three-stage, order-three explicit SSPRK scheme with nonnegative coefficients is

$$\begin{aligned} U^{(1)} &= U^n + \Delta t F(U^n), \\ U^{(2)} &= \frac{3}{4}U^n + \frac{1}{4}U^{(1)} + \frac{1}{4}\Delta t F(U^{(1)}), \\ U^{n+1} &= \frac{1}{3}U^n + \frac{2}{3}U^{(2)} + \frac{2}{3}\Delta t F(U^{(2)}), \end{aligned}$$

which has a CFL coefficient of 1 as well. The optimality of this scheme was later proved by Gottlieb and Shu [3]. This scheme is commonly called the *third-order TVD Runge–Kutta scheme*, but we will refer to it as SSP(3,3).

In [20], Ruuth and Spiteri derived a linear bound that can be used to prove that the optimal four-stage, order-three explicit SSPRK scheme with nonnegative coefficients is

$$\begin{aligned} U^{(1)} &= U^n + \frac{1}{2}\Delta t F(U^n), \\ U^{(2)} &= U^{(1)} + \frac{1}{2}\Delta t F(U^{(1)}), \\ U^{(3)} &= \frac{2}{3}U^n + \frac{1}{3}U^{(2)} + \frac{1}{6}\Delta t F(U^{(2)}), \\ U^{n+1} &= U^{(3)} + \frac{1}{2}\Delta t F(U^{(3)}), \end{aligned}$$

which has a CFL coefficient of 2. This observation appears in [25]. Following [25] we will refer to this scheme as SSP(4,3).

Moving on to methods with five stages and order 3 gives a numerically optimized scheme, SSP(5,3), with a CFL coefficient of approximately 2.65. It can be proven that this is also the optimal explicit SSPRK scheme with five stages and order 3 via the following line of reasoning. The CFL coefficient C of SSP(5,3) is equal to the *radius of absolute monotonicity* $R(A, b)$ for linear constant-coefficient problems [7]. Because $C \leq R(A, b)$ [2] and C (and $R(A, b)$) for nonlinear problems cannot exceed that for linear problems, we conclude that SSP(5,3) is the optimal five-stage, third-order explicit SSPRK scheme. A similar line of reasoning can be applied to prove the optimality of SSP(3,3), SSP(4,3), as well as the first- and second-order SSP schemes. Indeed, we have produced schemes of the form SSP(s,3), $s \leq 9$, with CFL coefficients equal to $R(A, b)$ for linear constant-coefficient problems; hence they are also optimal SSP schemes for nonlinear problems. It is worth mentioning, however, that this approach does not seem to be useful for proving the optimality of schemes of order greater than 3.

The main advantage offered by these high-stage schemes is that the additional computational cost incurred per step is more than offset by the increase in stable step size. For example, SSP(4,3) costs 33% more than SSP(3,3) but offers a 100% larger CFL coefficient. Thus for SSP(4,3), $C_{\text{eff}} = 2/4 = 1/2$, whereas for SSP(3,3), $C_{\text{eff}} = 1/3$. This translates into a $(1/2 - 1/3)/1/3 = 50\%$ increase in computational efficiency.

In [3], Gottlieb and Shu proved that it is impossible to have an explicit SSPRK method of order 4 in four stages having only nonnegative coefficients.[2] In section 3.3

[2]We remark that a proof that it is also impossible to have a fourth-order explicit Runge–Kutta method with four stages and $R(A, b) > 0$ appeared earlier and independently of this [8].

we prove the stronger result that it is in fact impossible to obtain an explicit SSPRK method of order 4 with any four (general) function evaluations. In [25] a five-stage, order-four explicit SSPRK scheme with nonnegative coefficients is given. It turns out that this scheme coincides with Kraaijevanger's optimal five-stage, order-four contractive scheme [8, 2]. A further study of explicit SSPRK methods of order 4 and $s = 6, 7, 8$ stages can be found in [24]. For the examples investigated in that paper, it was found that the increased stage number did lead to noteworthy improvements in practical performance. It is also worth mentioning that these high-stage schemes have modest storage requirements.

In [20] it is shown that explicit SSPRK schemes with nonnegative coefficients do not exist with order greater than 4. A similar restriction to orders 4 or less was proven for contractive schemes [8]. This means that the search for explicit schemes of order 5 and higher must involve evaluations of the downwinded operator $\tilde{F}(\cdot)$. In the remainder of the paper we present a unified treatment of explicit SSPRK schemes that use both upwinded and downwinded operators in terms of the effective CFL coefficient and prove the optimality of several lower-order schemes.

In section 4 we give the first fifth-order explicit SSPRK methods, optimized in terms of the effective CFL coefficient. A fifth-order explicit SSPRK method in the form (2.1) was thought to have been found in [22], based on a fifth-order explicit Runge–Kutta scheme on page 143 of [9], which was in turn based on a particular choice from a family of fifth-order explicit Runge–Kutta schemes that appeared in [12]. The family of schemes in question is described by the Butcher tableau

$$
\begin{array}{c|cccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\gamma & \gamma & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{4} - \frac{1}{32}\gamma^{-1} & \frac{1}{32}\gamma^{-1} & 0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} - 32\sigma - \frac{1}{8}\frac{1-64\sigma}{\gamma} & \frac{1}{8}\frac{1-64\sigma}{\gamma} & 32\sigma & 0 & 0 & 0 \\
\frac{3}{4} & -\frac{9}{16} + 24\sigma - \frac{6\sigma-\frac{3}{16}}{\gamma} & \frac{6\sigma-\frac{3}{16}}{\gamma} & \frac{3}{4} - 24\sigma & \frac{9}{16} & 0 & 0 \\
1 & \frac{11}{7} - \frac{384}{7}\sigma - \frac{1}{7}\frac{\frac{7}{2}-96\sigma}{\gamma} & \frac{1}{7}\frac{\frac{7}{2}-96\sigma}{\gamma} & \frac{384}{7}\sigma & -\frac{12}{7} & \frac{8}{7} & 0 \\
\hline
 & \frac{7}{90} & 0 & \frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90}
\end{array}
\tag{2.4}
$$

Unfortunately, although this family of explicit Runge–Kutta schemes (2.4) is indeed fifth order, there is an error in the particular member of this family upon which the reported fifth-order explicit SSPRK method was based. This proposed scheme has coefficient matrix A given by

$$
A = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\
0 & -\frac{3}{16} & \frac{3}{8} & \frac{9}{16} & 0 & 0 \\
\frac{1}{7} & \frac{4}{7} & \frac{6}{7} & -\frac{12}{7} & \frac{8}{7} & 0
\end{bmatrix}.
$$

This scheme was meant to correspond to the particular choice of $\sigma = \frac{1}{64}$ and (arbitrary) $\gamma = \frac{1}{2}$. However, it is easily verified that this scheme does not belong to the

family of explicit fifth-order schemes (2.4), differing in the coefficients a_{31} and a_{32}. In fact it is only *second order*; e.g., it is easily seen that the third-order condition $b^T Ac = \frac{1}{6}$ is not satisfied.

3. Optimal SSPRK methods. In this section we prove some optimality results in terms of the effective CFL number for some low-order explicit SSPRK methods ($p = 1, 2, 3$). Optimal schemes for high-order methods ($p = 4, 5$) are determined numerically. We begin by describing the form of the optimization problem solved in all cases. We then prove some existence and optimality results for some explicit SSPRK methods of up to order 4. Section 4 contains some numerical results for the first explicit SSPRK methods of order 5.

3.1. Formulation of the optimization problem. We seek to optimize an s-stage, order-p explicit SSPRK scheme by maximizing its effective CFL coefficient according to Theorem 2.3. That is, we seek the global maximum of the nonlinear programming problem

$$\max_{(\alpha_{ik},\beta_{ik})} \min \frac{\alpha_{ik}}{|\beta_{ik}|}, \tag{3.1}$$

where $\alpha_{ik},\ \beta_{ik},\ k = 0, 1, \ldots, i-1,\ i = 1, 2, \ldots, s$, are real and $0 \le \alpha_{ik} \le 1$. As noted in section 2, we insist that for each k and $i = k+1, k+2, \ldots, s$ that $\beta_{ik} \ge 0$ or $\beta_{ik} \le 0$ to ensure that the number of function evaluations corresponds to the number of stages. The case $\alpha_{ik} = \beta_{ik} = 0$ is defined as `NaN` in the sense that it is not included in the minimization process if it occurs. The objective function (3.1) is also subject to the constraints

$$\sum_{k=0}^{i-1} \alpha_{ik} = 1, \qquad i = 1, 2, \ldots, s, \tag{3.2}$$

$$\sum_{j=1}^{s} b_j \Phi_j(t) = \frac{1}{\gamma(t)}, \quad t \in T_q, \quad q = 1, 2, \ldots, p. \tag{3.3}$$

Here the functions $\Phi_j(t)$ are nonlinear constraints that are polynomial in $\alpha_{ik},\ \beta_{ik}$ and that correspond to the order conditions for a Runge–Kutta method to be of order p (see, e.g., [5]); i.e., T_q stands for the set of all rooted trees of order equal to q. The number of constraints represented by the Runge–Kutta order conditions is equal to

$$\sum_{q=1}^{p} \mathrm{card}(T_q),$$

where $\mathrm{card}(T_q)$ is the cardinality of T_q. Also, we use the notation b_j in the usual sense of the Butcher array representation of a Runge–Kutta method; again this would be a polynomial function of the coefficients α_{ik} and β_{ik}. It can be expected that the particular choice of coefficients $\alpha_{ik},\ \beta_{ik}$ that maximizes the quantity (2.2) for a given Runge–Kutta method will be naturally produced by the solution to this nonlinear programming problem; hence the result will be a sharp estimate of the CFL coefficient.

However, this formulation of the nonlinear programming problem does not lend itself easily to numerical solution; see [25] for further discussion. By introducing a

dummy variable z, the nonlinear programming problem can be reformulated as

$$\max_{(\alpha_{ik},\beta_{ik})} z \tag{3.4a}$$

subject to

$$\alpha_{ik} \geq 0, \tag{3.4b}$$

$$\beta_{k+1,k}, \beta_{k+2,k}, \ldots, \beta_{sk} \geq 0, \tag{3.4c}$$

$$\text{or} \quad \beta_{k+1,k}, \beta_{k+2,k}, \ldots, \beta_{sk} \leq 0, \quad k = 0, \ldots, s-1,$$

$$\sum_{k=0}^{i-1} \alpha_{ik} = 1, \qquad i = 1, 2, \ldots, s, \tag{3.4d}$$

$$\sum_{j=1}^{s} b_j \Phi_j(t) = \frac{1}{\gamma(t)}, \quad t \in T_q, \quad q = 1, 2, \ldots, p, \tag{3.4e}$$

$$\alpha_{ik} - z|\beta_{ik}| \geq 0, \quad k = 0, 1, \ldots, i-1, \ i = 1, 2, \ldots, s. \tag{3.4f}$$

Numerical optimization software may be applied to the reformulated problem (3.4) for various combinations of s and p. In our initial approach we considered using Matlab's Optimization Toolbox but found that it was nontrivial to determine an initial guess to start the nonlinear iteration. Subsequent efforts focused on BARON [26], a deterministic, global optimization software package that uses algorithms of the branch-and-bound type. This approach was found to be superior to Matlab's Optimization Toolbox in the sense that it is faster, gives improved optima, and satisfies active constraints to 15 decimal digits.

In each of the cases $s = 7, 8, 9$ numerically optimal fifth-order SSPRK schemes were found in less than 90 minutes on a (shared) cluster of 96 dual 1.2 GHz Athlon processors with BARON. See [19] for further details on applying BARON to the optimization of SSPRK schemes.

3.2. Optimality of some low-order methods. We now give new results on optimal effective CFL coefficients for some low-order explicit SSPRK methods. Previous results primarily focus on optimizing raw CFL coefficients for methods with nonnegative coefficients. Here we give existence and optimality results in the context of effective CFL coefficients for methods with no sign restriction on their coefficients.

THEOREM 3.1. *For $s = 1, 2, 3, \ldots$, the optimal s-stage explicit SSPRK method of order 1 has effective CFL coefficient 1 and can be represented in the form of SSP$(s, 1)$; i.e.,*

$$\alpha_{ik} = \begin{cases} 1, & k = i-1, \\ 0 & \textit{otherwise}, \end{cases} \qquad \beta_{ik} = \begin{cases} \frac{1}{s}, & k = i-1, \\ 0 & \textit{otherwise}, \end{cases} \qquad i = 1, 2, \ldots, s.$$

Before giving the proof of Theorem 3.1, we introduce the following notation and give two useful lemmas. We find it convenient to write the general s-stage explicit Runge–Kutta method in the following form (cf. [3]):

$$U^{(0)} = U^n, \tag{3.5a}$$

$$U^{(i)} = U^{(0)} + \Delta t \sum_{k=0}^{i-1} \kappa_{ik} \begin{cases} F(U^{(k)}) & \text{if } \kappa_{ik} \geq 0, \\ \tilde{F}(U^{(k)}) & \text{otherwise}, \end{cases} \qquad i = 1, 2, \ldots, s, \tag{3.5b}$$

$$U^{n+1} = U^{(s)}. \tag{3.5c}$$

Using the fact that the β_{ik} at a particular level are all of the same sign, the coefficients κ_{ik} are related to the coefficients $\alpha_{ik},\ \beta_{ik}$ recursively by

$$\kappa_{ik} = \beta_{ik} + \sum_{j=k+1}^{i-1} \alpha_{ij}\kappa_{jk}. \tag{3.6}$$

We remark that the coefficients κ_{ik} can be related to the Butcher array quantities $a_{ik},\ b_k$ by

$$\begin{aligned} a_{ik} &= \kappa_{i-1,k-1}, \qquad k = 1, 2, \ldots, i-1, \quad i = 1, 2, \ldots, s-1, \\ b_k &= \kappa_{s,k-1}, \qquad k = 1, 2, \ldots, s. \end{aligned}$$

It is also important to note that $\operatorname{sgn}(\kappa_{ik}) = \operatorname{sgn}(\beta_{ik})$, motivating the use of (3.5).

LEMMA 3.2. *If a method of the form* (2.1) *with* $\alpha_{ik} \geq 0$ *has a CFL coefficient* $c \geq m > 0$, *then* $0 \leq |\kappa_{ik}| \leq \frac{1}{m}$ *for all* $k = 0, 1, \ldots, i-1,\ i = 1, 2, \ldots, s$.

Proof. From Theorem 2.3, if $c \geq m > 0$, then $\alpha_{ik} \geq m|\beta_{ik}|$ or, equivalently, $|\beta_{ik}| \leq \frac{1}{m}\,\alpha_{ik}$ for all $i,\ k$ such that $\alpha_{ik} \neq 0$.

Now

$$\alpha_{ik} \geq 0, \quad \sum_{k=0}^{i-1} \alpha_{ik} = 1, \quad i = 1, 2, \ldots, s, \quad \Rightarrow \quad \alpha_{ik} \leq 1$$

for all $i,\ k$. Hence, $|\beta_{ik}| \leq \frac{1}{m}$ for all $i,\ k$. In particular, $|\kappa_{10}| = |\beta_{10}| \leq \frac{1}{m}$ for any valid explicit SSPRK method.

We now proceed by induction on stage ℓ of an s-stage method. Assume $|\kappa_{ij}| \leq \frac{1}{m}$ for $j = 0, 1, \ldots, \ell-1;\ i = 1, 2, \ldots, \ell$. (We have just shown that this result holds for $\ell = 1$.) Now consider stage $(\ell+1)$ of a valid explicit SSPRK method; i.e., consider coefficients $\kappa_{\ell+1,k}$ for $k = 0, 1, \ldots, \ell$ with

$$\sum_{k=0}^{\ell} \alpha_{\ell+1,k} = 1.$$

Then, using (3.6),

$$\begin{aligned} |\kappa_{\ell+1,0}| &= \left| \sum_{k=1}^{\ell} \alpha_{\ell+1,k}\kappa_{k0} + \beta_{\ell+1,0} \right| \\ &\leq \sum_{k=1}^{\ell} \alpha_{\ell+1,k}|\kappa_{k0}| + |\beta_{\ell+1,0}| \\ &\leq \frac{1}{m} \sum_{k=1}^{\ell} \alpha_{\ell+1,k} + \frac{1}{m}\,\alpha_{\ell+1,0} \\ &= \frac{1}{m}\,. \end{aligned}$$

Similar arguments can be used to show $|\kappa_{\ell+1,j}| \leq \frac{1}{m}$ for $j = 1, 2, \ldots, \ell$. The lemma now follows by induction. ∎

LEMMA 3.3. *Suppose a consistent s-stage explicit SSPRK method* (2.1) *has coefficients* $\beta_{ik} \le 0$ *at* ℓ *distinct stages; i.e.,* $\beta_{ik} \le 0$ *for all* i *and* $k = k_1, k_2, \ldots, k_\ell$ *with* $0 \le k_1 < k_2 < \cdots < k_\ell \le s-1$. *Then the CFL coefficient* C *of the method satisfies* $C \le s - \ell$.

Proof. Because the method is consistent, we have

$$\sum_{k=0}^{s-1} \kappa_{sk} = 1. \tag{3.7}$$

But by the definition of κ_{ik} it is clear that $\kappa_{sk_1}, \kappa_{sk_2}, \ldots, \kappa_{sk_\ell} \le 0$. Thus

$$\sum_{\substack{k=0 \\ k \ne k_1, k_2, \ldots, k_\ell}}^{s-1} \kappa_{sk} \ge 1. \tag{3.8}$$

The desired result $C \le s - \ell$ now follows immediately from applying Lemma 3.2 to (3.8). ∎

Proof of Theorem 3.1. For nonnegative coefficients $\{\beta_{ik}\}$ the result for the raw CFL coefficient has been shown in [25]. By Lemma 3.3, a method containing any $\beta_{ik} < 0$ must have a CFL coefficient $C \le s - 1 < s$, and thus we must have $C_{\text{eff}} < 1$. This completes the proof. ∎

Remark 1. As noted in [25], despite the increase in the raw CFL coefficient, these first-order methods do not offer a theoretical computational advantage.

THEOREM 3.4. *For* $s = 2, 3, 4, \ldots$, *the optimal s-stage explicit SSPRK method of order* 2 *has effective CFL coefficient* $\frac{s-1}{s}$ *and can be represented in the form of SSP*$(s, 2)$*; i.e.,*

$$\alpha_{ik} = \begin{cases} 1, & k = i-1, \\ 0 & \text{otherwise}, \end{cases} \qquad \beta_{ik} = \begin{cases} \frac{1}{s-1}, & k = i-1, \\ 0 & \text{otherwise}, \end{cases} \qquad i = 1, 2, \ldots, s-1,$$

$$\alpha_{ik} = \begin{cases} \frac{1}{s}, & k = 0, \\ \frac{s-1}{s}, & k = s-1, \\ 0 & \text{otherwise}, \end{cases} \qquad \beta_{ik} = \begin{cases} \frac{1}{s}, & k = s-1, \\ 0 & \text{otherwise}, \end{cases} \qquad i = s.$$

Proof. For nonnegative coefficients $\{\beta_{ik}\}$ the result for the raw CFL coefficient has been shown in [25]. By Lemma 3.3, any consistent, s-stage method with some $\beta_{ik} < 0$ must have a CFL coefficient $C \le s - 1$, and thus we must have $C_{\text{eff}} \le \frac{s-1}{s}$. This completes the proof. ∎

Remark 2. As noted in [25], in this case the theoretical increase in the raw CFL coefficient more than offsets the increased work per step, leading to an overall computational advantage with increasing s. However, the effective CFL coefficient is bounded above by 1.

We now give some specific optimality results for methods of order 3.

THEOREM 3.5. *The optimal three-stage explicit SSPRK method of order* 3 *has effective CFL coefficient* $C_{\text{eff}} = 1/3$, *and an optimal representation is given by SSP*(3,3).

Proof. For nonnegative coefficients $\{\beta_{ik}\}$ the result for the raw CFL coefficient has been shown in [3].

Now suppose we allow $\beta_{ik} < 0$ in an attempt to improve the CFL coefficient. From the third-order condition $b^T Ac = 1/6$, we have $\beta_{10}\beta_{21}\beta_{32} = 1/6 > 0$; so the scheme must have $\beta_{ik} \leq 0$ at exactly two levels. But then we may apply Lemma 3.3 to show that $C \leq 1$, and hence its $C_{\text{eff}} \leq 1/3$. ∎

THEOREM 3.6. *The optimal four-stage explicit SSPRK method of order* 3 *has effective CFL coefficient* $C_{\text{eff}} = 2/3$*, and an optimal representation is given by SSP*(4,3).

Proof. For nonnegative coefficients $\{\beta_{ik}\}$ the result for the raw CFL coefficient has been shown in [25]. By Lemma 3.3 it is clear that $C \leq 2$ if the $\beta_{ik} \leq 0$ at two or more levels. So the only possibility for an improvement in the CFL coefficient over SSP(4,3) is if $\beta_{ik} \leq 0$ at precisely one level. But then by the third-order condition $b^T Ac = 1/6$, one of the following must hold:

$$\begin{aligned}
\kappa_{43}\kappa_{32}\kappa_{21} &\geq 1/6,\\
\kappa_{43}\kappa_{32}\kappa_{20} &\geq 1/6,\\
\kappa_{43}\kappa_{31}\kappa_{10} &\geq 1/6,\\
\kappa_{42}\kappa_{21}\kappa_{10} &\geq 1/6.
\end{aligned}$$

Supposing that the CFL coefficient is greater than 2 in any of these statements leads to a condition of the form $\kappa_{4i}\kappa_{ij}\kappa_{jk} \leq 1/8$, $i = 2, 3$, $1 \leq j \leq i-1$, $0 \leq k \leq j-1$, by Lemma 3.2 and gives rise to a contradiction. Hence the optimal scheme must be SSP(4,3). ∎

3.3. A fourth-order result. In this section we demonstrate that, even allowing negative coefficients β_{ik}, there is no four-stage explicit SSPRK method of order 4. We begin with a lemma.

LEMMA 3.7. *If* $s = p$*, the* β_{ik} *at a particular level* k*, for some* $0 \leq k \leq s-1$*,* $k+1 \leq i \leq s$*, are all of the same sign, i.e.,* $\beta_{ik}\beta_{jk} \geq 0$ *for* $k+1 \leq i, j \leq s$*, and the CFL coefficient is positive, then* $\kappa_{ik} \neq 0$ *for* $k+1 \leq i \leq s$*.*

Proof. From the order conditions, we have $\prod_{i=1}^{p} \beta_{i,i-1} = \frac{1}{p!}$; so each $\beta_{i,i-1} \neq 0, 1 \leq i \leq s$. Since the CFL coefficient is positive, this implies each $\alpha_{i,i-1} > 0, 1 \leq i \leq s$. Expanding κ_{ij} in terms of the α and β coefficients (see, e.g., [3]) it is easily seen that $|\kappa_{ij}| \geq |\beta_{j+1,j} \prod_{k=j+1}^{i-1} \alpha_{k+1,k}| > 0$, proving our result. ∎

We note that Lemma 3.7 is only relevant for $s = p = 1, 2, 3, 4$.[3] In this section, we will, of course, be interested specifically in the case with $s = p = 4$.

In proving the main result of this section, we will make extensive use of the following lemma, which follows immediately from Lemma 3.7 and the definition of the κ_{ij}.

LEMMA 3.8. *If* $s = p$*, the* β_{ik} *at a particular level* k*, for some* $0 \leq k \leq s-1$*,* $k+1 \leq i \leq s$*, are all of the same sign, and the CFL coefficient is positive, then* $\kappa_{ik}, k+1 \leq i \leq s$ *are also all of that same sign and are nonzero.*

We now give the main result of this section.

THEOREM 3.9. *There is no four-stage explicit SSPRK method of order* 4 *with a positive CFL coefficient.*

Proof. General case. We proceed by contradiction. If two parameters u and v are such that $u \neq v$, $u \neq 0$, $u \neq 1/2$, $u \neq 1$, $v \neq 0$, $v \neq 1$, and $6uv - 4(u+v) + 3 \neq 0$,

[3] It is possible to have schemes with $s = p > 4$ for linear, constant-coefficient problems.

then the coefficients $\kappa_{ik} \neq 0$ may be written as functions of u and v [18]:

$$\begin{aligned}
\kappa_{10} &= u, \\
\kappa_{20} &= v - \kappa_{21}, \\
\kappa_{21} &= \frac{v(v-u)}{2u(1-2u)}, \\
\kappa_{30} &= 1 - \kappa_{31} - \kappa_{32}, \\
\kappa_{31} &= \frac{(1-u)[u+v-1-(2v-1)^2]}{2u(v-u)[6uv-4(u+v)+3]}, \\
\kappa_{32} &= \frac{(1-2u)(1-u)(1-v)}{v(v-u)[6uv-4(u+v)+3]}, \\
\kappa_{40} &= \frac{1}{2} + \frac{1-2(u+v)}{12uv}, \\
\kappa_{41} &= \frac{2v-1}{12u(v-u)(1-u)}, \\
\kappa_{42} &= \frac{1-2u}{12v(v-u)(1-v)}, \\
\kappa_{43} &= \frac{1}{2} + \frac{2(u+v)-3}{12(1-u)(1-v)}.
\end{aligned}$$

Similar to [3], there are five possibilities to consider:

1. $u < 0$. If $v < 0$, then $\kappa_{10}\kappa_{40} < 0$. Conversely, if $v > 0$, then $\kappa_{10}\kappa_{20} < 0$. Both results contradict Lemma 3.8.
2. $0 < u < \frac{1}{2}$ and $v < u$.
   $\kappa_{21}\kappa_{41} > 0$ implies that $v < 0$. But this implies $\kappa_{10}\kappa_{20} < 0$, contradicting Lemma 3.8.
3. $0 < u < \frac{1}{2}$ and $v > u$.
   $\kappa_{21}\kappa_{41} > 0$ requires $v > \frac{1}{2}$. $\kappa_{20} > 0$ requires $v < 3u - 4u^2 \leq \frac{9}{16}$. $\kappa_{32}\kappa_{42} > 0$ and $\kappa_{31}\kappa_{41} > 0$ require $u > 2 - 5v + 4v^2$. Since this is a decreasing function of v for $v \leq \frac{9}{16}$, we obtain $u > 2 - 5(3u - 4u^2) + 4(3u - 4u^2)^2$. Rearranging, we find that $0 > 2((2u-1)^2 + 4u^2)(2u-1)^2$, which is impossible.
4. $u > \frac{1}{2}$ and $v < u$. We can only have $\kappa_{32}\kappa_{42} > 0$ in one of two ways:
   (a) $1 - u > 0$ and $6uv - 4(u+v) + 3 > 0$.
   $\kappa_{21}\kappa_{41} > 0$ requires $0 < v < \frac{1}{2}$. Simple calculation yields

$$\kappa_{30} = \frac{(2 - 6u + 4u^2) + (-5 + 15u - 12u^2)v + (4 - 12u + 12u^2)v^2}{2uv(6uv - 4(u+v) + 3)};$$

   hence $\kappa_{30} > 0$ requires

$$\begin{aligned}
A + Bv + Cv^2 \equiv (2 - 6u + 4u^2) \\
&+ (-5 + 15u - 12u^2)v \\
&+ (4 - 12u + 12u^2)v^2 > 0.
\end{aligned}$$

   It is easy to show that when $\frac{1}{2} < u < 1$ we have $A < 0, B < 0$, and

$C > 0$. Thus for $0 < v < \frac{1}{2}$ we have

$$A + Bv + Cv^2 < \max\left(A, A + \frac{1}{2}B + \frac{1}{4}C\right)$$
$$= \max\left(A, \frac{1}{2}(1-2u)(1-u)\right) < 0,$$

resulting in a contradiction.

(b) $1 - u < 0$ and $6uv - 4(u+v) + 3 < 0$.
Suppose $v < 0$. Then $\kappa_{10}\kappa_{20} > 0$ implies $v < -u(4u-3) < -u$, and $\kappa_{21}\kappa_{31} > 0$ implies $u + v - 1 - (2v-1)^2 > 0$. Together these yield a contradiction.
Now suppose $v > 0$. $\kappa_{21}\kappa_{41} > 0$ implies $v > \frac{1}{2}$. $\kappa_{31}\kappa_{41} > 0$ requires $u + v - 1 - (2v-1)^2 < 0$, which implies

$$(1-4v)(1-v) = 4v^2 - 5v + 1 > u - 1 > 0.$$

Given the restrictions on v, this is true only if $v < \frac{1}{4}$, contradicting the requirement that $v > \frac{1}{2}$.

5. $u > \frac{1}{2}$ and $v > u$. $\kappa_{21}\kappa_{41} > 0$ requires $u > 1$. This implies $\kappa_{42} > 0$; so by Lemma 3.8, $\kappa_{32} > 0$. It is now easily seen that $\kappa_{10} > 0$, $\kappa_{21} < 0$, and $\kappa_{43} > 0$. Thus $\kappa_{10}\kappa_{21}\kappa_{32}\kappa_{43} < 0$, contradicting the fourth-order condition $\kappa_{10}\kappa_{21}\kappa_{32}\kappa_{43} = \frac{1}{4!}$.

If $6uv - 4(u+v) + 3 = 0$, $u = 0$, or $v = 0$, then this method is not fourth order [18].

Special cases. There remain three special cases [18, 5].

1. $u = \frac{1}{2},\ v = 0$; $\kappa_{42} = w \neq 0,\ \kappa_{40} = \frac{1}{6} - w,\ \kappa_{41} = \frac{2}{3}$, and $\kappa_{43} = \frac{1}{6}$.
2. $u = v = \frac{1}{2}$; $\kappa_{40} = \frac{1}{6},\ \kappa_{42} = w \neq 0,\ \kappa_{41} = \frac{2}{3} - w$, and $\kappa_{43} = \frac{1}{6}$.
3. $u = 1,\ v = \frac{1}{2}$; $\kappa_{43} = w \neq 0,\ \kappa_{41} = \frac{1}{6} - w,\ \kappa_{40} = \frac{1}{6}$, and $\kappa_{42} = \frac{2}{3}$.

In these cases, κ_{32} is obtained from

$$\kappa_{43}\kappa_{32} = \kappa_{42}(1-v).$$

The remaining coefficients κ_{21} and κ_{31} are then the solutions to the (nonsingular) linear system

$$\kappa_{42}\kappa_{21}uv + \kappa_{43}(\kappa_{31}u + \kappa_{32}v) = \frac{1}{8},$$
$$\kappa_{42}\kappa_{21} + \kappa_{43}\kappa_{31} = \kappa_{41}(1-u).$$

It is easily verified that in each case the κ_{ik} fail to have the same sign at each level whenever negative β_{ik} are considered. ∎

4. Fifth-order explicit SSPRK methods. In this section, we give the results of the numerical optimization procedure outlined in section 3. Examples of optimal explicit SSPRK methods of order 4 and up to eight stages with positive coefficients appear in [24]. We have also constructed optimal explicit SSPRK methods of order 3 and up to nine stages. Here we design the first optimized fifth-order explicit SSPRK methods. No formal proofs of optimality are given; however, the methods described here are the results of extensive numerical testing. We now give the coefficients of the Butcher tableaus for SSP(7,5), SSP(8,5), and SSP(9,5) in Tables 1–3, respectively. The Butcher tableau format is provided because this is the more advantageous format for implementation.

TABLE 1
Butcher tableau entries for SSP(7,5). *The CFL coefficient is* 1.178508348471858.

Entry	Value
$a(2,1)$	0.392382208054010
$a(3,1)$	0.310348765296963
$a(3,2)$	0.523846724909595
$a(4,1)$	0.114817342432177
$a(4,2)$	0.248293597111781
$a(4,3)$	0
$a(5,1)$	0.136041285050893
$a(5,2)$	0.163250087363657
$a(5,3)$	0
$a(5,4)$	0.557898557725281
$a(6,1)$	0.135252145083336
$a(6,2)$	0.207274083097540
$a(6,3)$	−0.180995372278096
$a(6,4)$	0.326486467604174
$a(6,5)$	0.348595427190109
$a(7,1)$	0.082675687408986
$a(7,2)$	0.146472328858960
$a(7,3)$	−0.160507707995237
$a(7,4)$	0.161924299217425
$a(7,5)$	0.028864227879979
$a(7,6)$	0.070259587451358
$b(1)$	0.110184169931401
$b(2)$	0.122082833871843
$b(3)$	−0.117309105328437
$b(4)$	0.169714358772186
$b(5)$	0.143346980044187
$b(6)$	0.348926696469455
$b(7)$	0.223054066239366

5. Numerical studies. In this section, we study the numerical behavior of our fifth-order schemes and the optimal known fifth-order SSP multistep method (1.3) for a few test problems designed to capture solution features that pose particular difficulties to numerical methods. Our focus here is to illustrate the stability behavior of various fifth-order schemes rather than to provide a detailed accuracy study. If a study of the relative error constants was desired it would be more appropriate to consider systems where the spatial discretization errors are dominated by the time-stepping error. Experiments using the standard implementation of Fehlberg's fifth-order explicit Runge–Kutta method [5] are also included because this method is commonly used in method-of-lines discretizations of hyperbolic conservation laws. We remark that Fehlberg's scheme does not have a positive CFL coefficient in its standard implementation (using only $F(\cdot)$) because SSP methods of order greater than 4 require evaluations of $\tilde{F}(\cdot)$ [20].

We remark that tests using the popular Dormand–Prince scheme [5] gave results very similar to Fehlberg's scheme. For clarity, we do not include these simulations in our plotted results.

5.1. Test problems. There are a variety of solution features in computational fluid dynamics that commonly cause numerical problems. For example, many numerical methods produce significant errors near sonic points (points where the wavespeed equals zero). Upwind methods, in particular, are forced to give sonic points special consideration since the upwind direction changes at sonic points. Shock waves, contact discontinuities, and expansion fans may also lead to a variety of serious problems,

TABLE 2
Butcher tableau entries for SSP(8,5). *The CFL coefficient is* 1.875684961641323.

Entry	Value
$a(2,1)$	0.276409720937984
$a(3,1)$	0.149896412080489
$a(3,2)$	0.289119929124728
$a(4,1)$	0.057048148321026
$a(4,2)$	0.110034365535150
$a(4,3)$	0.202903911101136
$a(5,1)$	0.169059298369086
$a(5,2)$	0.326081269617717
$a(5,3)$	0.450795162456598
$a(5,4)$	0
$a(6,1)$	0.061792381825461
$a(6,2)$	0.119185034557281
$a(6,3)$	0.199236908877949
$a(6,4)$	0.521072746262762
$a(6,5)$	−0.001094028365068
$a(7,1)$	0.111048724765050
$a(7,2)$	0.214190579933444
$a(7,3)$	0.116299126401843
$a(7,4)$	0.223170535417453
$a(7,5)$	−0.037093067908355
$a(7,6)$	0.228338214162494
$a(8,1)$	0.071096701602448
$a(8,2)$	0.137131189752988
$a(8,3)$	0.154859800527808
$a(8,4)$	0.043090968302309
$a(8,5)$	−0.163751550364691
$a(8,6)$	0.044088771531945
$a(8,7)$	0.102941265156393
$b(1)$	0.107263534301213
$b(2)$	0.148908166410810
$b(3)$	0.105268730914375
$b(4)$	0.124847526215373
$b(5)$	−0.068303238298102
$b(6)$	0.127738462988848
$b(7)$	0.298251879839231
$b(8)$	0.156024937628252

including oscillations, overshoots, and smearing that can spread discontinuities over several cells. In particular, contact discontinuities do not have any physical compression, and thus smearing increases progressively with the number of time steps. Even when approximating smooth solutions, most numerical methods exhibit obvious flaws. For example, many stable numerical methods continuously erode the solution, leading to amplitude and dissipation errors [11].

To investigate the behavior of our time-stepping schemes, we consider three of Laney's five test problems [11]. These three problems involve all of the important flow features identified above: shocks, contacts, expansion fans, sonic points, and smooth solutions. Similar to Laney, we focus on the behavior of the numerical scheme for interior regions rather than boundaries and impose periodic boundary conditions on the domain $[-1,1]$. It is known that sometimes a conventional (and intuitive) treatment of the boundary data (especially in the case of inflow boundary conditions) within the stages of a Runge–Kutta method can lead to a deterioration in the overall accuracy of the integration. We refer the reader to [1] and the references therein for a discussion of this problem and a method for its resolution. The spatial discretization

TABLE 3
Butcher tableau entries for SSP(9,5). *The CFL coefficient is* 2.695788289294857.

Entry	Value
$a(2,1)$	0.234806766829933
$a(3,1)$	0.110753442788106
$a(3,2)$	0.174968893063956
$a(4,1)$	0.050146926953296
$a(4,2)$	0.079222388746543
$a(4,3)$	0.167958236726863
$a(5,1)$	0.143763164125647
$a(5,2)$	0.227117830897242
$a(5,3)$	0.240798769812556
$a(5,4)$	0
$a(6,1)$	0.045536733856107
$a(6,2)$	0.071939180543530
$a(6,3)$	0.143881583463234
$a(6,4)$	0.298694357327376
$a(6,5)$	−0.013308014505658
$a(7,1)$	0.058996301344129
$a(7,2)$	0.093202678681501
$a(7,3)$	0.109350748582257
$a(7,4)$	0.227009258480886
$a(7,5)$	−0.010114159945349
$a(7,6)$	0.281923169534861
$a(8,1)$	0.114111232336224
$a(8,2)$	0.180273547308430
$a(8,3)$	0.132484700103381
$a(8,4)$	0.107410821979346
$a(8,5)$	−0.129172321959971
$a(8,6)$	0.133393675559324
$a(8,7)$	0.175516798122502
$a(9,1)$	0.096188287148324
$a(9,2)$	0.151958780732981
$a(9,3)$	0.111675915818310
$a(9,4)$	0.090540280530361
$a(9,5)$	−0.108883798219725
$a(9,6)$	0.112442122530629
$a(9,7)$	0.147949153045843
$a(9,8)$	0.312685695043563

Entry	Value
$b(1)$	0.088934582057735
$b(2)$	0.102812792947845
$b(3)$	0.111137942621198
$b(4)$	0.158704526123705
$b(5)$	−0.060510182639384
$b(6)$	0.197095410661808
$b(7)$	0.071489672566698
$b(8)$	0.151091084299943
$b(9)$	0.179244171360452

and the results of three test cases follow.

5.2. Spatial discretization. Similar to [22, 25], we choose finite-difference Shu–Osher methods (ENO) to spatially discretize the equations. These methods are derived using flux reconstruction and have a variety of desirable properties. For example, they naturally extend to an arbitrary order of accuracy in space, and they are independent of the time discretization, thus allowing experimentation with different time discretization methods. Moreover, educational codes are also freely available [11, 10], an attribute which is desirable for standardizing numerical studies. Since we are focusing on fifth-order Runge–Kutta methods we carry out our simulations using a fifth-order spatial discretization. We further note that flux splitting is carried out according to

$$f^+(U) = \frac{1}{2}(f(U) + \alpha^n_{i+1/2}U),$$
$$f^-(U) = \frac{1}{2}(f(U) - \alpha^n_{i+1/2}U),$$

where $\alpha^n_{i+1/2} = \max\{|f'(U^n_{i+1})|, |f'(U^n_i)|\}$. To evaluate $\tilde{F}(\cdot)$ we simply negate the discretization that arises when we apply the Shu–Osher finite difference method to the PDE evolved *backwards in time*[4] (see [21] for further details on the procedure). For further details on the underlying discretization, as well as code for the spatial discretization, see [11, 10].

It is noteworthy that high-order, fully TVD spatial discretization schemes are also available; see Osher and Chakravarthy [16]. In these numerical studies, we choose Shu–Osher spatial discretization schemes rather than TVD schemes because TVD schemes only obtain between first- and second-order accuracy at extrema and they have "been largely superseded by Shu and Osher's class of high-order ENO methods" [11].

It is also noteworthy that recent variations on Shu–Osher methods such as methods based on WENO reconstructions (e.g., [14, 6]) also naturally combine with SSPRK schemes. See [11] for detailed discussions on these and other spatial discretizations appropriate for hyperbolic conservation laws.

5.3. Test case 1: Linear advection of a sinusoid. In this test case, the smooth initial conditions

$$u(x, 0) = -\sin(\pi x)$$

are evolved to time $t = 30$ according to the linear advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

using a constant grid spacing of $\Delta x = 1/80$. Because this evolution causes the initial conditions to travel around the periodic domain $[-1, 1]$ exactly 15 times, it is clear that the exact solution is just $u(x, 30) = -\sin(\pi x)$. Test case 1 effectively illustrates the evolution of a smooth solution with no sonic points and is useful for verifying convergence rates for high-order schemes. Moreover, even on completely smooth solutions most numerical methods designed for hyperbolic conservation laws exhibit obvious flaws [11]. This test case is quite helpful for understanding phase and amplitude errors but should not be used to study dispersion because only one frequency is present in the exact solution. It is also informative to contrast these results with those derived for problems involving shocks and other discontinuities.

To quantify the accuracy of the computed solution, we use the logarithm of the l_1 errors, i.e.,

$$\log_{10}\left(\frac{1}{N}\sum_{i=1}^{N}|U_i - u(x_i, 30)|\right),$$

[4] To illustrate the procedure, consider a first-order spatial discretization of $\tilde{F}$ for Burgers's equation. To proceed we need to construct an upwind spatial discretization for Burgers's equation evolved backwards in time, i.e.,

$$u_t = uu_x.$$

Carrying out a first-order upwind discretization with a uniform discretization step size h gives $-\tilde{F}$:

$$-\tilde{F} = \begin{cases} U_j(U_{j+1} - U_j)/h & \text{if } U_j > 0, \\ U_j(U_j - U_{j-1})/h & \text{otherwise,} \end{cases}$$

from which $\tilde{F}$ is trivially obtained.

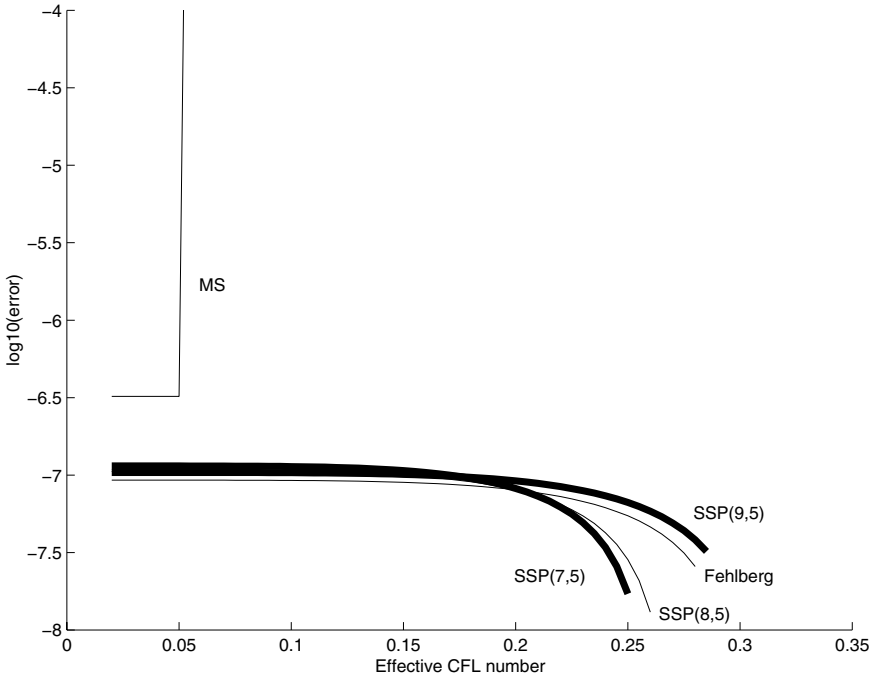


FIG. 1. l_1 *errors as a function of the effective CFL number for test case* 1.

where N is the number of grid points and x_i is the ith grid node. A plot of the error is given in Figure 1. To ensure a fair comparison for methods with a different number of stages, the error is plotted as a function of the effective CFL number[5] rather than the CFL number itself. This implies that for a particular plot the total number of function evaluations at a particular abscissa value will be the same for each scheme. We start calculating errors for an effective CFL number of 0.02 and continue until the numerical method is so unstable that a value of `NaN` is returned; i.e., the scheme has become completely unstable.

In this test example, the main conclusion is that Fehlberg's scheme and our new fifth-order explicit SSPRK schemes all outperform the multistep scheme (1.3) by more than 350%, with SSP(9,5) giving more than a 400% improvement. It is not surprising that Fehlberg's scheme performs well on this *smooth* problem because schemes based purely on a linear stability analysis are expected to perform well. SSP schemes are designed to outperform on problems involving discontinuities in the solution or its derivatives; so in this case there is no reason to expect that schemes derived using a nonlinear stability analysis will necessarily outperform classical schemes based on a linear stability analysis.

5.4. Test case 2: Linear advection of a square wave. In this test case, the discontinuous initial conditions

$$u(x,0) = \begin{cases} 1 & \text{for } |x| < 1/3, \\ 0 & \text{for } 1/3 < |x| \le 1 \end{cases}$$

[5]Similar to the definition of effective CFL coefficient, the *effective CFL number* of an SSPRK method is $\left(\frac{1}{s}\right)\frac{\Delta t}{\Delta x}$, where s is the number of stages required for one step of the method.

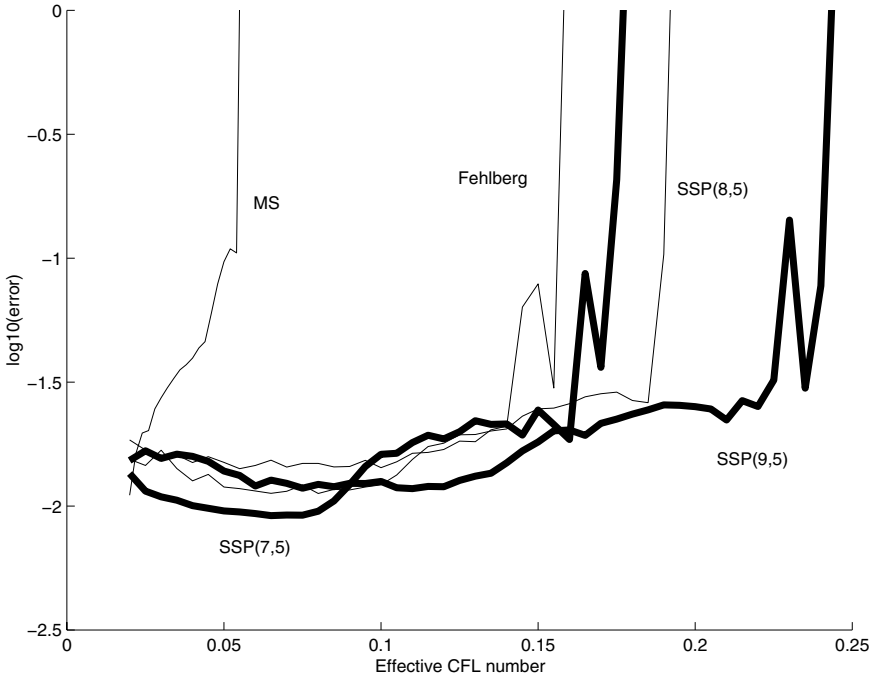


FIG. 2. l_1 *errors as a function of the effective CFL number for test case* 2.

are evolved to time $t = 4$ according to the linear advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

using a constant grid spacing of $\Delta x = 1/320$. Because this evolution causes the initial conditions to travel around the periodic domain $[-1, 1]$ exactly two times, it is clear that the exact solution at the final time is just $u(x, 4) = u(x, 0)$. Test case 2 exhibits two jump discontinuities in the solution that correspond to contact discontinuities. This test case nicely illustrates progressive contact smearing and dispersion.

The log of the l_1 errors as a function of the effective CFL number are plotted in Figure 2. Based on these plots, it is immediately clear that a material improvement in stability is obtained using our new fifth-order SSPRK schemes. Indeed, our schemes all outperform the multistep scheme (1.3) by 200% or more, with SSP(9,5) giving a 340% improvement. We also find that our schemes significantly outperform Fehlberg's scheme on this nonsmooth test. In particular, SSP(9,5) gives a 40% improvement over Fehlberg's scheme.

5.5. Test case 3: Evolution of a square wave by Burgers's equation. In this test case, the discontinuous initial conditions

$$u(x, 0) = \begin{cases} 1 & \text{for } |x| < 1/3, \\ -1 & \text{for } 1/3 < |x| \le 1 \end{cases}$$

are evolved to time $t = 0.3$ according to Burgers's equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}u^2\right) = 0$$

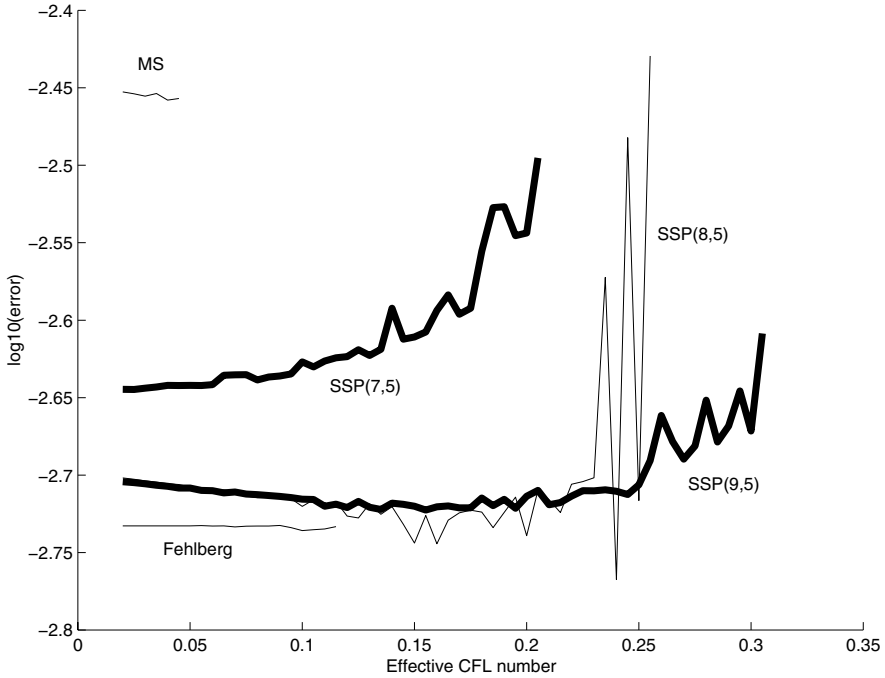


FIG. 3. l_1 *errors as a function of the effective CFL number for test case* 3.

using a constant grid spacing of $\Delta x = 1/320$. In this example, the jump at $x = -1/3$ creates a simple centered expansion fan, and the jump at $x = 1/3$ creates a steady shock. Until the shock and expansion fan intersect (at time $t = 2/3$), the exact solution is

$$u(x,t) = \begin{cases} -1 & \text{for } -\infty < x < b_1, \\ -1 + 2\frac{x-b_1}{b_2-b_1} & \text{for } b_1 < x < b_2, \\ 1 & \text{for } b_2 < x < b_{shock}, \\ -1 & \text{for } b_{shock} < x < \infty, \end{cases}$$

where $b_1 = -1/3 - t$, $b_2 = -1/3 + t$, and $b_{shock} = 1/3$ [11]. Test case 3 is particularly interesting because it illustrates the behaviors near sonic points ($u = 0$) that correspond to an expansion fan and a compressive shock.

The log of the l_1 errors as a function of the effective CFL number are plotted in Figure 3. In this nonlinear test case, we find a dramatic improvement for our new schemes over the multistep scheme (1.3). They all give more than a 350% improvement, with SSP(9,5) giving more than a 575% improvement. We also find that our schemes significantly outperform Fehlberg's scheme on this nonsmooth test. The SSP(9,5) scheme, in particular, gives more than a 150% improvement over Fehlberg's scheme.

6. Conclusions. We have studied high-order SSP explicit Runge–Kutta methods with downwind-biased spatial discretizations. We find that by requiring that the nonzero coefficients β_{ik} for a given k are all of the same sign we obtain a more appropriate description of what should be optimized. This leads to more efficient schemes with less smearing. When the order of the explicit Runge–Kutta method is less than or equal to 4 we prove in a variety of cases that there is no advantage in terms of the

effective CFL coefficient to using downwind-biased spatial discretizations. To achieve explicit SSPRK methods with fifth- or higher-order accuracy, however, downwind-biased discretizations are necessary. This paper provides the first examples of such schemes. We find that these new schemes are much more efficient than existing fifth-order explicit SSP multistep methods (both theoretically and in practice) and handily outperform classical explicit fifth-order schemes on nonsmooth problems. In particular, we found that in our marginally resolved test cases (involving shocks and contact discontinuities) larger time steps and improved efficiency were found as the effective CFL coefficient (and the number of stages) increased. In a well-resolved problem (test case 1), however, the practical performance of SSPRK schemes and classical Runge–Kutta schemes was very similar. This suggests that high-order SSPRK schemes with large effective CFL coefficients have the potential to provide high-order accuracy in smooth regions of the flow while still yielding large stable steps in marginally resolved regions. It is our hope that by providing numerically optimal schemes of this type we will stimulate further numerical studies and comparisons of SSPRK schemes against more classical approaches.

Acknowledgments. The authors thank J. Rusak for his help with the unconstrained global optimization. We also thank S. Gottlieb for interesting discussions on downwind-biased spatial discretizations.

REFERENCES

[1] S. Abarbanel, D. Gottlieb, and M. H. Carpenter, *On the removal of boundary errors caused by Runge–Kutta integration of nonlinear partial differential equations*, SIAM J. Sci. Comput., 17 (1996), pp. 777–782.

[2] L. Ferracina and M. N. Spijker, *Stepsize restrictions for the total-variation-diminishing property in general Runge–Kutta methods*, SIAM J. Numer. Anal., to appear.

[3] S. Gottlieb and C. W. Shu, *Total variation diminishing Runge-Kutta schemes*, Math. Comp., 67 (1998), pp. 73–85.

[4] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[5] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I, Springer-Verlag, Berlin, 1987.

[6] G.-S. Jiang and C. W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.

[7] J. F. B. M. Kraaijevanger, *Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems*, Numer. Math., 48 (1986), pp. 303–322.

[8] J. F. B. M. Kraaijevanger, *Contractivity of Runge-Kutta methods*, BIT, 31 (1991), pp. 482–528.

[9] J. D. Lambert, *Computational Methods in Ordinary Differential Equations*, John Wiley and Sons, London, New York, Sydney, 1973.

[10] C. Laney, *CFD Recipes: Software for Computational Gasdynamics*, http://capella.colorado.edu/∼laney/booksoft.htm (2000).

[11] C. Laney, *Computational Gasdynamics*, Cambridge University Press, Cambridge, UK, 1998.

[12] J. D. Lawson, *An order five Runge-Kutta process with extended region of stability*, SIAM J. Numer. Anal., 3 (1966), pp. 593–597.

[13] H. W. J. Lenferink, *Contractivity preserving explicit linear multistep methods*, Numer. Math., 55 (1989), pp. 213–223.

[14] X.-D. Liu, S. Osher, and T. Chan, *Weighted essentially nonoscillatory schemes*, J. Comput. Phys., 115 (1994), pp. 200–212.

[15] A. A. Medovikov, *High order explicit methods for parabolic equations*, BIT, 38 (1998), pp. 372–390.

[16] S. Osher and S. Chakravarthy, *Very high order accurate TVD schemes*, in Oscillation Theory, Computation, and Methods of Compensated Compactness, IMA Vol. Math. Appl. 2, C. Dafermos, J. Erikson, D. Kinderlehrer, and M. Slemrod, eds., Springer-Verlag, New York, 1986, pp. 229–274.

[17] S. J. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer-Verlag, New York, 2002.
[18] A. Ralston, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1965.
[19] S. J. Ruuth, *Global Optimization of Strong-Stability-Preserving Runge-Kutta Schemes*, manuscript, 2003.
[20] S. Ruuth and R. Spiteri, *Two barriers on strong-stability-preserving time discretization methods*, J. Sci. Comput., 17 (2002), pp. 211–220.
[21] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 1073–1084.
[22] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[23] M. N. Spijker, *Contractivity in the numerical solution of initial value problems*, Numer. Math., 42 (1983), pp. 271–290.
[24] R. J. Spiteri and S. J. Ruuth, *Nonlinear evolution using optimal fourth-order strong-stability-preserving Runge-Kutta methods*, Math. Comput. Simulation, 62 (2003), pp. 125–135.
[25] R. J. Spiteri and S. J. Ruuth, *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal., 40 (2002), pp. 469–491.
[26] M. Tawarmalani and N. V. Sahinidis, *Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming: Theory, Algorithms, Software, and Applications*, Nonconvex Optim. Appl. 65, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 997–1017

CONTINUOUS M-ESTIMATORS AND THEIR INTERPOLATION BY POLYNOMIALS*

JONG-SHI PANG† AND THOMAS P.-Y. YU‡

Abstract. Replacing the *median* by a general *M-estimator*, we construct in this paper a host of variants of the robust nonlinear pyramid transforms proposed by Donoho and Yu [*SIAM J. Math. Anal.*, 31 (2000) pp. 1030–1061]. Some of these new variants are more amenable to numerical implementations with provable properties when compared to the Donoho–Yu median-based pyramid transforms. At the crux of this generalized construction is the following result: the inverse problem of interpolating a univariate polynomial of degree n with $n+1$ prescribed values for any given continuous M-estimator on $n+1$ nonoverlapping intervals is a well-posed procedure. While the proof of this result is nonconstructive, we study the use of Newton methods for constructing such a polynomial interpolant and report numerical results in some test cases.

Key words. median, M-estimator, polynomial interpolation, Newton's method, convex analysis, nonlinear pyramid transforms, wavelets

AMS subject classifications. 58C15, 65D05, 65D10, 65T60, 34A34, 37L65

DOI. 10.1137/S0036142902412221

1. Introduction.

Donoho and Yu introduced in [3] a family of nonlinear pyramid transforms of signals which is reminiscent of linear biorthogonal wavelet transforms but has the advantage of being robust against non-Gaussian noise. Underlying their construction are the concept of measuring the median values of a signal over intervals at multiple scales and the accurate prediction of fine-scale median values from coarse-scale ones based on polynomial interpolation. The latter polynomial interpolation procedure, shown to be well posed by Goodman and Yu using a nonconstructive argument [6], involves the solution of a *nonsmooth* system of equations for which no existing equation solvers are shown to be convergent.

Instead of directly attacking the computationally challenging median interpolation problem, we introduce in this paper a broad class of "continuous M-estimators" that include the continuous median as a special case. These M-estimators naturally lead to new classes of nonlinear pyramid transforms that extend the original family introduced by Donoho and Yu, which is based on the continuous median. Extending the result of Goodman and Yu, we demonstrate that the interpolation problem of the continuous M-estimators by polynomials is also well posed. With the use of a "smooth kernel," the latter interpolation problem is equivalent to solving a system of differentiable equations, which can be accomplished by, say, a stabilized Newton method that can be demonstrated to be globally and quadratically convergent. In turn, this implies that the pyramid transforms can be implemented using a broad class of robust statistics with the aid of a provably convergent Newton method.

*Received by the editors July 25, 2002; accepted for publication (in revised form) December 22, 2003; published electronically July 14, 2004. This research was initiated during the second author's visit to The Johns Hopkins University in March 2002.

http://www.siam.org/journals/sinum/42-3/41222.html

†Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 and Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218-2682 (pangj@rpi.edu). The research of this author was partially supported by the National Science Foundation under grant CCR-0098013.

‡Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 (yut@rpi.edu). The research of this author was partially supported by a NSF CAREER Award (CCR 9984501).

In summary, the main contribution of our work is twofold: first, based on the class of continuous M-estimators, we have significantly broadened the original Donoho–Yu family of nonlinear pyramid transforms; more importantly, a large class of such transforms can be implemented by solving *smooth* systems of nonlinear equations by provably convergent numerical methods.

2. Continuous M-estimators. We begin with a brief review of the median value of a continuous function $f : \mathbb{R} \to \mathbb{R}$ on an interval I and the associated interpolation problem. Specifically, the latter value is defined as

$$\operatorname{med}(f|I) := \arg\min_{m\in\mathbb{R}} \int_I |\, f(t) - m \,|\, dt. \tag{2.1}$$

A key step in the Donoho–Yu proposal for accurately predicting fine-scale medians from coarse-scale ones [3] (see section 5 for more details) requires the inversion of the map

$$M : \Pi_n \to \mathbb{R}^{n+1}, \quad p \mapsto (\operatorname{med}(p|I_i))_{i=0}^n, \tag{2.2}$$

where I_i, $i = 0, \ldots, n$, are nonoverlapping intervals and Π_n is the $(n+1)$-dimensional vector space of polynomials of degree not exceeding n. While this nonlinear map is known to be a homeomorphism [6], implying that M^{-1} exists, computing $M^{-1}(a)$ for a given vector a is nevertheless not an easy task. In general, we have to resort to numerical methods. (In the case of $n = 2$, closed-form formulas for M^{-1} can be found in [3, section 2].) Newton's method applied to the system of nonlinear equations $M(p) = a$ is a prime candidate for this task. Nevertheless, since the map M is not everywhere differentiable, a classical Newton method for smooth systems (such as the ones in [2, 11]) is therefore not applicable. Although there exist provably convergent Newton methods for "semismooth" systems (see [5, Chapters 7 and 8] and the references therein), it is an open problem at this time whether M in (2.2) is semismooth in general.

Partly to broaden the continuous median and partly to alleviate the computational difficulty with inverting the nondifferentiable map M in (2.2), this paper introduces the class of continuous M-estimators defined with respect to triples (K, f, ρ) satisfying the following blanket specifications: (a) K is a solid, compact, connected set whose boundary, denoted ∂K, has measure zero; (b) f is a real-valued continuous function defined on K; and (c) $\rho : \mathbb{R} \to \mathbb{R}_+$ is a convex function with a unique minimizer at zero, and $\rho(0) = 0$. The solidness of K implies that its interior, denoted int K, is nonempty; hence K has positive measure. The convexity of ρ implies its continuity.

Corresponding to such a triple (K, f, ρ), we define the minimand $\theta(\cdot; K, f, \rho) : \mathbb{R} \to \mathbb{R}_+$ and the *continuous M-estimator* $m(f; K, \rho)$ as follows:

$$\begin{aligned} \theta(m; K, f, \rho) &:= \int_K \rho(f(x) - m)\, dx, \quad m \in \mathbb{R}, \\ m(f; K, \rho) &:= \arg\min_{m\in\mathbb{R}} \theta(m; K, f, \rho). \end{aligned}$$

When the pair (K, ρ) is clear from the context, we write $m(f)$ for $m(f; K, \rho)$. Implicit in the above definition of $m(f; K, \rho)$ is the assertion that this is a well-defined quantity (i.e., it exists and is unique); this will be justified in Theorem 2.2. Throughout the paper, we let $C(K)$ denote the set of continuous real-valued functions defined on K.

Before establishing properties of the continuous M-estimator, we give several choices of the convex function ρ. Among these choices all are continuously differentiable except the absolute value function and the third one, which are both piecewise linear. The nomenclature "continuous M-estimator" is coined as a generalization of the well-known robust Huber M-estimator [7], which is recovered from the fourth function.

1. $\rho(t) = |t|^p$ for $p \geq 1$. For $p = 1$, the resulting $m(f; K, \rho)$ is the continuous median of f on K:

$$\operatorname{med}(f|K) = \arg\min_m \int_K |f(x) - m| \, dx,$$

generalizing the case of an interval K. For $p = 2$, the resulting $m(f; K, \rho)$ is the average of f on K:

$$\operatorname{ave}(f|K) = \frac{1}{\operatorname{meas} K} \int_K f(x) \, dx,$$

where "meas" is the abbreviation of "measure."

2. $\rho_\varepsilon(t) = \sqrt{t^2 + \varepsilon^2} - \varepsilon$ for some $\varepsilon > 0$; this function ρ_ε is a strictly convex, C^∞ approximation to the absolute-value function $|\cdot|$. As we will prove in Proposition 4.4, $m(f; K, \rho_\varepsilon)$ converges to $\operatorname{med}(f|K)$ as $\varepsilon \downarrow 0$. Consequently, for $\varepsilon > 0$ sufficiently small, we expect $m(f; K, \rho_\varepsilon)$ to be as robust against outliers as the standard median. (A rigorous quantification of this claim is in the scope of robust statistics, which we will not get into in this paper.)

3. $\rho(t) = \alpha \max(t, 0) + (1 - \alpha) \max(-t, 0)$ for $\alpha \in (0, 1)$. The resulting $m(f; K, \rho)$ measures the continuous α-quantile of the function f over K, i.e., the unique value m, such that

$$\operatorname{meas}\{t \in K : f(t) \geq m\} \leq \alpha \operatorname{meas} K \quad \text{and} \quad \operatorname{meas}\{t \in K : f(t) \leq m\} \leq (1-\alpha)\operatorname{meas} K.$$

4. For a given $c > 0$,

$$\rho(t) = \begin{cases} \frac{1}{2} t^2 & \text{if } |t| \leq c, \\ c|t| - \frac{1}{2} c^2 & \text{if } |t| > c. \end{cases} \tag{2.3}$$

The resulting $m(f; K, \rho)$ is a continuous version of the Huber estimator for discrete empirical data.

2.1. Basic properties. We begin our study of the continuous M-estimator by stating the following result that summarizes several basic properties of the minimand $\theta(\cdot; K, f, \rho)$ and shows in particular that this function inherits many properties of ρ. In the proof of this and other results, we will freely use known properties of convex functions, which can all be found in the classic treatise by Rockafellar [10].

PROPOSITION 2.1. *Let the triple (K, f, ρ) satisfy the blanket assumptions. The following statements hold for the function $\theta := \theta(\cdot; K, \rho, f)$.*

(a) *θ is convex (thus continuous) and coercive on $\mathbb{R}$; coercivity means*

$$\lim_{|m| \to \infty} \theta(m) = \infty.$$

(b) *If ρ is strictly convex, then so is θ.*

(c) *The right and left derivatives of θ are equal to, respectively,*

$$\theta'_+(m) \;:=\; \lim_{h\to 0+} \frac{\varphi(t+h)-\varphi(t)}{h} \;=\; -\int_K \rho'_-(f(x)-m)\,dx,$$

$$\theta'_-(m) \;:=\; \lim_{h\to 0+} \frac{\varphi(t)-\varphi(t-h)}{h} \;=\; -\int_K \rho'_+(f(x)-m)\,dx.$$

(d) *If ρ is k-times continuously differentiable on $\mathbb{R}$ for some $k \geq 1$, then so is θ.*
(e) *If ρ is twice differentiable and $\rho''(t) > 0$ for all $t \in \mathbb{R}$, then so is θ.*
(f) *θ is differentiable at m if and only if the set Ω_m has measure zero, where*

$$\Omega_m := \{\, x \in K : \rho \text{ is not differentiable at } f(x)-m \,\}.$$

Proof. The convexity of θ follows from that of ρ; the coercivity of θ follows from the inequalities

$$\liminf_{|\,m\,|\to\infty} \frac{\theta(m)}{|\,m\,|} \;\geq\; \min(\,\rho(1),\rho(-1)\,) \liminf_{|m|\to\infty} \int_K \frac{|\,f(x)-m\,|}{|\,m\,|}\,dx \;>\; 0.$$

Since K has positive measure, (b) is obvious. To prove (c), let $h \in (0,1]$. We have

$$\frac{\theta(m+h)-\theta(m)}{h} \;=\; \int_K \left[\, \frac{\rho(f(x)-m-h)-\rho(f(x)-m)}{h} \,\right] dx.$$

Since $\rho'_-(t)$ exists for all t, by the compactness of K, it follows that for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x \in K$ and all $h \in (0,\delta]$,

$$\left|\, \frac{\rho(f(x)-m-h)-\rho(f(x)-m)}{h} \,\right| \;\leq\; |\,\rho'_-(f(x)-m)\,| + \varepsilon.$$

Since the right-hand side, as a function of x, is clearly integrable on K, it follows by the dominated convergence theorem that

$$\theta'_+(m) \;=\; \int_K -\rho'_-(f(x)-m)\,dx.$$

Similarly, we can establish the desired formula for the left derivative of θ at m. The differentiability of θ in parts (d) and (e) can be proved easily. The positivity of θ'' in part (e) follows from the formula

$$\theta''(m) \;=\; \int_K \rho''(f(x)-m)\,dx. \tag{2.4}$$

Finally, it is clear that if the set Ω_m has measure zero, then

$$\theta'_+(m) \;=\; \theta'_-(m) \;=\; -\int_{\mathcal{D}_m} \rho'(f(x)-m)\,dx,$$

where $\mathcal{D}_m$ is the complement of Ω_m in K. Conversely, suppose that Ω_m has positive measure. Since the set of nondifferentiable points of ρ is countable, it follows that there must exist a $t_0 \in \mathbb{R}$ such that the set

$$\Lambda := \{\, x \in K : f(x)-m = t_0 \,\}$$

has positive measure. Since $\rho'_+(f(x)-m) \geq \rho'_-(f(x)-m)$ for all $x \in K$, it follows that

$$\int_K [\,\rho'_+(f(x)-m) - \rho'_-(f(x)-m)\,]\,dx \;\geq\; (\,\rho'_+(t_0) - \rho'_-(t_0)\,)\,\text{meas}\,\Lambda.$$

Since the right-hand side is positive, it follows that $\theta'_-(m) > \theta'_+(m)$; (f) therefore holds. □

The next result summarizes various properties of the continuous M-estimator $m(f;K,\rho)$. Part (a) asserts the well-definedness of this minimizer and gives its variational characterization; part (b) says that this minimizer must belong to the range $f(\text{int } K)$; part (c) is a technical property that will be used subsequently.

THEOREM 2.2. *Let (K,ρ,f) satisfy the blanket assumptions. The following statements are valid.*

(a) *$m(f;K,\rho)$ exists and is unique; it is the unique scalar $\bar{m}$ satisfying*

$$\int_K \rho'_+(f(x)-\bar{m})\,dx \;\geq\; 0 \;\geq\; \int_K \rho'_-(f(x)-\bar{m})\,dx.$$

(b) *There exists $\bar{x} \in \text{int } K$ such that $f(\bar{x}) = m(f;K,\rho)$.*

(c) *If $m(f) = \min f :=$ minimum value of f on K, then* meas $\{x \in K : f(x) = m(f) = \min f\} > 0$.

Proof. The existence of a global minimizer of $\theta(\cdot;K,\rho,f)$ follows from its convexity (and thus continuity) and its coercivity on $\mathbb{R}$. Moreover, such a global minimizer $\bar{m}$ is characterized by the inclusion $0 \in \partial\theta(\bar{m};K,\rho,f)$, with the latter subgradient equal to the interval

$$\left[-\int_K \rho'_+(f(x)-\bar{m})dx,\; -\int_K \rho'_-(f(x)-\bar{m})dx \right].$$

Thus, except for the uniqueness of the minimizer of $\theta(\cdot;K,\rho,f)$, (a) holds. By way of contradiction, we assume that no $\bar{x} \in \text{int } K$ exists satisfying $\bar{m} = f(\bar{x})$. The function $f(x) - \bar{m}$ then never vanishes on int K. Without loss of generality, we may assume that $f(x) - \bar{m}$ is positive on int K. By convexity of ρ, it follows that $\rho'_-(f(x)-\bar{m})$ is positive on int K, which implies

$$\int_K \rho'_-(f(x)-\bar{m})\,dx \;>\; 0$$

because ∂K has measure zero. The above inequality contradicts the variational characterization of $\bar{m}$. It remains to show the uniqueness of $m(f;K,\rho)$. Suppose there are two distinct minimizers m_1 and m_2. Since $\theta(\cdot;K,\rho,f)$ is convex, $\frac{1}{2}m_1 + \frac{1}{2}m_2$ is also a minimizer. Therefore,

$$\int_K \left[\,\rho(f(x) - m_1/2 - m_2/2) - \tfrac{1}{2}\,\rho(f(x)-m_1) - \tfrac{1}{2}\,\rho(f(x)-m_2)\,\right]\,dx \;=\; 0.$$

Since the integrand on the left-hand side is continuous and nonpositive, it follows that

$$\rho(f(x) - m_1/2 - m_2/2) \;=\; \tfrac{1}{2}\,\rho(f(x)-m_1) + \tfrac{1}{2}\,\rho(f(x)-m_2)$$

for all $x \in K$. By what has been proved above, it follows that there exist $x^1 \neq x^2$ in K such that $f(x^i) = m_i$ for $i = 1,2$. Hence, by the connectedness of K, there exists $x^* \in K$ such that

$$f(x^*) \;=\; \tfrac{1}{2}\,m_1 + \tfrac{1}{2}\,m_2.$$

Since ρ is a nonnegative function, we deduce $\rho(f(x^*) - m_1) = 0 = \rho(f(x^*) - m_2)$, which yields $f(x^*) = m_1 = m_2$, a contradiction. Finally, to prove (c), assume for contradiction that the measure of the set in question is zero. We then have

$$\int_K \rho'_-(f(x) - \min f)\, dx \;=\; \int_{K_+} \rho'_-(f(x) - \min f)\, dx,$$

where $K_+ := \{x \in K : f(x) > \min f = m(f)\}$. By assumption, meas $K_+ =$ meas $K > 0$. Since $\rho'_-(f(x) - \min f) > 0$ on K_+, it follows that

$$\int_K \rho'_-(f(x) - \min f)\, dx \;>\; 0,$$

which contradicts the characterization of $m(f)$ because $\min f = m(f)$ by assumption. □

It is clear that for any constant $c > 0$, $m(f + c) = m(f) + c$. The next result identifies an important monotonicity property of the continuous M-estimator.

THEOREM 2.3. *Let f and g be in $C(K)$ such that $f \geq g$ on K. It holds that $m(f) \geq m(g)$; moreover, strict inequality holds if either* (a) *$f \neq g$, ρ is differentiable, and ρ' is strictly increasing or* (b) *$f > g$ in the interior of K.*

Proof. Assume for contradiction that $m(f) < m(g)$. We have

$$\int_K \rho'_+(g(x) - m(g))\, dx \;\geq\; 0 \;\geq\; \int_K \rho'_-(f(x) - m(f))\, dx.$$

Since ρ'_- is nondecreasing, we deduce

$$\int_K \rho'_-(f(x) - m(f))\, dx \;\geq\; \int_K \rho'_-(g(x) - m(f))\, dx$$

and

$$\int_K \rho'_+(g(x) - m(f))\, dx \;\geq\; \int_K \rho'_+(g(x) - m(g))\, dx.$$

Consequently,

$$\int_K \rho'_+(g(x) - m(f))\, dx \;\geq\; 0 \;\geq\; \int_K \rho'_-(g(x) - m(f))\, dx.$$

By the variational characterization and uniqueness of $m(g)$, it follows that $m(f) = m(g)$, which is a contradiction. If ρ is differentiable, then $m(f)$ and $m(g)$ are the unique scalars m_f and m_g that satisfy the equations

$$\int_K \rho'(f(x) - m(f))\, dx \;=\; 0 \quad \text{and} \quad \int_K \rho'(g(x) - m(g))\, dx \;=\; 0,$$

respectively. If $f \neq g$, then there must exist an open set $\mathcal{O}$ contained in the interior of K such that $f(x) > g(x)$ for all $x \in \mathcal{O}$. If $m(f) = m(g)$, the above characterizations of $m(f)$ and $m(g)$ immediately yield a contradiction, using the fact that $f > g$ on the open set $\mathcal{O} \subset K$. Consequently, strict monotonicity of m (i.e., $m(f) > m(g)$) holds under (a).

Assume (b). Writing $\theta_f := \theta(\cdot; K, \rho, f)$ and $\theta_g := \theta(\cdot; K, \rho, g)$, we divide the remaining proof into six steps.

Step 1. Since $f > g$ in the interior of K, we must have

$$\text{meas}\,\{\,x \in K : g(x) > m(g)\,\} \le \text{meas}\,\{\,x \in K : f(x) > m(g)\,\};$$

the claim is that strict inequality also holds:

$$\text{meas}\,\{x \in K : g(x) > m(g)\} < \text{meas}\,\{x \in K : f(x) > m(g)\}. \tag{2.5}$$

To prove this assertion, we distinguish between two cases (i) $m(g) = \min g$ and (ii) $m(g) > \min g$.

Step 2. Assume that $m(g) = \min g$. Since $\{x \in \text{int } K : f(x) > m(g)\} = \text{int } K$ and meas $\partial K = 0$, it follows that

$$\begin{aligned}
&\text{meas}\,\{\,x \in K : f(x) > m(g)\,\}\\
&\quad = \text{meas}\,\{\,x \in \text{int } K : f(x) > m(g)\,\} = \text{meas int } K = \text{meas } K\\
&\quad = \text{meas}\,\{\,x \in K : g(x) = \min g = m(g)\,\} + \text{meas}\,\{\,x \in K : g(x) > m(g)\,\}\\
&\quad > \text{meas}\,\{\,x \in K : g(x) > m(g)\,\},
\end{aligned}$$

where the last inequality follows from Theorem 2.2(c). Consequently, (2.5) holds in case (i).

Step 3. In case (ii), we have $\min g < m(g)$. Consider the sets

$$E := \{\,x \in K : g(x) \ge m(g)\,\} \quad \text{and} \quad U := \{\,x \in K : f(x) > m(g)\,\}.$$

Since E is contained in the closure of U, we must have meas $U \ge$ meas E. The desired inequality (2.5) follows readily if we can show meas $U >$ meas E. Assume for contradiction that meas $U =$ meas E. It then follows that meas $(U \setminus E) = 0$. But the only way the set $U \setminus E$, which is open in K, has zero measure is when it is empty. This means $U = E$. So $E = U$ is both open and closed in K, which is a connected set. Hence either $E = U = K$ or $E = U = \emptyset$. The former is not possible because $m(g) > \min g$; the latter is not possible because $m(g)$ belongs to the range $g(K)$. This establishes the claim, and thus (2.5) too.

Step 4. Inequality (2.5) implies

$$-\int_K \rho'_{\mp}(f(x) - m(g))\,dx \;<\; -\int_K \rho'_{\mp}(g(x) - m(g))\,dx.$$

Consequently,

$$(\theta_f)'_{\pm}(m(g)) \;<\; (\theta_g)'_{\pm}(m(g)). \tag{2.6}$$

Since $(\theta_f)'_+(m(f)) \ge 0$, to show $m(f) > m(g)$, it suffices to show

$$(\theta_f)'_+(m(g)) \;<\; 0. \tag{2.7}$$

Note that (2.6) is only good enough to imply that $(\theta_f)'_-(m(g)) < 0$, which is weaker than (2.7). If either θ_f or θ_g is differentiable at $m(g)$, then (2.7) follows readily. This is clear in the former case, whereas in the latter case we have $(\theta_f)'_+(m(g)) < (\theta_g)'(m(g)) = 0$.

Step 5. If θ_f and θ_g both fail to be differentiable at $m(g)$, let

$$f_\alpha := \alpha\, f + (1-\alpha)\, g \quad \text{and} \quad \theta_\alpha(m) := \int_K \rho(f_\alpha(x) - m)dx, \quad \alpha \in [0,1].$$

We claim that there exists $\alpha^* \in (0,1)$ such that θ_{α^*} is differentiable at $m(g)$. With this claim established, we can complete the proof as follows. Since $f > f_{\alpha^*} > g$ in the interior of K, together with the result in Step 4 and part (a) proved above, we arrive at $m(f) \geq m(f_{\alpha^*}) > m(g)$. Thus we are left only with the proof of the last claim.

Step 6. Let B be the set of points at which ρ fails to be differentiable. For each $t \in B$ and $\alpha \in (0,1)$, define

$$N_{\alpha,t} := \{\, x \in \text{int } K : f_\alpha(x) - m_g = t \,\}.$$

Note that for fixed $t \in B$ and $\alpha > \alpha'$, the sets $N_{\alpha,t}$ and $N_{\alpha',t}$ are disjoint. Since K has finite measure, for each $t \in B$ and each $n = 1, 2, \ldots$, only a finite number of α satisfies meas $N_{\alpha,t} > 1/n$. Hence the set $\{\alpha \in (0,1) : \text{meas } N_{\alpha,t} > 0\}$ is countable. Since B is countable, it follows that

$$\{\alpha \in (0,1) : \text{meas } N_{\alpha,t} > 0 \text{ for some } t \in B\} = \bigcup_{t \in B} \{\alpha \in (0,1) : \text{meas } N_{\alpha,t} > 0\}$$

is countable. But $(0,1)$ is uncountable, so there exists $\alpha^* \in (0,1)$ such that

$$\begin{aligned} 0 &= \text{meas} \bigcup_{t \in B} N_{\alpha^*,t} = \text{meas } \{\, x \in \text{int } K : f_{\alpha^*}(x) \in B \,\} \\ &= \text{meas } \{\, x \in K : f_{\alpha^*}(x) \in B \,\}. \end{aligned}$$

By part (f) of Proposition 2.1, it follows that θ_{α^*} is differentiable at $m(g)$. □

Remark. It is in general not true that if $f \geq g$ and $f \neq g$, then $m(f) > m(g)$, even if f is almost everywhere strictly greater than g. An example is if f and g are continuous functions defined on an interval I and are both strictly increasing, and $f(x) > g(x)$ except at the midpoint of I. With $\rho = |\cdot|$, it follows that $m(f) = \text{med}(f|I) = f(\text{midpoint of } I) = g(\text{midpoint of } I) = \text{med}(g|I) = m(g)$. □

Part (a) of the next result implies that the continuous M-estimator $m(f)$ is nonexpansive, and thus continuous, in its argument; the second part extends part (b) of Theorem 2.2.

COROLLARY 2.4. *The following two statements hold.*

(a) *The continuous M-estimator m is nonexpansive on $C(K)$; i.e.,*

$$|\, m(f) - m(g) \,| \leq \max_{x \in K} |\, f(x) - g(x) \,| \quad \forall\, f, g \in C(K).$$

(b) *If f and g in $C(K)$ are such that $m(f) = m(g)$, then $x \in \text{int } K$ exists such that $f(x) = g(x)$.*

Proof. Let $\sigma := \max_{x \in K} |f(x) - g(x)|$. We have $g - \sigma \leq f \leq g + \sigma$ on K. An application of the monotonicity of m yields

$$m(g) - \sigma = m(g - \sigma) \leq m(f) \leq m(g + \sigma) = m(g) + \sigma,$$

from which the desired nonexpansiveness of m follows readily.

To prove statement (b), assume for contradiction that $f(x) \neq g(x)$ for all $x \in \text{int } K$. It then follows that either $f > g$ on int K or $f < g$ on int K. Either case yields a contradiction to the assumption that $m(f) = m(g)$ by the strict monotonicity of m. □

The next result asserts a "strong minimizing" property of $m(f; K, \rho)$ associated with a C^2 function ρ with positive second derivatives. A consequence of this property

is that a global a posteriori error bound exists for the continuous M-estimator; see the last inequality in the proposition below.

PROPOSITION 2.5. *Let $f \in C(K)$ and $\rho \in C^2$ be given. If $\rho''(f(x) - m(f))$ does not vanish identically on K, then $\theta''(m(f); K, f, \rho) > 0$. Hence, positive constants c and δ exist such that*

$$|\, m - m(f)\,| \,\le\, \delta \;\Rightarrow\; \theta(m; K, f, \rho) - \theta(m(f); K, f, \rho) \,\ge\, c\,|\, m - m(f)\,|^2 \tag{2.8}$$

and

$$|\, m - m(f)\,| \,>\, \delta \;\Rightarrow\; \theta(m; K, f, \rho) - \theta(m(f); K, f, \rho) \,\ge\, c\,\delta\,|\, m - m(f)\,|. \tag{2.9}$$

Consequently, there exists $\eta > 0$ such that, for all $m \in \mathbb{R}$,

$$\begin{aligned}&|\, m - m(f)\,| \\ &\le\, \eta \max\left\{\, \theta(m; K, f, \rho) - \theta(m(f); K, f, \rho),\, \sqrt{\theta(m; K, f, \rho) - \theta(m(f); K, f, \rho)}\,\right\}.\end{aligned}$$

Proof. Write $\theta := \theta(\cdot; K, f, \rho)$. By part (d) of Proposition 2.1, θ'' exists. Moreover, the expression (2.4) shows that θ'' is nonnegative on $\mathbb{R}$. If $\theta''(m(f)) = 0$, then since f is continuous on K and ρ'' is nonnegative and continuous on $\mathbb{R}$, it follows that $\rho''(f(x) - m(f)) = 0$ for all $x \in K$. But this contradicts the assumption that the latter function does not vanish identically on K. Consequently, $\theta''(m(f)) > 0$.

The existence of δ and c satisfying (2.8) is a standard second-order consequence of the minimizing property of $m(f)$ and of the positivity of $\theta''(m(f); K, f, \rho)$. To prove (2.9), let m be such that $m - m(f) > \delta$. (The proof of the case $m - m(f) < -\delta$ is similar and omitted.) Define $m' := m(f) + \delta$. We then have

$$m' \,=\, \frac{\delta}{m - m(f)}\, m + \frac{m - m(f) - \delta}{m - m(f)}\, m(f),$$

which implies, by convexity of θ,

$$\theta(m') \,\le\, \frac{\delta}{m - m(f)}\, \theta(m) + \frac{m - m(f) - \delta}{m - m(f)}\, \theta(m(f)).$$

Since $\theta(m') \ge \theta(m(f)) + c(m' - m(f))^2 = \theta(m(f)) + c\delta^2$, (2.9) follows readily. The last assertion of the proposition is immediate from (2.8) and (2.9). ∎

3. Interpolation of continuous M-estimators. In this and the next section, we consider the interpolation of the continuous M-estimators of polynomials on given intervals. This section establishes the well-posedness of the interpolation problem using nonconstructive topological arguments, generalizing the main result in [6]. The next section discusses Newton's method for solving the interpolation problem with a smooth ρ.

Denote by Π_n the space of all polynomials of degree $\le n$. Let I be a given closed finite interval. For any nonzero $p \in \Pi_n$ and any $m \in \mathbb{R}$, the set $p^{-1}(m) \cap I$ has at most n elements. Hence, the set

$$\{\, t \in I : \rho \text{ is not differentiable at } p(t) - m \,\}$$

is countable; consequently, by part (f) of Proposition 2.1, the function $\theta(m)$ is differentiable everywhere on $\mathbb{R}$. Hence, for $p \in \Pi_n$ we have

$$\int_I \rho'_{\pm}(p(t) - m(p\,; I, \rho))\, dt \,=\, 0. \tag{3.1}$$

We introduce the map $\mathbf{M}_\rho : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ associated with $n+1$ *nonoverlapping* compact intervals I_i, for $i = 0, \ldots, n$, each with a nonempty interior, where nonoverlapping means $\operatorname{int} I_i \cap \operatorname{int} I_j = \emptyset$ for all $i \neq j$. For each vector $a \in \mathbb{R}^{n+1}$ with components a_i for $i = 0, \ldots, n$, let $\mathbf{M}_\rho(a)$ be the $(n+1)$-vector whose ith component, for $i = 0, \ldots, n$, is equal to $m(p; I_i, \rho)$, where p is the polynomial in Π_n with coefficients a_j; i.e.,

$$p(t) \equiv \sum_{j=0}^{n} a_j \, t^j, \quad t \in \mathbb{R}. \tag{3.2}$$

The goal of this section is to establish the following main result, which immediately implies the *well-posedness* of the interpolation problem of continuous M-estimators: given any values m_i, $i = 0, \ldots, n$, there exists a unique $p \in \Pi_n$ such that $m(p; I_i, \rho) = m_i$ for all i; moreover, such an interpolant p depends continuously on the data m_i.

THEOREM 3.1. *$\mathbf{M}_\rho$ is a homeomorphism from $\mathbb{R}^{n+1}$ onto $\mathbb{R}^{n+1}$.*

Proof. It suffices to show that $\mathbf{M}_\rho$ is continuous, injective, and norm-coercive [8, Theorem 5.3.8]. By Corollary 2.4(a), $\mathbf{M}_\rho$ is Lipschitz continuous. Indeed, for any two vectors a and b in $\mathbb{R}^{n+1}$ with associated polynomials p and q, we have, for every $i = 0, 1, \ldots, n$,

$$| \, m(p; I_i, \rho) - m(q; I_i, \rho) \, | \leq \sup_{t \in I_i} | \, p(t) - q(t) \, | \leq L_i \, \| \, a - b \, \|,$$

where $\| \cdot \|$ denotes the Euclidean norm in $\mathbb{R}^{n+1}$ and $L_i > 0$ is a constant that depends on the interval I_i and is independent of ρ and the vectors a and b. Letting $L = \max_{0 \leq i \leq n} L_i$, we deduce

$$\| \, \mathbf{M}_\rho(a) - \mathbf{M}_\rho(b) \, \| \leq L \, \| \, a - b \, \| \quad \forall \, \rho \text{ and } \quad \forall \, a, b \in \mathbb{R}^{n+1}. \tag{3.3}$$

Note that the Lipschitz constant L is independent of ρ; hence the family of maps $\{\mathbf{M}_\rho : \rho\}$ is *equi-Lipschitz continuous*. Injectivity of $\mathbf{M}_\rho$ follows easily from Corollary 2.4(b): If $p, q \in \Pi_n$ are such that $\mathbf{M}_\rho(p) = \mathbf{M}_\rho(q)$, then there exists, for every $i = 0, \ldots, n$, a point $t_i \in \operatorname{int} I_i$ such that $p(t_i) = q(t_i)$; this implies $p = q$. It remains to show norm-coerciveness; i.e., we need to show that

$$\lim_{\|a\| \to \infty} \| \, \mathbf{M}_\rho(a) \, \| = \infty. \tag{3.4}$$

We divide the remaining proof into several major steps.

Step 1. Define, for any scalar $c > 0$, the function $\rho_c : \mathbb{R} \to \mathbb{R}$ by $\rho_c(t) := \rho(ct)$ for all $t \in \mathbb{R}$. Observe that for any interval I and any function $f \in C(I)$, we have

$$\begin{aligned} m(f; I, \rho) &= \arg\min_m \int_I \rho(f(t) - m) \, dt = \arg\min_m \int_I \rho_c \left(\frac{f(t)}{c} - \frac{m}{c} \right) dt \\ &= c \, m \left(\frac{f}{c}; I, \rho_c \right). \end{aligned}$$

It follows that for a nonzero vector $a \in \mathbb{R}^{n+1}$,

$$\mathbf{M}_\rho(a) = \| \, a \, \| \, \mathbf{M}_{\rho_{\|a\|}} \left(a / \|a\| \right). \tag{3.5}$$

Let S^n denote the unit sphere in $\mathbb{R}^{n+1}$. Clearly, if

$$\inf \{ \, \| \, \mathbf{M}_{\rho_c}(a) \, \| \, : \, c \geq 1, \; a \in S^n \, \} > 0, \tag{3.6}$$

then (3.4) follows. It is clear that (3.6) implies

$$\inf\{\,\|\,\mathbf{M}_{\rho_c}(a)\,\|\,:\,c\,\geq\,1\,\}\,>0\quad\forall\,a\,\in\,S^n. \tag{3.7}$$

Thanks to the compactness of S^n and the equi-Lipschitz continuity of $\{\mathbf{M}_{\rho_c} : c > 0\}$, the converse implication also holds. Indeed, assume that (3.7) holds but (3.6) does not. There exist sequences $\{a^k\} \subset S^n$ and $\{c_k\} \in [1, \infty)$ such that $\mathbf{M}_{\rho_{c_k}}(a^k) \to 0$. Since S^n is compact, we may assume without loss of generality that the sequence $\{a^k\}$ converges to some $a^\infty \in S^n$. By (3.3), we have

$$\|\,\mathbf{M}_{\rho_{c_k}}(a^k) - \mathbf{M}_{\rho_{c_k}}(a^\infty)\,\|\,\leq\,L\,\|\,a^k - a^\infty\,\|\quad\forall\,k.$$

Consequently, it follows that

$$\lim_{k\to\infty}\,\mathbf{M}_{\rho_{c_k}}(a^\infty)\,=\,0,$$

which contradicts (3.7). Hence (3.6) $\Longleftrightarrow$ (3.7).

Step 2. Let $a \in \mathbb{R}^{n+1}$ be an arbitrary nonzero vector, and let $p \in \Pi_n$ be given by (3.2). Since p has no more than n roots and since there are $(n + 1)$ nonoverlapping intervals I_i, it follows that one of two cases must hold:

(i) there exists $j \in \{0, \ldots, n\}$ such that $p > 0$ or $p < 0$ on I_j, or

(ii) $n \geq 1$ and there exists $j \in \{0, \ldots, n-1\}$ such that I_j and I_{j+1} intersect at a common endpoint where p vanishes and $p(x)p(y) < 0$ for all $x \in \text{int } I_j$ and $y \in \text{int } I_{j+1}$.

To see this, note the following special property of a polynomial. Namely, if a polynomial does not change sign and does not vanish in an interval except at one interior point of the interval, then the latter point must be a root of the polynomial of multiplicity at least two. By this property and a straightforward counting argument, one can easily show that if neither of the above two cases hold, then, counting multiplicities, p must have at least $n + 1$ zeros, which is impossible. To complete the proof of the theorem, we show that (3.7) holds in either case (i) or (ii) above.

Step 3. In case (i) there exists $\delta > 0$ such that p or $-p \geq \delta$ on I_j. Since $m(p\,; I_j, \rho_c) = p(t_j)$ for some $t_j \in I_j$, it follows that $\|\mathbf{M}_{\rho_c}(a)\| \geq \delta$ for all $c > 0$.

Step 4. In case (ii) we prove (3.7) by contradiction. If (3.7) fails, then there exists a sequence $\{c_k\}$ in $[1, \infty)$ such that

$$\lim_{k\to\infty}\,m(p\,; I_i, \rho_{c_k})\,=\,0\quad\text{for } i = j \text{ and } j+1.$$

Without loss of generality, we may assume that

$$p\,<\,0 \text{ on int } I_j\quad\text{and}\quad p\,>\,0 \text{ on int } I_{j+1}.$$

For $i = j$, $j + 1$, write $m_{k,i} := m(p\,; I_i, \rho_{c_k})$ and $\theta_{k,i} := \theta(\cdot; I_i, p, \rho_{c_k})$. By part (c) of Theorem 2.2, it follows that $m_{k,j} < 0$ and $m_{k,j+1} > 0$ for all k.

By (3.1), we have the following optimality condition:

$$\int_{I_j}\,\rho'_{\pm}\,(c_k[p(t) - m_{k,j}])\,dt\,=\,0\,=\,\int_{I_{j+1}}\,\rho'_{\pm}\,(c_k[p(t) - m_{k,j+1}])\,dt\quad\forall\,k = 1, 2, \ldots\,. \tag{3.8}$$

Since meas$\{t \in I_i : p(t) = 0\} = 0$, an elementary property of measure gives

$$\lim_{m\uparrow 0}\,\text{meas}\,\{t \in I_j : p(t) \geq m\} = 0 = \lim_{m\downarrow 0}\,\text{meas}\,\{t \in I_{j+1} : p(t) \leq m\}. \tag{3.9}$$

By continuity of p, there exist $\varepsilon > 0$, $\delta > 0$, $J_i \subset I_i$ with meas $J_i \geq \varepsilon$ for $i = j, j+1$ such that $p \leq -\delta$ and $p \geq \delta$ on J_j and J_{j+1}, respectively. Pick k large enough such that $m_{k,j+1} - m_{k,j} < \delta$ (in particular, $-\delta < m_{k,j} < 0$ and $0 < m_{k,j+1} < \delta$),

$$\text{meas } \{t \in I_j : p(t) > m_{k,j}\} < \varepsilon \quad \text{and} \quad \text{meas } \{t \in I_{j+1} : p(t) < m_{k,j+1}\} < \varepsilon.$$

We then have

$$0 = \int_{I_j} (\rho_{c_k})'_+(p(t) - m_{k,j})\, dt = T_1 + T_2 + T_3,$$

where

$$\begin{aligned}
T_1 &:= \int_{\{t:p(t)\leq -\delta\}} (\rho_{c_k})'_+(p(t) - m_{k,j})\, dt \\
&\leq \int_{\{t:p(t)\leq -\delta\}} (\rho_{c_k})'_+(-\delta - m_{k,j})\, dt \leq \varepsilon(\rho_{c_k})'_+(-m_{k,j+1}) \\
T_2 &:= \int_{\{t:-\delta<p(t)<m_{k,j}\}} (\rho_{c_k})'_+(p(t) - m_{k,j})\, dt < 0 \\
T_3 &:= \int_{\{t:p(t)\geq m_{k,j}\}} (\rho_{c_k})'_+(p(t) - m_{k,j})\, dt \\
&\leq \int_{\{t:p(t)>m_{k,j}\}} (\rho_{c_k})'_+(-m_{k,j})\, dt \leq \varepsilon\, (\rho_{c_k})'_+(-m_{k,j}),
\end{aligned}$$

by the monotonicity of $(\rho_{c_k})'_+$. Hence

$$0 < (\rho_{c_k})'_+(-m_{k,j+1}) + (\rho_{c_k})'_+(-m_{k,j}).$$

Similarly, using

$$0 = \int_{I_{j+1}} (\rho_{c_k})'_+(p(t) - m_{k,j+1})\, dt,$$

we can deduce $0 > (\rho_{c_k})'_+(-m_{k,j}) + (\rho_{c_k})'_+(-m_{k,j+1})$, which is a contradiction. ∎

4. Newton's method for interpolation. By Theorem 3.1, the equation

$$\mathbf{M}_\rho(a) = b \tag{4.1}$$

has a unique solution for every vector $b \in \mathbb{R}^{n+1}$; moreover, such a solution depends continuously on b. A natural approach to solve (4.1) is by Newton's method, possibly with a globalization scheme [2]. In order for this method to be directly applicable, we take ρ to be twice continuously differentiable throughout this section. This smoothness assumption allows us to completely bypass the difficulty in dealing with the nonsmoothness of the original median-interpolation problem considered in [3, 6]. Moreover, by the theorem below, if $\rho'' > 0$, then $\mathbf{M}_\rho$ is a diffeomorphism on $\mathbb{R}^{n+1}$; i.e., both $\mathbf{M}_\rho$ and $(\mathbf{M}_\rho)^{-1}$ are differentiable maps. In particular, $\mathbf{M}_\rho$ has a nonsingular Jacobian matrix everywhere on $\mathbb{R}^{n+1}$.

THEOREM 4.1. *Let I_i, $i = 0, \ldots, n$, be $n+1$ nonoverlapping compact intervals, each with a nonempty interior. Let $\rho : \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable*

function with $\rho'' > 0$ on $\mathbb{R}$. If ρ has a minimizer at zero such that $\rho(0) = 0$, then $\mathbf{M}_\rho$ is a diffeomorphism from $\mathbb{R}^{n+1}$ onto itself.

Proof. For simplicity, we write $\mathbf{M}$ for $\mathbf{M}_\rho$. The positivity of ρ'' implies that ρ is strictly convex on $\mathbb{R}$. Hence, zero is the unique minimizer of ρ. For $i = 0, \ldots, n$, write

$$\theta_i(m, a) := \int_{I_i} \rho\left(\sum_{j=0}^{n} a_j\, t^j - m\right) dt. \tag{4.2}$$

As a function of the two arguments $(m, a) \in \mathbb{R}^{1+(n+1)}$, θ_i is twice continuously differentiable. For any fixed but arbitrary vector $a \in \mathbb{R}^{n+1}$, the continuous M-estimator $\mathbf{M}_i(a)$ is the unique root m of the equation

$$\frac{\partial \theta_i(m, a)}{\partial m} = 0,$$

which is a parametric, nonlinear equation with m as the primary variable and a as the parameter. By Proposition 2.5,

$$\frac{\partial^2 \theta_i(\mathbf{M}_i(a), a)}{\partial m^2} > 0.$$

Therefore, by the implicit-function theorem, it follows that $\mathbf{M}_i$ (and thus $\mathbf{M}$) is a differentiable function on $\mathbb{R}^{n+1}$. We claim that the Jacobian matrix of $\mathbf{M}$ at $a \in \mathbb{R}^{n+1}$, denoted $J\mathbf{M}(a)$, is nonsingular. Note that

$$\frac{\partial \mathbf{M}_i(a)}{\partial a_j} = \left(\frac{\partial^2 \theta_i(\mathbf{M}_i(a), a)}{\partial m^2}\right)^{-1} \left(\frac{\partial^2 \theta_i(\mathbf{M}_i(a), a)}{\partial m\, \partial a_j}\right). \tag{4.3}$$

If $e \in \mathbb{R}^{n+1}$ is such that $J\mathbf{M}(a)e = 0$, then we have

$$\sum_{j=0}^{n} \frac{\partial^2 \theta_i(\mathbf{M}_i(a), a)}{\partial m\, \partial a_j}\, e_j = 0 \quad \forall\, i = 0, \ldots, n.$$

By a direct differentiation, we can see that the above is equivalent to

$$\int_{I_i} \rho''\left(\sum_{\ell=0}^{n} a_\ell\, t^\ell - \mathbf{M}_i(a)\right) q(t)\, dt = 0, \quad i = 0, \ldots, n,$$

where $q(t) \equiv \sum_{j=0}^{n} e_j\, t^j$. Since ρ'' is everywhere positive, it follows that q has at least one zero in int I_i for every $i = 0, \ldots, n$. Consequently, the polynomial $q(t)$, which is of degree n, has $n + 1$ distinct real roots. This is not possible unless $q(t)$ is the zero polynomial. Hence $J\mathbf{M}(a)$ must be nonsingular. This implies that $\mathbf{M}^{-1}$ is differentiable; hence $\mathbf{M}$ is a diffeomorphism on $\mathbb{R}^{n+1}$. □

The condition $\rho'' > 0$ cannot be dispensed with in the above theorem. For example consider $\rho = |\cdot|^p$; then $\rho \in \mathrm{C}^2$ when $p \geq 2$. On the other hand, $\mathbf{M}_{|\cdot|^p}$ is homogeneous of degree 1; i.e.,

$$\mathbf{M}_{|\cdot|^p}(c\, a) = c\, \mathbf{M}_{|\cdot|^p}(a); \quad \text{thus} \quad D_v \mathbf{M}_{|\cdot|^p}(0) = \mathbf{M}_{|\cdot|^p}(v),$$

where D_v denotes the directional derivative operator in the direction v. This implies that $\mathbf{M}_{|\cdot|^p}$ is differentiable at the origin if and only if $\mathbf{M}_{|\cdot|^p}$ is a linear map; but an elementary calculation shows that (unless $n \leq 1$) the latter is true when and only when $p = 2$.

4.1. Method I. With the above preparation, we can formally state a globally convergent Newton method for solving (4.1); see [2, 11]. In essence, this is the damped Newton method that involves two main computational steps in each iteration: the first step is solving the Newton equation

$$\mathbf{M}_\rho(a^k) + J\mathbf{M}_\rho(a^k)\, da^k \;=\; b \tag{4.4}$$

for the direction da^k at the current iterate a^k; the second step is performing an Armijo line search [1, section 1.2, p. 29] on the merit function

$$\phi(a) \;:=\; \tfrac{1}{2}\,(\,\mathbf{M}_\rho(a) - b\,)^T(\,\mathbf{M}_\rho(a) - b\,)$$

at the current iterate along the computed Newton direction. Normally, given scalars σ' and β in $(0, 1)$, the latter line search calls for the determination of the smallest nonnegative integer i satisfying

$$\phi(a^k + \beta^i da^k) - \phi(a^k) \;\le\; \sigma'\,\beta^i\,\nabla\phi(a^k)^T da^k.$$

With the function ϕ on hand and the search direction da^k computed from (4.4), we have

$$\nabla\phi(a^k) \;=\; J\mathbf{M}_\rho(a^k)^T(\,\mathbf{M}_\rho(a^k) - b^k\,),$$

which yields $\nabla\phi(a^k)^T da^k = -2\phi(a^k)$. Noting the latter identity, we formulate the following algorithm.

Newton's Algorithm for Interpolation.

(a) (inputs). Let the function ρ and the $(n+1)$ intervals I_i satisfy the assumptions in Theorem 4.1. Let $b \in \mathbb{R}^{n+1}$ and $a^0 \in \mathbb{R}^{n+1}$ be given. Let σ and β be given scalars in $(0, 1)$. Set $k = 0$.

(b) (solving linear equations). Solve (4.4) for the Newton direction da^k.

(c) (Armijo line search). Let i_k be the smallest nonnegative integer i such that

$$\phi(a^k + \beta^i da^k) \;\le\; (\,1 - \sigma\,\beta^i\,)\,\phi(a^k).$$

Set $\tau_k := \beta^{i_k}$ and $a^{k+1} := a^k + \tau_k da^k$. Let $k \leftarrow k + 1$.

(d) If $\|\mathbf{M}_\rho(a^k) - b\| \le$ tolerance, stop. Otherwise return to step (b). □

While the convergence of Newton's method is well known (cf. [2, Chapter 6] or [5, Chapter 8], e.g.), for completeness we give a sketch of the proof of the following result, omitting some details.

Theorem 4.2. *Under the assumptions of Theorem* 4.1*, the algorithm as described above generates a well-defined sequence* $\{a^k\}$ *that converges Q-superlinearly to the unique solution of* (4.1)*. Moreover,* $i_k = 0$ *for all but finitely many* k*. Finally, if* ρ'' *is Lipschitz continuous on* $\mathbb{R}$*, then the convergence is Q-quadratic.*

Proof. The well-definedness of each search direction da^k is ensured by the nonsingularity of $J\mathbf{M}_\rho(a)$ on $\mathbb{R}^{n+1}$. Since the method is essentially a gradient-related line-search method applied to the unconstrained minimization of the merit function $\phi(a)$, standard results from nonlinear programming, such as [1, Proposition 1.2.1], guarantee that the sequence $\{a^k\}$ is well defined and that every accumulation point of the sequence is a stationary point of ϕ. At least one such point must exist because ϕ has bounded level sets; in turn, the latter is due to the diffeomorphism property of $\mathbf{M}_\rho$. Since $\nabla\phi(a) = J\mathbf{M}_\rho(a)^T(\mathbf{M}_\rho(a) - b)$ and $J\mathbf{M}_\rho(a)$ is a nonsingular matrix for all a, it follows that every accumulation point of the sequence $\{a^k\}$ satisfies (4.1).

Since the latter equation has a unique solution, it follows that $\{a^k\}$ converges to the unique solution of (4.1). Letting a^* be the unique zero of (4.1), we note that $\nabla^2\phi(a^*) = J\mathbf{M}_\rho(a^*)^T J\mathbf{M}_\rho(a^*)$ is positive definite. Therefore, by verifying the well-known Dennis–Moré condition (see [2, equation (6.3.10)]), Theorem 6.3.4 in the latter reference immediately yields the assertion about the ultimate attainment of a unit step size (i.e., $i_k = 0$ for all but finitely many k). This in turn readily implies the quadratic convergence statement. □

4.2. Method II. We present below an alternative formulation of the system of equations (4.1) which results in a total bypass of the evaluation of $\mathbf{M}_\rho(a)$. Nevertheless, we should caution the reader that there is a theoretical difference between the alternative formulation and the original equation (4.1) that will become clear in the following discussion.

Since $\mathbf{M}_i(a) \equiv (\mathbf{M}_\rho)_i(a)$ satisfies $\int_{I_i} \rho'(p(t) - \mathbf{M}_i(a))\,dt = 0$, it is clear that (4.1) is equivalent to

$$\int_{I_i} \rho'\left(\sum_{j=0}^{n} a_j t^j - b_i\right) dt = 0, \quad i = 0, \ldots, n. \tag{4.5}$$

Define the C^1 map $F_\rho := F : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$, where $F_i(a)$ is the left-hand side of the above equation. It is useful to clarify the difference between the two functions $\mathbf{M}$ and F in terms of the following function:

$$\psi_i(m, a) := \int_{I_i} \rho'\left(\sum_{j=0}^{n} a_j t^j - m\right) dt = -\frac{\partial \theta_i(m, a)}{\partial m}, \quad (m, a) \in \mathbb{R}^{1+(n+1)},$$

where $\theta_i(m, a)$ is given by (4.2). While $\mathbf{M}_i(a)$ is the (unique) zero of $\psi_i(\cdot, a)$ on $\mathbb{R}$, $F_i(a)$ is the value of $\psi_i(\cdot, a)$ at a given value $b_i \in \mathbb{R}$.

A similar Newton method can be applied to the equation

$$F(a) = 0. \tag{4.6}$$

It is easy to show that

$$\frac{\partial F_i(a)}{\partial a_j} = \left(\frac{\partial^2 \theta_i(b_i, a)}{\partial m\, \partial a_j}\right). \tag{4.7}$$

Under the assumptions of Theorem 4.1, we can show that $JF(a)$ is nonsingular for all $a \in \mathbb{R}^{n+1}$. Thus F is a local homeomorphism everywhere on $\mathbb{R}^{n+1}$. However, unlike $\mathbf{M}$, F is in general *not* a global homeomorphism. In fact, for a function ρ whose derivative ρ' is bounded (e.g., $\rho(t) \equiv \sqrt{t^2 + c} - \sqrt{c}$ for any $c > 0$), it is clear that F has a bounded range and therefore cannot be norm-coercive, and thus it cannot be a global homeomorphism.

In spite of the theoretical difference, Newton's method can be applied to (4.6). Omitting the details, we summarize the two main computational steps in each iteration. At the beginning of each iteration, an iterate $a^k \in \mathbb{R}^{n+1}$ is given. We then solve for da^k in the equation

$$F(a^k) + JF(a^k)\,da^k = 0$$

and next perform an Armijo line search on the merit function

$$\varphi(a) := \tfrac{1}{2} F(a)^T F(a)$$

starting at a^k and moving along the direction da^k. In the resulting scheme, there is no longer a need to evaluate $\mathbf{M}(a^k)$. The convergence of this alternative application of Newton's method is summarized below.

THEOREM 4.3. *Under the assumptions of Theorem* 4.1*, the algorithm as described above generates a well-defined sequence* $\{a^k\}$*. If the level set*

$$L(a^0) := \{ a \in \mathbb{R}^{n+1} : \| F(a) \|_2 \leq \| F(a^0) \|_2 \}$$

is bounded, then $\{a^k\}$ *is bounded, and all other conclusions of Theorem* 4.2 *remain valid.*

Proof. Since the sequence $\{a^k\}$ is contained in the level set $L(a^0)$, the boundedness of $\{a^k\}$ follows from that of the set. The rest of the proof is similar to that of Theorem 4.2. □

In summary, when Newton's method is applied to the two equivalent equations, (4.1) and (4.6), the resulting algorithms differ computationally and theoretically. The computational difference lies in the need to evaluate the continuous M-estimators $\mathbf{M}(a)$ at intermediate polynomials. The theoretical difference lies in the choice of initial iterate a^0. With (4.1), there is no restriction on a^0; with (4.6), a^0 should be such that the level set $L(a^0)$ is bounded.

4.3. Smoothed continuous medians. In addition to producing M-estimators that are of independent interest, the choice of a C^2 function ρ can be used to approximate the continuous median. One such family of "smoothed continuous medians" is obtained by letting

$$\rho_\varepsilon(t) := \sqrt{t^2 + \varepsilon^2} - \varepsilon, \quad t \in \mathbb{R},$$

where ε is a positive scalar presumed to be small. Note that ρ_ε is globally Lipschitz continuous on $\mathbb{R}$ and ρ_ε'' is positive everywhere. Notice that

$$0 \leq |t| - \rho_\varepsilon(t) < \varepsilon \quad \forall\, t \in \mathbb{R}.$$

By the next proposition (which is stated in a general context), the above inequality implies that $\{m(f; I_i, \rho_\varepsilon)\}$ converges to $\mathrm{med}(f|I_i)$, as $\varepsilon \downarrow 0$, for all $f \in C(I_i)$. This limit justifies the use of $m(f; I_i, \rho_\varepsilon)$ as an approximation of the continuous median $\mathrm{med}(f|I_i)$.

PROPOSITION 4.4. *Let* $\rho : \mathbb{R} \to \mathbb{R}_+$ *be a convex function with a unique minimizer at zero such that* $\rho(0) = 0$*. For every* $\varepsilon > 0$*, let* ρ_ε *be a convex function with the same properties as* ρ*. Assume that there exists a constant* $\eta > 0$ *such that, for every* $\varepsilon > 0$ *sufficiently small,*

$$|\rho(t) - \rho_\varepsilon(t)| \leq \eta\, \varepsilon \quad \forall\, t \in \mathbb{R};$$

then, for every $f \in C(K)$*,*

$$\lim_{\varepsilon \downarrow 0} m(f; K, \rho_\varepsilon) = m(f; K, \rho).$$

Proof. For every $\varepsilon > 0$, there exists $x(\varepsilon) \in \mathrm{int}\ K$ such that $m(f; K, \rho_\varepsilon) = f(x(\varepsilon))$. Consequently, it follows that

$$\sup_{\varepsilon > 0} |m(f; K, \rho_\varepsilon)| < \infty.$$

To establish the desired limit, it suffices to show that for every sequence of positive scalars $\{\varepsilon_k\}$ converging to zero, the sequence $\{m(f;K,\rho_{\varepsilon_k})\}$ converges to $m(f;K,\rho)$. In turn, it suffices to show that if

$$\lim_{k(\in\kappa)\to\infty} m(f;K,\rho_{\varepsilon_k}) \,=\, m_\infty,$$

where κ is an infinite subset of $\{1,2,\dots\}$, then

$$\int_K \rho(f(x)-m)\,dx \,\geq\, \int_K \rho(f(x)-m_\infty)\,dx \tag{4.8}$$

for all $m\in\mathbb{R}$. Writing $m_k := m(f;K,\rho_{\varepsilon_k})$, we have, for every k,

$$\int_K \rho_{\varepsilon_k}(f(x)-m)\,dx \,\geq\, \int_K \rho_{\varepsilon_k}(f(x)-m_k)\,dx.$$

For all k sufficiently large, we have

$$\int_K \rho_{\varepsilon_k}(f(x)-m)\,dx \,\leq\, \int_K \rho(f(x)-m)\,dx + \eta\,\varepsilon_k \,\text{meas}\,(K)$$

and

$$\int_K \rho_{\varepsilon_k}(f(x)-m_k)\,dx \,\geq\, \int_K \rho(f(x)-m_k)\,dx - \eta\,\varepsilon_k \,\text{meas}\,(K).$$

Passing to the limit $k(\in\kappa)\to\infty$ readily yields the desired inequality (4.8). □

4.4. Method II applied to interpolation of medians and smoothed medians. We have implemented a MATLAB solver `MEstimatorInterp` that applies Method II in section 4.2 to the M-estimator interpolation problem. We shall report and compare some of the numerical results based on this solver in the next subsection. In the implementation, we evaluate a polynomial using its Lagrange form. Specifically, for a polynomial L_i in Π_n, we write, relative to the interval I_i,

$$L_i(x) \,:=\, \left[\prod_{j=0,j\neq i}^{n} (\,x-x_i\,)\right]\left[\prod_{j=0,j\neq i}^{n} (\,x_j-x_i\,)\right]^{-1}, \quad x_j \,=\, \text{midpoint of } I_j.$$

The numerical result pertains to the following:

(i) `MEstimatorInterp` applied to the smoothed median-interpolation problem. In this case

$$F_i(a) \,=\, \int_{I_i} \rho'\left(\sum_{i=0}^{n} a_i L_i(t) - b_i\right)\,dt, \quad \rho'(t) = \frac{t}{\sqrt{t^2+0.01}}. \tag{4.9}$$

(ii) `MEstimatorInterp` applied to the median-interpolation problem. In this case

$$F_i(a) \,=\, \int_{I_i} \text{sign}\left(\sum_{i=0}^{n} a_i L_i(t) - b_i\right)\,dt. \tag{4.10}$$

Since $\rho = |\cdot|$ does not satisfy assumptions in Theorem 4.3, there is no guarantee that the method would converge.

(iii) `MedianInterp` in [4, section 2.1.4] applied to the median-interpolation problem. Based on a fixed-point iteration, an implementation of this solver, whose convergence has not been established, is freely available in WaveLab 802 at http://www-stat.stanford.edu/~wavelab/.

(iv) Same as (ii), except that the initial guess is chosen to be the fourth iterate computed from the fixed-point method in (iii).

Here are some implementation details:

Initial guess. Except in case (iv), we use the midpoint interpolant of the data $p_0 := \sum_{i=0}^{n} b_i L_i$ as the initial guess. This is motivated by the heuristic

$$\text{med}(p|I) \;\approx\; p(\text{midpoint of } I) \;\approx\; m(p; I, \sqrt{(\cdot)^2 + \varepsilon^2} - \varepsilon).$$

Computing $F(a)$. The integrals (4.9) and (4.10) are computed using `quad()` in Matlab 6.1. This routine is based on a recursive adaptive Simpson quadrature. The tolerance is set to $\texttt{tol} = 10^{-12}$. The evaluations of $\sum_i a_i L_i(t)$ in the integrands of (4.9) and (4.10) are performed using Neville's algorithm.

Computing $JF(a)$. For the computation of the Jacobian of F in the case of (4.9), one can in principle use (4.7), which basically involves an integral with ρ'' appearing in the integrand. This approach, however, turns out to be quite problematic for the purpose of smoothed median interpolation, since in this application $\rho'(x) \approx \text{sign}(x)$ and $\rho''(x) \approx$ the dirac function. Thus we approximate $\partial F_i(a)/\partial a_j$ based on a central divided difference:

$$\frac{\partial F_i(a)}{\partial a_j} \approx \frac{F_i(a + he_j) - F_i(a - he_j)}{2h}, \quad h = 10^{-8} \approx \sqrt{\text{machine precision}}, \tag{4.11}$$

where e_j is the jth vector in the standard basis of $\mathbb{R}^{n+1}$. This bypasses the singularity of ρ'' at zero. The finite difference (4.11) is applicable to the case of the median-interpolation problem; however, we remind the reader that at the time this article is written it is not known whether Method II applied to (4.10) would enjoy any convergence property.

Stopping criterion. We terminate a Newton iteration when the number of steps in a Armijo line search exceeds an upper bound `MAX_LINE_SEARCH` (chosen to be 10 in `MEstimatorInterp`.)

4.5. Numerical experiments.

Experiment I. $n = 4$,$(b_0, b_1, b_2, b_3, b_4) = (-1.2, 1.3, -0.9, 1.0, -0.8)$.

Experiment II. $n = 6$, $(b_0, b_1, b_2, b_3, b_4, b_5, b_6) = (0.05, 0.7, 0.6, -0.25, -0.38, -0.3, -1.5)$.

In both experiments, we take $I_i = [i, i + 1]$. Each of (i)–(iv) in section 4.4 is applied to both datasets above; in the following discussion we label the corresponding subexperiments as I(i)–I(iv) and II(i)–II(iv), respectively. The left panels of Figure 1 depict the error curves $\log(||F(a^k)||)$ versus k; on the right panels we graph the midpoint, median, and smoothed median (with smoothing factor $\varepsilon = 0.1$) interpolants of the data.

While Method II applied to median interpolation converges in Experiment II(ii), the method clearly fails in Experiment I(ii). Nevertheless, the latter failure of convergence is remedied by Experiment I(iv). As mentioned, there is currently no theoretical guarantee that the Newton method with line search applied to the nonsmooth systems in (ii) would enjoy any global (or even local) convergence property. Moreover, one can see from Figure 1(a) and (c) that Newton Method II applied to the nonsmooth median interpolation does not seem to offer fast convergence.

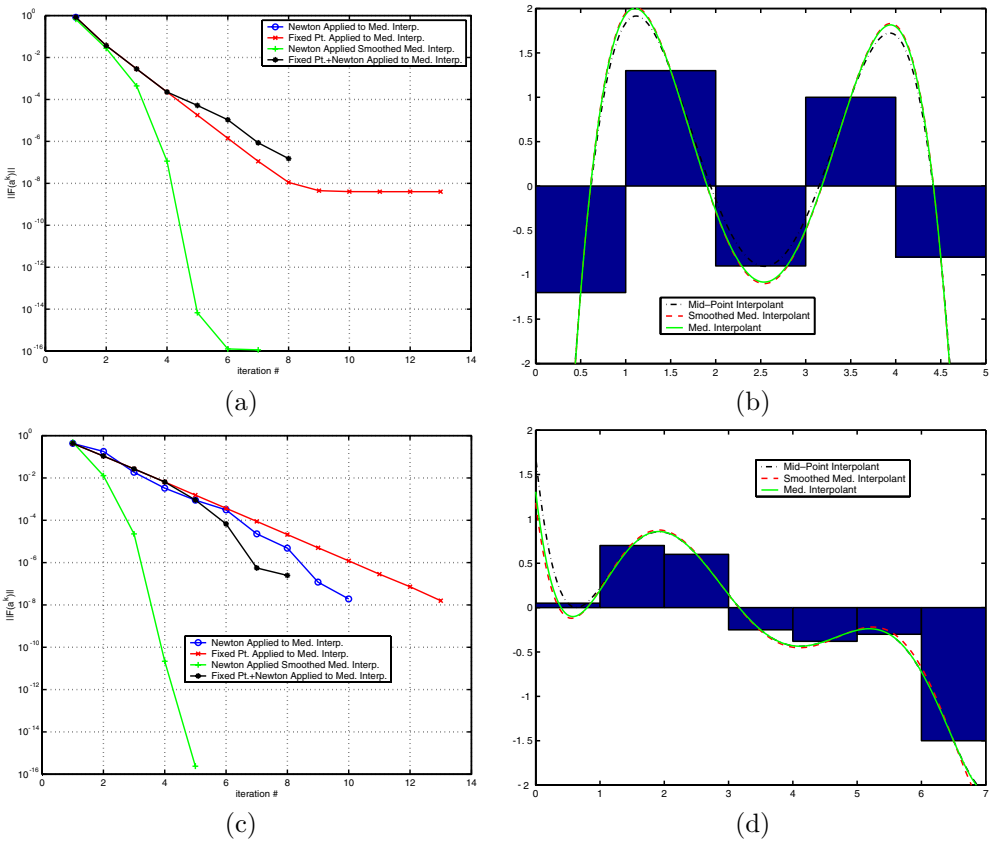


FIG. 1. *Error curves (left) and polynomial interpolants (right).*

On the other hand, Newton Method II applied to the smoothed median-interpolation problem exhibits superlinear convergence in Experiments I(i) and II(i), as expected from Theorem 4.3. As seen from Figure 1(a) and (c), the rates of convergence are noticeably faster than those in Experiment I(iii) and II(iii).

In summary the fixed-point algorithm used in `MedianInterp` for solving the median interpolation problem converges slower than Newton Method II applied to the smoothed median-interpolation problem. The latter method has a provable superlinear convergence property, while the former method has a conjectured linear convergence property. In terms of actual computational speed, one step of Newton iteration in our current implementation of `MEstimatorInterp` is much slower than one step of fixed-point iteration in `MedianInterp`; this is to be expected. Indeed, one can definitely improve the implementation of `MEstimatorInterp`.

During the revision of this article, we received the preprint by Qi [9], who studies the smoothness properties of $\mathbf{M}_\rho$ and F_ρ defined in section 4.1 and 4.2 in the case of the Huber M-Estimator (i.e., when ρ is given by (2.3).) The same article also illustrates a fundamental difficulty in the numerical solution of the median-interpolation problem: specifically, it is shown in section 4 of [9] that $F_{|\cdot|}$, unlike $M_{|\cdot|}$, is in general not even locally Lipschitz. This supports the slow convergence observed empirically in Experiments I(ii),(iv) and II(ii),(iv).

5. Nonlinear pyramid transforms based on M-estimators. Following the same strategy proposed in [3], one can now use the results in this paper to construct new nonlinear pyramid transforms: for a given signal $f : \mathbb{R} \to \mathbb{R}$, one measures the M-estimators of f over dyadic intervals $I_{j,k} = [2^{-j}k, 2^{-j}(k+1)]$, performs coarse-to-fine prediction using local polynomial interpolation, which has been shown to be a well-posed procedure by Theorem 3.1, and then defines the pyramid coefficients based on the errors of such predictions.

Nonlinear subdivision operator. We first define a new class of nonlinear subdivision operators which will serve the purpose of coarse-to-fine prediction. Denote by $l(\mathbb{Z})$ the vector space of all real sequences defined on $\mathbb{Z}$. Let ρ be a convex function with the standard assumptions. Let $L \geq 1$ be an integer. For a given $y \in l(\mathbb{Z})$ and $h > 0$, define $S_{\rho,L;h}(y) \in l(\mathbb{Z})$ as follows.

1. Interpolation: for each $k \in \mathbb{Z}$, let $p_k \in \Pi_{2L}$ be the unique polynomial such that
$$m(p_k; [h(k+l), h(k+l+1)], \rho) = y_{k+l}, \quad l = -L, \dots, L.$$
2. Imputation:
$$\begin{aligned} (S_{\rho,L;h}(y))_{2k} &:= m(p_k; [hk, h(k+1/2)], \rho), \\ (S_{\rho,L;h}(y))_{2k+1} &:= m(p_k; [h(k+1/2), h(k+1)], \rho), \quad k \in \mathbb{Z}. \end{aligned}$$

Since $m(f; K, \rho) = m(f(T\cdot); T^{-1}(K), \rho)$ for any invertible affine map T, the operator $S_{\rho,L;h}$ is independent of the scale parameter h; thus we drop the subscript h and write $S_{\rho,L} : l(\mathbb{Z}) \to l(\mathbb{Z})$. It is worth mentioning that Theorem 3.1 implies that $S_{\rho,L}$ is a bounded operator on $l^\infty(\mathbb{Z})$.

Pyramid transform. For a continuous signal $f : \mathbb{R} \to \mathbb{R}$, a ρ, and an integer $L \geq 1$, we define an M-estimator interpolating pyramid transform of f, denoted by $\textsc{MeIPT}(f; L, \rho)$, as follows.

1. Formation of M-estimators of f over dyadic blocks:
$$m_{j,k} := m(f; I_{j,k}, \rho), \quad j = j_0, j_0+1, \dots, \ k \in \mathbb{Z}.$$
2. Coarse-to-fine prediction: $\widetilde{m}_{j+1} = S_{\rho,L}((m_{j,k})_k)$.
3. Formation of detailed coefficients: $d_{j,k} = m_{j,k} - \widetilde{m}_{j,k}$.

$$\textsc{MeIPT}(f; L, \rho) := \{(m_{0,k})_{k\in\mathbb{Z}}, (d_{1,k})_{k\in\mathbb{Z}}, (d_{2,k})_{k\in\mathbb{Z}}, \dots\}.$$

There is also an inversion process for recovering f from $\textsc{MeIPT}(f; L, \rho)$ based on sequentially reversing the steps above. For this purpose we need also the observation that

$$\lim_{j\to\infty} \left\| f - \sum_{k\in\mathbb{Z}} m(f; I_{j,k}, \rho) 1_{[2^{-j}k, 2^{-j}(k+1))} \right\|_{L^\infty} = 0$$

when f is a bounded continuous function.

For $j = j_0, j_0+1, j_0+2, \ldots,$

1. $f_j := \sum_{k\in\mathbb{Z}} m_{j,k} 1_{[2^{-j}k, 2^{-j}(k+1))}$.
2. Coarse-to-fine prediction: $\widetilde{m}_{j+1} = S_{\rho,L}((m_{j,k})_k)$.
3. Recovery of scale $j+1$ M-estimators of f: $m_{j+1,k} = \widetilde{m}_{j+1,k} + d_{j+1,k}, \quad k \in \mathbb{Z}$.

$$(f_j)_j \to f \text{ uniformly on compact sets.} \tag{5.1}$$

Discussion. In practice, however, one typically recovers an *estimate* of f from a certain *perturbed* version of MeIPT$(f; L, \rho)$; in this case the convergence in (5.1) has to be reexamined, and also the *stability* of MeIPT becomes a very important issue. While these open problems are beyond the scope of the current paper, based on Theorem 3.1 we expect MeIPT to be a decent tool for signal compression. Following the formulation and arguments in [3, Proof of P3, p. 1055], the transform coefficients in MeIPT(f) can be shown to have good *sparsity* when f is piecewise smooth; and the sparsity improves gracefully as the smoothness of f improves. Such a property is attributable to the accurate coarse-to-fine prediction power of the operator $S_{\rho,L}$, thanks to a polynomial exactness property of $S_{\rho,L}$ guaranteed by Theorem 3.1.

Robust pyramid transform based on smoothed medians. Combining the ideas in this and the last section, one can now construct MeIPT based on smoothed continuous medians. Such pyramid transforms had been found experimentally to be as robust against outliers as the median-interpolating pyramid transforms considered in [3]. This is to be expected by virtue of Proposition 4.4.

Acknowledgments. The authors are grateful to Dan Naiman for his discussion in the initial phase of the work. The authors would like to thank two referees for constructive comments on the original version of the paper, which have led to an improvement of the presentation.

REFERENCES

[1] D.P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.
[2] J.E. Dennis and R.E. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983.
[3] D.L. Donoho and T. P.-Y. Yu, *Nonlinear pyramid transforms based on median-interpolation*, SIAM J. Math. Anal., 31 (2000), pp. 1030–1061.
[4] D.L. Donoho and T. P.-Y. Yu, *Nonlinear "Wavelet Transforms" Based on Median-Interpolation*, Technical report, Department of Statistics, Stanford University, Stanford, CA, 1997. Available online from http://stat.stanford.edu/reports/donoho/median.ps.Z.
[5] F. Facchinei and J.S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.
[6] T.N.T. Goodman and T. P.-Y. Yu, *Interpolation of medians*, Adv. Comput. Math., 11 (1999), pp. 1–10.
[7] P.J. Huber, *Robust Statistics*, John Wiley and Sons, New York, 1981.
[8] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[9] L. Qi, *Analytic Properties of the Huber and the Median M-Estimator Interpolation Problems*, preprint, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, 2003.
[10] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[11] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1018–1031

ON TIME-DOMAIN SIMULATION OF LOSSLESS TRANSMISSION LINES WITH NONLINEAR TERMINATIONS*

YAO-LIN JIANG†

Abstract. A time-domain approach is presented to solve nonlinear circuits with lossless transmission lines. Mathematically, the circuits are described by a special kind of nonlinear differential-algebraic equations (DAEs) with multiple constant delays. In order to directly compute these delay systems in time-domain, decoupling by waveform relaxation (WR) is applied to the systems. For the relaxation-based method we provide a new convergence proof. Numerical experiments are given to illustrate the novel approach.

Key words. nonlinear circuits, transmission lines, differential-algebraic equations with multiple delays, waveform relaxation, circuit simulation

AMS subject classifications. 65L80, 65Q05, 68W35, 65Y05, 37M05

DOI. 10.1137/S0036142902418886

1. Introduction.

A recent advance in VLSI technology has led to the development of high-speed integrated circuits in which the conductors must be regarded as transmission lines [1, 2]. At low bit rates, the lines may be considered as lossless.

The simulation task is to compute the transient response of a circuit consisting of nonlinear devices interconnected by transmission lines. These lines cause delay of the signals, and if they are terminated in nonlinear devices, they usually cause reflections as well.

In this paper we report a new method to simulate a circuit with lossless transmission lines. The circuit equations are nonlinear differential-algebraic equations (DAEs) with multiple constant delays. In time-domain waveform relaxation (WR) decoupling is applied to the system. The decoupled subsystems are some standard ordinary differential equations (ODEs) and algebraic equations (AEs). The general-purpose circuit simulators can be used to solve the subsystems. WR is a parallel process of computing transient solutions for large systems in time-domain [3, 4, 5, 6, 7]. Recent works on WR convergence for DAEs are reported in [8, 9, 10, 11, 12]. For a standard system of ODEs with usual delay, the papers [13, 14] considered simple WR methods. However, none of the works can be directly applied to this special kind of DAE with multiple delays in simulation of transmission lines. In this paper we provide a detailed proof of the proposed method. Three illustrative examples are also presented.

2. Models of lossless transmission lines.

In general, a transmission line is described by telegrapher's equations. For a lossless transmission line system shown in Figure 1, at time t $(0 \le t \le T_e)$ let $v(x,t)$ and $i(x,t)$ respectively be voltage and

*Received by the editors December 2, 2002; accepted for publication (in revised form) July 3, 2003; published electronically July 29, 2004. This research was supported by the Natural Science Foundation of China NSFC 10171080, the 863 Program of China 2001AA111042, and the Excellent Young Teachers Program of MOE 1887, P. R. China.
http://www.siam.org/journals/sinum/42-3/41888.html
†Institute of Information and System Sciences, School of Science, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, People's Republic of China (yljiang@mail.xjtu.edu.cn).

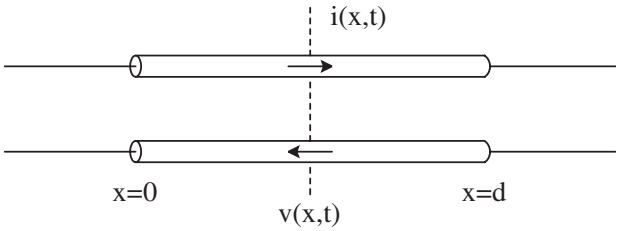


FIG. 1. *A segment of lossless transmission line.*

current at point x $(0 \le x \le d)$. The basic equations are

$$\begin{cases} \dfrac{\partial v(x,t)}{\partial x} = -L\dfrac{\partial i(x,t)}{\partial t}, \\ \dfrac{\partial i(x,t)}{\partial x} = -C\dfrac{\partial v(x,t)}{\partial t}, \end{cases} \tag{1}$$

where L is inductance and C is capacitance for unit length.

Differentiate the first part of (1) with respect to x and the second part of (1) with respect to t, and then combine them to obtain

$$\frac{\partial^2 v(x,t)}{\partial x^2} = LC\frac{\partial^2 v(x,t)}{\partial t^2}. \tag{2}$$

Similarly,

$$\frac{\partial i(x,t)}{\partial x^2} = LC\frac{\partial^2 i(x,t)}{\partial t^2}. \tag{3}$$

Let $LC = 1/\nu^2$, where ν is velocity of signal propagation. The solution of (2) and (3) has the form

$$\begin{cases} v(x,t) = f\left(t - \dfrac{x}{\nu}\right) + g\left(t + \dfrac{x}{\nu}\right), \\ i(x,t) = \dfrac{1}{z_0} f\left(t - \dfrac{x}{\nu}\right) - \dfrac{1}{z_0} g\left(t + \dfrac{x}{\nu}\right), \end{cases} \tag{4}$$

where $z_0 = \sqrt{L/C}$. By (4), we get

$$\begin{cases} f\left(t - \dfrac{x}{\nu}\right) = \dfrac{1}{2}[v(x,t) + z_0 i(x,t)], \\ g\left(t + \dfrac{x}{\nu}\right) = \dfrac{1}{2}[v(x,t) - z_0 i(x,t)]. \end{cases} \tag{5}$$

Let $\tau = \frac{d}{\nu}$, which is the delay of a signal going from $x = 0$ to $x = d$. Now at $x = d$ we have $2v(d,t) = [v(0, t-\tau) + z_0 i(0, t-\tau)] + [v(d,t) - z_0 i(d,t)]$ or

$$v(d,t) = v(0, t-\tau) + z_0 i(0, t-\tau) - z_0 i(d,t). \tag{6}$$

At $x = 0$, we have $2v(0,t) = [v(0,t) + z_0 i(0,t)] + [v(d, t-\tau) - z_0 i(d, t-\tau)]$ or

$$v(0,t) = v(d, t-\tau) - z_0 i(d, t-\tau) + z_0 i(0,t). \tag{7}$$

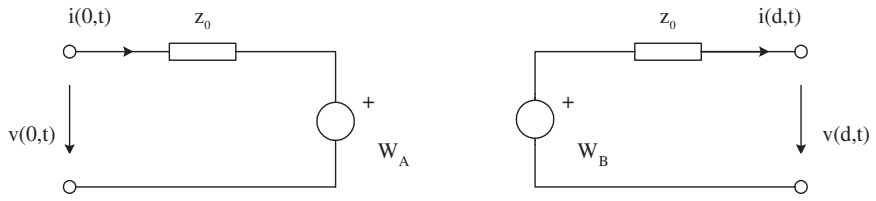


FIG. 2. *Characteristic 2-port of lossless transmission line system.*

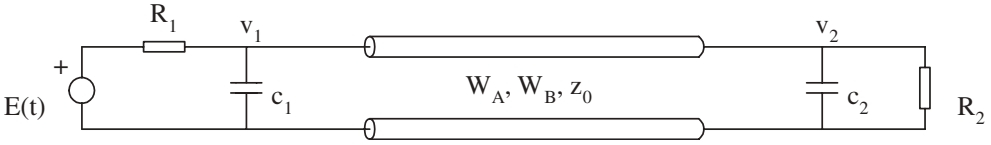


FIG. 3. *A linear circuit with transmission lines.*

Define two quantities

$$\begin{cases} W_A(t-\tau) = v(d,t-\tau) - z_0 i(d,t-\tau), \\ W_B(t-\tau) = v(0,t-\tau) + z_0 i(0,t-\tau). \end{cases} \tag{8}$$

Then (6) and (7) become

$$\begin{cases} v(0,t) = z_0 i(0,t) + W_A(t-\tau), \\ v(d,t) = -z_0 i(d,t) + W_B(t-\tau). \end{cases} \tag{9}$$

By (8) and (9), we have

$$\begin{cases} W_A(t) = 2v(d,t) - W_B(t-\tau), \\ W_B(t) = 2v(0,t) - W_A(t-\tau). \end{cases} \tag{10}$$

The transmission line is therefore characterized by

$$\begin{cases} i(0,t) = \dfrac{1}{z_0} v(0,t) - \dfrac{1}{z_0} - W_A(t-z), \\ 1pti(d,t) = \dfrac{1}{-z_0} v(d,t) + \dfrac{1}{z_0} + W_B(t-z). \end{cases} \tag{11}$$

Its equivalent circuit is shown in Figure 2.

For a linear circuit with the lossless transmission line system given in Figure 3, the circuit equations are

$$\begin{cases} c_1 \dfrac{dv_1(t)}{dt} = \dfrac{E(t)}{R_1} - \left(\dfrac{1}{R_1} + \dfrac{1}{z_0} \right) v_1(t) + \dfrac{1}{z_0} W_A(t-\tau), \\ c_2 \dfrac{dv_2(t)}{dt} = -\left(\dfrac{1}{R_2} + \dfrac{1}{z_0} \right) v_2(t) + \dfrac{1}{z_0} W_B(t-\tau), \\ W_A(t) = 2v_2(t) - W_B(t-\tau), \quad W_B(t) = 2v_1(t) - W_A(t-\tau). \end{cases} \tag{12}$$

Its equivalent circuit is also shown in Figure 4.

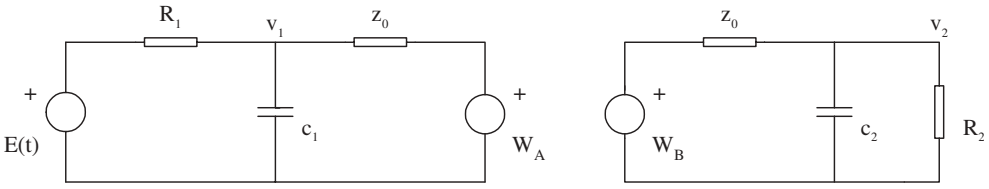


FIG. 4. *An equivalent circuit for a distributed network.*

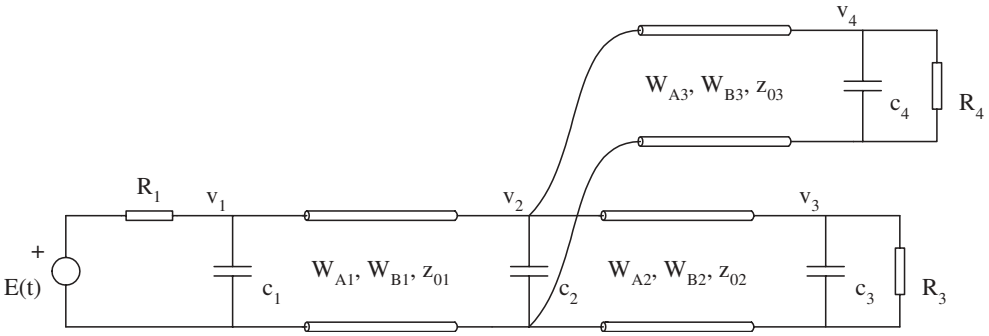


FIG. 5. *A linear circuit with multiple transmission lines.*

Another distributed circuit is given in Figure 5; its circuit equations are described by DAEs with multiple constant delays as follows:

$$
\begin{cases}
c_1\dfrac{dv_1(t)}{dt} = -\left(\dfrac{1}{R_1}+\dfrac{1}{z_{01}}\right)v_1(t) + \dfrac{1}{z_{01}}W_{A1}(t-\tau_1) + \dfrac{E(t)}{R_1}, \\
c_2\dfrac{dv_2(t)}{dt} = -\left(\dfrac{1}{z_{01}}+\dfrac{1}{z_{02}}+\dfrac{1}{z_{03}}\right)v_2(t) + \dfrac{1}{z_{01}}W_{B1}(t-\tau_1) + \dfrac{1}{z_{02}}W_{A2}(t-\tau_2) \\
\qquad\qquad + \dfrac{1}{z_{03}}W_{A3}(t-\tau_3), \\
c_3\dfrac{dv_3(t)}{dt} = -\left(\dfrac{1}{R_3}+\dfrac{1}{z_{02}}\right)v_3(t) + \dfrac{1}{z_{02}}W_{B2}(t-\tau_2), \\
c_4\dfrac{dv_4(t)}{dt} = -\left(\dfrac{1}{R_4}+\dfrac{1}{z_{03}}\right)v_4(t) + \dfrac{1}{z_{03}}W_{B3}(t-\tau_3), \\
W_{A1}(t) = 2v_2(t) - W_{B1}(t-\tau_1), \quad W_{B1}(t) = 2v_1(t) - W_{A1}(t-\tau_1), \\
W_{A2}(t) = 2v_3(t) - W_{B2}(t-\tau_2), \quad W_{B2}(t) = 2v_2(t) - W_{A2}(t-\tau_2), \\
W_{A3}(t) = 2v_4(t) - W_{B3}(t-\tau_3), \quad W_{B3}(t) = 2v_2(t) - W_{A3}(t-\tau_3).
\end{cases}
\tag{13}
$$

The systems of (12) and (13) have no standard form of ODEs with finite delays as studied in [13, 14]. To clearly understand this point, we may see the case of (12).

First, from the algebraic part of (12) we know that

$$
\begin{aligned}
W_A(t) &= 2v_2(t) - W_B(t-\tau) \\
&= 2v_2(t) - 2v_1(t-\tau) + W_A(t-2\tau)
\end{aligned}
\tag{14}
$$

and

$$
\begin{aligned}
W_B(t) &= 2v_1(t) - W_A(t-\tau) \\
&= 2v_1(t) - 2v_2(t-\tau) + W_B(t-2\tau).
\end{aligned}
\tag{15}
$$

Since $W_A(t)$ and $W_B(t)$ are defined on $[-\tau, T_e]$, where $\tau < +\infty$, we cannot rewrite (12) as a system of ODEs with delay. If the functions in (12) are defined on $(-\infty, T_e]$, by (14) and (15), $W_A(t-\tau)$ and $W_B(t-\tau)$ can be expressed as

$$W_A(t-\tau) = 2\sum_{j=1}^{\infty}[v_2(t-(2j-1)\tau) - v_1(t-2j\tau)] \tag{16}$$

and

$$W_B(t-\tau) = 2\sum_{j=1}^{\infty}[v_1(t-(2j-1)\tau) - v_2(t-2j\tau)]. \tag{17}$$

Substituting (16) and (17) into the differential part of (12), we have

$$\begin{cases} c_1\dfrac{dv_1(t)}{dt} = \dfrac{E(t)}{R_1} - \left(\dfrac{1}{R_1}+\dfrac{1}{z_0}\right)v_1(t) + \dfrac{2}{z_0}\displaystyle\sum_{j=1}^{\infty}[v_2(t-(2j-1)\tau) - v_1(t-2j\tau)], \\ c_2\dfrac{dv_2(t)}{dt} = -\left(\dfrac{1}{R_2}+\dfrac{1}{z_0}\right)v_2(t) + \dfrac{2}{z_0}\displaystyle\sum_{j=1}^{\infty}[v_1(t-(2j-1)\tau) - v_2(t-2j\tau)]. \end{cases} \tag{18}$$

However, it is now a system with infinite delays. The convergence conditions of WR in [13, 14] are still not suitable for simulation of transmission lines. This also further illustrates that for this kind of simulation problems it is better to directly use DAEs with delay.

3. WR of DAEs with multiple delays. Let us study the following circuit system of DAEs with multiple constant delays:

$$\begin{cases} C(t)\dfrac{dx(t)}{dt} + G(x(t),t) + DW(t-\tau) = b(t), \\ A(t)x(t) + W(t) + BW(t-\tau) = 0, \\ x(0) = x_0, \quad W(\theta) \equiv \varphi(\theta) \quad (-\tau \le \theta < 0), \quad t \in [0, T_e], \end{cases} \tag{19}$$

where $C(\cdot), A(\cdot)$ are two matrix-valued functions in $\mathbf{R}^{n\times n}$ and $\mathbf{R}^{2m\times n}$; D, B are two constant matrices in $\mathbf{R}^{n\times 2m}$ and $\mathbf{R}^{2m\times 2m}$; $G(\cdot,\cdot)$ is a nonlinear function; and for any t the functions $x(t) \in \mathbf{R}^{n\times 1}$, $W(t-\tau) \in \mathbf{R}^{2m\times 1}$ are to be computed, in which

$$W(t-\tau) = [y_1(t-\tau_1), z_1(t-\tau_1), \ldots, y_m(t-\tau_m), z_m(t-\tau_m)]^t. \tag{20}$$

We also define $\tau' = \min_{1\le i\le m}\{\tau_i\} (> 0)$.

For the above circuit system, $b(\cdot)$ is a known input function, x_0 is an initial value, and $\varphi(\theta)$ is an initial state of the transmission line system such that

$$\varphi(\theta) = [y_1(\theta_1), z_1(\theta_1), \ldots, y_m(\theta_m), z_m(\theta_m)]^t, \tag{21}$$

in which $-\tau_i \le \theta_i < 0$ $(1 \le i \le m)$. We also assume that the initial values x_0, $W_0 (= W(0))$ are consistent; that is, $A(0)x_0 + W_0 + BW(-\tau) = 0$. Further, by invoking the above characteristic of transmission lines, we assume that the form of B in (19) is a block diagonal matrix such that

$$B = \begin{bmatrix} I_d & & 0 \\ & \ddots & \\ 0 & & I_d \end{bmatrix} \in \mathbf{R}^{2m\times 2m}, \tag{22}$$

where $I_d = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$.

In (19), we assume that $C(\cdot)$, $\frac{dC(\cdot)}{dt}$, $A(\cdot)$, and $\frac{\partial G}{\partial x}$ are continuous. Moreover, $C(t)$ for all $t \in [0, T_e]$ are nonsingular and $\frac{\partial G}{\partial x}$ is bounded. In this paper we always assume that (19) has a unique solution for given initial value and state.

Let $F : (\mathbf{R}^n)^2 \times [0, T_e] \mapsto \mathbf{R}^n$ be a splitting function of G which satisfies

$$F(u, u, t) = G(u, t), \quad t \in [0, T_e], \tag{23}$$

where $u \in \mathbf{R}^n$. The WR decoupling of (19) is

$$\begin{cases} C_1(t)\dfrac{dx^{(k)}(t)}{dt} + F(x^{(k)}(t), x^{(k-1)}(t), t) = C_2(t)\dfrac{dx^{(k-1)}(t)}{dt} - DW^{(k-1)}(t-\tau) + b(t), \\ A_1(t)x^{(k)}(t) + W^{(k)}(t) = A_2(t)x^{(k-1)}(t) - BW^{(k-1)}(t-\tau), \\ x^{(k)}(0) = x_0, \quad W^{(k)}(\theta) \equiv \varphi(\theta) \quad (-\tau \le \theta < 0), \quad t \in [0, T_e], \quad k = 1, 2, \ldots, \end{cases} \tag{24}$$

where $C(\cdot) = C_1(\cdot) - C_2(\cdot)$, $A(\cdot) = A_1(\cdot) - A_2(\cdot)$, and $x^{(0)}, W^{(0)}$ are two initial guesses. Similarly, for $i = 1, 2$ we assume that $C_i(\cdot)$, $\frac{dC_i(\cdot)}{dt}$, $A_i(\cdot)$, and $\frac{\partial F}{\partial u_i}$ are continuous, where $\frac{\partial F}{\partial u_i}$ $(i = 1, 2)$ are the partial derivatives of $F(x^{(k)}(\cdot), x^{(k-1)}(\cdot), \cdot)$ with respect to the first two arguments. Moreover, we also assume that $C_1(t)$ for all $t \in [0, T_e]$ are nonsingular and $\frac{\partial F}{\partial u_i}$ $(i = 1, 2)$ are bounded. By the relaxation-based method (24), equation (19) is now decoupled into two independent subsystem ODEs and AEs for any fixed k.

Let the function pair (x, W) be the exact solution of (19). We define $\epsilon^{(l)}(\cdot) = x^{(l)}(\cdot) - x(\cdot)$ and $\delta^{(l)}(\cdot) = W^{(l)}(\cdot) - W(\cdot)$ for an index l such that $\epsilon^{(l)}(0) = 0$ and $\delta^{(l)}(\theta) \equiv 0$, where $-\tau \le \theta < 0$. By (19) and (24), according to the known mean theorem of vector functions we have

$$\begin{cases} C_1(t)\dfrac{d\epsilon^{(k)}(t)}{dt} + \dfrac{\partial F}{\partial u_1}\epsilon^{(k)}(t) + \dfrac{\partial F}{\partial u_2}\epsilon^{(k-1)}(t) = C_2(t)\dfrac{d\epsilon^{(k-1)}(t)}{dt} - D\delta^{(k-1)}(t-\tau), \\ A_1(t)\epsilon^{(k)}(t) + \delta^{(k)}(t) = A_2(t)\epsilon^{(k-1)}(t) - B\delta^{(k-1)}(t-\tau), \\ \epsilon^{(k)}(0) = 0, \quad \delta^{(k)}(\theta) \equiv 0 \quad (-\tau \le \theta < 0), \quad t \in [0, T_e], \quad k = 1, 2, \ldots. \end{cases} \tag{25}$$

We write the differential part of (25) as

$$\begin{cases} \dfrac{d\epsilon^{(k)}(t)}{dt} = -C_1^{-1}(t)\dfrac{\partial F}{\partial u_1}\epsilon^{(k)}(t) + C_1^{-1}(t)C_2(t)\dfrac{d\epsilon^{(k-1)}(t)}{dt} - C_1^{-1}(t)\dfrac{\partial F}{\partial u_2}\epsilon^{(k-1)}(t) \\ \qquad\qquad - C_1^{-1}(t)D\delta^{(k-1)}(t-\tau), \\ \epsilon^{(k)}(0) = 0, \quad t \in [0, T_e], \quad k = 1, 2, \ldots. \end{cases} \tag{26}$$

We now denote $X(\cdot) = -C_1^{-1}(\cdot)\frac{\partial F}{\partial u_1}$ and let the matrix-valued function $\Phi(\cdot)$ satisfy

$$\begin{cases} \dfrac{d\Phi(t)}{dt} = X(t)\Phi(t), \\ \Phi(0) = I, \quad t \in [0, T_e]. \end{cases} \tag{27}$$

Using the fundamental solution matrix $\Phi(\cdot)$ and doing some calculations as in [15],

for $t \in [0, T_e]$ we have

$$\begin{aligned}\epsilon^{(k)}(t) &= C_1^{-1}(t)C_2(t)\epsilon^{(k-1)}(t)\\ &- \Phi(t)\int_0^t \Phi^{-1}(s)\left[C_1^{-1}(s)\frac{\partial F}{\partial u_1}C_1^{-1}(s)C_2(s) + C_1^{-1}(s)\frac{\partial F}{\partial u_2} + \frac{d(C_1^{-1}C_2)(s)}{ds}\right]\epsilon^{(k-1)}(s)ds\\ &- \Phi(t)\int_0^t \Phi^{-1}(s)C_1^{-1}(s)D\delta^{(k-1)}(s-\tau)ds.\end{aligned} \tag{28}$$

For a positive parameter λ, multiplying both sides of (28) by $e^{-\lambda t}$, we have

$$\begin{aligned}e^{-\lambda t}\epsilon^{(k)}(t) &= C_1^{-1}(t)C_2(t)(e^{-\lambda t}\epsilon^{(k-1)}(t))\\ &- e^{-\lambda t}\Phi(t)\int_0^t e^{\lambda s}\Phi^{-1}(s)\left[C_1^{-1}(s)\frac{\partial F}{\partial u_1}C_1^{-1}(s)C_2(s) + C_1^{-1}(s)\frac{\partial F}{\partial u_2} + \frac{d(C_1^{-1}C_2)(s)}{ds}\right]\\ &\qquad \times (e^{-\lambda s}\epsilon^{(k-1)}(s))ds\\ &- e^{-\lambda t}\Phi(t)\int_0^t e^{\lambda s}\Phi^{-1}(s)C_1^{-1}(s)DJ(\tau)(e^{-\lambda(s-\tau)}\delta^{(k-1)}(s-\tau))ds, \quad t \in [0, T_e],\end{aligned} \tag{29}$$

where

$$\begin{aligned}&e^{-\lambda(s-\tau)}\delta^{(k-1)}(s-\tau)\\ = \;& [e^{-\lambda(s-\tau_1)}(y_1^{(k-1)}(s-\tau_1) - y_1(s-\tau_1)), e^{-\lambda(s-\tau_1)}(z_1^{(k-1)}(s-\tau_1) - z_1(s-\tau_1)), \ldots,\\ & e^{-\lambda(s-\tau_m)}(y_m^{(k-1)}(s-\tau_m) - y_m(s-\tau_m)), e^{-\lambda(s-\tau_m)}(z_m^{(k-1)}(s-\tau_m) - z_m(s-\tau_m))]^t\end{aligned}$$

and

$$J(\tau) = \begin{bmatrix} e^{-\lambda\tau_1}I & & 0\\ & \ddots & \\ 0 & & e^{-\lambda\tau_m}I\end{bmatrix} \in \mathbf{R}^{2m\times 2m},$$

in which $I = \left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]$.

All functions appearing in (29) are bounded due to the stated assumptions. There are two positive constants K_1 and K_2 such that

$$\begin{aligned}\|e^{-\lambda t}\epsilon^{(k)}(t)\| \;\le\; & \|C_1^{-1}(t)C_2(t)\|\|e^{-\lambda t}\epsilon^{(k-1)}(t)\| + e^{-\lambda t}K_1\int_0^t e^{\lambda s}\|e^{-\lambda s}\epsilon^{(k-1)}(s)\|ds\\ & + e^{-\lambda t}K_2\|J(\tau)\|\int_0^t e^{\lambda s}\|e^{-\lambda(s-\tau)}\delta^{(k-1)}(s-\tau)\|ds, \quad t \in [0, T_e],\end{aligned} \tag{30}$$

where the norm is adopted as one of $\|\cdot\|_p$, where $p = 1, 2, \infty$.

We define $\|u\|_t = \sup_{a\le s\le t}\{\|u(s)\|\}$ for a vector or matrix function $u(\cdot)$ on $[a, b]$. The function $\|u\|_t$ is monotonically increasing with respect to t on $[a, b]$. Because $\|J(\tau)\| = e^{-\lambda\tau'}$ and

$$e^{-\lambda t}\int_0^t e^{\lambda s}ds \le \frac{1}{\lambda}, \quad t \in [0, T_e],$$

we can rewrite (30) as

$$\begin{aligned}\|\epsilon^{(k)}\|_t \;&\le\; \|C_1^{-1}C_2\|_t\|\epsilon^{(k-1)}\|_t + \frac{K_1}{\lambda}\|\epsilon^{(k-1)}\|_t + \frac{K_2e^{-\lambda\tau'}}{\lambda}\|\delta^{(k-1)}\|_{t-\tau}\\ &\le\; \left(\|C_1^{-1}C_2\|_t + \frac{K_1}{\lambda}\right)\|\epsilon^{(k-1)}\|_t + \frac{K_2e^{-\lambda\tau'}}{\lambda}\|\delta^{(k-1)}\|_t, \quad t \in [0, T_e].\end{aligned} \tag{31}$$

We now write the algebraic part of (25) for a fixed k as

$$\delta^{(k)}(t) = -A_1(t)\epsilon^{(k)}(t) + A_2(t)\epsilon^{(k-1)}(t) - B\delta^{(k-1)}(t-\tau), \quad t \in [0, T_e]. \tag{32}$$

Using $e^{-\lambda t}$ to multiply its two sides, on $[0, T_e]$ we know that

$$e^{-\lambda t}\delta^{(k)}(t) = -A_1(t)(e^{-\lambda t}\epsilon^{(k)}(t)) + A_2(t)(e^{-\lambda t}\epsilon^{(k-1)}(t)) - e^{-\lambda\tau}B(e^{-\lambda(t-\tau)}\delta^{(k-1)}(t-\tau)), \tag{33}$$

where

$$e^{-\lambda\tau}B = \begin{bmatrix} e^{-\lambda\tau_1}I_d & & 0 \\ & \ddots & \\ 0 & & e^{-\lambda\tau_m}I_d \end{bmatrix} \in \mathbf{R}^{2m\times 2m}.$$

For $t \in [0, T_e]$, it follows that

$$\begin{array}{rcl} \|\delta^{(k)}\|_t & \le & \|A_1\|_t\|\epsilon^{(k)}\|_t + \|A_2\|_t\|\epsilon^{(k-1)}\|_t + \|e^{-\lambda\tau}B\|\|\delta^{(k-1)}\|_{t-\tau} \\ & \le & \|A_1\|_t\|\epsilon^{(k)}\|_t + \|A_2\|_t\|\epsilon^{(k-1)}\|_t + e^{-\lambda\tau'}\|\delta^{(k-1)}\|_t \end{array} \tag{34}$$

due to the fact $\|e^{-\lambda\tau}B\| = e^{-\lambda\tau'}$. Substituting (31) into the above inequality, we get

$$\begin{array}{rcl} \|\delta^{(k)}\|_t & \le & \left[\left(\|C_1^{-1}C_2\|_t + \dfrac{K_1}{\lambda}\right)\|A_1\|_t + \|A_2\|_t\right]\|\epsilon^{(k-1)}\|_t \\ & & + \left(e^{-\lambda\tau'} + \dfrac{K_2e^{-\lambda\tau'}}{\lambda}\|A_1\|_t\right)\|\delta^{(k-1)}\|_t, \quad t \in [0, T_e]. \end{array} \tag{35}$$

By (31) and (35), we have

$$\begin{bmatrix} \|\epsilon^{(k)}\|_{T_e} \\ \|\delta^{(k)}\|_{T_e} \end{bmatrix} \le M(1/\lambda) \begin{bmatrix} \|\epsilon^{(k-1)}\|_{T_e} \\ \|\delta^{(k-1)}\|_{T_e} \end{bmatrix}, \quad k = 1, 2, \ldots, \tag{36}$$

where

$$M(1/\lambda) = \begin{bmatrix} \|C_1^{-1}C_2\|_{T_e} + \frac{K_1}{\lambda} & \frac{K_2e^{-\lambda\tau'}}{\lambda} \\ (\|C_1^{-1}C_2\|_{T_e} + \frac{K_1}{\lambda})\|A_1\|_{T_e} + \|A_2\|_{T_e} & e^{-\lambda\tau'} + \frac{K_2e^{-\lambda\tau'}}{\lambda}\|A_1\|_{T_e} \end{bmatrix}. \tag{37}$$

It is a known conclusion that for any matrix $H(= (h_{ij}))$ its spectral radius $\rho(H)$ is a continuous function of h_{ij} for all i and j. Since the spectral radius of $M(0)$ is $\|C_1^{-1}C_2\|_{T_e}$, the spectral radius of $M(1/\lambda)$ will be less than one for some large λ if the quantity $\|C_1^{-1}C_2\|_{T_e}$ is less than one. Thus, we have proven the following theorem.

THEOREM 3.1. *For the nonlinear circuit system with lossless transmission lines* (19)*, its WR method* (24) *is convergent if*

$$\sup_{0\le t\le T_e} \|C_1^{-1}(t)C_2(t)\| < 1. \tag{38}$$

Let us denote $\alpha = \sup_{0\le t\le T_e} \|C_1^{-1}(t)C_2(t)\|$ and $\beta = \sup_{0\le t\le T_e} \|C_1^{-1}(t)D\|$. We rewrite the WR method (24) as

$$\begin{cases} \dfrac{dx^{(k)}(t)}{dt} = C_1^{-1}(t)C_2(t)\dfrac{dx^{(k-1)}(t)}{dt} - C_1^{-1}(t)F(x^{(k)}(t), x^{(k-1)}(t), t) \\ \qquad\qquad - C_1^{-1}(t)DW^{(k-1)}(t-\tau) + C_1^{-1}(t)b(t), \\ W^{(k)}(t) = -A_1(t)x^{(k)}(t) + A_2(t)x^{(k-1)}(t) - BW^{(k-1)}(t-\tau), \\ x^{(k)}(0) = x_0, \quad W^{(k)}(\theta) \equiv \varphi(\theta) \quad (-\tau \le \theta < 0), \quad t \in [0, T_e], \quad k = 1, 2, \ldots. \end{cases} \tag{39}$$

Now, we briefly define a basic 2×2 matrix as

$$N = \begin{bmatrix} \alpha & \beta \\ 0 & \|B\| \end{bmatrix}. \tag{40}$$

If one tries to use a known convergence result in [8, 12], the condition $\rho(N) < 1$ is necessary. However, under the assumption of Theorem 3.1 in the paper it is impossible because we know

$$\rho(N) = \max\{\alpha, \|B\|\} \tag{41}$$

and $\|B\| = 1$.

To speed up the WR convergence of (24), a windowing technique can be used for a long interval $[0, T_e]$. The decoupled subsystems by WR with windowing are separately solved on windows $[T_i, T_{i+1}]$ with time points T_i $(0 = T_0 < T_1 < \cdots < T_N = T_e)$. The computation of the WR solutions starts at T_0 and goes on from one window to another window until $T_N = T_e$ is reached. Namely,

$$\begin{cases} C_1(t)\dfrac{dx_i^{(k)}(t)}{dt} + F(x_i^{(k)}(t), x_i^{(k-1)}(t), t) = C_2(t)\dfrac{dx_i^{(k-1)}(t)}{dt} - DW_i^{(k-1)}(t-\tau) + b(t), \\ A_1(t)x_i^{(k)}(t) + W_i^{(k)}(t) = A_2(t)x_i^{(k-1)}(t) - BW_i^{(k-1)}(t-\tau), \quad t \in [T_i, T_{i+1}], \\ x_i^{(k)}(T_i) = x_{i-1}^{(k_{i-1})}(T_i), \quad k = 1, 2, \ldots, k_i, \quad i = 0, 1, \ldots, N-1, \end{cases} \tag{42}$$

where $x_{-1}^{(k_{-1})}(0) = x_0$, $W^{(l)}(t-\tau) \equiv \varphi(t-\tau)$ if $t \le \tau$, and the initial guesses $x_i^{(0)}(t) \equiv x_{i-1}^{(k_{i-1})}(T_i)$, $W_i^{(0)}(t-\tau) \equiv W_{i-1}^{(k_{i-1})}(T_i - \tau)$ on $[T_i, T_{i+1}]$ for all i.

4. Numerical experiments. Our numerical experiments are done for three simple circuits with transmission lines.

4.1. Example 1. The circuit is shown in Figure 3. For (12), its WR method is

$$\begin{cases} c_1\dfrac{dv_1^{(k)}(t)}{dt} + \left(\dfrac{1}{R_1} + \dfrac{1}{z_0}\right) v_1^{(k)}(t) = \dfrac{1}{z_0}W_A^{(k-1)}(t-\tau) + \dfrac{E(t)}{R_1}, \\ c_2\dfrac{dv_2^{(k)}(t)}{dt} + \left(\dfrac{1}{R_2} + \dfrac{1}{z_0}\right) v_2^{(k)}(t) = \dfrac{1}{z_0}W_B^{(k-1)}(t-\tau), \\ -2v_2^{(k)}(t) + W_A^{(k)}(t) = -W_B^{(k-1)}(t-\tau), \quad -2v_1^{(k)}(t) + W_B^{(k)}(t) = -W_A^{(k-1)}(t-\tau), \\ [v_1^{(k)}(0), v_2^{(k)}(0)]^t = [v_1(0), v_2(0)]^t, \\ [W_A^{(k)}(\theta), W_B^{(k)}(\theta)]^t \equiv [\varphi_A(\theta), \varphi_B(\theta)]^t \qquad (-\tau \le \theta < 0), \\ t \in [0, T_e], \quad k = 1, 2, \ldots. \end{cases} \tag{43}$$

We assume that the system starts from rest. The parameters are defined as $c_1 = c_2 = 10^{-3}$nF, $R_1 = 50\Omega$, $R_2 = 100\Omega$, $z_0 = 24.5\Omega$, $\tau = 0.244$ns, $T_e = 2.684$ns, and

$$E(t) = \begin{cases} \left(4 + \dfrac{t - 2 \times 10^{-10}}{2 \times 10^{-10}}\right) V, \quad 2 \times 10^{-10} < t \le 4 \times 10^{-10}; \\ 0 \quad \text{otherwise.} \end{cases}$$

The implicit Euler method, where the time-step is 0.0122ns, is taken to compute the decoupled ODEs. We define the iterative error as the sum of the squared differences of successive waveforms taken over all time points. The numerical results are given in Table 1. The far end voltage $v_2(t)$ is shown in Figure 6.

TABLE 1
WR of a linear circuit in Example 1.

Numbers of iterations	Iterative errors
1, 2, 3, 4, 5	$1.3820,\ \ 1.2130,\ \ 5.6123\times10^{-1},\ \ 2.7017\times10^{-1},\ \ 1.2604\times10^{-1}$
6, 7, 8, 9	$6.1703\times10^{-2},\ \ 2.9241\times10^{-2},\ \ 1.4721\times10^{-2},\ \ 6.5511\times10^{-3}$
10, 11	$1.9368\times10^{-3},\ \ 3.4737\times10^{-4}$

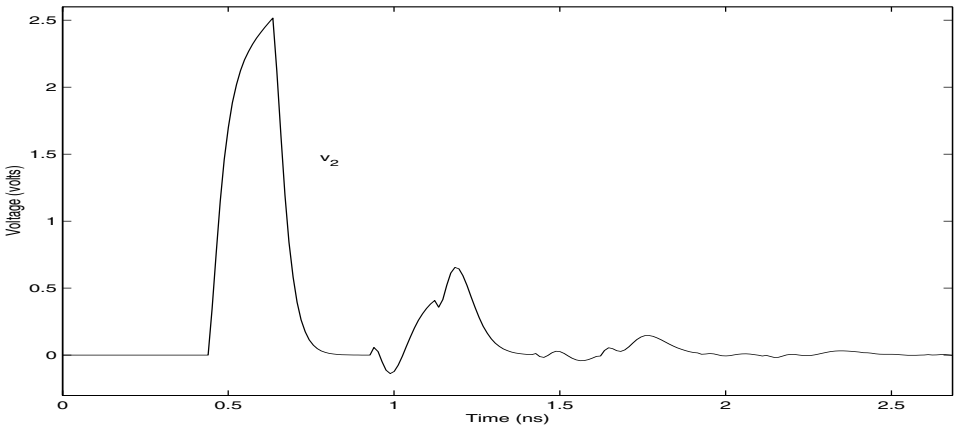


FIG. 6. *Far end voltage of a linear circuit in Example* 1.

4.2. Example 2. The circuit is shown in Figure 5. For (12), its WR method is

$$
\begin{cases}
c_1\dfrac{dv_1^{(k)}(t)}{dt}+\left(\dfrac{1}{R_1}+\dfrac{1}{z_{01}}\right)v_1^{(k)}(t)=\dfrac{1}{z_{01}}W_{A1}^{(k-1)}(t-\tau_1)+\dfrac{E(t)}{R_1},\\
c_2\dfrac{dv_2^{(k)}(t)}{dt}+\left(\dfrac{1}{z_{01}}+\dfrac{1}{z_{02}}+\dfrac{1}{z_{03}}\right)v_2^{(k)}(t)\ =\ \dfrac{1}{z_{01}}W_{B1}^{(k-1)}(t-\tau_1)+\dfrac{1}{z_{02}}W_{A2}^{(k-1)}(t-\tau_2)\\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad+\dfrac{1}{z_{03}}W_{A3}^{(k-1)}(t-\tau_3),\\
c_3\dfrac{dv_3^{(k)}(t)}{dt}+\left(\dfrac{1}{R_3}+\dfrac{1}{z_{02}}\right)v_3^{(k)}(t)=\dfrac{1}{z_{02}}W_{B2}^{(k-1)}(t-\tau_2),\\
c_4\dfrac{dv_4^{(k)}(t)}{dt}+\left(\dfrac{1}{R_4}+\dfrac{1}{z_{03}}\right)v_4^{(k)}(t)=\dfrac{1}{z_{03}}W_{B3}^{(k-1)}(t-\tau_3),\\
-2v_2^{(k)}(t)+W_{A1}^{(k)}(t)=-W_{B1}^{(k-1)}(t-\tau_1),\quad -2v_1^{(k)}(t)+W_{B1}^{(k)}(t)=-W_{A1}^{(k-1)}(t-\tau_1),\\
-2v_3^{(k)}(t)+W_{A2}^{(k)}(t)=-W_{B2}^{(k-1)}(t-\tau_2),\quad -2v_2^{(k)}(t)+W_{B2}^{(k)}(t)=-W_{A2}^{(k-1)}(t-\tau_2),\\
-2v_4^{(k)}(t)+W_{A3}^{(k)}(t)=-W_{B3}^{(k-1)}(t-\tau_3),\quad -2v_2^{(k)}(t)+W_{B3}^{(k)}(t)=-W_{A3}^{(k-1)}(t-\tau_3),\\
[v_1^{(k)}(0),v_2^{(k)}(0),v_3^{(k)}(0),v_4^{(k)}(0)]^t=[v_1(0),v_2(0),v_3(0),v_4(0)]^t,\\
[W_{A1}^{(k)}(\theta_1),W_{B1}^{(k)}(\theta_1),\ldots,W_{A3}^{(k)}(\theta_3),W_{B3}^{(k)}(\theta_3)]^t\\
\qquad\equiv[\varphi_{A1}(\theta_1),\varphi_{B1}(\theta_1),\ldots,\varphi_{A3}(\theta_3),\varphi_{B3}(\theta_3)]^t,\\
-\tau_i\le\theta_i\le 0,\quad 1\le i\le 3,\quad t\in[0,T_e],\quad k=1,2,\ldots.
\end{cases}
\tag{44}
$$

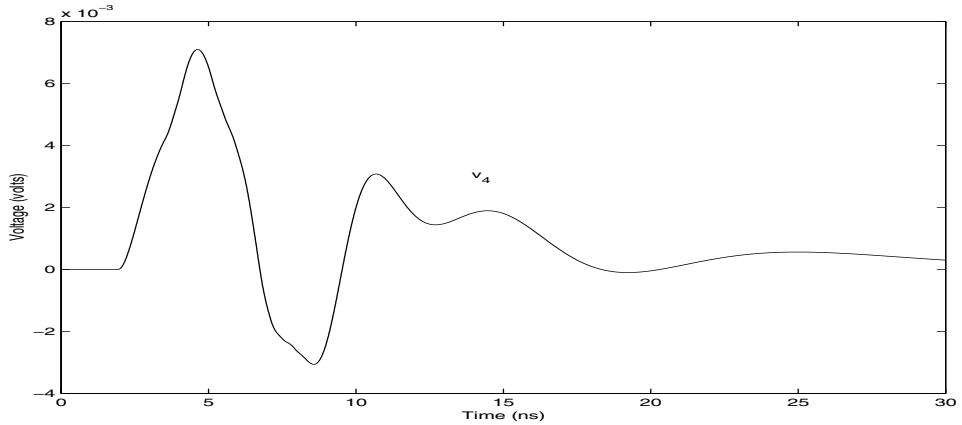


FIG. 7. *Far end voltage of a circuit with multiple transmission lines in Example* 2.

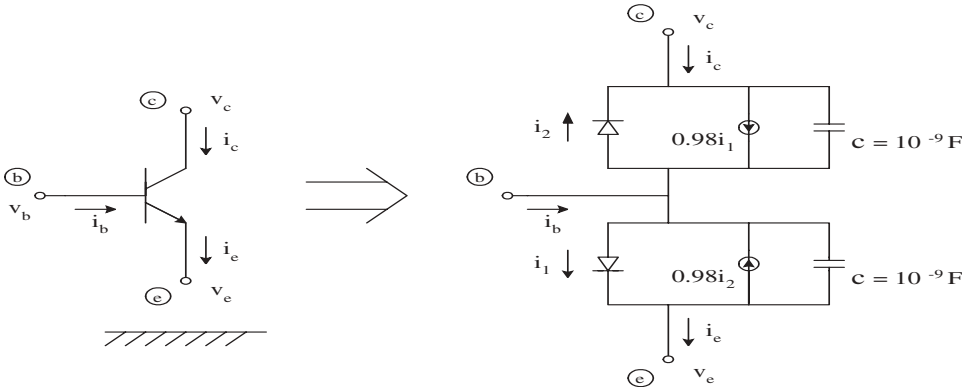


FIG. 8. *A transistor and its circuit.*

We also assume that the system starts from total rest for the example. The parameters are defined as $c_1 = c_4 = 0.1$nF, $c_2 = c_3 = 0.2$nF, $R_1 = 50\Omega$, $R_3 = 50\Omega$, $R_4 = 50\Omega$, $z_{01} = 10\Omega$, $z_{02} = 20\Omega$, $z_{03} = 30\Omega$, $\tau_1 = 1$ns, $\tau_2 = 0.9$ns, $\tau_3 = 0.8$ns, $T_e = 30$ns, and

$$E(t) = \begin{cases} 1\text{V}, & 0.1\text{ns} < t \le 0.2\text{ns}; \\ 0 & \text{otherwise.} \end{cases}$$

We let the time-step be 0.01ns. The far end voltage $v_4(t)$ is shown in Figure 7 after nine waveform iterations.

4.3. Example 3. Now we compute a circuit which includes transistors. The transistor is given in Figure 8. The transistor equations are

$$(45)\quad \begin{cases} i_c = 0.98i_1 - i_2 + c\dfrac{d(v_c - v_b)}{dt}, \\ i_e = -0.98i_2 + i_1 + c\dfrac{d(v_b - v_e)}{dt}, \\ i_b = i_e - i_c, \quad i_1 = 10^{-6}(e^{40(v_b - v_e)} - 1), \quad i_2 = 10^{-6}(e^{40(v_b - v_c)} - 1). \end{cases}$$

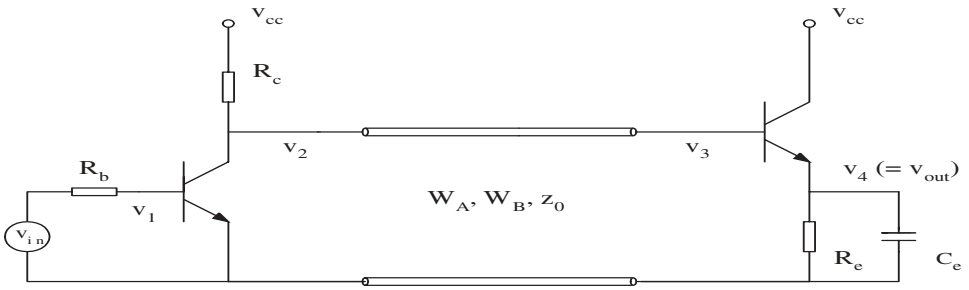


FIG. 9. *A nonlinear circuit with transistors and transmission lines.*

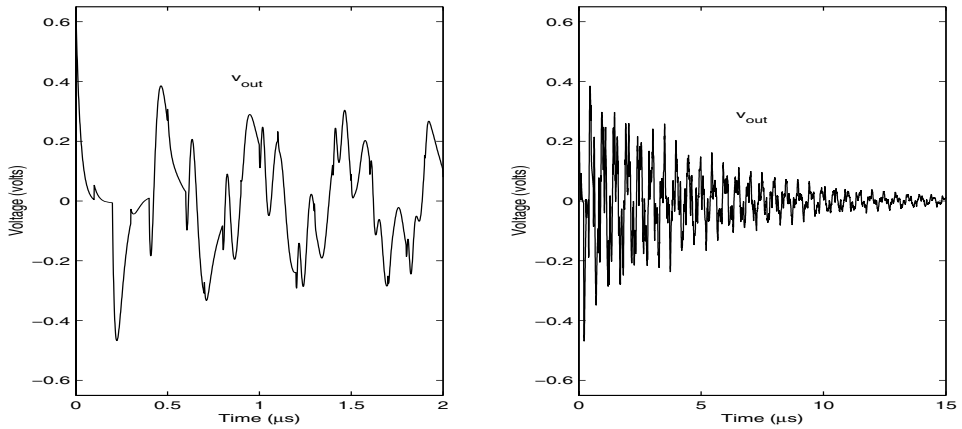


FIG. 10. *Far end voltage of a nonlinear circuit with transistors and transmission lines in Example* 3 *(left: computed result; right: long time behavior).*

The circuit is shown in Figure 9. Its equations are

$$
\begin{cases}
\dfrac{v_{in}-v_1}{R_b} = -0.98\times 10^{-6}(e^{40(v_1-v_2)}-1)+10^{-6}(e^{40v_1}-1) \\
\qquad\qquad +10^{-9}\dfrac{dv_1}{dt}-0.98\times 10^{-6}(e^{40v_1}-1)+10^{-6}(e^{40(v_1-v_2)}-1) \\
\qquad\qquad -10^{-9}\dfrac{d(v_2-v_1)}{dt}, \\
\dfrac{v_{cc}-v_2}{R_c} = \dfrac{v_2-W_A(t-\tau)}{z_0}+0.98\times 10^{-6}(e^{40v_1}-1)-10^{-6}(e^{40(v_1-v_2)}-1) \\
\qquad\qquad +10^{-9}\dfrac{d(v_2-v_1)}{dt}, \\
\dfrac{W_B(t-\tau)-v_3}{z_0}+0.98\times 10^{-6}(e^{40(v_3-v_4)}-1)-10^{-6}(e^{40(v_3-v_{cc})}-1)-10^{-9}\dfrac{dv_3}{dt} \\
\qquad = -0.98\times 10^{-6}(e^{40(v_3-v_{cc})}-1)+10^{-6}(e^{40(v_3-v_4)}-1)+10^{-9}\dfrac{d(v_3-v_4)}{dt}, \\
-0.98\times 10^{-6}(e^{40(v_3-v_{cc})}-1)+10^{-6}(e^{40(v_3-v_4)}-1)+10^{-9}\dfrac{d(v_3-v_4)}{dt} = \dfrac{v_4}{R_e}+c_e\dfrac{dv_4}{dt}, \\
W_A(t)=2v_3(t)-W_B(t-\tau), \quad W_B(t)=2v_2(t)-W_A(t-\tau).
\end{cases}
\tag{46}
$$

The above circuit parameters are set to be $v_{cc} = 5\text{V}$, $R_b = 50\Omega$, $R_c = 1000\Omega$, $R_e = 1000\Omega$, $c_e = 10\text{pF}$, $z_0 = 25\Omega$, $\tau = 0.1\mu\text{s}$, $T_e = 2\mu\text{s}$, $v_1(0) = v_2(0) = v_3(0) = 0.65\text{V}$, $\varphi_A(\theta) \equiv \varphi_B(\theta) \equiv 0$ $(-\tau \le \theta < 0)$, and

$$v_{in}(t) = \begin{cases} 0, & 0.05\mu\text{s} < t \le 0.06\mu\text{s}; \\ 0.8\text{V} & \text{otherwise.} \end{cases}$$

The Gauss–Seidel WR method of (46) is

$$\begin{cases} 2\times 10^{-9}\dfrac{dv_1^{(k)}}{dt} + \dfrac{v_1^{(k)}}{R_b} + 10^{-8}e^{40(v_1^{(k)}-v_2^{(k-1)})} + 10^{-8}e^{40v_1^{(k)}} \\ \qquad = 10^{-9}\dfrac{dv_2^{(k-1)}}{dt} + \dfrac{v_{in}}{R_b} + 2\times 10^{-8}, \\ -10^{-9}\dfrac{dv_1^{(k)}}{dt} + 10^{-9}\dfrac{dv_2^{(k)}}{dt} + \left(\dfrac{1}{R_c} + \dfrac{1}{z_0}\right)v_2^{(k)} + 0.98\times 10^{-6}e^{40v_1^{(k)}} - 10^{-6}e^{40(v_1^{(k)}-v_2^{(k)})} \\ \qquad = \dfrac{W_A^{(k-1)}(t-\tau)}{z_0} + \dfrac{v_{cc}}{R_c} - 2\times 10^{-8}, \\ 2\times 10^{-9}\dfrac{dv_3^{(k)}}{dt} + \dfrac{v_3^{(k)}}{z_0} + 10^{-8}e^{40(v_3^{(k)}-v_{cc})} + 10^{-8}e^{40(v_3^{(k)}-v_4^{(k-1)})} \\ \qquad = 10^{-9}\dfrac{dv_4^{(k-1)}}{dt} + \dfrac{W_B^{(k-1)}(t-\tau)}{z_0} + 2\times 10^{-8}, \\ -10^{-9}\dfrac{dv_3^{(k)}}{dt} + (1+c_e)\times 10^{-9}\dfrac{dv_4^{(k)}}{dt} + \dfrac{v_4^{(k)}}{R_e} - 10^{-6}e^{40(v_3^{(k)}-v_4^{(k)})} + 0.98\times 10^{-6}e^{40(v_3^{(k)}-v_{cc})} \\ \qquad = -2\times 10^{-8}, \\ -2v_3^{(k)}(t) + W_A^{(k)}(t) = -W_B^{(k-1)}(t-\tau), \quad -2v_2^{(k)}(t) + W_B^{(k)}(t) = -W_A^{(k-1)}(t-\tau), \\ [v_1^{(k)}(0), v_2^{(k)}(0), v_3^{(k)}(0), v_4^{(k)}(0)]^t = [v_1(0), v_2(0), v_3(0), v_4(0)]^t, \\ [W_A^{(k)}(\theta), W_B^{(k)}(\theta)]^t \equiv [\varphi_A(\theta), \varphi_B(\theta)]^t, \quad -\tau \le \theta < 0, \quad t \in [0, T_e], \quad k = 1, 2, \ldots. \end{cases} \tag{47}$$

The spectral radius of the Gauss–Seidel splitting on the coefficient matrix of the derivative term in (47) is 0.5. The Gauss–Seidel WR method (47) is convergent by our theorem of the paper. In our computations, we let the time-step be $10^{-3}\mu$s and the number of Newton iterations be 5. After 28 waveform iterations the computed voltage $v_{out}(t)(= v_4(t))$ is shown in Figure 10.

5. Conclusion. We have presented an interesting time-domain approach to solving a circuit with lossless distributed elements. The circuit is described by DAEs with multiple constant delays. The WR process decouples the system into ODEs and AEs. A detailed proof of convergence of the new method is given. This approach directly leads to the solution of the circuit in time-domain, and any of the general-purpose circuit simulators can be used to solve the decoupled subsystems.

Acknowledgments. The author wishes to thank Professor Omar Wing for his encouragement and for supplying the third example. He also thanks Professor J. R. Cash for his valuable suggestions.

REFERENCES

[1] O. Wing, *On VLSI interconnects*, in Proceedings of the China 1991 International Conference on Circuits and Systems, Shenzhen, China, 1991, pp. 991–996.

[2] R. Achar and M. S. Nakhla, *Simulation of high-speed interconnects*, Proc. IEEE, 89 (2001), pp. 693–728.

[3] U. Miekkala and O. Nevanlinna, *Convergence of dynamic iteration methods for initial value problems*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 459–482.

[4] J. K. White and A. L. Sangiovanni-Vincentelli, *Relaxation Techniques for the Simulation of VLSI Circuits*, Kluwer Academic Publishers, New York, 1987.

[5] K. Burrage, *Parallel and Sequential Methods for Ordinary Differential Equations*, Oxford University Press, New York, 1995.

[6] J. Janssen and S. Vandewalle, *Multigrid waveform relaxation on spatial finite element meshes: The continuous-time case*, SIAM J. Numer. Anal., 33 (1996), pp. 456–474.

[7] M. J. Gander and A. M. Stuart, *Space-time continuous analysis of waveform relaxation for the heat equation*, SIAM J. Sci. Comput., 19 (1998), pp. 2014–2031.

[8] Z. Jackiewicz and M. Kwapisz, *Convergence of waveform relaxation methods for differential-algebraic systems*, SIAM J. Numer. Anal., 33 (1996), pp. 2303–2317.

[9] G. D. Gristede, A. E. Ruehli, and C. A. Zukowski, *Convergence properties of waveform relaxation circuit simulation methods*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 45 (1998), pp. 726–738.

[10] Y. L. Jiang, W. S. Luk, and O. Wing, *Convergence-theoretics of classical and Krylov waveform relaxation methods for differential-algebraic equations*, IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, E80-A (1997), pp. 1961–1972.

[11] Y. L. Jiang and O. Wing, *Monotone waveform relaxation for systems of nonlinear differential-algebraic equations*, SIAM J. Numer. Anal., 38 (2000), pp. 170–185.

[12] Y. L. Jiang and O. Wing, *A note on convergence conditions of waveform relaxation algorithms for nonlinear differential-algebraic equations*, Appl. Numer. Math., 36 (2001), pp. 281–297.

[13] M. Bjørhus, *On dynamic iteration for delay differential equations*, BIT, 34 (1994), pp. 325–336.

[14] B. Zubik-Kowal, *Chebyshev pseudospectral method and waveform relaxation for differential and differential-functional parabolic equations*, Appl. Numer. Math., 34 (2000), pp. 309–328.

[15] Y. L. Jiang, R. M. M. Chen, and O. Wing, *Waveform relaxation of nonlinear second-order differential equations*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 48 (2001), pp. 1344–1347.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1032–1061

A POSTERIORI ERROR ESTIMATES FOR DISCONTINUOUS GALERKIN TIME-STEPPING METHOD FOR OPTIMAL CONTROL PROBLEMS GOVERNED BY PARABOLIC EQUATIONS*

WENBIN LIU†, HEPING MA‡, TAO TANG§, AND NINGNING YAN¶

Abstract. In this paper, we examine the discontinuous Galerkin (DG) finite element approximation to convex distributed optimal control problems governed by linear parabolic equations, where the discontinuous finite element method is used for the time discretization and the conforming finite element method is used for the space discretization. We derive a posteriori error estimates for both the state and the control approximation, assuming only that the underlying mesh in space is nondegenerate. For problems with control constraints of obstacle type, which are the kind most frequently met in applications, further improved error estimates are obtained.

Key words. optimal control, a posteriori error analysis, finite element approximation, discontinuous Galerkin method

AMS subject classifications. 49J20, 65N30

DOI. 10.1137/S0036142902397090

1. Introduction.

Optimal control or design is crucial to many engineering applications. Efficient numerical methods are essential to successful applications of optimal control. Nowadays, the finite element method seems to be the most widely used numerical method in computing optimal control problems, and the relevant literature is extensive. Some recent progress in this area has been made in, for example, [40, 41, 43]. Systematic introduction of the finite element method for PDEs and optimal control problems can be found in, for example, [10, 40, 43]. For instance, there have been extensive theoretical studies for finite element approximation of various optimal control problems; see [3, 15, 16, 18, 19], [20, 21, 22, 23, 24, 25, 26], and [37, 39, 44, 45]. For optimal control problems governed by linear elliptic or parabolic state equations, a priori error estimates of finite element approximation were established long ago; see, for example, [15, 18, 26, 37]. Furthermore a priori error estimates have been also established for some important flow control problems; see, e.g., [19, 20]. A priori error estimates have also been obtained for a class of state constrained con-

*Received by the editors February 8, 2002; accepted for publication (in revised form) October 8, 2003; published electronically July 29, 2004. This work was supported in part by Hong Kong Baptist University, Hong Kong Research Grants Council, and the British EPSRC GR/S11329.
http://www.siam.org/journals/sinum/42-3/39709.html

†Department of Mathematics, Xiangtan University, China, and CBS and Institute of Mathematics and Statistics, The University of Kent, Canterbury, England CT2 7NF (W.B.Liu@ukc.ac.uk).

‡Department of Mathematics, Shanghai University, Shanghai 200436, China (hpma@guomai.sh.cn). This author's research was also supported by the Special Funds for State Major Basic Research Projects of China G1999032804 and the Special Funds for Major Specialities of Shanghai Education Committee.

§Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong (ttang@math.hkbu.edu.hk).

¶Institute of System Sciences, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing, China (yan@staff.iss.ac.cn). This author's research was also supported by the Special Funds for Major State Basic Research Projects (G2000067102), National Natural Science Foundation of China (19931030), and Innovation Funds of the Academy of Mathematics and System Sciences, CAS.

trol problems in [44], although the state equation is assumed to be linear. In [32], the linear assumption has been removed by reformulating the control problem as an abstract optimization problem in some Banach spaces and then applying nonsmooth analysis. In fact, the state equation there can be a variational inequality.

In this paper, we examine an important class of finite element algorithms for a convex distributed optimal control problem governed by a linear parabolic equation, where the discontinuous polynomial base is used in time discretization and the conforming finite element method is used in space discretization. We present an a posteriori error analysis for this approximation.

Adaptive finite element approximation is among the most important means to boost the accuracy and efficiency of the finite element discretization. It ensures a higher density of nodes in certain areas of the given domain, where the solution is more difficult to approximate using an a posteriori error indicator. The decision about whether further refinement of meshes is necessary is based on the estimate of the discretization error. If further refinement is to be performed, then the error indicator is used as a guide to show how the refinement might be accomplished most efficiently. The literature in this area is huge. Some of the techniques directly relevant to our work can be found in [1, 5, 33, 36, 46]. It is our belief that adaptive finite element enhancement is one of the future directions to pursue in developing sophisticated numerical methods for optimal design problems.

Although adaptive finite element approximation is widely used in numerical simulations, it has not yet been *fully* utilized in optimal design. Initial attempts in this aspect have only been reported recently for some design problems (see, e.g., [2, 4, 38, 42]), and only a posteriori error indicators of a heuristic nature are used in most applications. For instance, in some existing work on adaptive finite element approximation of optimal design, the mesh refinement is guided by a posteriori error estimators based on a posteriori error estimates *solely* from the state equation for a fixed control. Thus error information from the approximation of the control (design) is not utilized. This strategy was found to be inefficient in recent numerical experiments (see [7, 27]). Although these methods may work well in some particular applications, they cannot be applied confidently in general. It is unlikely that the potential power of adaptive finite element approximation has been fully utilized due to the lack of more sophisticated a posteriori error indicators.

It is not straightforward to rigorously derive suitable a posteriori error estimators for general optimal control problems. In particular, it seems difficult to apply gradient recovery techniques since the control is normally not differentiable. Recovering approximation in function values is in general difficult. For a similar reason, it also seems difficult to apply the local solution strategy.

Very recently, some error indicators of residual type were developed in [6, 7, 27, 30, 34, 35, 36]. These error estimators are based on a posteriori estimation of the discretization error for the state *and* the control (design).

When there is no constraint in a control problem, normally the optimality conditions consist of coupled partial differential equations only. Consequently one may be able to write down the dual system of the *whole* optimality conditions, and then to apply the weighted a posteriori error estimation technique to obtain a posteriori estimators for objective functional approximation error of the control problem; see [6, 7]. Such estimators have indeed been derived for some unconstrained elliptic control problems, and have proved quite efficient in the numerical tests carried out in [6].

However, there frequently exist some constraints for the control in applications. In such cases, the optimality conditions often contain a variational inequality and then have some very different properties. For example, the dual system is generally unknown. Thus it does not seem to be always possible to apply the techniques used in [6, 7] to constrained control problems.

In our work, constrained cases are studied via residual estimation using the norms of energy type. A posteriori error estimators are derived for some constrained control problems governed by elliptic and parabolic equations; see [27, 34, 35, 36].

In recent years, the discontinuous Galerkin (DG) discretization has proved useful in computing time-dependent convection and diffusion equations; see [12, 13, 14] for the DG time-stepping method where only time discretization is discontinuous. It will be simply referred as to the DG method in this paper, although we are aware that there exist several DG discretization schemes in the literature. The DG has proved important in diffusion dominated equations, such as the heat equations, which govern our control problems to be examined in this paper. Furthermore the DG method has been found useful in computing optimal control of diffusion dominated systems; see [40]. However, there is a lack of an a posteriori error analysis for the DG approximation of the control systems, which is vital for further studies of mesh adaptivity of the control problems.

The purpose of this work is to extend the approaches in [12, 27, 34, 35, 36] and to derive a posteriori error estimates for the DG finite element approximation of distributed convex optimal problems governed by linear parabolic equations. Deriving such estimates for the DG finite element scheme is much more involved than for the backward-Euler scheme; see [36]. For example, some approaches applied in [12, 13, 14] have to be essentially modified for our purpose. Furthermore, novel approaches are needed to derive the improved estimates for the control with constraints of obstacle type. Optimal control with obstacle constraints is most frequently met in practical control problems. In fact, the majority of the existing research on constrained control concentrates on this type problem; see [28] and [43], for instance.

The plan of the paper is as follows. In section 2 we shall give a brief review of the finite element method and the discontinuous Galerkin discretization, and then construct the approximation schemes for the optimal control problem. In section 3, a posteriori error bounds are derived for the control problem. In section 4, some applications are discussed. In section 5, improved error estimates are derived for the problem with an obstacle constraint.

Let Ω and Ω_U be bounded open sets in $\mathbb{R}^n$ ($n \leq 3$) with Lipschitz boundaries $\partial\Omega$ and $\partial\Omega_U$. In this paper we adopt the standard notation $W^{m,q}(\Omega)$ for Sobolev spaces on Ω with norm $\|\cdot\|_{m,q,\Omega}$ and seminorm $|\cdot|_{m,q,\Omega}$. We denote $W^{m,2}(\Omega)$ by $H^m(\Omega)$ and set $H_0^1(\Omega) \equiv \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$.

We denote by $L^s(0,T;W^{m,q}(\Omega))$ the Banach space of all L^s integrable functions from $(0,T)$ into $W^{m,q}(\Omega)$ with norm $\|v\|_{L^s(0,T;W^{m,q}(\Omega))} = (\int_0^T \|v\|^s_{W^{m,q}(\Omega)} dt)^{\frac{1}{s}}$ for $s \in [1,\infty)$ and the standard modification for $s = \infty$. Similarly, we define the spaces $H^1(0,T;W^{m,q}(\Omega))$ and $C^l(0,T;W^{m,q}(\Omega))$. The details can be found in [29]. In addition c or C denotes a general positive constant independent of h.

2. Approximation scheme of optimal control problems governed by parabolic equations. In this section we study the finite element and the discontinuous Galerkin approximation of distributed convex optimal control problems, where the state is governed by a parabolic equation. In this paper, we shall take the state

space $W = L^2(0,T;Y)$ with $Y = H_0^1(\Omega)$ and the control space $X = L^2(0,T;U)$ with $U = L^2(\Omega_U)$ to fix the idea. Let B be a linear continuous operator from X to $L^2(0,T;Y')$ and K be a closed convex set in X. We are interested in the following optimal control problem:

$$\min_{u\in K}\int_0^T (g(y)+h(u))\,dt$$

subject to

$$\begin{cases} \partial_t y - \operatorname{div}(A\nabla y) = f + Bu, & x\in\Omega, \quad t\in(0,T],\\ y|_{\partial\Omega} = 0, & t\in[0,T],\\ y(x,0) = y_0(x), & x\in\Omega, \end{cases}$$

where $f \in L^2(0,T;Y')$, $y_0 \in H_0^1(\Omega)$, and

$$A(x) = (a_{ij}(x))_{n\times n} \in (C^\infty(\bar\Omega))^{n\times n}$$

such that there is a constant $c > 0$ satisfying

$$(A\xi)\cdot\xi \ge c|\xi|^2 \quad \forall \xi\in\mathbb{R}^n.$$

Let

$$\begin{aligned} a(v,w) &= \int_\Omega (A\nabla v)\cdot\nabla w && \forall v,w\in H^1(\Omega),\\ (f_1,f_2) &= \int_\Omega f_1 f_2 && \forall f_1,f_2\in L^2(\Omega),\\ (v,w)_U &= \int_{\Omega_U} vw && \forall v,w\in L^2(\Omega_U). \end{aligned}$$

It follows from the assumptions on A that there are constants c and $C > 0$ such that

$$a(v,v)\ge c\|v\|_{1,\Omega}^2, \quad |a(v,w)|\le C|v|_{1,\Omega}|w|_{1,\Omega} \quad \forall v,w\in Y.$$

Then a weak formulation of the convex optimal control problem reads as

$$\min_{u\in K}\int_0^T (g(y)+h(u))\,dt, \tag{1}$$

where $y\in W$ is subject to

$$\begin{cases} (\partial_t y, w) + a(y,w) = (f+Bu, w) & \forall w\in Y,\ t\in(0,T],\\ y(0) = y_0. \end{cases}$$

We assume that g is a convex functional which is continuously differentiable on $L^2(\Omega)$, and h is a strictly convex and continuously differentiable function on U. We further assume that $h(u)\to+\infty$ as $\|u\|_U\to\infty$ and that the functional $g(\cdot)$ is bounded below. This setting includes the most widely used quadratic control problem:

$$\min_{u\in K}\left\{\frac{1}{2}\int_0^T (\|y-z_d\|_{L^2(\Omega)}^2 + \|u\|_{L^2(\Omega_U)}^2)dt\right\}\ ,$$

where y, u are defined as above and z_d is a given state. It is well known (see, e.g., [28]) that the control problem (1) has a unique solution (y, u), and that a pair (y, u) is the solution of (1) if and only if there is a costate $p \in W$ such that the triplet (y, p, u) satisfies the following optimality conditions:

$$\text{(2)} \qquad \begin{cases} (\partial_t y, w) + a(y, w) = (f + Bu, w) & \forall w \in Y, \quad y(0) = y_0, \\ -(\partial_t p, q) + a(q, p) = (g'(y), q) & \forall q \in Y, \quad p(T) = 0, \\ \displaystyle\int_0^T (h'(u) + B^* p, v - u)_U \, dt \geq 0 & \forall v \in K, \end{cases}$$

where B^* is the adjoint operator of B.

Let us consider the finite element approximation of the control problem (1). Here we consider only n-simplices Lagrange elements.

Let Ω^h be a polygonal approximation to Ω with boundary $\partial\Omega^h$. Let T^h be a partitioning of Ω^h into disjoint regular n-simplex τ, so that $\bar{\Omega}^h = \cup_{\tau \in T^h} \bar{\tau}$. Each element has at most one face on $\partial\Omega^h$, and joint elements $\bar{\tau}$ and $\bar{\tau}'$ have either only one common vertex or a whole edge or face if τ and $\tau' \in T^h$. We further require that $P_i \in \partial\Omega^h$ implies $P_i \in \partial\Omega$, where $\{P_i\}$ $(i = 1, 2, \ldots, J)$ is the vertex set associated with the triangulation T^h. We assume that Ω is a convex polygon so that $\Omega = \Omega^h$. The convexity assumption is also important to have the H^2 a priori estimate for the dual equations in Lemma 3.4, which is used in deriving our L^2-L^2 and L^∞-L^2 a posteriori error estimates, although it is not needed for L^2-H^1 estimates. Without the convexity assumption, in general the order of our estimates for the state and costate approximation will be lower if $\partial\Omega$ is nonsmooth. We denote by h_τ the maximum diameter of the element τ in T^h.

Associated with T^h is a finite dimensional subspace S^h of $C(\bar{\Omega}^h)$ such that $w|_\tau$ are m-order polynomials $(m \geq 1)$ for all $w \in S^h$ and $\tau \in T^h$. Let $Y^h = S^h \cap H_0^1(\Omega)$, $W^h = L^2(0, T; Y^h)$; it is easy to see that $Y^h \subset Y$, $W^h \subset W$.

Similarly, we do a partitioning of Ω_U and use the following corresponding notations: T_U^h, τ_U, h_{τ_U}, P_i^U $(i = 1, 2, \ldots, J_U)$, and $\Omega_U^h = \Omega_U$.

Associated with T_U^h is another finite dimensional subspace U^h of $L^2(\Omega_U^h)$ such that $v|_{\tau_U}$ are m-order polynomials $(m \geq 0)$ for all $v \in U^h$ and $\tau_U \in T_U^h$. Here there is no requirement for the continuity. Let $X^h = L^2(0, T; U^h)$. It is easy to see that $U^h \subset U$ and $X^h \subset X$.

Let K^h be an approximation of K. Here we assume that $K^h \subset K$ and $K^h \subset X^h$ for ease of exposition. A nonconforming finite element method will be used later for the problem with the constraint of obstacle type. For more general cases, the readers are referred to [35]. Then a possible semidiscrete finite element approximation of (1) is as follows:

$$\text{(3)} \qquad \min_{u_h \in K^h} \int_0^T (g(y_h) + h(u_h)) \, dt$$

with $y_h \in W^h$ subject to

$$\begin{cases} (\partial_t y_h, w) + a(y_h, w) = (f + Bu_h, w) & \forall w \in Y^h, \quad t \in (0, T], \\ y_h(0) = y_0^h, \end{cases}$$

where K^h is a closed convex set in X^h, $y_0^h \in Y^h$ is an approximation of y_0.

It follows that the control problem (3) has a unique solution (y_h, u_h) and that a pair $(y_h, u_h) \in W^h \times K^h$ is the solution of (3) if and only if there is a costate $p_h \in W^h$

such that the triplet $(y_h, p_h, u_h) \in W^h \times W^h \times K^h$ satisfies the following optimality conditions:

$$(4)\quad \begin{cases} (\partial_t y_h, w) + a(y_h, w) = (f + Bu_h, w) & \forall w \in Y^h, \quad y_h(0) = y_0^h, \\ -(\partial_t p_h, q) + a(q, p_h) = (g'(y_h), q) & \forall q \in Y^h, \quad p_h(T) = 0, \\ \displaystyle\int_0^T (h'(u_h) + B^* p_h, v - u_h)_U \, dt \geq 0 & \forall v \in K^h. \end{cases}$$

The optimality conditions in (4) are the semidiscrete approximation to the problem (1). Now, we are going to consider the fully discrete approximation for the above semidiscrete problem by using the DG method.

Let $0 = t_0 < t_1 < \cdots < t_N = T$, $I_k = (t_{k-1}, t_k]$, $\Delta t_k = t_k - t_{k-1}$ $(k = 1, 2, \ldots, N)$. For $k = 1, 2, \ldots, N$, construct the finite element spaces $Y^{h,k} \in H_0^1(\Omega)$ (similar to Y^h) with the mesh $T^{h,k}$, and construct the finite element spaces $U^{h,k} \in L^2(\Omega_U)$ (similar to U^h) with the mesh $T_U^{h,k}$. Let h_{τ^k} $(h_{\tau_U^k})$ denote the maximum diameter of the element τ^k (τ_U^k) in $T^{h,k}$ $(T_U^{h,k})$. To simplify notation, we will regard a discrete quantity Q^k as $Q(t)$ such that $Q(t)|_{I_k} \equiv Q^k$, and we will denote $\tau(t)$, $\tau_U(t)$, $h_\tau(t)$, and $h_{\tau_U}(t)$ by τ, τ_U, h_τ, and h_{τ_U}, respectively. Let

$$\begin{aligned} &W^\delta = \left\{ w \mid w(x,t)|_{\Omega \times I_k} = \sum_{j=0}^{r} t^j \varphi_j(x), \quad \varphi_j \in Y^{h,k} \right\}, \quad r \geq 0, \\ &X^\delta = \{ v \mid v(x,t)|_{\Omega \times I_k} = \psi(x), \quad \psi \in U^{h,k} \}, \qquad K^\delta \subset (X^\delta \cap K), \\ &[w]_k = w_k^+ - w_k^-, \qquad w_k^\pm = \lim_{s \to 0^\pm} w(t_k + s). \end{aligned}$$

The fully discrete approximation scheme is to find $(y_\delta, u_\delta) \in W^\delta \times X^\delta$ such that

$$(5)\qquad \min_{u_\delta \in K^\delta} \int_0^T (g(y_\delta) + h(u_\delta)) \, dt$$

subject to

$$\begin{aligned} \int_0^T ((\partial_t y_\delta, w) + a(y_\delta, w)) \, dt + \sum_{k=1}^{N-1} ([y_\delta]_k, w_k^+) + ((y_\delta)_0^+ - y_0^h, w_0^+) \\ = \int_0^T (f + Bu_\delta, w) \, dt \quad \forall w \in W^\delta, \end{aligned}$$

where $y_0^h \in Y^{h,0}$ is the approximation to y_0. It follows that the control problem (5) has a unique solution (y_δ, u_δ), and that a pair $(y_\delta, u_\delta) \in W^\delta \times X^\delta$ is the solutions of (5) if and only if there is costate $p_\delta \in W^\delta$ such that the triplet $(y_\delta, p_\delta, u_\delta)$ satisfies

the following optimality conditions:

$$
(6)\quad \begin{cases} \displaystyle\int_0^T ((\partial_t y_\delta, w) + a(y_\delta, w))\, dt + \sum_{k=0}^{N-1} ([y_\delta]_k, w_k^+) & \\ \quad = \displaystyle\int_0^T (f + Bu_\delta, w)\, dt & \forall w \in W^\delta, \quad (y_\delta)_0^- = y_0^h, \\ \displaystyle\int_0^T (-(\partial_t p_\delta, q) + a(p_\delta, q))\, dt - \sum_{k=1}^{N} ([p_\delta]_k, q_k^-) & \\ \quad = \displaystyle\int_0^T (g'(y_\delta), q)\, dt & \forall q \in W^\delta, \quad (p_\delta)_N^+ = 0, \\ \displaystyle\int_0^T (h'(u_\delta) + B^* p_\delta, v - u_\delta)_U\, dt \geq 0 & \forall v \in K^\delta. \end{cases}
$$

This is a finite dimensional optimization problem and may be solved by existing mathematical programming methods. The above DG approximation of the control problem has been used in practical problems; see [40].

In order to obtain a numerical solution of acceptable accuracy for the optimal control problem, the finite element meshes have to be refined according to a mesh refinement scheme. Adaptive finite element approximation uses a posteriori error indicator to guide the mesh refinement procedure. In the following section we shall derive some a posteriori error estimates for the DG finite element approximation of the optimal control problem governed by parabolic equations, which can be used as such an error indicator in developing adaptive finite element schemes of the control problem.

3. A posteriori error estimates. In this section we derive a posteriori error estimates for the DG finite element approximation of the convex optimal problem governed by a parabolic equation. In general, analysis of the finite element approximation of a control problem governed by parabolic equations is more involved than is that of a control problem governed by elliptic equations. The main complication is due to the fact that the properties of the time variable and its discretization are quite different from those of the space (elliptic) variables. Thus different techniques are needed to handle the two groups of variables, and their interactions.

We now need more assumptions on B and g in deriving our estimates. We essentially assume that B is bounded from $L^2(0,T;L^2(\Omega_U))$ to $L^2(0,T;L^2(\Omega))$ so that differential operators are excluded. To derive L^∞ estimates, we need a continuity from $L^2(\Omega_U)$ to $L^2(\Omega)$ uniformly with respect to t, while we have embedded U into X. For g we assume that its derivative is Lipschitz continuous. Thus we make the following assumptions:

$$
(7)\qquad |(Bv, w)_X| = |(v, B^* w)| \leq C\|v\|_{0,\Omega_U}\|w\|_{0,\Omega} \qquad \forall\, v \in U, w \in Y,
$$

$$
(8)\qquad |(g'(v) - g'(w), q)| \leq C\|v - w\|_{0,\Omega}\|q\|_{0,\Omega} \qquad \forall\, v, w, q \in Y,
$$

and there is a constant $c > 0$ such that

$$
(9)\qquad (h'(v) - h'(w), v - w) \geq c\|v - w\|^2_{0,\Omega_U} \qquad \forall\, v, w \in U,
$$

$$
(10)\qquad (g'(v) - g'(w), v - w) \geq 0 \qquad \forall\, v, w \in Y,
$$

which are convex conditions on the functionals h and g. These conditions hold for the quadratic control problems where $\Omega = \Omega_U$ and $B = I$.

The following lemma is important in deriving residual type a posteriori error estimates.

LEMMA 3.1. *Let π_h be the average interpolation operator defined in* [21]. *For any $v \in W^{1,q}(\Omega^h)$ and $1 \le q \le \infty$,*

$$\|v - \pi_h v\|_{l,q,\tau} \le C \sum_{\bar{\tau}' \cap \bar{\tau} \neq \emptyset} h_\tau^{m-l} |v|_{m,q,\tau'}, \qquad v \in W^{m,q}(\tau'), \qquad l = 0, 1, \quad l \le m \le 2.$$

REMARK 3.1. *One of the key steps in deriving a posteriori error estimates for the discontinuous Galerkin method is to construct a suitable L^2 stable approximation of the solution of the dual equation. In* [12], *this approximation is defined to be the space-time L^2-projection of the solution. However, for this selection the spatial projection error cannot be bounded locally due to the global nature of the projecting onto continuous piecewise polynomial functions. This leads to the inconvenience restriction in* [12] *on the mesh used in the approximation: the change in the size of the elements in the mesh must be very smooth, which may be unrealistic in an adaptive finite element implementation. We shall define this approximation to be the L^2-projection of the solution of the dual equation in time, but the quasi-interpolant of the solution in space as defined in* [21]. *It follows from Lemma* 3.1 *that this approximation is L^2 stable. Furthermore, optimal approximation results hold on local patches surrounding a particular element. It is then possible to derive a posteriori error estimates assuming only nondegeneracy of the mesh.*

LEMMA 3.2 (see [25]). *For all $v \in W^{1,q}(\Omega)$, $1 \le q \le \infty$,*

$$\|v\|_{0,q,\partial\tau} \le C(h_\tau^{-1/q} \|v\|_{0,q,\tau} + h_\tau^{1-1/q} |v|_{1,q,\tau}). \tag{11}$$

3.1. $L^2(L^2)$ error estimates. First, let us present a lemma which is essential for our a posteriori error estimate analysis. Assuming that one can find an element v in K^δ to approximate the optimal control in an appropriate way, the approximation error in the control is then shown to be represented by an a posteriori error estimator, plus the approximation error in the costate. For constraints of obstacle type, this assumption can be verified for piecewise constant control approximation by taking v to be the integral average of the optimal control; see Examples 3.1 and 3.2.

LEMMA 3.3. *Let (y, p, u) and $(y_\delta, p_\delta, u_\delta)$ be the solutions of* (2) *and* (6). *Assume that* (9), (10), *and* (7) *hold; $K^\delta \subset K$; for all $1 \le k \le N$, $(h'(u_\delta) + B^* p_\delta)|_{\tau_U^k \times I_k} \in H^1(\tau_U^k \times I_k)$; and there is a $v \in K^\delta$ such that*

(12)
$$\begin{aligned} &\left| \int_{I_k} (h'(u_\delta) + B^* p_\delta, v - u)_U \, dt \right| \\ \le\ & C \int_{I_k} \sum_{\tau_U \in T_U^{h,k}} (h_{\tau_U} |h'(u_\delta) + B^* p_\delta|_{1,\tau_U} + \Delta t_k \|\partial_t (h'(u_\delta) + B^* p_\delta)\|_{0,\tau_U}) \|u - u_\delta\|_{0,\tau_U} \, dt. \end{aligned}$$

Then we have

$$\|u - u_\delta\|^2_{L^2(0,T;L^2(\Omega_U))} \le C \left(\eta_1^2 + \|p_\delta - p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))} \right), \tag{13}$$

where

$$\eta_1^2 = \sum_{k=1}^N \sum_{\tau_U \in T_U^{h,k}} \int_{I_k} (h_{\tau_U}^2 |h'(u_\delta) + B^* p_\delta|^2_{1,\tau_U} + \Delta t_k^2 \|\partial_t (h'(u_\delta) + B^* p_\delta)\|^2_{0,\tau_U}) \, dt$$

and $(y^{u_\delta}, p^{u_\delta}) \in W \times W$ *is defined by the following system:*

$$\begin{cases} (\partial_t y^{u_\delta}, w) + a(y^{u_\delta}, w) = (f + Bu_\delta, w) \quad \forall w \in Y, \quad t \in (0,T], \\ y^{u_\delta}(0) = y_0, \end{cases} \tag{14}$$

$$\begin{cases} -(\partial_t p^{u_\delta}, q) + a(q, p^{u_\delta}) = (g'(y^{u_\delta}), q) \quad \forall q \in Y, \quad t \in [0,T), \\ p^{u_\delta}(T) = 0. \end{cases} \tag{15}$$

Proof. It follows from (9), $(2)_3$, and $(6)_3$ that, for any $v \in K^\delta$,

$$\begin{aligned} & c\|u - u_\delta\|^2_{L^2(0,T;L^2(\Omega_U))} \le \int_0^T (h'(u), u - u_\delta)_U \, dt - \int_0^T (h'(u_\delta), u - u_\delta)_U \, dt \\ \le & -\int_0^T (B^*p, u - u_\delta)_U \, dt - \int_0^T (h'(u_\delta), u - u_\delta)_U \, dt + \int_0^T (h'(u_\delta) + B^*p_\delta, v - u_\delta)_U \, dt \\ \le & \int_0^T (h'(u_\delta) + B^*p_\delta, v - u)_U \, dt + \int_0^T (B^*(p_\delta - p^{u_\delta}), u - u_\delta)_U \, dt \\ & + \int_0^T (B^*(p^{u_\delta} - p), u - u_\delta)_U \, dt, \end{aligned} \tag{16}$$

where p^{u_δ} is defined in (15). It is easy to see from (2), (14), and (15) that

$$(\partial_t(y^{u_\delta} - y), w) + a(y^{u_\delta} - y, w) = (B(u_\delta - u), w) \qquad \forall w \in Y, \tag{17}$$

$$-(\partial_t(p^{u_\delta} - p), q) + a(q, p^{u_\delta} - p) = (g'(y^{u_\delta}) - g'(y), q) \quad \forall q \in Y. \tag{18}$$

Taking $w = p^{u_\delta} - p$ in (17) and $q = y^{u_\delta} - y$ in (18) and using $(y^{u_\delta} - y)|_{t=0} = (p^{u_\delta} - p)|_{t=T} = 0$ and (10) lead to

$$\begin{aligned} \int_0^T (B(u_\delta - u), p^{u_\delta} - p) \, dt &= (y^{u_\delta} - y, p^{u_\delta} - p)\Big|_0^T \\ &\quad + \int_0^T (g'(y^{u_\delta}) - g'(y), y^{u_\delta} - y) \, dt \ge 0. \end{aligned} \tag{19}$$

Let v be the function satisfying (12). Then by (12), (7), and (19),

$$c\|u - u_\delta\|^2_{L^2(0,T;L^2(\Omega_U))} \le C\left(\eta_1^2 + \|p_\delta - p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))}\right) + \frac{c}{2}\|u - u_\delta\|^2_{L^2(0,T;L^2(\Omega_U))}, \tag{20}$$

which completes the proof. ∎

The assumption (12) is related to approximation properties of the convex set K. For instance, it always holds for unconstrained control, where $K = U$. For constraints of obstacle type, this assumption can also be verified.

We shall use the following dual equations: For given $f \in L^2(0,T;L^2(\Omega))$,

$$\begin{cases} \partial_t \varphi - \operatorname{div}(A\nabla\varphi) = f, & (x,t) \in \Omega \times (0,T], \\ \varphi|_{\partial\Omega} = 0, \quad t \in [0,T], & \varphi(x,0) = 0, \quad x \in \Omega, \end{cases} \tag{21}$$

and

$$\begin{cases} -\partial_t \psi - \operatorname{div}(A^*\nabla\psi) = f, & (x,t) \in \Omega \times [0,T), \\ \psi|_{\partial\Omega} = 0, \quad t \in [0,T], & \psi(x,T) = 0, \ x \in \Omega. \end{cases} \tag{22}$$

A similar idea is used in [21] for a Lagrange–Galerkin method.

LEMMA 3.4 (see [21]). *Assume that Ω is a convex domain. Let φ and ψ be the solutions of* (21) *and* (22)*, respectively. Then, for $v = \varphi$ or $v = \psi$,*

$$\begin{aligned}
&\|v\|_{L^\infty(0,T;L^2(\Omega))} \le C\|f\|_{L^2(0,T;L^2(\Omega))},\\
&\|\nabla v\|_{L^2(0,T;L^2(\Omega))} \le C\|f\|_{L^2(0,T;L^2(\Omega))},\\
&\|D^2 v\|_{L^2(0,T;L^2(\Omega))} \le C\|f\|_{L^2(0,T;L^2(\Omega))},\\
&\|\partial_t v\|_{L^2(0,T;L^2(\Omega))} \le C\|f\|_{L^2(0,T;L^2(\Omega))},
\end{aligned}$$

where $D^2 v = \max_{1\le i,j\le n} |\partial^2 v/\partial x_i \partial x_j|$.

In the following we deal with the error $\|p_\delta - p^{u_\delta}\|_{L^2(0,T;L^2(\Omega))}$ to derive the final estimates. Let $\partial T^{h,k}$ be the set consisting of all the faces l of any $\tau^k \in T^{h,k}$ such that l is not on $\partial\Omega$. The A-normal derivative jump over the interior face l is defined by

$$[(A\nabla v)\cdot \mathbf{n}]_l = ((A\nabla v)|_{\partial\tau_l^1} - (A\nabla v)|_{\partial\tau_l^2})\cdot \mathbf{n},$$

where $\mathbf{n}$ is the unit outer normal vector of τ_l^1 on $l = \bar\tau_l^1 \cap \bar\tau_l^2$. Let h_l be the maximum diameter of the face l.

LEMMA 3.5. *Let (y,p,u), $(y_\delta, p_\delta, u_\delta)$, and p^{u_δ} be the solutions of* (2), (6), *and* (15)*, respectively. Under the conditions of Lemma* 3.4 *and* (8)*,*

$$\|p_\delta - p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))} \le C \sum_{i=0,2\text{–}7} \eta_i^2,$$

where

$$\begin{aligned}
\eta_0^2 &= \|y_0^h - y_0\|^2_{0,\Omega},\\
\eta_2^2 &= \sum_{k=1}^N \sum_{\tau\in T^{h,k}} \int_{I_k} h_\tau^4 \left\| \partial_t p_\delta + g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta) + \frac{[p_\delta]_k}{\Delta t_k}\right\|^2_{0,\tau} dt,\\
\eta_3^2 &= \sum_{k=1}^N \sum_{\tau\in T^{h,k}} \int_{I_k} \Delta t_k^2 \,\|(\pi_k - I)(g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta))\|^2_{0,\tau}\, dt,\\
\eta_4^2 &= \sum_{k=1}^N \sum_{\tau\in T^{h,k}} \int_{I_k} h_\tau^4 \left\| \partial_t y_\delta - f - Bu_\delta - \operatorname{div}(A\nabla y_\delta) + \frac{[y_\delta]_{k-1}}{\Delta t_k}\right\|^2_{0,\tau} dt,\\
\eta_5^2 &= \sum_{k=1}^N \sum_{\tau\in T^{h,k}} \int_{I_k} \Delta t_k^2 \,\|(\pi_k - I)(f + \operatorname{div}(A\nabla y_\delta))\|^2_{0,\tau}\, dt,\\
\eta_6^2 &= \sum_{k=1}^N \sum_{l\in\partial T^{h,k}} \int_{I_k} h_l^3 (\|[(A\nabla y_\delta)\cdot\mathbf{n}]\|^2_{0,l} + \|[(A^*\nabla p_\delta)\cdot\mathbf{n}]\|^2_{0,l})\, dt,\\
\eta_7^2 &= \sum_{k=1}^N \sum_{\tau\in T^{h,k}} \Delta t_k (\|[y_\delta]_{k-1}\|^2_{0,\Omega} + \|[p_\delta]_k\|^2_{0,\Omega}),
\end{aligned}$$

where $\pi_k : L^2(I_k) \to \mathbb{P}_r(I_k)$ is the L^2-projection operator on the variable t.

Proof. Let φ be the solution of (21) with $f = p_\delta - p^{u_\delta}$ and $\varphi_I \in X^\delta$ be the interpolation of φ such that

$$\varphi_I|_{\Omega\times I_k} = \pi_{h,k}\pi_k\varphi, \qquad k = 1,2,\dots,N, \tag{23}$$

where $\pi_{h,k}$ is defined in Lemma 3.1 corresponding to the partitioning $T^{h,k}$ and $\pi_k : L^2(I_k) \to \mathbb{P}_r(I_k)$ is the L^2-projection operator on the variable t. Then it follows from (21), (15), (6), and Green's formula that

$$\begin{aligned}
&\|p_\delta - p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))} = \int_0^T (p_\delta - p^{u_\delta}, f)\, dt \\
= &\int_0^T (p_\delta - p^{u_\delta}, \partial_t\varphi - \operatorname{div}(A\nabla\varphi))\, dt \\
= &\int_0^T \left(-(\partial_t(p_\delta - p^{u_\delta}), \varphi) + a(\varphi, p_\delta - p^{u_\delta})\right)\, dt - \sum_{k=1}^N ([p_\delta]_k, \varphi_k^-) \\
= &\int_0^T \left(-(\partial_t p_\delta + g'(y^{u_\delta}), \varphi) + a(\varphi, p_\delta) - a(\varphi_I, p_\delta) + (\partial_t p_\delta + g'(y_\delta), \varphi_I)\right)\, dt \\
&+ \sum_{k=1}^N ([p_\delta]_k, (\varphi_I - \varphi)_k^-)\,,
\end{aligned}$$

which leads to

$$\begin{aligned}
&\|p_\delta - p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))} \\
= &\sum_{k=1}^N \int_{I_k} -\left(\partial_t p_\delta + g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta) + \frac{[p_\delta]_k}{\Delta t_k}, \varphi - \varphi_I\right)\, dt \\
&+ \int_0^T (g'(y_\delta) - g'(y^{u_\delta}), \varphi)\, dt + \int_0^T \sum_{l\in\partial T^h} \int_l [(A^*\nabla p_\delta)\cdot\mathbf{n}](\varphi - \varphi_I)\, dt \\
&+ \sum_{k=1}^N \int_{I_k} \left(\frac{[p_\delta]_k}{\Delta t_k}, (\varphi_I)_k^- - \varphi_I + \varphi - \varphi_k^-\right) dt \\
:= &\sum_{i=1}^4 I_i.
\end{aligned} \tag{24}$$

For simplicity, let

$$r_p(x,t)\Big|_{\Omega\times I_k} := \partial_t p_\delta + g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta) + \frac{[p_\delta]_k}{\Delta t_k}.$$

By Lemmas 3.1 and 3.4,

$$\begin{aligned}
&I_1 = \sum_{k=1}^N \int_{I_k} (r_p, (\pi_{h,k} - I)\pi_k\varphi + (\pi_k - I)\varphi)\, dt \\
= &\sum_{k=1}^N \int_{I_k} ((r_p, (\pi_{h,k} - I)\pi_k\varphi) - ((\pi_k - I)(g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta)), (\pi_k - I)\varphi))\, dt \\
\le &C\sum_{k=1}^N \sum_{\tau\in T^{h,k}} \int_{I_k} \left(h_\tau^4 \|r_p\|^2_{0,\tau} + \Delta t_k^2 \|(\pi_k - I)(g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta))\|^2_{0,\tau}\right)\, dt \\
&+ \sigma(\|D^2(\pi_k\varphi)\|^2_{L^2(0,T;L^2(\Omega))} + \|\partial_t\varphi\|^2_{L^2(0,T;L^2(\Omega))}) \\
\le &C(\eta_2^2 + \eta_3^2) + C\sigma\|p^{u_\delta} - p_\delta\|^2_{L^2(0,T;L^2(\Omega))}.
\end{aligned} \tag{25}$$

It is easy to see that from (8) and Lemma 3.4,

$$
\begin{aligned}
I_2 &= \int_0^T (g'(y_\delta) - g'(y^{u_\delta}), \varphi)\, dt \\
&\le C\|y_\delta - y^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))} + \sigma\|p^{u_\delta} - p_\delta\|^2_{L^2(0,T;L^2(\Omega))}.
\end{aligned} \tag{26}
$$

Similarly, by Lemmas 3.1, 3.2, and 3.4,

$$
\begin{aligned}
I_3 &= \int_0^T \sum_{l\in\partial T^h} \int_l [(A^*\nabla p_\delta)\cdot \mathbf{n}](\varphi - \varphi_I)\, dt \\
&= \sum_{k=1}^N \sum_{l\in\partial T^{h,k}} \int_{I_k}\int_l [(A^*\nabla p_\delta)\cdot \mathbf{n}](\varphi - \pi_{h,k}\varphi)\, dt \\
&\le C\sum_{k=1}^N \sum_{l\in\partial T^{h,k}} \int_{I_k} h_l^3 \|[(A^*\nabla p_\delta)\cdot \mathbf{n}]\|^2_{0,l}\, dt + \sigma\|D^2\varphi\|^2_{L^2(0,T;L^2(\Omega))} \\
&\le C\eta_6^2 + C\sigma\|p^{u_\delta} - p_\delta\|^2_{L^2(0,T;L^2(\Omega))}.
\end{aligned} \tag{27}
$$

It follows from Lemma 3.4 and the Schwarz inequality that

$$
\begin{aligned}
I_4 &= \sum_{k=1}^N \int_{I_k} \left(\frac{[p_\delta]_k}{\Delta t_k}, (\varphi_I)_k^- - \varphi_I + \varphi - \varphi_k^- \right) dt \\
&\le \sum_{k=1}^N \Delta t_k \|[p_\delta]_k\|^2_{0,\Omega} + \sigma\left(\|\partial_t\varphi_I\|^2_{L^2(0,T;L^2(\Omega))} + \|\partial_t\varphi\|^2_{L^2(0,T;L^2(\Omega))} \right) \\
&\le C\eta_7^2 + C\sigma\|p^{u_\delta} - p_\delta\|^2_{L^2(0,T;L^2(\Omega))}.
\end{aligned} \tag{28}
$$

Thus, the above estimates give

$$
\|p_\delta - p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))} \le C \sum_{i=2,3,6,7} \eta_i^2 + C\|y_\delta - y^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))}. \tag{29}
$$

Similarly, let ψ be the solution of (22) with $f = y_\delta - y^{u_\delta}$ and $\psi_I \in X^\delta$ be the interpolation of ψ such that

$$
\psi_I|_{\Omega\times I_k} = \pi_{h,k}\pi_k\psi, \qquad k = 1, 2, \ldots, N. \tag{30}
$$

Then, by Lemma 3.4, (14), (6), and Green's formula,

(31)
$$\begin{aligned}
&\|y_\delta - y^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))} = \int_0^T (y_\delta - y^{u_\delta}, f)\,dt = \int_0^T (y_\delta - y^{u_\delta}, -\partial_t\psi - \operatorname{div}(A^*\psi))\,dt\\
=\ & \int_0^T ((\partial_t(y_\delta - y^{u_\delta}),\psi) + a(y_\delta - y^{u_\delta},\psi))\ dt + \sum_{k=1}^{N-1}([y_\delta]_k,\psi_k^+) + ((y_\delta - y^{u_\delta})_0^+,\psi_0^+)\\
=\ & \int_0^T ((\partial_t y_\delta - f - Bu_\delta,\psi) + a(y_\delta,\psi) - a(y_\delta,\psi_I) - (\partial_t y_\delta - f - Bu_\delta,\psi_I))\ dt\\
&+ \sum_{k=0}^{N-1}([y_\delta]_k,(\psi-\psi_I)_k^+) + ((y_\delta - y^{u_\delta})_0^+,\psi_0^+) - ([y_\delta]_0,\psi_0^+)\\
=\ & \sum_{k=1}^{N}\int_{I_k}\left(\partial_t y_\delta - f - Bu_\delta - \operatorname{div}(A\nabla y_\delta) + \frac{[y_\delta]_{k-1}}{\Delta t_k},\psi-\psi_I\right)\ dt\\
&+ \int_0^T \sum_{l\in\partial T^h}\int_l [(A\nabla y_\delta)\cdot\mathbf{n}](\psi-\psi_I)\,dt\\
&+ \sum_{k=1}^{N}\int_{I_k}\left(\frac{[y_\delta]_{k-1}}{\Delta t_k},\psi_{k-1}^+ - \psi + \psi_I - (\psi_I)_{k-1}^+\right)\ dt\\
&+ ((y_\delta)_0^- - (y^{u_\delta})_0^+,\psi_0^+) := \sum_{i=1}^{4} J_i.
\end{aligned}$$

Let

$$r_y(x,t)\Big|_{I_k} := \partial_t y_\delta - f - Bu_\delta - \operatorname{div}(A\nabla y_\delta) + \frac{[y_\delta]_{k-1}}{\Delta t_k}.$$

Then, as in (25), (27), and (28),

(32)
$$\begin{aligned}
J_1 &= \sum_{k=1}^{N}\int_{I_k}(r_y,(\pi_{h,k}-I)\pi_k\psi + (\pi_k - I)\psi)\,dt\\
&= \sum_{k=1}^{N}\int_{I_k}((r_y,(\pi_{h,k}-I)\pi_k\psi) + ((\pi_k - I)(f+\operatorname{div}(A\nabla y_\delta)),(\pi_k-I)\psi))\ dt\\
&\le C\sum_{k=1}^{N}\sum_{\tau\in T^{h,k}}\int_{I_k}\left(h_\tau^4\|r_y\|^2_{0,\tau} + \Delta t_k^2\|(\pi_k - I)(f+\operatorname{div}(A\nabla y_\delta))\|^2_{0,\tau}\right)\ dt\\
&\quad + \sigma(\|D^2(\pi_k\psi)\|^2_{L^2(0,T;L^2(\Omega))} + \|\partial_t\psi\|^2_{L^2(0,T;L^2(\Omega))})\\
&\le C(\eta_4^2+\eta_5^2) + C\sigma\|y^{u_\delta} - y_\delta\|^2_{L^2(0,T;L^2(\Omega))},
\end{aligned}$$

(33)
$$J_2 = \int_0^T \sum_{l\in\partial T^h}\int_l [(A\nabla y_\delta)\cdot\mathbf{n}](\psi-\psi_I)\,dt \le C\eta_6^2 + \sigma\|y^{u_\delta} - y_\delta\|^2_{L^2(0,T;L^2(\Omega))},$$

(34)
$$\begin{aligned}
J_3 &= \sum_{k=1}^{N}\int_{I_k}\left(\frac{[y_\delta]_{k-1}}{\Delta t_k},\psi_{k-1}^+ - \psi + \psi_I - (\psi_I)_{k-1}^+\right)\ dt\\
&\le C\eta_7^2 + \sigma\|y^{u_\delta} - y_\delta\|^2_{L^2(0,T;L^2(\Omega))},
\end{aligned}$$

and

$$J_4 = ((y_\delta)_0^- - (y^{u_\delta})_0^+, \psi_0^+) \le C\eta_0^2 + \sigma \|y^{u_\delta} - y_\delta\|^2_{L^2(0,T;L^2(\Omega))}.$$

Hence

$$\|y_\delta - y^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))} \le C \sum_{i=0,4\text{-}7} \eta_i^2. \tag{35}$$

We complete the proof by combining the estimates (29) and (35). ∎

From Lemmas 3.3 and 3.5, we have the following a posteriori error estimates.

THEOREM 3.1. *Let* (y, p, u) *and* $(y_\delta, p_\delta, u_\delta)$ *be the solutions of* (2) *and* (6). *Assume that the conditions in Lemmas* 3.3–3.5 *are valid; then*

$$\|u - u_\delta\|^2_{L^2(0,T;L^2(\Omega))} + \|y - y_\delta\|^2_{L^2(0,T;L^2(\Omega))} + \|p - p_\delta\|^2_{L^2(0,T;L^2(\Omega))} \le C \sum_{i=0}^{7} \eta_i^2,$$

where η_i *are defined in Lemmas* 3.3 *and* 3.5.

Proof. We obtain from (13), (35), and (29) that

$$\|u - u_\delta\|^2_{L^2(0,T;L^2(\Omega))} + \|y^{u_\delta} - y_\delta\|^2_{L^2(0,T;L^2(\Omega))} + \|p^{u_\delta} - p_\delta\|^2_{L^2(0,T;L^2(\Omega))} \le C \sum_{i=0}^{7} \eta_i^2.$$

Then the desired results follows from the triangle inequality and

$$\|p - p^{u_\delta}\|_{L^2(0,T;L^2(\Omega))} \le C\|y - y^{u_\delta}\|_{L^2(0,T;L^2(\Omega))} \le C\|u - u_\delta\|_{L^2(0,T;L^2(\Omega))}, \tag{36}$$

which can be derived from (17) and (18). ∎

It seems to be difficult to derive any lower error bounds for the control problem. As matter of fact, there seem to be no good lower a posteriori error bounds in the literature even for the full backward-Euler finite element approximation of linear parabolic equations. The main difficulty seems to be that the properties of the time variable and its discretization are quite different from those of the space variables. Novel techniques are yet to be developed to derive lower bounds for such mixed approximations.

REMARK 3.2. *It is clear that the above a posteriori error estimator consists of two parts. The* η_1^2 *part results from the approximation error of the inequality in the optimality condition* (2). *The other (more familiar) part* (η_i^2 $(i = 0, 2, \ldots, 7)$) *is contributed from the approximation error of the state and costate equations and in this sense is more or less standard. Among them,* η_1^2 *mainly indicates the approximation error for the control, and the other part mainly reflects the approximation error for the state and costate.*

The part (η_i^2 $(i = 0, 2, \ldots, 7)$) *can be further divided into two parts: one from the approximation error of the state equation and the other from that of the costate equation. Clearly, a posteriori error estimators obtained solely from the state equation, which only present the part contributed from the state equation, may fail to reflect the main approximation error of the optimal control problem and thus fail to yield efficient mesh refinements.*

The above error estimates are applicable to a wide range of control problems. It may be possible to further improve them in some individual cases, as will be seen in the next section. To this end, it is clear that one needs to derive improved error

estimates for the approximation of the inequality in (2)*, and thus one requires explicit information on the structure of* K*.*

REMARK 3.3. *It is generally difficult to know the exact bounding constant* C *in Theorem* 3.1*, as is true for most a posteriori error estimates of residual type. The constant is contributed from those in the interpolation results (e.g., Lemmas* 3.1–3.2*), the stability results (e.g., Lemmas* 3.3–3.4*), and the Sobolev embedding theorems. For simpler situations, it may be possible to trace down all those constants and to give the bounding constant good upper bounds; see* [9] *for some of the latest advances on this aspect. Generally this is a complex procedure. On the other hand, a posteriori estimators of residual type can be (actually have widely been) used to guide mesh refinements without having exact knowledge on the bounding constants, provided they are not too large. It seems that the magnitude of the bounding constants does not cause any serious problems in guiding mesh refinements for elliptic and parabolic equations, although it does bring up serious concerns in CFD (see* [23]*), since it can indeed be extremely large there.*

In our case, it seems that the bounding constant in Theorems 3.1–3.2 *will have a similar magnitude as those for the standard parabolic equation case, as the only new contribution here is from the constant* C *in Lemma* 3.3*. This constant can be traced down in Examples* 3.1–3.2*, which in turns depends on the bounding constant for the integral averaging interpolator* $\pi^a_{\delta,k}$*. It is known that the bounding constant associated with* $\pi^a_{\delta,k}$ *will not be very large; see* [9] *for the details.*

REMARK 3.4. *It is not straightforward to develop suitable implementation techniques for (x-t) mesh adaptivity of parabolic control problems. To the best of our knowledge, there seems to be no existing work in the literature, even using the same meshes for the state and the control. For instance, it seems impossible to simply extend the mesh adaptivity techniques developed for evolutional equations (e.g., parabolic or Navier–Stokes equations) to the control problem that we have just studied. Although the state equation is evolutional, the* optimal control problem *itself is clearly not. It is impossible to solve the control problem step by step in time, although this is possible for the state equation. This calls for new implementation techniques on mesh adaptivity for the optimal control governed by evolutional state equations. From the above analysis of* η_1^2 *(*η_i^2*), it is also clear that the most suitable implementation, and thus the optimal mesh refinements will greatly depend on what is the most important quantity to be computed in a particular control problem. It also depends on the structure of the meshes used in the computations. Furthermore, as some large discretized optimization problems may need to be repeatedly solved, one may have to use a suitable multigrids method together with mesh adaptivity. Issues like which items in the estimator are more important and how to pick up the constant* C *are also important. It is clear that a systematic study of this is much needed. These issues will be investigated in our future research.*

3.2. $L^\infty(L^2)$ error estimates. In some adaptive schemes, it is more desirable to have $L^\infty(L^2)$ estimates. In this subsection, we give error estimates in $L^\infty(L^2)$-norm. Concretely, we shall use the norm of the following form:

$$\|v\|_{I_k,Q} = \left\{ \frac{1}{\Delta t_k} \int_{I_k} \|v(t)\|^2_{0,Q}\, dt \right\}^{1/2}, \qquad Q = \Omega_U, \tau_U, \Omega, \tau, l.$$

We now need to consider the following dual equations for any $1 \le k \le N-1$:

$$\begin{cases} \partial_t \varphi - \operatorname{div}(A\nabla\varphi) = 0, & (x,t) \in \Omega \times (t_k, T], \\ \varphi|_{\partial\Omega} = 0, \quad t \in [t_k, T], & \varphi(x,t_k) = \varphi_*(x), \quad x \in \Omega, \end{cases} \tag{37}$$

and

$$\begin{cases} -\partial_t \psi - \operatorname{div}(A^*\nabla\psi) = 0, & (x,t) \in \Omega \times [0, t_k), \\ \psi|_{\partial\Omega} = 0, \quad t \in [0, t_k], & \psi(x,t_k) = \psi_*(x), \ x \in \Omega. \end{cases} \tag{38}$$

We have the following stability results [12].

LEMMA 3.6. *Assume that Ω is a convex domain. Let φ and ψ be the solutions of* (37) *and* (38)*, respectively. Then*

$$\begin{aligned} &\|\varphi\|_{L^\infty(t_k,T;L^2(\Omega))} \le C\|\varphi_*\|_{L^2(\Omega)}, \\ &\|\varphi\|_{L^2(t_k,t_k+\varepsilon;L^2(\Omega))} \le C\sqrt{\varepsilon}\|\varphi_*\|_{L^2(\Omega)}, \quad 0 < \varepsilon < T - t_k, \\ &\|\nabla\varphi\|_{L^2(t_k,T;L^2(\Omega))} \le C\|\varphi_*\|_{L^2(\Omega)}, \\ &\|\sqrt{t-t_k}\,|D^2\varphi|\|_{L^2(t_k,T;L^2(\Omega))} \le C\|\varphi_*\|_{L^2(\Omega)}, \\ &\|\sqrt{t-t_k}\,\partial_t\varphi\|_{L^2(t_k,T;L^2(\Omega))} \le C\|\varphi_*\|_{L^2(\Omega)} \end{aligned}$$

and

$$\begin{aligned} &\|\psi\|_{L^\infty(0,t_k;L^2(\Omega))} \le C\|\psi_*\|_{L^2(\Omega)}, \\ &\|\psi\|_{L^2(t_k-\varepsilon,t_k;L^2(\Omega))} \le C\sqrt{\varepsilon}\|\psi_*\|_{L^2(\Omega)}, \quad 0 < \varepsilon < t_k, \\ &\|\nabla\psi\|_{L^2(0,t_k;L^2(\Omega))} \le C\|\psi_*\|_{L^2(\Omega)}, \\ &\|\sqrt{t_k-t}\,|D^2\psi|\|_{L^2(0,t_k;L^2(\Omega))} \le C\|\psi_*\|_{L^2(\Omega)}, \\ &\|\sqrt{t_k-t}\,\partial_t\psi\|_{L^2(0,t_k;L^2(\Omega))} \le C\|\psi_*\|_{L^2(\Omega)}, \end{aligned}$$

where $D^2 v = \max_{1\le i,j\le n} |\partial^2 v/\partial x_i \partial x_j|$.

THEOREM 3.2. *Let (y,p,u) and (y_h,p_h,u_h) be the solutions of* (2) *and* (6)*, respectively. Assume that the conditions in Theorem* 3.1 *and Lemma* 3.6 *are valid; then*

$$\max_{1\le k\le N} \left(\|u-u_h\|^2_{I_k,\Omega_U} + \|y-y_h\|^2_{I_k,\Omega} + \|p-p_h\|^2_{I_k,\Omega}\right) \le C\sum_{i=0}^{8} \mathfrak{N}_i^2,$$

where

$$\begin{aligned} \mathfrak{N}_0^2 &= \|y_0^h - y_0\|^2_{0,\Omega}, \\ \mathfrak{N}_1^2 &= \max_{1\le k\le N} \sum_{\tau_U \in T_U^{h,k}} \left(h^2_{\tau_U}\|\nabla(h'(u_\delta) + B^*p_\delta)\|^2_{I_k,\tau_U} + \Delta t_k^2 \|\partial_t(h'(u_\delta) + B^*p_\delta)\|^2_{I_k,\tau_U}\right), \\ \mathfrak{N}_2^2 &= \max_{1\le k\le N} \sum_{\tau\in T^{h,k}} h_\tau^2(\Delta t_k + L_N h_\tau^2) \left\| \partial_t p_\delta + g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta) + \frac{[p_\delta]_k}{\Delta t_k}\right\|^2_{I_k,\tau}, \\ \mathfrak{N}_3^2 &= \max_{1\le k\le N} \sum_{\tau\in T^{h,k}} \Delta t_k^2 \left\|(\pi_k - I)(g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta))\right\|^2_{I_k,\tau}, \\ \mathfrak{N}_4^2 &= \max_{1\le k\le N} \sum_{l\in\partial T^{h,k}} h_l(\Delta t_k + L_N h_l^2) \left\|[(A^*\nabla p_\delta)\cdot \mathbf{n}]\right\|^2_{I_k,l}, \end{aligned}$$

$$
\begin{aligned}
\mathfrak{N}_5^2 &= \max_{1\le k\le N}\sum_{\tau\in T^{h,k}} h_\tau^2(\Delta t_k + L_N h_\tau^2)\left\|\partial_t y_\delta - f - Bu_\delta - \operatorname{div}(A\nabla y_\delta) + \frac{[y_\delta]_{k-1}}{\Delta t_k}\right\|_{I_k,\tau}^2,\\
\mathfrak{N}_6^2 &= \max_{1\le k\le N}\sum_{\tau\in T^{h,k}} \Delta t_k^2\,\|(\pi_k - I)(f + \operatorname{div}(A\nabla y_\delta))\|_{I_k,\tau}^2,\\
\mathfrak{N}_7^2 &= \max_{1\le k\le N}\sum_{l\in \partial T^{h,k}} h_l(\Delta t_k + L_N h_l^2)\,\|[(A\nabla y_\delta)\cdot \mathbf{n}]\|_{I_k,l}^2,\\
\mathfrak{N}_8^2 &= \max_{1\le k\le N}(\|[y_\delta]_{k-1}\|_{0,\Omega}^2 + \|[p_\delta]_k\|_{0,\Omega}^2),
\end{aligned}
$$

where

$$
L_N = \max\left\{\max_{1\le k\le N-2}\sum_{k'=k+2}^{N}\frac{\Delta t_{k'}}{t_{k'-1}-t_k},\quad \max_{2\le k\le N}\sum_{k'=1}^{k-1}\frac{\Delta t_{k'}}{t_k - t_{k'}}\right\}.
$$

Proof. We first consider $\|u-u_h\|_{L^2(I_k;L^2(\Omega_U))}$. As in (16) and (20), for any $v\in K^\delta$, we have

$$
\begin{aligned}
& c\|u-u_\delta\|_{L^2(I_k;L^2(\Omega_U))}^2 \le \int_{I_k}(h'(u),u-u_\delta)_U\,dt - \int_{I_k}(h'(u_\delta),u-u_\delta)_U\,dt\\
\le\ & \int_{I_k}(h'(u_\delta)+B^*p_\delta, v-u)_U\,dt + \int_{I_k}(B^*(p_\delta-p),u-u_\delta)_U\,dt\\
\le\ & C\int_{I_k}\left(h_{\tau_U}^2|h'(u_\delta)+B^*p_\delta|_{1,\tau_U}^2 + \Delta t_k^2\|\partial_t(h'(u_\delta)+B^*p_\delta)\|_{0,\tau_U}^2\right)dt\\
& + C\left(\|p_\delta - p^{u_\delta}\|_{L^2(I_k;L^2(\Omega))}^2 + \|p^{u_\delta}-p\|_{L^2(I_k;L^2(\Omega))}^2\right) + \frac{c}{2}\|u-u_\delta\|_{L^2(I_k;L^2(\Omega_U))}^2.
\end{aligned}
$$

It is easy to see from (18) and (8) that

$$
\begin{aligned}
\|p^{u_\delta}-p\|_{I_k,\Omega} &\le \|p^{u_\delta}-p\|_{L^\infty(0,T;L^2(\Omega))}^2 \le C\|y^{u_\delta}-y\|_{L^2(0,T;L^2(\Omega))}^2\\
&\le C\|u-u_\delta\|_{L^2(0,T;L^2(\Omega_U))}^2.
\end{aligned}
$$

We thus obtain

$$
\|u-u_\delta\|_{I_k,\Omega_U}^2 \le C\left(\mathfrak{N}_1^2 + \|p_\delta - p^{u_\delta}\|_{I_k,\Omega}^2\right) + C\|u-u_\delta\|_{L^2(0,T;L^2(\Omega_U))}^2. \tag{39}
$$

The last term above has been estimated in Theorem 3.1.

We consider $\|p_\delta - p^{u_\delta}\|_{I_k,\Omega}^2$ for any $1\le k\le N$. Let φ be the solution of the dual problem

$$
\begin{cases}
\partial_t\varphi - \operatorname{div}(A\nabla\varphi) = p_\delta - p^{u_\delta}, & (x,t)\in\Omega\times I_k,\\
\varphi|_{\partial\Omega} = 0,\quad t\in I_k, & \varphi(x,t_{k-1}) = 0,\quad x\in\Omega,
\end{cases}
$$

and let φ_I be defined as in (23). Then, similarly to (24),

$$\begin{aligned}
&\|p_\delta - p^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))} = \int_{I_k} (p_\delta - p^{u_\delta}, \partial_t\varphi - \operatorname{div}(A\nabla\varphi))\,dt\\
=\ & \int_{I_k} (-(\partial_t(p_\delta - p^{u_\delta}),\varphi) + a(\varphi, p_\delta - p^{u_\delta}))\ dt + (p_\delta - p^{u_\delta},\varphi)^-_k\\
=\ & \int_{I_k} (-(\partial_t p_\delta + g'(y^{u_\delta}),\varphi) + a(\varphi,p_\delta) - a(\varphi_I,p_\delta) + (\partial_t p_\delta + g'(y_\delta),\varphi_I))\ dt\\
&+([p_\delta]_k,(\varphi_I)^-_k) + (p_\delta - p^{u_\delta},\varphi)^-_k\\
=\ & \int_{I_k} \left(\partial_t p_\delta + g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta) + \frac{[p_\delta]_k}{\Delta t_k}, \varphi_I - \varphi\right) dt + \int_{I_k} (g'(y_\delta) - g'(y^{u_\delta}),\varphi)\,dt\\
&+\int_{I_k}\sum_{l\in\partial T^h}\int_l [(A^*\nabla p_\delta)\cdot\mathbf{n}](\varphi-\varphi_I)\,dt + \int_{I_k}\left(\frac{[p_\delta]_k}{\Delta t_k}, (\varphi_I)^-_k - \varphi_I + \varphi - \varphi^+_{k-1}\right) dt\\
&+(p_\delta - p^{u_\delta},\varphi)^-_k := \sum_{i=1}^5 \mathcal{I}_i.
\end{aligned}$$

It is easy to see that $\mathcal{I}_i$ ($i = 1$–4) can be estimated in the same way as in (25)–(28) such that

$$\begin{aligned}
(40)\quad \mathcal{I}_1 &\le C\sum_{\tau\in T^h}\int_{I_k}\left(h_\tau^4\|r_p\|^2_{0,\tau} + \Delta t_k^2\|(\pi_k - I)(g'(y_\delta) + \operatorname{div}(A^*\nabla p_\delta))\|^2_{0,\tau}\right) dt\\
&\quad +\sigma\|p_\delta - p^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))},\\
\mathcal{I}_2 &\le C\|y_\delta - y^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))} + \sigma\|p_\delta - p^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))},\\
(41)\quad \mathcal{I}_3 &\le C\sum_{l\in\partial T^h}\int_{I_k} h_l^3\|[(A^*\nabla p_\delta)\cdot\mathbf{n}]\|^2_{0,l}\,dt + \sigma\|p_\delta - p^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))},\\
(42)\quad \mathcal{I}_4 &\le \Delta t_k\|[p_\delta]_k\|^2_{0,\Omega} + \sigma\|p_\delta - p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))}.
\end{aligned}$$

We bound $\mathcal{I}_5$ by

$$(43)\qquad \begin{aligned}
\mathcal{I}_5 &\le \|(p_\delta - p^{u_\delta})^-_k\|_{0,\Omega}\sqrt{\Delta t_k}\,\|\partial_t\varphi\|_{L^2(I_k;L^2(\Omega))}\\
&\le C\Delta t_k\|(p_\delta - p^{u_\delta})^-_k\|^2_{0,\Omega} + \sigma\|p_\delta - p^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))}.
\end{aligned}$$

Thus, the above estimates give

$$(44)\qquad \|p_\delta - p^{u_\delta}\|^2_{I_k,\Omega} \le C\left(\sum_{i=2\text{–}4,8}\mathfrak{N}_i^2 + \|y_\delta - y^{u_\delta}\|^2_{I_k,\Omega} + \|(p_\delta - p^{u_\delta})^-_k\|^2_{0,\Omega}\right).$$

We then consider $\|y_\delta - y^{u_\delta}\|^2_{I_k,\Omega}$. Let ψ be the solution of the dual problem

$$\begin{cases}
-\partial_t\psi - \operatorname{div}(A^*\nabla\psi) = y_\delta - y^{u_\delta}, & (x,t)\in\Omega\times I_k,\\
\psi|_{\partial\Omega} = 0, \quad t\in I_k, & \psi(x,t_k) = 0,\ x\in\Omega,
\end{cases}$$

and let ψ_I be defined as in (30). Then, similarly to (31), for $1 \le k \le N$,

$$\begin{aligned}
&\|y_\delta - y^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))} = \int_{I_k} (y_\delta - y^{u_\delta}, -\partial_t\psi - \operatorname{div}(A^*\psi))\,dt\\
=\ & \int_{I_k} ((\partial_t(y_\delta - y^{u_\delta}),\psi) + a(y_\delta - y^{u_\delta},\psi))\ dt + (y_\delta - y^{u_\delta},\psi)^+_{k-1}\\
=\ & \int_{I_k} ((\partial_t y_\delta - f - Bu_\delta,\psi) + a(y_\delta,\psi) - a(y_\delta,\psi_I) - (\partial_t y_\delta - f - Bu_\delta,\psi_I))\ dt\\
& -([y_\delta]_{k-1},(\psi_I)^+_{k-1}) + (y_\delta - y^{u_\delta},\psi)^+_{k-1}\\
=\ & \int_{I_k} \left(\partial_t y_\delta - f - Bu_\delta - \operatorname{div}(A\nabla y_\delta) + \frac{[y_\delta]_{k-1}}{\Delta t_k}, \psi - \psi_I\right) dt\\
& + \int_{I_k} \sum_{l\in\partial T^h} \int_l [(A\nabla y_\delta)\cdot\mathbf{n}](\psi - \psi_I)\,dt + \int_{I_k} \left(\frac{[y_\delta]_{k-1}}{\Delta t_k}, \psi_k^- - \psi + \psi_I - (\psi_I)^+_{k-1}\right) dt\\
& + (y_\delta - y^{u_\delta},\psi)^+_{k-1} := \sum_{i=1}^4 \mathcal{J}_i,
\end{aligned}$$

where $\mathcal{J}_i$ ($i = 1$–3) can be estimated as in (40)–(43) so that

$$\begin{aligned}
\mathcal{J}_1 \le\ & C\int_{I_k} \left(h_\tau^4\|r_y\|^2_{0,\tau} + \Delta t_k^2\|(\pi_k - I)(f + \operatorname{div}(A\nabla y_\delta))\|^2_{0,\tau}\right) dt\\
& + \sigma\|y_\delta - y^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))},\\
\mathcal{J}_2 \le\ & C\sum_{l\in\partial T^h} \int_{I_k} h_l^3\|[(A\nabla y_\delta)\cdot\mathbf{n}]\|^2_{0,l}\,dt + \sigma\|y_\delta - y^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))},\\
\mathcal{J}_3 \le\ & \Delta t_k\|[y_\delta]_{k-1}\|^2_{0,\Omega} + \sigma\|y_\delta - y^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))},\\
\mathcal{J}_4 \le\ & C\Delta t_k\|(y_\delta - y^{u_\delta})^+_{k-1}\|^2_{0,\Omega} + \sigma\|y_\delta - y^{u_\delta}\|^2_{L^2(I_k;L^2(\Omega))}.
\end{aligned}$$

Therefore,

$$\|y_\delta - y^{u_\delta}\|^2_{I_k,\Omega} \le C\left(\sum_{i=5\text{–}8} \mathfrak{N}_i^2 + \|(y_\delta - y^{u_\delta})^+_{k-1}\|^2_{0,\Omega}\right). \tag{45}$$

We need to further consider $\|(p_\delta - p^{u_\delta})^-_k\|^2_{0,\Omega}$ and $\|(y_\delta - y^{u_\delta})^+_{k-1}\|^2_{0,\Omega}$ ($1 \le k \le N$). We note that $\|(p_\delta - p^{u_\delta})^-_N\|^2 = \|[p_\delta]_N\|^2_{0,\Omega} \le \mathfrak{N}_8^2$. For any $1 \le k \le N-1$, let φ be the solution of (37) with $\varphi_* = (p_\delta - p^{u_\delta})^-_k$ and φ_I be defined as in (23). Then, by (37), (15), and (6),

$$\begin{aligned}
&\|(p_\delta - p^{u_\delta})^-_k\|^2_{0,\Omega} = ((p_\delta - p^{u_\delta})^-_k,\varphi_*) - ((p_\delta - p^{u_\delta})^+_k,\varphi_*) + (p_\delta - p^{u_\delta},\varphi)^+_k\\
=\ & \int_{t_k}^T (-(\partial_t(p_\delta - p^{u_\delta}),\varphi) + a(\varphi, p_\delta - p^{u_\delta}))\ dt - \sum_{k'=k+1}^N ([p_\delta]_{k'},\varphi^-_{k'}) - ([p_\delta]_k,\varphi_*)\\
=\ & \int_{t_k}^T (-(\partial_t p_\delta + g'(y^{u_\delta}),\varphi) + a(\varphi,p_\delta) - a(\varphi_I,p_\delta) + (\partial_t p_\delta + g'(y_\delta),\varphi_I))\ dt\\
& + \sum_{k'=k+1}^N ([p_\delta]_{k'},(\varphi_I - \varphi)^-_{k'}) - ([p_\delta]_k,\varphi_*)
\end{aligned}$$

$$
\begin{aligned}
= & \sum_{k'=k+1}^{N} \int_{I_{k'}} \Big(\partial_t p_\delta + g'(y_\delta) + \operatorname{div}(A^* \nabla p_\delta) + \frac{[p_\delta]_{k'}}{\Delta t_{k'}}, \varphi_I - \varphi \Big) \\
& + \int_{t_k}^{T} (g'(y_\delta) - g'(y^{u_\delta}), \varphi)\, dt \\
& + \int_{t_k}^{T} \sum_{l \in \partial T^h} \int_l [(A^* \nabla p_\delta) \cdot \mathbf{n}](\varphi - \varphi_I)\, dt \\
& + \sum_{k'=k+1}^{N} \int_{I_{k'}} \left(\frac{[p_\delta]_{k'}}{\Delta t_{k'}}, (\varphi_I)_{k'}^- - \varphi_I + \varphi - \varphi_{k'}^- \right) \\
& - ([p_\delta]_k, \varphi_*) := \sum_{i=1}^{5} \mathit{I}_i.
\end{aligned}
$$

We have to treat the cases in which t_k is near T and away from T differently. For simplicity, let $c_k = 1$ for $1 \le k \le N-2$ and $c_{N-1} = 0$. We decompose I_1 as follows:

$$
\begin{aligned}
\mathit{I}_1 &= \left(\sum_{k'=k+1} + c_k \sum_{k'=k+2}^{N} \right) \int_{I_{k'}} ((r_p, (\pi_{h,k} - I)\pi_k \varphi) + ((\pi_k - I) r_p, (\pi_k - I)\varphi))\, dt \\
&:= \mathit{I}_{11} + c_k \mathit{I}_{12}.
\end{aligned}
$$

By Lemmas 3.1 and 3.6, we have

$$
\begin{aligned}
\mathit{I}_{11} &\le C \int_{I_{k+1}} \sum_{\tau \in T^h} h_\tau \|r_p\|_{0,\tau} |\pi_k \varphi|_{1,\tau}\, dt + \int_{I_{k+1}} \|(\pi_k - I) r_p\|_{0,\Omega} \|\varphi\|_{0,\Omega}\, dt \\
&\le C \int_{I_{k+1}} \sum_{\tau \in T^h} h_\tau^2 \|r_p\|_{0,\tau}^2\, dt + \sigma \int_{I_{k+1}} |\varphi|_{1,\Omega}^2\, dt \\
&\quad + C \Delta t_{k+1} \int_{I_{k+1}} \|(\pi_k - I) r_p\|_{0,\Omega}^2\, dt + \sigma \|\varphi\|_{L^\infty(I_{k+1}; L^2(\Omega))}^2 \\
&\le C(\mathfrak{N}_2^2 + \mathfrak{N}_3^2) + C\sigma \|(p_\delta - p^{u_\delta})_k^-\|_{0,\Omega}^2,
\end{aligned} \tag{46}
$$

and

$$
\begin{aligned}
\mathit{I}_{12} &\le C \sum_{k'=k+2}^{N} \left(\int_{I_{k'}} \sum_{\tau \in T^h} h_\tau^2 \|r_p\|_{0,\tau} |\pi_k \varphi|_{2,\tau}\, dt \right. \\
&\qquad \left. + \Delta t_{k'} \|(\pi_k - I) r_p\|_{L^2(I_{k'}; L^2(\Omega))} \|\partial_t \varphi\|_{L^2(I_{k'}; L^2(\Omega))} \right) \\
&\le C \sum_{k'=k+2}^{N} \int_{I_{k'}} (t_{k'-1} - t_k)^{-1} \sum_{\tau \in T^h} h_\tau^4 \|r_p\|_{0,\tau}^2\, dt + \sigma \int_{t_{k+1}}^{T} (t - t_k) \|D^2 \varphi\|_{0,\Omega}^2\, dt \\
&\quad + C \sum_{k'=k+2}^{N} \Delta t_{k'} \|(\pi_k - I) r_p\|_{L^2(I_{k'}; L^2(\Omega))} \frac{1}{\sqrt{t_{k'-1} - t_k}} \|\sqrt{t - t_k} \partial_t \varphi\|_{L^2(I_{k'}; L^2(\Omega))} \\
&\le C L_N \max_{k+2 \le k' \le N} \sum_{\tau \in T^h} \left(h_\tau^4 \|r_p\|_{I_k, \tau}^2 + \Delta t_{k'}^2 \|(\pi_k - I) r_p\|_{I_k, \tau}^2 \right)
\end{aligned} \tag{47}
$$

$$
\begin{aligned}
&+ \sigma \int_{t_{k+1}}^{T} |t - t_k| \left(\|D^2 \varphi\|_{0,\Omega}^2 + \|\partial_t \varphi\|_{0,\Omega}^2 \right) dt \\
&\le C(\mathfrak{N}_2^2 + \mathfrak{N}_3^2) + C\sigma \|(p_\delta - p^{u_\delta})_k^- \|_{0,\Omega}^2.
\end{aligned}
$$

It follows from (8) and Lemma 3.6 that

$$
\mathit{\Pi}_2 \le C\|y_\delta - y^{u_\delta}\|_{L^2(0,T;L^2(\Omega))}^2 + \sigma \|(p_\delta - p^{u_\delta})_k^- \|_{0,\Omega}^2.
$$

By using (11) and Lemma 3.1, we can estimate $\mathit{\Pi}_3$ in the same way as for $\mathit{\Pi}_1$ such that

(48)

$$
\begin{aligned}
\mathit{\Pi}_3 &= \left(\int_{t_k}^{t_{k+1}} + c_k \int_{t_{k+1}}^{T} \right) \sum_{l \in \partial T^h} \int_l [(A^* \nabla p_\delta) \cdot \mathbf{n}](\varphi - \pi_{h,k}\varphi)\, dt \\
&\le C \int_{t_k}^{t_{k+1}} \sum_{l \in \partial T^h} h_l \|[(A^* \nabla p_\delta) \cdot \mathbf{n}]\|_{0,l}^2\, dt + \sigma \int_{t_k}^{t_{k+1}} |\varphi|_{1,\Omega}^2\, dt \\
&\quad + Cc_k \int_{t_{k+1}}^{T} |t - t_k|^{-1} \sum_{l \in \partial T^h} h_l^3 \|[(A^* \nabla p_\delta) \cdot \mathbf{n}]\|_{0,l}^2\, dt + \sigma \int_{t_{k+1}}^{T} |t - t_k| \|D^2 \varphi\|_{0,\Omega}^2\, dt \\
&\le C\mathfrak{N}_4^2 + C\sigma \|(p_\delta - p^{u_\delta})_k^- \|_{0,\Omega}^2.
\end{aligned}
$$

We rewrite $\mathit{\Pi}_4$ as

$$
\mathit{\Pi}_4 = \left(\sum_{k'=k+1} + c_k \sum_{k'=k+2}^{N} \right) \int_{I_{k'}} \left(\frac{[p_\delta]_{k'}}{\Delta t_{k'}}, (\varphi_I)_{k'}^- - \varphi_I + \varphi - \varphi_{k'}^- \right) dt := \mathit{\Pi}_{41} + c_k \mathit{\Pi}_{42}.
$$

We then use Lemma 3.6 again to obtain

(49)

$$
\begin{aligned}
\mathit{\Pi}_{41} &= ([p_\delta]_{k+1}, (\varphi_I - \varphi)_{k+1}^-) + \int_{t_k}^{t_{k+1}} \left(\frac{[p_\delta]_{k+1}}{\Delta t_{k+1}}, \varphi - \pi_{h,k}\varphi \right) dt \\
&\le C\|[p_\delta]_{k+1}\|_{0,\Omega} (\Delta t_{k+1}^{-1/2} \|\varphi\|_{L^2(I_{k+1};L^2(\Omega))} + \|\varphi\|_{L^\infty(I_{k+1};L^2(\Omega))}) \\
&\le C\|[p_\delta]_{k+1}\|_{0,\Omega}^2 + \sigma \|(p^{u_\delta} - p_\delta)_k^- \|_{0,\Omega}^2,
\end{aligned}
$$

(50)

$$
\begin{aligned}
\mathit{\Pi}_{42} &= \sum_{k'=k+2}^{N} \int_{k'} \left(\frac{[p_\delta]_{k'}}{\Delta t_{k'}}, (\varphi_I)_{k'}^- - \varphi_I + \varphi - \varphi_{k'}^- \right) dt \\
&\le C \sum_{k'=k+2}^{N} \|[p_\delta]_{k'}\|_{0,\Omega} \sqrt{\Delta t_{k'}} \|\partial_t \varphi\|_{L^2(I_{k'};L^2(\Omega))} \\
&\le C \sum_{k'=k+2}^{N} \frac{\Delta t_{k'}}{t_{k'-1} - t_k} \|[p_\delta]_{k'}\|_{0,\Omega}^2 + \sigma \|\sqrt{t - t_k}\, \partial_t \varphi\|_{L^2(t_{k+1},T;L^2(\Omega))}^2 \\
&\le CL_N \max_{k+2 \le k' \le N} \|[p_\delta]_{k'}\|_{0,\Omega}^2 + C\sigma \|(p^{u_\delta} - p_\delta)_k^- \|_{0,\Omega}^2,
\end{aligned}
$$

and

$$
\mathit{\Pi}_5 \le C\|[p_\delta]_k\|_{0,\Omega}^2 + \sigma \|(p^{u_\delta} - p_\delta)_k^- \|_{0,\Omega}^2. \tag{51}
$$

We thus have shown that

$$
\|(p^{u_\delta} - p_\delta)_k^- \|_{0,\Omega}^2 \le C \sum_{i=2\text{–}4,8} \mathfrak{N}_i^2 + C\|y_\delta - y^{u_\delta}\|_{L^2(0,T;L^2(\Omega))}^2. \tag{52}
$$

The last term above has been estimated in Theorem 3.1.

It remains to estimate $\|(y_\delta - y^{u_\delta})_k^+\|_{0,\Omega}^2$ $(0 \le k \le N-1)$. Since

$$\|(y_\delta - y^{u_\delta})_0^+\|_{0,\Omega}^2 \le \|[y_\delta]_0\|_{0,\Omega}^2 + \|(y_\delta)_0^- - (y^{u_\delta})_0^+\|_{0,\Omega}^2 \le \mathfrak{N}_8^2 + \mathfrak{N}_0^2,$$

we need only to consider the cases of $1 \le k \le N-1$. Let ψ be the solution of (38) with $\psi_* = (y_\delta - y^{u_\delta})_k^+$ and ψ_I be defined as in (30). Then, by (38) and (14),

$$\begin{aligned}
&\|(y_\delta - y^{u_\delta})_k^+\|_{0,\Omega}^2 = ((y_\delta - y^{u_\delta})_k^+, \psi_*) - (y_\delta - y^{u_\delta}, \psi)_k^- + (y_\delta - y^{u_\delta}, \psi)_k^- \\
= &\int_0^{t_k} ((\partial_t(y_\delta - y^{u_\delta}), \psi) + a(y_\delta - y^{u_\delta}, \psi))\, dt + \sum_{k'=0}^{k-1} ([y_\delta]_{k'}, \psi_{k'}^+) \\
&+ (y_0^h - y_0, \psi_0^+) + ([y_\delta]_k, \psi_*) \\
= &\int_0^{t_k} ((\partial_t y_\delta - f - Bu_\delta, \psi) + a(y_\delta, \psi) - a(y_\delta, \psi_I) - (\partial_t y_\delta - f - Bu_\delta, \psi_I))\, dt \\
&+ \sum_{k'=1}^{k} ([y_\delta]_{k'-1}, (\psi - \psi_I)_{k'-1}^+) + (y_0^h - y_0, \psi_0^+) + ([y_\delta]_k, \psi_*) \\
= &\sum_{k'=1}^{k} \int_{I_{k'}} \left(\partial_t y_\delta - f - Bu_\delta - \operatorname{div}(A\nabla y_\delta + \frac{[y_\delta]_{k'-1}}{\Delta t_{k'}}, \psi - \psi_I\right) \\
&+ \int_0^{t_k} \left(\sum_{l \in \partial T^h} \int_l [(A\nabla y_\delta) \cdot \mathbf{n}](\psi - \psi_I) \right) dt \\
&+ \sum_{k'=1}^{k} \int_{I_{k'}} \left(\frac{[y_\delta]_{k'-1}}{\Delta t_{k'}}, \psi_{k'-1}^+ - \psi + \psi_I - (\psi_I)_{k'-1}^+ \right) \\
&+ (y_0^h - y_0, \psi_0^+) + ([y_\delta]_k, \psi_*) := \sum_{i=1}^{5} \mathcal{J}_i.
\end{aligned}$$

Let $c_1 = 0$ and $c_k = 1$ for $2 \le k \le N-1$. Then, as in (46)–(51),

$$\begin{aligned}
\mathcal{J}_1 &= \left(c_k \sum_{k'=1}^{k-1} + \sum_{k'=k} \right) \int_{I_{k'}} \{(r_y, (\pi_{h,k} - I)\pi_k \psi) + ((\pi_k - I) r_y, (\pi_k - I)\psi)\}\, dt \\
&\le C(\mathfrak{N}_5^2 + \mathfrak{N}_6^2) + \sigma \|(y^{u_\delta} - y_\delta)_k^+\|_{0,\Omega}^2, \\
\mathcal{J}_2 &= \left(c_k \sum_{k'=1}^{k-1} + \sum_{k'=k} \right) \int_{I_{k'}} \sum_{l \in \partial T^h} \int_l [(A\nabla y_\delta) \cdot \mathbf{n}](\psi - \pi_{h,k}\psi)\, dt \\
&\le C\mathfrak{N}_7^2 + \sigma \|(y^{u_\delta} - y_\delta)_k^+\|_{0,\Omega}^2, \\
\mathcal{J}_3 &= \left(c_k \sum_{k'=1}^{k-1} + \sum_{k'=k} \right) \int_{I_{k'}} \left(\frac{[y_\delta]_{k'-1}}{\Delta t_{k'}}, (\psi_I)_{k'}^- - \psi_I + \psi - \psi_{k'}^- \right) dt \\
&\le C\mathfrak{N}_8^2 + \sigma \|(y^{u_\delta} - y_\delta)_k^+\|_{0,\Omega}^2, \\
\mathcal{J}_4 &\le C\|y_0^h - y_0\|_{0,\Omega}^2 + \sigma \|(y^{u_\delta} - y_\delta)_k^+\|_{0,\Omega}^2, \\
\mathcal{J}_5 &\le C\|[y_\delta]_k\|_{0,\Omega}^2 + \sigma \|(y^{u_\delta} - y_\delta)_k^+\|_{0,\Omega}^2.
\end{aligned}$$

Hence

$$\|(y_\delta - y^{u_\delta})_k^+\|_{0,\Omega}^2 \le C \sum_{i=0,5-8} \mathfrak{N}_i^2. \tag{53}$$

We complete the proof by combining the estimates (39), (44), (45), (52), and (53) and the result of Theorem 3.1. ∎

In the rest of the section, we apply the results obtained to some model control problems. We only consider the piecewise constant finite element space for the approximation of the control.

Example 3.1. Consider the case $K = \{v \in X : v \geq \phi_0\}$, where ϕ_0 is a constant. Let $K^\delta = \{v \in X^\delta : v \geq \phi_0\}$. Then it is easy to see that $K^\delta \subset K$. Let v in Lemma 3.3 be such that $v|_{\tau_U^k \times I_k} = \pi^a_{\delta,k} u$, where $\pi^a_{\delta,k} u$ is the integral average of u on $\tau_U^k \times I_k$. Then $v = \pi^a_{\delta,k} u \in K^\delta$, and for $1 \leq k \leq N$,

$$\begin{aligned}
&\left| \int_{I_k} (h'(u_\delta) + B^* p_\delta, v - u)_U \, dt \right| = \left| \int_{I_k} (h'(u_\delta) + B^* p_\delta, \pi^a_{\delta,k} u - u)_U \, dt \right| \\
&= \left| \int_{I_k} ((\pi^a_{\delta,k} - I)(h'(u_\delta) + B^* p_\delta), (\pi^a_{\delta,k} - I)(u - u_\delta))_U \, dt \right| \\
&\leq C \int_{I_k} \sum_{\tau_U \in T_U^{h,k}} (h_{\tau_U} |h'(u_\delta) + B^* p_\delta|_{1,\tau_U} + \Delta t_k \|\partial_t (h'(u_\delta) + B^* p_\delta)\|_{0,\tau_U}) \|u - u_\delta\|_{0,\tau_U} \, dt.
\end{aligned}$$

Hence, the condition (12) in Lemma 3.3 is satisfied. Consequently the estimates obtained in Theorems 3.1–3.2 are applicable.

Example 3.2. Consider the case $K = \{v \in X : \int_{\Omega_U} v \geq 0\}$. Let $K^\delta = \{v \in X^\delta : \int_{\Omega_U} v \geq 0\}$. Then it is easy to see that $K^\delta \subset K$. Let v in Lemma 3.3 be defined as in Example 3.1. Then, the condition (12) in Lemma 3.3 is also satisfied.

4. Improved error estimates for the constraint of obstacle type. It seems to be difficult to further improve the estimates obtained in Theorems 3.1 and 3.2 without having structure information on the constraint set K. In this section, we consider a case where the constraint set is of obstacle type, which is met very frequently in real applications. We are then able to derive improved error estimates for the DG scheme of the finite element approximation to the parabolic optimal control problem (6). As mentioned in section 3, the essential step is to derive improved estimates for the approximation of the inequality in (2), via utilizing the structure information of K. Such improved estimates are found to be useful in computing elliptic control problems; see [27]. We shall only examine piecewise constant or piecewise linear control approximation.

We assume that the constraint on the control is an obstacle such that

$$K = \{v \in X : \ v \geq \phi \ \text{ a.e. in } \Omega_U \times (0, T]\},$$

where $\phi \in X$. We define the coincidence set (contact set) $\Omega_U^-(t)$ and the noncoincidence set (noncontact set) $\Omega_U^+(t)$ as follows:

$$\Omega_U^-(t) := \{x \in \Omega_U : u(x,t) = \phi(x,t)\}, \qquad \Omega_U^+(t) := \{x \in \Omega_U : u(x,t) > \phi(x,t)\}.$$

Let

$$K^\delta = \{v \in X^\delta : \ v \geq \phi^\delta \text{ in } \Omega_U \times (0, T]\}, \tag{54}$$

where $\phi^\delta \in X^\delta$ is an approximation to ϕ satisfying $\phi^\delta \geq \phi$. Hence, we have that $K^\delta \subset K$. In this section, we assume that

$$h(u) = \int_{\Omega_U} j(u),$$

where $j(\cdot)$ is a convex continuously differentiable function on $\mathbb{R}$. Then, it is easy to see that

$$\int_0^T (h'(u), v)_U = \int_0^T (j'(u), v)_U = \int_0^T \int_{\Omega_U} j'(u)v.$$

We shall assume the following uniform convexity condition:

$$(j'(t) - j'(s))(t - s) \geq c(t - s)^2 \quad \forall s, t \in \mathbb{R}.$$

It can be seen that the inequality in (2) is now equivalent to the following:

$$j'(u) + B^*p \geq 0, \quad u \geq \phi, \quad (j'(u) + B^*p)(u - \phi) = 0, \quad \text{a.e. in } \Omega_U \times (0, T]. \tag{55}$$

In order to have the improved a posteriori error estimate, we divide $\Omega_U \times (0, T]$ into the following three subsets:

$$\begin{aligned}
\Omega_\phi &= \{(x, t) \in \Omega_U \times (0, T] : \ (B^*p_\delta)(x, t) \leq -j'(\phi^\delta)\}, \\
\Omega_\phi^0 &= \{(x, t) \in \Omega_U \times (0, T] : \ (B^*p_\delta)(x, t) > -j'(\phi^\delta), \ u_\delta = \phi^\delta\}, \\
\Omega_\phi^+ &= \{(x, t) \in \Omega_U \times (0, T] : \ (B^*p_\delta)(x, t) > -j'(\phi^\delta), \ u_\delta > \phi^\delta\}.
\end{aligned}$$

Then, it is easy to see that the above three subsets do not overlap each other, and

$$\bar{\Omega}_U \times (0, T] = \bar{\Omega}_\phi \cup \bar{\Omega}_\phi^0 \cup \bar{\Omega}_\phi^+.$$

We shall show that $h'(u_\delta) + B^*p_\delta$ can be replaced by $(j'(u_\delta) + B^*p_\delta)|_{\Omega_\phi}$ in the error estimates. Note that $j'(u) + B^*p = 0$ when $u > \phi$. Thus in a sense, the set Ω_ϕ is an approximation of the noncoincidence set $\{(x, t) : x \in \Omega_U^+(t), \ t \in (0, T]\}$.

THEOREM 4.1. *Let (y, p, u) and $(y_\delta, p_\delta, u_\delta)$ be the solutions of (2) and (6), respectively. Assume that all the conditions of Lemma 3.5 hold, and K^δ is defined in (54) with $\phi \in L^2(0, T; L^2(\Omega_U))$. Moreover, assume that $j'(\cdot)$ and $g'(\cdot)$ are locally Lipschitz continuous. Then*

$$\|u_\delta - u\|^2_{L^2(0,T;L^2(\Omega_U))} + \|y_\delta - y\|^2_{L^2(0,T;L^2(\Omega))} + \|p_\delta - p\|^2_{L^2(0,T;L^2(\Omega))} \leq C \sum_{i=0}^{8} \hat{\eta}_i^2,$$

where $\hat{\eta}_i^2 = \eta_i^2$ ($i = 0, 2$–7) are given in Lemma 3.5 and

$$\begin{aligned}
\hat{\eta}_1^2 &= \int_{\Omega_\phi} |j'(u_\delta) + B^*p_\delta|^2, \\
\hat{\eta}_8^2 &= \|\phi - \phi^\delta\|^2_{0,\Omega_\phi^0}.
\end{aligned}$$

Proof. We consider $\|u_\delta - u\|^2_{L^2(0,T;L^2(\Omega_U))}$. From the uniform convexity of j, we

have that

$$
\begin{aligned}
&c\|u-u_\delta\|^2_{L^2(0,T;L^2(\Omega_U))} \le \int_0^T (j'(u)-j'(u_\delta),u-u_\delta)_U\\
&=\int_0^T (j'(u)+B^*p,u-u_\delta)_U+\int_0^T (j'(u_\delta)+B^*p_\delta,u_\delta-u)_U\\
&\quad+\int_0^T (B^*(p_\delta-p^{u_\delta}),u-u_\delta)_U+\int_0^T (B^*(p^{u_\delta}-p),u-u_\delta)_U \qquad (56)\\
&=\int_0^T (j'(u)+B^*p,u-u_\delta)_U+\int_0^T (j'(u_\delta)+B^*p_\delta,u_\delta-u)_U\\
&\quad+\int_0^T (B^*(p_\delta-p^{u_\delta}),u-u_\delta)_U+\int_0^T (y^{u_\delta}-y,y-y^{u_\delta})\\
&\le\int_0^T (j'(u)+B^*p,u-u_\delta)_U+\int_0^T (j'(u_\delta)+B^*p_\delta,u_\delta-u)_U\\
&\quad+\int_0^T (B^*(p_\delta-p^{u_\delta}),u-u_\delta)_U:=\sum_1^3 I_i.
\end{aligned}
$$

We first estimate I_1. Note that

$$
\begin{aligned}
&\int_0^T (j'(u)+B^*p,u-u_\delta)_U \qquad (57)\\
&=\int_{\Omega_\phi\cup\Omega_\phi^+}(j'(u)+B^*p)(u-u_\delta)+\int_{\Omega_\phi^0}(j'(u)+B^*p)(u-\phi^\delta).
\end{aligned}
$$

Let

$$
w=\begin{cases} u_\delta, & (x,t)\in\Omega_\phi\cup\Omega_\phi^+,\\ u, & (x,t)\in\Omega_\phi^0.\end{cases}
$$

Then, $w\in K$, and hence

$$
\int_{\Omega_\phi\cup\Omega_\phi^+}(j'(u)+B^*p)(u-u_\delta)=\int_0^T\int_{\Omega_U}(j'(u)+B^*p)(u-w)\le 0. \qquad (58)
$$

Note that $(j'(u)+B^*p)(u-\phi)=0$. We have that

$$
\begin{aligned}
\int_{\Omega_\phi^0}(j'(u)+B^*p)(u-\phi^\delta)&=\int_{\Omega_\phi^0}(j'(u)+B^*p)((u-\phi)+(\phi-\phi^\delta))\\
&=\int_{\Omega_\phi^0}(j'(u)+B^*p)(\phi-\phi^\delta). \qquad (59)
\end{aligned}
$$

It follows from (57)–(59) that

$$
I_1=\int_0^T (j'(u)+B^*p,u-u_\delta)_U\le\int_{\Omega_\phi^0}(j'(u)+B^*p)(\phi-\phi^\delta). \qquad (60)
$$

Next we estimate I_2. It is clear that

$$\begin{aligned}&\int_0^T (j'(u_\delta)+B^*p_\delta, u_\delta-u)_U\\ &=\int_{\Omega_\phi}(j'(u_\delta)+B^*p_\delta)(u_\delta-u)+\int_{\Omega_\phi^+}(j'(u_\delta)+B^*p_\delta)(u_\delta-u)\\ &\quad+\int_{\Omega_\phi^0}(j'(\phi_h)+B^*p_\delta)(\phi^\delta-u). \qquad (61)\end{aligned}$$

First it is easy to see that

$$\begin{aligned}\int_{\Omega_\phi}(j'(u_\delta)+B^*p_\delta)(u_\delta-u)&\le C\int_{\Omega_\phi}(j'(u_\delta)+B^*p_\delta)^2+C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}\\ &=C\hat\eta_1^2+C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}. \qquad (62)\end{aligned}$$

Second, let $\tau_U\times(t_i,t_{i+1}]$ be such that $u_\delta|_{\tau_U\times(t_i,t_{i+1}]}>\phi^\delta$; it follows from (6) that there exist $\epsilon>0$ and $\psi\in X^\delta$, such that $\psi\ge 0$, $\|\psi\|_{L^\infty(t_i,t_{i+1};L^\infty(\tau_U))}=1$, and

$$\int_{t_i}^{t_{i+1}}\int_{\tau_U}(j'(u_\delta)+B^*p_\delta)(u_\delta-(u_\delta-\epsilon\psi))=\epsilon\int_{t_i}^{t_{i+1}}\int_{\tau_U}(j'(u_\delta)+B^*p_\delta)\psi\le 0.$$

Note that on Ω_ϕ^+, $(j'(u_\delta)+B^*p_\delta)>(j'(\phi^\delta)+B^*p_\delta)>0$. We have that

$$\begin{aligned}&\int_{(\tau_U\times(t_i,t_{i+1}])\cap\Omega_\phi^+}|j'(u_\delta)+B^*p_\delta|\psi=\int_{(\tau_U\times(t_i,t_{i+1}])\cap\Omega_\phi^+}(j'(u_\delta)+B^*p_\delta)\psi\\ &\le-\int_{(\tau_U\times(t_i,t_{i+1}])\cap\Omega_\phi}(j'(u_\delta)+B^*p_\delta)\psi\le\int_{(\tau_U\times(t_i,t_{i+1}])\cap\Omega_\phi}|j'(u_\delta)+B^*p_\delta|.\end{aligned}$$

Let $\hat\tau_{Ut_i}$ be the reference element of $\tau_U\times(t_i,t_{i+1}]$, $\tau^0_{Ut_i}=(\tau_U\times(t_i,t_{i+1}])\cap\Omega_\phi^+$, and $\hat\tau^0_{Ut_i}\subset\hat\tau_{Ut_i}$ be the image of $\tau^0_{Ut_i}$. Let n be the dimension of Ω_U and $k_i=t_{i+1}-t_i$. Note that $j'(\cdot)$ is locally Lipschitz continuous. It follows from the equivalence of the norm in a finite dimensional space that

$$\begin{aligned}&\int_{\tau^0_{Ut_i}}|j'(u_\delta)+B^*p_\delta|^2\le Ch^n_{\tau_U}k_i\int_{\hat\tau^0_{Ut_i}}|j'(u_\delta)+B^*p_\delta|^2\\ &\le Ch^n_{\tau_U}k_i\left(\int_{\hat\tau^0_{Ut_i}}|j'(u_\delta)+B^*p_\delta|\psi\right)^2\le Ch^{-n}_{\tau_U}k_i^{-1}\left(\int_{\tau^0_{Ut_i}}|j'(u_\delta)+B^*p_\delta|\psi\right)^2\\ &\le Ch^{-n}_{\tau_U}k_i^{-1}\left(\int_{\tau_{Ut_i}\cap\Omega_\phi}|j'(u_\delta)+B^*p_\delta|\right)^2\le C\int_{\tau_{Ut_i}\cap\Omega_\phi}|j'(u_\delta)+B^*p_\delta|^2.\end{aligned}$$

Therefore,

$$\begin{aligned}&\int_{\Omega_\phi^+}(j'(u_\delta)+B^*p_\delta)(u_\delta-u)\\ &\le C\int_{\Omega_\phi^+}(j'(u_\delta)+B^*p_\delta)^2+C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}\\ &\le C\int_{\Omega_\phi}(j'(u_\delta)+B^*p_\delta)^2+C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))} \qquad (63)\\ &=C\hat\eta_1^2+C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}.\end{aligned}$$

It follows from the definition of Ω_ϕ^0 that $(j'(\phi^\delta)+B^*p_\delta)>0$ on Ω_ϕ^0. Then we have

$$\begin{aligned}\int_{\Omega_\phi^0}(j'(\phi_h)+B^*p_\delta)(\phi^\delta-u) &= \int_{\Omega_\phi^0}(j'(\phi^\delta)+B^*p_\delta)((\phi^\delta-\phi)+(\phi-u))\\ &\le \int_{\Omega_\phi^0}(j'(u_\delta)+B^*p_\delta)(\phi^\delta-\phi). \end{aligned}\tag{64}$$

Thus it follows from (61)–(64) that

$$\begin{aligned}I_2=\int_0^T(j'(u_\delta)+B^*p_\delta,u_\delta-u)_U &\le C\hat\eta_1^2+\int_{\Omega_\phi^0}(j'(u_\delta)+B^*p_\delta)(\phi^\delta-\phi)\\ &\quad +C\delta\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}.\end{aligned}\tag{65}$$

Then it follows from (60) and (65) that

$$\begin{aligned}&I_1+I_2=\\ &\int_0^T(j'(u)+B^*p,u-u_\delta)_U+\int_0^T(j'(u_\delta)+B^*p_\delta,u_\delta-u)_U\\ &\le C\hat\eta_1^2+\int_{\Omega_\phi^0}(j'(u)+B^*p-j'(u_\delta)-B^*p_\delta)(\phi-\phi^\delta)\\ &\quad +C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}\\ &\le C(\hat\eta_1^2+\|\phi-\phi^\delta\|^2_{0,\Omega_\phi^0})+C\sigma(\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}\\ &\quad +\|j'(u_\delta)-j'(u)\|^2_{L^2(0,T;L^2(\Omega_U))}+\|B^*(p_\delta-p^{u_\delta})\|^2_{L^2(0,T;L^2(\Omega_U))}\\ &\quad +\|B^*(p^{u_\delta}-p)\|^2_{L^2(0,T;L^2(\Omega_U))})\\ &\le C(\hat\eta_1^2+\hat\eta_8^2)+C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}+C\|p_\delta-p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))}.\end{aligned}\tag{66}$$

Here we used the inequalities

$$\|j'(u_\delta)-j'(u)\|^2_{L^2(0,T;L^2(\Omega_U))}\le C\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))},$$

$$\|B^*(p_\delta-p^{u_\delta})\|^2_{L^2(0,T;L^2(\Omega_U))}\le C\|p_\delta-p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega_U))},$$

and

$$\|B^*(p^{u_\delta}-p)\|^2_{L^2(0,T;L^2(\Omega_U))}\le C\|p^{u_\delta}-p\|^2_{L^2(0,T;L^2(\Omega))}\le C\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}.$$

Finally for I_3, it is easy to show that

$$\begin{aligned}I_3&=\int_0^T(B^*(p_\delta-p^{u_\delta}),u-u_\delta)_U\\ &\le C\|B^*(p_\delta-p^{u_\delta})\|^2_{L^2(0,T;L^2(\Omega))}+C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}\\ &\le C\|p_\delta-p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))}+C\sigma\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}.\end{aligned}\tag{67}$$

Thus, we obtain from (56), (66), and (67) that

$$\|u_\delta-u\|^2_{L^2(0,T;L^2(\Omega_U))}\le C(\hat\eta_1+\hat\eta_8+\|p_\delta-p^{u_\delta}\|^2_{L^2(0,T;L^2(\Omega))}).$$

The remainder of the proof is the same as for Lemma 3.5 and Theorem 3.1. ∎

REMARK 4.1. *By the same argument, we can obtain a similar estimate in the* $L^\infty(L^2)$ *norm considered in Theorem* 3.2. *It is worth noting that there may be different approaches to derive sharp a posteriori error bounds for the obstacle constraints. Noticeably, it may be possible to design some penalty schemes to solve the optimality system, and then apply the techniques used in* [8, 17, 22] *to derive sharp bounds.*

REMARK 4.2. *Here the key idea is to remove some inactive data in the coincidence set and to thus obtain sharper error estimates for the approximation of the inequality in* (2). *In fact, as seen in the above proof, only the part where* $j'(u_\delta) + B^* p_\delta \le 0$ *needs to be left in the estimator* $\hat{\eta}_1^2$. *Let us define*

$$\hat{\Omega}_\phi = \{(x,t) \in \Omega_U \times (0,T] : \ (B^* p_\delta)(x,t) \le -j'(u_\delta)\}.$$

In a sense, the set $\hat{\Omega}_\phi$ *is an approximation of the noncoincidence set. It follows that* $(j'(u_\delta) + B^* p_\delta)|_{\hat{\Omega}_\phi} \le 0$, *while* $j'(u) + B^* p \ge 0$. *Thus on* $\hat{\Omega}_\phi$, $j'(u_\delta) + B^* p_\delta$ *truly indicates the error. In fact, we have*

$$\int_{\hat{\Omega}_\phi} |j'(u_\delta) + B^* p_\delta|^2 \le \int_{\hat{\Omega}_\phi} |j'(u_\delta) + B^* p_\delta - (j'(u) + B^* p)|^2$$

$$\le C(\|u - u_\delta\|^2_{L^2(0,T;L^2(\Omega_U))} + \|p - p_\delta\|^2_{L^2(0,T;L^2(\Omega))}).$$

For ease of computation, we have used the set Ω_ϕ, *which is a little larger than* $\hat{\Omega}_\phi$. *However, we still have*

$$\hat{\eta}_1^2 \le C(\|u - u_\delta\|^2_{L^2(0,T;L^2(\Omega_U))} + \|p - p_\delta\|^2_{L^2(0,T;L^2(\Omega))} + \hat{\eta}_8^2).$$

On the coincidence set, $u = \phi$. *Therefore the error should be indicated by* $\hat{\eta}_8$, *and the term* $j'(u_\delta) + B^* p_\delta$ *should not appear there.*

REFERENCES

[1] M. AINSWORTH AND J. T. ODEN, *A posteriori error estimation in finite element analysis*, Comput. Methods Appl. Mech. Engrg., 142 (1997), pp. 1–88.

[2] P. ALOTTO, P. GIRDINIO, P. MOLFINO, AND M. NERVI, *Mesh adaption and optimization techniques in magnet design*, IEEE Trans. Magnetics, 32 (1996), pp. 2954–2957.

[3] W. ALT AND U. MACKENROTH, *Convergence of finite element approximation to state constrained convex parabolic boundary control problems*, SIAM J. Control Optim., 27 (1989), pp. 718–736.

[4] N. V. BANICHUK, F. J. BARTHOLD, A. FALK, AND E. STEIN, *Mesh refinement for shape optimization*, Structural Optimisation, 9 (1995), pp. 45–51.

[5] R. E. BANK AND A. WEISER, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[6] R. BECKER AND H. KAPP, *Optimization in PDE Models with Adaptive Finite Element Discretization*, Report 98-20 (SFB 359), University of Heidelberg, Heidelberg, Germany, 1998.

[7] R. BECKER, H. KAPP, AND R. RANNACHER, *Adaptive finite element methods for optimal control of partial differential equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.

[8] M. BOMAN, *A Posteriori Error Analysis in the Maximum Norm for a Penalty Finite Element Method for the Time Dependent Obstacle Problem*, Preprint, Department of Mathematics, Chalmers University of Technology and Goteburg University, 2000.

[9] C. CARSTENSEN AND S. A. FUNKEN, *Constants in Clement-interpolation error and residual based a posteriori estimates in finite element methods*, East-West J. Numer. Math., 8 (2000), pp. 153–175.

[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[11] B. Cockburn and C. W. Shu, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.
[12] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* I: *A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.
[13] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* II: *Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.
[14] K. Eriksson, C. Johnson, and V. Thomee, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Model. Math. Anal. Numer., 19 (1985), pp. 611–643.
[15] F. S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.
[16] D. A. French and J. T. King, *Approximation of an elliptic control problem by the finite element method*, Numer. Funct. Anal. Appl., 12 (1991), pp. 299–315.
[17] D. French, S. Larsson, and R. Nochetto, *A posteriori error estimates for a finite element approximation of the obstacle problem in L^∞*, Comput. Methods Appl. Math., 1 (2001), pp. 18–38.
[18] T. Geveci, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numer., 13 (1979), pp. 313–328.
[19] M. D. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for stationary Navier-Stokes equations with Dirichlet controls*, RAIRO Model. Math. Anal. Numer., 25 (1991), pp. 711–748.
[20] L. Hou and J. C. Turner, *Analysis and finite element approximation of an optimal control problem in electrochemistry with current density controls*, Numer. Math., 71 (1995), pp. 289–315.
[21] P. Houston and E. Süli, *A Posteriori Error Analysis for Linear Convection-Diffusion Problems under Weak Mesh Regularity Assumptions*, Report 97/03, Oxford University Computing Laboratory, Oxford, UK, 1997.
[22] C. Johnson, *Adaptive finite element methods for the obstacle problem*, Math. Models Methods Appl. Sci., 2 (1992), pp. 483–487.
[23] C. Johnson, R. Rannacher, and M. Boman, *Numerics and hydrodynamic stability: Toward error control in computational fluid dynamics*, SIAM J. Numer. Anal., 32 (1995), pp. 1058–1079.
[24] G. Knowles, *Finite element approximation of parabolic time optimal control problems*, SIAM J. Control Optim., 20 (1982), pp. 414–427.
[25] A. Kufner, O. John, and S. Fucik, *Function Spaces*, Nordhoff, Leyden, The Netherlands, 1977.
[26] I. Lasiecka, *Ritz-Galerkin approximation of the time optimal boundary control problem for parabolic systems with Dirichlet boundary conditions*, SIAM J. Control Optim., 22 (1984), pp. 477–499.
[27] R. Li, W. B. Liu, H. P. Ma, and T. Tang, *Adaptive finite element approximation for distributed elliptic optimal control problems*, SIAM J. Control Optim., 41 (2002), pp. 1321–1349.
[28] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.
[29] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972.
[30] W. B. Liu, *Recent advances in mesh adaptivity for optimal control problems*, in Fast Solution of Discretized Optimization Problems, Internat. Ser. Numer. Math. 138, Birkhäuser, Basel, 2001, pp. 154–166.
[31] W. B. Liu and J. E. Rubio, *Optimality conditions for strongly monotone variational inequalities*, Appl. Math. Optim., 27 (1993), pp. 291–312.
[32] W. B. Liu and D. Tiba, *Error estimates approximation of optimization problems governed by nonlinear operators*, Numer. Funct. Anal. Optim., 22 (2001), pp. 953–972.
[33] W. B. Liu and N. Yan, *Quasi-norm local error estimators for p-Laplacian*, SIAM J. Numer. Anal., 39 (2001), pp. 100–127.
[34] W. B. Liu and N. Yan, *A posteriori error analysis for convex boundary control problems*, SIAM J. Numer. Anal., 39 (2001), pp. 73–99.
[35] W. B. Liu and N. Yan, *A posteriori error analysis for convex distributed optimal control problems*, Adv. Comput. Math., 15 (2001), pp. 285–309.

[36] W. B. Liu and N. Yan, *A posteriori error estimates for optimal control problems governed by parabolic equations*, Numer. Math., 93 (2003), pp. 497–521.
[37] K. Malanowski, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems*, Appl. Math. Optim., 8 (1981), pp. 69–95.
[38] K. Maute, S. Schwarz, and E. Ramm, *Adaptive topology optimization of elastoplastic structures*, Structural Optimization, 15 (1998), pp. 81–91.
[39] R. S. McKnight and W. E. Bosarge, Jr., *The Ritz–Galerkin procedure for parabolic control problems*, SIAM J. Control Optim., 11 (1973), pp. 510–524.
[40] P. Neittaanmaki and D. Tiba, *Optimal Control of Nonlinear Parabolic Systems: Theory, Algorithms and Applications*, Marcel Dekker, New York, 1994.
[41] O. Pironneau, *Optimal Shape Design for Elliptic System*, Springer-Verlag, Berlin, 1984.
[42] A. Schleupen, K. Maute, and E. Ramm, *Adaptive FE-procedures in shape optimization*, Structural and Multidisciplinary Optimization, 19 (2000), pp. 282–302.
[43] D. Tiba, *Lectures on the Optimal Control of Elliptic Equations*, University of Jyvaskyla Press, Jyvaskyla, Finland, 1995.
[44] D. Tiba and F. Troltzsch, *Error estimates for the discretization of state constrained convex control problems*, Numerical Funct. Anal. Optim., 17 (1996), pp. 1005–1028.
[45] F. Troltzsch, *Semidiscrete Ritz-Galerkin approximation of nonlinear parabolic boundary control problems—Strong convergence of optimal control*, Appl. Math. Optim., 29 (1994), pp. 309–329.
[46] R. Verfurth, *A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement*, Wiley-Teubner, Stuttgart, 1996.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1062–1072

A NEW DISCONTINUOUS FINITE VOLUME METHOD FOR ELLIPTIC PROBLEMS*

XIU YE†

Abstract. We develop and analyze a new discontinuous finite volume method for second order elliptic problems with tensor coefficients. An optimal order error estimate is obtained in a mesh dependent norm. An L^2-error estimate is also obtained.

Key words. finite element methods, discontinuous Galerkin method, finite volume methods, elliptic problems

AMS subject classifications. 65N15, 65N30, 76D07, 35B45, 35J50

DOI. 10.1137/S0036142902417042

1. Introduction.

Like the finite element method and the finite difference method, the finite volume method is a discretization technique for solving partial differential equations. The integral formulation of a finite volume scheme for a partial differential equation (PDE) is obtained by integrating the PDE over a control volume, and it represents in general the conservation of a quantity of interest, such as mass, momentum, or energy. Due to this natural association, finite volume methods are widely used in practical problems, such as fluid mechanics computations [18, 19, 20]. Recently, Chou, Kwak, and Vassilevski [10, 12, 13, 11] applied finite volume element methods involving nonconforming trial functions to diffusion, diffusion-reaction, and Stokes problems and obtained optimal order error estimates in a discrete H^1 norm.

Unlike the standard conforming and nonconforming finite element methods, the discontinuous Galerkin method does not require continuity of the approximation functions across the interelement boundary but instead enforces the connection between elements by adding a penalty term. These methods came from the idea of enforcing Dirichlet boundary conditions through penalties (see Lions [17], Aubin [3], Babuska [4], and Nitsche [21]). Because of the use of discontinuous functions, discontinuous Galerkin methods have the advantages of a high order of accuracy, high parallelizability, localizability, and easy handling of complicated geometries. Due to these advantages, the study of discontinuous Galerkin methods has been an active research area since its introduction by Reed and Hill [22]. The discontinuous Galerkin methods have been used to solve hyperbolic and elliptic equations by many researchers. For example, see [5, 6, 7, 8, 9, 14, 15, 16, 23]. Recently, Arnold et al. [2] provided a framework for the analysis of a large class of discontinuous Galerkin methods for second order elliptic problems. Most literature concerning discontinuous Galerkin methods for finite element approximations can be found in the references given in [2].

However, very little has been done using discontinuous functions for the finite volume approximation. In this paper, we developed a new discontinuous finite volume method (DFV) in which discontinuous piecewise polynomials are used for the trial functions. It is natural to assume that the advantages of using discontinuous functions in finite element methods should apply to finite volume methods. The nature of the

*Received by the editors October 29 2002; accepted for publication (in revised form) October 14, 2003; published electronically July 29, 2004.
http://www.siam.org/journals/sinum/42-3/41704.html
†Department of Mathematics and Statistics, University of Arkansas at Little Rock, 2801 South University, Little Rock, AR 72204 (xxye@ualr.edu).

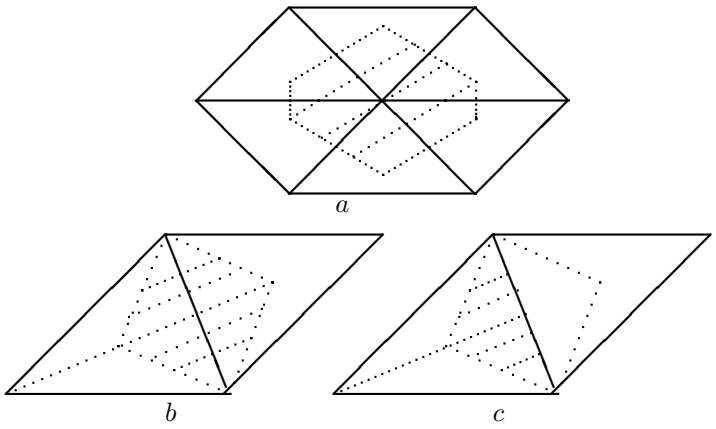


FIG. 1. *Dual partitions.*

discontinuity of the trial function is such that the elements in the corresponding dual partition have the smallest support, as compared with the cases when conforming (classical finite volume methods) and nonconforming elements (more recent finite volume methods [10]) are used for the trial functions. This discontinuous finite volume scheme will be analyzed for the elliptic problems and an optimal order error estimate will be obtained in a discrete H^1 norm. However, our method and analysis can be applied to solve more complex problems, like the Stokes equations, the system of linear elasticity, and plate problems.

In Figure 1, we compare dual partitions associated with conforming elements, nonconforming elements, and our new discontinuous elements. The shaded region in Figure 1(a) represents a typical element in the dual partition when conforming elements are used for trial functions. This region involves six elements in the primary partition. This case is the classical finite volume scheme, where the continuous piecewise linear functions are used for the trial functions. The shaded quadrilateral in Figure 1(b) is an element in the dual partition associated with nonconforming elements [10] which involves two elements of the primary partition. Figure 1(c) shows that the elements of the dual partition associated with our new discontinuous finite volume method have the best local ability when only one element in the primary partition is involved. The localizability of the discontinuous element and its dual partition in our method should provide an advantage for parallel computing.

Since the discontinuous functions are used in the approximation, the number of unknowns is larger. However, the small support of the control volume for this method makes the method more suitable to the domain decomposition such that the information can be updated triangle by triangle in the primary partition.

We consider the model problems

$$-\nabla \cdot B\nabla u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \tag{1.1}$$

where Ω is a bounded convex polygon in R^2. $B = (b_{ij})_{2\times 2} \in W^{1,\infty}(\Omega)^4$ is a symmetric matrix value function and satisfies the following condition: there exists a constant $\beta > 0$ such that

$$\beta\xi^T\xi \le \xi^T B\xi \quad \forall \xi \in R^2. \tag{1.2}$$

We will use the standard definitions for the Sobolev spaces $H^s(K)$ and their

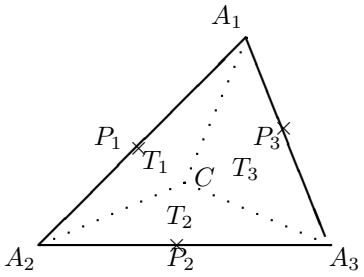


FIG. 2. *A triangular partition and its dual.*

associated inner products $(\cdot,\cdot)_{s,K}$, norms $\|\cdot\|_{s,K}$, and seminorms $|\cdot|_{s,K}$, $s \geq 0$. The space $H^0(K)$ coincides with $L^2(K)$, in which case the norm and inner product are denoted by $\|\cdot\|_K$ and $(\cdot,\cdot)_K$, respectively. If $K = \Omega$, we drop K.

This paper is organized as follows. In section 2, we derive a discontinuous finite volume formulation for the elliptic problems. In section 3, an optimal order error estimate is obtained in a mesh dependent norm and a first order L^2-error estimate is derived.

2. Discontinuous finite volume formulation. The simplicity of the finite volume method is due to its use of piecewise constant functions as test functions. To keep the same dimension for the spaces of the trial functions and test functions, two different triangulations are needed. Let $\mathcal{R}_h$ be a triangulation of Ω with $\text{diam}(K) \leq h,\ K \in \mathcal{R}_h$. Assume that the triangulation $\mathcal{R}_h$ is quasi-uniform. We define the dual partition $\mathcal{T}_h$ of $\mathcal{R}_h$ for the test function space as follows. We divide each $K \in \mathcal{R}_h$ into three triangles by connecting the barycenter and the three corners of the triangle as shown in Figure 2. Let $\mathcal{T}_h$ consist of all these triangles T_j.

We define the finite dimensional space associated with $\mathcal{R}_h$ for the trial functions as

$$V_h = \{v \in L^2(\Omega) :\ v|_K \in P_1(K)\ \ \forall K \in \mathcal{R}_h\} \tag{2.1}$$

and define the finite dimensional space P_h for test functions associated with the dual partition $\mathcal{T}_h$ as

$$P_h = \{q \in L^2(\Omega) :\ q|_T \in P_0(T)\ \ \forall T \in \mathcal{T}_h\},$$

where $P_l(T)$ consists of all the polynomials with degree less than or equal to l defined on T.

Let $V(h) = V_h + H^2(\Omega) \cap H_0^1(\Omega)$. Define a mapping $\gamma :\ V(h) \to P_h$ as

$$\gamma v|_T = \frac{1}{h_e}\int_e v|_T ds, \quad T \in \mathcal{T}_h,$$

as shown in Figure 3. The above idea of connecting the trial function and test function spaces in the Petrov–Galerkin method through an operator was introduced in [10] in the context of elliptic problems.

Multiplying (1.1) by $q \in P_h$, we have

$$-\sum_{T\in\mathcal{T}_h}\int_{\partial T} B\nabla u \cdot \mathbf{n} q ds = (f,\ q), \tag{2.2}$$

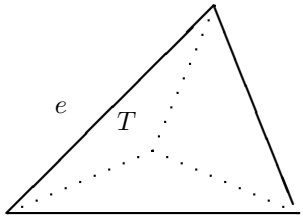


FIG. 3. *Element $T \in \mathcal{T}_h$ and one of its edges e.*

where $\mathbf{n}$ is the unit outward normal vector on ∂T. Let $T_j \in \mathcal{T}_h$ $(j = 1, 2, 3)$ be three triangles in $K \in \mathcal{R}_h$. Then we have

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T} B\nabla u\cdot\mathbf{n}q ds = \sum_{K\in\mathcal{R}_h}\sum_{j=1}^{3}\int_{\partial T_j} B\nabla u\cdot\mathbf{n}q ds \tag{2.3}$$

$$= \sum_{K\in\mathcal{R}_h}\sum_{j=1}^{3}\int_{A_{j+1}CA_j} B\nabla u\cdot\mathbf{n}q ds + \sum_{K\in\mathcal{R}_h}\int_{\partial K} B\nabla u\cdot\mathbf{n}q ds,$$

where $A_4 = A_1$.

Let e be an interior edge shared by two elements K_1 and K_2 in $\mathcal{R}_h$, and let $\mathbf{n}_1$ and $\mathbf{n}_2$ be unit normal vectors on e pointing exterior to K_1 and K_2, respectively. We define the average $\{\cdot\}$ and jump $[\cdot]$ on e for scalar q and vector $\mathbf{w}$, respectively, as (see [2])

$$\{q\} = \frac{1}{2}(q|_{\partial T_1} + q|_{\partial T_2}), \quad [q] = q|_{\partial T_1}\mathbf{n}_1 + q|_{\partial T_2}\mathbf{n}_2,$$

$$\{\mathbf{w}\} = \frac{1}{2}(\mathbf{w}|_{\partial T_1} + \mathbf{w}|_{\partial T_2}), \quad [\mathbf{w}] = \mathbf{w}|_{\partial T_1}\cdot\mathbf{n}_1 + \mathbf{w}|_{\partial T_2}\cdot\mathbf{n}_2.$$

If e is a edge on the boundary of Ω, we define

$$\{q\} = q, \quad [\mathbf{w}] = \mathbf{w}\cdot\mathbf{n}.$$

Let Γ denote the union of the boundaries of the triangle K of $\mathcal{R}_h$ and $\Gamma_0 := \Gamma\backslash\partial\Omega$. A straightforward computation gives

$$\sum_{K\in\mathcal{R}_h}\int_{\partial K} q\mathbf{v}\cdot\mathbf{n} ds = \sum_{e\in\Gamma}\int_e [q]\cdot\{\mathbf{v}\} ds + \sum_{e\in\Gamma_0}\int_e \{q\}[\mathbf{v}] ds. \tag{2.4}$$

Using (2.4) and the fact that $[B\nabla u] = 0$, (2.3) becomes

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T} B\nabla u\cdot\mathbf{n}q ds = \sum_{K\in\mathcal{R}_h}\sum_{j=1}^{3}\int_{A_{j+1}CA_j} B\nabla u\cdot\mathbf{n}q ds + \sum_{e\in\Gamma}\int_e [q]\cdot\{B\nabla u\} ds.$$

Our discontinuous finite volume approximation problem for (1.1) is to find $u_h \in V_h$ such that

$$a(u_h, q) = (f,\ q) \qquad \forall q \in P_h \tag{2.5}$$

with

$$a(v,q) = -\sum_{K\in\mathcal{R}_h}\sum_{j=1}^{3}\int_{A_{j+1}CA_j} B\nabla u\cdot\mathbf{n}q ds - \sum_{e\in\Gamma}\int_e [q]\cdot\{B\nabla v\}ds + \alpha\sum_{e\in\Gamma}[\gamma v]_e\cdot[q]_e,$$

where α is a real number that will be determined later. In the definition of $a(\cdot,\cdot)$, γv and q are piecewise constant functions and so are their jump functions $[\gamma v]_e$ and $[q]_e$. Thus the penalty term in the definition of $a(\cdot,\cdot)$ is the same as the penalty term in many discontinuous Galerkin formulation for the finite element methods [2].

Define the following bilinear form:

$$A(v,\ w) = a(v,\ \gamma w) \quad \forall v, w \in V(h).$$

Then the approximation problem (2.5) becomes the following: find $u_h \in V_h$ such that

$$A(u_h, v) = (f,\ \gamma v) \quad \forall v \in V_h. \tag{2.6}$$

Since $[\gamma u]_e = 0$, it is easy to see that u, the solution of (1.1), satisfies

$$A(u, v) = (f,\ \gamma v) \quad \forall v \in V_h. \tag{2.7}$$

Let

$$A_1(v,w) = -\sum_{K\in\mathcal{R}_h}\sum_{j=1}^{3}\int_{A_{j+1}CA_j} B\nabla u\cdot\mathbf{n}\gamma w ds.$$

Thus

$$A(v,w) = A_1(v,w) - \sum_{e\in\Gamma}\int_e [\gamma w]\cdot\{B\nabla v\}ds + \alpha\sum_{e\in\Gamma}[\gamma v]_e\cdot[\gamma w]_e.$$

Let $\nabla_h v$ be the function whose restriction to each element $K \in \mathcal{R}_h$ is equal to ∇v. Let

$$\bar{B}|_K = \frac{1}{\text{meas}(K)}\int_K B(\mathbf{x})d\mathbf{x} \quad \forall K \in \mathcal{R}_h.$$

We define a norm $|||\cdot|||$ for $V(h)$ as follows:

$$|||v|||^2 = |v|_{1,h}^2 + \sum_e [\gamma v]_e^2, \tag{2.8}$$

where $|v|_{1,h}^2 = \sum_K |v|_{1,K}^2$. It is easy to see that $|||\cdot|||$ defines a norm. (We use C with or without subscripts in this note to denote a generic positive constant, possibly different at different occurrences, that is independent of the mesh size h but may depend on the domain Ω.)

The following trace inequality can be found in [1]. For $w \in H^2(K)$ and for an edge e of K,

$$\|w\|_e^2 \le C(h_e^{-1}|w|_K^2 + h_e|w|_{1,K}^2), \tag{2.9}$$

where C depends only on the minimum angle of K.

LEMMA 2.1. *For any $v, w \in V(h)$, we have*

$$A_1(v,w) = (B\nabla_h v, \nabla_h w) + \sum_K \int_{\partial K} (\gamma w - w) B\nabla v \cdot \mathbf{n} ds + \sum_K (\nabla \cdot B\nabla v, w - \gamma w)_K. \tag{2.10}$$

Furthermore, if $v, w \in V_h$, then

$$A_1(v,w) \geq (B\nabla_h v, \nabla_h w) - C_1 h |\!|\!| v |\!|\!| |\!|\!| w |\!|\!|. \tag{2.11}$$

Proof. Using the divergence theorem on each triangle T_j for $v \in V_h$, we have

$$\begin{aligned}
A_1(v,w) &= -\sum_K \sum_{j=1}^3 \int_{A_{j+1}CA_j} B\nabla v \cdot \mathbf{n} \gamma w ds = -\sum_K \sum_{j=1}^3 \gamma w \int_{A_{j+1}CA_j} B\nabla v \cdot \mathbf{n} ds \\
&= \sum_K \sum_{j=1}^3 \gamma w \int_{A_j A_{j+1}} B\nabla v \cdot \mathbf{n} ds - \sum_K \sum_{T_j} (\nabla \cdot B\nabla v, \gamma w)_{T_j} \\
&= \sum_K \sum_{j=1}^3 \int_{A_j A_{j+1}} (\gamma w - w) B\nabla v \cdot \mathbf{n} ds + \sum_K \int_{\partial K} w B\nabla v \cdot \mathbf{n} ds \\
&- \sum_K \sum_{T_j} (\nabla \cdot B\nabla v, \gamma w)_{T_j} \\
&= \sum_K \int_{\partial K} (\gamma w - w) B\nabla v \cdot \mathbf{n} ds + \sum_K (B\nabla v, \nabla w)_K + \sum_K (\nabla \cdot B\nabla v, w)_K \\
&- \sum_K \sum_{T_j} (\nabla \cdot B\nabla v, \gamma w)_{T_j} \\
&= \sum_K \int_{\partial K} (\gamma w - w) B\nabla v \cdot \mathbf{n} ds + (B\nabla_h v, \nabla_h w) + \sum_K (\nabla \cdot B\nabla v, w - \gamma w)_K.
\end{aligned}$$

If $v, w \in V_h$, the inverse inequality and (2.9) imply

$$\begin{aligned}
&\sum_K \int_{\partial K} (\gamma w - w) B\nabla v \cdot \mathbf{n} ds = \sum_K \int_{\partial K} (\gamma w - w)(B - \bar{B}) \nabla v \cdot \mathbf{n} ds \\
&\leq Ch \|B\|_{1,\infty} \sum_K (h^{-1} \|w - \gamma w\|_K^2 + h |w - \gamma w|_{1,K}^2)^{\frac{1}{2}} \cdot (h^{-1} |v|_{1,K}^2 + h |v|_{2,K}^2)^{\frac{1}{2}} \\
&\leq Ch |\!|\!| v |\!|\!| |\!|\!| w |\!|\!|, \qquad (2.12)
\end{aligned}$$

where $|w - \gamma w|_{1,K}^2 = \sum_{j=1}^3 |w - \gamma w|_{1,T_j}^2$. Let $\nabla(\nabla v) = (\nabla v_{x_1}, \nabla v_{x_2})$, a two by two matrix. Then

$$\sum_K (\nabla \cdot B\nabla v, w - \gamma w)_K = \sum_K (\nabla \cdot B \cdot \nabla v + B : \nabla(\nabla v), w - \gamma w)_K.$$

Since $\nabla(\nabla v) = 0$ for $v \in V_h$,

$$\sum_K (\nabla \cdot B\nabla v, w - \gamma w)_K \leq Ch \|B\|_{1,\infty} |\!|\!| v |\!|\!| |\!|\!| w |\!|\!|. \tag{2.13}$$

Equations (2.12) and (2.13) imply (2.11).

LEMMA 2.2. *There is a constant C independent of h such that for α large enough and h small enough*

$$A(v,v) \geq C|||v|||^2 \qquad \forall v \in V_h. \tag{2.14}$$

Proof. The trace inequality (2.9) and the inverse inequality give that for $v \in V_h$

$$\begin{aligned}
&\sum_e \int_e [\gamma v] \cdot \{B\nabla_h v\} ds \\
&\leq C\|B\|_{1,\infty} \left(\sum_K (|v|_{1,K}^2 + h^2 |v|_{2,K}^2) \right)^{\frac{1}{2}} \left(\sum_e h_e^{-1} \int_e [\gamma v]^2 ds \right)^{\frac{1}{2}} \\
&\leq C|||v||| \left(\sum_e [\gamma v]_e^2 \right)^{\frac{1}{2}}.
\end{aligned}$$

Using the inequality above and Lemma 2.1, we have

$$\begin{aligned}
A(v,v) &\geq (B\nabla_h v, \nabla_h v) + \alpha \sum_e [\gamma v]_e^2 - \sum_e \int_e [\gamma v] \cdot \{B\nabla v\} ds - C_1 h |||v|||^2 \\
&\geq |v|_{1,h}^2 + \alpha \sum_e [\gamma v]_e^2 - C|||v||| \left(\sum_e [\gamma v]_e^2 \right)^{\frac{1}{2}} - C_1 h |||v|||^2 \\
&\geq C|||v|||^2 - C_1 h |||v||| \geq C|||v|||^2.
\end{aligned}$$

The last inequality is obtained by using the generalized arithmetic-geometric mean inequality and choosing α large enough and h small enough .

In the following, we will assume that α is large enough and h small enough so that (2.14) holds.

LEMMA 2.3. *For $v, w \in V(h)$, we have*

$$A(v,w) \leq C \left(|||v||| + \left(\sum_K h^2 |v|_{2,K}^2 \right)^{\frac{1}{2}} \right) |||w|||. \tag{2.15}$$

If $v, w \in V_h$, then

$$A(v,w) \leq C|||v|||||w|||. \tag{2.16}$$

Proof. By Lemma 2.1, the trace inequality (2.9), and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
|A_1(v,w)| &\leq |(B\nabla_h v, \nabla_h w)| \\
&\quad + \left| \sum_K \int_{\partial K} (w - \gamma w) B\nabla v \cdot \mathbf{n} ds \right| + \left| \sum_K (\nabla \cdot B\nabla v, w - \gamma w)_K \right| \\
&\leq C\|B\|_{1,\infty} (|v|_{1,h}|w|_{1,h} + \sum_K (h^{-1}\|w - \gamma w\|_K^2 + h|w - \gamma w|_{1,K}^2)^{\frac{1}{2}} \cdot (h^{-1}|v|_{1,K}^2 + h|v|_{2,K}^2)^{\frac{1}{2}} \\
&\quad + \sum_K h(|v|_{1,K} + |v|_{2,K})|w|_{1,K}) \\
&\leq C \left(|v|_{1,h}|w|_{1,h} + \left(\sum_K h^2 |v|_{2,K}^2 \right)^{\frac{1}{2}} |w|_{1,h} \right).
\end{aligned}$$

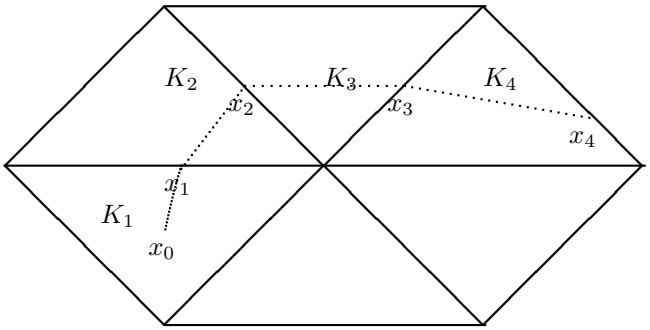


FIG. 4. *A path.*

The definition of $A(v,w)$ and the inequality above imply that

$$\begin{aligned} A(v,w) &\le C(|v|_{1,h}|w|_{1,h} + \left(\sum_K h^2|v|^2_{2,K}\right)^{\frac{1}{2}} |w|_{1,h} \\ &+ \left(\sum_K (|v|^2_{1,K} + h^2|v|^2_{2,K})\right)^{\frac{1}{2}} \left(\sum_e [\gamma w]^2_e\right)^{\frac{1}{2}} \\ &+ \alpha \left(\sum_e [\gamma v]^2_e\right)^{\frac{1}{2}} \left(\sum_e [\gamma w]^2_e)^{\frac{1}{2}}\right) \\ &\le C(|||v||| + \left(\sum_K h^2|v|^2_{2,K})^{\frac{1}{2}}\right) |||w|||. \end{aligned}$$

Equation (2.15) and the inverse inequality imply (2.16).

3. Error estimates. We will derive an optimal order error estimate in the norm $|||\cdot|||$ defined in (2.8) and a first order error estimate in the L^2-norm. We start with the following lemma.

LEMMA 3.1. *There exists a constant C independent of h such that*

$$\|w\| \le C|||w||| \qquad \forall w \in V_h.$$

Proof. The proof is similar to the proof of Lemma 2.1 in [12]. Let $|e|$ denote the length of edge e. Since $[w]$ is continuous on $e_i \in \Gamma$, there exists $\mathbf{x}_i \in e_i$ such that

$$\int_{e_i} [w]ds = [w](\mathbf{x}_i)|e_i|, \tag{3.1}$$

where $[w](\mathbf{x}_i) = w|_{K_1}(\mathbf{x}_i) - w|_{K_2}(\mathbf{x}_i)$ and K_1 and K_2 are the two triangles that share e_i. If e_i is an edge on the boundary, then $\mathbf{x}_i \in e_i$ is a point such that $\int_{e_i} wds = w(\mathbf{x}_i)|e_i|$. For any $\mathbf{x} = \mathbf{x}_0 \in K \in \mathcal{R}_h$, we can find a path from $\mathbf{x}_0$ to $\mathbf{x}_l$, a point on the boundary, by joining a sequence of $\mathbf{x}_i$ as shown in Figure 4, where $\mathbf{x}_i$ $(i = 1, \ldots, l)$ satisfy (3.1). Let C_0 be a constant such that $lh \le C_0$. Let $\{K_i\}_{i=1}^l$ be the sequence of triangles in $\mathcal{R}_h$ containing $\mathbf{x}_i$ $(i = 0, \ldots, l)$ as shown in Figure 4.

Define $w(\mathbf{x}_0^1) = w(\mathbf{x}_0)$, $w(\mathbf{x}_1^2) = w|_{K_1}(\mathbf{x}_1)$, and $w(\mathbf{x}_1^1) = w|_{K_2}(\mathbf{x}_1)$. In general, $w(\mathbf{x}_i^2) = w|_{K_i}(\mathbf{x}_i)$ and $w(\mathbf{x}_i^1) = w|_{K_{i+1}}(\mathbf{x}_i)$. The mean value theorem, the Cauchy–Schwarz inequality, and (3.1) give

$$|w(\mathbf{x})|^2 = |w(\mathbf{x}_0)|^2 = \left| \sum_{i=1}^{l} (w(\mathbf{x}_{i-1}^1) - w(\mathbf{x}_i^2)) + \sum_{i=1}^{l} [w](\mathbf{x}_i) \right|^2 \tag{3.2}$$

$$\leq Cl \left(\sum_{i=1}^{l} (\nabla w(\bar{\mathbf{x}}_i)^2 (\mathbf{x}_{i-1} - \mathbf{x}_i)^2 + \sum_{i=1}^{l} [\gamma w]_{e_i}^2 \right), \tag{3.3}$$

where $\bar{\mathbf{x}}_i \in K_i$ is a point between $\mathbf{x}_{i-1}$ and $\mathbf{x}_i$. As in [12], we have

$$|\nabla w(\bar{\mathbf{x}}_i)|^2 h^2 \leq C |\nabla w|_{K_i}^2. \tag{3.4}$$

Equation (3.4) implies

$$|w(\mathbf{x})|^2 \leq Cl \left(\sum_{i=1}^{l} |\nabla w|_{K_i}^2 + \sum_{i=1}^{l} [\gamma w]_{e_i}^2 \right). \tag{3.5}$$

Integrating (3.5) over K gives

$$\int_K |w(\mathbf{x})|^2 \leq Clh^2 \left(\sum_{i=1}^{l} |\nabla w|_{K_i}^2 + \sum_{i=1}^{l} [\gamma w]_{e_i}^2 \right). \tag{3.6}$$

For each $\mathbf{x} \in \Omega$, choose the path such that, when adding all the $K \in \mathcal{R}_h$, the same K_i appears at most l times. Then using the fact that $lh \leq C_0$, we have

$$\|w\|^2 = \int_\Omega |w(\mathbf{x})|^2 \leq C \left(|w|_{1,h}^2 + \sum_e [\gamma w]_e^2 \right) \leq C |\!|\!| w |\!|\!|^2. \tag{3.7}$$

This completes the proof. □

Let $u_I \in V_h$ be the interpolation of u. It is well known that

$$|u - u_I|_{s,K} \leq Ch^{2-s} |u|_{2,K} \quad \forall K \in \mathcal{R}_h, \ s = 0, 1, \tag{3.8}$$

where C depends only on the angle of K. The Cauchy–Schwarz inequality implies

$$[\gamma v]_e^2 = \left(\frac{1}{h_e} \int_e [v] ds \right)^2 \leq \left(\frac{1}{h_e} \right)^2 \int_e [v]^2 ds \int_e ds$$

$$= \int_e \frac{1}{h_e} [v]^2 ds. \tag{3.9}$$

The definitions of the norm $|\!|\!| \cdot |\!|\!|$, (3.9), (2.9), and (3.8) give

$$|\!|\!| u - u_I |\!|\!|^2 = |u - u_I|_{1,h}^2 + \sum_e [\gamma u - \gamma u_I]_e^2 \tag{3.10}$$

$$\leq C \left(|u - u_I|_{1,h}^2 + \sum_e \int_e h_e^{-1} [u - u_I]^2 ds \right)$$

$$\leq C \left(|u - u_I|_{1,h}^2 + \sum_K h^{-2} \|u - u_I\|_K^2 \right)$$

$$\leq Ch^2 |u|_2^2$$

and

$$\left(\sum_K h^2|u-u_I|_{2,K}^2\right)^{\frac{1}{2}} \le Ch|u|_2. \tag{3.11}$$

THEOREM 3.2. *Let $u_h \in V_h$ and $u \in H^2(\Omega)\cap H_0^1(\Omega)$ be the solutions of* (2.5) *and* (1.1)*, respectively; then there exists a constant C independent of h such that*

$$|||u-u_h||| \le Ch|u|_2 \tag{3.12}$$

and

$$\|u-u_h\| \le Ch|u|_2. \tag{3.13}$$

Proof. Subtracting (2.6) from (2.7) gives

$$A(u-u_h,v)=0 \quad \forall v \in V_h. \tag{3.14}$$

Using (2.14), (3.14), and (2.15), we have

$$\begin{aligned} |||u_h-u_I|||^2 &\le CA(u_h-u_I,u_h-u_I) \\ &= CA(u-u_I,u_h-u_I) \\ &\le C(|||u-u_I||| + \left(\sum_K h^2|u-u_I|_{2,K}^2\right)^{\frac{1}{2}})\,|||u_h-u_I|||. \end{aligned} \tag{3.15}$$

Using (3.15), (3.10), and (3.11), we have

$$|||u_h-u_I||| \le Ch|u|_2. \tag{3.16}$$

The triangle inequality and (3.10) imply (3.12). Using Lemma 3.1 and (3.16), we have

$$\|u_h-u_I\| \le C|||u_h-u_I||| \le Ch|u|_2.$$

Equation (3.8) and the triangle inequality imply (3.13). We have completed the proof. ∎

REFERENCES

[1] D. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] J. AUBIN, *Approximation des problemes aux limites non homogenes pour des operateurs non lineaires*, J. Math. Anal. Appl., 30 (1970), pp. 510–521.

[4] I. BABUSKA, *The finite element method with penalty*, Math. Comp., 27 (1973), pp. 221–228.

[5] I. BABŬSKA AND M. ZLÁMAL, *Nonconforming elements in the finite element method with penalty*, SIAM J. Numer. Anal., 10 (1973), pp. 863–875.

[6] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[7] C. E. BAUMANN AND J. T. ODEN, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[8] Z. CHEN, B. COCKBURN, C. GARDNER, AND J. JEROME, *Quantum hydrodynamic simulation of hysteresis in the resonant tunneling diode*, J. Comput. Phys., 117 (1995), pp. 274–280.

[9] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV: *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

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

[11] S. H. Chou and P. S. Vassilevski, *A general mixed co-volume framework for constructing conservative schemes for elliptic problems*, Math. Comp., 68 (1999), pp. 991–1011.

[12] S. H. Chou and D. Y. Kwak, *A covolume method based on rotated bilinears for the generalized stokes problem*, SIAM J. Numer. Anal., 35 (1998), pp. 494–507.

[13] S. H. Chou and D. Y. Kwak, *Analysis and convergence of a MAC scheme for the generalized Stokes problem*, Numer. Methods Partial Differential Equations, 13 (1997), pp. 147–162.

[14] B. Cockburn, G. E. Karniadakis, and C.W. Shu, eds., *The Discontinuous Galerkin Methods: Theory, Computation and Applications*, Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin, 2000.

[15] B. Cockburn and C. W. Shu, *The local discontinuous Galerkin finite element method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[16] J. Douglas, Jr. and T. Dupont, *Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods*, Lecture Notes Phy. 58, Springer-Verlag, Berlin, 1976.

[17] J.-L. Lions, *Problemes aus limites non homogenes a donees irregulieres: Une methode d'approximation*, in Numerical Analysis of Partial Differential Equations (C.I.M.E. 2 Ciclo, Ispra, 1967), Edizioni Cremonese, Rome, 1968, pp. 283–292.

[18] R. Lazarov, I. Michev, and P. Vassilevski, *Finite volume methods for convection-diffusion problems*, SIAM J. Numer. Anal., 33 (1996), pp. 31–55.

[19] C. Liu and S. McCormick, *The finite volume element method (FVE) for planar cavity flow*, in Proceedings of the 11th International Conference on CFD, Williamsburg, VA, 1988.

[20] R. A. Nicolaides, T. A. Porsching, and C. A. Hall, *Covolume methods in computational fluid dynamics*, in Computational Fluid Dynamics Review, M. Hafez and K. Oshima, eds., Wiley, New York, 1995, pp. 279–299.

[21] J. A. Nitsche, *Uber ein Variationsprinzip zur Losung Dirichlet-Problemen bei Verwendung von Teilraumen, die keinen Randbedingungen unteworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.

[22] W. H. Reed and T. R. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[23] B. Riviere, M. F. Wheeler, and V. Girault, *Improved Energy Estimates for Interior Penalty, Constrained and Discontinuous Galerkin Methods for Elliptic Problems, Part* I, Tech. Report 99-09, TICAM, 1999.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1073–1093

STEPSIZE RESTRICTIONS FOR THE TOTAL-VARIATION-DIMINISHING PROPERTY IN GENERAL RUNGE–KUTTA METHODS*

L. FERRACINA† AND M. N. SPIJKER†

Abstract. Much attention has been paid in the literature to total-variation-diminishing (TVD) numerical processes in the solution of nonlinear hyperbolic differential equations. For special Runge–Kutta methods, conditions on the stepsize were derived that are sufficient for the TVD property; see, e.g., Shu and Osher [*J. Comput. Phys.*, 77 (1988), pp. 439–471] and Gottlieb and Shu [*Math. Comp.*, 67 (1998), pp. 73–85]. Various basic questions are still open regarding the following issues: 1. the extension of the above conditions to more general Runge–Kutta methods; 2. simple restrictions on the stepsize which are not only sufficient but at the same time necessary for the TVD property; and 3. the determination of optimal Runge–Kutta methods with the TVD property.

In this paper we propose a theory by means of which we are able to clarify the above questions. Moreover, by applying our theory, we settle analogous questions regarding the related strong-stability-preserving (SSP) property (see, e.g., Gottlieb, Shu, and Tadmor [*SIAM Rev.*, 43 (2001), pp. 89–112] and Shu [*Collected Lectures on the Preservation of Stability under Discretization*, D. Estep and S. Tavener, eds., SIAM, Philadelphia, 2002]). Our theory can be viewed as a variant to a theory of Kraaijevanger [*BIT*, 31 (1991), pp. 482–528] on the contractivity of Runge–Kutta methods.

Key words. initial value problem, conservation law, method of lines, Runge–Kutta formula, total-variation-diminishing (TVD), strong-stability-preserving (SSP), monotonicity

AMS subject classifications. Primary, 65M20; Secondary, 65L05, 65L06

DOI. 10.1137/S0036142902415584

1. Introduction.

1.1. The purpose of the paper. In this paper we shall address some natural questions arising in the numerical solution of certain partial differential equations (PDEs). In order to formulate these questions, we consider an initial value problem for a system of ordinary differential equations (ODEs) of the form

$$\frac{d}{dt}U(t) = F(U(t)) \quad (t \geq 0), \quad U(0) = u_0. \tag{1.1}$$

We assume that (1.1) results from an application of the method of lines to a Cauchy problem for a PDE of the form

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}f(u(x,t)) = 0 \quad (t \geq 0, \quad -\infty < x < \infty).$$

Here f stands for a given (possibly nonlinear) scalar function, so that the PDE is a simple instance of a conservation law; cf., e.g., Kröner (1997) and LeVeque (2002).

The solution $U(t)$ to (1.1) stands for a (time dependent) vector in $\mathbb{R}^\infty = \{y : y = (\ldots, \eta_{-1}, \eta_0, \eta_1, \ldots)$ with $\eta_j \in \mathbb{R}$ for $j = 0, \pm 1, \pm 2, \ldots\}$. The components $U_j(t)$ of $U(t)$ are to approximate the desired true solution values $u(j\Delta x, t)$ (or cell averages

*Received by the editors October 4, 2002; accepted for publication (in revised form) September 26, 2003; published electronically July 29, 2004.
http://www.siam.org/journals/sinum/42-3/41558.html
†Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands (ferra@math.leidenuniv.nl, spijker@math.leidenuniv.nl). The research of the first author was partially supported by a Padova University grant, Dipartimento di Matematica Pura e Applicata, Università di Padova.

thereof); here Δx denotes a (positive) mesh-width. Furthermore, F stands for a function from $\mathbb{R}^\infty$ into $\mathbb{R}^\infty$; it depends on the given function f as well as on the process of semidiscretization being used. Finally, $u_0 \in \mathbb{R}^\infty$ depends on the initial data of the original Cauchy problem.

Any Runge–Kutta method, when applied to problem (1.1), yields approximations u_n to $U(n\Delta t)$, where $\Delta t > 0$ denotes the time step and $n = 1, 2, 3, \ldots$. Since $\frac{d}{dt}U(t) = F(U(t))$ stands for a semidiscrete version of a conservation law, it is desirable that the (fully discrete) process be *total-variation-diminishing (TVD)* in the sense that

$$\|u_n\|_{TV} \le \|u_{n-1}\|_{TV}; \tag{1.2}$$

here the function $\|.\|_{TV}$ is defined by

$$\|y\|_{TV} = \sum_{j=-\infty}^{+\infty} |\eta_j - \eta_{j-1}| \quad (\text{for } y \in \mathbb{R}^\infty \text{ with components } \eta_j).$$

For an explanation of the importance of the TVD property, particularly in the numerical solution of nonlinear conservation laws, see, e.g., Harten (1983), Laney (1998), Toro (1999), LeVeque (2002), and Hundsdorfer and Verwer (2003).

By Shu and Osher (1988) (see also, e.g., Gottlieb, Shu, and Tadmor (2001) and Shu (2002)) a simple but very useful approach was described for obtaining (high order) Runge–Kutta methods leading to TVD numerical processes. They considered explicit m-stage Runge–Kutta methods, written in the special form

$$\begin{aligned} y_1 &= u_{n-1}, \\ y_i &= \sum_{j=1}^{i-1} [\lambda_{ij}\, y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (2 \le i \le m+1), \\ u_n &= y_{m+1}. \end{aligned} \tag{1.3}$$

Here λ_{ij}, μ_{ij} are real coefficients specifying the Runge–Kutta method, and y_i are intermediate vectors in $\mathbb{R}^\infty$, depending on u_{n-1}, used for computing u_n (for $n = 1, 2, 3, \ldots$). Theorem 1.1 will state one of the conclusions formulated in the three papers just mentioned. It applies to the situation where the semidiscretization of the conservation law has been carried out in such a manner that the forward Euler method, applied to $\frac{d}{dt}U(t) = F(U(t))$, yields a fully discrete process which is TVD, when the stepsize Δt is suitably restricted, i.e.,

$$\|v + \Delta t\, F(v)\|_{TV} \le \|v\|_{TV} \quad (\text{whenever } 0 < \Delta t \le \tau_0 \text{ and } v \in \mathbb{R}^\infty). \tag{1.4}$$

Furthermore, in the theorem it is assumed that

$$\lambda_{i1} + \lambda_{i2} + \cdots + \lambda_{i,i-1} = 1 \quad (2 \le i \le m+1), \tag{1.5a}$$

$$\lambda_{ij} \ge 0, \quad \mu_{ij} \ge 0 \quad (1 \le j < i \le m+1), \tag{1.5b}$$

and the following notation is used:

$$c_{ij} = \lambda_{ij}/\mu_{ij} \ \ (\text{for } \mu_{ij} \ne 0), \quad c_{ij} = \infty \ \ (\text{for } \mu_{ij} = 0), \tag{1.6a}$$

$$c = \min_{i,j} c_{ij}. \tag{1.6b}$$

THEOREM 1.1 (Shu and Osher). *Assume* (1.5)*, and let* c *be defined by* (1.6)*. Suppose* (1.4) *holds, and*

$$0 < \Delta t \le c \cdot \tau_0. \tag{1.7}$$

Then process (1.3) *is TVD; i.e.,* (1.2) *holds whenever* u_n *is computed from* u_{n-1} *according to* (1.3).

It was remarked, notably in Shu and Osher (1988) and Gottlieb, Shu, and Tadmor (2001), that, under the assumptions (1.5), (1.6), the above theorem can be generalized. Let $\mathbb{V}$ be an arbitrary linear subspace of $\mathbb{R}^\infty$ and let $\|.\|$ denote any corresponding seminorm (i.e., $\|u+v\| \le \|u\| + \|v\|$ and $\|\lambda v\| = |\lambda| \cdot \|v\|$ for all $\lambda \in \mathbb{R}$ and $u, v \in \mathbb{V}$). A straightforward generalized version of Theorem 1.1 says that if $F : \mathbb{V} \to \mathbb{V}$ and

$$\|v + \Delta t F(v)\| \le \|v\| \quad (\text{whenever } 0 < \Delta t \le \tau_0 \text{ and } v \in \mathbb{V}), \tag{1.8}$$

then (1.7) still implies that

$$\|u_n\| \le \|u_{n-1}\|, \tag{1.9}$$

when u_n is computed from $u_{n-1} \in \mathbb{V}$ according to (1.3). In the last mentioned paper, time discretization methods for which a positive constant c exists such that (1.7), (1.8) always imply (1.9) were called *strong-stability-preserving (SSP)*. Property (1.9) is important, also with seminorms different from $\|.\|_{TV}$, and also when solving certain differential equations different from conservation laws—see, e.g., Dekker and Verwer (1984), LeVeque (2002), and Hundsdorfer and Verwer (2003).

Clearly, it would be awkward if the factor c, defined in (1.6), would be so small that (1.7) would reduce to a stepsize restriction which is too severe for any practical purposes—in fact, the less restrictions on Δt, the better. One might thus be tempted to take the magnitude of c into account when comparing the effectiveness of different Runge–Kutta processes (1.3), (1.5) to each other. However, it is evident that such a use of c, defined by (1.6), could be quite misleading if, for a given process (1.3), (1.5), the conclusion in Theorem 1.1 would also be valid with some factor c which is (much) larger than the one given by (1.6).

For any given method (1.3) satisfying (1.5), the question thus arises what is the largest factor c, *not necessarily defined via* (1.6), such that the conclusion in Theorem 1.1 is still valid. Moreover, a second question is of whether there exists a positive constant c such that (1.4), (1.7) imply (1.2), also for methods (1.3) satisfying (1.5a) but violating (1.5b). Two analogous questions arise in connection with the generalized version of Theorem 1.1, related to the SSP property, mentioned above.

The purpose of this paper is to propose a general theory which allows us to answer the above questions, as well as related ones.

1.2. Outline of the rest of the paper. In section 2 we present our general theory, just mentioned at the end of section 1.1. Section 2.1 contains notations and definitions which are basic for the rest of our paper. We review here the concept of *monotonicity*, which generalizes the TVD-property (1.2) in the context of arbitrary vector spaces $\mathbb{V}$, with seminorms $\|.\|$, and of general Runge–Kutta schemes (A, b). Furthermore, we introduce the notion of a *stepsize-coefficient* for monotonicity, which formalizes and generalizes the property of the coefficient c as stated in Theorem 1.1. In section 2.2 we recall the concept of irreducibility for general Runge–Kutta schemes (A, b), and we review the crucial quantity $R(A, b)$, introduced by Kraaijevanger (1991). In section 2.3 we present (without proof) our main result, Theorem 2.5. This theorem can be regarded as a variant to a theorem, on contractivity of Runge–Kutta methods, of Kraaijevanger (1991). Theorem 2.5 is relevant to arbitrary irreducible Runge–Kutta schemes (A, b); it tell us that, in the important situations specified by (2.9), (2.10), (2.11), respectively, the largest stepsize-coefficient for monotonicity is equal to $R(A, b)$.

In section 3 we apply Theorem 2.5 to a generalized version of process (1.3). After the introductory section 3.1, we clarify in the sections 3.2 and 3.3, respectively, the questions raised at the end of section 1.1 regarding the TVD and SSP properties of process (1.3). Section 3.4 gives two examples illustrating the superiority of the quantity $R(A, b)$ (to the factor c, given by (1.6)) as a guide to stepsize restrictions for the TVD and SSP properties.

Section 4 is mainly devoted to explicit Runge–Kutta schemes which are optimal, in the sense of their stepsize-coefficients for monotonicity. After the introductory section 4.1, we review, in section 4.2, conclusions of Kraaijevanger (1991) regarding the optimization of $R(A, b)$, in various classes of explicit Runge–Kutta schemes (A, b). Combining these conclusions and our Theorem 2.5, we are able to extend and shed new light on (recent) results in the literature about the optimization of c defined by (1.6). In section 4.3 we describe an algorithm for computing $R(A, b)$, which may be useful in determining further optimal Runge–Kutta methods. Section 4.4 contains a brief discussion of a few important related issues.

In order to look at our main result in the right theoretical perspective, we give in the final section, section 5, not only the formal proof of Theorem 2.5, but we present a short account of related material from Kraaijevanger (1991) as well. In section 5.1 we review Kraaijevanger's theorem mentioned above, and we compare it with our Theorem 2.5. In section 5.2 we give the proof of our main result.

We have framed our paper purposefully in the way just described: the reader who is primarily interested in our Theorem 2.5 and its applications (rather than in the underlying theory) will not be hampered by unnecessary digressions when reading sections 2, 3, and 4.

2. A general theory for monotonic Runge–Kutta processes.

2.1. Stepsize-coefficients for monotonicity in a general context. We want to study properties like (1.2) and (1.9) in a general setting. For that reason, we assume that $\mathbb{V}$ is an arbitrary real vector space, and that $F(v)$ is a given function, defined for all $v \in \mathbb{V}$, with values in $\mathbb{V}$. We consider a formal generalization of (1.1),

$$\frac{d}{dt}U(t) = F(U(t)) \quad (t \geq 0), \quad U(0) = u_0, \tag{2.1}$$

where u_0 and $U(t)$ stand for vectors in $\mathbb{V}$.

The general Runge–Kutta method with m stages, (formally) applied to the abstract problem (2.1), provides us with vectors $u_1, u_2, u_3, \ldots$ in $\mathbb{V}$ (see, e.g., Dekker and Verwer (1984), Butcher (1987), and Hairer and Wanner (1996)). Here u_n is related to u_{n-1} by the formula

$$u_n = u_{n-1} + \Delta t \sum_{j=1}^{m} b_j F(y_j), \tag{2.2a}$$

where the vectors y_j in $\mathbb{V}$ satisfy

$$y_i = u_{n-1} + \Delta t \sum_{j=1}^{m} a_{ij} F(y_j) \quad (1 \leq i \leq m). \tag{2.2b}$$

In these formulas, $\Delta t > 0$ denotes the stepsize and b_j, a_{ij} are real parameters, specifying the Runge–Kutta method. We always assume that $b_1 + b_2 + \cdots + b_m = 1$. If

$a_{ij} = 0$ (for $j \geq i$), the Runge–Kutta method is called *explicit*. Defining the $m \times m$ matrix A by $A = (a_{ij})$ and the column vector $b \in \mathbb{R}^m$ by $b = (b_1, b_2, b_3, \ldots, b_m)^T$, we can identify the Runge–Kutta method with the *coefficient scheme* (A, b).

Let $\|.\|$ denote an arbitrary seminorm on $\mathbb{V}$ (i.e., $\|u+v\| \leq \|u\| + \|v\|$ and $\|\lambda v\| = |\lambda| \cdot \|v\|$ for all real λ and $u, v \in \mathbb{V}$). The following inequality generalizes (1.2) and (1.9):

$$\|u_n\| \leq \|u_{n-1}\|. \tag{2.3}$$

We shall say that the Runge–Kutta method is *monotonic* (for the stepsize Δt, function F, and seminorm $\|.\|$) if (2.3) holds whenever the vectors u_{n-1} and u_n in $\mathbb{V}$ are related to each other as in (2.2). Our use of the term "monotonic" is nicely in agreement with earlier use of this term, e.g., by Burrage and Butcher (1980), Dekker and Verwer (1984, p. 263), Spijker (1986), Butcher (1987, p. 392), and Hundsdorfer, Ruuth, and Spiteri (2003). Property (2.3) is related to what sometimes is called *practical stability* or *strong stability*; see, e.g., Morton (1980) and Gottlieb, Shu, and Tadmor (2001).

In order to study stepsize restrictions for monotonicity, we start from a given stepsize $\tau_0 \in (0, \infty)$. We shall deal with the situation where F is a function from $\mathbb{V}$ into $\mathbb{V}$, satisfying

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad (\text{for all } v \in \mathbb{V}). \tag{2.4}$$

The last inequality implies, for $0 < \Delta t \leq \tau_0$, that $\|v + \Delta t F(v)\| = \|(1 - \Delta t/\tau_0)v + (\Delta t/\tau_0)(v + \tau_0 F(v))\| \leq \|v\|$. Consequently, (2.4) is equivalent to the following generalized version of (1.4) and (1.8):

$$\|v + \Delta t F(v)\| \leq \|v\| \quad (\text{whenever } 0 < \Delta t \leq \tau_0 \text{ and } v \in \mathbb{V}).$$

Let a Runge–Kutta method (A, b) be given. We shall study monotonicity of the method under arbitrary stepsize restrictions of the form

$$0 < \Delta t \leq c \cdot \tau_0. \tag{2.5}$$

Definition 2.1 (stepsize-coefficient for monotonicity). *A value $c \in (0, \infty]$ is called a stepsize-coefficient for monotonicity (with respect to $\mathbb{V}$ and $\|.\|$) if the Runge–Kutta method is monotonic, as in* (2.3)*, whenever F is a function from $\mathbb{V}$ to $\mathbb{V}$ satisfying* (2.4)*, and Δt is a (finite) stepsize satisfying* (2.5)*.*

It is easily verified that this definition is independent of the above value τ_0: if c is a stepsize-coefficient for monotonicity, with respect to $\mathbb{V}$ and $\|.\|$, using one particular value $\tau_0 > 0$, then c will have the same property when using any other value, say $\tau_0' > 0$.

The concept of a stepsize-coefficient as introduced in the above definition, corresponds to what is sometimes called a *CFL coefficient* in the context of discretizations for hyperbolic problems; see, e.g., Gottlieb and Shu (1998) and Shu (2002).

In subsection 2.3 we shall give maximal stepsize-coefficients for monotonicity with respect to various spaces $\mathbb{V}$ and seminorms $\|.\|$.

2.2. Irreducible Runge–Kutta schemes and the quantity $R(A, b)$. In this subsection we give some definitions which will be needed when we formulate our results, in subsection 2.3, about maximal stepsize-coefficients c. We start with the fundamental concepts of reducibility and irreducibility.

DEFINITION 2.2 (reducibility and irreducibility). *An m-stage Runge–Kutta scheme (A, b) is called reducible if (at least) one of the following two statements* (i), (ii) *is true; it is called irreducible if neither* (i) *nor* (ii) *is true.*

(i) *There exist nonempty, disjoint index sets M, N with $M \cup N = \{1, 2, \ldots, m\}$ such that $b_j = 0$ (for $j \in N$) and $a_{ij} = 0$ (for $i \in M$, $j \in N$);*

(ii) *there exist nonempty, pairwise disjoint index sets $M_1, M_2, \ldots, M_r$, with $1 \leq r < m$ and $M_1 \cup M_2 \cup \cdots \cup M_r = \{1, 2, \ldots, m\}$, such that $\sum_{k \in M_q} a_{ik} = \sum_{k \in M_q} a_{jk}$ whenever $1 \leq p \leq r$, $1 \leq q \leq r$, and $i, j \in M_p$.*

In case the above statement (i) is true, the vectors y_j in (2.2) with $j \in N$ have no influence on u_n, and the Runge–Kutta method is equivalent to a method with less than m stages. Also in case of (ii), the Runge–Kutta method essentially reduces to a method with less then m stages; see, e.g., Dekker and Verwer (1984) or Hairer and Wanner (1996). Clearly, for all practical purposes, it is enough to consider only Runge–Kutta schemes which are irreducible.

Next, we turn to a very useful characteristic quantity for Runge–Kutta schemes introduced by Kraaijevanger (1991). Following this author, we shall denote his quantity by $R(A, b)$, and in defining it, we shall use, for real ξ, the notations

$$A(\xi) = A(I - \xi A)^{-1}, \qquad b(\xi) = (I - \xi A)^{-T} b,$$
$$e(\xi) = (I - \xi A)^{-1} e, \qquad \varphi(\xi) = 1 + \xi b^T (I - \xi A)^{-1} e.$$

Here $^{-T}$ stands for transposition after inversion, I denotes the identity matrix of order m, and e stands for the column vector in $\mathbb{R}^m$, all of whose components are equal to 1. We shall focus on values $\xi \leq 0$ for which

$$I - \xi A \text{ is invertible}, \quad A(\xi) \geq 0, \quad b(\xi) \geq 0, \quad e(\xi) \geq 0, \quad \text{and} \quad \varphi(\xi) \geq 0. \tag{2.6}$$

The first inequality in (2.6) should be interpreted entrywise, the second and the third ones componentwise. Similarly, all inequalities for matrices and vectors occurring below are to be interpreted entrywise and componentwise, respectively.

DEFINITION 2.3 (the quantity $R(A, b)$). *Let (A, b) be a given coefficient scheme. In case $A \geq 0$ and $b \geq 0$, we define*

$$R(A, b) = \sup\{r : \ r \geq 0 \ \text{ and } (2.6) \text{ holds for all } \xi \text{ with } -r \leq \xi \leq 0\}.$$

In case (at least) one of the inequalities $A \geq 0$, $b \geq 0$ is violated, we define $R(A, b) = 0$.

Definition 2.3 suggests that it may be difficult to determine $R(A, b)$ for given coefficient schemes (A, b). However, in section 4 we shall see that (for explicit Runge–Kutta methods) *a simple algorithm exists for computing $R(A, b)$*. Moreover, Kraaijevanger (1991, p. 497) gave the following simple criterion (2.7) for determining whether $R(A, b) = 0$ or $R(A, b) > 0$. For any given $k \times l$ matrix $B = (b_{ij})$, we define the corresponding $k \times l$ *incidence matrix* by

$$\mathrm{Inc}(B) = (c_{ij}), \quad \text{with } c_{ij} = 1 \text{ (if } b_{ij} \neq 0) \text{ and } c_{ij} = 0 \text{ (if } b_{ij} = 0).$$

THEOREM 2.4 (about positivity of $R(A, b)$). *Let (A, b) be a given irreducible coefficient scheme. Then $R(A, b) > 0$ if and only if*

$$A \geq 0, \quad b > 0, \quad \textit{and} \quad \mathrm{Inc}(A^2) \leq \mathrm{Inc}(A). \tag{2.7}$$

Proof. For ξ sufficiently close to zero, the matrix $I - \xi A$ is invertible and $e(\xi) \geq 0$, $\varphi(\xi) \geq 0$. Therefore, it is sufficient to analyze the inequalities $A(\xi) \geq 0$ and $b(\xi) \geq 0$. With no loss of generality, we assume $A \geq 0$, $b \geq 0$.

For ξ close to zero, we have

$$A(\xi) = (A + \xi A^2)\sum_{k=0}^{\infty}(\xi A)^{2k} \quad \text{and} \quad b(\xi)^T = (b^T + \xi b^T A)\sum_{k=0}^{\infty}(\xi A)^{2k}.$$

From these two expressions, one easily sees that there exists a positive r, with

$$A(\xi) \geq 0 \quad \text{and} \quad b(\xi)^T \geq 0 \quad (\text{for } -r \leq \xi \leq 0)$$

if and only if $\text{Inc}(A^2) \leq \text{Inc}(A)$ and $\text{Inc}(b^T A) \leq \text{Inc}(b^T)$. Since statement (i) in Definition 2.2 is *not* true, we conclude that the last inequality is equivalent to $b > 0$. □

We note that, in Kraaijevanger (1991), one can find various other interesting properties related to $R(A, b)$, among them characterizations different from Definition 2.3.

2.3. Formulation of our main theorem. In this subsection we shall determine maximal stepsize-coefficients (Definition 2.1) with respect to general spaces $\mathbb{V}$ and seminorms $\|.\|$. Moreover, we shall pay special attention to the particular (semi)norms

$$\|y\|_\infty = \sup_{-\infty<j<\infty} |\eta_j|, \quad \|y\|_1 = \sum_{-\infty}^{\infty} |\eta_j|, \quad \|y\|_{TV} = \sum_{-\infty}^{\infty} |\eta_j - \eta_{j-1}|$$

for $y = (\ldots, \eta_{-1}, \eta_0, \eta_1, \ldots) \in \mathbb{R}^\infty$. Furthermore, for integers $s \geq 1$ and vectors $y \in \mathbb{R}^s$ with components η_j $(1 \leq j \leq s)$, we shall focus on the (semi)norms

$$\|y\|_\infty = \max_{1\leq j\leq s} |\eta_j|, \quad \|y\|_1 = \sum_{j=1}^{s} |\eta_j|, \quad \|y\|_{TV} = \sum_{j=2}^{s} |\eta_j - \eta_{j-1}|$$

(where $\sum_{j=2}^{s} |\eta_j - \eta_{j-1}| = 0$ for $s = 1$). In our Theorem 2.5, the following inequality will play a prominent part:

$$c \leq R(A, b). \tag{2.8}$$

Here is our main theorem, about stepsize-coefficients of irreducible Runge–Kutta schemes (Definitions 2.1 and 2.2).

THEOREM 2.5 (relating monotonicity to $R(A, b)$). *Consider an arbitrary irreducible Runge–Kutta scheme (A, b). Let c be a given value with $0 < c \leq \infty$. Choose one of the three (semi)norms $\|.\|_\infty$, $\|.\|_1$, or $\|.\|_{TV}$, and denote it by $|\!|.|\!|$. Then each of the following three statements is equivalent to* (2.8).

(2.9) *c is a stepsize-coefficient for monotonicity, with respect to all vector spaces $\mathbb{V}$ and seminorms $\|.\|$ on $\mathbb{V}$;*

(2.10) *c is a stepsize-coefficient for monotonicity, with respect to the special space $\mathbb{V} = \{y : \ y \in \mathbb{R}^\infty \text{ and } |\!|y|\!| < \infty\}$ and seminorm $\|.\| = |\!|.|\!|$;*

(2.11) *c is a stepsize-coefficient for monotonicity, with respect to the finite dimensional space $\mathbb{V} = \mathbb{R}^s$ and seminorm $\|.\| = |\!|.|\!|$ for $s = 1, 2, 3, \ldots$.*

Clearly, (2.9) is a priori a stronger statement than (2.10) or (2.11). Accordingly, the essence of Theorem 2.5 is that the (algebraic) property (2.8) implies the (strong) statement (2.9), whereas already either of the (weaker) statements (2.10) or (2.11) implies (2.8).

The above theorem highlights the importance of Kraaijevanger's quantity $R(A, b)$. Theorem 2.5 shows that, with respect to each of the three situations specified in (2.9), (2.10), and (2.11), *the maximal stepsize-coefficient for monotonicity is equal to* $R(A, b)$.

The above theorem will be compared with a theorem on nonlinear contractivity of Kraaijevanger (1991) in section 5.1, and it will be proved in section 5.2.

3. The application of our main theorem to the questions raised in subsection 1.1.

3.1. The equivalence of (a generalized version of) process (1.3) to method (2.2). In section 3 we study time stepping processes producing numerical approximations $u_n \in \mathbb{R}^\infty$ to $U(n\Delta t)$ (for $n \geq 1$), where $U(t) \in \mathbb{R}^\infty$ satisfies (1.1). We focus on processes of the form

$$y_1 = u_{n-1}, \tag{3.1a}$$

$$y_i = \sum_{j=1}^{m} [\lambda_{ij} y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (2 \leq i \leq m), \tag{3.1b}$$

$$u_n = \sum_{j=1}^{m} [\lambda_{m+1,j} y_j + \Delta t \cdot \mu_{m+1,j} F(y_j)]. \tag{3.1c}$$

Here λ_{ij}, μ_{ij} are arbitrary real coefficients with

$$\lambda_{i1} + \lambda_{i2} + \cdots + \lambda_{im} = 1 \quad (2 \leq i \leq m+1). \tag{3.2a}$$

Clearly, if $\lambda_{ij} = \mu_{ij} = 0$ (for $j \geq i$), the above process reduces to algorithm (1.3). Moreover, process (3.1) is sufficiently general to also cover other algorithms, such as the one in Gottlieb, Shu, and Tadmor (2001, p. 109), which was considered recently for solving (1.1).

In order to relate (3.1) to a Runge–Kutta method in the standard form (2.2), we define $\lambda_{ij} = \mu_{ij} = 0$ (for $i = 1$ and $1 \leq j \leq m$), and we introduce the $(m+1) \times m$ matrices $L = (\lambda_{ij})$, $M = (\mu_{ij})$. The $m \times m$ submatrices composed of the first m rows of L and M, respectively, will be denoted by L_0 and M_0. Furthermore, the last rows of L and M—that is, $(\lambda_{m+1,1}, \ldots, \lambda_{m+1,m})$ and $(\mu_{m+1,1}, \ldots, \mu_{m+1,m})$, respectively—will be denoted by L_1 and M_1, so that

$$L = \begin{pmatrix} L_0 \\ L_1 \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} M_0 \\ M_1 \end{pmatrix}. \tag{3.2b}$$

We assume that

$$\text{the } m \times m \text{ matrix } I - L_0 \text{ is invertible.} \tag{3.2c}$$

We shall now show that the relations (3.1) imply (2.2), with matrix $A = (a_{ij})$ and column vector $b = (b_i)$ specified by

$$A = (I - L_0)^{-1} M_0 \quad \text{and} \quad b^T = M_1 + L_1 A. \tag{3.3}$$

We denote the entries of the matrix $(I - L_0)^{-1}$ by γ_{ij}, and note that the relations (3.1a), (3.1b) can be rewritten as

$$\sum_{k=1}^{m} (\delta_{jk} - \lambda_{jk}) y_k = \delta_{j,1} u_{n-1} + \sum_{k=1}^{m} \mu_{jk} F_k \quad (\text{for } 1 \leq j \leq m), \tag{3.4}$$

where δ_{jk} is the Kronecker index and $F_k = \Delta t \cdot F(y_k)$. Multiplying (3.4) by γ_{ij} and summing over $j = 1, 2, \ldots, m$, we obtain, for $1 \le i \le m$, the equality $y_i = (\sum_{j=1}^{m} \gamma_{ij}\delta_{j,1})u_{n-1} + \sum_{k=1}^{m}(\sum_{j=1}^{m} \gamma_{ij}\mu_{jk})F_k$. In view of (3.2a), the first sum in the right-hand member of the last equality is equal to 1; hence (2.2b) holds with $(a_{ij}) = (I - L_0)^{-1}M_0$. Furthermore, in view of (3.1c), we easily arrive at (2.2a) with $(b_1, b_2, \ldots, b_m) = M_1 + L_1 A$.

Similarly to the above, the relations (2.2), (3.3) can be proved to imply (3.1), so that the following conclusion is valid.

LEMMA 3.1. *Let λ_{ij} and μ_{ij} be given coefficients satisfying* (3.2a), (3.2b), (3.2c). *Define the Runge–Kutta scheme (A, b) by* (3.3). *Then the relations* (3.1) *are equivalent to* (2.2).

In the following subsections, we shall use this lemma for relating the monotonicity properties of process (3.1) to those of the corresponding Runge–Kutta scheme (A, b) given by (3.3).

3.2. The total-variation-diminishing property of process (3.1). Our following Theorem 3.2 gives a stepsize restriction guaranteeing the TVD-property for the general process (3.1). Since (3.1) is more general than process (1.3), our theorem is highly relevant to (1.3). In the theorem, we shall use the notation

$$\mathbb{R}^{\infty}_{TV} = \{y : \ y \in \mathbb{R}^{\infty} \text{ with } \|y\|_{TV} < \infty\},$$

where $\|.\|_{TV}$ has the same meaning as in subsection 1.1. We shall deal with functions F from $\mathbb{R}^{\infty}_{TV}$ into $\mathbb{R}^{\infty}_{TV}$, satisfying

$$\|v + \tau_0 F(v)\|_{TV} \le \|v\|_{TV} \quad (\text{whenever } v \in \mathbb{R}^{\infty}_{TV}), \tag{3.5}$$

and with stepsize restrictions of the form

$$0 < \Delta t \le R(A, b) \cdot \tau_0 \tag{3.6}$$

(see Definition 2.3).

THEOREM 3.2 (optimal stepsize restriction for the TVD-property in process (3.1)). *Let λ_{ij} and μ_{ij} be given coefficients satisfying* (3.2a), (3.2b), (3.2c). *Define the matrix A and the vector b by* (3.3), *and suppose that the coefficient scheme (A, b) is irreducible (Definition* 2.2). *Let F be a function from $\mathbb{R}^{\infty}_{TV}$ into $\mathbb{R}^{\infty}_{TV}$ satisfying* (3.5), *and let Δt be a (finite) stepsize satisfying* (3.6).

Then, process (3.1) *is TVD; i.e., the inequality* (1.2) *holds whenever $u_{n-1}, u_n \in \mathbb{R}^{\infty}_{TV}$ are related to each other as in* (3.1).

Proof. We apply Lemma 3.1, and consider the Runge–Kutta scheme (A, b) specified by the lemma. Next, we apply Theorem 2.5: choosing $c = R(A, b)$, we have (2.8) so that (2.10) must be fulfilled with $|\!|.|\!| = \|.\|_{TV}$. An application of Definition 2.1 completes the proof of the theorem. □

Remark 3.3. The above theorem has a *wider scope than Theorem* 1.1. The class of numerical methods (3.1) satisfying (3.2a), (3.2b), (3.2c) encompasses all processes (1.3) satisfying (1.5a), as well as other (implicit) procedures. Specifically, unlike Theorem 1.1, the above Theorem 3.2 is relevant to processes (1.3) satisfying (1.5a) but violating (1.5b)—see Example 3.7 in subsection 3.4 for an illustration.

Remark 3.4. The above theorem, when applied to any process (1.3) satisfying (1.5a), (1.5b), gives a *stronger conclusion than Theorem* 1.1. By Theorem 2.5, property (2.10) with $|\!|.|\!| = \|.\|_{TV}$ implies inequality (2.8). Therefore the coefficient c, given by Theorem 1.1, satisfies $c \le R(A, b)$; this means that the stepsize restriction (3.6) of

Theorem 3.2 is, in general, less severe than the restriction (1.7) of Theorem 1.1—see Example 3.8 in subsection 3.4 for an illustration.

Remark 3.5. Theorem 3.2 gives a stepsize restriction which is *optimal* in that the conclusion of the theorem would no longer be valid if the factor $R(A,b)$ in (3.6) would be replaced by any factor $c > R(A,b)$. This follows again from Theorem 2.5.

3.3. The strong-stability-preserving property of process (3.1). Let $\mathbb{V}$ be an arbitrary linear subspace of $\mathbb{R}^\infty$, and let $\|.\|$ denote any seminorm on $\mathbb{V}$. For functions $F : \mathbb{V} \longrightarrow \mathbb{V}$ satisfying

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad (\text{whenever } v \in \mathbb{V}), \tag{3.7}$$

we shall consider process (3.1) under a stepsize restriction of the form

$$0 < \Delta t \leq c \cdot \tau_0. \tag{3.8}$$

Following the terminology of Gottlieb, Shu, and Tadmor (2001), already reviewed in subsection 1.1, we shall say that process (3.1) is *strong-stability-preserving* (SSP) if a positive constant c exists (only depending on λ_{ij} and μ_{ij}) such that (1.9) holds whenever (3.1), (3.7), (3.8) are fulfilled.

THEOREM 3.6 (criterion for the SSP property of process (3.1)). *Let λ_{ij} and μ_{ij} be given coefficients satisfying* (3.2a), (3.2b), (3.2c). *Define the matrix A and vector b by* (3.3), *and suppose that the coefficient scheme (A,b) is irreducible (Definition* 2.2). *Then process* (3.1) *is SSP if and only if* (2.7) *holds.*

Proof. By Lemma 3.1 and Theorem 2.5, process (3.1) is SSP if and only if $R(A,b) > 0$. According to Theorem 2.4, the last inequality is equivalent to (2.7). ∎

It is clear that the above Theorem 3.6, similarly as Theorem 3.2, is highly relevant to all numerical processes (1.3) satisfying (1.5a); see Examples 3.7 and 3.8 below for illustrations.

3.4. Illustrations to Theorems 3.2 and 3.6. We give two examples illustrating Theorems 3.2 and 3.6.

Example 3.7. Consider process (1.3), with $m = 3$ and coefficients λ_{ij}, μ_{ij} given by the relations

$$\begin{pmatrix} \lambda_{21} & & \\ \lambda_{31} & \lambda_{32} & \\ \lambda_{41} & \lambda_{42} & \lambda_{43} \end{pmatrix} = \begin{pmatrix} 1 & & \\ \frac{1}{4} & \frac{3}{4} & \\ 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} \mu_{21} & & \\ \mu_{31} & \mu_{32} & \\ \mu_{41} & \mu_{42} & \mu_{43} \end{pmatrix} = \begin{pmatrix} 1 & & \\ -\frac{1}{2} & \frac{1}{4} & \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{pmatrix}.$$

Since $\mu_{31} < 0$, condition (1.5b) is violated; therefore Theorem 1.1 does not apply.

For the corresponding matrix $A = (a_{ij})$ and vector $b = (b_i)$ (see (3.3)), we have $a_{ij} = 0$ $(j \geq i)$, $a_{21} = 1$, $a_{31} = a_{32} = 1/4$ and $b_1 = b_2 = 1/6$, $b_3 = 2/3$, respectively. It is very easy to see that (2.7) holds; by virtue of Theorem 3.6, the numerical process is thus SSP. Moreover, according to Kraaijevanger (1991, Theorem 9.4), for this process we have $R(A,b) = 1$. By Theorem 3.2 we conclude that the process is TVD, under the assumption (3.5) if $0 < \Delta t \leq \tau_0$. We note that essentially the same numerical process was presented earlier by Shu and Osher (1988); we shall come back to it in section 4.2 (Remark 4.4; $m = p = 3$).

Example 3.8. Consider process (1.3), with $m = 2$ and

$$\begin{pmatrix} \lambda_{21} & \\ \lambda_{31} & \lambda_{32} \end{pmatrix} = \begin{pmatrix} 1 & \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} \mu_{21} & \\ \mu_{31} & \mu_{32} \end{pmatrix} = \begin{pmatrix} 1/2 & \\ 1/2 & 1/2 \end{pmatrix}.$$

The conditions (1.5a), (1.5b) are neatly fulfilled, but the coefficient c, defined by (1.6), is equal to 0.

For the corresponding Runge–Kutta scheme (A, b), defined by (3.3), we have $a_{ij} = 0$ $(j \geq i)$, $a_{21} = 1/2$ and $b_1 = b_2 = 1/2$. Clearly, (2.7) is fulfilled, guaranteeing the SSP property (see Theorem 3.6). Moreover, according to Kraaijevanger (1991, Theorem 9.2), we have $R(A, b) = 2$. Therefore, by Theorem 3.2, the numerical process is TVD, under assumption (3.5), if $0 < \Delta t \leq 2 \cdot \tau_0$. We note that the same method was presented by Spiteri and Ruuth (2002); we shall come back to it in section 4.2 (Remark 4.4; $m = 2$, $p = 1$).

4. Optimal Runge–Kutta methods.

4.1. Preliminaries. For integer values $m \geq 1$ and $p \geq 1$, we shall denote by $E_{m,p}$ the class of all explicit m-stage Runge–Kutta methods (A, b) with (classical) order of accuracy at least p. Considerable attention has been paid, in the literature, to identifying methods of class $E_{m,p}$ of the special form (1.3), (1.5) which are optimal in the sense of the coefficient c given by (1.6); see notably Shu and Osher (1988), Gottlieb and Shu (1998), Ruuth and Spiteri (2002), Shu (2002), and Spiteri and Ruuth (2002). Independently of this work, Kraaijevanger (1991) dealt with the optimization, in the full class $E_{m,p}$, of his quantity $R(A, b)$. Our theory (section 2) can be used to relate his conclusions to the work just mentioned about optimization of c defined in (1.6).

In section 4.2 we shall briefly review some of Kraaijevanger's conclusions so as to arrive at extensions and completions of the material, referred to above, on optimality in the sense of c (1.6). Furthermore, we shall consider scaled stepsize-coefficients which reflect the efficiency of the methods better than the unscaled coefficients; in Table 4.1 we shall display optimal scaled stepsize-coefficients. Next, in section 4.3, we shall focus on an algorithm for computing $R(A, b)$; the authors feel that it can be useful in (future) calculations for determining, numerically, optimal Runge–Kutta methods. Finally, in section 4.4 we touch upon a few important related issues.

4.2. Optimal methods in the class $E_{m,p}$. We start with the following fundamental lemma, which gives a simple upper bound for $R(A, b)$ in the class $E_{m,p}$.

LEMMA 4.1 (Kraaijevanger (1991, p. 517)). *Let $1 \leq p \leq m$, and consider an arbitrary Runge–Kutta method (A, b) of class $E_{m,p}$. Then $R(A, b) \leq m - p + 1$.*

Remark 4.2. Ruuth and Spiteri (2002, Theorem 3.1) showed that, for Runge–Kutta methods in class $E_{m,p}$ of the special form (1.3), (1.5), the coefficient c defined by (1.6) satisfies $c \leq m-p+1$. Clearly, a combination of the above lemma and our theory (section 2) yields an extension and improvement over the last bound on c: for *any* Runge–Kutta method of class $E_{m,p}$, *any* stepsize-coefficient for monotonicity, say c', and *any* of the situations covered by (2.9), (2.10), or (2.11), we have $c' \leq m - p + 1$.

The following theorem specifies methods (A, b) for which the upper bound $R(A, b) \leq m - p + 1$ of Lemma 4.1 becomes an equality.

THEOREM 4.3 (Kraaijevanger (1991, pp. 518–520)).

(a) *Let $p = 1 \leq m$. Then there is a unique method (A, b) of class $E_{m,p}$ with $R(A, b) = m$; it is given by $a_{ij} = 1/m$ $(1 \leq j < i \leq m)$ and $b_i = 1/m$ $(1 \leq i \leq m)$.*

(b) *Let $p = 2 \leq m$. Then there is a unique method (A, b) of class $E_{m,p}$ with $R(A, b) = m - 1$; it is given by $a_{ij} = 1/(m - 1)$ $(1 \leq j < i \leq m)$ and $b_i = 1/m$ $(1 \leq i \leq m)$.*

(c) *Let $p = 3$, $m = 3$. Then there is a unique method (A, b) of class $E_{m,p}$ with $R(A, b) = 1$; it is given by $a_{21} = 1$, $a_{31} = a_{32} = 1/4$, $b_1 = b_2 = 1/6$, and $b_3 = 2/3$.*

(d) *Let $p = 3$, $m = 4$. Then there is a unique method (A, b) of class $E_{m,p}$ with $R(A, b) = 2$; it is given by $a_{21} = a_{31} = a_{32} = b_4 = 1/2$ and $a_{4,i} = b_i = 1/6$ $(1 \le i \le 3)$.*

Remark 4.4. Essentially the same methods as specified in the above theorem, for $m = p = 2$ and $m = p = 3$, were already found by Shu and Osher (1988) in a search for methods in $E_{m,p}$, of the special type (1.3), (1.5), with maximal c (defined in (1.6)); Gottlieb and Shu (1998) proved optimality for these two methods with respect to c, (1.6). In an analogous search, Spiteri and Ruuth (2002) arrived at all other methods specified by the theorem, and proved optimality in the sense of c, (1.6). Similarly as in Remark 4.2, our theory (section 2) can be used here to conclude that all methods given in Theorem 4.3 are optimal (with respect to their stepsize-coefficients for monotonicity) in a *stronger sense*, and over a *larger class* of Runge–Kutta methods, than can be concluded from the three papers just mentioned.

Kraaijevanger (1991) did not specify analytically any methods (A, b) in $E_{m,p}$ with maximal $R(A, b)$, for pairs p, m different from those in Theorem 4.3. However, he arrived at interesting (negative) conclusions: if method (A, b) is of class $E_{m,p}$ and $p = 3$, $m \ge 5$, then $R(A, b) < m-p+1$; and if (A, b) belongs to $E_{m,p}$ with $p = m = 4$ or $p \ge 5$, then $R(A, b) = 0$. Moreover, by combining Kraaijevanger (1986, Theorem 5.1), Spijker (1983), and our Theorem 2.5, one can conclude that $R(A, b) < m - p + 1$ also for all (A, b) in $E_{m,p}$ with $p = 4$, $m \ge 6$. A combination of these conclusions and our theory (section 2) amounts to a far-reaching extension of related results obtained in Ruuth and Spiteri (2002).

Kraaijevanger (1991, pp. 522–523) constructed numerically an explicit 5-stage method (A, b) of order 4, with $R(A, b) \approx 1.508$. It is interesting to note that the same method was found by Spiteri and Ruuth (2002) in a numerical search within the class of methods (1.3) satisfying (1.5). By a similar search, the last authors also found a 5-stage method of order 3 with $c \approx 2.651$ (given by (1.6)). In view of Kraaijevanger (1986, Theorem 5.3), Spijker (1983), and our Theorem 2.5, we can conclude that this method has a value $R(A, b) \approx 2.651$, and is optimal in a *stronger sense* and over a *larger class* of methods than follows from Spiteri and Ruuth (2002).

Clearly, when comparing two explicit Runge–Kutta methods to each other, one cannot simply say that the one with the largest value $R(A, b)$ is the most efficient one. However, assuming that the stepsize Δt, used for solving (1.1) over some interval $[0, T]$, is governed by monotonicity (TVD) demands, it seems reasonable to use the quantity $m \cdot T/R(A, b)$ as a measure of the amount of computational labor of a Runge–Kutta method (A, b) with m stages—cf. Jeltsch and Nevanlinna (1981), Kraaijevanger (1986), and Spiteri and Ruuth (2002) for related considerations. In line with the terminology in the first two of these papers, we shall refer to the ratio $R(A, b)/m$ as a *scaled stepsize-coefficient*. The above mentioned measure, for the amount of computational labor, is inversely proportional to $R(A, b)/m$, so the scaled stepsize-coefficient is a more realistic guide than $R(A, b)$ for comparing the efficiency of different methods to each other.

In Table 4.1 we display scaled stepsize-coefficients of Runge–Kutta methods (A, b), which were reviewed above and are optimal in $E_{m,p}$ with respect to $R(A, b)$.

From the table, one may conclude that, for given p, it is advantageous to use optimal methods with relatively large m. Clearly, this conclusion is (only) justifiable under the above assumption about Δt being determined by monotonicity demands. For related numerical experiments, see, e.g., Gottlieb and Shu (1998) and Spiteri and Ruuth (2002).

TABLE 4.1
Scaled stepsize-coefficients $R(A,b)/m$ for optimal Runge–Kutta methods in $E_{m,p}$.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$p=1$	1	1	1	1	1
$p=2$		0.500	0.667	0.750	0.800
$p=3$			0.333	0.500	0.530
$p=4$					0.302

4.3. An algorithm for computing $R(A,b)$, for methods of class $E_{m,p}$. Below we will describe a simple algorithm for computing $R(A,b)$ whenever (A,b) is an irreducible Runge–Kutta scheme of class $E_{m,p}$. The following lemma plays a fundamental role in the algorithm.

LEMMA 4.5 (Kraaijevanger (1991, pp. 497–498)). *Let (A,b) be an irreducible coefficient scheme and r a positive real number. Then $R(A,b) \geq r$ if and only if $A \geq 0$ and the conditions* (2.6) *are fulfilled at $\xi = -r$.*

It was noted by Kraaijevanger (1991) that the above lemma simplifies calculating $R(A,b)$ if $A \geq 0$: for checking the conditions (2.6) on the whole of an interval $[-r,0]$, it is sufficient to consider only the left endpoint $\xi = -r$.

Let Test1 and Test2(x) be boolean functions defined by

$$\mathsf{Test1} = \begin{cases} \mathsf{true} & \text{if (2.7) holds,} \\ \mathsf{false} & \text{otherwise;} \end{cases} \qquad \mathsf{Test2(x)} = \begin{cases} \mathsf{true} & \text{if (2.6) holds at } \xi = \mathsf{x}, \\ \mathsf{false} & \text{otherwise.} \end{cases}$$

From Lemma 4.1 we know that if (A,b) is a coefficient scheme of class $E_{m,p}$, then $R(A,b) \leq m - p + 1$. In view of the last inequality, Theorem 2.4, and Lemma 4.5, we can calculate $R(A,b)$ with the wanted precision Tol, by using the above boolean functions as well as two pointers LeftExtr and RightExtr. The following algorithm finds $R(A,b)$ with error $\leq$ Tol.

```
x=0
if Test1
   LeftExtr=-(m-p+1),   RightExtr=0,   x=LeftExtr
   while (RightExtr-LeftExtr ≥ 2·Tol)
      if Test2(x)
         RightExtr=x,   x=(LeftExtr+RightExtr)/2
      else
         LeftExt=x,   x=(LeftExtr+RightExtr)/2
      end
   end
end
R(A,b)=−x.
```

4.4. Final remarks. For completeness, we note that Gottlieb and Shu (1998), Shu (2002), and Spiteri and Ruuth (2002) gave useful results regarding the optimization of c, (1.6), over classes of low-storage schemes of the (special) form (1.3), (1.5). Furthermore, Kennedy, Carpenter, and Lewis (2000) obtained interesting related results regarding the optimization of $R(A,b)$ over general classes of low-storage schemes (A,b). Clearly, our theory (section 2) is fit to put also this work in a wider perspective.

Above, in section 4, we dealt exclusively with explicit Runge–Kutta schemes. However, in Kraaijevanger (1991) also (a few) results were obtained, regarding the size of $R(A,b)$, relevant to implicit schemes—see below. A combination of these results

with our Theorem 2.5 immediately leads to interesting conclusions about stepsize-coefficients for monotonicity.

For arbitrary (possibly implicit) schemes (A, b) of order p, the following general results were obtained in Kraaijevanger (1991, pp. 514, 516): if $p \geq 2$, then $R(A, b) < \infty$; and if $p \geq 7$, then $R(A, b) = 0$. Moreover (on p. 516 of that article), a notable implicit method (A, b) was given, with a value $R(A, b)$ exceeding the upper bound of Lemma 4.1: the method with $m = 2$, $a_{1,1} = a_{1,2} = 0$, $a_{2,1} = a_{2,2} = 3/8$, $b_1 = 1/3$, $b_2 = 2/3$ is of order $p = 2$ and has a value $R(A, b) = 8/3$. The last value is considerably larger than the optimal value $m - p + 1 = 1$, which can be achieved in $E_{2,2}$ (cf. section 4.2); but this advantage should of course be balanced against the additional amount of work per step due to the implicitness of the method.

We think that it would be very useful to perform a systematic search for implicit methods which are optimal, for given m and p, in the sense of $R(A, b)$. Because such a search is beyond the scope of our present work, we do not go further into this matter here.

Finally, we note that our algorithm in section 4.3 can easily be adapted so as to compute $R(A, b)$ also for methods (A, b), of order at least 2, which are implicit: we still base the algorithm on Lemma 4.5, and (instead of using Lemma 4.1) we start with $\mathsf{LeftExtr} = \xi$, where ξ is a negative value at which (2.6) is violated; in view of the bound $R(A, b) < \infty$, such a ξ can be found, e.g., by a simple doubling process.

5. Kraaijevanger's theory and our proof of Theorem 2.5.

5.1. A theorem of Kraaijevanger on contractivity. Kraaijevanger (1991) presented an interesting theory, relevant to method (2.2) in the situation where F is a function from $\mathbb{R}^s$ into $\mathbb{R}^s$, and $\|.\|$ is a norm on $\mathbb{R}^s$. The focus in his paper is on numerical processes which, for given F, $\|.\|$, and Δt, are *contractive* in the sense that

$$\|\tilde{u}_n - u_n\| \leq \|\tilde{u}_{n-1} - u_{n-1}\| \tag{5.1}$$

whenever both the vectors u_{n-1}, u_n and the vectors $\tilde{u}_{n-1}, \tilde{u}_n$ are related to each other as in (2.2). Kraaijevanger studied property (5.1) for functions F satisfying

$$\|F(\tilde{v}) - F(v) + \rho(\tilde{v} - v)\| \leq \rho\|\tilde{v} - v\| \quad (\text{for all } v, \tilde{v} \in \mathbb{R}^s). \tag{5.2}$$

Here ρ is a positive constant; in the literature on numerical ODEs one often refers to (5.2) as a *circle condition* (with radius ρ) on the function F—cf. Kraaijevanger (1991).

In order to be able to reformulate one of Kraaijevanger's main results in such a way that it can easily be compared to our Theorem 2.5, we consider stepsize-restrictions of the form

$$0 < \Delta t \leq c/\rho. \tag{5.3}$$

Furthermore, adapting our Definition 2.1 to the situation at hand, we arrive at the following definition.

DEFINITION 5.1 (stepsize-coefficient for contractivity). *A value $c \in (0, \infty]$ is a stepsize-coefficient for contractivity (with respect to $\mathbb{R}^s$ and $\|.\|$) if the Runge–Kutta method is contractive, as in* (5.1)*, whenever $F : \mathbb{R}^s \to \mathbb{R}^s$ satisfies* (5.2) *and Δt is a (finite) stepsize satisfying* (5.3).

The subsequent theorem is an easy consequence of Kraaijevanger (1991, Theorem 5.4); it relates stepsize-coefficients for contractivity to the inequality

$$c \leq R(A, b). \tag{5.4}$$

THEOREM 5.2 (relating contractivity to $R(A,b)$). *Consider an arbitrary irreducible Runge–Kutta scheme (A,b). Let c be a given value with $0 < c \leq \infty$. Then both of the following statements are equivalent to* (5.4)*:*

(5.5) *c is a stepsize-coefficient for contractivity, with respect to $\mathbb{R}^s$ and $\|.\|$ for each $s \geq 1$ and each norm $\|.\|$ on $\mathbb{R}^s$;*

(5.6) *c is a stepsize-coefficient for contractivity, with respect to $\mathbb{R}^s$ and the special norm $\|.\|_\infty$ for each $s \geq 1$.*

Since condition (5.2) is equivalent to requiring that the forward Euler method with stepsize $\tau_0 = 1/\rho$ is contractive, there is a close resemblance between (5.2) and (2.4) (with $\mathbb{V} = \mathbb{R}^s$). Accordingly, one might think that (part of) our Theorem 2.5 is a simple consequence of Theorem 5.2. However, the following three remarks indicate that the relation between the two theorems is far from being that simple.

Remark 5.3. Let c be as in statement (2.11), with seminorm $\|.\| = \|.\|_1$ or $\|.\| = \|.\|_{TV}$. Theorem 2.5 claims that this coefficient c must satisfy $c \leq R(A,b)$. This claim cannot be expected to follow from the above Theorem 5.2; at best, it might follow from a version of that theorem in which the norm $\|.\|_\infty$ (in (5.6)) would simply be replaced by $\|.\|_1$ or $\|.\|_{TV}$. However, it is not known whether such a version is actually valid—Kraaijevanger's proof, underlying Theorem 5.2 as formulated above, makes an essential use of a specific (geometric) property of the norm $\|.\|_\infty$ which is *not* valid for $\|.\|_1$ or $\|.\|_{TV}$; cf. Kraaijevanger (1991, p. 505) and Schönbeck (1967, Theorem 2.4) for more details.

Remark 5.4. Let c be as in (2.11), with $\|.\| = \|.\|_\infty$. Even in this more convenient situation, it is not evident how the inequality $c \leq R(A,b)$, claimed by Theorem 2.5, could follow from Theorem 5.2. The fact is that (2.11) (with $\|.\| = \|.\|_\infty$) does not imply (5.6), because, in general, monotonicity does *not* imply contractivity.

Remark 5.5. Suppose $c \leq R(A,b)$. Then Theorem 2.5 claims that (2.9) is valid so that c would certainly be a stepsize-coefficient for monotonicity, with respect to $\mathbb{R}^s$ and any norm on $\mathbb{R}^s$. Even this last property of c does not follow from a simple application of Theorem 5.2, because it is no obvious consequence of (5.5)—note that (2.4) (with $\mathbb{V} = \mathbb{R}^s$) does *not* imply (5.2) (with $\rho = 1/\tau_0$).

The above three remarks make clear that our Theorem 2.5 can be viewed as a variant of Theorem 5.2 covering essentially new situations.

5.2. The proof of Theorem 2.5.

5.2.1. Preliminaries. Throughout this section 5.2 we assume, unless specified otherwise, that (A,b), c, and $|\!|.|\!|$ are as explained at the beginning of Theorem 2.5. With no loss of generality, we assume that c is finite. Below we shall prove the theorem by showing that the following five implications are valid: (2.8) $\Longrightarrow$ (2.9), (2.9) $\Longrightarrow$ (2.10), (2.10) $\Longrightarrow$ (2.11), [(2.11) with $|\!|.|\!| = \|.\|_{TV}$] $\Longrightarrow$ [(2.11) with $|\!|.|\!| = \|.\|_1$], and finally [(2.11) with $|\!|.|\!| = \|.\|_1$ or $\|.\|_\infty$] $\Longrightarrow$ (2.8).

The first implication will be proved in section 5.2.2, using arguments which are analogous to arguments for proving that (5.4) implies (5.5) (see Kraaijevanger (1991, pp. 502–504)).

The second implication is trivial, whereas the third and fourth implication will be proved in section 5.2.3. The proofs, in this section, are *not* related to arguments used in Kraaijevanger (1991), but are based on Lemma 5.6. This lemma gives a general framework in which the property of c being a stepsize-coefficient for monotonicity

can be carried over from a space $\mathbb{Y}$ with seminorm $\|.\|_{\mathbb{Y}}$ to another space $\mathbb{X}$ with seminorm $\|.\|_{\mathbb{X}}$.

The proof of the fifth implication will be given in section 5.2.4.

In that section we shall first deal with a linear variant of process (2.2). Lemma 5.7 tells us that a monotonicity property of that variant implies (2.8); the lemma is relevant to the norms $\|.\|_p$, with $p = 1$ and $p = \infty$. This lemma, with value $p = \infty$, was used implicitly by Kraaijevanger (1991, pp. 507–508) in a proof related to the implication (5.6) $\Longrightarrow$ (5.4) (cf. Theorem 5.2).

Next, we shall give Lemma 5.8, which states that property (2.11), with $|\!|.|\!| = \|.\|_p$ and $p = 1$ or $p = \infty$, implies the monotonicity property of the linear variant considered in Lemma 5.7. A combination of Lemmas 5.7 and 5.8 proves the fifth implication. Our proof of Lemma 5.8 has no relation to arguments in Kraaijevanger (1991); it makes use, among other things, of arguments employed earlier in Spijker (1986).

For completeness we mention that no counterpart of Lemma 5.8 is known to the authors which is relevant to contractivity with respect to $\mathbb{R}^s$ and $\|.\|_1$—cf. Remark 5.3 and Kraaijevanger (1991, p. 505).

5.2.2. Statement (2.8) $\Longrightarrow$ statement (2.9). We start this subsection by introducing some notation relevant to the vector space $\mathbb{V}$. For any vectors $v_1, v_2, \ldots, v_m$ in $\mathbb{V}$, we shall denote the vector in $\mathbb{V}^m$ with components v_j by

$$v = [v_j] = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \in \mathbb{V}^m.$$

Furthermore, for any (real) $l \times m$ matrix $B = (b_{ij})$, we define a corresponding linear operator $B_{\mathbb{V}}$, from $\mathbb{V}^m$ to $\mathbb{V}^l$, by $B_{\mathbb{V}}(v) = w$, for $v = [v_j] \in \mathbb{V}^m$, where $w = [w_i] \in \mathbb{V}^l$ with $w_i = \sum_{j=1}^m b_{ij} v_j$ $(1 \le i \le l)$. Clearly, if B and C are $l \times m$ matrices and D is an $m \times k$ matrix, then $(B + C)_{\mathbb{V}} = B_{\mathbb{V}} + C_{\mathbb{V}}$, $(\lambda B)_{\mathbb{V}} = \lambda \cdot B_{\mathbb{V}}$, and $(BD)_{\mathbb{V}} = B_{\mathbb{V}} \cdot D_{\mathbb{V}}$. Here, the addition and multiplications occurring in the last three left-hand members stand for the usual algebraic operations for matrices, whereas the addition and multiplications in the right-hand members apply to linear operators. The last three equalities will underlie part of our subsequent calculations.

Assume (2.8), and let F be a function from $\mathbb{V}$ to $\mathbb{V}$ satisfying (2.4). We have to prove that c is a stepsize-coefficient for monotonicity; i.e., $0 < \Delta t \le c \cdot \tau_0$ implies $\|u_n\| \le \|u_{n-1}\|$ whenever u_n and u_{n-1} are related to each other by (2.2).

Assuming (2.2), with $0 < \Delta t \le c \cdot \tau_0$, we obtain

$$u_n = u_{n-1} + \sum_{j=1}^m b_j w_j, \tag{5.7a}$$

$$y_i = u_{n-1} + \sum_{j=1}^m a_{ij} w_j \quad (1 \le i \le m), \tag{5.7b}$$

where $w_j = \Delta t F(y_j)$. Putting $\gamma = \Delta t/\tau_0$, we have $\|w_i + c y_i\| = \gamma \|(c/\gamma) y_i + \tau_0 F(y_i)\| \le \gamma\{(c/\gamma - 1)\|y\| + \|y_i + \tau_0 F(y_i)\|\}$. Therefore, in view of (2.4),

$$\|w_i + c y_i\| \le c \|y_i\|. \tag{5.8}$$

Defining $y = [y_i] \in \mathbb{V}^m$, $w = [w_i] \in \mathbb{V}^m$, and $e = (1, \ldots, 1)^T \in \mathbb{R}^m$, we can rewrite (5.7) as

$$u_n = u_{n-1} + \mathbf{b}^T w, \tag{5.9a}$$

$$y = \mathbf{e}u_{n-1} + \mathbf{A}w, \tag{5.9b}$$

where $\mathbf{b}^T = (b^T)_{\mathbb{V}}$, $\mathbf{e} = (e)_{\mathbb{V}}$, and $\mathbf{A} = A_{\mathbb{V}}$. Denoting the identity in $\mathbb{V}^m$ by $\mathbf{I}$, we see from (5.9b) that $(\mathbf{I} + c\mathbf{A})y = \mathbf{e}u_{n-1} + \mathbf{A}w + c\mathbf{A}y = \mathbf{e}u_{n-1} + \mathbf{A}(w + cy)$. From Lemma 4.5, we conclude that (2.6) holds with $\xi = -c$ and that $A \geq 0$. Therefore, $\mathbf{I} + c\mathbf{A}$ is invertible and

$$y = (\mathbf{I} + c\mathbf{A})^{-1}\mathbf{e}u_{n-1} + \mathbf{A}(\mathbf{I} + c\mathbf{A})^{-1}(w + cy). \tag{5.10}$$

Since $(I + cA)^{-1}e = e(-c) \geq 0$ and $A(I + cA)^{-1} = A(-c) \geq 0$ we arrive at the inequality $[\|y_i\|] \leq \|u_{n-1}\|(I + cA)^{-1}e + A(I + cA)^{-1}[\|w_i + cy_i\|]$. In view of (5.8), there follows $[\|y_i\|] \leq \|u_{n-1}\|(I + cA)^{-1}e + cA(I + cA)^{-1}[\|y_i\|]$, which is the same as $(I + cA)^{-1}[\|y_i\|] \leq \|u_{n-1}\|(I + cA)^{-1}e$. Multiplying the last inequality by the matrix $I + cA \geq 0$, we can conclude that

$$\|y_i\| \leq \|u_{n-1}\| \quad (1 \leq i \leq m). \tag{5.11}$$

Using (5.9a), (5.10), we obtain

$$\begin{aligned} u_n &= u_{n-1} + \mathbf{b}^T w = u_{n-1} - c\mathbf{b}^T y + \mathbf{b}^T(w + cy) \\ &= u_{n-1} - c\mathbf{b}^T\{(\mathbf{I} + c\mathbf{A})^{-1}\mathbf{e}u_{n-1} + \mathbf{A}(\mathbf{I} + c\mathbf{A})^{-1}(w + cy)\} + \mathbf{b}^T(w + cy) \\ &= \{1 - cb^T(I + cA)^{-1}e\}u_{n-1} + \mathbf{b}^T(\mathbf{I} + c\mathbf{A})^{-1}(w + cy). \end{aligned}$$

Since $\varphi(-c) \geq 0$, $b(-c) \geq 0$, and (5.8), (5.11) are valid, we see from the last expression for u_n that

$$\begin{aligned} \|u_n\| &\leq \{1 - cb^T(I + cA)^{-1}e\}\|u_{n-1}\| + b^T(I + cA)^{-1}[\|w_i + cy_i\|] \\ &\leq \{1 - cb^T(I + cA)^{-1}e\}\|u_{n-1}\| + (cb^T(I + cA)^{-1}e)\|u_{n-1}\| = \|u_{n-1}\|. \end{aligned}$$

This completes the proof of (2.9).

5.2.3. Statement (2.10) $\Longrightarrow$ statement(2.11); and statement (2.11) with $|.| = \|.\|_{TV} \Longrightarrow$ statement (2.11) with $|.| = \|.\|_1$. We start this subsection by giving Lemma 5.6. The lemma deals with a general situation where

(5.12a) $\mathbb{X}$ and $\mathbb{Y}$ are vector spaces, with seminorms $\|.\|_{\mathbb{X}}$ and $\|.\|_{\mathbb{Y}}$, respectively,

(5.12b) $S : \mathbb{X} \to \mathbb{Y}$ is a linear operator,

(5.12c) $Sx = 0$ only for $x = 0$, and

(5.12d) $\|x\|_{\mathbb{X}} = \|Sx\|_{\mathbb{Y}}$ (for all $x \in \mathbb{X}$).

LEMMA 5.6. *Assume* (5.12) *and let c be a stepsize-coefficient for monotonicity, with respect to $\mathbb{Y}$ and $\|.\|_{\mathbb{Y}}$. Then c is also a stepsize-coefficient for monotonicity, with respect to $\mathbb{X}$ and $\|.\|_{\mathbb{X}}$.*

Proof. Let Δt be a stepsize with $0 < \Delta t \leq c \cdot \tau_0$, and let $F : \mathbb{X} \to \mathbb{X}$ with

$$\|x + \tau_0 F(x)\|_{\mathbb{X}} \leq \|x\|_{\mathbb{X}} \quad (\text{on } \mathbb{X}). \tag{5.13a}$$

Suppose the relations (2.2) are fulfilled. We have to prove that

$$\|u_n\|_{\mathbb{X}} \leq \|u_{n-1}\|_{\mathbb{X}}. \tag{5.13b}$$

We define the subspace $\mathbb{Y}_0 = \{y : \ y = Sx \text{ for some } x \in \mathbb{X}\}$ and we introduce a linear transformation T, from $\mathbb{Y}_0$ onto $\mathbb{X}$, by $Ty = x$ (for $y = Sx \in \mathbb{Y}_0$).

In view of (2.2), the vector $v_n = Su_n$ is generated from $v_{n-1} = Su_{n-1}$ by applying the Runge–Kutta method to the function $G_0 : \mathbb{Y}_0 \to \mathbb{Y}_0$, defined by $G_0(y) = SFT(y)$ (for $y \in \mathbb{Y}_0$). Using (5.12d) and (5.13a), one easily sees that $\|y + \tau_0 G_0(y)\|_{\mathbb{Y}} \le \|y\|_{\mathbb{Y}}$ (for all $y \in \mathbb{Y}_0$).

We define $G : \mathbb{Y} \to \mathbb{Y}$ by $G(y) = G_0(y)$ (for $y \in \mathbb{Y}_0$) and $G(y) = 0$ (for $y \in \mathbb{Y} \setminus \mathbb{Y}_0$). Clearly $\|y + \tau_0 G(y)\|_{\mathbb{Y}} \le \|y\|_{\mathbb{Y}}$ (for all $y \in \mathbb{Y}$). Moreover, the vector v_n can be viewed as being generated from v_{n-1} by applying the Runge–Kutta method, with stepsize Δt, to the function G. Consequently, $\|v_n\|_{\mathbb{Y}} \le \|v_{n-1}\|_{\mathbb{Y}}$. Combining this inequality and (5.12d), we arrive at (5.13b). □

Now assume (2.10). We shall prove (2.11) by applying Lemma 5.6.

We define $\mathbb{X} = \mathbb{R}^s$, $\mathbb{Y} = \{y : \ y \in \mathbb{R}^\infty$, and $\left|y\right| < \infty\}$, and $\|x\|_{\mathbb{X}} = \left|x\right|$, $\|y\|_{\mathbb{Y}} = \left|y\right|$ (for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, respectively). Furthermore, we introduce the operator S by

$$Sx = \begin{cases} (\dots, 0, 0, x_1, x_2, .., x_s, 0, 0 \dots) & \text{if } \ \left|.\right| = \|.\|_\infty \text{ or } \|.\|_1, \\ (\dots, x_1, x_1, x_1, x_2, \dots, x_s, x_s, x_s \dots) & \text{if } \ \left|.\right| = \|.\|_{TV} \end{cases}$$

for $x = (x_1, x_2, \dots, x_s) \in \mathbb{X}$.

With these definitions, the conditions (5.12) are fulfilled. In view of (2.10), we can apply Lemma 5.6 so as to conclude that (2.11) holds.

Finally assume (2.11) with $\left|.\right| = \|.\|_{TV}$. Let $s \ge 1$ and $\mathbb{X} = \mathbb{R}^s$, $\|x\|_{\mathbb{X}} = \|x\|_1$ (for $x \in \mathbb{X}$). We want to prove that c is a stepsize-coefficient for monotonicity with respect to $\mathbb{X}$ and $\|.\|_{\mathbb{X}}$.

In order to be able to apply Lemma 5.6 to the situation at hand, we define $\mathbb{Y} = \mathbb{R}^{s+1}$, $\|y\|_{\mathbb{Y}} = \|y\|_{TV}$ (for $y \in \mathbb{Y}$). Furthermore, for $x = (x_1, x_2, \dots, x_s) \in \mathbb{X}$ we define $Sx = (y_1, \dots, y_{s+1})$ with $y_1 = 0$ and $y_i = x_1 + x_2 + \cdots + x_{i-1}$ (for $2 \le i \le s+1$).

One easily sees that, with the above definitions, all assumptions of Lemma 5.6 are fulfilled. Hence, c has the required property.

5.2.4. (2.11) with $\left|.\right| = \|.\|_1$ or $\|.\|_\infty \implies$ (2.8). Throughout this subsection we shall use, for $p = 1, \infty$ and $s \times s$ matrices G, the notation $\|G\|_p = \max \|Gv\|_p / \|v\|_p$, where the maximum is over all nonzero vectors v in $\mathbb{R}^s$. Furthermore, we shall denote the $s \times s$ identity matrix by I.

Let $G_1, G_2, \dots, G_m$ be given $s \times s$ matrices. We consider a linear variant of (2.2) (with $n = 1$, $u_0 \in \mathbb{V} = \mathbb{R}^s$) in which all vectors $F(y_j)$ are replaced by $G_j\, y_j$. Furthermore, we consider the following linear variant of condition (2.4): $\|I + \tau_0 G_i\|_p \le 1$ $(1 \le i \le m)$.

Choose $\Delta t = c\,\tau_0$ and write $Z_i = \Delta t G_i$. Then the above linear variants of (2.2) and (2.4), respectively, can be written in the form

$$u_1 = u_0 + \sum_{j=1}^{m} b_j Z_j y_j, \tag{5.14a}$$

$$y_i = u_0 + \sum_{j=1}^{m} a_{ij} Z_j y_j \quad (1 \le i \le m), \tag{5.14b}$$

and

$$\|cI + Z_i\|_p \le c \quad (1 \le i \le m). \tag{5.15}$$

In the following we shall focus on ordered m-tuples $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_m)$, where the Z_i are $s \times s$ matrices, such that (5.15) holds and the system of equations (5.14b) has a unique solution $y_1, y_2, \ldots, y_m$. The set consisting of all of these $\mathbf{Z}$ will be denoted by $\mathscr{D}_p(c, s)$.

For any $\mathbf{Z}$ in $\mathscr{D}_p(c, s)$, the vector u_1 in (5.14) depends uniquely and linearly on u_0; we denote the $s \times s$ matrix transforming u_0 into u_1 by $\mathbf{K}(\mathbf{Z})$. We thus have

$$u_1 = \mathbf{K}(\mathbf{Z})u_0 \text{ whenever } \mathbf{Z} \in \mathscr{D}_p(c, s) \text{ and } u_0, u_1 \in \mathbb{R}^s \text{ satisfy (5.14).} \tag{5.16}$$

The inequality

$$\|\mathbf{K}(\mathbf{Z})\|_p \leq 1 \text{ (for all } \mathbf{Z} \in \mathscr{D}_p(c, s) \text{ and } s \geq 1) \tag{5.17}$$

amounts to a monotonicity condition on process (5.14). It will be related to (2.8) and to (2.11) in Lemmas 5.7 and 5.8, respectively.

LEMMA 5.7. *Consider an arbitrary irreducible Runge–Kutta scheme* (A, b)*, and let* $p = 1$ *or* $p = \infty$*. Let* $0 < c < \infty$*, and assume condition* (5.17) *is fulfilled. Then* c *satisfies* (2.8).

Proof. In Kraaijevanger (1991) this lemma was proved (implicitly) for $p = \infty$. The proof in that paper is long and technical but is presented in a very clear way. Therefore, we do not repeat it here but note that the actual proof (given on pp. 507–508 of the paper) consists in a combination of conclusions regarding absolute monotonicity (on pp. 485–496) with Lemma 5.10 (on p. 505). The conclusions stated on pp. 485–496 are independent of the norm in $\mathbb{R}^s$, whereas Lemma 5.10 is tuned to the special norm $\|.\|_\infty$. It is not difficult to adapt the proof of the last mentioned lemma to the norm $\|.\|_1$ so as to conclude that Lemma 5.10 is verbatim valid for $\|.\|_1$ as well. As a result, the arguments in Kraaijevanger (1991, pp. 507–508) prove our Lemma 5.7 also for $p = 1$. □

A combination of the following lemma and Lemma 5.7 immediately leads to the desired implication ((2.11) with $|\!|.|\!| = \|.\|_1$ or $\|.\|_\infty \Longrightarrow$ (2.8)).

LEMMA 5.8. *Consider an arbitrary irreducible Runge–Kutta schema* (A, b)*, and let* $p = 1$ *or* $p = \infty$*. Let* $0 < c < \infty$*, and assume* (2.11) *with* $|\!|.|\!| = \|.\|_p$*. Then condition* (5.17) *is fulfilled.*

Proof. The proof will be given in three steps.

Step 1. Let

$$s \geq 1,\ u_0 \in \mathbb{R}^s,\ \mathbf{Z} = (Z_1, \ldots, Z_m) \in \mathscr{D}_p(c, s), \tag{5.18}$$

and assume that the corresponding vectors y_i, defined by (5.14b), satisfy

$$y_i \neq y_j \quad (\text{for } i \neq j). \tag{5.19}$$

We shall prove that

$$\|\mathbf{K}(\mathbf{Z})u_0\|_p \leq \|u_0\|_p. \tag{5.20}$$

Choose any $\tau_0 > 0$, and define $F : \mathbb{R}^s \to \mathbb{R}^s$ by $F(v) = (c\tau_0)^{-1} Z_i y_i$ (for $v = y_i$) and $F(v) = 0$ (for all other $v \in \mathbb{R}^s$). In view of (5.15), the function F satisfies (2.4) with $\mathbb{V} = \mathbb{R}^s$, $\|.\| = \|.\|_p$. Furthermore, we see from (5.14), (5.16) that the vector $\mathbf{K}(\mathbf{Z})u_0$ is generated from u_0 by applying the Runge–Kutta method with stepsize $\Delta t = c\tau_0$ to the function F. By virtue of (2.11) (with $|\!|.|\!| = \|.\|_p$), we conclude that (5.20) holds.

Step 2. Due to the restriction (5.19) in Step 1, the proof of (5.17) is not yet complete. Below, in Step 3, we shall get rid of this restriction by using (real) values γ_i, η_i (for $1 \le i \le m$) with the following properties:

$$0 < \gamma_i < c \quad (1 \le i \le m); \tag{5.21a}$$

$$\text{the } m \times m \text{ matrix } I + A \cdot \mathrm{diag}(\gamma_i) \text{ is invertible;} \tag{5.21b}$$

$$\eta_i = 1 - \sum_{j=1}^{m} a_{ij}\gamma_j\eta_j \quad (1 \le i \le m); \tag{5.21c}$$

$$\eta_i \ne \eta_j \quad (\text{whenever } i \ne j). \tag{5.21d}$$

In this (second) step we shall prove the existence of γ_i, η_i satisfying (5.21).

Since (A, b) is irreducible, statement (ii) (of Definition 2.2) is not true. It follows that the polynomials $p_i(t) = \sum_{j=1}^{m} a_{ij}t^j$ are different from each other. Therefore, there is a positive t_0 with $p_i(t_0) \ne p_j(t_0)$ (for all $i \ne j$). Writing $t_i = (t_0)^i$, we thus have

$$\sum_{k=1}^{m} a_{ik}t_k \ne \sum_{k=1}^{m} a_{jk}t_k \quad (\text{whenever } i \ne j).$$

Let $\gamma_i = \lambda t_i$, with $\lambda > 0$. We choose λ sufficiently small to guarantee (5.21a) and (5.21b). The corresponding values $\eta_i = \eta_i(\lambda)$, solving (5.21c), satisfy

$$\eta_i(\lambda) = 1 - \lambda \sum_{k=1}^{m} a_{ik}t_k + O(\lambda^2) \quad (\text{for } \lambda \downarrow 0).$$

Choosing λ sufficiently small, we conclude that γ_i, η_i exist satisfying (5.21).

Step 3. Assume (5.18). We shall prove (5.20).

Let y_i satisfy (5.14b), and choose any γ_i, η_i as in (5.21). We choose $\varepsilon > 0$ and define

$$u_0^* = \begin{pmatrix} u_0 \\ \varepsilon \end{pmatrix}, \qquad Z_i^* = \begin{pmatrix} Z_i & 0 \\ 0 & -\gamma_i \end{pmatrix}, \qquad y_i^* = \begin{pmatrix} y_i \\ \varepsilon\eta_i \end{pmatrix}.$$

Since $\mathbf{Z} \in \mathscr{D}_p(c, s)$ and (5.21a), (5.21b) hold, the m-tuple $\mathbf{Z}^* = (Z_1^*, Z_2^*, \ldots, Z_m^*)$ belongs to $\mathscr{D}_p(c, s+1)$. Furthermore, $y_i^* = u_0^* + \sum_{j=1}^{m} a_{ij}Z_j^*y_j^*$ $(1 \le i \le m)$ and $y_i^* \ne y_j^*$ (for $i \ne j$). Consequently, the conclusion of the above Step 1 can be applied (to $u_0^* \in \mathbb{R}^{s+1}$ and $\mathbf{Z}^* \in \mathscr{D}_p(c, s+1)$) so as to obtain $\|\mathbf{K}(\mathbf{Z}^*)u_0^*\|_p \le \|u_0^*\|_p$.

Since $\|\mathbf{K}(\mathbf{Z})u_0\|_p \le \|\mathbf{K}(\mathbf{Z}^*)u_0^*\|_p$ and $\|u_0^*\|_p \le \|u_0\|_p + \varepsilon$, we arrive at (5.20) by letting $\varepsilon \to 0$. □

Acknowledgments. The authors are most thankful to Dr. W. H. Hundsdorfer for useful discussions and information regarding the topic of this paper. Moreover, they are indebted to three anonymous referees for constructive criticism regarding an earlier version of the paper.

REFERENCES

K. Burrage and J. C. Butcher (1980), *Non-linear stability of a general class of differential equation methods*, BIT, 20, pp. 185–203.

J. C. Butcher (1987), *The Numerical Analysis of Ordinary Differential Equations*, John Wiley, Chichester, UK.

K. Dekker and J. G. Verwer (1984), *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North–Holland, Amsterdam.

S. Gottlieb and C.-W. Shu (1998), *Total-variation-diminishing Runge-Kutta schemes*, Math. Comp., 67, pp. 73–85.

S. Gottlieb, C.-W. Shu, and E. Tadmor (2001), *Strong-stability-preserving high-order time discretization methods*, SIAM Rev., 43, pp. 89–112.

E. Hairer and G. Wanner (1996), *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, 2nd revised ed., Springer-Verlag, Berlin.

A. Harten (1983), *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49, pp. 357–393.

W. Hundsdorfer, S. J. Ruuth, and R. J. Spiteri (2003), *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal., 41, pp. 605–623.

W. Hundsdorfer and J. G. Verwer (2003), *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Ser. Comput. Math. 33, Springer-Verlag, Berlin.

R. Jeltsch and O. Nevanlinna (1981), *Stability of explicit time discretizations for solving initial value problems*, Numer. Math., 37, pp. 61–91.

C. K. Kennedy, M. H. Carpenter, and R. M. Lewis (2000), *Low-storage, explicit Runge-Kutta schemes for the compressible Navier-Stokes equations*, Appl. Numer. Math, 35, pp. 177–219.

J. F. B. M Kraaijevanger (1991), *Contractivity of Runge-Kutta methods*, BIT, 31, pp. 482–528.

J. F. B. M. Kraaijevanger (1986), *Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems*, Numer. Math., 48, pp. 303–322.

D. Kröner (1997), *Numerical Schemes for Conservation Laws*, Wiley, Chichester, UK, and Teubner, Stuttgart, Germany.

C. B. Laney (1998), *Computational Gasdynamics*, Cambridge University Press, Cambridge, UK.

R. J. LeVeque (2002), *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK.

K. W. Morton (1980), *Stability of difference approximations to a diffusion-convection equation*, Internat. J. Numer. Methods Engrg., 15, pp. 677–683.

S. Ruuth and R. Spiteri (2002), *Two barriers on strong-stability-preserving time discretization methods*, J. Sci. Comput., 17, pp. 211–220.

S. O. Schönbeck (1967), *On the extension of Lipschitz maps*, Ark. Mat., 7, pp. 201–209.

C.-W. Shu (2002), *A survey of strong stability preserving high-order time discretizations*, in Collected Lectures on the Preservation of Stability Under Discretization, D. Estep and S. Tavener, eds., SIAM, Philadelphia, pp. 51–65.

C.-W. Shu and S. Osher (1988), *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77, pp. 439–471.

M. N. Spijker (1983), *Contractivity in the numerical solution of initial value problems*, Numer. Math., 42, pp. 271–290.

M. N. Spijker (1986), *Monotonicity and boundedness in implicit Runge-Kutta methods*, Numer. Math., 50, pp. 97–109.

R. Spiteri and S. Ruuth (2002), *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal., 40, pp. 469–491.

E. F. Toro (1999), *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 2nd ed., Springer-Verlag, Berlin.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1094–1109

CONVERGENCE OF BINOMIAL TREE METHODS FOR EUROPEAN/AMERICAN PATH-DEPENDENT OPTIONS*

LISHANG JIANG† AND MIN DAI‡

Abstract. The binomial tree method, first proposed by Cox, Ross, and Rubinstein [*Journal of Financial Economics,* 7 (1979), pp. 229–263], is one of the most popular approaches to pricing options. By introducing an additional path-dependent variable, such methods can be readily extended to the valuation of path-dependent options. In this paper, using numerical analysis and the notion of viscosity solutions, we present a unifying theoretical framework to show the uniform convergence of binomial tree methods for European/American path-dependent options, including arithmetic average options, geometric average options, and lookback options.

Key words. binomial tree method, European/American path-dependent options, convergence

AMS subject classifications. 90A09, 91B28, 93E20

DOI. 10.1137/S0036142902414220

1. Introduction.

Path-dependent options are options whose payoffs depend on historical values of the underlying asset over a given time period as well as its current price. Well-known examples are Asian arithmetic/geometric average options, lookback options, etc. The binomial tree method (BTM), first proposed by Cox, Ross, and Rubinstein [6], has become one of the most popular approaches to pricing vanilla options due to its simplicity and flexibility. By introducing an additional path-dependent variable at each node, BTM can be readily extended to the valuation of path-dependent options.

Many authors have shown that the prices of European vanilla options computed from BTM converge to their corresponding continuous-time model values (see [12] and references therein). Amin and Khanna [1] and Jiang and Dai [14] produce the convergence proofs for American vanilla options by using the probabilistic approach and the partial differential equation (PDE) approach, respectively. In this paper, using the PDE approach, a unifying framework is given to show uniform convergence of BTMs for both European and American path-dependent options, including Asian arithmetic/geometric average options and lookback options. The basic idea stems from the result of Barles and Souganidis [4], which essentially says that any stable, monotone, and consistent numerical scheme converges, provided that one has a strong comparison principle in the sense of viscosity solution for the limiting equation. For Asian options and lookback options, the BTMs are clearly monotone and the needed strong comparison principles can be deduced from Crandall, Ishii, and Lions [8] and Barles, Daher, and Romano [2]. Hence, in addition to showing consistency, the key point of the proof is to prove stability, namely, to obtain uniform estimates of bounds of the approximate solutions sequences computed by BTMs. We arrive at this by two steps: first, it is shown that values of lookback options (computed

*Received by the editors September 10, 2002; accepted for publication (in revised form) October 28, 2003; published electronically July 29, 2004.
http://www.siam.org/journals/sinum/42-3/41422.html
†Institute of Mathematics, Tongji University, Shanghai 200092, China (jianglsk@online.sh.cn). This author's work was supported by CNSF grant 10171078.
‡Corresponding author. LMAM and Department of Financial Mathematics, School of Mathematical Sciences, Peking University, Beijing 100871, China (mindai@math.pku.edu.cn). This author's work was partially supported by CNSF grant 10301002.

from BTMs) are the most expensive among those path-dependent options; second, by constructing a suitable auxiliary function, we give a uniform estimate of bounds for the price functions of lookback options. On the basis of the estimates, we then make use of the notion of viscosity solutions and numerical analysis to prove the uniform convergence.

Throughout this paper we only consider continuously monitored path-dependent options. Actually, all results can be generalized to the case of discrete monitoring because the key proof of boundedness follows from the fact that all prices of the options with discrete monitoring are not greater than that of the corresponding continuously monitored lookback option.

The outline for this paper is as follows. In the next section we recall algorithms of BTMs for arithmetic average, geometric average, and lookback options, respectively. Section 3 is devoted to the consistency of BTMs and PDEs in each case. In section 4 we establish the relationship of BTMs and finite difference methods. In sections 5 and 6 we compare prices of the above three path-dependent options and present bounds of solutions of BTMs. We prove the convergence of BTMs in section 7.

2. Algorithms. As is common in the risk neutral world, the underlying asset price S is assumed to follow the lognormal diffusion process

$$dS = rSdt + \sigma SdW, \tag{2.1}$$

where dW is a Wiener process and r and σ represent the interest rate and volatility, respectively. Consider a path-dependent option with the lifetime $[0, T]$ and the payoff

$$\Lambda(S, A) = \begin{cases} (S - A)^+ \text{ for floating strike call,} \\ (A - S)^+ \text{ for floating strike put,} \\ (A - X)^+ \text{ for fixed strike call,} \\ (X - A)^+ \text{ for fixed strike put,} \end{cases}$$

where A is the path-dependent variable and X is the strike price.

If N is the number of discrete time points, we have time points $t_n = n\Delta t$, $n = 0, 1, \ldots, N$, with $\Delta t = T/N$. Let $V^n(S, A)$ be the option price at time t_n with underlying asset value S and path-dependent variable A. Here we might as well assume

$$A = \begin{cases} \frac{1}{n}\Sigma_{i=1}^{n} S_{t_i} \text{ for arithmetic average,} \\ (\Pi_{i=1}^{n} S_{t_i})^{1/n} \text{ for geometric average,} \\ \max_{0\le i\le n} S_{t_i} \text{ for floating (fixed) strike lookback put (call) and } S \le A, \\ \min_{0\le i\le n} S_{t_i} \text{ for floating (fixed) strike lookback call (put) and } S \ge A. \end{cases} \tag{2.2}$$

S_{t_i} stands for the underlying asset value of such path at time t_i, $i = 0, 1, \ldots, n$ (note $S_{t_n} = S$). It is assumed that S will either jump up to Su with probability p or down to Sd with probability $1 - p$ at time t_{n+1}. Consequently, A will become either A^u or A^d, where

$$A^u = \begin{cases} \frac{nA+Su}{n+1} \text{ for arithmetic average,} \\ (A^n Su)^{1/(n+1)} \text{ for geometric average,} \\ \max(A, Su) \text{ for floating (fixed) strike lookback put (call) and } S \le A, \\ A \text{ for floating (fixed) strike lookback call (put) and } S \ge A \end{cases} \tag{2.3}$$

and

$$A^d = \begin{cases} \frac{nA+Sd}{n+1} \text{ for arithmetic average,} \\ (A^n Sd)^{1/(n+1)} \text{ for geometric average,} \\ A \text{ for floating (fixed) strike lookback put (call) and } S \le A, \\ \min(A, Sd) \text{ for floating (fixed) strike lookback call (put) and } S \ge A. \end{cases}$$

(2.4)

By no-arbitrage argument, one has for European path-dependent options

$$V^n(S,A) = e^{-r\Delta t}[pV^{n+1}(Su, A^u) + (1-p)V^{n+1}(Sd, A^d)], \tag{2.5}$$

where $p = \frac{e^{r\Delta t}-d}{u-d}$. Setting $ud = 1$ and combining with stochastic differential equation (2.1), we get

$$u = e^{\sigma\sqrt{\Delta t}},\ d = e^{-\sigma\sqrt{\Delta t}}$$

and thus

$$p = \frac{e^{r\Delta t} - e^{\sigma\sqrt{\Delta t}}}{e^{\sigma\sqrt{\Delta t}} - e^{-\sigma\sqrt{\Delta t}}}.$$

At expiration time $T = N\Delta t$, we have

$$V^N(S,A) = \Lambda(S,A). \tag{2.6}$$

Using the backward induction (2.5)–(2.6), option prices can be calculated. This is the so-called binomial tree model.

For American path-dependent options, (2.5) is replaced by

$$V^n(S,A) = \max\{e^{-r\Delta t}[pV^{n+1}(Su, A^u) + (1-p)V^{n+1}(Sd, A^d)], \Lambda(S,A)\}. \tag{2.7}$$

3. Consistency. For the continuous model, the path-dependent variable is given as follows:

$$A_t = \begin{cases} \frac{1}{t}\int_0^t S(\tau)d\tau \text{ for arithmetic average,} \\ \exp(\frac{1}{t}\int_0^t \ln S(\tau)d\tau) \text{ for geometric average,} \\ \max_{0\le\tau\le t} S(\tau) \text{ for floating (fixed) strike lookback put (call),} \\ \min_{0\le\tau\le t} S(\tau) \text{ for floating (fixed) strike lookback call (put).} \end{cases}$$

Let $V(S,A,t)$ be the path-dependent option value. Note that S, A, and t are mutually independent from the view point of PDEs. The pricing model of European path-dependent options is (see Kwok [16] or Wilmott, Dewynne, and Howison [17])

$$\frac{\partial V}{\partial t} + \mathcal{L}V = 0, \qquad t \in (0,T),\ (S,A) \in \mathcal{D}, \tag{3.1}$$

with the final value condition

$$V(S,A,T) = \Lambda(S,A), \tag{3.2}$$

where

$$\mathcal{L}V = \begin{cases} \frac{1}{t}(S-A)\frac{\partial V}{\partial A} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV \text{ for arithmetic average,} \\ \frac{A}{t}(\ln S - \ln A)\frac{\partial V}{\partial A} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV \text{ for geometric average,} \\ \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV \text{ for lookback} \end{cases}$$

and

$$\mathcal{D} = \begin{cases} (0,\infty)\times(0,\infty) \text{ for arithmetic or geometric average,} \\ \{(S,A): 0<S<A<\infty\} \text{ for floating (fixed) strike lookback put (call),} \\ \{(S,A): 0<A<S<\infty\} \text{ for floating (fixed) strike lookback call (put).} \end{cases}$$

In addition, for lookback options, one has an additional boundary condition

$$\frac{\partial V}{\partial A}(S,S,t) = 0. \tag{3.3}$$

REMARK 1. *Note that $\mathcal{L}V$ is not well defined at $t = 0$ for Asian options. To remove the singularity, we can take the transformation*

$$I = \begin{cases} tA \text{ for arithmetic average,} \\ t\ln A \text{ for geometric average} \end{cases} \tag{3.4}$$

to get

$$\mathcal{L}V = \begin{cases} S\frac{\partial V}{\partial I} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV \text{ for arithmetic average,} \\ \ln S\frac{\partial V}{\partial I} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV \text{ for geometric average,} \end{cases} \tag{3.5}$$

where $I \in (0,\infty)$ for the arithmetic average and $I \in (-\infty,\infty)$ for the geometric average.

REMARK 2. *One does not need to give boundary conditions at $S = 0$ that reduce to $x = -\infty$ by the transformation $x = \ln S$. Similarly, noting* (3.5) *and the directions of the characteristic lines, we do not impose boundary conditions at $A = 0$ (i.e., $I = 0$ or $-\infty$) for Asian options. We always assume that option values do not grow too fast at $S = \infty$ and $A = \infty$.*

For American options, (3.1) is replaced by a variational inequality

$$\min\left\{-\frac{\partial V}{\partial t} - \mathcal{L}V, V - \Lambda\right\} = 0, \qquad t \in (0,T),\ (S,A) \in \mathcal{D}. \tag{3.6}$$

with the final condition (3.2) (and boundary condition (3.3) for lookback options).

REMARK 3. *For American Asian options, even if the transformation* (3.4) *is employed, one cannot remove the singularity of* (3.6) *at $t = 0$ because $\Lambda = (S-A)^+ = (S - \frac{I}{t})^+$ (floating strike call, for example) is, as of yet, not well defined at $t = 0$. The financial background gives $S = A$ (i.e. $I = 0$) at $t = 0$. However, S and A are mutually independent variables in* (3.6) *and the behavior of the solution at the point $(S,S,0)$ remains to be studied further. Throughout this paper we always confine ourselves to the interval $(0,T]$ instead of $[0,T]$, except for special claim options.*

In what follows, we will show the consistency of binomial tree methods and PDEs.

THEOREM 3.1. *The binomial tree methods* (2.5) *(resp.,* (2.7)*) are consistent with the corresponding PDE* (3.1) *(resp.,* (3.6)*).*

Proof. We only take the European type arithmetic average option as an example since it is similar for other cases. We need to show that for sufficiently smooth function $\phi(S,A,t)$ and $(S_0,A_0,t_0) \in \mathcal{D}\times(0,T)$,

$$\lim_{\substack{\Delta t\to 0 \\ (S,A,t)\to(S_0,A_0,t_0)}} \frac{1}{\Delta t}(\phi - F_{\Delta t}\phi)(S,A,t) = -\frac{\partial V}{\partial t} - \mathcal{L}V\Big|_{(S_0,A_0,t_0)},$$

where

$$F_{\Delta t}\phi(S,A,t) = e^{-r\Delta t}[p\phi(Su, A^u, t) + (1-p)\phi(Sd, A^d, t)], \tag{3.7}$$

$$A^u = \frac{(t-\Delta t)A + Su\Delta t}{t} \text{ and } A^d = \frac{(t-\Delta t)A + Sd\Delta t}{t}.$$

By Taylor expansions and the identities

$$e^{-r\Delta t}[p(u-1) + (1-p)(d-1)] = r\Delta t + O(\Delta t^2),$$
$$e^{-r\Delta t}[p(u-1)^2 + (1-p)(d-1)^2] = \sigma^2\Delta t + O\left(\Delta t^2\right),$$
$$e^{-r\Delta t}[p(u-1)^3 + (1-p)(d-1)^3] = O\left(\Delta t^2\right),$$

(3.7) reduces to

$$\begin{aligned}
&(\phi - F_{\Delta t}\phi)(S,A,t)\\
&= -\left[\frac{\partial\phi}{\partial t}(S,A,t) + rS\frac{\partial\phi}{\partial S}(S,A,t) + \frac{1}{2}\sigma^2S^2\frac{\partial^2\phi}{\partial S^2}(S,A,t) - r\phi(S,A,t)\right]\Delta t\\
&\quad -e^{-r\Delta t}[p(A^u - A) + (1-p)(A^d - A)]\frac{\partial\phi}{\partial A}(S,A,t)\\
&\quad -e^{-r\Delta t}[p(u-1)(A^u - A) + (1-p)(d-1)(A^d - A)]S\frac{\partial^2\phi}{\partial S\partial A}(S,A,t)\\
&\quad +O(\Delta t^2) + O((A^u - A)\Delta t) + O((A^d - A)\Delta t) + O((A^u - A)^2) + O((A^d - A)^2).
\end{aligned} \tag{3.8}$$

Noting that $A^u - A = \frac{Su-A}{t}\Delta t$ and $A^d - A = \frac{Sd-A}{t}\Delta t$, we have

$$e^{-r\Delta t}[p(A^u - A) + (1-p)(A^d - A)] = \frac{S-A}{t}\Delta t + O(\Delta t^2),$$
$$e^{-r\Delta t}[\,p(u-1)(A^u - A) + (1-p)(d-1)(A^d - A)] = O(\Delta t^2).$$

Then we get

$$\begin{aligned}
\frac{1}{\Delta t}(\phi - F_{\Delta t}\phi)(S,A,t) = &-\left.\frac{\partial\phi}{\partial t} - \frac{1}{t}(S-A)\frac{\partial\phi}{\partial A} - \frac{1}{2}\sigma^2S^2\frac{\partial^2\phi}{\partial S^2} - rS\frac{\partial\phi}{\partial S} + r\phi\right|_{(S,A,t)}\\
&+ O(\Delta t).
\end{aligned} \tag{3.9}$$

The proof is complete.

REMARK 4. *For lookback options, the consistency of BTMs and the boundary condition* (3.3) *in the sense of the viscosity solution will be shown implicitly in the convergence proof of section* 7.

4. Relationship between BTM and finite difference method. It has been pointed out by many authors that, for vanilla options, the BTM is equivalent to certain explicit difference schemes. In this section we establish the relationship between BTMs and finite difference methods for path-dependent options.

To illustrate the basic idea, we confine ourselves to European arithmetic average options. The governing equation is

$$\frac{\partial V}{\partial t} + \frac{1}{t}(S-A)\frac{\partial V}{\partial A} + \frac{1}{2}\sigma^2S^2\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0.$$

Consider the characteristic line of $\frac{\partial V}{\partial t} + \frac{1}{t}(S-A)\frac{\partial V}{\partial A} = 0$ in $[t_n, t_{n+1}]$

$$\begin{cases} \frac{dt}{1} = \frac{dA}{\frac{1}{t}(S-A)}, & t_n \le t \le t_{n+1}, \\ A(t_n) = A_n, \end{cases}$$

whose solution is

$$A(t) = S - \frac{t_n}{t}(S - A_n).$$

The governing equation is thereby rewritten as

$$\frac{dV}{dt}\left(S, S - \frac{t_n}{t}(S-A_n), t\right) + \left(\frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV\right)\Big|_{A=S+\frac{t_n}{t}(S-A_n)} = 0, \quad t_n \le t \le t_{n+1}.$$

By adding the following three small terms to the above equation at $(S, S + \frac{t_n}{t}(S - A_n), t)$,

$$\frac{\sigma^2}{2}\left[\frac{t-t_n}{t}S\right]^2\frac{\partial^2 V}{\partial A^2} + \sigma^2 S\frac{t-t_n}{t}S\frac{\partial^2 V}{\partial A \partial S} + \left(r - \frac{\sigma^2}{2}\right)S\frac{t-t_n}{t}\frac{\partial V}{\partial A} \quad (t_n \le t \le t_{n+1}),$$

we have

$$\begin{aligned} &\frac{d}{dt}V\left(S, S - \frac{t_n}{t}(S-A_n), t\right) + \frac{\sigma^2}{2}S\frac{d}{dS}\left(S\frac{d}{dS}V\left(S, S - \frac{t_n}{t}(S-A_n), t\right)\right) \\ &+ \left(r - \frac{\sigma^2}{2}\right)S\frac{d}{dS}V\left(S, S - \frac{t_n}{t}(S-A_n), t\right) - rV\left(S, S - \frac{t_n}{t}(S-A_n), t\right) = 0, \\ &\qquad t_n \le t \le t_{n+1}. \end{aligned} \tag{4.1}$$

Noting that $\frac{d}{dS}$ is a total differential operator, (4.1) can be regarded as a Black–Scholes equation in $[t_n, t_{n+1}]$. By taking the explicit difference scheme for (4.1), we can get

$$V(S, A_n, t_n) = \frac{1}{1 + r\Delta t}[aV(Su, A_n^u, t_{n+1}) + (1-a)V(Sd, A_n^d, t_{n+1})],$$

where

$$a = \frac{1}{2} + \frac{\sqrt{\Delta t}}{2\sigma}\left(r - \frac{\sigma^2}{2}\right).$$

Since $e^{r\Delta t} = 1 + r\Delta t + O(\Delta t^2)$ and

$$p = a + O(\Delta t^{3/2}),$$

we conclude that by neglecting a high order of Δt, BTM is equivalent to the above explicit difference scheme with method of characteristic line.

REMARK 5. *For geometric average options and lookback options, we have similar results.*

5. Comparison of path-dependent options prices. In this section we will compare prices of arithmetic average options, geometric average options, and lookback options computed from the binomial tree approximation (2.6)–(2.7). To illustrate this method, we will consider the American floating strike put option and the fixed strike call option.

For Δt given and $0 \le n \le N = T/\Delta t$, we can compute $V^n(S, A)$ for all $(S, A) \in D$ by (2.6)–(2.7). In the following, $V^n(S, A)$ is regarded as a function defined by (2.6)–(2.7) in $\mathcal{D}$. In addition, we always suppose

$$0 < p < 1, \tag{5.1}$$

which is a fact for sufficiently small Δt. Under the assumption (5.1), BTMs are monotone schemes.

LEMMA 5.1. *Let $V^n(S, A)$ be the function defined by (2.6)–(2.7) in $\mathcal{D}$ for the American floating strike put option (or fixed strike call option) with payoff $(A - S)^+$ (or $(A - X)^+$). If $A_1 \le A_2$, then*

$$V^n(S, A_1) \le V^n(S, A_2)$$

for all $0 \le n \le N$.

Proof. The proof is obvious.

LEMMA 5.2. *Let $V_G^n(S, A)$, $V_A^n(S, A)$, and $V_L^n(S, A)$ be the functions defined by (2.6)–(2.7) in $\mathcal{D}$ for American floating strike geometric average, arithmetic average, and lookback put options (or corresponding fixed strike call options) with payoffs $(A - S)^+$ (or $(A - X)^+$).*

(1) *For all $0 \le n \le N$, we have*

$$V_G^n(S, A) \le V_A^n(S, A) \le V_L^n(S, \max(S, A)). \tag{5.2}$$

(2) *Let A_g, A_a, and A_l be values of the path-dependent variable for an identical path. Then for all $0 \le n \le N$*

$$V_G^n(S, A_g) \le V_A^n(S, A_a) \le V_L^n(S, \max(S, A_l)). \tag{5.3}$$

Proof. We take floating strike put options for example.

(1) Suppose (5.2) is true for $n + 1$:

$$\begin{aligned} V_A^n(S, A) &= \max\{e^{-r\Delta t}[pV_A^{n+1}(Su, A_A^u) + (1 - p)V_A^{n+1}(Sd, A_A^d)], (A - S)^+\} \\ &\ge \max\{e^{-r\Delta t}[pV_G^{n+1}(Su, A_A^u) + (1 - p)V_G^{n+1}(Sd, A_A^d)], (A - S)^+\}. \end{aligned}$$

Here

$$A_A^u = \frac{nA + Su}{n + 1} \ge (A^n Su)^{1/(n+1)} = A_G^u,$$

and similarly

$$A_A^d \ge A_G^d;$$

it follows from Lemma 5.1 that

$$\begin{aligned} V_A^n(S, A) &\ge \max\{e^{-r\Delta t}[pV_G^{n+1}(Su, A_G^u) + (1 - p)V_G^{n+1}(Sd, A_G^d)], (A - S)^+\} \\ &= V_G^{n+1}(S, A), \end{aligned}$$

which is the desired result. Combining with Lemma 5.1 and noticing that $V_L^n(S,A)$ is defined for $S \le A$, the right inequality follows similarly.

(2) For an identical path, by the definition (2.2), one has

$$A_g \le A_a \le A_l,$$

which yields (5.3) due to (5.2) and Lemma 5.1.

REMARK 6. *Lemmas* 5.1 and 5.2 *remain valid for European path-dependent options. Similar results also hold for floating strike call and fixed strike put options.*

6. Boundedness. In this section we will present bounds of solutions of BTMs, which is crucial in the proof of convergence.

LEMMA 6.1. *Let $V^n(S,A)$ be the function defined by* (2.6)–(2.7) *in $\mathcal{D}$ for an American fixed strike put option with payoff $(X-A)^+$. Then*

$$V^n(S,A) \le X$$

for all $0 \le n \le N$.

Proof. By induction, the result is obvious.

LEMMA 6.2. *Let $V^n(S,A)$ be the function defined by* (2.6)–(2.7) *in $\mathcal{D}$ for an American floating strike call option with payoff $(S-A)^+$. Then*

$$V^n(S,A) \le S$$

for all $0 \le n \le N$.

Proof. Let $V^n(S,A) = S\overline{V}^n(S,A)$ for all $0 \le n \le N$. It suffices to show that for all $0 \le n \le N$

$$\overline{V}^n(S,A) \le 1. \tag{6.1}$$

Clearly for $n < N$,

$$\overline{V}^n(S,A) = \max\left\{e^{-r\Delta t}[pu\overline{V}^{n+1}(Su,A^u) + (1-p)d\overline{V}^{n+1}(Sd,A^d)], \left(1-\frac{A}{S}\right)^+\right\}.$$

Since $\overline{V}^N(S,A) = (1-\frac{A}{S})^+ \le 1$, one might as well assume that (6.1) holds for $n+1$ and hence

$$\begin{aligned}\overline{V}^n(S,A) &\le \max\left\{e^{-r\Delta t}[pu+(1-p)d], \left(1-\frac{A}{S}\right)^+\right\}\\ &= \max\left\{1, \left(1-\frac{A}{S}\right)^+\right\} \le 1,\end{aligned}$$

which arrives at the conclusion.

LEMMA 6.3. *Let $V_L^n(S,A)$ be the function defined by* (2.6)–(2.7) *in $\mathcal{D}$ for an American floating strike lookback put option (or fixed strike call option) with payoff $(A-S)^+$ (or $(A-X)^+$). Let $\alpha > 0$ and $W^n(S,A)$ be the solution to the following problem:*

$$\begin{cases} W^n(S,A) \\ \quad = \max\{e^{-r\Delta t}[pW^{n+1}(Su,\max(Su,A)) + (1-p)W^{n+1}(Sd,A)], e^{\alpha(N-n)\Delta t}A\}, & S \le A,\\ W^N(S,A) = A. \end{cases} \tag{6.2}$$

Then for all $0 \le n \le N$ *and* $S \le A$

$$V_L^n(S,A) \le W^n(S,A). \tag{6.3}$$

Proof. We take floating strike put options for example. Since $W^N(S,A) = A \ge V_L^N(S,A)$, we may suppose (6.3) holds for $n+1$. Because $\alpha > 0$,

$$\begin{aligned} W^n(S,A) &= \max\{e^{-r\Delta t}[pW^{n+1}(Su,\max(Su,A)) + (1-p)W^{n+1}(Sd,A)], e^{\alpha(N-n)\Delta t}A\} \\ &\ge \max\{e^{-r\Delta t}[pV_L^{n+1}(Su,\max(Su,A)) + (1-p)V_L^{n+1}(Sd,A)], (A-S)^+\} \\ &= V_L^{n+1}(S,A). \end{aligned}$$

The proof is complete.

7. Convergence. In this section, we will employ the notion of viscosity solutions to show the convergence of binomial tree method. Let us first recall the notion of viscosity solutions. For convenience, we use the following notations:

$$H(V,S,A,t) = \begin{cases} -\frac{\partial V}{\partial t} - \mathcal{L}V \text{ for European options,} \\ \min\{-\frac{\partial V}{\partial t} - \mathcal{L}V, V - \Lambda\} \text{ for American options,} \end{cases}$$

$$B(V,S,A,t) = \begin{cases} -\frac{\partial V}{\partial A} \text{ for floating (fixed) strike lookback put (call),} \\ \frac{\partial V}{\partial A} \text{ for floating (fixed) strike lookback call (put),} \end{cases} \tag{7.1}$$

and

$$\overline{\mathcal{D}} = \mathcal{D} \cup \partial\mathcal{D},\ \partial\mathcal{D} = \begin{cases} \emptyset \text{ for Asian options,} \\ \{(S,A) : 0 < S = A < \infty\} \text{ for lookback options.} \end{cases}$$

REMARK 7. *In* (7.1)*, the sign before* $\frac{\partial V}{\partial A}$ *is determined by the outward unit normal to* $\partial\mathcal{D} \times (0,T)$ *(see* [8]*).*

DEFINITION 7.1. *A function* $V \in USC(\overline{\mathcal{D}} \times (0,T])$ *(resp.,* $LSC(\overline{\mathcal{D}} \times (0,T])$*) is a viscosity subsolution (resp., supersolution) of the problem* (3.6)*,* (3.2) *(and* (3.3) *for lookback options) if* $V(S,A,T) \le \Lambda(x)$ *(resp.,* $V(S,A,T) \ge \Lambda(x)$*), and whenever* $\phi \in C^{2,1}(\overline{\mathcal{D}} \times (0,T))$*,* $V - \phi$ *attains its local maximum (resp., local minimum) at* $(S_0,A_0,t_0) \in \overline{\mathcal{D}} \times (0,T)$ *and* $(V-\phi)(S_0,A_0,t_0) = 0$*, we have*

$$H(\phi,S_0,A_0,t_0) \le 0 \text{ for } (S_0,A_0,t_0) \in \mathcal{D} \times (0,T)$$

(resp.,

$$H(\phi,S_0,A_0,t_0) \ge 0 \text{ for } (S_0,A_0,t_0) \in \mathcal{D} \times (0,T)),$$

and (only for lookback option)

$$\min\{H(\phi,S_0,A_0,t_0), B(\phi,S_0,A_0,t_0)\} \le 0 \text{ for } (S_0,A_0,t_0) \in \partial\mathcal{D} \times (0,T)$$

(resp.,

$$\max\{H(\phi,S_0,A_0,t_0), B(\phi,S_0,A_0,t_0)\} \ge 0 \text{ for } (S_0,A_0,t_0) \in \partial\mathcal{D} \times (0,T)).$$

We call $V \in C(\overline{\mathcal{D}} \times (0,T])$ *a viscosity solution of* (3.6)*,* (3.2) *(and* (3.3) *for lookback options) if it is both a viscosity subsolution and a supersolution.*

The convergence proof needs the strong comparison principle that holds for Asian options (see Remark 1 and [8], [9], and [11] and references therein). For lookback options where the oblique derivative boundary condition is involved, Barles, Daher, and Romano have shown that the strong comparison principle still remains valid (see [2] and [3]). Then we get the following.

LEMMA 7.2. *The strong comparison principle holds for problem* (3.6), (3.2) *(and* (3.3) *for lookback options); namely, if u and v are the viscosity subsolution and supersolution of the problem, respectively, then $u \le v$.*

Let $V^n(S,A)$ be the function defined by (2.6)–(2.7) in $\mathcal{D}$ for American path-dependent option. We now define the extension function $V_{\Delta t}(S,A,t)$ as follows: for $t \in [n\Delta t, (n+1)\Delta t]$, $n = 0, 1, \ldots, N-1$,

$$V_{\Delta t}(S,A,t) = \frac{(n+1)\Delta t - t}{\Delta t} V^n(S,A) + \frac{t - n\Delta t}{\Delta t} V^{n+1}(S,A).$$

THEOREM 7.3. *Suppose that $V(S,A,t)$ is the viscosity solution to the problem* (3.6), (3.2) *(and* (3.3) *for lookback options). Then, as $\Delta t \to 0$, we have $V_{\Delta t}(S,A,t)$ converges uniformly to $V(S,A,t)$ in any bounded closed subdomain of $\mathcal{D} \times (0,T)$.*

In order to prove this theorem, we have to show $V^*(S,A,t)$ and $V_*(S,A,t)$ are well defined at first, where

$$V^*(S,A,t) = \limsup_{\Delta t \to 0, (x,y,z) \to (S,A,t)} V_{\Delta t}(x,y,z),$$
$$V_*(S,A,t) = \liminf_{\Delta t \to 0, (x,y,z) \to (S,A,t)} V_{\Delta t}(x,y,z).$$

In fact, due to Lemmas 6.1 and 6.2, it is true for fixed strike put options and floating strike call options. As for fixed strike call options and floating strike put options, by Lemma 6.3, it suffices to show the following.

LEMMA 7.4. *Let $W^n(S,A)$ be the solution to* (6.2) *with $\alpha > 0$. Then we have*

$$W^n(S,A) \le e^{\alpha T} \left(\max\left(A, \left(\frac{\lambda_-(\lambda_+ - 1)}{\lambda_+(\lambda_- - 1)} \right)^{1/(\lambda_- - \lambda_+)} S \right) + 1 \right) \tag{7.2}$$

for sufficiently small Δt, where

$$\lambda_\pm = \frac{r}{\sigma^2} + \frac{1}{2} \pm \sqrt{\left(\frac{r}{\sigma^2} + \frac{1}{2} \right)^2 + \frac{2\alpha}{\sigma^2}}. \tag{7.3}$$

REMARK 8. *In Lemma* 7.4, *$\alpha > 0$ guarantees* $(\frac{\lambda_-(\lambda_+ - 1)}{\lambda_+(\lambda_- - 1)})^{1/(\lambda_- - \lambda_+)} < \infty$.

Before proving Lemma 7.4 we inquire into some properties of the solution to the problem (6.2). By transformations

$$x = \ln \frac{A}{S} \text{ and } \overline{W}^n(x) = e^{-\alpha(N-n)\Delta t} \frac{W^n(S,A)}{S}, \tag{7.4}$$

the numerical scheme (6.2) is reduced to

$$\begin{cases} \overline{W}^n(x) = \max\{e^{-(r+\alpha)\Delta t}[p u\overline{W}^{n+1}((x - \sigma\sqrt{\Delta t})^+) + (1-p) d\overline{W}^{n+1}(x + \sigma\sqrt{\Delta t})], e^x\}, \\ \qquad x \ge 0, \\ \overline{W}^N(x) = e^x. \end{cases} \tag{7.5}$$

LEMMA 7.5. *Let* $\overline{W}^n(x)$ *be the solution to* (7.5). *Then we have*
(a) $\overline{W}^{n+1}(x) \le \overline{W}^n(x)$,
(b) $\overline{W}^n(x_1) \le \overline{W}^n(x_2)$ *if* $x_1 \le x_2$,
(c) *for each* $n \le N$,

$$\overline{W}^n(x) = e^x \text{ if } x \ge (N-n)\sigma\sqrt{\Delta t}. \tag{7.6}$$

Proof. (a) and (b) are obvious. In order to prove (c), we use induction. Suppose (7.6) holds for $n = k+1$, namely $\overline{W}^{k+1}(x) = e^x$ for $x \ge (N-k-1)\sigma\sqrt{\Delta t}$. If $x \ge (N-k)\sigma\sqrt{\Delta t}$, then

$$\begin{aligned}\overline{W}^k(x) &= \max\{e^{-(r+\alpha)\Delta t}[pu\overline{W}^{k+1}((x-\sigma\sqrt{\Delta t})^+) + (1-p)d\overline{W}^{k+1}(x+\sigma\sqrt{\Delta t})], e^x\}\\ &= \max\{e^{-(r+\alpha)\Delta t}[pue^{x-\sigma\sqrt{\Delta t}} + (1-p)de^{x+\sigma\sqrt{\Delta t}}], e^x\}\\ &= \max\{e^{-(r+\alpha)\Delta t}e^x, e^x\} = e^x,\end{aligned}$$

which is the desired result.

To simplify notation, (7.5) will also be written as

$$\overline{W}^n(x) = F(\Delta t)\overline{W}^{n+1}(x). \tag{7.7}$$

LEMMA 7.6. *For* Δt *given, there exists unique element* $\overline{W}_{\Delta t}(x)$ *satisfying* $\overline{W}_{\Delta t}(x) - e^x \in L^\infty(R^+)$ *such that*

$$\overline{W}_{\Delta t}(x) = F(\Delta t)\overline{W}_{\Delta t}(x). \tag{7.8}$$

In addition, $\overline{W}_{\Delta t}(x)$ *is a monotone function of* x *and*

$$\overline{W}^n(x) \le \overline{W}_{\Delta t}(x). \tag{7.9}$$

Proof. Let $\widetilde{W}^n(x) = \overline{W}^n(x) - e^x$. Then $\widetilde{W}^n(x)$ satisfies

$$\widetilde{W}^n(x) = F(\Delta t)(\widetilde{W}^{n+1}(x) + e^x) - e^x \widehat{=} G(\Delta t)\widetilde{W}^{n+1}(x).$$

By (7.6), $\widetilde{W}^n(x) \in L^\infty(R^+)$. Hence $G(\Delta t)$ can be regarded as a mapping from $L^\infty(R^+)$ to $L^\infty(R^+)$. Next we will show $G(\Delta t)$ is a contraction mapping. Let $U(x)$, $V(x) \in L^\infty(R^+)$. Then

$$\begin{aligned}&\|G(\Delta t)U(x) - G(\Delta t)V(x)\|_\infty\\ &= \|F(\Delta t)(U(x)+e^x) - F(\Delta t)(V(x)+e^x)\|_\infty\\ &\le e^{-(r+\alpha)\Delta t}[pu + (1-p)d]\,\|U(x) - V(x)\|_\infty\\ &= e^{-\alpha\Delta t}\,\|U(x) - V(x)\|_\infty.\end{aligned}$$

Therefore, there exists a unique element $\widetilde{W}_{\Delta t}(x) \in L^\infty(R^+)$ such that $\widetilde{W}_{\Delta t}(x) = G(\Delta t)\widetilde{W}_{\Delta t}(x)$. Owing to Lemma 7.5, $\widetilde{W}_{\Delta t}(x)$ is a monotone function of x and

$$\widetilde{W}^n(x) \le \widetilde{W}_{\Delta t}(x).$$

This completes the proof by denoting $\overline{W}_{\Delta t}(x) = \widetilde{W}_{\Delta t}(x) + e^x$.

Proof of Lemma 7.4. The idea of the proof stems from Dai [10]. Let $\Delta x = \sigma\sqrt{\Delta t}$, $x_j = j\Delta x$, and $u^j = e^{x_j}$, $j = 0, 1, \ldots$. It is not hard to see that $\overline{W}_{\Delta t}(x_j)$ satisfies

$$\begin{cases} \overline{W}_{\Delta t}(x_j) = \max\{e^{-(r+\alpha)\Delta t}[pu\overline{W}_{\Delta t}(x_{j-1}) + (1-p)d\overline{W}_{\Delta t}(x_{j+1})], u^j\}, \ j \geq 1, \\ \overline{W}_{\Delta t}(x_0) = e^{-(r+\alpha)\Delta t}[pu\overline{W}_{\Delta t}(x_0) + (1-p)d\overline{W}_{\Delta t}(x_1)], \end{cases}$$

which is equivalent to a free boundary problem of a difference equation as follows:

$$\overline{W}_{\Delta t}(x_j) = e^{-(r+\alpha)\Delta t}[pu\overline{W}_{\Delta t}(x_{j-1}) + (1-p)d\overline{W}_{\Delta t}(x_{j+1})] \text{ for } 1 \leq j < j_\infty,$$

$$\overline{W}_{\Delta t}(x_0) = e^{-(r+\alpha)\Delta t}[pu\overline{W}_{\Delta t}(x_0) + (1-p)d\overline{W}_{\Delta t}(x_1)], \tag{7.10}$$

$$\overline{W}_{\Delta t}(x_{j_\infty}) = u^{j_\infty}, \qquad \overline{W}_{\Delta t}(x_{j_\infty+1}) = u^{j_\infty+1}. \tag{7.11}$$

Here j_∞ is the point of free boundary to be determined. We claim

$$\overline{W}_{\Delta t}(x_j) = C_1\xi_1^j + C_2\xi_2^j \text{ for } 0 \leq j \leq j_\infty + 1, \tag{7.12}$$

where ξ_1, ξ_2 are two real roots of the equation $\xi = e^{-(r+\alpha)\Delta t}(pu + (1-p)d\xi^2)$, namely,

$$\xi_{1,2} = \frac{e^{(r+\alpha)\Delta t} \pm \sqrt{e^{2(r+\alpha)\Delta t} - 4p(1-p)}}{2(1-p)d}. \tag{7.13}$$

To determine constants C_1, C_2, and j_∞, we make use of boundary condition (7.10) and free boundary condition (7.11); we have

$$\frac{C_1}{C_2} = \frac{(e^{(r+\alpha)\Delta t} - pu) - (1-p)d\xi_2}{(1-p)d\xi_1 - (e^{(r+\alpha)\Delta t} - pu)}, \tag{7.14}$$

$$\overline{W}_{\Delta t}(x_{j_\infty}) = C_1\xi_1^{j_\infty} + C_2\xi_2^{j_\infty} = u^{j_\infty}, \tag{7.15}$$

$$\overline{W}_{\Delta t}(x_{j_\infty+1}) = C_1\xi_1^{j_\infty+1} + C_2\xi_2^{j_\infty+1} = u^{j_\infty+1}. \tag{7.16}$$

By solving (7.14)–(7.16), we get

$$C_1 = \frac{\xi_2 u^{j_\infty} - u^{j_\infty+1}}{\xi_1^{j_\infty}(\xi_2 - \xi_1)}, \qquad C_2 = \frac{\xi_1 u^{j_\infty} - u^{j_\infty+1}}{\xi_2^{j_\infty}(\xi_1 - \xi_2)}, \tag{7.17}$$

and

$$j_\infty = \frac{1}{\ln\xi_2 - \ln\xi_1} \ln\left(-\frac{(e^{(r+\alpha)\Delta t} - pu) - (1-p)d\xi_2}{(1-p)d\xi_1 - (e^{(r+\alpha)\Delta t} - pu)} \frac{\xi_1 - u}{\xi_2 - u}\right) \tag{7.18}$$

Noticing that $\overline{W}_{\Delta t}(x)$ is monotone with respect to x and combining with (7.12), we have

$$\overline{W}_{\Delta t}(x) \leq \max(e^{x+\Delta x}, e^{j_\infty \Delta x}).$$

By symbol operation, one gets

$$\lim_{\Delta t \to 0} j_\infty \Delta x = \frac{1}{\lambda_- - \lambda_+} \ln\frac{\lambda_-(\lambda_+ - 1)}{\lambda_+(\lambda_- - 1)} < \infty,$$

where $\lambda_\pm$ are given by (7.3). Then for sufficiently small Δt,

$$\overline{W}_{\Delta t}(x) \leq \max\left(e^x, \left(\frac{\lambda_-(\lambda_+ - 1)}{\lambda_+(\lambda_- - 1)}\right)^{1/(\lambda_- - \lambda_+)}\right) + 1.$$

Together with (7.4) and (7.9), this implies (7.2), which completes the proof of Lemma 7.4.

We now prove Theorem 7.3. The idea is based on [4] and [14].

Proof of Theorem 7.3. Since V^* and V_* are well defined, it is obvious that $V^* \in USC$ and $V_* \in LSC$, and $V_*(S,A,t) \le V^*(S,A,t)$. If we show that V^* and V_* are the viscosity subsolution and supersolution of (3.6), respectively, then in terms of the comparison principle (Lemma 7.2), we deduce $V^*(S,A,t) \le V_*(S,A,t)$ and thus $V^*(S,A,t) = V_*(S,A,t) = V(S,A,t)$, which is the desired conclusion.

We need only to show that V^* is a subsolution of (3.6), (3.2) (and (3.3) for lookback options). It can be shown that $V^*(S,A,T) \le \Lambda(S,A)$ (see [11]). Suppose that for $\phi \in C^{2,1}(\mathcal{D} \times (0,T])$, $V^* - \phi$ attains a local maximum at $(S_0,A_0,t_0) \in \mathcal{D} \times (0,T)$ and $(V^* - \phi)(S_0,A_0,t_0) = 0$. We might as well assume that (S_0,A_0,t_0) is a strict local maximum on $B_r = \{t_0 \le t \le t_0 + r, |S - S_0| \le r, |A - A_0| \le r\}$, $r > 0$. By the definition of V^*, there exists a sequence $u_{\Delta t_k}(S_k,A_k,t_k)$ such that $\Delta t_k \to 0$, $(S_k,A_k,t_k) \to (S_0,A_0,t_0)$, $V_{\Delta t_k}(S_k,A_k,t_k) \to V^*(S_0,A_0,t_0)$ when $k \to \infty$. Assuming that $(\widehat{S}_k,\widehat{A}_k,\widehat{t}_k)$ is a global maximum point of $V_{\Delta t_k} - \phi$ on B_r, we can deduce that there is a subsequence $V_{\Delta t_{k_i}}(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i})$ such that

$$\begin{aligned} &\Delta t_{k_i} \to 0,\ (\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) \to (S_0,A_0,t_0), \\ &(V_{\Delta t_{k_i}} - \phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) \to (V^* - \phi)(S_0,A_0,t_0) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{as } k_i \to \infty. \end{aligned} \tag{7.19}$$

Indeed, suppose $(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) \to (\widehat{S},\widehat{A},\widehat{t})$; then

$$\begin{aligned} (V^* - \phi)(S_0,A_0,t_0) &= \lim_{k_i \to \infty} (V_{\Delta t_{k_i}} - \phi)(S_{k_i},A_{k_i},t_{k_i}) \\ &\le \lim_{k_i \to \infty} (V_{\Delta t_{k_i}} - \phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) \le (V^* - \phi)(\widehat{S},\widehat{A},\widehat{t}), \end{aligned}$$

which forces $(\widehat{S},\widehat{A},\widehat{t}) = (S_0,A_0,t_0)$ since (S_0,A_0,t_0) is a local strict maximum point of $V^* - \phi$. Therefore

$$(V_{\Delta t_{k_i}} - \phi)(\cdot,\cdot,\widehat{t}_{k_i} + \Delta t_{k_i}) \le (V_{\Delta t_{k_i}} - \phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) \ \ in\ B_r;$$

that is,

$$V_{\Delta t_{k_i}}(\cdot,\cdot,\widehat{t}_{k_i} + \Delta t_{k_i}) \le \phi(\cdot,\cdot,\widehat{t}_{k_i} + \Delta t_{k_i}) + (V_{\Delta t_{k_i}} - \phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) \quad \text{in } B_r.$$

Then

$$\begin{aligned} &V_{\Delta t_{k_i}}(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) \\ &= \max\{F_{\Delta t_{k_i}} V_{\Delta t_{k_i}}(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}), \Lambda(\widehat{S}_{k_i},\widehat{A}_{k_i})\} \\ &\le \max\{F_{\Delta t_{k_i}} \phi(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) + e^{-r\Delta t_{k_i}}(V_{\Delta t_{k_i}} - \phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}), \Lambda(\widehat{S}_{k_i},\widehat{A}_{k_i})\}, \end{aligned}$$

namely,

$$\begin{aligned} \min\{&(\phi - F_{\Delta t_{k_i}}\phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) + (1 - e^{-r\Delta t_{k_i}})(V_{\Delta t_{k_i}} - \phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}), \\ &V(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) - \Lambda(\widehat{S}_{k_i},\widehat{A}_{k_i})\} \le 0. \end{aligned} \tag{7.20}$$

Here the operator $F_{\Delta t_{k_i}}$ is given by (3.7). Dividing the first argument in the min by

$\Delta t e^{-r\Delta t_{k_i}}$, letting $k_i \to \infty$, and noticing that

$$\frac{1-e^{-r\Delta t_{k_i}}}{\Delta t e^{-r\Delta t_{k_i}}}(V_{\Delta t_{k_i}}-\phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) \to (V^*-\phi)(S_0,A_0,t_0)=0, \tag{7.21}$$

we get by consistency and (7.19) that

$$\min\left\{-\frac{\partial\phi}{\partial t}-\mathcal{L}\phi, V^*-\Lambda\right\}_{(S_0,A_0,t_0)} \le 0,$$

which yields the desired result because of $V^*(S_0,A_0,t_0)=\phi(S_0,A_0,t_0)$.

For lookback options (fixed strike call, for example), if $(S_0,A_0,t_0)\in\partial\mathcal{D}\times(0,T)$ and (7.19) holds, we might as well assume either $(\widehat{S}_{k_i},\widehat{A}_{k_i})\in\mathcal{D}$ for all k_i or $(\widehat{S}_{k_i},\widehat{A}_{k_i})\in\partial\mathcal{D}$ for all k_i. If it is the former, we can use the same argument as before to get $\min\{-\frac{\partial\phi}{\partial t}-\mathcal{L}\phi,\phi-\Lambda\}_{(S_0,A_0,t_0)}\le 0$. If $(\widehat{S}_{k_i},\widehat{A}_{k_i})\in\partial\mathcal{D}$, i.e., $\widehat{S}_{k_i}=\widehat{A}_{k_i}$, then

$$\widehat{A}^u_{k_i}=\widehat{S}_{k_i}u \text{ and } \widehat{A}^d_{k_i}=\widehat{S}_{k_i} \tag{7.22}$$

Using (7.22) and (3.8), we get

$$\begin{aligned}(\phi-F_{\Delta t_{k_i}}\phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}) &= \left(-\frac{\partial\phi}{\partial t}-\mathcal{L}\phi\right)\Delta t_{k_i}-\frac{\sigma\widehat{S}_{k_i}}{2}\frac{\partial\phi}{\partial A}\Delta t_{k_i}^{1/2}+O(\Delta t_{k_i}^2)\\ &= -\frac{\sigma\widehat{S}_{k_i}}{2}\frac{\partial\phi}{\partial A}\Delta t_{k_i}^{1/2}+O(\Delta t_{k_i}).\end{aligned}$$

Combining with (7.20), which can also be similarly obtained in this case, we deduce

$$\min\left\{-\frac{\sigma\widehat{S}_{k_i}}{2}\frac{\partial\phi}{\partial A}\Delta t_{k_i}^{1/2}+O(\Delta t_{k_i})+(1-e^{-r\Delta t_{k_i}})(V_{\Delta t_{k_i}}-\phi)(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i}),\right.$$
$$\left. V(\widehat{S}_{k_i},\widehat{A}_{k_i},\widehat{t}_{k_i})-\Lambda(\widehat{S}_{k_i},\widehat{A}_{k_i})\right\}\le 0.$$

Dividing the first argument in the min by $\frac{\sigma\widehat{S}_{k_i}}{2}\Delta t^{1/2}e^{-r\Delta t_{k_i}}$, letting $k_i\to\infty$, and noticing (7.21), we then get by (7.19)

$$\min\left\{-\frac{\partial\phi}{\partial A},\phi-\Lambda\right\}_{(S_0,A_0,t_0)}\le 0.$$

Hence, in either case, we have

$$\min\left\{\min\left\{-\frac{\partial\phi}{\partial t}-\mathcal{L}\phi,\phi-\Lambda\right\},-\frac{\partial\phi}{\partial A}\right\}_{(S_0,A_0,t_0)}\le 0.$$

The proof is complete.

Theorem 7.3 indicates that BTMs for American path-dependent options are locally uniformly convergent. It is clear that V^* and V_* are well defined for European path-dependent options because the prices of European options computing by BTMs are always less than those of the corresponding American options. Similar arguments

also give the convergence of BTMs for European path-dependent options. As a result, we assert the following.

THEOREM 7.7. *Binomial tree methods for European/American path-dependent options are uniformly convergent in any bounded closed domain of* $\mathcal{D} \times (0, T)$.

REMARK 9. *Clearly the convergence proof of BTMs remains valid at* $t = 0$ *for lookback options. By virtue of Remark* 1*, the convergence result at* $t = 0$ *for European Asian options is not too difficult an extension. However, for American Asian options, the convergence at* $t = 0$ *is currently not available because we cannot prove the strong comparison principle in* $\mathcal{D} \times [0, T)$ *(see also Remark* 3*).*

Due to Lemma 5.2 and Theorems 7.3 and 7.7, we have the following.

COROLLARY 7.8. *Let* $V_G(S, A, t)$*,* $V_A(S, A, t)$*, and* $V_L(S, A, t)$ *be the solutions of the continuous models for floating strike geometric average, arithmetic average, and lookback put options (or corresponding fixed strike call options); then we have*

$$V_G(S, A, t) \le V_A(S, A, t) \le V_L(S, \max(S, A), t).$$

We conclude the paper with the following remark.

REMARK 10. *It is well known that the BTM is not feasible for pricing arithmetic average options because the number of possible arithmetic average values increases exponentially with the number of timesteps. Barraquand and Pudet* [5] *and Hull and White* [13] *present modified BTMs that restrict the possible average values to a set of predetermined values. Our technique can also be applied to prove the convergence of their methods. We refer interested readers to* [15] *for details.*

REFERENCES

[1] K. AMIN AND A. KHANNA, *Convergence of American option values from discrete- to continuous-time financial models*, Math. Finance, 4 (1994), pp. 289–304.

[2] G. BARLES, CH. DAHER, AND M. ROMANO, *Optimal control of the* L^∞ *norm of a diffusion process*, SIAM J. Control Optim., 32 (1994), pp. 612–634.

[3] G. BARLES, CH. DAHER, AND M. ROMANO, *Convergence of numerical schemes for parabolic equations arising in finance theory*, Math. Models Methods Appl. Sci., 5 (1995), pp. 125–143.

[4] G. BARLES AND P.E. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 1 (1991), pp. 271–283.

[5] J. BARRAQUAND AND T. PUDET, *Pricing of American path-dependent contingent claims*, Math. Finance, 6 (1996), pp. 17–51.

[6] J.C. COX, S.A. ROSS, AND M. RUBINSTEIN, *Option pricing: A simple approach*, Journal of Financial Economics, 7 (1979), pp. 229–263.

[7] T.H.F. CHEUCK AND T.C.F. VORST, *Currency lookback options and observation frequency: A binomial approach*, Journal of International Money and Finance, 16 (1997), pp. 173–187.

[8] M.G. CRANDALL, H. ISHII, AND P.L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[9] M.G. CRANDALL, P.L. LIONS, AND P.E. SOUGANIDIS, *Maximal solutions and universal bounds for some quasilinear evolution equations of parabolic type*, Arch. Ration. Mech. Anal., 105 (1989), pp. 163–190.

[10] M. DAI, *A closed form solution to perpetual American floating strike lookback option*, Journal of Computational Finance, 4 (2000), pp. 63–68.

[11] W.H. FLEMING AND H.M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, Berlin, 1993.

[12] H. HE, *Convergence from discrete- to continuous-time contingent claims prices*, Review of Financial Studies, 3 (1990), pp. 523–546.

[13] J. HULL AND A. WHITE, *Efficient procedures for valuing European and American path-dependent options*, Journal of Derivatives, 1 (1993), pp. 21–31.

[14] L. Jiang and M. Dai, *Convergence of binomial tree method for American options*, in Partial Differential Equations and Their Applications, H. Chen and L. Rodino, eds., World Scientific, River Edge, NJ, 1999, pp. 106–118.
[15] L. Jiang and M. Dai, *Convergence of the Forward Shooting Grid Method for American-Style Asian Options*, working paper, Peking University, Beijing, China, 2000.
[16] Y.K. Kwok, *Mathematical Models of Financial Derivatives*, Springer-Verlag, Singapore, 1998.
[17] P. Wilmott, J. Dewynne, and S. Howison, *Option Pricing: Mathematical Models and Computation*, Oxford Financial Press, Oxford, UK, 1993.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1110–1127

CONVERGENCE OF PETVIASHVILI'S ITERATION METHOD FOR NUMERICAL APPROXIMATION OF STATIONARY SOLUTIONS OF NONLINEAR WAVE EQUATIONS*

DMITRY E. PELINOVSKY† AND YURY A. STEPANYANTS‡

In memory of Vladimir I. Petviashvili (1936–1993)*, who made an outstanding contribution to the theory of nonlinear waves.*

Abstract. We analyze a heuristic numerical method suggested by V.I. Petviashvili in 1976 for approximation of stationary solutions of nonlinear wave equations. The method is used to construct numerically the solitary wave solutions, such as solitons, lumps, and vortices, in a space of one and higher dimensions. Assuming that the stationary solution exists, we find conditions when the iteration method converges to the stationary solution and when the rate of convergence is the fastest. The theory is illustrated with examples of physical interest such as generalized Korteweg–de Vries, Benjamin–Ono, Zakharov–Kuznetsov, Kadomtsev–Petviashvili, and Klein–Gordon equations.

Key words. nonlinear evolution equations, solitary waves, numerical approximations, iteration methods, convergence and stability, linearized operators

AMS subject classifications. 65N35, 35Q51, 35Q53, 35P20

DOI. 10.1137/S0036142902414232

1. Introduction.

Nonlinear waves and vortices are often described by partial differential equations, whose solutions cannot be found analytically even in a space of one dimension. Numerical computations are used to approximate various solutions, including stationary solutions. An effective numerical method for computing solitary wave solutions in a space of two dimensions was proposed by V. I. Petviashvili in the context of the Kadomtsev–Petviashvili equation with positive dispersion (KPI equation) [P76]. The numerical method was shown to converge to a stationary solution, but no analysis or proof was given. One year later, the very same solution was found analytically [MZ77], referred to as the two-dimensional soliton or lump. After the pioneering work [P76], Petviashili's numerical method was applied to numerous nonlinear problems in modern mathematical physics [PP92].

In this paper, we prove the convergence theorem for Petviashvili's numerical method in a context of a nonlinear scalar wave equation with power nonlinearity. We assume that the stationary solution exists in a suitable function space, when the method is well defined. The method clearly diverges in the cases when no stationary solution exists in such spaces. We derive conditions on parameters of the numerical method and on the spectrum of a linearized operator associated with the stationary solution, when the method converges to the stationary solution.

We start with a nonlinear scalar wave equation with power nonlinearity in one dimension:

$$u_t - (\mathcal{L}u)_x + pu^{p-1}u_x = 0, \tag{1.1}$$

where $u : \mathbb{R} \times \mathbb{R}_+ \mapsto \mathbb{R}$, $p > 1$, and $\mathcal{L}$ is a linear self-adjoint nonnegative pseudodiffer-

*Received by the editors September 9, 2002; accepted for publication (in revised form) November 12, 2003; published electronically September 18, 2004.
http://www.siam.org/journals/sinum/42-3/41423.html

†Department of Mathematics, McMaster University, Hamilton, ON, Canada, L8S 4K1 (dmpeli@math.mcmaster.ca). The work of this author was supported by NSERC grant 5-36694.

‡Environment at ANSTO, PMB 1, Menai, NSW, 2234, Australia (ysx@ansto.gov.au).

ential operator in x with constant coefficients, such that

$$\langle u, \mathcal{L}u\rangle = \langle \mathcal{L}u, u\rangle \geq 0, \qquad \langle f, g\rangle = \int_{-\infty}^{\infty} \bar{f}(x)g(x)\,dx. \tag{1.2}$$

Stationary solutions of (1.1) are of the form $u(x,t) = \Phi(x - ct)$, where c is an eigenvalue and $\Phi(x)$ is a bound state of the boundary-value problem on $x \in \mathbb{R}$,

$$c\Phi + \mathcal{L}\Phi = \Phi^p, \tag{1.3}$$

such that $\lim_{|x|\to\infty} \Phi(x) = 0$. The paramater c, which is typically continuous, has a physical meaning of a speed of the stationary wave. The bound state $\Phi(x)$ belongs to the function space $X(\mathbb{R})$, defined in Assumption 1.1.

We employ the Fourier transform,

$$u(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{u}(k)e^{ikx}dk, \quad \hat{u}(k) = \int_{-\infty}^{\infty} u(x)e^{-ikx}\,dx \tag{1.4}$$

and rewrite the boundary–value problem (1.3) in the form

$$[c + v(k)]\,\hat{\Phi}(k) = \widehat{\Phi^p}(k), \tag{1.5}$$

where $v(k)$ is the range of $\mathcal{L}$ in the Fourier space. If $\mathcal{L}$ is a nonnegative pseudodifferential operator of order m, the function $v(k)$ is an mth order polynomial of $|k|$, such that $v(k) \geq 0$. The function $v(k)$ has meaning of phase velocity of linear waves (infinitesimal perturbations) of the scalar wave equation (1.1). Resonance between nonlinear bound states and linear waves is excluded if $c + v(k) \neq 0$ for any $k \in \mathbb{R}$. For the nonnegative operator $\mathcal{L}$ with $v(k) \geq 0$, the resonance is excluded for $c > 0$.

Assumption 1.1. Let m be the order of a linear pseudodifferential operator $\mathcal{L}$, $p > 1$, $v(k) \geq 0$, and $c > 0$. There exists a real analytical solution of the boundary-value problem (1.5) in the function space

$$X = L^2(\mathbb{R}) \cap L^{p+1}(\mathbb{R}) \cap H^{m/2}(\mathbb{R}). \tag{1.6}$$

A naive iterative algorithm for numerical approximation of $\hat{\Phi}(k)$ in the problem (1.5) can be proposed in the form

$$\widehat{u}_{n+1}(k) = \frac{\widehat{u_n^p}(k)}{c + v(k)}, \tag{1.7}$$

where $\widehat{u}_n(k)$ is the Fourier transform of $u_n(x)$ and $u_n(x)$ is the nth iteration of the numerical solution. However, this algorithm usually diverges, even if a fixed point $\hat{\Phi}(k)$ exists in the nonlinear problem (1.5). A modified iterative procedure is proposed by introducing the stabilizing factor M_n [P76],

$$\widehat{u}_{n+1}(k) = M_n^{\gamma}\frac{\widehat{u_n^p}(k)}{c + v(k)}, \tag{1.8}$$

where the stabilizing factor M_n is computed as

$$M_n = M_n[\widehat{u}_n] = \frac{\int_{-\infty}^{\infty} [c + v(k)]\,[\widehat{u}_n(k)]^2\,dk}{\int_{-\infty}^{\infty} \widehat{u}_n(k)\widehat{u_n^p}(k)\,dk}, \tag{1.9}$$

and γ is a free parameter, which must be chosen for convergence of the sequence $\{u_n(x)\}_{n=0}^{\infty}$. The fixed points of the iterative map (1.8)–(1.9) are the same as the bound states $\hat{\Phi}(k)$ of the nonlinear boundary-value problem (1.5).

LEMMA 1.2. *A set of fixed points of the iteration map* (1.8)–(1.9) *coincides with a set of bound states* $\hat{\Phi}(k)$ *of the boundary-value problem* (1.5)*, provided that* $\gamma \neq 1+2n$, $n \in \mathbb{Z}$.

Proof. If $\widehat{u}_n(k) = \hat{\Phi}(k)$ is a solution of the boundary-value problem (1.5), then $M_n = 1$ from (1.9) and $\widehat{u}_{n+1}(k) = \hat{\Phi}(k)$ from (1.8). Therefore, the solution $\hat{\Phi}(k)$ is a fixed point of the iteration map (1.8)–(1.9). In the other direction, let $\widehat{u}_*(k)$ be a fixed point of the iteration map (1.8)–(1.9). Multiplying (1.8) by $[c+v(k)]\widehat{u}_*(k)$ and integrating over k, we find $M_* = M_*^{\gamma}$. When $\gamma \neq 1+2n$, $n \in \mathbb{Z}$, there exist only two solutions: $M_* = 0$ or $M_* = 1$. Since $c + v(k) > 0$ for any $k \in \mathbb{R}$, the former solution is equivalent to a trivial zero fixed point: $\widehat{u}_*(k) = 0$. The fixed point of (1.8) with $M_* = 1$ satisfies the boundary-value problem (1.5), such that $\widehat{u}_*(k) = \hat{\Phi}(k)$. □

When $\gamma = 0$, the iterative method (1.8) is the same as in (1.7) and it diverges in most cases as was mentioned above. Nevertheless, a nonempty range for γ can be found empirically, when the method converges to the bound state $\Phi(x)$, starting with $u_0 \in X(\mathbb{R})$ such that $u_n \in X(\mathbb{R})$, $\lim_{n\to\infty} u_n(x) = \Phi(x)$, and $\lim_{n\to\infty} M_n = 1$. For $p = 2$ (quadratic nonlinearity), Petviashvili has found empirically that the iteration method (1.8)–(1.9) converges for $1 < \gamma < 3$, with the fastest rate of convergence at $\gamma = 2$ [PP92]. He also noticed that the fastest rate of convergence occurs when the degree of the uniformity of the right-hand side of (1.8) is zero with respect to $\widehat{u}_n(k)$. The convergence results do not depend on the actual dependence $v(k)$, provided that $c + v(k) > 0$ [PP92].

In this paper, we prove that the iteration method (1.8)–(1.9) converges for $1 < \gamma < (p+1)/(p-1)$ under some additional assumptions on the spectrum of a linearized operator associated with the bound state $\Phi(x)$. The fastest rate of convergence occurs for $\gamma = \gamma_* = p/(p-1)$.

From a practical point, the iteration procedure can be stopped when $|M_n - 1| \leq \varepsilon$ for any given small $\varepsilon > 0$. Therefore, parameter ε defines the distance between $u_n(x)$ and $\Phi(x)$ that measures the numerical error in the sense of the integrals in (1.9). Two additional sources of numerical errors come from the use of spectral methods, such as (i) the truncation of the integration domain $k \in \mathbb{R}$ by a finite interval $k \in [-K, K]$ and (ii) the discretization of the integrals at a finite number of grid points.

The paper is organized as follows. Section 2 describes properties of the linearized operator associated with the scalar wave equation (1.1) and also formulates the main convergence theorem. Section 3 presents the proof of the convergence theorem. Convergence of the special sequences, which are self-similar to the bound states, is considered in section 4. Examples of the iteration method (1.8)–(1.9) in one and two dimensions are studied in sections 5 and 6.

2. Spectral properties of the linearized operator. Here we study properties of the linearized operator associated with the nonlinear wave equation (1.1) at $u = \Phi(x - ct)$,

$$\mathcal{H} = c + \mathcal{L} - p\Phi^{p-1}(x), \tag{2.1}$$

such that $\mathcal{H} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ and $\langle f, \mathcal{H}g\rangle = \langle \mathcal{H}f, g\rangle$. Since the operator $\mathcal{H}$ is self-adjoint in $L^2(\mathbb{R})$, its spectrum is real, eigenvalues of the discrete spectrum have equal geometric and algebraic multiplicities, and the spectral decomposition of $L^2(\mathbb{R})$ is orthogonal.

The continuous spectrum of $\mathcal{H}$ is positive and bounded away from zero under Assumption 1.1. The null-space of $\mathcal{H}$ is not empty and includes at least one eigenfunction: $\mathcal{H}\Phi'(x) = 0$, since the nonlinear equation (1.1) has the translation symmetry: $u(x,t) \to u(x-x_0,t)$. The negative spectrum of $\mathcal{H}$ is not empty, since $\mathcal{H}\Phi(x) = (1-p)\Phi^p(x)$ and

$$\begin{aligned}\langle \mathcal{H}\Phi, \Phi\rangle = -(p-1)\langle \Phi^p, \Phi\rangle &= -\frac{(p-1)}{2\pi}\int_{-\infty}^{\infty} \widehat{\Phi}(k)\widehat{\Phi^p}(k)dk \\ &= -\frac{(p-1)}{2\pi}\int_{-\infty}^{\infty} [c+v(k)]\left[\hat{\Phi}(k)\right]^2 dk \ < 0. \end{aligned} \tag{2.2}$$

The analysis does not depend on the number and type of positive eigenvalues of $\mathcal{H}$. We summarize the main properties of the spectrum of $\mathcal{H}$ in the following assumption.

Assumption 2.1. The spectrum of $\mathcal{H}$ in $L^2(\mathbb{R})$ consists of eigenvalues μ of the discrete spectrum for $\mu < c$ and the continuous spectrum for $\mu \geq c$. The null-space of $\mathcal{H}$ is one dimensional with the eigenfunction $\Phi'(x)$. The negative space of $\mathcal{H}$ has dimension $n(\mathcal{H}) \geq 1$.

Two linear eigenvalue problems are associated with the linearized operator $\mathcal{H}$ on $x \in \mathbb{R}$:

$$\text{Problem I:} \qquad \partial_x \mathcal{H} U = \lambda U \tag{2.3}$$

and

$$\text{Problem II:} \qquad \mathcal{H} U = \lambda(c + \mathcal{L})U. \tag{2.4}$$

Problem I occurs in the linearization of the nonlinear wave equation (1.1) with a small perturbation to the bound state: $u = \Phi(x-ct) + U(x-ct)e^{\lambda t}$. The nonzero spectrum of $\partial_x \mathcal{H}$ is defined in the constrained function space $X_c(\mathbb{R})$,

$$X_c = \{U \in L^2(\mathbb{R}) : \ \langle \Phi, U\rangle = 0\}, \tag{2.5}$$

since $\lambda\langle \Phi, U\rangle = \langle \Phi, \partial_x \mathcal{H} U\rangle = -\langle \mathcal{H}\Phi', U\rangle = 0$. The spectrum of $\partial_x \mathcal{H}$ in $X_c(\mathbb{R})$ gives stability or instability of the bound state $\Phi(x)$ in the time evolution of the nonlinear wave equation (1.1). If there exists $\lambda \in \mathbb{C}$ such that $\mathrm{Re}(\lambda) > 0$, the bound state is spectrally unstable and the perturbations grow exponentially in time. If the spectrum is located at the axis $\mathrm{Re}(\lambda) = 0$, the bound state is weakly spectrally stable and the perturbation may grow at most as powers of time. The spectral stability-instability theorem for the scalar wave equation (1.1) can be formulated as follows.

THEOREM 2.2 ([BSS87, PW92]). *. Let $P_s(c) = \langle \Phi, \Phi\rangle$ be a C^1 function of c for $c > 0$ and Assumptions* 1.1 *and* 2.1 *be satisfied. The bound state $\Phi(x)$ is weakly spectrally stable with respect to the time evolution problem* (2.3) *if $n(\mathcal{H}) = 1$ and $P_s'(c) > 0$. The bound state $\Phi(x)$ is spectrally unstable if $n(\mathcal{H}) = 1$ and $P_s'(c) < 0$. The negative space of the operator $\mathcal{H}$ in the constrained function space $X_c(\mathbb{R})$ has the dimension $n(\mathcal{H}) - 1$ if $P_s'(c) > 0$ and the dimension $n(\mathcal{H})$ if $P_s'(c) < 0$.*

Spectral stability of the bound state $\Phi(x)$ occurs if the negative space of $\mathcal{H}$ is empty in the constrained function space $X_c(\mathbb{R})$ and the spectral instability occurs if the negative space of $\mathcal{H}$ is one dimensional in $X_c(\mathbb{R})$. On the contrary, the convergence of the iteration method (1.8)–(1.9) does not depend on spectral stability or instability of bound states. Convergence of the iteration method is related to the spectrum of Problem II, which occurs in the linearization of the iteration method (1.8)–(1.9); see (3.5) below.

We consider the spectrum of the operator $(c+\mathcal{L})^{-1}\mathcal{H}$ in a different constrained space $X_p(\mathbb{R})$,

$$X_p = \{U \in L^2(\mathbb{R}): \ \langle \Phi^p, U\rangle = 0\}. \tag{2.6}$$

The spectrum of $(c+\mathcal{L})^{-1}\mathcal{H}$ diagonalizes simultaneously two linear operators: $\mathcal{H}$ and $(c+\mathcal{L})$. Since $(c+\mathcal{L})$ is positive, all eigenvalues λ are real and the algebraic multiplicity of eigenvalues equals to their geometric multiplicity. Therefore, the spectral decomposition of $L^2(\mathbb{R})$ is orthogonal with respect to the positive weighted inner product $\langle U, (c+\mathcal{L})U\rangle$. In particular, due to the constraint (2.6), the eigenfunction $U(x)$ is orthogonal with respect to $(c+\mathcal{L})$ to $\Phi(x)$, which is the eigenfunction of Problem II for $\lambda = 1-p < 0$.

Before formulating our main result (Theorem 2.8), we study the spectrum of $(c+\mathcal{L})^{-1}\mathcal{H}$ in $X_p(\mathbb{R})$ under Assumption 2.1. Our analysis appears similar to the Birman–Schwinger principle for Schrödinger operators in quantum mechanics [BS87].

LEMMA 2.3. *The negative space of $\mathcal{H}$ in $X_p(\mathbb{R})$ has the dimension $n(\mathcal{H})-1$.*

Proof. The number of eigenvalues of $\mathcal{H}$ in the constrained function space $X_p(\mathbb{R})$ can be found from the constrained eigenvalue problem

$$\mathcal{H}\psi = \mu\psi - \nu\Phi^p(x), \tag{2.7}$$

where (μ,ψ) is the eigenvalue-eigenfunction pair of $\mathcal{H}$ in $X_p(\mathbb{R})$ and ν is the Lagrange multiplier defined from the constraint $\langle \Phi^p, \psi\rangle = 0$. The operator $\mathcal{H}-\mu$ is invertible for any μ not in the spectrum of $\mathcal{H}$, where the spectral decomposition for $\psi(x)$ takes the form

$$\psi(x) = \nu\left[\sum_{\mu_k<0}\frac{\langle u_k,\Phi^p\rangle}{\mu-\mu_k}u_k(x) + \sum_{\mu_k>0}\frac{\langle u_k,\Phi^p\rangle}{\mu-\mu_k}u_k(x)\right]. \tag{2.8}$$

Here (μ_k,u_k) is the eigenvalue-eigenfunction pair of $\mathcal{H}$ in $L^2(\mathbb{R})$ and the formal sum $\sum_{\mu_k>0}$ includes also the integral over the positive continuous spectrum of $\mathcal{H}$. The set of eigenfunctions $\{u_k(x)\}_k$ is assumed to be orthogonal and normalized. The set of eigenvalues μ of $\mathcal{H}$ in $X_p(\mathbb{R})$ consists of two subsets. The first subset is given by eigenvalues μ_k, whose eigenfunctions $u_k(x)$ belong to $X_p(\mathbb{R})$. The other subset is defined by zeros of the function

$$F(\mu) = \frac{1}{\nu}\langle\Phi^p,\psi\rangle = \sum_{\mu_k<0}\frac{|\langle\Phi^p,u_k\rangle|^2}{\mu-\mu_k} + \sum_{\mu_k>0}\frac{|\langle\Phi^p,u_k\rangle|^2}{\mu-\mu_k}. \tag{2.9}$$

We study zeros of $F(\mu)$ by direct application of the theory of constrained variational problems [P04]. The function $F(\mu)$ is monotonically decreasing for $\mu\le 0$ and $\mu\neq\mu_k$. Assume for simplicity that $\mu=\mu_k$ is a single eigenvalue. The function $F(\mu)$ has a jump from negative infinity at $\mu=\mu_k-0$ to positive infinity at $\mu=\mu_k+0$, if the eigenfunction $u_k(x)$ at $\mu=\mu_k$ does not belong to constrained function space $X_p(\mathbb{R})$. Otherwise, i.e., if $u_k(x)$ lies in $X_p(\mathbb{R})$, the function $F(\mu)$ is continuous at $\mu=\mu_k$. The function $F(\mu)$ approaches -0 in the limit $\mu\to-\infty$ and it approaches a positive value in the limit $\mu\to 0$,

$$F(0) = -\langle\Phi^p,\mathcal{H}^{-1}\Phi^p\rangle = \frac{1}{p-1}\langle\Phi^p,\Phi\rangle > 0, \tag{2.10}$$

where we have used the Parseval identity (2.2). The number of negative eigenvalues μ of operator $\mathcal{H}$ in $X_p(\mathbb{R})$ equals the number of zeros of the function $F(\mu)$ for $\mu<0$ and

the number of eigenfunctions $u_k(x)$ of operator $\mathcal{H}$ that belongs to the space $X_p(\mathbb{R})$ for $\mu_k < 0$. By continuity of the decreasing function $F(\mu)$ between $\mu \in [\mu_k, \mu_{k+1}]$ and by counting the jump discontinuity of $F(\mu)$ at $\mu = \mu_k$ [P04], we conclude that the number of negative eigenvalues μ of $\mathcal{H}$ in $X_p(\mathbb{R})$ equals $n(\mathcal{H}) - 1$. □

LEMMA 2.4. *The spectrum of* $(c+\mathcal{L})^{-1}\mathcal{H}$ *in* $X_p(\mathbb{R})$ *has* $n(\mathcal{H})-1$ *negative eigenvalues* λ.

Proof. By Sylvester's inertial theorem [M88, P04], the dimension of the negative space of the quadratic form $\langle U, \mathcal{H}U\rangle$ is invariant in any orthogonal basis of $X_p(\mathbb{R})$ that diagonalizes $\langle U, \mathcal{H}U\rangle$ with respect to a positive weighted inner product. One orthogonal basis for $X_p(\mathbb{R})$ is given by the eigenfunctions $\psi(x)$ of the constrained problem (2.7). The other orthogonal basis with respect to $(c + \mathcal{L})$ is defined by the eigenvalue problem (2.4). By invariance of the negative index of $\mathcal{H}$ in $X_p(\mathbb{R})$, we have $n(\mathcal{H}) - 1$ negative eigenvalues λ in Problem II. □

LEMMA 2.5. *The positive spectrum of* $(c+\mathcal{L})^{-1}\mathcal{H}$ *in* $X_p(\mathbb{R})$ *consists of infinitely many discrete eigenvalues* λ *in the interval* $0 < \lambda < 1$, *accumulating to* $\lambda \to 1^-$. *If* $\Phi^{p-1}(x) \geq 0$ *for* $x \in \mathbb{R}$, *no eigenvalues* λ *exists for* $\lambda > 1$. *If there exists* $x_0 \in \mathbb{R}$ *such that* $\Phi^{p-1}(x_0) < 0$, *the spectrum of* $(c+\mathcal{L})^{-1}\mathcal{H}$ *also includes infinitely many discrete eigenvalues in the interval* $1 < \lambda \leq \lambda_{\max}$, *accumulating to* $\lambda \to 1^+$, *where*

$$\lambda_{\max} < 1 + \frac{p}{c}\left|\min_{x\in\mathbb{R}} \Phi^{p-1}(x)\right| < \infty. \tag{2.11}$$

Proof. Positive eigenvalues λ can be estimated from (2.4) rewritten in the form

$$(c+\mathcal{L})U - \frac{p}{1-\lambda}\Phi^{p-1}(x)U = 0. \tag{2.12}$$

Since $(c + \mathcal{L})$ is positive, no continuous spectrum of the problem (2.12) exists. It was proved in [CM99] for a similar spectral problem that the spectrum of the problem (2.12) is discrete since $\mathrm{tr}M^2 < \infty$, where $M = (c+\mathcal{L})^{-1/2}\Phi^{p-1}(c+\mathcal{L})^{-1/2}$ is a bounded operator. Since the spectrum of $\Phi^{p-1}(x)$ is infinite-dimensional, the spectrum of the bounded operator M cannot be a finite rank [C01]. The potential term in (2.12) becomes singular in the limit $\lambda \to 1$ and therefore the point $\lambda = 1$ is an accumulation point of the discrete eigenvalues. If $\Phi^{p-1}(x) \geq 0$ for any $x \in \mathbb{R}$, the positive part of M and the spectrum of the problem (2.12) in the interval $0 < \lambda < 1$ are infinite-dimensional, with eigenvalues accumulating to $\lambda \to 1^-$. In this case, no eigenvalues exist for $\lambda > 1$, since

$$\lambda = 1 - p\frac{\langle U, \Phi^{p-1}U\rangle}{\langle U, (c+\mathcal{L})U\rangle} < 1. \tag{2.13}$$

If $\Phi^{p-1}(x)$ changes sign on $x \in \mathbb{R}$, the negative part of M and the spectrum of the problem (2.12) for $\lambda > 1$ are infinite-dimensional, with eigenvalues accumulating to $\lambda \to 1^+$ [C01]. Since $\langle U, \mathcal{L}U\rangle \geq 0$ and

$$\langle U, \Phi^{p-1}U\rangle > -\left(\min_{x\in\mathbb{R}} |\Phi^{p-1}(x)|\right)\ \langle U, U\rangle,$$

the largest positive eigenvalue $\lambda = \lambda_{\max}$ is bounded from above by (2.11). The spectrum of $(c + \mathcal{L})^{-1}\mathcal{H}$ is shown schematically on Figure 1. □

COROLLARY 2.6. *The spectrum of* $(c+\mathcal{L})^{-1}\mathcal{H}$ *is located below* $\lambda < 1$ *if and only if* p *is odd or the bound state of the nonlinear problem* (1.3) *is nonnegative,* $\Phi(x) \geq 0$ *on* $x \in \mathbb{R}$.

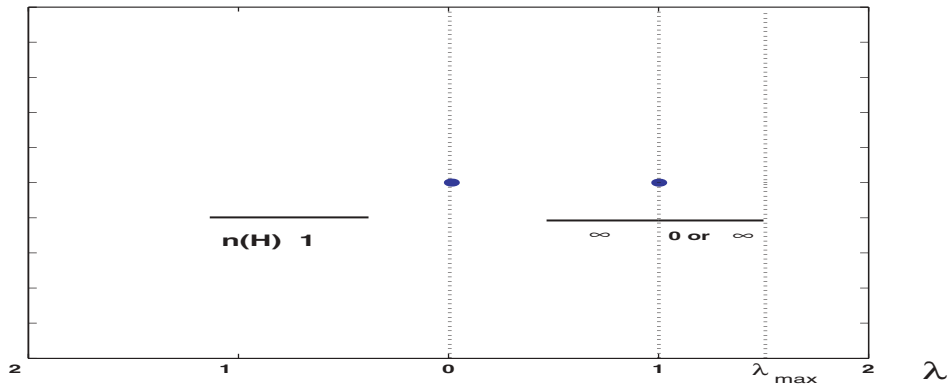


FIG. 1. *Schematical representation of the spectrum of the operator* $(c+\mathcal{L})^{-1}\mathcal{H}$.

Assumption 2.7. Either $\Phi^{p-1}(x) \geq 0$ on $x \in \mathbb{R}$ or $\lambda_{\max} < 2$.

Our main theorem prescribes convergence or divergence of the iteration method (1.8)–(1.9).

THEOREM 2.8. *Let* $\hat{\Phi}(k)$ *be a solution of the boundary-value problem* (1.5) *and Assumptions* 1.1 *and* 2.1 *be satisfied. The iteration method* (1.8)–(1.9) *converges to* $\hat{\Phi}(k)$ *in a small open neighborhood of* $\hat{\Phi}(k)$ *if* (i) $1 < \gamma < (p+1)/(p-1)$, (ii) $n(\mathcal{H}) = 1$, *and* (iii) *Assumption* 2.7 *is met. The fastest rate of convergence occurs for* $\gamma = \gamma_* \equiv p/(p-1)$. *If any of the three conditions are not met, the iteration method* (1.8)–(1.9) *diverges from* $\hat{\Phi}(k)$.

3. Contraction of the iterative method near the fixed point. Our proof of Theorem 2.8 is based on the spectral analysis of the iteration operator (1.8)–(1.9), linearized at $\hat{\Phi}(k)$, and on the application of the contraction mapping principle for nonlinear operators [HP80].

PROPOSITION 3.1. *The iteration operator* (1.8)–(1.9), *linearized at* $\hat{\Phi}(k)$, *has a spectral radius smaller than one if and only if* (i) $1 < \gamma < (p+1)/(p-1)$, (ii) $n(\mathcal{H}) = 1$, *and* (iii) *Assumption* 2.7 *is met.*

Proof. Consider $\widehat{w}_0(k) = \widehat{u}_0(k) - \hat{\Phi}(k)$ be a small perturbation to $\hat{\Phi}(k)$, such that $\langle \Phi', w_0 \rangle = 0$. The sequence $\widehat{w}_n(k) = \widehat{u}_n(k) - \hat{\Phi}(k)$ is generated by the iteration operator (1.8), linearized at $\hat{\Phi}(k)$,

$$\widehat{w}_{n+1}(k) = \gamma m_n \hat{\Phi}(k) + p \frac{\widehat{\Phi^{p-1}} * \widehat{w}_n(k)}{c + v(k)}, \tag{3.1}$$

where $*$ is the convolution operator and $m_n = M_n - 1$. The correction m_n is generated by the stabilizing factor (1.9), linearized at $\hat{\Phi}(k)$,

$$m_n = (1-p) \frac{\int_{-\infty}^{\infty} \widehat{\Phi^p}(k) \widehat{w}_n(k)\, dk}{\int_{-\infty}^{\infty} \widehat{\Phi^p}(k) \hat{\Phi}(k)\, dk}. \tag{3.2}$$

The correction term $w_n(x)$ can be decomposed explicitly as

$$w_n = a_n \Phi(x) + q_n(x), \qquad q_n \in X_p(\mathbb{R}), \tag{3.3}$$

where $X_p(\mathbb{R})$ is defined by (2.6). It follows from (3.1) and (3.2) that $m_n = (1-p)a_n$ and m_n solves the linear map

$$m_{n+1} = [p - \gamma(p-1)]\, m_n. \tag{3.4}$$

On the other hand, the correction term $q_n(x)$ solves the homogeneous part of the problem (3.1), which is equivalently rewritten on $x \in \mathbb{R}$ as

$$q_{n+1}(x) = q_n(x) - (c + \mathcal{L})^{-1}\mathcal{H}q_n(x). \tag{3.5}$$

If $1 < \gamma < (p+1)/(p-1)$, then $\lim_{n\to\infty} m_n = 0$, such that $\lim_{n\to\infty} M_n = 1$ for the stabilizing factor $M_n = 1 + m_n$. Therefore, the first term in the decomposition (3.3) vanishes as $n \to \infty$. The second term $q_n(x)$ may, however, remain finite or even grow with the number of iterations. We derive the conditions when $q_n(x)$ converges to zero as $n \to \infty$.

If $w_0(x)$ is orthogonal to Φ', then $\langle \Phi', q_0 \rangle = 0$. It follows from (3.5) that $\langle \Phi', q_n \rangle = 0$, $\forall n$. We apply, therefore, the spectral decomposition of $X_p(\mathbb{R})$, described in Lemmas 2.4 and 2.5. The sequence $\{q_n(x)\}_{n=0}^{\infty}$ is decomposed through eigenfunctions $U_k(x)$ of the operator $(c + \mathcal{L})^{-1}\mathcal{H}$ as follows:

$$q_n(x) = \sum_{k=1}^{n(\mathcal{H})-1} \alpha_k^{(n)} U_k(x) + \sum_{0<\lambda_k<1} \beta_k^{(n)} U_k(x) + \sum_{1<\lambda_k\le\lambda_{\max}} \gamma_k^{(n)} U_k(x), \tag{3.6}$$

where the first sum represents the finite-dimensional negative space of $X_p(\mathbb{R})$, the second sum represents the infinite-dimensional positive space of $X_p(\mathbb{R})$ for $0 < \lambda < 1$, and the third sum represents the infinite-dimensional positive space of $X_p(\mathbb{R})$ for $1 < \lambda \le \lambda_{\max}$, if the latter exists. The linear maps for coefficients of expansions are

$$\alpha_k^{(n+1)} = (1 + |\lambda_k|)\alpha_k^{(n)}, \quad \lambda_k < 0, \tag{3.7}$$

$$\beta_k^{(n+1)} = (1 - \lambda_k)\beta_k^{(n)}, \quad 0 < \lambda_k < 1, \tag{3.8}$$

and

$$\gamma_k^{(n+1)} = (1 - \lambda_k)\gamma_k^{(n)}, \quad 1 < \lambda_k \le \lambda_{\max}. \tag{3.9}$$

Iterations for coefficients $\alpha_k^{(n)}$ diverge for any $\lambda_k < 0$. Iterations for coefficients $\beta_k^{(n)}$ converges for any $0 < \lambda_k < 1$. Iterations for coefficients $\gamma_k^{(n)}$ diverge for any $\lambda_k \ge 2$ and converge for $1 < \lambda_k < 2$. In the limit $n \to \infty$, the correction $q_n(x)$ uniformly converges to zero if the negative space of $X_p(\mathbb{R})$ is empty, i.e., $n(\mathcal{H}) = 1$, and the positive space of $X_p(\mathbb{R})$ is empty for $\lambda_k \ge 2$. The latter condition is satisfied under Assumption 2.7, i.e., when either the third sum in (3.6) is absent (p is odd or $\Phi(x) \ge 0$ on $x \in \mathbb{R}$) or $\lambda_{\max} < 2$. We note that $\lambda_{\max}$ is bounded from above by (2.11). □

Remark 3.2. In the proof of Proposition 3.1, we have assumed that $\langle \Phi', w_0 \rangle = 0$. If $w_0(x)$ does not satisfy the constraint, iterations of the linearized operator (3.1)–(3.2) converge to the eigenfunction $\Phi'(x)$ of the kernel of $\mathcal{H}$, which simply translates the bound state $\Phi(x)$ in x.

When the kernel of $\mathcal{H}$ has dimension greater than one, the corresponding eigenfunctions translate the bound state $\Phi(x)$ to some other solutions, which typically implies bifurcations of the bound states. It is expected that the iteration method (1.8)–(1.9) selects only one branch of solutions beyond the bifurcation, i.e., the other branches of solutions have the negative index $n(\mathcal{H}) > 1$. We eliminate the bifurcation cases by Assumption 2.1, which ensures that the kernel of $\mathcal{H}$ is one dimensional.

Remark 3.3. The rate of convergence of the stabilizing factor $M_n = 1 + m_n$ becomes superlinear if $\gamma = \gamma_* = p/(p-1)$, see (3.4). However, the corrections $q_n(x)$

still converge with the linear rate at $\gamma = \gamma_*$; see (3.5). Thus, the fastest but linear rate of convergence occurs at $\gamma = \gamma_*$. This conclusion confirms the Petviashvili's conjecture on the fastest rate of convergence [PP92].

PROPOSITION 3.4. *The iteration operator* (1.8)–(1.9)*, linearized at the sequence* $\{\hat{\phi}_n(k)\}_{n=0}^{\infty}$*, is continuous in a small open neighborhood of* $\hat{\Phi}(k)$*.*

Proof. Consider a difference $\delta\hat{u}_n(k) = \hat{u}_n(k) - \hat{\phi}_n(k)$ between any two sequences $\{\hat{u}_n(k)\}_{n=0}^{\infty}$ and $\{\hat{\phi}_n(k)\}_{n=0}^{\infty}$ generated by the iteration operator (1.8)–(1.9). The sequence $\delta\hat{u}_n(k)$ is defined by the iteration operator (1.8), linearized at $\hat{\phi}_n(k)$,

$$\delta\hat{u}_{n+1}(k) = \gamma\frac{\delta M_n}{M_n}\hat{\phi}_{n+1}(k) + pM_n^{\gamma}\frac{\widehat{\phi_n^{p-1}} * \delta\hat{u}_n(k)}{c + v(k)}, \tag{3.10}$$

where $M_n = M_n[\hat{\phi}_n]$ and $\delta M_n = M_n[\hat{\phi}_n + \delta\hat{u}_n] - M_n[\hat{\phi}_n]$. The correction δM_n is generated by the stabilizing factor (1.9), linearized at $\hat{\phi}_n(k)$,

$$\delta M_n = \frac{2\int_{-\infty}^{\infty}[c + v(k)]\,\hat{\phi}_n(k)\delta\hat{u}_n(k)\,dk - (1+p)M_n\int_{-\infty}^{\infty}\widehat{\phi_n^p}(k)\delta\hat{u}_n(k)\,dk}{\int_{-\infty}^{\infty}\widehat{\phi_n^p}(k)\hat{\phi}_n(k)\,dk}. \tag{3.11}$$

The linearized iteration operator (3.10)–(3.11) is continuous with respect to $\phi_n \in X(\mathbb{R})$, where $X(\mathbb{R})$ is defined by (1.6). □

Proof of Theorem 2.8. The iteration method (1.8)–(1.9) represents a nonlinear operator $\hat{u}_{n+1} = \mathcal{A}(\hat{u}_n)$ in function space $X(\mathbb{R})$. The operator $\mathcal{A}(\hat{u}_n)$ has a continuous Frechet derivative $\mathcal{A}'(\hat{u}_n)$ in small open neighborhood of $\hat{\Phi}$ in $X(\mathbb{R})$. Under the three conditions of Proposition 3.1, the spectral radius of $\mathcal{A}'(\hat{\Phi})$ is smaller than one, i.e., $||\mathcal{A}'(\hat{\Phi})|| < 1$. By continuity of the Frechet derivative, for any ϵ with $0 < \epsilon < 1 - ||\mathcal{A}'(\hat{\Phi})||$, there is a small open ball $S(\hat{\Phi}, \delta) \in X(\mathbb{R})$ centered at $\hat{\Phi}(k)$ with the radius $\delta = \delta(\epsilon)$, such that

$$q = \sup_{\hat{u}_n \in S(\hat{\Phi},\delta)} ||\mathcal{A}'(\hat{u}_n)|| < 1. \tag{3.12}$$

It follows from [HP80, Lemma 4.4.7] that

$$||\mathcal{A}(\hat{f}) - \mathcal{A}(\hat{g})|| \leq q||\hat{f} - \hat{g}|| \tag{3.13}$$

for any $\hat{f}, \hat{g} \in S(\hat{\Phi}, \delta)$. Then, the contraction mapping theorem [HP80, Theorem 4.3.4] applies and the nonlinear operator $\mathcal{A}(\hat{u}_n)$ has a unique asymptotically stable fixed point for $\hat{u}_n \in S(\hat{\Phi}, \delta)$. Moreover, the asymptotic rate of convergence is determined by the Frechet derivative at $\hat{\Phi}$ as follows:

$$||\hat{u}_n - \hat{\Phi}|| \leq \left(||\mathcal{A}'(\hat{\Phi})|| + \epsilon\right)^n ||\hat{u}_0 - \hat{\Phi}||. \tag{3.14}$$

See [HP80, Lemma 4.4.8] for further details. □

4. Convergence of self-similar sequences. Here we derive conditions for convergence of a special sequence $\{x_n\hat{\Phi}(k)\}_{n=0}^{\infty}$, which is self-similar to $\hat{\Phi}(k)$ module to amplitude scaling. We also consider convergence of a general sequence in the small open neighborhood of $\{x_n\hat{\Phi}(k)\}_{n=0}^{\infty}$.

PROPOSITION 4.1. *Let* $\hat{\Phi}(k)$ *be a solution of the boundary-value problem* (1.5) *and Assumption* 1.1 *be satisfied. There exists a sequence* $\{x_n\hat{\Phi}(k)\}_{n=0}^{\infty}$ *in the iteration map* (1.8)–(1.9)*, which converges to* $\hat{\Phi}(k)$ *for any* $x_0 > 0$ *if* $1 < \gamma < (p+1)/(p-1)$*.*

Proof. Define $\hat{u}_0(k) = x_0\hat{\Phi}(k)$ for any $x_0 > 0$. Then, it follows from (1.5), (1.8), and (1.9) that $\hat{u}_n(k) = x_n\hat{\Phi}(k)$ for any $n \geq 0$, where x_n is defined by the power iteration map

$$x_{n+1} = M_n^\gamma x_n^p = x_n^{p-\gamma(p-1)}, \tag{4.1}$$

where $M_n = x_n^{1-p}$. The iteration map converges for $1 < \gamma < (p+1)/(p-1)$ with the limit $\lim_{n\to\infty} x_n = 1$. As a result, $\lim_{n\to\infty} \hat{u}_n(k) = \hat{\Phi}(k)$. ∎

Remark 4.2. The rate of convergence of the power iteration map (4.1) is linear for $\gamma \neq \gamma_*$, where $\gamma_* = p/(p-1)$. When $\gamma = \gamma_*$, the convergence occurs in a single iteration: $\hat{u}_1(k) = \hat{\Phi}(k)$ for any $x_0 > 0$. The starting value $\hat{u}_0(k)$ is self-similar to the bound state $\hat{\Phi}(k)$ module to amplitude scaling. The special sequence $\{\hat{u}_n(k)\}_{n=0}^\infty$ exists in the iteration map (1.8)–(1.9) due to the power nonlinearity. The special sequence does not exist for general nonlinear functions.

PROPOSITION 4.3. *Let $\hat{\Phi}(k)$ be a solution of the boundary-value problem* (1.5) *and Assumptions* 1.1 *and* 2.1 *be satisfied. Let $\{x_n\hat{\Phi}(k)\}_{n=0}^\infty$ be a self-similar sequence, where x_n is generated by the power iteration map* (4.1) *with any $x_0 > 0$. The iteration operator* (1.8)–(1.9)*, linearized at $\{x_n\hat{\Phi}(k)\}_{n=0}^\infty$, has a spectral radius smaller than one if and only if* (i) $1 < \gamma < (p+1)/(p-1)$, (ii) $n(\mathcal{H}) = 1$, *and* (iii) *Assumption* 2.7 *is met.*

Proof. We use the linear map (3.10)–(3.11) with $\hat{\phi}_n(k) = x_n\hat{\Phi}(k)$, where x_n solves the power iteration map (4.1). As a result, we find that $M_n = x_n^{1-p}$. The linear map (3.10)–(3.11) is then equivalent to the linear map (3.1)–(3.2) with the relations

$$\hat{w}_n(k) = \frac{\delta\hat{u}_n(k)}{x_n}, \quad m_n = \frac{\delta M_n}{x_n^p}.$$

Thus, Proposition 4.3 is equivalent to Proposition 3.1. ∎

5. Examples in one dimension. Here we discuss two examples of the scalar wave equation (1.1), where the iteration method (1.8)–(1.9) can be used for finding stationary solutions such as solitary waves.

Example 5.1 (generalized Korteweg–de Vries (KdV) equations). A family of generalized KdV equations is defined for $\mathcal{L} = -\partial_x^2$, such that $v(k) = k^2 \geq 0$ and $m = 2$. The bound state solutions of the boundary-value problem (1.3) exist for $p > 1$ in the analytical form (see, e.g., [PW92])

$$\Phi(x) = \left[\sqrt{\frac{(p+1)c}{2}}\,\operatorname{sech}\left(\frac{p-1}{2}\sqrt{c}x\right)\right]^{\frac{2}{p-1}}. \tag{5.1}$$

It follows from (5.1) that the bound state $\Phi(x)$ decays exponentially as

$$\lim_{|x|\to\infty} \Phi(x)e^{\sqrt{c}|x|} = a_\infty, \quad a_\infty = [2(p+1)c]^{\frac{1}{p-1}}. \tag{5.2}$$

The function $\Phi(x)$ belongs to $X(\mathbb{R})$ of Assumption 1.1. Since $\Phi(x) \geq 0$ on $x \in \mathbb{R}$, it also satisfies Assumption 2.7. The linearized operator $\mathcal{H}$ becomes a Schrödinger operator with a solvable potential,

$$\mathcal{H} = c - \partial_x^2 - \frac{p(p+1)c}{2}\operatorname{sech}^2\left(\frac{p-1}{2}\sqrt{c}x\right). \tag{5.3}$$

The Schrödinger operator (5.3) satisfies Assumption 2.1. Since $\mathcal{H}\Phi'(x) = 0$ and $\Phi(x)$ has no nodes on $x \in \mathbb{R}$, the Sturm oscillation theorem predicts only one negative eigenvalue of $\mathcal{H}$, i.e., $n(\mathcal{H}) = 1$. As a result, Theorem 2.8 applies and the iteration method (1.8)–(1.9) converges to the bound state $\hat{\Phi}(k)$ in the generalized KdV equation for any value of $p > 1$ if $1 < \gamma < (p+1)/(p-1)$.

Remark 5.2. In accordance with Theorem 2.2, the bound state $\Phi(x)$ is weakly spectrally stable with respect to the time evolution problem for $p < 5$ and spectrally unstable for $p \geq 5$ (see also [BSS87, PW92]). On the other hand, the iteration method (1.8)–(1.9) converges for any $p > 1$, irrelevantly to the stability of bound states in the time evolution problem. For instance, the interval of convergence with $p = 5$ is $1 < \gamma < 3/2$ and the interval shrinks to zero when $p \to \infty$.

Example 5.3 (generalized Benjamin–Ono (BO) equations). A family of generalized BO equations is defined for $\mathcal{L} = -\partial_x H$, where $H(u)$ is the Hilbert transform of $u(x)$,

$$H(u) = \frac{1}{\pi} \wp \int_{-\infty}^{\infty} \frac{u(z)dz}{z - x}, \tag{5.4}$$

and the symbol $\wp$ denotes the principal value of the integral. In this case, $v(k) = |k| \geq 0$ and $m = 1$. The bound state solutions of the nonlinear problem (1.3) are unknown in the analytical form except for the case $p = 2$, when

$$\Phi(x) = \frac{2c}{1 + c^2 x^2}. \tag{5.5}$$

Using the asymptotic representation for $\Phi(x) \in L^1(\mathbb{R})$,

$$H(\Phi) = -\frac{1}{\pi x} \int_{-\infty}^{\infty} \Phi(z)dz + \mathrm{O}\left(\frac{1}{x^2}\right),$$

and the balance of inverse powers of x in the problem (1.3), we derive the algebraic decay of $\Phi(x)$ at infinity,

$$\lim_{|x| \to \infty} x^2 \, \Phi(x) = a_{-2}, \quad a_{-2} = \frac{1}{\pi c} \int_{-\infty}^{\infty} \Phi(x)dx. \tag{5.6}$$

The function $\Phi(x)$ has sufficient decay at infinity to belong to $X(\mathbb{R})$ of Assumption 1.1, if it exists for $p > 1$. Since $\Phi(x) \geq 0$ on $x \in \mathbb{R}$ as follows from our numerical approximations (see Figure 2), Assumption 2.7 is satisfied. The linearized operator $\mathcal{H}$ becomes a nonlocal operator,

$$\mathcal{H} = c - \partial_x H - p\Phi^{p-1}(x). \tag{5.7}$$

It was proved in [CK80] for $p = 2$ that the nonlocal linearized operator (5.7) satisfies Assumption 2.1 and has only one negative eigenvalue, i.e., $n(\mathcal{H}) = 1$. As a result, Theorem 2.8 states that the iteration method (1.8)–(1.9) converges to the bound state $\hat{\Phi}(k)$ for the case $p = 2$ if $1 < \gamma < 3$.

We have computed the bound states $\Phi(x)$ for $p = 2, 3, 4, 5$ from the iteration method (1.8)–(1.9) starting with the Gaussian approximation $u_0(x) = \exp(-x^2)$ for $c = 1$ (see also [AS87]). The numerical approximations are plotted on Figure 2, where dots for $p = 2$ show the exact values from (5.5). Figure 3 shows convergence of the stabilizing factor M_n in the iteration method (1.8)–(1.9) with $p = 2$, for three

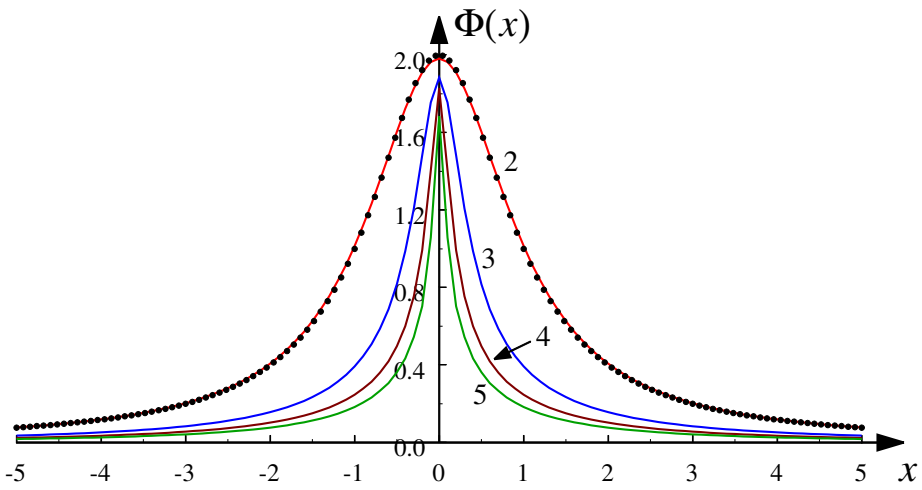


FIG. 2. *Numerical approximations of the bound states* $\Phi(x)$ *of the generalized BO equation for* $p = 2, 3, 4, 5$. *Dots on curve* 2 *show exact values from the analytical solution* (5.5).

different values of γ: $\gamma = 2$ (dots), when the rate of convergence is the fastest; $\gamma = 1.1$ (triangles), near the left boundary of the convergence interval; and $\gamma = 2.9$ (crosses), near the right boundary of the convergence interval. We conclude that the iteration method (1.8)–(1.9) converges to the bound state of the generalized BO equation for $p = 2, 3, 4, 5$ if $1 < \gamma < (p + 1)/(p - 1)$. Moreover, numerical computations show convergence of the method to a positive-definite bound state $\Phi(x)$ for any $p > 1$, including noninteger values of p.

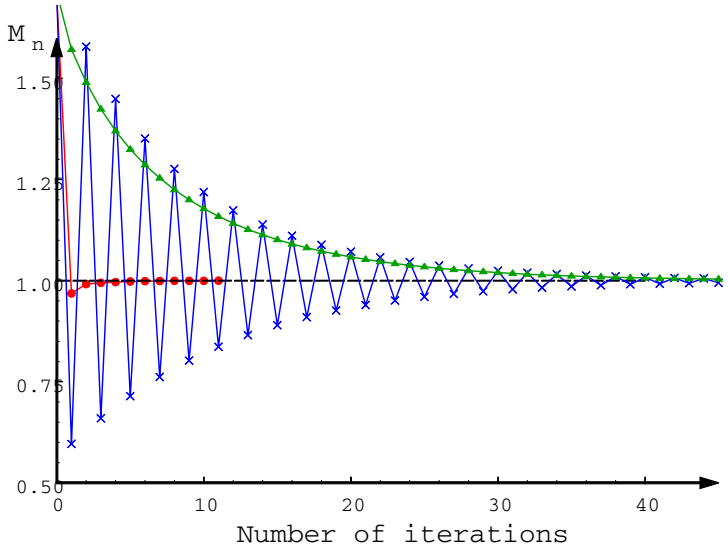


FIG. 3. *Stabilizing factor* M_n *versus* n *for the iteration method* (1.8)–(1.9) *with* $p = 2$ *for* $\gamma = 2$ *(dots),* $\gamma = 1.1$ *(triangles), and* $\gamma = 2.9$ *(crosses).*

Remark 5.4. The bound state $\Phi(x)$ is spectrally stable with respect to the time evolution problem for $p = 2$ and spectrally unstable for $p \geq 3$ [BSS87, CK80]. Under assumption that $n(\mathcal{H}) = 1$ and $\Phi(x) \geq 0$ on $x \in \mathbb{R}$ for any $p > 1$, the iteration method (1.8)–(1.9) converges to the bound state for any $p > 1$, irrelevantly to the stability of bound states in the time evolution problem. Therefore, the method becomes useful for numerical approximations of the bound states in the generalized BO equation, when exact analytical expressions are not available. In fact, the iteration method (1.8)–(1.9) was successfully used for numerical approximations of soliton solutions in the generalized BO and KdV equations in [AS87]. Another numerical method is developed with the help of Newton iteration algorithms but the Newton iterations have convergence problems as pointed out in [BK03]. We notice that Petviashvili's iteration method (1.8)–(1.9) is not sensitive to the choice of a starting function, which is its great advantage compared to the Newton's iteration method.

6. Examples in two dimensions. We finish the article with generalizations of the iteration method (1.8)–(1.9) for the scalar wave equation in space of two dimensions,

$$u_t - (\mathcal{L}u)_x + pu^{p-1}u_x = 0, \tag{6.1}$$

where $u : \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$, $p > 1$, and $\mathcal{L}$ is a linear self-adjoint nonnegative pseudo-differential operator in x and y with constant coefficients. If the Fourier transform (1.4) is replaced by the double Fourier transform in $L^2(\mathbb{R}^2)$, the iteration method (1.8)–(1.9) can be applied to the scalar wave equation (6.1) in two dimensions. The only modification is required for Assumption 2.1, since the kernel of $\mathcal{H} = c + \mathcal{L} - p\Phi^{p-1}(x, y)$ has at least two eigenfunctions $\partial_x\Phi(x, y)$ and $\partial_y\Phi(x, y)$.

Assumption 6.1. The spectrum of $\mathcal{H}$ in $L^2(\mathbb{R}^2)$ consists of eigenvalues μ of the discrete spectrum for $\mu < c$ and the continuous spectrum for $\mu \geq c$. The null-space of $\mathcal{H}$ is two dimensional with the eigenfunctions $\partial_x\Phi(x, y)$ and $\partial_y\Phi(x, y)$. The negative space of $\mathcal{H}$ has dimension $n(\mathcal{H}) \geq 1$.

With this modification, we formulate the results of sections 2 and 3 as the following theorem.

THEOREM 6.2. *Let $\hat{\Phi}(k)$ be a solution of the boundary-value problem* (1.5) *and Assumptions* 1.1 *and* 6.1 *be satisfied. The iteration method* (1.8)–(1.9) *converges to $\hat{\Phi}(k)$ in a small open neighborhood of $\hat{\Phi}(k)$ if* (i) $1 < \gamma < (p+1)/(p-1)$, (ii) $n(\mathcal{H}) = 1$, *and* (iii) *Assumption* 2.7 *is met. The fastest rate of convergence occurs for $\gamma = \gamma_* \equiv p/(p-1)$. If any of the three conditions are not met, the iteration method* (1.8)–(1.9) *diverges from $\hat{\Phi}(k)$.*

Here we discuss three examples of the scalar wave equation (6.1) in two dimensions, where the iteration method (1.8)–(1.9) can be used for finding stationary solutions such as solitary waves.

Example 6.3 (generalized Zakharov–Kuznetsov (ZK) equations). The generalized KdV equations of Example 5.1 are extended to the two-dimensional ZK equations, when $\mathcal{L}$ is an isotropic operator,

$$\mathcal{L} = -(\partial_x^2 + \partial_y^2), \tag{6.2}$$

such that $v(k) = k_x^2 + k_y^2 \geq 0$. The bound state $u = \Phi(x - ct, y)$ satisfies the nonlinear problem

$$c\Phi - \Delta\Phi = \Phi^p. \tag{6.3}$$

Existence and uniqueness of positive solutions of the nonlinear elliptic problem (6.3) was proved for any $p > 1$ [GNN81, K89] such that $\Phi(x, y) = \Phi(r)$ is radially symmetric, where $r = \sqrt{x^2 + y^2}$, and satisfies the limiting decay

$$\lim_{r\to\infty} e^{\sqrt{c}r} r^{1/2} \Phi(r) = a_\infty > 0. \tag{6.4}$$

The positive solutions $\Phi(r)$ satisfy Assumptions 1.1 and 2.7. The linearized operator $\mathcal{H}$ becomes the Schrödinger operator with the radially symmetric potential

$$\mathcal{H} = c - \partial_x^2 - \partial_y^2 + p\Phi^{p-1}(r). \tag{6.5}$$

Assumption 6.1 is satisfied for the Schrödinger operator (6.5) and the negative index of $\mathcal{H}$ for the positive ground state $\Phi(r)$ is one, i.e., $n(\mathcal{H}) = 1$ [S9, p. 63]. Therefore, iterations of the numerical method (1.8)–(1.9) converge for $1 < \gamma < (p + 1)/(p - 1)$, according to Theorem 6.2 for any $p > 1$. This result justifies the use of the iteration method (1.8)–(1.9) for numerical approximation of bound states of the generalized ZK equations.

Example 6.4 (generalized Kadomtsev–Petviashvili (KP) equations). The generalized KdV equations of Example 5.1 are extended to the two-dimensional KP equations, when $\mathcal{L}$ is an anisotropic operator,

$$\mathcal{L} = -\partial_x^2 + \partial_x^{-2}\partial_y^2, \tag{6.6}$$

such that $v(k) = k_x^2 + k_x^{-2}k_y^2 \geq 0$. The linear operator $\mathcal{L}$ in (6.6) corresponds to the KPI equation with two-dimensional solitons, called lumps. The nonlocal ∂_x^{-1} operator is well posed subject to the constraint on $u(x, y, t)$

$$\int_{-\infty}^{\infty} u(x, y, t)\, dx = 0. \tag{6.7}$$

The bound state $u = \Phi(x - ct, y)$ satisfies the nonlinear problem

$$c\Phi - \Phi_{xx} + \partial_x^{-2}\Phi_{yy} = \Phi^p. \tag{6.8}$$

The exact analytical solution for $\Phi(x, y)$ exists for $p = 2$ [MZ77],

$$\Phi(x, y) = 12c\frac{3 + c^2y^2 - cx^2}{(3 + c^2y^2 + cx^2)^2}. \tag{6.9}$$

The bound state $\Phi(x, y)$ is sign-indefinite due to the constraint (6.7). Existence of sign-indefinite bound states in the nonlinear problem (6.8) was proved for $p = 3, 4$ by using constrained minimization [BS97]. It was also shown that the solution exists only for $p < 5$ and $p = p_1/p_2$, where p_1 is any even integer and p_2 is any odd integer [LW97]. Bound states $\Phi(x, y)$ satisfy Assumption 1.1.

It can be shown with the Riemann–Hilbert inverse scattering method [PS00] that the spectrum of $\mathcal{H}$ for $p = 2$ satisfies Assumption 6.1 with $n(\mathcal{H}) = 1$. Since the bound states $\Phi(x, y)$ are nonpositive, they satisfy Assumption 2.7 only if $\lambda_{\max} < 2$. It follows from (6.9) for $p = 2$ that

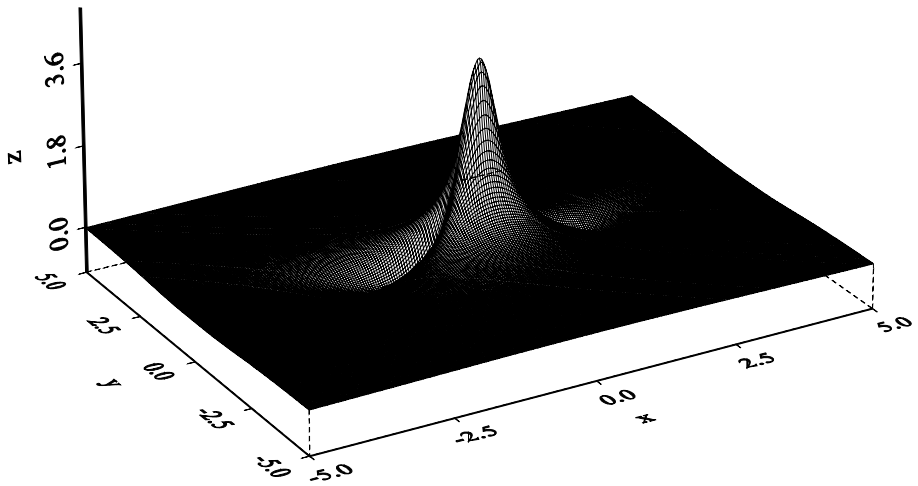


FIG. 4. *A numerical approximation of the bound state $\Phi(x,y)$ of the generalized KPI equation with $p=3$.*

$$\min_{(x,y)\in\mathbb{R}^2} \Phi(x,y) = \Phi\left(\pm\frac{3}{\sqrt{c}},0\right) = -\frac{c}{2}.$$

Therefore, the upper bound (2.11) applies with $\lambda_{\max} < 1+1 = 2$, i.e., Assumption 2.7 is also satisfied. Theorem 6.2 states that the iteration method (1.8)–(1.9) converges to $\Phi(x,y)$ for $p=2$ if $1<\gamma<3$. This analysis justifies the use of the numerical iteration method (1.8)–(1.9), proposed originally by Petviashvili [P76].

We have computed the bound states $\Phi(x,y)$ for $p=2,3,4$ from the iteration method (1.8)–(1.9) starting with the lump solution (6.9) with $c=1$. The constraint (6.7) is built into the algorithm as zero Fourier mode with $k_x=0$. The final solution $\Phi(x,y)$ is shown on Figure 4 for $p=3$ (see also [AS87]). Cross-sections $\Phi(x,0)$ and $\Phi(0,y)$ are shown on Figure 5(a,b) for $p=2,3,4$, where dots for $p=2$ show exact values from (6.9). Figure 6 shows convergence of the stabilizing factor M_n in the iteration method (1.8)–(1.9) with the fastest rate $\gamma=p/(p-1)$ for $p=2,3,4$. We conclude that the iteration method (1.8)–(1.9) converges to the bound state of the generalized KP equation for $p=2,3,4$ if $1<\gamma<(p+1)/(p-1)$.

Remark 6.5. Nonpositive bound states of the generalized KP equations may consist of several individual lumps. Multilump solutions of the KPI equation with $p=2$ were discovered both numerically [AS85] and analytically [PS93]. However, a discrepancy occurs between the numerical and analytical solutions for a double-lump; the analytical solution is unique for the double-lump [PS93], while the numerical solution represents a continuous family with a free parameter of the distance between the two lumps [AS85]. This discrepancy is likely to be explained by low accuracy of the numerical procedure in [AS85], i.e., low resolution of the numerical mesh and small grid size. Since the negative index of $\mathcal{H}$ for multilump solutions typically exceeds one, the iteration method (1.8)–(1.9) must diverge in the neighborhood of multilump solutions, according to Theorem 6.2. Numerical approximations obtained in [AS85] are likely supported by the truncation of the domain on $\mathbb{R}^2$ and the discretization

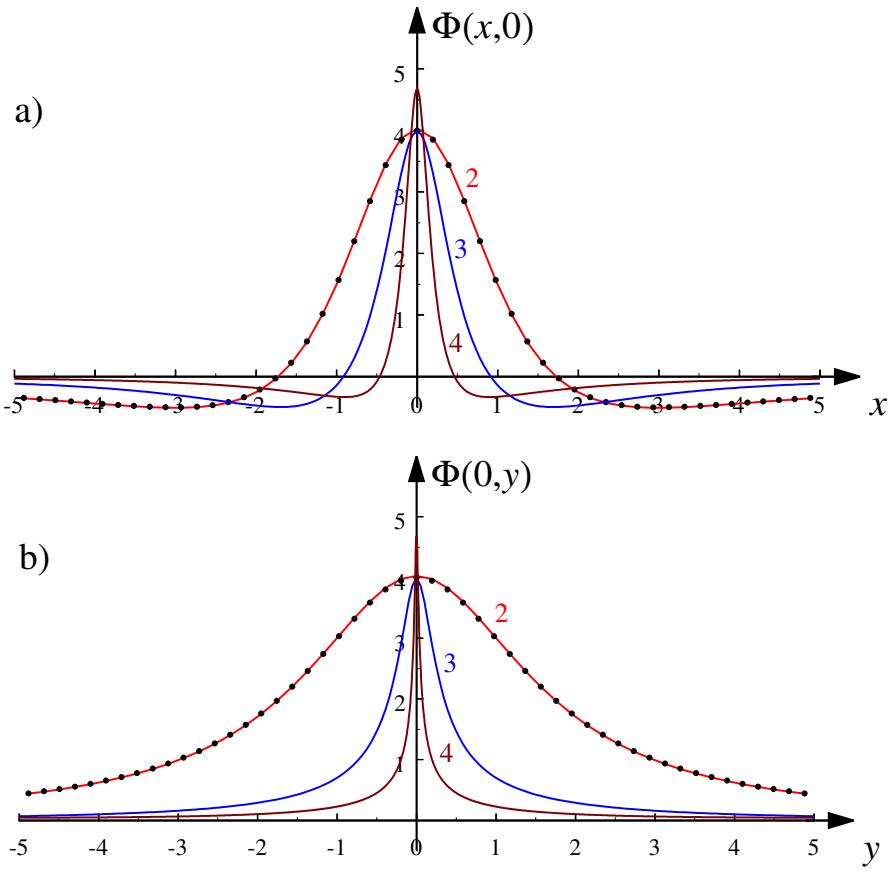


FIG. 5. *Cross sections* $\Phi(x,0)$ *and* $\Phi(0,y)$ *of numerical approximations of the bound states* $\Phi(x,y)$ *of the generalized KPI equation with* $p = 2,3,4$. *Dots on curves* 2 *show exact values from the analytical solution* (6.9).

of the numerical grid (x,y). This example shows a danger of the direct use of the iteration method (1.8)–(1.9) without analysis of the three conditions of convergence in Theorems 2.8 and 6.2.

Example 6.6 (generalized Klein–Gordon (KG) equations). Our last example shows that the iteration method (1.8)–(1.9) can be used for other nonlinear problems, such as the generalized KG equation,

$$u_{tt} - c_0^2(u_{xx} + u_{yy}) + u = u^p. \tag{6.10}$$

Travelling wave solutions of (6.10) are of the form $u(x,y,t) = \Phi(x - ct, y)$, where $\Phi(x,y)$ satisfies the boundary–value problem

$$\Phi - (c_0^2 - c^2)\Phi_{xx} - c_0^2\Phi_{yy} = \Phi^p. \tag{6.11}$$

If $|c| < c_0$, the boundary–value problem (6.11) can be reduced to the form (6.3) of Example 6.3 with a simple rescaling of variables x and y.

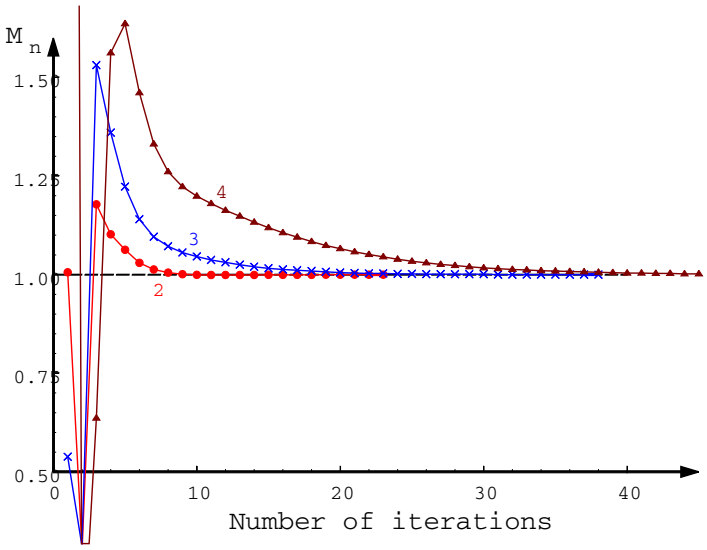


FIG. 6. *Stabilizing factor M_n versus n in the iteration method* (1.8)–(1.9) *with the fastest rate* $\gamma = p/(p-1)$ *for* $p = 2, 3, 4$.

Acknowledgments. The authors are thankful to H. Kalisch and A. N. Notik for collaboration on the early stage of the work. One of the authors (D. P.) appreciates useful discussions with R. Gadyl'shin, A. Pushnitski, A. Scheel, and V. Vougalter.

REFERENCES

[AS85] L. A. ABRAMYAN AND YU. A. STEPANYANTS, *Two-dimensional multisolitons: stationary solutions of the Kadomtsev–Petviashvili equation*, Radiophysics and Quantum Electronics, 28 (1985), pp. 20–26.

[AS87] L. A. ABRAMYAN AND YU. A. STEPANYANTS, *Structure of two-dimensional solitons in the context of a generalized Kadomtsev–Petviashvili equation*, Radiophysics and Quantum Electronics, 30 (1987), pp. 861–865.

[BS87] M. SH. BIRMAN AND M. Z. SOLOMYAK, *Spectral Theory of Self-Adjoint Operators in Hilbert Space*, Reidel, Dordrecht, The Netherlands, 1987.

[BK03] J. BONA AND H. KALISCH, *Singularity Formation in the Generalized Benjamin–Ono Equation*, preprint, 2003.

[BSS87] J. L. BONA, P. E. SOUGANIDIS, AND W. A. STRAUSS, *Stability and instability of solitary waves of Korteweg-de Vries type*, Proc. Roy. Soc. London Ser. A, 411 (1987), pp. 395–412.

[BS97] A. DE BOUARD AND J. C. SAUT, *Solitary waves of generalized Kadomtsev–Petviashvili equations*, Ann. Inst. H. Poincare, Anal. Non Lineaire, 14 (1997), pp. 211–236.

[CK80] H. H. CHEN AND D. J. KAUP, *Linear stability of internal wave solitons in a deep stratified fluid*, Phys. Fluids, 23 (1980), pp. 235–238.

[C01] A. CONSTANTIN, *On the scattering problem for the Camassa–Holm equation*, Proc. R. Soc. Lond. Ser. A Mat. Phys. Eng. Sci. 457 (2001), pp. 953–970.

[CM99] A. CONSTANTIN AND H. P. MCKEAN, *A shallow water equation on the circle*, Comm. Pure Appl. Math., 52 (1999), pp. 949–982.

[GNN81] B. GIDAS, W. M. NI, AND L. NIRENBERG, *Symmetry of positive solutions of nonlinear elliptic equations in* $\mathbb{R}^N$, in Mathematical Analysis and Applications, Adv. in Math. Suppl. Studies 7A, L. Nachbin, ed., Academic Press, New York, 1981, pp. 369–402.

[HP80] V. HUTSON AND J. S. PYM, *Applications of Functional Analysis and Operator Theory*, Academic Press, London, 1980.

[K89] M. K. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$ *in* $\mathbb{R}^N$, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[LW97] Y. Liu and X. P. Wang, *Nonlinear stability of solitary waves of a generalized Kadomtsev–Petviashvili equation*, Comm. Math. Phys., 183 (1997), pp. 253–266.

[M88] J. H. Maddocks, *Restricted quadratic forms, inertia theorems, and the Schur complement*, Linear Algebra Appl., 108 (1988), pp. 1–36.

[MZ77] S. V. Manakov, V. E. Zakharov, L. A. Bordag, A. R. Its, and V. B. Matveev, *Two-dimensional solitons of the Kadomtsev–Petviashvili equation and their interaction*, Phys. Lett. A, 63 (1977), pp. 205–206.

[PW92] R. L. Pego and M. I. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Phil. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.

[P04] D. Pelinovsky, *Inertia law for spectral stability of solitary waves in coupled nonlinear Schrödinger equations*, preprint, Proc. Roy. Soc. Lond. Ser. A, (2004).

[PS93] D. E. Pelinovsky and Yu. A. Stepanyants, *New multisoliton solutions of the Kadomtsev–Petviashvili equations*, JETP Lett., 57 (1993), pp. 24–28.

[PS00] D. E. Pelinovsky and C. Sulem, *Eigenfunctions and eigenvalues for a scalar Riemann–Hilbert problem associated to inverse scattering*, Comm. Math. Phys., 208 (2000), pp. 713–760.

[P76] V. I. Petviashvili, *Equation of an extraordinary soliton*, Plasma Physics, 2 (1976), p. 469.

[PP92] V. I. Petviashvili and O. V. Pokhotelov, *Solitary Waves in Plasmas and in the Atmosphere*, Gordon and Breach, Philadelphia, 1992.

[S9] W. A. Strauss, *Nonlinear Wave Equations*, CBMS Reg. Conf. Ser. Math. 73, AMS, Providence, RI, 1989.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1128–1145

Γ-CONVERGENCE OF DISCRETE FUNCTIONALS WITH NONCONVEX PERTURBATION FOR IMAGE CLASSIFICATION*

GILLES AUBERT†, LAURE BLANC-FÉRAUD‡, AND RICCARDO MARCH§

Abstract. The purpose of this paper is to show the theoretical soundness of a variational method proposed in image processing for supervised classification. Based on works developed for phase transitions in fluid mechanics, the classification is obtained by minimizing a sequence of functionals. The method provides an image composed of homogeneous regions with regular boundaries, a region being defined as a set of pixels belonging to the same class. In this paper, we show the Γ-convergence of the sequence of functionals which differ from the ones proposed in fluid mechanics in the sense that the perturbation term is not quadratic but has a finite asymptote at infinity, corresponding to an edge-preserving regularization term in image processing.

Key words. Γ-convergence, finite elements, image processing, phase transitions

AMS subject classifications. 49M25, 49J45, 68U10

DOI. 10.1137/S0036142902412336

1. Introduction.

Image classification consists of assigning a label to each site of an image to produce a partition of the image into homogeneous labelled areas. The classification problem concerns many applications as, for instance, land use management in remote sensing.

Based on results conducted in the Van der Waals–Cahn–Hilliard theory framework for phase transitions in fluid mechanics [2, 4, 13, 18, 20], we have recently proposed a sequence of functionals for image classification [19]. The soundness of such a method relies upon Γ-convergence theory. The purpose of this paper is to prove the Γ-convergence of the sequence of functionals we use, which differs from the one used in fluid mechanics in the sense that the perturbation term is not quadratic, but it is an edge-preserving regularization term as defined in image processing.

Let Ω be an open bounded subset of $\mathbf{R}^2$, $I : \Omega \to \mathbf{R}$ the observed data to classify, $I \in L^\infty(\Omega)$. A class is characterized by parameters of the spatial distribution of intensity, i.e., the mean and standard deviation for Gaussian hypothesis. This work takes place in the general framework of supervised classification, which means that the number n of classes and the parameters of the Gaussian distribution of the classes (a_i, σ_i) are known a priori. These values either are given by an expert or are precomputed by using a fuzzy C means algorithm with an entropy term (see [16], for instance). Knowing (a_i, σ_i), $i = 1, \ldots, n$, the question is now to find a partition of Ω based on the observed image, where a component is the set of pixels in class i. We also add a regularity constraint on the partition. In order to assign a class i to each

*Received by the editors August 1, 2002; accepted for publication (in revised form) October 22, 2003; published electronically September 18, 2004.
http://www.siam.org/journals/sinum/42-3/41233.html
†Laboratoire J. A. Dieudonné, Umr 6621 du Cnrs, Université de Nice Sophia Antipolis, 06108 Nice Cedex 2, France (gaubert@math.unice.fr).
‡Projet Ariana, laboratoire I3S (CNRS/UNSA) and INRIA Sophia Antipolis, Inria, 2004 route des Lucioles, BP 93, 06902 Sophia Antipolis cedex, France (blancf@sophia.inria.fr).
§Istituto per le Applicazioni del Calcolo, CNR, Viale del Policlinico 137, 00161 Roma, Italy (march@iac.rm.cnr.it).

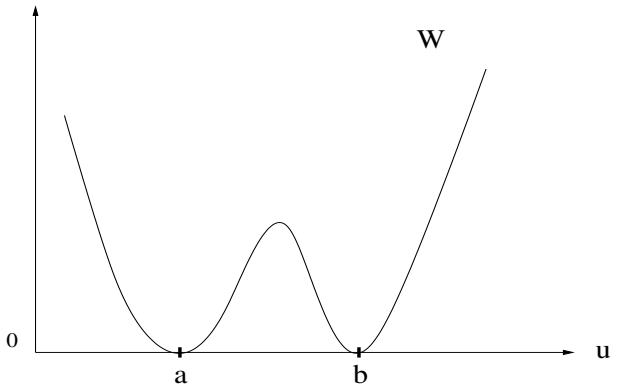


FIG. 1.1. *Example of double-well potential W, in the case of two classes with $a_1 = a, a_2 = b$.*

pixel x, we have proposed in [19] the sequence of functionals

$$F_\varepsilon(u) = \underbrace{\int_\Omega |u(x) - I(x)|^2 \, dx}_{\text{data term}} + \varepsilon \underbrace{\int_\Omega \varphi(|\nabla u(x)|) dx}_{\text{restoration term}} + \frac{1}{\varepsilon} \underbrace{\int_\Omega W(u(x)) dx}_{\text{classification}}, \tag{1.1}$$

and the associated problem consists of finding u_0 such that

$$u_0 = \lim_{\varepsilon \to 0^+} \Big[\arg\min_u F_\varepsilon(u) \Big]. \tag{1.2}$$

Let us first consider the functional with a fixed ε. The first two terms of (1.1) are standard for noisy image restoration by nonquadratic regularization [10, 3]. Function φ is a smoothing function that will be defined later.

The third term of (1.1) is a level constraint such that $W : \mathbf{R} \to \mathbf{R}^+$ attracts the values of $u(x)$ towards the mean a_i of class i, taking into account the standard deviation σ_i. W has n minima at a_i such that $W(a_i) = 0 \; \forall \; i = 1, \ldots, n$. W is quadratic around each minimum (from the Gaussian distribution hypothesis), i.e., around the a_i, $W(t) = (\frac{t - a_i}{\sigma_i})^2$, and is piecewise parabolic between the wells (see Figure 1.1).

Considering a sequence of energies F_ε when $\varepsilon \to 0$ is inspired from works conducted in the Van der Waals–Cahn–Hilliard theory framework for phase transitions in fluid mechanics [2, 4, 13, 18, 20] using Γ-convergence.

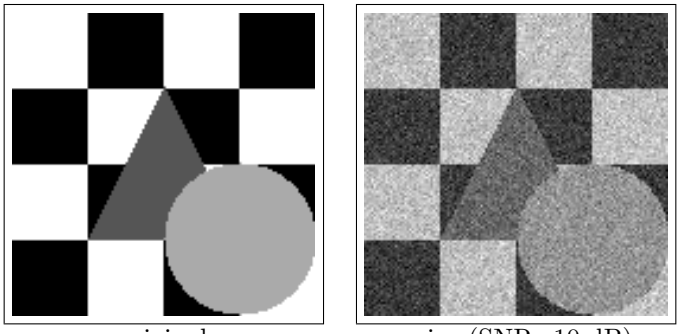
We recall the definition and some properties of Γ-convergence (see [12]). Let X be a metric space, and let $f_\varepsilon : X \to [0, +\infty]$ be a family of functions indexed by $\varepsilon > 0$. We say that f_ε Γ-converge as $\varepsilon \to 0^+$ to $f : X \to [0, +\infty]$ if the two conditions

$$\forall x_\varepsilon \to x, \quad \liminf_{\varepsilon \to 0^+} f_\varepsilon(x_\varepsilon) \geq f(x) \tag{1.3}$$

and

$$\exists x_\varepsilon \to x, \quad \limsup_{\varepsilon \to 0^+} f_\varepsilon(x_\varepsilon) \leq f(x) \tag{1.4}$$

are fulfilled for every $x \in X$. The Γ-limit, if it exists, is unique and lower semicontinuous. The Γ-convergence is stable under continuou perturbations, that is, $(f_\varepsilon + v)$



original noisy (SNR=10 dB)

FIG. 1.2. *Synthetic "check" image.*

Γ-converge to $(f+v)$ if f_ε Γ-converge to f and v is continuous. The most important property of Γ-convergence is the following: if $\{x_\varepsilon\}_\varepsilon$ is asymptotically minimizing, i.e.,

$$\lim_{\varepsilon \to 0^+} \left(f_\varepsilon(x_\varepsilon) - \inf_X f_\varepsilon \right) = 0, \tag{1.5}$$

and if $\{x_{\varepsilon_h}\}_h$ converge to x for some sequence $\varepsilon_h \to 0$, then x minimizes f.

The minimization problem (1.2) relies upon Γ-convergence arguments. If $\varphi(t) = t^2$, then it can be shown from [4] that the sequence of functionals (1.1) Γ-converges to

$$F_0(u) = \begin{cases} \displaystyle\sum_{i=1}^{n} \int_{A_i} |a_i - I|^2 dx \quad + \displaystyle\sum_{i,l=1}^{n} |\kappa_{i,l}| \mathcal{H}^1(\partial^* A_i \cap \partial^* A_l \cap \Omega) \\ \qquad\qquad\qquad\qquad \text{if } u \in BV(\Omega; \{a_1, \ldots, a_n\}), \\ +\infty \qquad\qquad\qquad\qquad \text{elsewhere in} L^2(\Omega), \end{cases} \tag{1.6}$$

where $BV(\Omega)$ is the space of functions of bounded variation [1], $\mathcal{H}^1$ is the one-dimensional Hausdorff measure, and $\partial^* A_i$ is the essential boundary of the subset A_i. For $u \in BV(\Omega; \{a_1, \ldots, a_n\})$, $A_i = \{x \in \Omega : u(x) = a_i\}$ for any $i = 1, \ldots, n$. Then the sets $A_1, \ldots, A_n$ define a partition of Ω into sets with finite perimeter. This partition is the classification result. The weight $\kappa_{i,l}$ is defined by

$$\kappa_{i,l} = \int_{a_i}^{a_l} \sqrt{W(t)} dt. \tag{1.7}$$

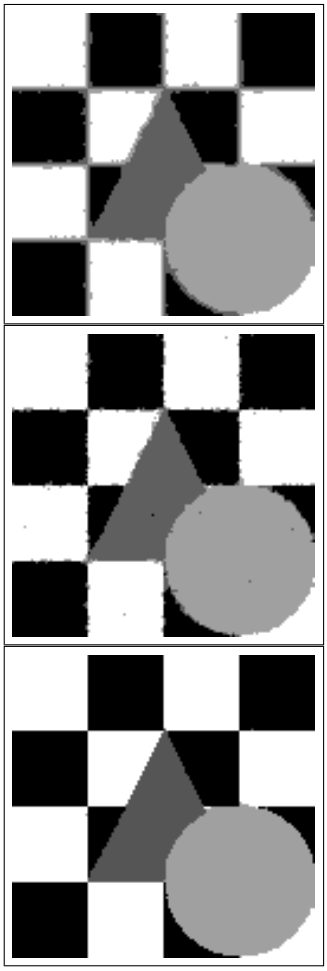
From Γ-convergence and compactness results, we know that the sequence of minimizers u_ε of $F_\varepsilon(u)$ converges (up to a subsequence) to a minimizer of F_0. So u_0 defines a partition of Ω according to the predefined classes, with minimal interfaces with respect to the weighted length (1.6), (1.7).

From the numerical point of view, when ε decreases, the functional turns from a restoration process (the third term in (1.1) is negligible) into a classification process.

We do not use the quadratic function for φ but an edge-preserving regularizing function $\varphi(t) = \frac{t^2}{1+\mu t^2}$ because, numerically, it gives better results by preserving high gradients which represent edges [10]. This is illustrated on a synthetic image of size 128×128 pixels ("check" image), containing four classes.

The white Gaussian noise introduced is such that the signal-to-noise ratio (SNR) given by $SNR = 10 \log_{10} \frac{nonnoisy\ signal\ variance}{noise\ variance}$ is 10 dB. Figure 1.2 presents the

synthetic image (nonnoisy and noisy). From the noisy one, we compute the classification as in (1.2) for different φ-functions (see [19] for the detailed algorithm). The results are presented in Figure 1.3. For a Tikhonov regularization ($\varphi(t) = t^2$), edges are oversmoothed. With a convex φ, there are still many misclassified pixels on the boundaries. Best results are provided with the use of the nonconvex function $\varphi(t) = \frac{t^2}{1+\mu t^2}$, with $\mu > 0$.



$\varphi(t) = t^2$: convex (Tikhonov)

$\varphi(t) = \log(\cosh(t))$: convex (Green)

$\varphi(t) = \frac{t^2}{1+\mu t^2}$: nonconvex (Geman & McClure)

FIG. 1.3. *Classification of "check" image with different functions φ. Nonconvex functions provide better results than convex functions which lead to oversmooth results: we get damaged edges.*

Before stating the result shown in this paper, we observe that the family of functionals F_ε in (1.1) does not Γ-converge to the limit F_0 given in (1.6) if $\varphi(t) = \frac{t^2}{1+\mu t^2}$ as it does when $\varphi(t) = t^2$.

Let, for instance, $n = 2$ and $u_0 \in BV(\Omega; \{a_1, a_2\})$ with $a_1 < a_2$. Let A_ε be the tubular neighborhood of S_{u_0}, the set of jumps of u_0, defined by $A_\varepsilon = \{x \in \Omega : \operatorname{dist}(x, S_{u_0}) < \lambda_\varepsilon\}$, with $\lambda_\varepsilon > 0$. Then let $u_\varepsilon \in W^{1,2}(\Omega)$ be a function which makes a sharp transition between the values a_1 and a_2 in the set A_ε and takes the values a_1 and a_2 outside of A_ε.

If we neglect the term $\int_\Omega |u_\varepsilon(x) - I(x)|^2 \, dx$, since it is a continuous perturbation,

it is easy to check that the remaining part of the energy, denoted by $E_\varepsilon(u_\varepsilon)$, is bounded by const. $|A_\varepsilon|/\varepsilon$, where $|\cdot|$ denotes the Lebesgue measure. As λ_ε can be chosen in such a way that $|A_\varepsilon| \to 0$ as fast as we want with ε, then $E_\varepsilon(u_\varepsilon)$ converges to 0 as $\varepsilon \to 0$. So u_ε is a counterexample to the lower inequality of the Γ-convergence. This example shows that too-sharp transitions make the proof of the Γ-convergence fail. In order to obtain the Γ-convergence with the nonconvex function φ, we have to consider the subspace of $W^{1,2}(\Omega)$ of finite elements and to use a method introduced by Chambolle and Dal Maso in [9]. The meshsize of the discretization will limit the sharpness of the transitions.

The paper is organized as follows. In section 2 we define the sequence of functionals and we state the Γ-convergence result. The proof of Γ-convergence is given in sections 3 and 4. Section 5 is devoted to the compactness of the minimizers for the sequence of functionals. In section 6 we show that the evaluation of the discrete functionals via the vertex quadrature rule does not affect the Γ-convergence and compactness results.

2. Mathematical preliminaries and statement of the result. In the following, $|A|$ denotes the two-dimensional Lebesgue measure of a set $A \subset \mathbf{R}^2$, and $\mathcal{H}^1(\partial A)$ denotes the one-dimensional Hausdorff measure of ∂A.

Let $\Omega \subset \mathbf{R}^2$ be a bounded open set. We will use standard notation for the Lebesgue and Sobolev spaces $L^p(\Omega)$ and $W^{1,p}(\Omega)$. We say that $u \in L^1(\Omega)$ is a function of bounded variation in Ω, and we write $u \in BV(\Omega)$, if the distributional derivative Du of u is a vector-valued Radon measure with finite total variation in Ω. We denote by $|Du|$ the total variation of Du and by ∇u the density of the absolutely continuous part of Du with respect to the Lebesgue measure. It can be proved [1] that ∇u coincides almost everywhere with the approximate differential of u. We denote by $u^-(x)$, $u^+(x)$ the approximate lower and upper limit of u at the point x, and we denote by S_u the discontinuity set of u in an approximate sense, defined as

$$S_u = \{x \in \Omega : u^-(x) < u^+(x)\}.$$

We say that a Borel set $A \subset \mathbf{R}^2$ is a set with finite perimeter in Ω if $\chi_A \in BV(\Omega)$, where χ_A denotes the characteristic function of A. We denote by $\partial^* A$ the essential boundary of A, i.e., the set of points where A does not have density 0 or 1. The perimeter of A in Ω is then given by $|D\chi_A|(\Omega) = \mathcal{H}^1(\partial^* A \cap \Omega)$.

In the following, $\Omega \subset \mathbf{R}^2$ will denote an open polygonal domain. Let θ_0 be an angle such that $0 < \theta_0 \leq \pi/3$, and let $\nu(h)$ be a function such that $\nu(h) \geq h$ for any $h > 0$ and $\nu(h) = O(h)$ as $h \to 0^+$. Let us denote by $\{\mathbf{T}_h\}_h$ a family of triangulations of Ω made of triangles whose edges, for any $h > 0$, have length between h and $\nu(h)$, and whose angles are all greater than or equal to θ_0.

We denote by $V_h(\Omega) \subset W^{1,2}(\Omega) \cap C^0(\overline{\Omega})$ the linear finite element space

$$V_h(\Omega) = \{u : \Omega \to \mathbf{R} \ : \ u \text{ continuous}, u|_T \in P_1(T) \ \forall T \in \mathbf{T}_h\},$$

where T denotes a triangle of $\mathbf{T}_h$, $u|_T$ denotes the restriction of u to T, and $P_1(T)$ denotes the space of polynomials of degree 1 on T. We denote by $\pi_h : C^0(\overline{\Omega}) \to V_h(\Omega)$ the Lagrange interpolation operator.

Let $\{a_1, \ldots, a_n\} \subset \mathbf{R}$ with $a_1 < a_2 < \cdots < a_n$. Let $W : \mathbf{R} \to \mathbf{R}$ be a function with the following properties:

(i) W is $\mathcal{C}^1(\mathbf{R})$ with Lipschitz continuous derivative;

(ii) W is $\mathcal{C}^2$ in a neighborhood of a_i for any $i \in \{1, \ldots, n\}$;

(iii) $W(t) > 0$ for any $t \notin \{a_1, \dots, a_n\}$ and

$$W(a_i) = 0, \quad W'(a_i) = 0, \quad W''(a_i) > 0 \qquad \forall i \in \{1, \dots, n\}; \tag{2.1}$$

(iv) $W(t)$ is monotone increasing for $t \geq a_n$, and monotone decreasing for $t \leq a_1$. For any $i, l \in \{1, \dots, n\}$ we set

$$\kappa_{i,l} = \int_{a_i}^{a_l} \sqrt{W(t)}dt.$$

For any $I \in L^\infty(\Omega)$ such that $\|I\|_{L^\infty(\Omega)} \leq K < +\infty$, we set

$$+\infty > M > \max\{|a_1|, |a_n|, K\}. \tag{2.2}$$

For any $h > 0$ and any $\varepsilon > 0$ we define the functional $E_{\varepsilon,h} : L^2(\Omega) \to [0, +\infty]$ by

$$E_{\varepsilon,h}(u) = \begin{cases} \varepsilon \displaystyle\int_\Omega \frac{|\nabla u|^2}{1 + \mu_{\varepsilon,h}|\nabla u|^2} dx + \frac{1}{\varepsilon} \int_\Omega W(u)dx & \text{if } u \in \mathcal{D}(E_{\varepsilon,h}), \\ +\infty & \text{elsewhere in } L^2(\Omega), \end{cases}$$

where $\mu_{\varepsilon,h} > 0$ and $\mathcal{D}(E_{\varepsilon,h}) = \{u \in V_h(\Omega) : \|u\|_{L^\infty(\Omega)} \leq M\}$.

We say that n Borel sets $A_1, \dots, A_n$ define a partition of Ω if

$$A_i \cap A_l = \emptyset \quad \forall i, l \in \{1, \dots, n\},\ i \neq l, \qquad |\Omega \setminus \cup_{i=1}^n A_i| = 0.$$

Let $u \in BV(\Omega; \{a_1, \dots, a_n\})$ and let $A_i = \{x \in \Omega : u(x) = a_i\}$ for any $i = 1, \dots, n$. Then the sets $A_1, \dots, A_n$ define a partition of Ω into sets with finite perimeter.

Then we define the functional $E_0 : L^2(\Omega) \to [0, +\infty]$ by

$$E_0(u) = \begin{cases} 2 \displaystyle\sum_{\substack{i,l=1 \\ i<l}}^n \kappa_{i,l} \mathcal{H}^1(\partial^* A_i \cap \partial^* A_l \cap \Omega) & \text{if } u \in BV(\Omega; \{a_1, \dots, a_n\}), \\ +\infty & \text{elsewhere in } L^2(\Omega). \end{cases}$$

Finally we state the main result of the paper. We define

$$F_{\varepsilon,h}(u) = \int_\Omega (u - I)^2 dx + E_{\varepsilon,h}(u),$$

and we will prove the following theorems.

THEOREM 2.1. *Assume that $h = o(\varepsilon|\log\varepsilon|^{-1})$ and that $\mu_{\varepsilon,h} = o(\varepsilon h)$. Then the family $\{F_{\varepsilon,h}\}_\varepsilon$ Γ-converges to the functional*

$$\int_\Omega (u - I)^2 dx + E_0(u)$$

in the $L^2(\Omega)$-topology as $\varepsilon \to 0^+$.

Since the term $\int_\Omega (u - I)^2 dx$ is a continuous perturbation with respect to the strong-$L^2(\Omega)$ topology, in order to prove the theorem it will be enough to prove that the family of functionals $\{E_{\varepsilon,h}\}_\varepsilon$ Γ-converges to the functional E_0.

THEOREM 2.2. *Assume that $h = o(\varepsilon|\log\varepsilon|^{-1})$ and that $\mu_{\varepsilon,h} = o(\varepsilon h)$. Then any family $\{u_{\varepsilon,h}\}_\varepsilon$ of absolute minimizers of $F_{\varepsilon,h}$ is relatively compact in $L^2(\Omega)$, and each of its limit points minimizes the functional*

$$\int_\Omega (u - I)^2 dx + E_0(u).$$

3. Lower inequality. In this section we investigate the Γ-convergence lower inequality (1.3) with $f_\varepsilon = E_{\varepsilon,h}$ and $f = E_0$.

THEOREM 3.1. *Let $\mu_{\varepsilon,h} = o(\varepsilon h)$ and $h = h(\varepsilon)$ with $\lim_{\varepsilon\to 0^+} h(\varepsilon) = 0$. Then, for every function $u_0 \in L^2(\Omega)$ and for every sequence $\{u_{\varepsilon,h}\}_\varepsilon \subset L^2(\Omega)$ converging to u_0 in $L^2(\Omega)$ as $\varepsilon \to 0^+$, we have*

$$\liminf_{\varepsilon\to 0^+} E_{\varepsilon,h}(u_{\varepsilon,h}) \geq E_0(u_0).$$

We need the following lemma.

LEMMA 3.2. *Assume that $\mu_{\varepsilon,h} = o(\varepsilon h)$. Then, for every $\varepsilon > 0$ and for every $u \in V_h(\Omega)$ $(0 < h < 1)$, there exists $v \in BV(\Omega)$ such that*

$$E_{\varepsilon,h}(u) \geq (1-\delta_h)\varepsilon \int_\Omega |\nabla v|^2 dx + \frac{1}{\varepsilon}\int_\Omega W(v)dx + 2\kappa_{1,n}\mathcal{H}^1(S_v), \tag{3.1}$$

$$\frac{h}{c}E_{\varepsilon,h}(u) \geq |\{x \in \Omega : v(x) \neq u(x)\}|, \tag{3.2}$$

where $v(x) = a_1$ for any $x \in \Omega$ such that $v(x) \neq u(x)$, $\{\delta_h\}_h$ is a sequence of positive numbers converging to zero, and c is a constant independent of h.

The proof of the lemma is essentially the same of Proposition 3.3 in [9], with some slight modifications which can be found in the appendix.

Proof of Theorem 3.1. Up to the extraction of a subsequence, we may assume that $\{u_{\varepsilon,h}\}_\varepsilon \subset \mathcal{D}(E_{\varepsilon,h})$, and

$$+\infty > \liminf_{\varepsilon\to 0^+} E_{\varepsilon,h}(u_{\varepsilon,h}) = \lim_{\varepsilon\to 0^+} E_{\varepsilon,h}(u_{\varepsilon,h}); \tag{3.3}$$

otherwise the result is trivial. To simplify the notation we set $u_\varepsilon = u_{\varepsilon,h(\varepsilon)}$ and we assume that u_ε converges a.e. to u_0 as $\varepsilon \to 0^+$.

Using (3.3) and Fatou's lemma, we deduce that $\int_\Omega W(u_0)dx = 0$; thus $W(u_0(x)) = 0$ a.e. in Ω. Then, using (2.1), there exists a partition $\{A_i\}_{i=1,\dots,n}$ of Ω into measurable subsets such that $u_0(x) = \sum_{i=1}^n a_i \chi_{A_i}(x)$.

For any $\varepsilon > 0$, Lemma 3.2 provides a function $v_\varepsilon \in BV(\Omega)$ which satisfies (3.1) and (3.2). Since $\|v_\varepsilon\|_{L^\infty(\Omega)} \leq M$, we have $|v_\varepsilon - u_0| \leq 2M$ a.e. in Ω. Then, using (3.2) and (3.3), we have $|\{v_\varepsilon \neq u_\varepsilon\}| \to 0$ as $\varepsilon \to 0^+$, from which we deduce that v_ε converges to u_0 in $L^2(\Omega)$ as $\varepsilon \to 0^+$.

Let $\widehat{v}_\varepsilon$, with $a_1 \leq \widehat{v}_\varepsilon \leq a_n$, denote the truncated function

$$\widehat{v}_\varepsilon = \max\{a_1, \min\{v_\varepsilon, a_n\}\}.$$

We have that $\widehat{v}_\varepsilon$ converges to u_0 in $L^2(\Omega)$ as $\varepsilon \to 0^+$. Now we define the two functions

$$g(y) = \int_{a_1}^y \sqrt{W(t)}dt \tag{3.4}$$

and $\psi_\varepsilon(x) = g(\widehat{v}_\varepsilon(x))$. Since g is Lipschitz continuous, we have $\psi_\varepsilon \to g(u_0)$ in $L^2(\Omega)$ and

$$g(u_0(x)) = \kappa_{1,l} \quad \text{if } u_0(x) = a_l \quad \forall l \in \{1,\dots,n\}. \tag{3.5}$$

For any ε we have $\psi_\varepsilon \in BV(\Omega)$, $S_{\psi_\varepsilon} \subseteq S_{v_\varepsilon} = S_{v_\varepsilon}$, and the approximate differential is given by

$$\nabla\psi_\varepsilon(x) = \sqrt{W(\widehat{v}_\varepsilon(x))}\nabla\widehat{v}_\varepsilon(x).$$

The following estimate then holds for the total variation:

$$|D\psi_\varepsilon|(\Omega) = \int_\Omega |\nabla\psi_\varepsilon|dx + \int_{S_{\psi_\varepsilon}} |\psi_\varepsilon^+ - \psi_\varepsilon^-|d\mathcal{H}^1 \le \int_\Omega |\nabla\psi_\varepsilon|dx + \kappa_{1,n}\mathcal{H}^1(S_{v_\varepsilon}).$$

Using (3.1) and the above estimate we have

$$\begin{aligned} E_{\varepsilon,h}(u_\varepsilon) &\ge (1-\delta_h)\varepsilon\int_\Omega |\nabla v_\varepsilon|^2dx + \frac{1}{\varepsilon}\int_\Omega W(v_\varepsilon)dx + 2\kappa_{1,n}\mathcal{H}^1(S_{v_\varepsilon}) \\ &\ge (1-\delta_h)\varepsilon\int_\Omega |\nabla\widehat{v}_\varepsilon|^2dx + \frac{1}{\varepsilon}\int_\Omega W(\widehat{v}_\varepsilon)dx + 2\kappa_{1,n}\mathcal{H}^1(S_{v_\varepsilon}) \\ &\ge 2(1-\delta_h)^{1/2}\int_\Omega \sqrt{W(\widehat{v}_\varepsilon)}|\nabla\widehat{v}_\varepsilon|dx + 2\kappa_{1,n}\mathcal{H}^1(S_{v_\varepsilon}) \\ &\ge 2(1-\delta_h)^{1/2}|D\psi_\varepsilon|(\Omega), \end{aligned} \tag{3.6}$$

from which, using (3.3) and the compactness theorem in BV [14], it follows that $g(u_0) \in BV(\Omega)$. Then the sets A_i have finite perimeter, so that $u_0 \in BV(\Omega;\{a_1,\dots,a_n\})$ and $E_0(u_0) < +\infty$.

Using (3.6) and the lower semicontinuity of the total variation, we find

$$\liminf_{\varepsilon\to 0^+} E_{\varepsilon,h}(u_\varepsilon) \ge 2\lim_{\varepsilon\to 0^+}(1-\delta_h)^{1/2}\liminf_{\varepsilon\to 0^+}|D\psi_\varepsilon|(\Omega) \ge 2|Dg(u_0)|(\Omega).$$

Then, using (3.5), we have

$$|Dg(u_0)|(\Omega) = \int_{S_{g(u_0)}} |g(u_0)^+ - g(u_0)^-|d\mathcal{H}^1 = \sum_{\substack{i,l=1\\ i<l}}^n \kappa_{i,l}\mathcal{H}^1(\partial^*A_i\cap\partial^*A_l\cap\Omega), \tag{3.7}$$

which concludes the proof. □

4. Upper inequality. In this section we investigate the Γ-convergence upper inequality (1.4) with $f_\varepsilon = E_{\varepsilon,h}$ and $f = E_0$.

THEOREM 4.1. *Assume that $h = o(\varepsilon|\log\varepsilon|^{-1})$. Then, for every function $u_0 \in L^2(\Omega)$ there exists a sequence $\{u_{\varepsilon,h}\}_\varepsilon \subset L^2(\Omega)$ converging to u_0 in $L^2(\Omega)$ as $\varepsilon \to 0^+$ such that*

$$\limsup_{\varepsilon\to 0^+} E_{\varepsilon,h}(u_{\varepsilon,h}) \le E_0(u_0).$$

First we need the following lemma.

LEMMA 4.2. *For any $i,l \in \{1,\dots,n\}$ with $i < l$, there exists a sequence of functions $\{\gamma_\varepsilon^{(i,l)}\}_\varepsilon \subset \mathcal{C}^1(\mathbf{R})$ with the following properties:*

(i) *$\gamma_\varepsilon^{(i,l)}(t) = a_l$ for $t \ge \rho_\varepsilon$, $\gamma_\varepsilon^{(i,l)}(t) = a_i$ for $t \le -\rho_\varepsilon$, with $\rho_\varepsilon > 0$ independent of the pair (i,l) and $\rho_\varepsilon = O(\varepsilon|\log\varepsilon|)$;*

(ii)

$$\|d\gamma_\varepsilon^{(i,l)}/dt\|_{L^\infty(-\rho_\varepsilon,\rho_\varepsilon)} = O\left(\frac{1}{\varepsilon}\right), \qquad \|d^2\gamma_\varepsilon^{(i,l)}/dt^2\|_{L^\infty(-\rho_\varepsilon,\rho_\varepsilon)} = O\left(\frac{1}{\varepsilon^2}\right);$$

(iii)

$$\lim_{\varepsilon\to 0^+}\int_{-\infty}^{+\infty}\left[\varepsilon\left(\frac{d\gamma_\varepsilon^{(i,l)}}{dt}\right)^2+\frac{1}{\varepsilon}W(\gamma_\varepsilon^{(i,l)})\right]dt=2\kappa_{i,l}.$$

In the proof of this lemma a standard construction of Γ-convergence theory, used for a double-well potential (see, for instance, [20]), is extended to the case when the potential W has multiple wells on the real axis. The resulting construction of the functions $\gamma_\varepsilon^{(i,l)}$ is sketched in the appendix.

In order to prove Theorem 4.1 we need the following density result.

LEMMA 4.3. *Let $u_0\in BV(\Omega;\{a_1,\dots,a_n\})$; then there exists a sequence $\{u_\varepsilon\}_\varepsilon\subset BV(\Omega;\{a_1,\dots,a_n\})$ such that*

(i) *the set $A_i^\varepsilon=\{x\in\Omega : u_\varepsilon(x)=a_i\}$ is polygonal and $\mathcal{H}^1(\partial A_i^\varepsilon\cap\partial\Omega)=0$ for any $i=1,\dots n$ and for any $\varepsilon>0$;*

(ii) *$u_\varepsilon\to u_0$ in $L^2(\Omega)$ as $\varepsilon\to 0^+$;*

(iii)

$$\lim_{\varepsilon\to 0^+}E_0(u_\varepsilon)=E_0(u_0).$$

Proof. This approximation lemma is due to Baldo [4, Lemma 3.1]. Parts (i) and (ii) are stated exactly as above in Baldo [4]. Part (iii) needs some developments. We first set out Baldo's result and then we explain how we can use it in our context.

Let W be given as in (2.1) and let us define on $\mathbf{R}$ the metric

$$d(\xi_1,\xi_2)=\inf\left\{\int_0^1\sqrt{W(\gamma(t))}|\gamma'(t)|dt\quad;\quad\gamma(0)=\xi_1,\,\gamma(1)=\xi_2,\,\gamma\in\mathcal{C}^1\left([0,1];\mathbf{R}\right)\right\}.$$

Then let us set $g_i(\xi)=d(a_i,\xi)$, and let us define the Borel measures

$$\mu_i(B)=\int_B|Dg_i(u_0)|,\qquad\mu_i^\varepsilon(B)=\int_B|Dg_i(u_\varepsilon)|,$$

where B is a Borel set. In [4], Baldo proved the following result:

$$\lim_{\varepsilon\to 0^+}\left(\bigvee_{i=1}^n\mu_i^\varepsilon\right)(\Omega)=\left(\bigvee_{i=1}^n\mu_i\right)(\Omega)=\frac{1}{2}\sum_{i,j=1}^n d(a_i,a_j)\mathcal{H}^1(\partial^*A_i\cap\partial^*A_j\cap\Omega),\tag{4.1}$$

where the symbol $\bigvee$ denotes the supremum of a family of measures. In what follows we show that (4.1) is nothing else than part (iii) of Lemma 4.3.

First, one can prove that for any $i,j=1,\dots,n$, $i<j$, we have

$$d(a_i,a_j)=\kappa_{i,j}=\int_{a_i}^{a_j}\sqrt{W(y)}dy.$$

Hence, if $\{\mu_\alpha\}_{\alpha\in A}$ is a family of regular positive Borel measures, the supremum of $\{\mu_\alpha\}_{\alpha\in A}$ is defined as follows: let E be any subset of Ω; then

$$\left(\bigvee_{\alpha\in A}\mu_\alpha\right)(E)=\sup\left\{\sum_{\alpha\in A'}\mu_\alpha(E_\alpha);\ E_\alpha\text{ disjoint open sets in }\Omega, E=\bigcup_{\alpha\in A'}E_\alpha\right\},$$

where A' is any finite or countable subfamily of A. For any open subset $\Omega' \subset \Omega$ we have

$$\left(\bigvee_{i=1}^{n} \mu_i^\varepsilon\right)(\Omega') = \sup\left\{\sum_i \mu_i^\varepsilon(\Omega_i);\ \Omega_i \text{ open disjoint},\ \Omega' = \bigcup \Omega_i\right\}.$$

But with the same computations used in (3.7) we get

$$\begin{aligned}\sum_i \mu_i^\varepsilon(\Omega_i) &= \sum_i \sum_{j<k} \kappa_{jk}\mathcal{H}^1(\partial A_k^\varepsilon \cap \partial A_j^\varepsilon \cap \Omega_i) = \sum_{j<k} \kappa_{jk} \sum_i \mathcal{H}^1(\partial A_k^\varepsilon \cap \partial A_j^\varepsilon \cap \Omega_i)\\ &\le \sum_{j<k} \kappa_{jk}\mathcal{H}^1(\partial A_k^\varepsilon \cap \partial A_j^\varepsilon \cap \Omega').\end{aligned}$$

Since the sets A_i^ε are polygonal, there exists a partition of Ω' into open subsets Ω_i such that $\mathcal{H}^1(\partial A_k^\varepsilon \cap \partial A_j^\varepsilon \cap \partial\Omega_i) = 0$ for any $i, j, k \in \{1, \dots, n\}$, so that the above inequality becomes an equality. Thus for all $\Omega' \subset \Omega$ and for any $\varepsilon > 0$, we obtain

$$\left(\bigvee_{i=1}^{n} \mu_i^\varepsilon\right)(\Omega') = \sum_{j<k} \kappa_{jk}\mathcal{H}^1(\partial A_k^\varepsilon \cap \partial A_j^\varepsilon \cap \Omega');$$

i.e., the supremum with respect to i "disappears." In particular for $\Omega' = \Omega$ Baldo's result (4.1) reads as

$$\lim_{\varepsilon\to 0^+} \sum_{j<k} \kappa_{jk}\mathcal{H}^1(\partial A_k^\varepsilon \cap \partial A_j^\varepsilon) = \sum_{j<k} \kappa_{jk}\mathcal{H}^1(\partial^* A_k \cap \partial^* A_j \cap \Omega),$$

which exactly means that

$$\lim_{\varepsilon\to 0^+} E_0(u_\varepsilon) = E_0(u_0),$$

i.e., part (iii) of Lemma 4.3. □

We can now prove the upper inequality.

Proof of Theorem 4.1. In the following we use a method developed by Bellettini, Paolini, and Verdi in [6]. Since most of the estimates we need can be proved in the same way as in [6, proof of Theorem 2.1], we will omit the details.

We assume $u_0 \in BV(\Omega; \{a_1, \dots, a_n\})$; otherwise the inequality is trivial. Using Lemma 4.3 and a diagonal argument, we can suppose that the set $A_i = \{x \in \Omega : u_0(x) = a_i\}$ is a polygonal domain with $\mathcal{H}^1(\partial A_i \cap \partial\Omega) = 0$ for any $i = 1, \dots, n$. We have

$$S_{u_0} = \bigcup_{\substack{i,l=1\\ i<l}}^{n} (\partial A_i \cap \partial A_l \cap \Omega).$$

We denote by $\mathcal{P}_i$ the set of the vertices of the polygon A_i, and we set $\mathcal{P} = \cup_{i=1}^n \mathcal{P}_i$; hence $\mathcal{P}$ is a finite set of points. We denote by ω the minimum angle between the edges of S_{u_0}.

Following [6], we introduce the following notations. We set

$$\begin{aligned}(S_{u_0})_{\rho_\varepsilon} &= \{x \in \Omega : \operatorname{dist}(x, S_{u_0}) \le \rho_\varepsilon\};\\ \Pi_{S_{u_0}}(x) &= \{y \in S_{u_0} : |y - x| = \operatorname{dist}(x, S_{u_0})\};\\ Q_\varepsilon &= \left\{x \in (S_{u_0})_{\rho_\varepsilon} : \operatorname{dist}(\Pi_{S_{u_0}}(x), \mathcal{P}) \le \cot\left(\frac{\omega}{2}\right)\rho_\varepsilon\right\};\\ Q_{\varepsilon,h} &= \bigcup\{T \in \mathbf{T}_h : T \cap Q_\varepsilon \ne \emptyset\}.\end{aligned}$$

We denote by $\mathcal{M}$ the set of all the pairs of integers (i,l) such that $i,l \in \{1,\dots,n\}$, $i<l$, and $\mathcal{H}^1(\partial A_i \cap \partial A_l) > 0$. Then for any $(i,l) \in \mathcal{M}$ we set

$$d_{i,l}(x) = \begin{cases} \operatorname{dist}(x, \partial A_i \cap \partial A_l) & \text{if } x \in A_l, \\ -\operatorname{dist}(x, \partial A_i \cap \partial A_l) & \text{if } x \in A_i; \end{cases}$$
$$L_\varepsilon^{(i,l)} = \{x \in \Omega : |d_{i,l}(x)| \le \rho_\varepsilon\};$$
$$L_{\varepsilon,h}^{(i,l)} = \bigcup \{T \in \mathbf{T}_h : T \cap L_\varepsilon^{(i,l)} \neq \emptyset\}.$$

Using the above definitions, we have for any $(i,l) \in \mathcal{M}$

$$|Q_\varepsilon| = O\left(\varepsilon^2 |\log \varepsilon|^2\right), \qquad |L_\varepsilon^{(i,l)}| = O\left(\varepsilon |\log \varepsilon|\right). \tag{4.2}$$

Then we define the following function u_ε on $\Omega \setminus Q_\varepsilon$:

$$u_\varepsilon(x) = \begin{cases} u_0(x) & \text{if } x \in \Omega \setminus (S_{u_0})_{\rho_\varepsilon}, \\ \gamma_\varepsilon^{(i,l)}(d_{i,l}(x)) & \text{if } x \in L_\varepsilon^{(i,l)} \setminus Q_\varepsilon \quad \forall\ (i,l) \in \mathcal{M}. \end{cases} \tag{4.3}$$

Using the properties (i) and (ii) of Lemma 4.2, we have that u_ε is Lipschitz continuous in $\Omega \setminus Q_\varepsilon$ with $\operatorname{Lip}(u_\varepsilon) = O(\varepsilon^{-1})$. Then u_ε can be extended [6] on the whole Ω as a Lipschitz continuous function with $\operatorname{Lip}(u_\varepsilon) = O(\varepsilon^{-1})$. Moreover, we have $u_\varepsilon \to u_0$ in $L^2(\Omega)$ as $\varepsilon \to 0^+$.

Now we define $u_{\varepsilon,h} = \pi_h(u_\varepsilon)$. Using (2.2), we have $u_{\varepsilon,h} \in \mathcal{D}(E_{\varepsilon,h})$ for any ε small enough. Then, using the properties of the Lagrange interpolation operator [11] and the condition $h = o(\varepsilon |\log \varepsilon|^{-1})$, we have $u_{\varepsilon,h} \to u_0$ in $L^2(\Omega)$ (see [6, proof of Theorem 2.1]). Now we observe that

$$\begin{aligned} E_{\varepsilon,h}(u_{\varepsilon,h}) &= \varepsilon \int_\Omega \frac{|\nabla u_{\varepsilon,h}|^2}{1+\mu_{\varepsilon,h}|\nabla u_{\varepsilon,h}|^2} dx + \frac{1}{\varepsilon}\int_\Omega W(u_{\varepsilon,h}) dx \\ &\le \varepsilon \int_\Omega |\nabla u_{\varepsilon,h}|^2 dx + \frac{1}{\varepsilon}\int_\Omega W(u_{\varepsilon,h}) dx, \end{aligned}$$

and, following the method of [6], we split the functional on the right-hand side as follows:

$$\begin{aligned} E_{\varepsilon,h}(u_{\varepsilon,h}) \le{}& \varepsilon \int_\Omega |\nabla u_\varepsilon|^2 dx + \frac{1}{\varepsilon}\int_\Omega W(u_\varepsilon) dx + \varepsilon \int_\Omega \left(|\nabla u_{\varepsilon,h}|^2 - |\nabla u_\varepsilon|^2\right) dx \\ &+ \frac{1}{\varepsilon}\int_\Omega \left(W(u_{\varepsilon,h}) - W(u_\varepsilon)\right) dx = I_{\varepsilon,h} + II_{\varepsilon,h} + III_{\varepsilon,h}. \end{aligned}$$

First we prove that $\limsup_{\varepsilon \to 0^+} I_{\varepsilon,h} \le E_0(u_0)$. Using (4.3), we have that the contribution of the integrals on the set $\Omega \setminus (S_{u_0})_{\rho_\varepsilon}$ is zero. Moreover, since $\operatorname{Lip}(u_\varepsilon) = O(\varepsilon^{-1})$, using (4.2), we have

$$\varepsilon \int_{Q_\varepsilon} |\nabla u_\varepsilon|^2 dx + \frac{1}{\varepsilon}\int_{Q_\varepsilon} W(u_\varepsilon) dx = O\left(\varepsilon |\log \varepsilon|^2\right). \tag{4.4}$$

For any $(i,l) \in \mathcal{M}$ we set

$$J_\varepsilon^{(i,l)} = \varepsilon \int_{L_\varepsilon^{(i,l)} \setminus Q_\varepsilon} |\nabla u_\varepsilon|^2 dx + \frac{1}{\varepsilon}\int_{L_\varepsilon^{(i,l)} \setminus Q_\varepsilon} W(u_\varepsilon) dx.$$

A standard computation in Γ-convergence theory [6, 17, 20] yields

$$\limsup_{\varepsilon\to 0^+} J_\varepsilon^{(i,l)} \le \mathcal{H}^1(\partial A_i \cap \partial A_l \cap \Omega) \lim_{\varepsilon\to 0^+} \int_{-\infty}^{+\infty} \left[\varepsilon \left(\frac{d\gamma_\varepsilon^{(i,l)}}{dt} \right)^2 + \frac{1}{\varepsilon} W(\gamma_\varepsilon^{(i,l)}) \right] dt,$$

from which, using (4.4) and the property (iii) of Lemma 4.2, we obtain

$$\limsup_{\varepsilon\to 0^+} I_{\varepsilon,h} = \limsup_{\varepsilon\to 0^+} \sum_{(i,l)\in\mathcal{M}} J_\varepsilon^{(i,l)} \le 2 \sum_{\substack{i,l=1 \\ i<l}}^{n} \kappa_{i,l} \mathcal{H}^1(\partial A_i \cap \partial A_l \cap \Omega) = E_0(u_0).$$

Now we prove that the terms $II_{\varepsilon,h}$ and $III_{\varepsilon,h}$ vanish as $\varepsilon \to 0^+$. In [6, proof of Theorem 2.1], the following estimate has been proved:

$$II_{\varepsilon,h} \le \frac{C_1}{\varepsilon} |Q_{\varepsilon,h}| + C_2 h \sum_{(i,l)\in\mathcal{M}} |L_{\varepsilon,h}^{(i,l)} \setminus Q_{\varepsilon,h}| \, \|\nabla^2 u_\varepsilon\|_{L^\infty(L_\varepsilon^{(i,l)} \setminus Q_\varepsilon)} . \tag{4.5}$$

Since $d_{i,l}$ is a distance function from polygonal boundaries, we have $\nabla^2 d_{i,l} = 0$ on $L_\varepsilon^{(i,l)} \setminus Q_\varepsilon$, so that, using (4.3) and the property (ii) of Lemma 4.2, it follows

$$\|\nabla^2 u_\varepsilon\|_{L^\infty(L_\varepsilon^{(i,l)} \setminus Q_\varepsilon)} \le \|d^2 \gamma_\varepsilon^{(i,l)}/dt^2\|_{L^\infty(-\rho_\varepsilon,\rho_\varepsilon)} = O\left(\frac{1}{\varepsilon^2}\right).$$

Then, using (4.2), (4.5), and the condition $h = o(\varepsilon|\log\varepsilon|^{-1})$, we obtain that the term $II_{\varepsilon,h}$ vanishes.

Using again the results obtained in the proof of Theorem 2.1 of [6], the following estimate holds:

$$III_{\varepsilon,h} \le C \frac{h}{\varepsilon} \|\nabla u_\varepsilon\|_{L^\infty(\Omega)} \sum_{(i,l)\in\mathcal{M}} |L_{\varepsilon,h}^{(i,l)}|. \tag{4.6}$$

Since $\|\nabla u_\varepsilon\|_{L^\infty(\Omega)} = O(1/\varepsilon)$, using (4.2), (4.6), and the condition $h = o(\varepsilon|\log\varepsilon|^{-1})$, we obtain that also the term $III_{\varepsilon,h}$ vanishes, concluding the proof. □

Theorem 2.1 then follows from Theorem 3.1, Theorem 4.1, and the fact that the term $\int_\Omega |u(x) - I(x)|^2 \, dx$ is a continuous perturbation.

5. Convergence of minimizers. In this section we prove Theorem 2.2 stated in section 2.

Proof of Theorem 2.2. The existence of a minimizer u_ε of $F_{\varepsilon,h(\varepsilon)}$ is obtained easily since in fact we search for a minimizer in a compact subset of the space V_h which is of finite dimension. Moreover, there exists a constant $C > 0$ such that

$$F_{\varepsilon,h(\varepsilon)}(u_\varepsilon) \le C. \tag{5.1}$$

For any $\varepsilon > 0$, Lemma 3.2 provides a function $v_\varepsilon \in BV(\Omega)$ which satisfies (3.1) and (3.2). Let $\psi_\varepsilon(x) = g(v_\varepsilon(x))$, where g is the function defined by (3.4). We have $\psi_\varepsilon \in BV(\Omega)$ and $S_{\psi_\varepsilon} \subseteq S_{v_\varepsilon}$.

Set $c_M = \int_{-M}^{M} \sqrt{W(t)} dt$. Since $\|v_\varepsilon\|_{L^\infty(\Omega)} \le M$, the following estimates hold for any $\varepsilon > 0$:

$$\|\psi_\varepsilon\|_{L^\infty(\Omega)} < c_M, \qquad |D\psi_\varepsilon|(\Omega) \le \int_\Omega |\nabla\psi_\varepsilon| dx + c_M \mathcal{H}^1(S_{v_\varepsilon}). \tag{5.2}$$

Arguing as in the proof of Theorem 3.1, see (3.6), and using (5.1) and (3.1), we find

$$\begin{aligned} C \geq E_{\varepsilon,h}(u_\varepsilon) &\geq (1-\delta_h)\varepsilon \int_\Omega |\nabla v_\varepsilon|^2 dx + \frac{1}{\varepsilon}\int_\Omega W(v_\varepsilon)dx + 2\kappa_{1,n}\mathcal{H}^1(S_{v_\varepsilon}) \\ &\geq 2(1-\delta_h)^{1/2}\left[\int_\Omega |\nabla \psi_\varepsilon| dx + \kappa_{1,n}\mathcal{H}^1(S_{v_\varepsilon})\right] \\ &\geq 2\frac{\kappa_{1,n}}{c_M}(1-\delta_h)^{1/2}\left[\int_\Omega |\nabla \psi_\varepsilon| dx + c_M\mathcal{H}^1(S_{v_\varepsilon})\right], \end{aligned}$$

where we have used $c_M \geq \kappa_{1,n}$; see (2.2). Using (5.2), it follows that ψ_ε is uniformly bounded in $BV(\Omega)$ with respect to ε, for ε small enough. Then, using the compactness theorem in BV [14], there exists a subsequence $\{\psi_{\varepsilon_j}\}_j$ converging in $L^2(\Omega)$ to a function ψ_0.

Set $u_0 = g^{-1}(\psi_0)$. Since the function g is monotone increasing, using the properties of the potential W, it follows that the inverse function g^{-1} is bounded and uniformly continuous on compact subsets of $\mathbf{R}$. Then

$$v_{\varepsilon_j} = g^{-1} \circ \psi_{\varepsilon_j} \to u_0 \quad \text{in } L^2(\Omega) \text{ as } \varepsilon_j \to 0^+;$$

see also [18, proof of Proposition 3]. Reasoning as in the proof of Theorem 3.1 and using Lemma 3.2, we find that $u_{\varepsilon_j} \to u_0$ in $L^2(\Omega)$ as $\varepsilon_j \to 0^+$.

Hence, the statement of Theorem 2.2 follows from Theorem 2.1 and the property (1.5) of Γ-convergence. □

The condition $h = o(\varepsilon|\log\varepsilon|^{-1})$ deserves a discussion about the numerical implementation of the Γ-convergent approximation. According to such a condition in practical computations, h should be much smaller than ε, so that a very fine mesh has to be used. Nevertheless, this criterion is required only close to the discontinuity curves, so that fine meshing is necessary only across the boundaries of the classification.

For small ε, a minimizer u_ε of $F_{\varepsilon,h(\varepsilon)}$ is essentially constant, with values a_i in large regions corresponding to the sets A_i, whereas it exhibits sharp transitions in narrow strips across the jump set S_{u_0}. Hence an accurate reconstruction of the boundaries can be achieved by resorting to adaptive mesh generation, i.e., refining the mesh locally within the strips surrounding the transitions of the function u_ε. Such an approach has been numerically implemented for functionals with convex perturbation in [7]. In the application to image processing, the input image I has to be opportunely regularized (see next section), the parameter ε should be set equal to the width of the pixel, and the mesh locally refined at the subpixel level by taking $h \ll \varepsilon$. A different approach based on adaptive mesh optimization can be found in [8] for the numerical approximation of the Mumford–Shah functional.

The problem of numerical implementation that arises in the Γ-convergent approximation is related to a crucial problem of image processing: the computation of the length of curves in digitized images. This problem has been addressed in [15], where it has been shown that the computations must become nonlocal as the digitization gets finer and finer. The mesh refined across the discontinuity curves yields asymptotically a nonlocal computation in order to recover the length of such curves in the continuum limit.

6. Numerical integration. In this section we show that the numerical approximation of the lower order terms in the energy via the vertex quadrature rule does not change the results previously obtained (see [5, 6]). More precisely, for any $\varepsilon > 0$

let $I_\varepsilon \in \mathcal{C}_0^\infty(\Omega)$ approximate the function $I \in L^\infty(\Omega)$ so that [5] $I_\varepsilon \to I$ in $L^2(\Omega)$, $\|I_\varepsilon\|_{L^\infty(\Omega)} \le \|I\|_{L^\infty(\Omega)}$, and $\|\nabla I_\varepsilon\|_{L^\infty(\Omega)} \le C/\varepsilon$.

For any $h > 0$ and any $\varepsilon > 0$ we define the functional $\widehat{E}_{\varepsilon,h}$ by

$$\widehat{E}_{\varepsilon,h}(u) = \varepsilon \int_\Omega \frac{|\nabla u|^2}{1 + \mu_{\varepsilon,h}|\nabla u|^2} dx + \frac{1}{\varepsilon} \int_\Omega \pi_h(W(u)) dx$$

and the functional $\widehat{F}_{\varepsilon,h} : L^2(\Omega) \to [0, +\infty]$ by

$$\widehat{F}_{\varepsilon,h}(u) = \begin{cases} \widehat{E}_{\varepsilon,h}(u) + \displaystyle\int_\Omega \pi_h((u - I_\varepsilon)^2) dx & \text{if } u \in \mathcal{D}(\widehat{F}_{\varepsilon,h}), \\ \\ +\infty & \text{elsewhere in } L^2(\Omega), \end{cases}$$

where $\mathcal{D}(\widehat{F}_{\varepsilon,h}) = \{u \in V_h(\Omega) : \|u\|_{L^\infty(\Omega)} \le M\}$. The integrals in $\widehat{F}_{\varepsilon,h}$ can be evaluated via the vertex quadrature rule, which is exact for piecewise linear functions.

Let $u \in V_h(\Omega)$ and let $\mathcal{N}_h$ denote the set of all nodes of the triangulation $\mathbf{T}_h$. Define the function $\tilde{u} \in V_h(\Omega)$ in the following way: for any $q \in \mathcal{N}_h$ set $\tilde{u}(q) = u(q)$ if $|u(q)| \le M$, $\tilde{u}(q) = M$ if $u(q) > M$, and $\tilde{u}(q) = -M$ if $u(q) < -M$. Since the function $\varphi(t) = \frac{t^2}{1+\mu t^2}$ is monotone increasing for $t \ge 0$, by using the property (iv) of the potential W, we have

$$\widehat{E}_{\varepsilon,h}(\tilde{u}) + \int_\Omega \pi_h((\tilde{u} - I_\varepsilon)^2) dx \le \widehat{E}_{\varepsilon,h}(u) + \int_\Omega \pi_h((u - I_\varepsilon)^2) dx.$$

It follows that any absolute minimizer $u_{\varepsilon,h}$ of the above energy belongs to the domain $\mathcal{D}(\widehat{F}_{\varepsilon,h})$.

We prove the following proposition.

PROPOSITION 6.1. *Assume that $h = o(\varepsilon|\log\varepsilon|^{-1})$ and that $\mu_{\varepsilon,h} = o(\varepsilon h)$. Then the family $\{\widehat{F}_{\varepsilon,h}\}_\varepsilon$ Γ-converges to the functional*

$$F_0(u) = \int_\Omega (u - I)^2 dx + E_0(u)$$

in the $L^2(\Omega)$-topology as $\varepsilon \to 0^+$.

Moreover, any family $\{u_{\varepsilon,h}\}_\varepsilon$ of absolute minimizers of $\widehat{F}_{\varepsilon,h}$ is relatively compact in $L^2(\Omega)$, and each of its limit points minimizes F_0.

The proof of the proposition, which is based on Lemma 3.2 and on the estimates obtained in [5, 6], is given in the appendix.

For instance, let be $\Omega = [0,1]^2$, and let $\mathbf{T}_h$ consist of a uniform mesh of triangles formed by dividing Ω into uniform squares of size $h \times h$ and dividing each square into two triangles by cutting along the (1,–1) direction. The resulting discrete scheme is then equivalent to the finite difference method used in [19] for the classification of real images.

Appendix. First we prove Lemma 3.2. Since the proof is essentially the same as that of Proposition 3.3 in [9], we omit the details and underline the slight modifications which are useful in proving Theorem 3.1.

Proof of Lemma 3.2. For any $\delta_h \in (0,1)$ we have

$$\frac{t^2}{1 + \mu_{\varepsilon,h} t^2} \ge \min\left\{(1 - \delta_h)t^2, \frac{\delta_h}{\mu_{\varepsilon,h}}\right\} \qquad \forall t \ge 0. \tag{A.1}$$

Let ∇u_T denote the constant value of the gradient of u on each triangle $T \in \mathbf{T}_h$, and let $\mathbf{T}_h^1 = \{T \in \mathbf{T}_h : (1-\delta_h)|\nabla u_T|^2 > \frac{\delta_h}{\mu_{\varepsilon,h}}\}$. We define the function v in the following way:

$$v(x) = \begin{cases} a_1 & \text{on every } T \in \mathbf{T}_h^1, \\ u(x) & \text{on every } T \in \mathbf{T}_h \setminus \mathbf{T}_h^1. \end{cases}$$

Clearly we have $v \in BV(\Omega)$. Using inequality (A.1), the definition of v, and the same method of proof of Proposition 3.3 of [9], we find

$$E_{\varepsilon,h}(u) \geq (1-\delta_h)\varepsilon \int_\Omega |\nabla v|^2 dx + \frac{1}{\varepsilon}\int_\Omega W(v)dx + \frac{\delta_h \varepsilon}{\mu_{\varepsilon,h}} \sum_{T \in \mathbf{T}_h^1} |T|. \tag{A.2}$$

By the assumptions on the triangulation, the following inequality has been proved in [9]:

$$\sum_{T \in \mathbf{T}_h^1} |T| \geq \frac{1}{6} \cdot h \sin\theta_0 \cdot \mathcal{H}^1(S_v). \tag{A.3}$$

We now set

$$\delta_h = c\frac{\mu_{\varepsilon,h}}{\varepsilon h}, \qquad c = \frac{12\kappa_{1,n}}{\sin\theta_0}.$$

The estimate (3.1) then follows by using (A.2) and (A.3). The inequality (A.2) implies

$$E_{\varepsilon,h}(u) \geq \frac{c}{h} \sum_{T \in \mathbf{T}_h^1} |T|,$$

from which the estimate (3.2) follows. □

Now we prove Lemma 4.2. The extension of the standard construction method for a double-well potential [20] to the case of the potential W does not create particular difficulties, because the wells lie on the real axis. Hence we give the main steps of the construction omitting the details.

Proof of Lemma 4.2. Fix $m \in \{i, \dots, l-1\}$ and consider the solution $\eta^{(m,m+1)}$ of the ordinary differential equation:

$$\frac{d\eta^{(m,m+1)}}{dt} = \sqrt{W(\eta^{(m,m+1)})}, \qquad \eta^{(m,m+1)}(0) = \frac{1}{2}(a_m + a_{m+1}). \tag{A.4}$$

Using the properties of the function W and arguing as in section 1-B of [20], the solution can be defined on all of $\mathbf{R}$, and it is a monotone increasing function such that $a_m < \eta^{(m,m+1)}(t) < a_{m+1}$ for any t, and having the following asymptotic behavior:

$$\lim_{t \to +\infty} \frac{a_{m+1} - \eta^{(m,m+1)}(t)}{\exp(-\sqrt{\alpha_{m+1}}t)} = B_{m,m+1}, \qquad \lim_{t \to -\infty} \frac{\eta^{(m,m+1)}(t) - a_m}{\exp(\sqrt{\alpha_m}t)} = C_{m,m+1}, \tag{A.5}$$

where $2\alpha_m = W''(a_m)$, $2\alpha_{m+1} = W''(a_{m+1})$, and $B_{m,m+1}$, $C_{m,m+1}$ are positive constants. Now we set $t_{m,\varepsilon} = \varepsilon|\log\varepsilon|/\sqrt{\alpha_m}$ and we define $\eta_\varepsilon^{(m,m+1)} : \mathbf{R} \to \mathbf{R}$ in the

following way:

$$\eta_\varepsilon^{(m,m+1)}(t) = \begin{cases} \eta^{(m,m+1)}\left(\dfrac{t}{\varepsilon}\right) & \text{if } -t_{m,\varepsilon} \le t \le t_{m+1,\varepsilon}, \\ q_{m,\varepsilon}(t) & \text{if } -2t_{m,\varepsilon} \le t \le -t_{m,\varepsilon}, \\ a_m & \text{if } t \le -2t_{m,\varepsilon}, \\ p_{m+1,\varepsilon}(t) & \text{if } t_{m+1,\varepsilon} \le t \le 2t_{m+1,\varepsilon}, \\ a_{m+1} & \text{if} t \ge 2t_{m+1,\varepsilon}, \end{cases}$$

where $q_{m,\varepsilon}$, $p_{m+1,\varepsilon}$ are cubic polynomials chosen in such a way that $\eta_\varepsilon^{(m,m+1)} \in \mathcal{C}^1(\mathbf{R})$ for any $\varepsilon > 0$. One can verify that

$$\|d\eta_\varepsilon^{(m,m+1)}/dt\|_{L^\infty(-2t_{m,\varepsilon},2t_{m+1,\varepsilon})} = O\left(\frac{1}{\varepsilon}\right), \tag{A.6}$$

$$\|d^2\eta_\varepsilon^{(m,m+1)}/dt^2\|_{L^\infty(-2t_{m,\varepsilon},2t_{m+1,\varepsilon})} = O\left(\frac{1}{\varepsilon^2}\right). \tag{A.7}$$

Now we set

$$I_\varepsilon^{(m,m+1)} = \int_{-\infty}^{+\infty} \left[\varepsilon\left(\frac{d\eta_\varepsilon^{(m,m+1)}}{dt}\right)^2 + \frac{1}{\varepsilon}W(\eta_\varepsilon^{(m,m+1)})\right] dt.$$

By using (A.4) and (A.5), a standard computation of Γ-convergence applied to phase transition problems [20] yields

$$\lim_{\varepsilon\to 0^+} I_\varepsilon^{(m,m+1)} = 2\kappa_{m,m+1}. \tag{A.8}$$

The function $\gamma_\varepsilon^{(i,l)} : \mathbf{R} \to \mathbf{R}$ is then constructed by means of translations of the functions $\eta_\varepsilon^{(m,m+1)}$ in such a way that the smooth transitions between the values a_m, a_{m+1} do not overlap. We set

$$\rho_\varepsilon = \varepsilon|\log\varepsilon| \sum_{m=1}^{n-1} \left(\frac{1}{\sqrt{\alpha_m}} + \frac{1}{\sqrt{\alpha_{m+1}}}\right),$$

and we define a partition of the interval $[-\rho_\varepsilon, \rho_\varepsilon]$ into closed subintervals with disjoint interiors:

$$[-\rho_\varepsilon, \rho_\varepsilon] = \bigcup_{m=i}^{l-1} [\xi_{m,\varepsilon}, \xi_{m+1,\varepsilon}], \quad \xi_{m+1,\varepsilon} - \xi_{m,\varepsilon} = 2(t_{m,\varepsilon} + t_{m+1,\varepsilon}) \quad \text{for} i \le m < l-1,$$

with $\xi_{i,\varepsilon} = -\rho_\varepsilon$ and $\xi_{l,\varepsilon} = \rho_\varepsilon$. In each subinterval $[\xi_{m,\varepsilon}, \xi_{m+1,\varepsilon}]$ we set $\gamma_\varepsilon^{(i,l)}$ equal to $\eta_\varepsilon^{(m,m+1)}$ translated in such a way that $\gamma_\varepsilon^{(i,l)}(\xi_{m,\varepsilon}) = a_m$ and $\gamma_\varepsilon^{(i,l)}(\xi_{m+1,\varepsilon}) = a_{m+1}$. It is easy to check that such a function $\gamma_\varepsilon^{(i,l)}$ satisfies the property (i) of the statement of the lemma. The property (ii) then follows from (A.6) and (A.7).

By the construction of $\gamma_\varepsilon^{(i,l)}$, using (A.8), we obtain

$$\begin{aligned} \lim_{\varepsilon\to 0^+} \int_{-\infty}^{+\infty} \left[\varepsilon\left(\frac{d\gamma_\varepsilon^{(i,l)}}{dt}\right)^2 + \frac{1}{\varepsilon}W(\gamma_\varepsilon^{(i,l)})\right] dt &= \lim_{\varepsilon\to 0^+} \sum_{m=i}^{l-1} I_\varepsilon^{(m,m+1)} \\ &= \sum_{m=i}^{l-1} \lim_{\varepsilon\to 0^+} I_\varepsilon^{(m,m+1)} = \sum_{m=i}^{l-1} 2\kappa_{m,m+1} = 2\kappa_{i,l}, \end{aligned}$$

and we have proved the property (iii). ∎

Proof of Proposition 6.1. We prove first the lower inequality: for every function $u_0 \in L^2(\Omega)$ and for every sequence $\{u_{\varepsilon,h}\}_\varepsilon \subset L^2(\Omega)$ converging to u_0 in $L^2(\Omega)$ we have

$$\liminf_{\varepsilon\to 0^+} \widehat{F}_{\varepsilon,h}(u_{\varepsilon,h}) \geq \int_\Omega (u_0 - I)^2 dx + E_0(u_0). \tag{A.9}$$

We can suppose, possibly extracting a subsequence, that $\{u_{\varepsilon,h}\}_\varepsilon \subset \mathcal{D}(\widehat{F}_{\varepsilon,h})$ and $\liminf_{\varepsilon\to 0^+} \widehat{F}_{\varepsilon,h}(u_{\varepsilon,h}) = \lim_{\varepsilon\to 0^+} \widehat{F}_{\varepsilon,h}(u_{\varepsilon,h}) = L < +\infty$; otherwise (A.9) is obvious. Following the method of [5, 6], we split $\widehat{F}_{\varepsilon,h}(u_{\varepsilon,h})$ as follows:

$$\begin{aligned}\widehat{F}_{\varepsilon,h}(u_{\varepsilon,h}) &= F_{\varepsilon,h}(u_{\varepsilon,h}) + \frac{1}{\varepsilon}\int_\Omega [\pi_h(W(u_{\varepsilon,h})) - W(u_{\varepsilon,h})]dx \\ &\quad + \int_\Omega [\pi_h((u_{\varepsilon,h} - I_\varepsilon)^2) - (u_{\varepsilon,h} - I)^2]dx = F_{\varepsilon,h}(u_{\varepsilon,h}) + I_{\varepsilon,h} + II_{\varepsilon,h}.\end{aligned} \tag{A.10}$$

In view of Theorem 2.1, in order to show (A.9) it will be enough to prove that $\lim_{\varepsilon\to 0^+} I_{\varepsilon,h} = \lim_{\varepsilon\to 0^+} II_{\varepsilon,h} = 0$. Arguing as in the proof of (A.2), we have for a fixed $\delta \in (0,1)$ and ε small enough

$$L + 1 \geq (1-\delta)\varepsilon \int_{\mathcal{A}_{\varepsilon,h}} |\nabla u_{\varepsilon,h}|^2 dx + \frac{\delta\varepsilon}{\mu_{\varepsilon,h}} |\Omega \setminus \mathcal{A}_{\varepsilon,h}|, \tag{A.11}$$

where $\mathcal{A}_{\varepsilon,h} = \bigcup_{T\in\mathbf{T}_h\setminus\mathbf{T}_h^1} T$. Using the estimates obtained in the proof of Theorem 2.1 of [6] and the properties of the potential W, we have

$$|I_{\varepsilon,h}| \leq C_1 \mathrm{Lip}(W')\frac{h^2}{\varepsilon}\int_{\mathcal{A}_{\varepsilon,h}} |\nabla u_{\varepsilon,h}|^2 dx + \frac{C_2}{\varepsilon}|\Omega \setminus \mathcal{A}_{\varepsilon,h}|. \tag{A.12}$$

Hence, using (A.11), (A.12), and the conditions $h = o(\varepsilon|\log\varepsilon|^{-1})$ and $\mu_{\varepsilon,h} = o(\varepsilon h)$, we find that the term $I_{\varepsilon,h}$ vanishes. Analogously we have

$$|II_{\varepsilon,h}| \leq \int_{\mathcal{A}_{\varepsilon,h}} |\pi_h((u_{\varepsilon,h} - I_\varepsilon)^2) - (u_{\varepsilon,h} - I)^2| dx + C_2|\Omega \setminus \mathcal{A}_{\varepsilon,h}|. \tag{A.13}$$

Arguing as in the proof of Theorem 4.1 of [5], it follows that the first term in the right-hand side of (A.13) converges to zero. Hence, using (A.11), we find that also the term $II_{\varepsilon,h}$ vanishes. This concludes the proof of (A.9).

We now prove the upper inequality. Let $u_0 \in BV(\Omega;\{a_1,\dots,a_n\})$, and let $\{u_{\varepsilon,h}\}_\varepsilon$ be the sequence converging to u_0 constructed in the proof of Theorem 4.1 and such that $\limsup_{\varepsilon\to 0^+} E_{\varepsilon,h}(u_{\varepsilon,h}) \leq E_0(u_0)$. The proof of Theorem 4.1 shows that the following estimate holds:

$$\int_\Omega |\nabla u_{\varepsilon,h}|^2 dx \leq \frac{C}{\varepsilon}. \tag{A.14}$$

Let us split $\widehat{F}_{\varepsilon,h}(u_{\varepsilon,h})$ as in (A.10). Then the results obtained in the proof of the lower inequality and the estimate (A.14) guarantee that the terms $I_{\varepsilon,h}$ and $II_{\varepsilon,h}$ again vanish as $\varepsilon \to 0^+$. Hence

$$\limsup_{\varepsilon\to 0^+} \widehat{F}_{\varepsilon,h}(u_{\varepsilon,h}) = \limsup_{\varepsilon\to 0^+} F_{\varepsilon,h}(u_{\varepsilon,h}) \leq \int_\Omega (u_0 - I)^2 dx + E_0(u_0),$$

which yields the upper inequality and concludes the proof of Γ-convergence.

Finally, the convergence of the minimizers follows by splitting again the functional $\widehat{F}_{\varepsilon,h}(u_{\varepsilon,h})$ as in (A.10) and arguing as in the proof of Theorem 2.2. □

REFERENCES

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Oxford University Press, New York, 2000.

[2] S. Angenent and M. E. Gurtin, *Multiphase thermomechanics with interfacial structure* II. *Evolution of an isothermal interface*, Arch. Ration. Mech. Anal., 108 (1989), pp. 323–391.

[3] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing*, Appl. Math. Sci. 147, Springer-Verlag, New York, 2002.

[4] S. Baldo, *Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluids*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 67–90.

[5] G. Bellettini and A. Coscia, *Discrete approximation of a free discontinuity problem*, Numer. Funct. Anal. Optim., 15 (1994), pp. 201–224.

[6] G. Bellettini, M. Paolini, and C. Verdi, *Γ-convergence of discrete approximations to interfaces with prescribed mean curvature*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 1 (1990), pp. 317–328.

[7] G. Bellettini, M. Paolini, and C. Verdi, *Numerical minimization of geometrical type problems related to calculus of variations*, Calcolo, 27 (1990), pp. 251–278.

[8] B. Bourdin and A. Chambolle, *Implementation of an adaptive finite-element approximation of the Mumford-Shah functional*, Numer. Math., 85 (2000), pp. 609–646.

[9] A. Chambolle and G. Dal Maso, *Discrete approximation of the Mumford-Shah functional in dimension two*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 651–672.

[10] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, *Deterministic edge-preserving regularization in computed imaging*, IEEE Trans. Image Process., 6 (1997), pp. 298–311.

[11] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[12] G. Dal Maso, *An Introduction to Γ-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Boston, 1993.

[13] I. Fonseca and L. Tartar, *The gradient theory of phase transitions for systems with two potential wells*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 89–102.

[14] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Verlag, Basel, Switzerland, 1984.

[15] S. R. Kulkarni, S. K. Mitter, T. J. Richardson, and J. N. Tsitsiklis, *Local versus nonlocal computation of length of digitized curves*, IEEE Trans. Pattern Analysis and Machine Intelligence, 16 (1994), pp. 711–718.

[16] A. Lorette, X. Descombes, and J. Zerubia, *Texture analysis through a Markovian modelling and fuzzy classification: Application to urban area extraction from satellite images*, Internat. J. Computer Vision, 36 (2000), pp. 221–236.

[17] L. Modica and S. Mortola, *Un esempio di* Γ^-*-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.

[18] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Ration. Mech. Anal., 98 (1987), pp. 123–142.

[19] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia, *A variational model for image classification and restoration*, IEEE Trans. Pattern Analysis and Machine Intelligence, 22 (2000), pp. 460–472.

[20] P. Sternberg, *The effect of a singular perturbation on nonconvex variational problems*, Arch. Ration. Mech. Anal., 101 (1988), pp. 209–260.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1146–1166

POINTWISE ERROR ESTIMATES OF DISCONTINUOUS GALERKIN METHODS WITH PENALTY FOR SECOND-ORDER ELLIPTIC PROBLEMS*

ZHANGXIN CHEN† AND HONGSEN CHEN‡

Abstract. In this paper discontinuous Galerkin methods with penalty for solving second-order elliptic problems are considered. Error estimates are studied for these methods. In particular, optimal localized pointwise error estimates are established on quasi-uniform grids in R^N $(N \geq 2)$.

Key words. discontinuous Galerkin method, pointwise error estimate, elliptic problem

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

DOI. 10.1137/S0036142903421527

1. Introduction.

Considerable attention has been recently paid to the development and analysis of discontinuous Galerkin (DG) methods for second-order elliptic problems. The DG methods with penalty are nowadays widely used in many problems and applications [11]. In a DG method the approximation space typically consists of discontinuous piecewise polynomials, with boundary conditions and continuity on interelement boundaries weakly imposed through a bilinear form. For second-order elliptic problems these methods trace back originally to the work of Nitsche [15], which was further developed and analyzed by Douglas and Dupont [12], Baker [4], Wheeler [23], and Arnold [2]. More recent developments of the DG methods for elliptic problems can be found in Oden, Babuška, and Baumann [16], Brezzi et al. [5], Castillo et al. [6], Rivière, Wheeler, and Girault [18], Chen and Chen [7], and Chen, Chen, and Li [8] (also see Cockburn, Karniadakis, and Shu [11]). We refer to Chen [9] and Arnold et al. [3] for a review on the relationships and properties of different DG methods. Optimal error estimates for the DG methods with penalty have been obtained in [2, 18] in energy and L^2 norms.

The aim of this paper is to derive localized pointwise error estimates for the DG methods. The results in this paper represent an improvement and extension to those obtained by Kanschat and Rannacher [14]. Our analysis is based on the technique developed by Schatz [19, 20] for the standard continuous finite element methods for second-order elliptic problems. Therefore, the new pointwise error estimates obtained in this paper indicate a more localized dependence of the error on the derivatives of the true solution. We will give a more detailed description on our results and on the difference between our results and those of [14]. The error estimate of optimal order for the DG methods with penalty takes the form

$$\|u - u_h\|_{L^2(\Omega)} + h|||u - u_h|||_{H_h^1(\Omega)} \leq Ch^{1+r}\|u\|_{H^{1+r}(\Omega)},$$

where u and u_h are the true and approximate solutions, respectively; r is the order of polynomials used in the finite element space; and $|||\cdot|||_{H_h^1(\Omega)}$ is a special energy

*Received by the editors January 13, 2003; accepted for publication (in revised form) January 6, 2004; published electronically September 18, 2004. This research is supported in part by NSF grants DMS-9972147 and INT-9901498.

http://www.siam.org/journals/sinum/42-3/42152.html

†Research Center for Science, Xi'an Jiaotong University, Xi'an 710049, China, and Department of Mathematics, Southern Methodist University, Dallas, TX 75275-0156 (zchen@mail.smu.edu).

‡Department of Mathematics, University of Wyoming, Laramie, WY 82070 (hchen@uwyo.edu).

norm involving measures on the jumps of discontinuous functions on the boundaries of interelements. In [14] the local pointwise error estimate was obtained:

$$|(u-u_h)(z)| \le Ch^2 \ln h^{-1} \|D^2 u\|_{L^\infty(B_a)} + Ch^2 \|D^2 u\|_{L^2(\Omega)}, \tag{1.1}$$

where B_a is a ball centered at $z \in \Omega$ and with a fixed radius $a = \mathcal{O}(1)$. Estimate (1.1) was proved for the model Laplacian equation in a two-dimensional domain and for the space of (discontinuous) piecewise linear functions. The proof of (1.1) in [14] relies on the technique of the discrete Green's function developed by Frehse and Rannacher [13] and Rannacher and Scott [17]. In this paper we derive pointwise error estimates for general second-order elliptic problems defined in a domain of R^N and for general finite element spaces. Moreover, our analysis is based on the technique developed by Schatz [19, 20] and Schatz and Wahlbin [21] for the standard conforming (continuous) finite element methods. Therefore, more localized and sharper results are obtained in this paper. In fact, the new pointwise error estimates obtained in this paper, which are for the symmetric DG method, are the following:

$$\begin{aligned} &|(u-u_h)(z)| \le Ch^{1+r} \left(\ln h^{-1}\right)^{\bar{s}} \|u\|_{W^{1+r,\infty}(\Omega),z,s}, \quad 0 \le s \le r-1, \\ &|\nabla(u-u_h)(z)| \le Ch^{r} \left(\ln h^{-1}\right)^{\bar{\bar{s}}} \|u\|_{W^{1+r,\infty}(\Omega),z,s}, \quad 0 \le s \le r, \end{aligned} \tag{1.2}$$

where $\bar{s} = 0$ if $0 \le s < r-1$ and $\bar{s} = 1$ if $s = r-1$; $\bar{\bar{s}} = 0$ if $0 \le s < r$ and $\bar{\bar{s}} = 1$ if $s = r$. The norm $\|\cdot\|_{W^{1+r,\infty}(\Omega),z,s}$ is a weighted Sobolev norm with the weight function $\sigma_{z,h}^s = (h/(|z-x|+h))^s$. The gradient operator in (1.2) is understood elementwise. According to the estimates in (1.2), the higher order of the finite element approximation, the more localized dependence of the errors on the true solution. As special consequences of (1.2), we have

$$\begin{aligned} &\|u-u_h\|_{L^\infty(\Omega)} \le Ch^{1+r} \left(\ln h^{-1}\right)^{\bar{r}} \|u\|_{W^{1+r,\infty}(\Omega)}, \\ &\|\nabla(u-u_h)\|_{L^\infty(\Omega)} \le Ch^{r} \|u\|_{W^{1+r,\infty}(\Omega)}, \end{aligned}$$

where $\bar{r} = 0$ if $r > 1$ and $\bar{r} = 1$ if $r = 1$. The pointwise error estimates obtained in this paper will be bounded by global estimates. Local bounds will be studied in a forthcoming paper.

2. Preliminaries. We consider the following homogeneous Dirichlet boundary value problem:

$$\begin{aligned} \mathcal{L}u \equiv -\sum_{i,j=1}^{N} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{N} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x) \quad &\text{in } \Omega, \\ u = 0 \quad &\text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $\Omega \subset R^N$ $(N \ge 2)$ is a bounded domain with smooth boundary $\partial\Omega$ and f is a given function. For the sake of simplicity, we assume that the coefficients a_{ij}, b_i, and c are in $C^\infty(\Omega)$ and the operator $\mathcal{L}$ is uniformly elliptic in the sense that there exists a constant $C_L > 0$ such that

$$C_L \sum_{i=1}^{N} \zeta_i^2 \le \sum_{i,j=1}^{N} a_{ij}(x)\zeta_i\zeta_j \qquad \forall\ \zeta \in R^N \ \text{ and } \ x \in \Omega. \tag{2.2}$$

We will use the standard notation for the Sobolev spaces and their norms. For any open subset $D \subset \Omega$, nonnegative integer ℓ, and real number $1 \le p \le \infty$, denote the Sobolev spaces by $W^{\ell,p}(D) = \{v :\ \|v\|_{W^{\ell,p}(D)} < \infty\}$, with

$$\|v\|^p_{W^{\ell,p}(D)} = \sum_{i=0}^{\ell} |v|^p_{W^{i,p}(D)},\ |v|^p_{W^{i,p}(D)} = \sum_{|\alpha|=i} \int_D \left| \frac{\partial^\alpha v(x)}{\partial x^\alpha} \right|^p \, dx.$$

(When $p = \infty$, a standard modification applies to these norms [1].) Let $W_0^{\ell,p}(D)$ be the completion of $C_0^\infty(D)$ according to the norm $\|\cdot\|_{W^{\ell,p}(D)}$, where $C_0^\infty(D)$ represents the space of functions with continuous derivatives of arbitrary order and compact support in D. We also adopt the usual notation for the Hilbert H^ℓ and L^p spaces:

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

Denote by $(\cdot,\cdot)$ the inner product in $L^2(\Omega)$: $(u,v) = \int_\Omega u(x)v(x)\,dx$. For $\ell \geq 0$, the negative norm $\|\cdot\|_{H^{-\ell}(D)}$ is defined as follows:

$$\|v\|_{H^{-\ell}(D)} = \sup_{\varphi \in C_0^\infty(D),\ \|\varphi\|_{H^\ell(D)}=1} (v,\varphi).$$

To introduce the DG methods, let $\mathcal{J}_h$ denote a partition of the domain Ω into a finite number N_h of open subdomains K_j, $j = 1, 2, \ldots, N_h$, such that

$$\bar{\Omega} = \bigcup_{K_j \in \mathcal{J}_h} \bar{K}_j \quad \text{and} \quad K_i \cap K_j = \emptyset \quad \text{if } i \neq j.$$

We assume that the partition $\mathcal{J}_h$ is shape-regular. To be more precise, let $B_\theta(z)$ denote the ball centered at $z \in R^N$ and with radius θ. We set

$$h_K = \operatorname{diam}(K), \quad \theta_K = \max\{\theta :\ B_\theta(z) \subset K, z \in K\}, \quad h = \max_{K \in \mathcal{J}_h} h_K.$$

There are constants $C_1 > 0$ and $C_2 > 0$ independent of h and $K \in \mathcal{J}_h$ such that

$$h_K/\theta_K \leq C_1, \quad h \leq C_2 h_K, \qquad \forall\ K \in \mathcal{J}_h.$$

Note that the mesh $\mathcal{J}_h$ is not required to be conforming; i.e., a vertex of an element may lie on the boundary of another element and there may be hanging nodes.

Furthermore, let Γ_h denote the set of $(N-1)$-dimensional open subsets e_j, $j = 1, 2, \ldots, N_h^e$, such that

$$\bigcup_{j=1}^{N_h} \partial K_j = \bigcup_{j=1}^{N_h^e} \bar{e}_j \quad \text{and} \quad e_i \cap e_j = \emptyset \quad \text{if} \quad i \neq j,$$

and let

$$\Gamma_h^0 = \{e \in \Gamma_h :\ e \cap \partial\Omega = \emptyset\}.$$

We assume that for each $e \in \Gamma_h^0$, $e \subset \partial K \cap \partial K'$ for some K, $K' \in \mathcal{J}_h$. For each $e \in \Gamma_h$, we define $h_e = (h_K + h_{K'})/2$ if $e \subset \partial K \cap \partial K'$ and $h_e = h_K$ if $e \in \partial K \cap \partial\Omega$, and for $K \in \mathcal{J}_h$, let $n_K \in R^N$ denote the unit outward normal vector to ∂K.

Define the discontinuous Sobolev space

$$W_h^{\ell,p}(D) = \{v :\ v \in W^{\ell,p}(K \cap D) \ \text{ for each } K \in \mathcal{J}_h \ \text{ and } \ \|v\|_{W_h^{\ell,p}(D)} < \infty\},$$

equipped with the norm

$$\|v\|^p_{W_h^{\ell,p}(D)} = \sum_{i=0}^{\ell} |v|^p_{W_h^{i,p}(D)},\ |v|^p_{W_h^{i,p}(D)} = \sum_{K\in\mathcal{J}_h} |v|^p_{W^{i,p}(K\cap D)}.$$

When $p = 2$, we set $H_h^\ell(D) = W_h^{\ell,2}(D)$.

For any $v \in H_h^\ell(\Omega)$ $(\ell > 1/2)$, we define its average and jump on an intersection of any two elements as follows: For any $e \in \Gamma_h$, we define

$$\{\!\{v\}\!\}|_e = \begin{cases} \dfrac{1}{2}(v|_K + v|_{K'}) & \text{if } e \in \Gamma_h^0 \text{ and } e \subset \partial K \cap \partial K', \\ v|_K & \text{if } e \in \Gamma_h\backslash\Gamma_h^0 \text{ and } e \in \partial K, \end{cases}$$

$$[\![v]\!]|_e = \begin{cases} v|_K\, n_K + v|_{K'}\, n_{K'} & \text{if } e \in \Gamma_h^0 \text{ and } e \subset \partial K \cap \partial K', \\ v|_K\, n_K & \text{if } e \in \Gamma_h\backslash\Gamma_h^0 \text{ and } e \in \partial K. \end{cases}$$

It is clear that $[\![v]\!] \in R^N$ is a vector with components $[\![v]\!]_i = v|_K\, n_{K,i} + v|_{K'}\, n_{K',i}$ if $e \subset \partial K \cap \partial K'$ and $[\![v]\!]_i = v|_K\, n_{K,i}$ if $e \in \partial K \cap (\Gamma_h\backslash\Gamma_h^0)$ for $1 \le i \le N$.

Multiplying (2.1) by $v \in H_h^2(\Omega)$, integrating over $K \in \mathcal{J}_h$, using the continuity of u and its fluxes across each $e \in \Gamma_h^0$, and adding on all $K \in \mathcal{J}_h$, it follows that

$$a(u, v) - b(u, v) = (f, v), \tag{2.3}$$

where

$$a(u,v) = \sum_{K\in\mathcal{J}_h} \int_K \left(\sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + \sum_{i=1}^N b_i \frac{\partial u}{\partial x_i} v + cuv \right) dx,$$

$$b(u,v) = \sum_{e\in\Gamma_h} \int_e \sum_{i,j=1}^N \left\{\!\!\left\{ a_{ij} \frac{\partial u}{\partial x_i} \right\}\!\!\right\} [\![v]\!]_j \, ds.$$

Here and throughout this paper, a differential operator is defined elementwise when it is not valid globally. To enforce stability of the DG methods, a term $b(v, u)$ involving continuity of the true solution at interelement interfaces and a penalty term $\lambda(u, v)$ are added to the left-hand side of (2.3) to obtain the bilinear form $A(\cdot, \cdot)$:

$$A(u, v) = a(u, v) - b(u, v) - \tau b(v, u) + \lambda(u, v), \tag{2.4}$$

where $\tau = 1$ or -1 and

$$\lambda(u, v) = \sum_{e\in\Gamma_h} \frac{\lambda_e}{h_e} \int_e [\![u]\!][\![v]\!] \, ds. \tag{2.5}$$

Here, for each $e \in \Gamma_h$, λ_e is a real number satisfying

$$C_\lambda \le \lambda_e < \infty \qquad \forall\ e \in \mathcal{J}_h. \tag{2.6}$$

The constant C_λ in (2.6) will be determined later and should be large enough to enforce the stability of the DG methods.

Let $S^h \subset W_h^{2,\infty}(\Omega)$ be a finite dimensional subspace. For simplicity, we assume that S^h consists of piecewise polynomials of degree $r \geq 1$:

$$S^h = \{v \in L^\infty(\Omega) : \ v|_K \in \mathcal{S}(K), \ K \in \mathcal{J}_h\},$$

where $\mathcal{P}_r(K) \subset \mathcal{S}(K) \subset \mathcal{P}_{r_1}(K)$ for some integers $1 \leq r \leq r_1$ and $\mathcal{P}_r(K)$ denotes the set on K of all polynomials of degree less than or equal to r. We define the finite element approximation $u_h \in S^h$ of $u \in H_0^1(\Omega)$ to be the solution of

$$A(u_h, v) = (f, v) \qquad \forall \ v \in S^h. \tag{2.7}$$

Noting that the bilinear form $A(\cdot, \cdot)$ is *consistent* with (2.1) in the sense that

$$A(u, v) = (f, v) \qquad \forall \ v \in H_h^2(\Omega), \tag{2.8}$$

we have the following error equation:

$$A(u - u_h, v) = 0 \qquad \forall \ v \in S^h. \tag{2.9}$$

The cases $\tau = 1$ and $\tau = -1$ correspond to the symmetric and nonsymmetric interior penalty DG methods, respectively.

For any $D \subset \Omega$, we will need a special norm $|||\cdot|||_{W_h^{1,p}(D)}$ defined by

$$\begin{aligned} |||v|||_{W_h^{1,p}(D)} = \|v\|_{W_h^{1,p}(D)} &+ \left(\sum_{e \in \Gamma_h} h_e^{1-p} \int_{e \cap D} |[\![v]\!]|^p \ ds \right)^{1/p} \\ &+ \left(\sum_{e \in \Gamma_h} h_e \int_{e \cap D} \sum_{i=1}^N \left| \left\{\!\!\left\{ \frac{\partial v}{\partial x_i} \right\}\!\!\right\} \right|^p \ ds \right)^{1/p}, \end{aligned} \tag{2.10}$$

where $1 \leq p < \infty$. For $p = \infty$, we have the modification

$$\begin{aligned} |||v|||_{W_h^{1,\infty}(D)} = \|v\|_{W_h^{1,\infty}(D)} &+ \max_{e \in \Gamma_h} h_e^{-1} \|[\![v]\!]\|_{L^\infty(e \cap D)} \\ &+ \max_{e \in \Gamma_h} \|\{\!\{\nabla v\}\!\}\|_{L^\infty(e \cap D)}. \end{aligned} \tag{2.11}$$

When $p = 2$, we use $|||v|||_{H_h^1(D)} = |||v|||_{W_h^{1,2}(D)}$. We also need a weighted analogue of $|||\cdot|||_{W_h^{1,p}(D)}$. To this end, following Schatz [19] we introduce the weight function

$$\sigma_{z,h}^s(x) = \left(\frac{h}{|x - z| + h} \right)^s. \tag{2.12}$$

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

$$\begin{aligned} |||v|||_{W_h^{1,p}(D),z,s}^p = \|v\|_{W_h^{1,p}(D),z,s}^p &+ \sum_{e \in \Gamma_h} h_e^{1-p} \int_{e \cap D} |\sigma_{z,h}^s [\![v]\!]|^p \ ds \\ &+ \sum_{e \in \Gamma_h} h_e \int_{e \cap D} \sum_{i=1}^N \left| \sigma_{z,h}^s \left\{\!\!\left\{ \frac{\partial v}{\partial x_i} \right\}\!\!\right\} \right|^p \ ds, \end{aligned} \tag{2.13}$$

where $\|v\|_{W_h^{1,p}(D),z,s}^p = \|\sigma_{z,h}^s v\|_{L^p(D)}^p + \|\sigma_{z,h}^s \nabla v\|_{L^p(D)}^p$. For $p = \infty$, a modification similar to (2.11) can be made.

Next, we will collect some well-known results about approximation properties, inverse properties, and superapproximation properties. For the proof of Propositions 2.1 and 2.2 below, refer to [10]. The proof of Proposition 2.4 can be found in Schatz and Wahlbin [21]. Below we assume that $\kappa > 0$ is a fixed constant.

PROPOSITION 2.1. *Let $1 \le p \le \infty$ and $0 \le i \le 1 \le j \le 1 + r$.*

(i) *If $v \in W^{j,p}(K)$, then there exists a $\chi \in S^h$ such that for any $K \in \mathcal{J}_h$,*

$$\|v - \chi\|_{W^{i,p}(K)} \le Ch^{j-i}\|v\|_{W^{j,p}(K)},$$

where the constant C is independent of v, h, and K.

(ii) *If $D_1 \subset D_2$ satisfies $dist(D_1, \partial D_2 \backslash \partial\Omega) \ge \kappa h$ and $v \in W_h^{j,p}(D_2)$, then there exists a $\chi \in S^h(D_2)$ such that*

$$\|v - \chi\|_{W_h^{i,p}(D_1)} \le Ch^{j-i}\|v\|_{W_h^{j,p}(D_2)},$$

where the constant C is independent of v, h, D_1, and D_2.

PROPOSITION 2.2. *Let $1 \le q \le p \le \infty$ and $0 \le i \le j \le 1 + r$.*

(i) *If $v \in S^h$, then*

$$\|v\|_{W^{i,p}(K)} \le Ch^{i-j+N(1/q-1/p)}\|v\|_{W^{j,q}(K)}, \quad K \in \mathcal{J}_h,$$

where the constant C is independent of v, h, and K.

(ii) *If $D_1 \subset D_2$ with $dist(D_1, \partial D_2 \backslash \partial\Omega) \ge \kappa h$ and $v \in S^h(D_2)$, then*

$$\|v\|_{W_h^{i,p}(D_1)} \le Ch^{i-j+N(1/q-1/p)}\|v\|_{W_h^{j,q}(D_2)}$$

and

$$|||v|||_{H_h^1(D_1)} \le Ch^{-1}\|v\|_{L^2(D_2)},$$

where the constant C is independent of v, h, D_1, and D_2.

A superapproximation property in $|||\cdot|||_{H_h^1}$ is needed in the proof of interior error estimates. To state it, we need additional notation. For any subsets $D_1 \subset \Omega$ and $D_2 \subset \Omega$, by $D_1 \Subset D_2$ we mean $D_1 \subset D_2$ and $\mathrm{dist}(D_1, \partial D_2 \backslash \partial\Omega) > 0$. Moreover, for any $D \subset \Omega$, $C_<^\infty(D)$ denotes the subspace of $C^\infty(D)$ defined by

$$C_<^\infty(D) = \{v \in C^\infty(D) : \ \mathrm{supp}(v) \Subset D\}.$$

PROPOSITION 2.3. *Let $D_0 \subset D_1 \subset D_2 \subset \Omega$, with the conditions $dist(D_0, \partial D_1 \backslash \partial\Omega) \ge \kappa h$, $dist(D_1, \partial D_2 \backslash \partial\Omega) \ge \kappa h$, and $\omega \in C_<^\infty(D_1)$. Then, for any $v \in S^h(D_2)$ there exists a $\chi \in S^h(D_2)$ such that* $\mathrm{supp}(\chi) \Subset D_2$ *and*

$$|||\omega v - \chi|||_{H_h^1(D_2)} \le Ch|||v|||_{H_h^1(D_2)}.$$

Proof. It suffices to show this property on each element $K \in \mathcal{J}_h$:

$$|||\omega v - \chi|||_{H^1(K)} \le Ch|||v|||_{H^1(\hat{K})},$$

which is apparently true (see Ciarlet [10] and Schatz and Wahlbin [21]). Here $\hat{K}$ denotes the union of elements whose boundary intersects with ∂K. ∎

The last proposition is about a scaling property. This property is used to obtain an explicit dependence on the distance between two subdomains in the local error estimates, which is crucial for the proof of pointwise error estimates.

PROPOSITION 2.4. *Let $x_0 \in \bar{\Omega}$ and $d \ge kh$. The linear transformation $y = (x - x_0)/d$ maps the set $B_d(x_0) = \{x \in \Omega : \ |x - x_0| < d\}$ into a new set $\hat{B}_1(x_0)$ and the space $S^h(B_d(x_0))$ into a new space $\hat{S}^{h/d}(\hat{B}_1(x_0))$. Moreover, $\hat{S}^{h/d}(\hat{B}_1(x_0))$ satisfies Propositions* 2.1–2.3 *with h replaced by h/d. The constants occurring in these propositions remain unchanged, in particular independent of d.*

3. Stability and boundedness. In this section we discuss the stability and boundedness of $A(\cdot,\cdot)$. By *stability* we mean the coercivity of $A(\cdot,\cdot)$ in $H_h^1(\Omega)$ (see (3.5) in Lemma 3.2). This coercivity is essential to guarantee existence and uniqueness of the finite element solution u_h of (2.7). The boundedness of $A(\cdot,\cdot)$ in $|||\cdot|||_{W_h^{1,p}(\Omega)}$ is proved in Lemma 3.3. These results will be used in the proofs in the fourth and fifth sections. It is well known that for any $v \in H_0^1(\Omega)$, $\|v\|_{L^2(\Omega)} \le C\|\nabla v\|_{L^2(\Omega)}$. The result in the next lemma is a discontinuous version of this inequality.

LEMMA 3.1. *There is a constant $C > 0$ such that for any $v \in S^h$ we have*

$$\|v\|_{L^2(\Omega)}^2 \le C \sum_{e\in\Gamma_h} h_e^{-1} \int_e |[\![v]\!]|^2 \, ds + C|v|_{H_h^1(\Omega)}^2. \tag{3.1}$$

Proof. For any $K \in \mathcal{J}_h$ and $x \in K$, choose

$$\{K_1, K_2, \ldots, K_{i_K}\} \subset \mathcal{J}_h \quad \text{and} \quad \{e_1, e_2, \ldots, e_{i_K}\} \subset \Gamma_h$$

such that $K_1 = K$, $i_k \le Ch^{-1}$, $e_i \subset \partial K_i \cap \partial K_{i+1}$ for $1 \le i \le i_K - 1$, and $e_{i_K} \subset \partial K_{i_K} \cap \partial\Omega$. Then, for any $s \in e_1$ and $y \in K_2$,

$$v(x) = (v|_{K_1}(s) - v|_{K_2}(s)) + (v(x) - v|_{K_1}(s)) + (v|_{K_2}(s) - v(y)) + v(y),$$

which implies

$$\begin{aligned} |v(x)| &\le |[\![v]\!](s)| + Ch\|\nabla v\|_{L^\infty(K_1\cup K_2)} + |v(y)| \\ &\le |[\![v]\!](s)| + Ch^{1-N/2}\|\nabla v\|_{L^2(K_1\cup K_2)} + |v(y)|. \end{aligned} \tag{3.2}$$

Integrating on e_1 with respect to s, it follows from (3.2) that

$$\begin{aligned} |v(x)| &\le Ch^{1-N} \int_{e_1} |[\![v]\!]| \, ds + Ch^{1-N/2}\|\nabla v\|_{L^2(K_1\cup K_2)} + |v(y)| \\ &\le Ch^{(1-N)/2} \left(\int_{e_1} |[\![v]\!]|^2 \, ds \right)^{1/2} + Ch^{1-N/2}\|\nabla v\|_{L^2(K_1\cup K_2)} + |v(y)|. \end{aligned}$$

Reasoning in the same way, we have for each $1 \le i \le i_K - 1$, $x \in K_i$, and $y \in K_{i+1}$,

$$|v(x)| \le Ch^{(1-N)/2} \left(\int_{e_i} |[\![v]\!]|^2 \, ds \right)^{1/2} + Ch^{1-N/2}\|\nabla v\|_{L^2(K_i\cup K_{i+1})} + |v(y)|.$$

Using the fact that $[\![v]\!] = v$ on $e_{i_K} \subset \partial\Omega$, for any $y \in K_{i+1}$ we have

$$|v(y)| \le Ch^{(1-N)/2} \left(\int_{e_{i_K}} |[\![v]\!]|^2 \, ds \right)^{1/2} + Ch^{1-N/2}\|\nabla v\|_{L^2(K_{i_K})}.$$

Thus we conclude that for $x \in K$,

$$|v(x)| \le Ch^{1-N/2} \sum_{i=1}^{i_K} \left(h_e^{-1} \int_{e_i} |[\![v]\!]|^2 \, ds \right)^{1/2} + Ch^{1-N/2} \sum_{i=1}^{i_K} \|\nabla v\|_{L^2(K_i)},$$

which further yields

$$|v(x)|^2 \le Ch^{1-N} \sum_{i=1}^{i_K} h_e^{-1} \int_{e_i} |[\![v]\!]|^2 \, ds + Ch^{1-N} \sum_{i=1}^{i_K} \|\nabla v\|_{L^2(K_i)}^2. \tag{3.3}$$

Integrating (3.3) on K with respect to x, we obtain

$$\|v\|^2_{L^2(K)} \le Ch \sum_{i=1}^{i_K} h_e^{-1} \int_{e_i} |[\![v]\!]|^2 \, ds + Ch \sum_{i=1}^{i_K} \|\nabla v\|^2_{L^2(K_i)}. \tag{3.4}$$

Taking summation over $K \in \mathcal{J}_h$ in (3.4) and noting that the integrals on the right-hand side of (3.4) may repeat at most Ch^{-1} times, the result (3.1) follows. ∎

LEMMA 3.2. *For sufficiently large constants C_L and C_λ defined in* (2.2) *and* (2.6)*, there is a constant $C > 0$ such that*

$$C|||v|||^2_{H_h^1(\Omega)} \le A(v,v) \qquad \forall\ v \in S^h. \tag{3.5}$$

Proof. For any $v \in S^h$, by the definition of the bilinear form $A(\cdot,\cdot)$, we have

$$\begin{aligned} A(v,v) &= a(v,v) - (\tau+1)b(v,v) + \lambda(v,v) \\ &\ge C_L |v|^2_{H_h^1(\Omega)} + C_\lambda \sum_{e \in \mathcal{J}_h} h_e^{-1} \int_e |[\![v]\!]|^2 \, ds \\ &\quad + \int_\Omega \sum_{i=1}^N b_i \frac{\partial v}{\partial x_i} v \, dx - (\tau+1) b(v,v). \end{aligned} \tag{3.6}$$

In virtue of (3.1) and Hölder's inequality, we see that

$$\int_\Omega \sum_{i=1}^N b_i \frac{\partial v}{\partial x_i} v \, dx \le C|v|_{H_h^1(\Omega)} \|v\|_{L^2(\Omega)} \le C|v|^2_{H_h^1(\Omega)} + C \sum_{e \in \Gamma_h} h_e^{-1} \int_e |[\![v]\!]|^2 \, ds. \tag{3.7}$$

By the definition of $b(\cdot,\cdot)$, we have

$$b(v,v) \le C \sum_{e \in \Gamma_h} h_e \int_e \sum_{i=1}^N \left| \left\{ \frac{\partial v}{\partial x_i} \right\} \right|^2 \, ds + C \sum_{e \in \Gamma_h} h_e^{-1} \int_e |[\![v]\!]|^2 \, ds. \tag{3.8}$$

We recall the trace inequality

$$\int_e \sum_{i=1}^N \left| \left\{ \frac{\partial v}{\partial x_i} \right\} \right|^2 \, ds \le C h_e^{-1} |v|^2_{H^1(K)} + C h_e |v|^2_{H^2(K)}, \tag{3.9}$$

where $K \in \mathcal{J}_h$ and $e \subset \partial K$. Inserting (3.9) into (3.8) and using an inverse inequality, we have

$$b(v,v) \le C|v|^2_{H_h^1(\Omega)} + C \sum_{e \in \Gamma_h} h_e^{-1} \int_e |[\![v]\!]|^2 \, ds. \tag{3.10}$$

Applying (3.7) and (3.10) in (3.6) and choosing C_L and C_λ sufficiently large, we obtain

$$A(v,v) \ge C \left(|v|^2_{H_h^1(\Omega)} + \sum_{e \in \Gamma_h} h_e^{-1} \int_e |[\![v]\!]|^2 \, ds \right). \tag{3.11}$$

Using the trace inequality (3.9), an inverse inequality, and (3.1), we can easily see that

$$|||v|||^2_{H_h^1(\Omega)} \le C \left(|v|^2_{H_h^1(\Omega)} + \sum_{e \in \Gamma_h} h_e^{-1} \int_e |[\![v]\!]|^2 \, ds \right),$$

which, along with (3.11), shows the desired result (3.5). ∎

From now on, we assume that the constants C_L and C_λ are sufficiently large so that inequality (3.5) of Lemma 3.2 holds. Using the result in Lemma 3.2, we obtain the unique solvability of problem (2.7). To obtain error estimates, we further need the following boundedness of the bilinear form $A(\cdot,\cdot)$:

LEMMA 3.3. *For $1 \le p \le \infty$, there is a constant $C > 0$ such that for any $v \in W_h^{2,p}(\Omega)$ and $w \in W_h^{2,p'}(\Omega)$, it holds that*

$$A(v,w) \le C|||v|||_{W_h^{1,p}(\Omega)}|||w|||_{W_h^{1,p'}(\Omega)}, \tag{3.12}$$

where p' is the conjugate of p, i.e., $1/p + 1/p' = 1$.

Proof. From the definition (2.4) of $A(\cdot,\cdot)$, we see that it suffices to bound each of $a(v,w)$, $b(v,w)$, $b(w,v)$, and $\lambda(v,w)$ by the right-hand side of (3.12). In fact, by Hölder's inequality, we have

$$a(v,w) \le C\|v\|_{W_h^{1,p}(\Omega)}\|w\|_{W_h^{1,p'}(\Omega)},$$

$$b(v,w) \le C\left(\sum_{e\in\Gamma_h} h_e \int_e \sum_{i=1}^N \left|\left\{\left\{\frac{\partial v}{\partial x_i}\right\}\right\}\right|^p \, ds\right)^{1/p} \left(\sum_{e\in\Gamma_h} h_e^{1-p'} \int_e |[\![w]\!]|^{p'} \, ds\right)^{1/p'},$$

$$b(w,v) \le C\left(\sum_{e\in\Gamma_h} h_e^{1-p} \int_e |[\![v]\!]|^{p} \, ds\right)^{1/p} \left(\sum_{e\in\Gamma_h} h_e \int_e \sum_{i=1}^N \left|\left\{\left\{\frac{\partial w}{\partial x_i}\right\}\right\}\right|^{p'} \, ds\right)^{1/p'},$$

$$\lambda(v,w) \le C\left(\sum_{e\in\Gamma_h} h_e^{1-p} \int_e |[\![v]\!]|^{p} \, ds\right)^{1/p} \left(\sum_{e\in\Gamma_h} h_e^{1-p'} \int_e |[\![w]\!]|^{p'} \, ds\right)^{1/p'}.$$

Substituting the above four inequalities into (2.4) and recalling the definition of the norm (2.10) completes the proof of this lemma. ∎

LEMMA 3.4. *Suppose that $u \in H_0^1(\Omega)$ and $u_h \in S^h$ satisfy (2.9). Then*

$$|||u - u_h|||_{H_h^1(\Omega)} \le C \inf_{\chi\in S^h} |||u - \chi|||_{H_h^1(\Omega)}. \tag{3.13}$$

Proof. Inequality (3.13) follows immediately from the error equation (2.9), the stability estimate (3.5), the boundedness result (3.12), and Hölder's inequality. ∎

4. Interior error estimates. To prepare for the proof of pointwise error estimates, we show local error estimates in the energy and L^2 norms for the error of the finite element approximation and local a priori estimates for the solution of the elliptic problem (2.1). The result in Lemma 4.1 indicates that the local error of the finite element solution measured in the H_h^1 norm is bounded by the local approximation property of the finite element space in this norm plus the error measured in the weaker L^2 norm.

LEMMA 4.1. *Let $\Omega_0 \subset \Omega_1 \subset \Omega$ with $d = dist(\Omega_0, \partial\Omega_1\backslash\partial\Omega) \ge 4\kappa h$. If $u \in H_0^1(\Omega)$ and $u_h \in S^h$ satisfy (2.9), then*

$$|||u - u_h|||_{H_h^1(\Omega_0)} \le C \inf_{\chi\in S^h(\Omega_1)} |||u - \chi|||_{H_h^1(\Omega_1)} + C\|u - u_h\|_{L^2(\Omega_1)}, \tag{4.1}$$

where the constant $C > 0$ depends on d but is independent of h and u.

Proof. Let $\Omega_0 \subset \Omega_2 \subset \Omega_3 \subset \Omega_4 \subset \Omega_1$ satisfy

$$\operatorname{dist}(\Omega_0, \partial\Omega_2 \backslash \partial\Omega) = \operatorname{dist}(\Omega_2, \partial\Omega_3 \backslash \partial\Omega) = \operatorname{dist}(\Omega_3, \partial\Omega_4 \backslash \partial\Omega) = \operatorname{dist}(\Omega_4, \partial\Omega_1 \backslash \partial\Omega) = d/4.$$

Also, let $\omega \in C_0^\infty(\Omega_2)$ satisfy $\omega \equiv 1$ on Ω_0. Then, by Proposition 2.3, we choose $\eta \in S^h(\Omega_3)$ such that $\operatorname{supp}(\eta) \not\subset \Omega_3$ and

$$|||\omega(\chi - u_h) - \eta|||_{H_h^1(\Omega_3)} \le Ch|||\chi - u_h|||_{H_h^1(\Omega_3)}. \tag{4.2}$$

Using the triangle inequality, (3.5), and (4.2), we have

$$\begin{aligned} |||\chi - u_h|||_{H_h^1(\Omega_0)} &\le |||\omega(\chi - u_h)|||_{H_h^1(\Omega)} \\ &\le |||\eta|||_{H_h^1(\Omega)} + |||\omega(\chi - u_h) - \eta|||_{H_h^1(\Omega_3)} \\ &\le C\left(\sqrt{A(\eta,\eta)} + h|||\chi - u_h|||_{H_h^1(\Omega_2)}\right). \end{aligned} \tag{4.3}$$

With some straightforward manipulations, we have the identity

$$\begin{aligned} A(\eta,\eta) = {} & A(\omega(\chi - u_h), \omega(\chi - u_h)) + A(\eta - \omega(\chi - u_h), \eta - \omega(\chi - u_h)) \\ & + A(\omega(\chi - u_h), \eta - \omega(\chi - u_h)) + A(\eta - \omega(\chi - u_h), \omega(\chi - u_h)). \end{aligned} \tag{4.4}$$

From (3.12), (4.2), and (4.4), it follows that

$$\begin{aligned} A(\eta,\eta) \le {} & A(\omega(\chi - u_h), \omega(\chi - u_h)) + Ch^2|||\chi - u_h|||^2_{H_h^1(\Omega_3)} \\ & + \varepsilon|||\omega(\chi - u_h)|||^2_{H_h^1(\Omega)}, \end{aligned} \tag{4.5}$$

where $\varepsilon > 0$ is an arbitrary but fixed real number, which will be determined later in this proof. We will now estimate the first term on the right-hand side of (4.5). Some simple calculations lead to the following equation:

$$\begin{aligned} A(\omega(\chi - u_h), \omega(\chi - u_h)) = {} & A(\chi - u_h, \omega^2(\chi - u_h)) + I_\omega(\chi - u_h) \\ & - (\tau + 1) \sum_{e \in \Gamma_h} \int_e \sum_{i,j=1}^N a_{ij} \frac{\partial \omega}{\partial x_i} \{\!\{\chi - u_h\}\!\} [\![\omega(\chi - u_h)]\!]_j \, ds, \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} I_\omega(v) = {} & \int_\Omega \sum_{i,j=1}^N a_{ij} \left(v \frac{\partial \omega}{\partial x_i} \frac{\partial(\omega v)}{\partial x_j} - v \frac{\partial \omega}{\partial x_j} \frac{\partial(\omega v)}{\partial x_i} + v^2 \frac{\partial \omega}{\partial x_i} \frac{\partial \omega}{\partial x_j} \right) dx \\ & + \int_\Omega \omega v^2 \sum_{i=1}^N b_i \frac{\partial \omega}{\partial x_i} \, dx. \end{aligned} \tag{4.7}$$

By (4.7) and Hölder's inequality, $I_\omega(\chi - u_h)$ can be estimated as follows:

$$I_\omega(\chi - u_h) \le C\|\chi - u_h\|^2_{L^2(\Omega_2)} + \varepsilon\|\omega(\chi - u_h)\|^2_{H_h^1(\Omega)}. \tag{4.8}$$

To estimate the first term in the right-hand side of (4.6), we use the superapproximation property in Proposition 2.3 and choose a function $\eta_1 \in S^h(\Omega_3)$ such that $\operatorname{supp}(\eta_1) \not\subset \Omega_3$ and

$$|||\omega^2(\chi - u_h) - \eta_1|||_{H_h^1(\Omega_3)} \le Ch|||\chi - u_h|||_{H_h^1(\Omega_3)}. \tag{4.9}$$

Then, from (2.9), inequality (3.12), and estimate (4.9), it follows that

$$
\begin{aligned}
(4.10)\qquad & A(\chi-u_h,\omega^2(\chi-u_h))\\
&= A(\chi-u,\omega^2(\chi-u_h))+A(u-u_h,\omega^2(\chi-u_h)-\eta_1)\\
&\le C|||u-\chi|||_{H_h^1(\Omega_2)}|||\omega(\chi-u_h)|||_{H_h^1(\Omega)}\\
&\quad +Ch|||u-u_h|||_{H_h^1(\Omega_3)}|||\chi-u_h|||_{H_h^1(\Omega_3)}\\
&\le C|||u-\chi|||^2_{H_h^1(\Omega_3)}+\varepsilon|||\omega(\chi-u_h)|||^2_{H_h^1(\Omega)}+Ch|||\chi-u_h|||^2_{H_h^1(\Omega_3)}.
\end{aligned}
$$

For the last term in the right-hand side of (4.6), applying Hölder's inequality, a trace theorem, and an inverse inequality, we have

$$
\begin{aligned}
(4.11)\qquad & \sum_{e\in\Gamma_h}\int_e\sum_{i,j=1}^N a_{ij}\frac{\partial\omega}{\partial x_i}\{\chi-u_h\}[\![\omega(\chi-u_h)]\!]_j\,ds\\
&\le C\left(\sum_{e\in\Gamma_h}h_e^{-1}\int_e|[\![\omega(\chi-u_h)]\!]|^2\,ds\right)^{1/2}\left(\sum_{e\in\Gamma_h}h_e\int_e|\{\chi-u_h\}|^2\,ds\right)^{1/2}\\
&\le \varepsilon|||\omega(\chi-u_h)|||^2_{H_h^1(\Omega)}+C\|\chi-u_h\|^2_{L^2(\Omega_3)}.
\end{aligned}
$$

Now, using estimates (4.3), (4.5), (4.6), (4.8), (4.10), and (4.11), we obtain

$$
\begin{aligned}
(4.12)\qquad |||\omega(\chi-u_h)|||^2_{H_h^1(\Omega)} &\le C\varepsilon|||\omega(\chi-u_h)|||^2_{H_h^1(\Omega)}+Ch|||\chi-u_h|||^2_{H_h^1(\Omega_3)}\\
&\quad +C|||u-\chi|||^2_{H_h^1(\Omega_3)}+C\|u-u_h\|^2_{L^2(\Omega_3)}.
\end{aligned}
$$

Next, choosing ε sufficiently small so that the term $\varepsilon|||\omega(\chi-u_h)|||^2_{H_h^1(\Omega)}$ can be kicked back to the left-hand side of (4.12) and then using (4.3), we get

$$
\begin{aligned}
(4.13)\qquad & |||\chi-u_h|||^2_{H_h^1(\Omega_0)}\\
&\le C\left(h|||\chi-u_h|||^2_{H_h^1(\Omega_3)}+|||u-\chi|||^2_{H_h^1(\Omega_3)}+\|u-u_h\|^2_{L^2(\Omega_3)}\right).
\end{aligned}
$$

Using (4.13) with Ω_0 replaced by Ω_3, we also have

$$
\begin{aligned}
(4.14)\qquad & |||\chi-u_h|||^2_{H_h^1(\Omega_3)}\\
&\le C\left(h|||\chi-u_h|||^2_{H_h^1(\Omega_4)}+|||u-\chi|||^2_{H_h^1(\Omega_4)}+\|u-u_h\|^2_{L^2(\Omega_4)}\right).
\end{aligned}
$$

Inserting (4.14) into (4.13) and applying an inverse inequality, we obtain

$$
\begin{aligned}
(4.15)\qquad & |||\chi-u_h|||^2_{H_h^1(\Omega_0)}\\
&\le C\left(h^2|||\chi-u_h|||^2_{H_h^1(\Omega_4)}+|||u-\chi|||^2_{H_h^1(\Omega_4)}+\|u-u_h\|^2_{L^2(\Omega_4)}\right)\\
&\le C|||u-\chi|||^2_{H_h^1(\Omega_1)}+C\|u-u_h\|^2_{L^2(\Omega_1)}.
\end{aligned}
$$

Estimate (4.15) and the triangle inequality imply (4.1). □

An explicit dependence of the bound in (4.1) on the distance d between Ω_0 and Ω_1 can be determined through Proposition 2.4 and a scaling argument.

LEMMA 4.2. *Let $\Omega_0\subset\Omega_1\subset\Omega$ with $d=dist(\Omega_0,\partial\Omega_1\backslash\partial\Omega)\ge\kappa h$. If $u\in H_0^1(\Omega)$ and $u_h\in S^h$ satisfy* (2.9)*, then*

$$
(4.16)\qquad |||u-u_h|||_{H_h^1(\Omega_0)}\le Ch^r\|u\|_{H^{1+r}(\Omega_1)}+Cd^{-1}\|u-u_h\|_{L^2(\Omega_1)}.
$$

Proof. Without loss of generality, we assume that Ω is the unit ball in R^N. It suffices to show (4.16) with Ω_0 and Ω_1 being the spheres of radii $d/2$ and d, respectively, with centers at x_0. Assume that $x_0 = 0$ and x denotes the variable on Ω. Let $\tilde{x} = x/d$ be the new variable on the transferred regions $\tilde{\Omega}_0$ and $\tilde{\Omega}_1$. Then $\operatorname{dist}(\tilde{\Omega}_0, \partial\tilde{\Omega}_1) = 1/2$. Set $\tilde{u}(\tilde{x}) = u(\tilde{x}d)$ and $\tilde{u}_h = u_h(\tilde{x}d)$. Then we see that

$$\tilde{A}(\tilde{u} - \tilde{u}_h, \tilde{\chi}) = 0 \qquad \forall\ \tilde{\chi} \in \tilde{S}^h,$$

where $\tilde{S}^h$ is the transferred space of S^h,

$$\tilde{A}(\tilde{v}, \tilde{w}) = \tilde{a}(\tilde{v}, \tilde{w}) - \tilde{b}(\tilde{v}, \tilde{w}) - \tau\tilde{b}(\tilde{w}, \tilde{v}) + \tilde{\lambda}(\tilde{v}, \tilde{w}),$$

and

$$\tilde{a}(\tilde{v}, \tilde{w}) = \sum_{\tilde{K} \in \tilde{\mathcal{J}}_h} \int_{\tilde{K}} \left(\sum_{i,j=1}^{N} \tilde{a}_{ij}(\tilde{x}) \frac{\partial \tilde{v}}{\partial \tilde{x}_i} \frac{\partial \tilde{w}}{\partial \tilde{x}_j} + d \sum_{i}^{N} \tilde{b}(\tilde{x}) \frac{\partial \tilde{v}}{\partial \tilde{x}_i} \tilde{w} + d^2 \tilde{c}(\tilde{x}) \tilde{v}\tilde{w} \right) d\tilde{x},$$

$$\tilde{b}(\tilde{v}, \tilde{w}) = \sum_{\tilde{e} \in \tilde{\Gamma}_h} \int_{\tilde{e}} \sum_{i,j=1}^{N} \left\{ \tilde{a}_{ij} \frac{\partial \tilde{v}}{\partial \tilde{x}_i} \right\} [\![\tilde{w}]\!]_j \, d\tilde{s},$$

$$\tilde{\lambda}(\tilde{v}, \tilde{w}) = \sum_{\tilde{e} \in \tilde{\Gamma}_h} \frac{\lambda_e}{h_e/d} \int_{\tilde{e}} [\![\tilde{v}]\!] [\![\tilde{w}]\!] \, d\tilde{s}.$$

Since the coefficients $\tilde{a}_{ij}$ in the bilinear form $\tilde{A}(\cdot, \cdot)$ satisfy inequality (2.2) with the same constant C_L, and the upper bounds of the derivatives of the transferred coefficients in this form are reduced compared to those of the original coefficients in $A(\cdot, \cdot)$, using Proposition 2.4 and estimate (4.1) for the error $\tilde{u} - \tilde{u}_h$ we deduce that

$$|||\tilde{u} - \tilde{u}_h|||_{H_h^1(\tilde{\Omega}_0)} \le C \inf_{\tilde{\chi} \in \tilde{S}^h} |||\tilde{u} - \tilde{\chi}|||_{H_h^1(\tilde{\Omega}_1)} + C\|\tilde{u} - \tilde{u}_h\|_{L^2(\tilde{\Omega}_1)}. \tag{4.17}$$

Using (4.17) and the approximation properties in space $\tilde{S}^h$, we see that

$$\begin{aligned} |||\tilde{u} - \tilde{u}_h|||_{H_h^1(\tilde{\Omega}_0)} &\le C(h/d)^r |\tilde{u}|_{H^{1+r}(\tilde{\Omega}_1)} + C\|\tilde{u} - \tilde{u}_h\|_{L^2(\tilde{\Omega}_1)} \\ &\le Ch^r d^{1-N/2} |u|_{H^{1+r}(\Omega_1)} + Cd^{-N/2} \|u - u_h\|_{L^2(\Omega_1)}. \end{aligned} \tag{4.18}$$

From (4.18) and the inequality

$$|||u - u_h|||_{H_h^1(\Omega_0)} \le Cd^{N/2-1} |||\tilde{u} - \tilde{u}_h|||_{H_h^1(\tilde{\Omega}_0)},$$

we obtain the desired result (4.16). ∎

Without loss of generality we assume *in the rest of the paper* that $\operatorname{diam}(\Omega) \le 1$ and define

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

and for any fixed $x \in \bar{\Omega}$, set

$$\begin{aligned} \Omega_j &= \{x \in \Omega :\ d_{j+1} < |x - z| < d_j\}, \\ \Omega_j^{(1)} &= \{x \in \Omega :\ d_{j+2} < |x - z| < d_{j-1}\}, \\ \Omega_j^{(2)} &= \{x \in \Omega :\ d_{j+3} < |x - z| < d_{j-2}\}. \end{aligned} \tag{4.19}$$

LEMMA 4.3. *For $\rho \in L^\infty(\Omega)$, let $g \in H_0^1(\Omega)$ be the solution of*

$$\mathcal{L}g = \rho \quad \text{in } \Omega$$

and $g_h \in S^h$ be the finite element approximation of g:

$$A(g - g_h, v) = 0 \qquad \forall\ v \in S^h.$$

If ρ has compact support in $B_{Mh}(z)$ for some $M > 1$ and $\|\rho\|_{L^2(B_{Mh}(z))} \le Ch^{-N/2}$,

$$|||g - g_h|||_{H_h^1(\Omega_j)} \le Ch^r d_j^{1-r-N/2} + d_j^{-1}\|g - g_h\|_{L^2(\Omega_j^{(1)})}. \tag{4.20}$$

Proof. Using (4.16), we have

$$|||g - g_h|||_{H_h^1(\Omega_j)} \le Ch^r \|g\|_{H^{1+r}(\Omega_j^{(1)})} + d_j^{-1}\|g - g_h\|_{L^2(\Omega_j^{(1)})}. \tag{4.21}$$

For any $x \in \Omega_j^{(1)}$, let G_x be Green's function for problem (2.1) with singularity at x. Then we have (see Solonnikov [22])

$$g(x) = \int_\Omega G_x(y)\rho(y)\ dy \tag{4.22}$$

and

$$\left|\frac{\partial^{\alpha+\beta} G_x(y)}{\partial x^\alpha \partial y^\beta}\right| \le C|x - y|^{2-N-|\alpha|-|\beta|} \quad \text{for } |\alpha| + |\beta| > 0. \tag{4.23}$$

Differentiating (4.22) with respect to x, for $x \in \Omega_j^{(1)}$ and $|\alpha| \le 1 + r$ we have

$$\begin{aligned}\left|\frac{\partial^\alpha g(x)}{\partial x^\alpha}\right| &= \left|\int_\Omega \frac{\partial^\alpha G_x(y)}{\partial x^\alpha}\rho(y)\ dy\right| \\ &\le C\int_{B_{Mh}(z)} |x - y|^{2-N-|\alpha|}\,|\rho(y)|\ dy \\ &\le Cd_j^{1-N-r}h^{N/2}\|\rho\|_{L^2(B_{Mh}(z))} \le Cd_j^{1-N-r}.\end{aligned} \tag{4.24}$$

Integrating (4.24) over $\Omega_j^{(1)}$ gives

$$\|g\|_{H^{1+r}(\Omega_j^{(1)})} \le Cd_j^{1-N/2-r}.$$

Substituting this into (4.21) implies the desired result (4.20). ∎

LEMMA 4.4. *For $\varphi \in C_0^\infty(B_{Mh}(z))$ satisfying $\|\varphi\|_{H^1(B_{Mh}(z))} = 1$, let $g \in H_0^1(\Omega)$ be the solution of*

$$\mathcal{L}g = -h^{-N/2-1}\frac{\partial \varphi}{\partial x_i} \quad \text{in } \Omega$$

and $g_h \in S^h$ be the finite element approximation of g:

$$A(g - g_h, v) = 0 \qquad \forall\ v \in S^h.$$

Then it holds that

$$|||g-g_h|||_{H_h^1(\Omega_j)} \le Ch^r d_j^{-r-N/2} + d_j^{-1}\|g-g_h\|_{L^2(\Omega_j^{(1)})}. \tag{4.25}$$

Proof. The proof is the same as that of Lemma 4.3. The only difference is that instead of (4.22), for $x \in \Omega_j$ and $|\alpha| = 1+r$ we have

$$\begin{aligned}\left|\frac{\partial^\alpha g(x)}{\partial x^\alpha}\right| &= h^{-N/2-1}\left|\int_\Omega \frac{\partial^\alpha G_x(y)}{\partial x^\alpha}\frac{\partial\varphi(y)}{\partial y_i}\,dy\right| \\ &= h^{-N/2-1}\left|\int_{B_{Mh}(z)} \frac{\partial}{\partial y_i}\frac{\partial^\alpha G_x(y)}{\partial x^\alpha}\varphi(y)\,dy\right| \\ &\le Cd_j^{-N-r}h^{-1}\|\varphi\|_{L^2(B_{Mh}(z))}.\end{aligned} \tag{4.26}$$

Using the fact that $\varphi \in C_0^\infty(B_{Mh}(z))$, we have

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

Inserting this into (4.26) and integrating over Ω_j, we conclude that

$$\|g\|_{H^{1+r}(\Omega_j)} \le Cd_j^{-N/2-r}, \tag{4.27}$$

which completes the proof. ∎

LEMMA 4.5. *For $\varphi \in C_0^\infty(\Omega_j^{(1)})$ satisfying $\|\varphi\|_{L^2(\Omega)} = 1$, let $\Phi \in H_0^1(\Omega)$ be the solution of $\mathcal{L}\Phi = \varphi$ in Ω. Then it holds that*

$$\|\Phi\|_{W^{1+r,\infty}(\Omega\setminus\Omega_j^{(2)})} \le Cd_j^{1-r-N/2}. \tag{4.28}$$

Proof. Estimate (4.28) follows immediately from differentiating the representation

$$\Phi(x) = \int_{\Omega)} G_x(y)\varphi(y)\,dy$$

and using inequality (4.23). ∎

5. Pointwise error estimates. In this section we prove our main results: the optimal localized pointwise error estimates. These results are stated in Theorems 5.1 and 5.3. We concentrate on the case $\tau = 1$, i.e., the symmetric interior penalty DG method. The technique used here (see (5.5) and (5.6) below) cannot be easily extended to the nonsymmetric case. Pointwise error estimates for this case will be investigated in the future.

THEOREM 5.1. *Let $u \in W_0^{1,\infty}(\Omega)$ and $u_h \in S^h$ satisfy (2.9) and $0 \le s \le r-1$. Then there is a constant $C > 0$ such that for any $z \in \bar\Omega$,*

$$|(u-u_h)(z)| \le Ch\left(\ln h^{-1}\right)^{\bar s}\inf_{\chi\in S^h}|||u-\chi|||_{W_h^{1,\infty}(\Omega),z,s}, \tag{5.1}$$

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

Proof. Let $K_z \in \mathcal{J}_h$ be such that $z \in \bar K_z$. By choosing appropriately a $\psi \in S^h$ according to Proposition 2.1, the triangle inequality and Proposition 2.2 yield

$$\begin{aligned}|(u-u_h)(z)| &\le |(u-\psi)(z)| + Ch^{-N/2}\|\psi-u_h\|_{L^2(K_z)} \\ &\le C\|u-\psi\|_{L^\infty(K_z)} + Ch^{-N/2}\|u-u_h\|_{L^2(K_z)} \\ &\le Ch\|u\|_{W_h^{1,\infty}(\Omega),z,s} + Ch^{-N/2}\|u-u_h\|_{L^2(K_z)}.\end{aligned} \tag{5.2}$$

Obviously, (5.2) holds also for u replaced by $u-\chi$ for any $\chi \in S^h$. Therefore, noting that $u-\chi-(u-\chi)_h = u-u_h$, (5.2) implies

$$|(u-u_h)(z)| \le Ch\|u-\chi\|_{W_h^{1,\infty}(\Omega),z,s} + Ch^{-N/2}\|u-u_h\|_{L^2(K)}. \tag{5.3}$$

Define

$$\rho(x) = h^{-N/2}(u-u_h)(x)/\|u-u_h\|_{L^2(K_z)} \tag{5.4}$$

and let $g_z \in H_0^1(\Omega)$ be the solution of

$$\mathcal{L}^* g_z = \rho, \tag{5.5}$$

where $\mathcal{L}^*$ is the adjoint operator of $\mathcal{L}$. Furthermore, let $g_{z,h}$ be the finite element approximation of g_z satisfying

$$A(v, g_z - g_{z,h}) = 0 \qquad \forall\ v \in S^h. \tag{5.6}$$

Then, from (2.9), (3.12), and (5.6), for any $\chi \in S^h$ it follows that

$$\begin{aligned} h^{-N/2}\|u-u_h\|_{L^2(K_z)} &= (\rho, u-u_h) = A(u-\chi, g_z - g_{z,h}) \\ &\le C|||u-\chi|||_{W_h^{1,\infty}(\Omega),z,s}|||g_z - g_{z,h}|||_{W_h^{1,1}(\Omega),z,-s}, \end{aligned} \tag{5.7}$$

which, along with (5.8) in Lemma 5.2 below, deduces the desired result (5.1). ∎

LEMMA 5.2. *Let $g_z \in H_0^1(\Omega)$ and $g_{z,h} \in S^h$ satisfy* (5.5) *and* (5.6). *Then it holds that for $0 \le s \le r-1$,*

$$|||g_z - g_{z,h}|||_{W_h^{1,1}(\Omega),z,-s} \le Ch\left(\ln h^{-1}\right)^{\bar{s}}, \tag{5.8}$$

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

Proof. Let $M > 1$ be a real number to be determined later and J be an integer such that $Mh = 2^{-J}$. Then $J \le C\ln(1/h)$. For notational convenience, set $E = g_z - g_{z,h}$. In view of $\Omega = B_{Mh}(z) \cup (\cup_{j=0}^J \Omega_j)$, it follows that

$$|||E|||_{W_h^{1,1}(\Omega),z,-s} \le |||E|||_{W_h^{1,1}(B_{Mh}(z)),z,-s} + \sum_{j=0}^{J} |||E|||_{W_h^{1,1}(\Omega_j),z,-s}. \tag{5.9}$$

Recall the definition of $|||E|||_{W_h^{1,1}(\Omega_j),z,-s}$:

$$\begin{aligned} |||E|||_{W_h^{1,1}(\Omega_j),z,-s} &= \|E\|_{W_h^{1,1}(\Omega_j),z,-s} + \sum_{e\in\Gamma_h}\int_{e\cap\Omega_j} |\sigma_{z,h}^s [\![E]\!]|\ ds \\ &\quad + \sum_{e\in\Gamma_h} h_e \int_{e\cap\Omega_j} \sum_{i=1}^{N} \left|\sigma_{z,h}^s \left\{\!\!\left\{\frac{\partial E}{\partial x_i}\right\}\!\!\right\}\right|\ ds. \end{aligned} \tag{5.10}$$

We will handle each of the three terms on the right-hand side of (5.10) separately. By Hölder's inequality, we see that

$$\|E\|_{W_h^{1,1}(\Omega_j),z,-s} \le C d_j^{N/2+s} h^{-s} \|E\|_{H_h^1(\Omega_j)}, \tag{5.11}$$

$$\sum_{e\in\Gamma_h}\int_{e\cap\Omega_j}|\sigma_{z,h}^s[\![E]\!]|\,ds \le d_j^s h^{(N-1)/2-s}\sum_{e\in\Gamma_h}\left(\int_{e\cap\Omega_j}|[\![E]\!]|^2\,ds\right)^{1/2} \le Cd_j^{N/2+s}h^{-s}\left(\sum_{e\in\Gamma_h}h_e^{-1}\int_{e\cap\Omega_j}|[\![E]\!]|^2\,ds\right)^{1/2}, \tag{5.12}$$

$$\sum_{e\in\Gamma_h}h_e\int_{e\cap\Omega_j}\sum_{i=1}^N\sigma_{z,h}^s\left|\left\{\!\!\left\{\frac{\partial E}{\partial x_i}\right\}\!\!\right\}\right|\,ds \le Cd_j^{N/2+s}h^{-s}\left(\sum_{e\in\Gamma_h}h_e\int_{e\cap\Omega_j}\sum_{i=1}^N\left|\left\{\!\!\left\{\frac{\partial E}{\partial x_i}\right\}\!\!\right\}\right|^2\,ds\right)^{1/2}. \tag{5.13}$$

Inserting (5.11)–(5.13) into (5.10) gives

$$|||E|||_{W_h^{1,1}(\Omega_j),z,-s} \le Cd_j^{N/2+s}h^{-s}|||E|||_{H_h^1(\Omega_j)}. \tag{5.14}$$

Similar to (5.14) (with Ω_j replaced by $B_{Mh}(z)$), it can easily be seen that

$$|||E|||_{W_h^{1,1}(B_{Mh}(z)),z,-s} \le CM^{N/2+s}h^{N/2}|||E|||_{H_h^1(B_{Mh}(z))}. \tag{5.15}$$

Using Lemma 3.4, the approximation properties in Proposition 2.1, and an elliptic regularity on g_z, we have

$$|||E|||_{H_h^1(B_{Mh}(z))} \le Ch\|g_z\|_{H^2(\Omega)} \le Ch\|\rho\|_{L^2(K_z)} \le Ch^{1-N/2},$$

which, along with (5.15), yields

$$|||E|||_{W_h^{1,1}(B_{Mh}(z)),z,-s} \le CM^{N/2+s}h. \tag{5.16}$$

Inserting (5.14) and (5.16) into (5.9), we have

$$|||E|||_{W_h^{1,1}(\Omega),z,-s} \le CM^{N/2+s}h + CL, \tag{5.17}$$

where

$$L = \sum_{j=0}^J d_j^{N/2+s}h^{-s}|||E|||_{H_h^1(\Omega_j)}.$$

To estimate L, we use (4.20) to obtain

$$L \le C\sum_{j=0}^J h^{r-s}d_j^{1+s-r} + C\sum_{j=0}^J d_j^{s+N/2-1}h^{-s}\|E\|_{L^2(\Omega_j^{(1)})} = Ch\Theta(r-1-s) + \sum_{j=0}^J d_j^{s+N/2-1}h^{-s}\|E\|_{L^2(\Omega_j^{(1)})}, \tag{5.18}$$

where $\Theta(\gamma)$ is the following function:

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

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

$$\Theta(\gamma) = \sum_{j=0}^{J} \left(\frac{h}{d_j}\right)^{\gamma} \le C \begin{cases} \ln h^{-1} & \text{if } \gamma = 0, \\ \dfrac{1}{M^{\gamma}(1-2^{-\gamma})} & \text{if } \gamma > 0. \end{cases} \tag{5.19}$$

To estimate L, we now estimate $\|E\|_{L^2(\Omega_j^{(1)})}$ for each $0 \le j \le J$. By duality, we have

$$\|E\|_{L^2(\Omega_j^{(1)})} = \sup_{\varphi \in C_0^{\infty}(\Omega_j^{(1)}),\ \|\varphi\|_{L^2(\Omega)}=1} (E, \varphi). \tag{5.20}$$

For each $\varphi \in C_0^{\infty}(\Omega_j^{(1)})$, let $\Phi \in H_0^1(\Omega)$ be the solution of $\mathcal{L}\Phi = \varphi$. Then

$$(E, \varphi) = A(\Phi, E) = A(\Phi - \eta, E) = I_1 + I_2, \tag{5.21}$$

where

$$I_1 = A_{\Omega \setminus \Omega_j^{(2)}}(\Phi - \eta, E), \quad I_2 = A_{\Omega_j^{(2)}}(\Phi - \eta, E).$$

We will now estimate I_1 and I_2 separately. For I_1, by inequality (3.12), the approximation properties in Proposition 2.1, and (4.28) in Lemma 4.5, we see that

$$\begin{aligned} I_1 &\le C|||\Phi - \eta|||_{W_h^{1,\infty}(\Omega \setminus \Omega_j^{(2)})} |||E|||_{W_h^{1,1}(\Omega \setminus \Omega_j^{(2)})} \\ &\le Ch^r \|\Phi\|_{W^{1+r,\infty}(\Omega \setminus \Omega_j^{(2)})} |||E|||_{W_h^{1,1}(\Omega \setminus \Omega_j^{(2)})} \le Ch^r d_j^{1-r-N/2} |||E|||_{W_h^{1,1}(\Omega)}. \end{aligned} \tag{5.22}$$

For I_2, from (3.12) and Proposition 2.1, we have

$$\begin{aligned} I_2 &\le C|||\Phi - \eta|||_{H_h^1(\Omega_j^{(2)})} |||E|||_{H_h^1(\Omega_j^{(2)})} \le Ch\|\Phi\|_{H^2(\Omega)} |||E|||_{H_h^1(\Omega_j^{(2)})} \\ &\le Ch|||E|||_{H_h^1(\Omega_j^{(2)})}. \end{aligned} \tag{5.23}$$

Using (5.20)–(5.23), we obtain

$$\|E\|_{L^2(\Omega_j^{(1)})} \le Ch^r d_j^{1-r-N/2} |||E|||_{W_h^{1,1}(\Omega)} + Ch|||E|||_{H_h^1(\Omega_j^{(2)})}, \tag{5.24}$$

which leads to

$$\begin{aligned} &\sum_{j=0}^{J} d_j^{s+N/2-1} h^{-s} \|E\|_{L^2(\Omega_j^{(1)})} \\ &\quad \le C \sum_{j=0}^{J} (h/d_j)^{r-s} |||E|||_{W_h^{1,1}(\Omega)} + C \sum_{j=0}^{J} d_j^{s+N/2-1} h^{1-s} |||E|||_{H_h^1(\Omega_j^{(2)})} \\ &\quad \le C\Theta(r-s) |||E|||_{W_h^{1,1}(\Omega)} + CM^{N/2+s} h + CL/M. \end{aligned}$$

Hence, according to inequality (5.18) and choosing M sufficiently large, we have

$$L \le CM^{N/2+s} h + Ch\Theta(r-1-s) + C\Theta(r-s) |||E|||_{W_h^{1,1}(\Omega)}. \tag{5.25}$$

Substituting this into (5.17), we get

$$|||E|||_{W_h^{1,1}(\Omega),z,-s} \le CM^{N/2+s} h + Ch\Theta(r-1-s) + C\Theta(r-s) |||E|||_{W_h^{1,1}(\Omega)}. \tag{5.26}$$

The particular case where $s = 0$ in inequality (5.26) implies

$$|||E|||_{W_h^{1,1}(\Omega)} \le CM^{N/2}h + Ch\Theta(r-1) + C\Theta(r)|||E|||_{W_h^{1,1}(\Omega)}. \tag{5.27}$$

By choosing M large enough such that $C\Theta(r) < 1/2$, it follows from (5.27) that

$$|||E|||_{W_h^{1,1}(\Omega)} \le CM^{N/2}h + Ch\Theta(r-1) \le Ch\Theta(r-1). \tag{5.28}$$

Inserting (5.28) into (5.26), we obtain

$$\begin{aligned} |||E|||_{W_h^{1,1}(\Omega),z,-s} &\le CM^{N/2+s}h + Ch\Theta(r-1-s) + Ch\Theta(r-1)\Theta(r-s) \\ &\le Ch\left(\ln 1/h\right)^{\bar{\bar{s}}}, \end{aligned}$$

which proves the desired result (5.8). ∎

THEOREM 5.3. *Let $u \in W_0^{1,\infty}(\Omega)$ and $u_h \in S^h$ satisfy (2.9) and $0 \le s \le r$. Then there is a constant $C > 0$ such that for any $z \in \bar{\Omega}$,*

$$|\nabla(u-u_h)(z)| \le C\left(\ln h^{-1}\right)^{\bar{\bar{s}}} \inf_{\chi\in S^h} |||u-\chi|||_{W_h^{1,\infty}(\Omega),z,s}, \tag{5.29}$$

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

Proof. For any $x \in \bar{\Omega}$, let $z \in K_z$ for some $K_z \in \mathcal{J}_h$. Following a similar procedure as in the derivation of (5.2), we have

$$\begin{aligned} \left|\frac{\partial}{\partial x_i}(u-u_h)(z)\right| \le{}& \left\|\frac{\partial}{\partial x_i}(u-\chi)\right\|_{L^\infty(K_z)} \\ &+ Ch^{-N/2-1}\left\|\frac{\partial}{\partial x_i}(u-u_h)\right\|_{H^{-1}(K_z)}, \end{aligned} \tag{5.30}$$

where $\chi \in S^h$ is any function. Using integration by parts, it follows that

$$\begin{aligned} &h^{-N/2-1}\left\|\frac{\partial}{\partial x_i}(u-u_h)\right\|_{H^{-1}(K_z)} \\ &\quad= \sup_{\varphi\in C_0^\infty(K_z),\ \|\varphi\|_{H^1(K_z)}=1} \left(h^{-N/2-1}\frac{\partial}{\partial x_i}(u-u_h), \varphi\right) \\ &\quad= \sup_{\varphi\in C_0^\infty(K_z),\ \|\varphi\|_{H^1(K_z)}=1} \left(u-u_h, -h^{-N/2-1}\frac{\partial\varphi}{\partial x_i}\right). \end{aligned} \tag{5.31}$$

For any $\varphi \in C_0^\infty(K_z)$ satisfying $\|\varphi\|_{H^1(K_z)} = 1$, let $\hat{g}_z \in H_0^1(\Omega)$ be the solution of

$$\mathcal{L}^*\hat{g}_z = -h^{-N/2-1}\frac{\partial\varphi}{\partial x_i}. \tag{5.32}$$

Then, in view of (2.9) and (5.32), we have

$$\left(u-u_h, -h^{-N/2-1}\frac{\partial\varphi}{\partial x_i}\right) = A(u-u_h, \hat{g}_z) = A(u-\chi, \hat{g}_z - \hat{g}_{z,h}), \tag{5.33}$$

where $\hat{g}_{z,h} \in S^h$ is the finite element solution of $\hat{g}_z$

$$A(\chi, \hat{g}_z - \hat{g}_{z,h}) = 0 \qquad \forall\ \chi \in S^h. \tag{5.34}$$

Using (5.33), inequality (3.12), and the result in Lemma 5.4, we have

$$\begin{aligned}(5.35)\qquad &\left(u-u_h,-h^{-N/2-1}\frac{\partial\varphi}{\partial x_i}\right)\\ &\le C|||u-\chi|||_{W_h^{1,\infty}(\Omega),z,s}|||\hat{g}_z-\hat{g}_{z,h}|||_{W_h^{1,1}(\Omega),z,s}\\ &\le C\left(\ln h^{-1}\right)^{\bar{\bar{s}}}|||u-\chi|||_{W_h^{1,\infty}(\Omega),z,s}.\end{aligned}$$

From (5.30), (5.31), (5.35), and (5.36), the desired estimate (5.29) follows. ∎

LEMMA 5.4. *Let $\hat{g}_z\in H_0^1(\Omega)$ and $\hat{g}_{z,h}\in S^h$ satisfy* (5.32) *and* (5.34). *Then, for $0\le s\le r$ it holds that*

$$(5.36)\qquad |||\hat{g}_z-\hat{g}_{z,h}|||_{W_h^{1,1}(\Omega),z-s}\le C\left(\ln h^{-1}\right)^{\bar{\bar{s}}},$$

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

Proof. Let $M>1$ and J be as before. Again, for notational convenience, set $\hat{E}=\hat{g}_z-\hat{g}_{z,h}$. Similar to (5.17), we have

$$(5.37)\qquad |||\hat{E}|||_{W_h^{1,1}(\Omega),z,-s}\le CM^{N/2+s}+C\hat{L},$$

where

$$\hat{L}=\sum_{j=0}^{J}d_j^{N/2+s}h^{-s}|||\hat{E}|||_{H_h^1(\Omega_j)}.$$

Using (4.16), we obtain

$$(5.38)\qquad \hat{L}\le C\Theta(r-s)+C\sum_{j=0}^{J}d_j^{s+N/2-1}h^{-s}\|\hat{E}\|_{L^2(\Omega_j^{(1)})}.$$

The norm $\|\hat{E}\|_{L^2(\Omega_j^{(1)})}$ can be estimated in the same way as for $\|E\|_{L^2(\Omega_j^{(1)})}$ in (5.24). Thus we have

$$(5.39)\qquad \|\hat{E}\|_{L^2(\Omega_j^{(1)})}\le Ch^r d_j^{1-r-N/2}|||\hat{E}|||_{W_h^{1,1}(\Omega)}+Ch|||\hat{E}|||_{H_h^1(\Omega_j^{(2)})},$$

which implies

$$\begin{aligned}&\sum_{j=0}^{J}d_j^{s+N/2-1}h^{-s}\|\hat{E}\|_{L^2(\Omega_j^{(1)})}\\ &\quad\le C\sum_{j=0}^{J}(h/d_j)^{r-s}\,|||\hat{E}|||_{W_h^{1,1}(\Omega)}+C\sum_{j=0}^{J}d_j^{s+N/2-1}h^{1-s}|||\hat{E}|||_{H_h^1(\Omega_j^{(2)})}\\ &\quad\le C\Theta(r-s)|||\hat{E}|||_{W_h^{1,1}(\Omega)}+CM^{N/2+s}+CL/M.\end{aligned}$$

Hence, using the definition of $\hat{L}$ and choosing M sufficiently large, we see that

$$\hat{L}\le CM^{N/2+s}+C\Theta(r-s)+C\Theta(r-s)|||\hat{E}|||_{W_h^{1,1}(\Omega)}.$$

Substituting this into (5.37), we get

$$(5.40)\qquad |||\hat{E}|||_{W_h^{1,1}(\Omega),z,-s}\le CM^{N/2+s}+C\Theta(r-s)+C\Theta(r-s)|||\hat{E}|||_{W_h^{1,1}(\Omega)}.$$

Inequality (5.40) with $s = 0$ implies

$$|||\hat{E}|||_{W_h^{1,1}(\Omega)} \le CM^{N/2} + C\Theta(r) + C\Theta(r)|||E|||_{W_h^{1,1}(\Omega)}. \tag{5.41}$$

By choosing M large enough such that $C\Theta(r) < 1/2$, it follows from (5.41) that

$$|||\hat{E}|||_{W_h^{1,1}(\Omega)} \le CM^{N/2} + C\Theta(r) \le C. \tag{5.42}$$

Inserting (5.42) into (5.40), we obtain

$$|||\hat{E}|||_{W_h^{1,1}(\Omega),z,-s} \le CM^{N/2+s} + C\Theta(r-s) + C\Theta(r-s) \le C\left(\ln 1/h\right))^{\bar{\bar{s}}},$$

which proves the desired result (5.36). ∎

Using the results in Theorems 5.1 and 5.3, we have the following corollaries.

COROLLARY 5.5. *Let $u \in W_0^{1,\infty}(\Omega) \cap W^{1+r}(\Omega)$, $u_h \in S^h$ satisfy* (2.9)*, and $0 \le s \le r$. Then there is a constant $C > 0$ such that for any $z \in \bar{\Omega}$,*

$$|(u - u_h)(z)| \le Ch^{1+r}\left(\ln h^{-1}\right)^{\bar{s}} \|u\|_{W^{1+r,\infty}(\Omega),z,s}, \tag{5.43}$$

where $\bar{s}$ is the same as in Theorem 5.1.

Proof. Estimate (5.43) follows immediately from (5.1) and the property

$$\inf_{\chi \in S^h} |||u - \chi|||_{W_h^{1+r,\infty}(\Omega),z,s} \le Ch^r \|u\|_{W^{1+r,\infty}(\Omega),z,s}, \tag{5.44}$$

which can be easily obtained from Proposition 2.1. ∎

COROLLARY 5.6. *Let $u \in W_0^{1,\infty}(\Omega) \cap W^{1+r,\infty}(\Omega)$, $u_h \in S^h$ satisfy* (2.9)*, and $0 \le s \le r$. Then there is a constant $C > 0$ such that for any $z \in \bar{\Omega}$,*

$$|\nabla(u - u_h)(z)| \le Ch^r\left(\ln h^{-1}\right)^{\bar{\bar{s}}} \|u\|_{W^{1+r,\infty}(\Omega),z,s}, \tag{5.45}$$

where $\bar{\bar{s}}$ is the same is in Theorem 5.3.

Proof. Estimate (5.45) follows immediately from (5.29) and (5.44). ∎

REFERENCES

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

[2] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[3] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARTIN, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[4] G. A. BAKER, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

[5] F. BREZZI, G. MANZINI, D. MARINI, P. PIETRA, AND A. RUSSO, *Discontinuous Galerkin approximations for elliptic problems*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 365–378.

[6] P. CASTILLO, B. COCKBURN, I. PERUGIA, AND D. SCHÖTZAU, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[7] H. CHEN AND Z. CHEN, *Stability and convergence of mixed discontinuous finite element methods for second order differential problems*, J. Numer. Math., 11 (2003), pp. 253–287.

[8] H. CHEN, Z. CHEN, AND B. LI, *Numerical study of the hp version of mixed discontinuous finite element methods for reaction diffusion problems: 1D case*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 525–553.

[9] Z. CHEN, *On the relationship of various discontinuous finite element methods for second-order elliptic equations*, East-West, Numer. Math., 9 (2001), pp. 99–122.

[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[11] B. Cockburn, G. E. Karniadakis, and C. W. Shu, *Discontinuous Galerkin Methods, Theory, Computation, and Applications*, Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin, 2000.

[12] J. Douglas and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences, Lecture Notes in Phys. 58, R. Glowinski and J. L. Lions, eds., Springer-Verlag, Berlin, 1976, pp. 207–216.

[13] J. Frehse and R. Rannacher, *Eine L^1-fehlerabschätzung für diskrete grundlösungen in der methode der finiten elemente*, Tagungsband finite elemente, 89 (1976), pp. 92–114.

[14] G. Kanschat and R. Rannacher, *Local error analysis of the interior penalty discontinuous Galerkin method for second order elliptic problems*, J. Numer. Math., 10 (2002), pp. 249–274.

[15] J. A. Nitsche, *Über ein variationsprinzip zur lösung von Dirichlet-problemen bei verwendung von teilräumen, die keinen Randbedingungen unteworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.

[16] J. T. Oden, I. Babuška and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.

[17] R. Rannacher and L. R. Scott, *Some optimal error estimates for piecewise linear finite element approximations*, Math. Comp., 38 (1982), pp. 437–445.

[18] B. Rivière, M. F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.

[19] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids:* I, Global Estimates. Math. Comp., 67 (1998), pp. 877–899.

[20] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids:* II, Interior Estimates. SIAM J. Numer. Anal., 38 (2000), pp. 1269–1293.

[21] A. H. Schatz and L. B. Wahlbin, *Interior maximum norm-estimates for finite element methods*, II, Math. Comp., 31 (1977), pp. 414–442.

[22] V. Solonnikov, *On Green's matrices for elliptic boundary value problems*, I, Proc. Steklov Inst. Math., 110 (1970), pp. 123–170.

[23] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1167–1188

STABILITY AND CONVERGENCE OF COLLOCATION SCHEMES FOR RETARDED POTENTIAL INTEGRAL EQUATIONS*

PENNY J. DAVIES† AND DUGALD B. DUNCAN‡

Abstract. Time domain boundary integral formulations of transient scattering problems involve retarded potential integral equations. Solving such equations numerically is both complicated and computationally intensive, and numerical methods often prove to be unstable. Collocation schemes are easier to implement than full finite element formulations, but little appears to be known about their stability and convergence. Here we derive and analyze some new stable collocation schemes for the single layer equation for transient acoustic scattering, and use (spatial) Fourier and (temporal) Laplace transform techniques to demonstrate that such stable schemes are second order convergent.

Key words. convergence, stability, retarded potential, boundary integral

AMS subject classifications. 65M12, 65R20, 78M05, 78M15

DOI. 10.1137/S0036142901395321

1. Introduction.

The scalar integral equation for $u(\boldsymbol{x},t)$ on $\Gamma\times(0,T)$

$$\int_\Gamma \frac{u(\boldsymbol{x}',t-|\boldsymbol{x}'-\boldsymbol{x}|)}{|\boldsymbol{x}'-\boldsymbol{x}|}\,d\boldsymbol{x}' = a(\boldsymbol{x},t) \tag{1.1}$$

is the single layer potential equation for transient acoustic scattering from the two-dimensional surface $\Gamma\subset\mathbb{R}^3$ [27, sect. 2.3]. Here a is given on $\Gamma\times(0,T)$ for fixed $T>0$, and u and a satisfy the causality condition

$$u\equiv 0,\quad a\equiv 0 \quad \text{for all } t\le 0. \tag{1.2}$$

Once the potential u has been calculated on Γ, the scattered field can be computed anywhere in $\mathbb{R}^3$. The time argument of the integrand in (1.1) is delayed or retarded, and such equations are commonly called retarded potential integral equations (RPIEs). They also arise in boundary integral formulations of electromagnetic scattering problems [2, 21, 28, 29, 30, 31].

Existence, uniqueness, and well-posedness results for (1.1) are given in [3, 19, 20, 27]. A similar argument to that used by Lubich [27, sect. 2.3] in the case that Γ is a smooth, closed surface (based on results of Bamberger and Ha-Duong [3, Prop. 3]) can be used to deduce the following result from [19] when Γ is a flat plate. We use the notation

$$H^m_*(0,T) = \big\{ f\,\big|_{(0,T)} : f\in H^m(\mathbb{R}) \text{ with } f\equiv 0 \text{ on } (-\infty,0)\big\},$$

(this space is called H^m_0 in [27, Chap. 2]), where $H^m(\mathbb{R})$ denotes the usual Sobolev space of order m [1, Chap. 6].

PROPOSITION 1.1 (Ha-Duong [19, Thm. 3], Lubich [27, sect. 2.3]). *For temporally smooth data $a(\cdot,t)\in H^{1/2}(\Gamma)$ which vanish near $t=0$, the RPIE* (1.1) *has*

*Received by the editors September 17, 2001; accepted for publication (in revised form) November 2, 2003; published electronically September 18, 2004.
http://www.siam.org/journals/sinum/42-3/39532.html
†Department of Mathematics, University of Strathclyde, 26 Richmond St., Glasgow, G1 1XH, UK (penny@maths.strath.ac.uk).
‡Department of Mathematics, Heriot-Watt University, Riccarton, Edinburgh, EH14 4AS, UK (D.B.Duncan@ma.hw.ac.uk).

a unique smooth solution $u(\cdot,\, t) \in H^{-1/2}(\Gamma)$. *Moreover there exists a constant* C *depending only on* T *and* Γ *such that*

$$\|u\|_{H_*^m\left(0,T;H^{-1/2}(\Gamma)\right)} \leq C\, \|a\|_{H_*^{m+1}\left(0,T;H^{1/2}(\Gamma)\right)} \quad (m \in \mathbb{R})\,.$$

The spaces $H_*^m\,(0,T;X)$ and their norms are as defined by Lions and Magenes [25, Chaps. 1.1, 4.2]; namely

$$\|f\|^2_{H_*^m(0,T;X)} = \sum_{k=0}^{m} \|f^{(k)}\|^2_{L^2(0,T;X)}, \tag{1.3}$$

where $f^{(k)} = \partial^k f/\partial t^k$ and

$$\|f\|_{L^2(0,T;X)} = \left(\int_0^T \|f\|_X^2\, dt\right)^{1/2}.$$

Various numerical methods for computing u have been reported in the literature. Bamberger and Ha-Duong [3] describe a variational method for the problem when Γ is closed and smooth, one that is based on the coercivity of a bilinear form corresponding to a full Galerkin approximation in time and space. This approach has been extended to deal with the case when Γ is a flat surface by Ha-Duong [19], who also gives a comprehensive survey of the numerical analysis of such schemes in [20]. However, the variational method is complicated (and costly) to implement since it involves calculating five-dimensional integrals over $\Gamma\times\Gamma\times(0,T)$, and collocation schemes are frequently used for RPIEs in electromagnetic scattering problems [28, 29, 31]. In both approaches it takes $O(N_T\, N_S^2)$ flops to compute the solution up to time $T = N_T\,\Delta t$, where N_S is the number of spatial degrees of freedom used in the approximation, so RPIE algorithms are highly computationally intensive. Recently Michielssen and co-workers [15, 16, 26] have introduced "fast methods" for time-dependent boundary integral equations (BIEs) such as (1.1) that reduce the operation count to $O(N_T\, N_S^{3/2}\, \log N_S)$ (for a two-level scheme), or $O(N_T\, N_S\, \log^2 N_S)$ (multilevel). Although complicated to implement, these make the BIE approach for time-dependent scattering problems viable compared to methods based on solving PDEs in three-dimensional space.

The usefulness of collocation methods is often limited by the fact that they tend to exhibit numerical instabilities (see, e.g., [22, sect. 5]). Fourier analysis [6, 7, 10] indicates that the most likely cause of instability is the inaccurate approximation of (1.1). Here we present two new *stable* collocation methods for the problem (1.1)–(1.2). Our other main result is a proof that these schemes converge. The proof relies on the spatial Fourier transform of (1.1) being a convolution equation in time, and we use the Laplace and Z transform techniques of Lubich [27] to bound the Fourier transform of the approximation error. We then use classical estimates derived by Bramble and Hilbert [4] and Thomée [33] to bound the discrete norm of the error as the mesh-size tends to zero. We believe that this is the first convergence proof for an actual collocation RPIE scheme.

2. Preliminaries. We now describe the notation and some basic results used in the manuscript. The stability and convergence analysis in sections 4–5 is for the scalar RPIE (1.1) posed on an infinite flat surface, i.e., for

$$\int_{\mathbb{R}^2} \frac{u(\boldsymbol{x}',\, t-|\boldsymbol{x}'-\boldsymbol{x}|)}{|\boldsymbol{x}'-\boldsymbol{x}|}\, d\boldsymbol{x}' = a(\boldsymbol{x},t) \quad \text{on } \mathbb{R}^2\times(0,T)\,, \tag{2.1}$$

where u and a satisfy (1.2).

The singularity in the integrand can be removed by the polar coordinate transformation $\boldsymbol{x}' = \boldsymbol{x} + R\,\boldsymbol{e}_\theta$, where $\boldsymbol{e}_\theta = (\cos\theta,\, \sin\theta)$ (see also [5, 9]). When $\Gamma = \mathbb{R}^2$ causality (1.2) results in the RPIE

$$\int_0^t \int_0^{2\pi} u(\boldsymbol{x} + R\,\boldsymbol{e}_\theta,\, t-R)\; d\theta\; dR = a(\boldsymbol{x}, t). \tag{2.2}$$

If Γ is finite, then the integral is over the appropriate region of (R, θ)-space (which depends on $\boldsymbol{x}$).

2.1. Continuous and discrete spatial Fourier transforms. The continuous Fourier transform (CFT) of a function $g \in L^2(\mathbb{R}^2)$ is $\widehat{g} \in L^2(\mathbb{R}^2)$ defined by

$$\widehat{g}(\boldsymbol{\omega}) \equiv \int_{\mathbb{R}^2} g(\boldsymbol{x})\, e^{-i\boldsymbol{x}\cdot\boldsymbol{\omega}}\, d\boldsymbol{x},$$

and the inverse transform is

$$g(\boldsymbol{x}) = \frac{1}{4\pi^2} \int_{\mathbb{R}^2} \widehat{g}(\boldsymbol{\omega})\, e^{i\boldsymbol{x}\cdot\boldsymbol{\omega}}\, d\boldsymbol{\omega}.$$

Note that this definition of the CFT is that used by Bramble and Hilbert [4] and differs from that of [1] by a factor of 2π. The CFT can be used to define the norm in $H^r(\mathbb{R}^2)$ when $r \geq 0$:

$$\|g\|_r = \|(1+\omega)^r\, \widehat{g}\|_{\mathcal{F}} \equiv \frac{1}{2\pi} \left(\int_{\mathbb{R}^2} |(1+\omega)^r\, \widehat{g}(\boldsymbol{\omega})|^2\; d\boldsymbol{\omega} \right)^{1/2}, \tag{2.3}$$

where $\omega = |\boldsymbol{\omega}|$ (see [27, sect. 2.1]). When $r = 0$ this is the Parseval–Plancherel identity. The discrete Fourier transform (DFT) of a function g evaluated at the nodes of a uniform $h \times h$ space mesh in $\mathbb{R}^2$ is denoted by $\tilde{g}$ and defined by

$$\tilde{g}(\boldsymbol{\omega}) = h^2 \sum_{j,k=-\infty}^{\infty} g(\boldsymbol{x}_{j,k}) e^{-i\boldsymbol{\omega}\cdot\boldsymbol{x}_{j,k}} \tag{2.4}$$

for $\boldsymbol{\omega} \in S_h = \{(\omega_1, \omega_2) : |\omega_1|, |\omega_2| \leq \pi/h\}$, where $(j,k) \in \mathbb{Z}^2$ and $\boldsymbol{x}_{j,k} = (jh, kh)$. The function $\tilde{g}$ is $2\pi/h$ periodic in each component of $\boldsymbol{\omega}$. The DFT is defined for $g \in H^r(\mathbb{R}^2)$ with $r > 1$ [4, sect. 4] and satisfies the discrete analogue of Parseval's identity:

$$\|\tilde{g}\|_{\mathcal{F}_h} = \|g\|_h, \tag{2.5}$$

where

$$\|\tilde{g}\|_{\mathcal{F}_h} = \left(\frac{1}{4\pi^2} \int_{S_h} |\tilde{g}(\boldsymbol{\omega})|^2\, d\boldsymbol{\omega} \right)^{1/2}$$

is the discrete Fourier norm and

$$\|g\|_h = \left(h^2 \sum_{j,k} |g(\boldsymbol{x}_{j,k})|^2 \right)^{1/2}$$

is the discrete L^2 norm.

The following results due to Bramble and Hilbert [4] link the discrete and continuous Fourier transforms of a function.

PROPOSITION 2.1 (see [4, Theorem 5]). *Let $g \in H^r(\mathbb{R}^2)$ for $r > 1$. Then there exists a constant C independent of h and g such that*

$$\|\tilde{g} - \widehat{g}\|_{\mathcal{F}_h} \leq Ch^r \|g\|_r. \tag{2.6}$$

PROPOSITION 2.2 (Poisson sum formula [4, Theorem 6]). *Let $g \in H^r(\mathbb{R}^2)$ for $r > 1$. Then*

$$\tilde{g}(\boldsymbol{\omega}) = \sum_{j,k} \widehat{g}(\boldsymbol{\omega} + 2\pi(j,k)/h) \quad a.e. \tag{2.7}$$

2.2. Laplace and *Z* transforms in time. The Laplace transform of the causal function $f(t)$ (i.e., $f(t) \equiv 0$, $t < 0$) is

$$\bar{f}(s) = \int_0^\infty f(t) e^{-st}\, dt,$$

where $s = \sigma + i\eta$ with $\sigma > 0$ and $\eta \in \mathbb{R}$. Throughout the paper σ is always assumed to be the **same** fixed positive constant. The Parseval Laplace identity is

$$\|e^{-\sigma t} f(t)\|_{L^2(\mathbb{R}^+)} = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{\infty} |\bar{f}(\sigma + i\eta)|^2\, d\eta \right)^{1/2}. \tag{2.8}$$

This is equivalent to the one-dimensional version of (2.3) applied to the causal function $e^{-\sigma t} f(t)$ with $r = 0$. It follows that if $f \in H^m_*(\mathbb{R}^+)$, then

$$C \int_{-\infty}^{\infty} (1+|s|)^{2m} |\bar{f}|^2\, d\eta \leq 2\pi \sum_{k=0}^{m} \left\| \frac{\partial^k}{\partial t^k} \left(e^{-\sigma t} f(t) \right) \right\|^2_{L^2(\mathbb{R}^+)} \leq \int_{-\infty}^{\infty} (1+|s|)^{2m} |\bar{f}|^2\, d\eta, \tag{2.9}$$

where the constant C depends only on σ and m.

The Z transform is the discrete version of the Laplace transform defined by

$$Zf(s) = \sum_{n=0}^{\infty} f(n\Delta t) e^{-sn\Delta t}, \tag{2.10}$$

where again $s = \sigma + i\eta$, but now $\eta \in [-\pi/\Delta t, \pi/\Delta t]$. The inversion formula is

$$f(n\Delta t) = \frac{\Delta t}{2\pi i} \int_{-\pi/\Delta t}^{\pi/\Delta t} e^{n\Delta t(\sigma + i\eta)} Zf(\sigma + i\eta)\, d\eta \tag{2.11}$$

for $n \in \mathbb{N}$. The Z and Laplace transforms are related by the Poisson sum formula

$$\Delta t\, Zf(s) = \sum_{k=-\infty}^{\infty} \bar{f}\left(s + i\, \frac{2\pi k}{\Delta t} \right), \tag{2.12}$$

a one-dimensional version of (2.7), valid for $e^{-\sigma t} f(t) \in H^r(\mathbb{R}^+)$ with $r > 1/2$.

2.3. Fourier transformed RPIE. We suppose that $a(\cdot,t)$, $u(\cdot,t) \in L^2(\mathbb{R}^2)$ for $t \in (0,T)$ and take the CFT of the RPIE (2.2). This gives the first kind convolution Volterra integral equation

$$2\pi \int_0^t \widehat{u}(\boldsymbol{\omega}, t-R)\, J_0(\omega R)\, dR = \widehat{a}(\boldsymbol{\omega}, t) \quad \text{for } \boldsymbol{\omega} \in \mathbb{R}^2,\ t \in (0,T), \tag{2.13}$$

where J_0 is the first kind Bessel function of order zero. We use the identity [18, sect. 8.41]

$$J_0(z) = \frac{1}{2\pi} \int_0^{2\pi} e^{iz\sin\theta}\, d\theta \tag{2.14}$$

to obtain the integral equation.

The results of [27, sect. 2.1] apply to give the following result for the infinite flat surface, analogous to Proposition 1.1 for the finite surface.

LEMMA 2.3. *Suppose that* $a \in H_*^{m+1}\left(0,T; H^{r+1}(\mathbb{R}^2)\right)$ *for integer* $m \geq 0$ *and* $r \in [0,\infty)$. *Then the solution* $u(\boldsymbol{x},t)$ *defined by* (2.17) *satisfies* $u \in H_*^m\left(0,T; H^r(\mathbb{R}^2)\right)$.

Proof. We essentially use the operator version of [27, Lemma 2.1] to obtain this result. We first extend the range of definition of a in time from $(0,T)$ to $(0,\infty)$ so that

$$\|a\|_{H_*^{m+1}(\mathbb{R}^+; H^{r+1}(\mathbb{R}^2))} \leq C \|a\|_{H_*^{m+1}(0,T; H^{r+1}(\mathbb{R}^2))} \tag{2.15}$$

(see, e.g., [1, Thm. 6.3.5]). Then extending the definition of the convolution (2.13) to $\mathbb{R}^+$ and taking the Laplace transform in time gives

$$\frac{2\pi}{\sqrt{\omega^2+s^2}}\, \bar{\widehat{u}}(\boldsymbol{\omega}, s) = \bar{\widehat{a}}(\boldsymbol{\omega}, s), \tag{2.16}$$

where the overbar denotes Laplace transform in t, and s is the Laplace transform parameter. Hence

$$\bar{\widehat{u}}(\boldsymbol{\omega}, s) = \frac{\sqrt{\omega^2+s^2}}{2\pi}\, \bar{\widehat{a}}(\boldsymbol{\omega}, s) \tag{2.17}$$

and so

$$|\bar{\widehat{u}}(\boldsymbol{\omega}, s)|^2 \leq \frac{\omega^2+|s|^2}{4\pi^2}\, |\bar{\widehat{a}}(\boldsymbol{\omega}, s)|^2 \leq \frac{(1+\omega)^2(1+|s|)^2}{4\pi^2}\, |\bar{\widehat{a}}(\boldsymbol{\omega}, s)|^2. \tag{2.18}$$

It follows from definition (1.3) that

$$\|u\|^2_{H_*^m(0,T;H^r(\mathbb{R}^2))} \leq e^{2\sigma T} \sum_{k=0}^m \int_0^\infty e^{-2\sigma t} \|u^{(k)}(\cdot,t)\|^2_{H^r(\mathbb{R}^2)}\, dt \equiv I_1.$$

The characterization (2.3) of $H^r(\mathbb{R}^2)$ in terms of Fourier transforms gives

$$I_1 = \frac{e^{2\sigma T}}{4\pi^2} \sum_{k=0}^m \int_0^\infty \int_{\mathbb{R}^2} e^{-2\sigma t} (1+\omega)^{2r} |\widehat{u}^{(k)}(\boldsymbol{\omega}, t)|^2\, d\boldsymbol{\omega}\, dt$$

and reversing the order of integration and using the Laplace Parseval equality (2.8) result in

$$I_1 = \frac{e^{2\sigma T}}{8\pi^3} \sum_{k=0}^m \int_{\mathbb{R}^2} \int_{\mathbb{R}} (1+\omega)^{2r} |s|^{2k} |\bar{\widehat{u}}(\boldsymbol{\omega}, s)|^2\, d\eta\, d\boldsymbol{\omega}.$$

Now using the inequality (2.18) and reversing the steps above we get

$$I_1 \le \frac{e^{2\sigma T}}{8\pi^5} \sum_{k=0}^{m+1} \int_{\mathbb{R}^2} \int_{\mathbb{R}} (1+\omega)^{2r+2} |s|^{2k} |\bar{\widehat{a}}(\boldsymbol{\omega}, s)|^2 \, d\eta \, d\boldsymbol{\omega}$$
$$= \frac{e^{2\sigma T}}{\pi^2} \sum_{k=0}^{m+1} \int_0^\infty e^{-2\sigma t} \|a^{(k)}(\cdot, t)\|^2_{H^{r+1}(\mathbb{R}^2)} \, dt \le \frac{e^{2\sigma T}}{\pi^2} \|a\|^2_{H^{m+1}_*(\mathbb{R}^+; H^{r+1}(\mathbb{R}^2))}.$$

Finally we use the extension result (2.15) to get

$$\|u\|_{H^m_*(0,T;H^r(\mathbb{R}^2))} \le \sqrt{I_1} \le C \, \|a\|_{H^{m+1}_*(0,T;H^{r+1}(\mathbb{R}^2))},$$

where C depends only on m, r, σ, and T, and the result follows. ∎

We also require the following pointwise bound on $\widehat{u}$.

LEMMA 2.4. *Under the conditions of the previous lemma, there exists a constant C such that*

$$|\widehat{u}(\boldsymbol{\omega}, t)| \le \frac{e^{\sigma T}}{\sqrt{2\pi}} \|e^{-\sigma t} \widehat{u}(\boldsymbol{\omega}, \cdot)\|_{H^1(\mathbb{R}^+)} \le C \, (1+\omega) \, \|\widehat{a}(\boldsymbol{\omega}, \cdot)\|_{H^2(\mathbb{R}^+)}$$

for $t \in (0, T)$.

Proof. The first inequality follows from the standard result

$$|f(t)| \le \frac{1}{\sqrt{2\pi}} \|f\|_{H^1(\mathbb{R}^+)} \tag{2.19}$$

[1, Ex. 6.4.5] applied with $f(t) = e^{-\sigma t} \widehat{u}(\boldsymbol{\omega}, t)$. Multiplying (2.18) by $(1+|s|)^2$, where $s = \sigma + i\eta$, and using the norm equivalence (2.9), gives

$$\|e^{-\sigma t} \widehat{u}(\boldsymbol{\omega}, t)\|_{H^1(\mathbb{R}^+)} \le C_1 \, (1+\omega) \|e^{-\sigma t} \widehat{a}(\boldsymbol{\omega}, t)\|_{H^2(\mathbb{R}^+)} \le C_2 \, (1+\omega) \|\widehat{a}(\boldsymbol{\omega}, \cdot)\|_{H^2(\mathbb{R}^+)}$$

for constants C_1 and C_2, which results in the second inequality. ∎

3. Algorithms. Because we are primarily interested in the analysis of RPIE algorithms here, we concentrate on the case $\Gamma = \mathbb{R}^2$. The restriction to finite Γ should be obvious. The RPIE (2.1) is approximated on a square space grid of side h and uniformly spaced time levels $t^n = n\Delta t$ for $n \in \mathbb{Z}^+$ in terms of piecewise constant or linear space and time basis functions, i.e., the approximate solution is expanded as

$$u(\boldsymbol{x}, t) \approx U(\boldsymbol{x}, t) = \sum_{m \ge 1} \sum_{j,k} U^m_{j,k} \, \phi^{[\alpha]}_j(x) \, \phi^{[\alpha]}_k(y) \, \psi^{[\beta]}_m(t)$$

for $\boldsymbol{x} = (x, y) \in \mathbb{R}^2$, where $\alpha, \beta \in \{0, 1\}$ indicate the orders of the space and time basis functions respectively. The spatial basis functions are defined by

$$\phi^{[\alpha]}_j(x) = \phi^{[\alpha]}(x/h - j),$$

where

$$\phi^{[0]}(z) = \begin{cases} 1 & \text{if } |z| < 1/2, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \phi^{[1]}(z) = \begin{cases} 1 - |z| & \text{if } |z| < 1, \\ 0 & \text{otherwise} \end{cases}$$

are the standard constant "pulse" and linear "hat" basis functions. The basis functions in time are

$$\psi^{[0]}_m(t) = \phi^{[0]}(t/\Delta t - m + 1/2) \quad \text{and} \quad \psi^{[1]}_m(t) = \phi^{[1]}(t/\Delta t - m).$$

When the temporal basis functions are piecewise linear (resp., constant) the approximate solution $U(\boldsymbol{x},t)$ is evaluated at time $t=t^n$ (resp., $t=t^{n-1/2}$), where n is an integer. Hence the coefficients $U^n_{j,k}$ correspond to the approximate solution at time $t=(n-(1-\beta)/2)\Delta t$ and

$$U(\boldsymbol{x},t^{n-(1-\beta)/2})=\sum_{j,k}U^n_{j,k}\,\phi^{[\alpha]}_j(x)\,\phi^{[\alpha]}_k(y)\,.$$

Note that the approximate solution automatically satisfies the causality condition $U(x,t)=0$ for $t\le 0$.

We shall consider the four schemes denoted by SαTβ, for $\alpha,\,\beta\in\{0,1\}$ to indicate the degree of the basis functions in space ("S") and time ("T"). They are obtained by substituting U for u in the RPIE (2.1), evaluating (collocating) at each space mesh node $\boldsymbol{x}=\boldsymbol{x}_{p,q}$ and time level $t=t^n$, and carrying out all the required integrations *exactly*. This can be written as

$$a(\boldsymbol{x}_{p,q},t^n)=\int_{\mathbb{R}^2}\frac{U(\boldsymbol{x}'+\boldsymbol{x}_{p,q},t^n-|\boldsymbol{x}'|)}{|\boldsymbol{x}'|}\,d\boldsymbol{x}'=\sum_{m=0}^{n-1}\sum_{j,k}C^m_{j,k}\,U^{n-m}_{p+j,q+k},\tag{3.1}$$

where the coefficients

$$C^m_{j,k}=\int_{\mathbb{R}^2}\frac{\phi^{[\alpha]}_j(x')\,\phi^{[\alpha]}_k(y')\,\psi^{[\beta]}_m(|\boldsymbol{x}'|)}{|\boldsymbol{x}'|}\,d\boldsymbol{x}'\tag{3.2}$$

are evaluated exactly. Because of the finite support of the spatial and temporal basis functions, $C^m_{j,k}$ is zero unless $\big||\boldsymbol{x}_{j,k}|-t^{m-(1-\beta)/2}\big|\le(1+\beta)\Delta t/2+(1+\alpha)h/\sqrt{2}$. Also, it follows from the definition of the basis functions that

$$C^m_{j,k}=C^m_{k,j}=C^m_{-j,k}=C^m_{j,-k}.$$

The approximation scheme can hence be written as

$$\sum_{m=0}^{n-1}\mathbb{Q}^m U^{n-m}_{p,q}=a(\boldsymbol{x}_{p,q},t^n),\tag{3.3}$$

where $\mathbb{Q}^m=\sum_{j,k}C^m_{j,k}\,S^j_x\,S^k_y$ for $m\ge 0$ are discrete operators written in terms of unit shift operators S_x and S_y defined by $S^j_x\,U_{p,q}=U_{p+j,q}$, $S^k_y\,U_{p,q}=U_{p,q+k}$.

The sum can be rearranged to give

$$\mathbb{Q}^0U^n_{p,q}=a(\boldsymbol{x}_{p,q},t^n)-\sum_{m=1}^{n-1}\mathbb{Q}^{n-m}U^m_{p,q}\qquad\text{for } n\ge 1$$

and solved at successive time-levels, provided the difference operator $\mathbb{Q}^0$ is invertible. We examine this and other aspects of these schemes in the next section.

4. Stability. We use Fourier methods developed in [6, 7, 10] to analyze the stability of each of the schemes of the previous section. The analysis is for the RPIE (2.1) on an infinite uniform space mesh with uniform time steps, and is analogous to a von Neumann stability analysis for a PDE approximation. Results for the more general RPIE (1.1) approximated on nonuniform grids cannot be obtained this way. However it is clear that infinite mesh stability is necessary for a scheme to be stable in more general circumstances as the mesh is refined [6, sect. 4].

4.1. DFT of the schemes. Using definition (2.4), the DFT of the difference equation (3.3) over the space mesh node points is

$$\sum_{m=0}^{n-1} q_m(\boldsymbol{\omega})\, \tilde{U}^{n-m}(\boldsymbol{\omega}) = \tilde{a}(\boldsymbol{\omega}, t^n) \tag{4.1}$$

for all $\boldsymbol{\omega} \in S_h$ and $n \geq 1$, where the functions $q_m(\boldsymbol{\omega})$ are the discrete transforms of the difference operators $\mathbb{Q}^m$ and are given by

$$q_m(\boldsymbol{\omega}) = \sum_{j,k} C^m_{j,k}\, e^{ih(j\omega_1 + k\omega_2)} \quad \text{for } m \geq 0, \tag{4.2}$$

where the $C^m_{j,k}$ are defined in (3.2). If $q_0(\boldsymbol{\omega}) \neq 0$ then the solution of the scalar convolution sum equation (4.1) is

$$\tilde{U}^n(\boldsymbol{\omega}) = \frac{1}{q_0(\boldsymbol{\omega})} \sum_{m=1}^{n} p_m(\boldsymbol{\omega})\, \tilde{a}(\boldsymbol{\omega}, t^{n-m+1}), \tag{4.3}$$

where the coefficients p_n are defined recursively for all $\boldsymbol{\omega} \in S_h$ by

$$p_1(\boldsymbol{\omega}) = 1\ , \quad p_n(\boldsymbol{\omega}) = \frac{-1}{q_0(\boldsymbol{\omega})} \sum_{m=1}^{n-1} q_m(\boldsymbol{\omega})\, p_{n-m}(\boldsymbol{\omega}) \quad \text{for } n \geq 2. \tag{4.4}$$

The assumption that $q_0(\boldsymbol{\omega}) \neq 0$ for all $\boldsymbol{\omega} \in S_h$ is equivalent to the invertibility of the difference operator $\mathbb{Q}^0$ [8, 14]. The following two lemmas provide more information about q_0 and the other q_m.

LEMMA 4.1. *The coefficients q_m for scheme $S\alpha T\beta$ defined in* (4.2) *satisfy*

$$q_m(\boldsymbol{\omega}) = 2\pi \sum_{j,k} \Phi^{[\alpha]}(h\omega_1 + 2\pi j)\, \Phi^{[\alpha]}(h\omega_2 + 2\pi k)\, I^m_{j,k}(\boldsymbol{\omega}), \tag{4.5}$$

where

$$\Phi^{[0]}(z) = 2\sin(z/2)/z\ , \quad \Phi^{[1]}(z) = 2(1-\cos z)/z^2 \tag{4.6}$$

are the Fourier transforms of the basis functions $\phi^{[\alpha]}(x)$ defined in section 3 *and*

$$I^m_{j,k}(\boldsymbol{\omega}) = \int_0^\infty \psi^{[\beta]}_{m^*}(R)\, J_0\left(R\,|\boldsymbol{\omega} + 2\pi(j,k)/h|\right)\, dR \quad \textit{for } m \geq 0, \tag{4.7}$$

where $m^* = m + 1 - \beta$.

Proof. Recall the labelling of the schemes used in section 3: $\alpha, \beta \in \{0, 1\}$ indicate the order of the space and time basis functions respectively. We first substitute the approximate solution U for u in the left-hand side of the RPIE (2.2) at time $t = t^n$ and use this to define

$$A(\boldsymbol{x}, t^n) = \sum_{m=0}^{n-1} \sum_{j,k} U^{n-m}_{j,k} \int_0^\infty \psi^{[\beta]}_{m^*}(R) \int_0^{2\pi} \phi^{[\alpha]}_j(x + R\cos\theta)\, \phi^{[\alpha]}_k(y + R\sin\theta)\, d\theta\, dR.$$

Note that it follows from the numerical scheme (3.1) that $A(\boldsymbol{x}_{p,q}, t^n) = a(\boldsymbol{x}_{p,q}, t^n)$ on the grid, resulting in

$$\tilde{A} = \tilde{a}. \tag{4.8}$$

Taking the CFT of A with respect to $\boldsymbol{x}$ gives

$$\sum_{m=0}^{n-1}\sum_{j,k} U_{j,k}^{n-m}\int_0^\infty \psi_{m^*}^{[\beta]}(R)\int_0^{2\pi}\Phi^{[\alpha]}(h\omega_1)\,\Phi^{[\alpha]}(h\omega_2)\,e^{i\boldsymbol{\omega}\cdot(h(j,k)+R\boldsymbol{e}_\theta)}\,d\theta\,dR = \widehat{A}(\boldsymbol{\omega},t^n),$$

which can be rearranged as

$$2\pi\sum_{m=0}^{n-1}\tilde{U}^{n-m}(\boldsymbol{\omega})\Phi^{[\alpha]}(h\omega_1)\,\Phi^{[\alpha]}(h\omega_2)\int_0^\infty \psi_{m^*}^{[\beta]}(R)J_0(\omega R)dR = \widehat{A}(\boldsymbol{\omega},t^n)$$

using the definition (2.4) and the Bessel function identity (2.14). Finally we apply the Poisson sum formula (2.7) to both sides of this equation and use (4.8) and the periodicity of the DFT $\tilde{U}^m(\boldsymbol{\omega}+2\pi(j,k)/h)=\tilde{U}^m(\boldsymbol{\omega})$ to obtain

$$2\pi\sum_{j,k}\sum_{m=0}^{n-1}\tilde{U}^{n-m}(\boldsymbol{\omega})\Phi^{[\alpha]}(h\omega_1+2\pi j)\,\Phi^{[\alpha]}(h\omega_2+2\pi k)I_{j,k}^m(\boldsymbol{\omega}) = \tilde{a}(\boldsymbol{\omega},t^n).$$

The result follows by comparing this with (4.1). ∎

LEMMA 4.2. *The coefficient $q_0(\boldsymbol{\omega})\ge C\Delta t$ for all $\boldsymbol{\omega}\in S_h$ where $C>0$ depends only on the mesh ratio $\Delta t/h$ (which is a fixed number in the scheme).*

Proof. We first show that each term in the summation (4.5) for q_0 is nonnegative for each of the four schemes under consideration. Clearly $\Phi^{[\alpha]}(h\omega_1+2\pi j)\ge 0$ for all $j\in\mathbb{Z}$, $\boldsymbol{\omega}\in S_h$ by definition (4.6). Also, (4.7) with $m=0$ can be written as $I_{j,k}^0=\omega_{j,k}^{-1}F^{[\beta]}(\Delta t\omega_{j,k})$, where $\omega_{j,k}=|\boldsymbol{\omega}+2\pi(j,k)/h|$,

$$F^{[0]}(t)=\int_0^t J_0(s)\,ds \quad\text{and}\quad F^{[1]}(t)=\int_0^t(1-s/t)J_0(s)ds=\int_0^t s^{-1}J_1(s)\,ds.$$

It follows from results in [32, sect. 5] that $F^{[\beta]}(t)>0$ for $\beta\in\{0,1\}$ and all $t>0$, and hence each term in the summation (4.5) for q_0 is nonnegative.

Pulling out the term with $j=k=0$ and using the definition (4.6) then gives

$$q_0(\boldsymbol{\omega})\ge 2\pi\Phi^{[\alpha]}(h\omega_1)\,\Phi^{[\alpha]}(h\omega_2)\,I_{0,0}^0(\boldsymbol{\omega})\ge 2\pi(2/\pi)^{2(\alpha+1)}\,I_{0,0}^0(\boldsymbol{\omega})$$

for $\boldsymbol{\omega}\in S_h$, $\beta\in\{0,1\}$ where $I_{0,0}^0(\boldsymbol{\omega})=\omega^{-1}F^{[\beta]}(\omega\Delta t)$. The turning points of the functions $F^{[\beta]}$ occur at the zeros $z_{\beta,l}$ of the Bessel function J_β, and following [32, sect. 5], it can be shown that $F^{[\beta]}(t)\ge F^{[\beta]}(z_{\beta,2})$ for all $t\ge z_{\beta,1}$. After a little manipulation we have $I_{0,0}^0(\boldsymbol{\omega})\ge \Delta t F^{[\beta]}(z_{\beta,2})/\max(z_{\beta,1},\sqrt{2}\,\pi\,\Delta t/h)$ where $\Delta t\omega\le\sqrt{2}\,\pi\,\Delta t/h$ for all $\boldsymbol{\omega}\in S_h$. ∎

4.2. Stability results. To define stability we follow [6] and investigate the growth of perturbations in the solution of the homogeneous problem for which $a\equiv 0$. Because of linearity, it is enough to consider the propagation of nonzero initial data $U^1\ne 0$. The homogeneous stability problem is thus (3.3) with $a\equiv 0$ and U^1 a given, nonzero mesh function, i.e.,

$$\mathbb{Q}^0U_{p,q}^n = -\sum_{m=1}^{n-1}\mathbb{Q}^mU_{p,q}^{n-m} \tag{4.9}$$

for $n\ge 2$ with $U_{p,q}^1\not\equiv 0$.

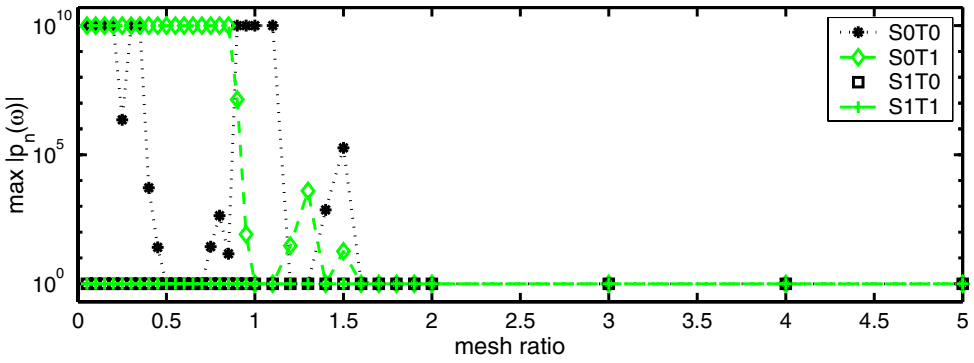


FIG. 4.1. *Stability plot for each of the four schemes* $S\alpha T\beta$. *The graph shows* $\min(\max_{n,j,k}\{|p_n(\boldsymbol{\omega}_{j,k})|\},\ 1e10)$ *plotted against the mesh ratio* $\rho = \Delta t/h$, *where the maximum is taken over timesteps* $n \le \min(1000, 1000/\rho)$ *and frequencies* $\boldsymbol{\omega}_{j,k} = 0.1\,\pi\,(j,k)/h$ *for* $0 \le j,\ k \le 10$.

DEFINITION 4.3. *The numerical scheme* (4.9) *is said to be stable on* $(0,T)$ *if there exists a constant* C *independent of* n *and* h *such that*

$$\|U^n\|_h \le C\,\|U^1\|_h$$

whenever $t^n < T$, *for all functions* U^1 *for which* $\|U^1\|_h < \infty$.

It is straightforward to show that stability corresponds to the existence of a constant C such that $|p_n(\boldsymbol{\omega})| \le C$ for all n and all $\boldsymbol{\omega} \in S_h$ (details are given in [6]). Unfortunately there appears to be no obvious way to check this condition by analysis, and we resort to testing it numerically for many individual frequencies $\boldsymbol{\omega} \in S_h$ to determine the stability of the four schemes. Results are shown in Figure 4.1 and indicate that the two schemes based on piecewise constant spatial basis functions (S0T0 and S0T1) are unstable for many values of mesh ratio, whereas the two schemes based on piecewise linear spatial basis functions (S1T0 and S1T1) appear stable over the range of mesh ratios tested. Stability over a wide range of mesh ratios is very important, since practical calculations over general surfaces may involve space mesh elements of vastly different sizes. Hence we do not consider schemes S0T0 and S0T1 further here.

It is shown in [11] that removing the singularity in the RPIE integrals (1.1) by using local polar coordinates (see also [5, 9]) can also lead to stable collocation schemes. The polar approximation based on the trapezoidal rule in R and arbitrarily accurate integration in θ for which the temporal and spatial basis functions are piecewise linear also appears stable over all values of mesh ratio considered [11]. The disadvantage of this scheme is that the transformed region of integration has a complicated shape that depends on $\boldsymbol{x}$ when Γ is finite and so the scheme is not straightforward to implement in practice. We note also that the collocation RPIE scheme due to Rynne and Smith [31] (which uses piecewise constant basis functions in space, piecewise linears in time, and the midpoint quadrature rule to evaluate the coefficients $C^m_{j,k}$) can be made stable at any value of mesh ratio by averaging in time [7, 10, 31] (which filters out high frequency instabilities). However, this is not entirely satisfactory because, for example, electromagnetic scattering problems involve more complicated RPIEs and hence are harder to stabilize [8]. We believe that a minimum requirement for a scalar RPIE scheme to be generally useful is that it should be stable over a wide range of mesh ratio when applied on an infinite flat plate without recourse to any filtering.

4.3. Further properties of the Fourier transformed schemes. This subsection lays the groundwork for the convergence analysis of S1T0 and S1T1, which appear stable for a wide range of mesh ratio values. We make precise the relationship between the stability coefficients q_m for the schemes and appropriate quadrature approximations of the Fourier transformed RPIE (2.13). The connection between q_m for piecewise linear in time RPIE schemes (like S1T1) and the trapezoidal rule approximation of (2.13) was first described in [10]. The q_m for the piecewise constant in time scheme S1T0 are similarly connected to the midpoint rule approximation of (2.13).

Letting $\widehat{u}_{\Delta t}^{m+1/2}(\boldsymbol{\omega})$ denote the approximation of $\widehat{u}(\boldsymbol{\omega}, t^{m+1/2})$ obtained by using the composite midpoint rule for (2.13) with spacing Δt, we have

$$2\pi\Delta t \sum_{m=0}^{n-1} J_0(\omega t^{n-m-1/2})\, \widehat{u}_{\Delta t}^{m+1/2}(\boldsymbol{\omega}) = \widehat{a}(\boldsymbol{\omega}, t^n).$$

Comparing this with the DFT equation (4.1) and matching the q_m and Bessel function terms gives $q_m(\boldsymbol{\omega}) \sim 2\pi\Delta t J_0(\omega t^{m+1/2})$ for $m \geq 0$. Similarly, comparing the coefficients for S1T1 with the trapezoidal rule approximation of (2.13) gives $q_0(\boldsymbol{\omega}) \sim \pi\Delta t J_0(0)$ and $q_m(\boldsymbol{\omega}) \sim 2\pi\Delta t J_0(\omega t^m)$ for $m \geq 1$ [10]. To see just how close this match is we define $\alpha_m(\boldsymbol{\omega})$ for each scheme by

$$\alpha_0(\boldsymbol{\omega}) \equiv J_0(\omega\, t^{(1-\beta)/2})/(\beta+1) - q_0(\boldsymbol{\omega})/(2\pi\Delta t), \tag{4.10a}$$

$$\alpha_m(\boldsymbol{\omega}) \equiv J_0(\omega\, t^{m+(1-\beta)/2}) - q_m(\boldsymbol{\omega})/(2\pi\Delta t) \quad \text{for } m \geq 1, \tag{4.10b}$$

where we recall that $\beta = 0$ for S1T0 and $\beta = 1$ for S1T1. The following result states the small $h\omega$ behavior of the α_m.

LEMMA 4.4. *There exists a constant C independent of h, $\boldsymbol{\omega}$, and m such that the coefficients α_m for S1T0 and S1T1 satisfy*

$$|\alpha_m(\boldsymbol{\omega})| \leq C(h\omega)^2 \tag{4.11}$$

for all $m \geq 0$ and $\boldsymbol{\omega} \in S_h$.

Proof. We prove the result for scheme S1T0 and note that the details for S1T1 are similar. Substituting the q_m equation (4.5) for S1T0 into the definition (4.10) of α_m gives

$$\begin{aligned}\alpha_m(\boldsymbol{\omega}) &= J_0(\omega\, t^{m+1/2}) - \frac{1}{\Delta t}\sum_{j,k} \Phi^{[1]}(h\omega_1 + 2\pi j)\, \Phi^{[1]}(h\omega_2 + 2\pi k)\, I_{j,k}^m(\boldsymbol{\omega}) \\ &= (T_1 - T_2 - T_3 - T_4)/\Delta t,\end{aligned}$$

where, using the definition (4.6) of $\Phi^{[1]}$,

$$\begin{aligned}T_1 &= (1 - \Phi^{[1]}(h\omega_1)\, \Phi^{[1]}(h\omega_2))\, I_{0,0}^m(\boldsymbol{\omega}) + 2\Delta t J_0(t^{m+1/2}\omega) - I_{0,0}^m(\boldsymbol{\omega}), \\ T_2 &= 2\, \Phi^{[1]}(h\omega_1)\, (1 - \cos(h\omega_2)) \sum_{k\neq 0} \frac{I_{0,k}^m(\boldsymbol{\omega})}{(h\omega_2 + 2\pi k)^2}, \\ T_3 &= 2\, \Phi^{[1]}(h\omega_2)\, (1 - \cos(h\omega_1)) \sum_{j\neq 0} \frac{I_{j,0}^m(\boldsymbol{\omega})}{(h\omega_1 + 2\pi j)^2}, \quad \text{and} \\ T_4 &= 4\, (1 - \cos(h\omega_1))\, (1 - \cos(h\omega_2)) \sum_{j,k\neq 0} \frac{I_{j,k}^m(\boldsymbol{\omega})}{(h\omega_1 + 2\pi j)^2 (h\omega_2 + 2\pi k)^2}.\end{aligned}$$

Since $|\psi_m^{[0]}(z)| \leq 1$ and $|J_0(z)| \leq 1$ for all $z \in \mathbb{R}$ it follows from definition (4.7) that $|I_{j,k}^m(\boldsymbol{\omega})| \leq \Delta t$ for all $(j,k) \in \mathbb{Z}^2$ and $m \geq 0$. It then follows from the inequalities $|1 - \cos z| \leq z^2/2$ and $|\Phi^{[1]}(z)| \leq 1$, and the boundedness of the sum $\sum_{k \neq 0} k^{-2}$ that $|T_2|$, $|T_3| \leq C\, h^2\omega^2\Delta t$ and $|T_4| \leq C\,(h\omega_1)^2(h\omega_2)^2\Delta t$, and hence is also bounded by $C\, h^2\omega^2\Delta t$ for $\boldsymbol{\omega} \in S_h$.

Using standard results for the midpoint quadrature rule [12] gives

$$I_{0,0}^m(\boldsymbol{\omega}) = \Delta t J_0(t^{m+1/2}\omega) + \omega^2\Delta t^3 J_0''(\omega R_m)/24$$

for some $R_m \in (t^m, t^{m+1})$, and hence

$$\left| I_{0,0}^m(\boldsymbol{\omega}) - \Delta t J_0(t^{m+1/2}\omega) \right| \leq C\Delta t (h\omega)^2$$

since $|J_0''(z)|$ is bounded for all $z \in \mathbb{R}$. It thus follows from the triangle inequality and the additional bound $|1 - \Phi^{[1]}(z)| \leq z^2/12$ that $|T_1| \leq C\, h^2\omega^2\Delta t$. ∎

This result means that the scaled coefficients $q_m(\boldsymbol{\omega})$ (which are DFTs of the difference operators $\mathbb{Q}^m$) are second order accurate approximations of the Bessel functions in the Fourier transformed RPIE (2.13). We use this to establish convergence of the schemes S1T0 and S1T1 in the next section.

The midpoint and trapezoidal quadrature rules are both known to give stable schemes for Volterra equations like (2.13), although the leading error term for the trapezoidal rule solution is oscillatory [17, 23, 24]. It is also known [17, 24] that all higher order Newton–Cotes quadrature rules give rise to unstable approximations of (2.13). Hence one would need to be careful in constructing approximations of (2.1) that use temporal basis functions of higher degree, in case they give rise to the same instabilities.

5. Convergence. In this section we demonstrate that the schemes S1T0 and S1T1 for the infinite flat plate problem (2.1) are convergent for values of the mesh ratio $\Delta t/h$ at which they are stable. We work with spatially Fourier transformed quantities and also use Laplace and Z transforms in time to obtain the results. The proof relies on the Fourier transformed RPIE (2.13) being a convolution equation in time and we use techniques due originally to Lubich [27] to obtain bounds for the Fourier transform of the approximation error. We then use arguments similar to those used to prove convergence of approximation schemes for a linear PDE by Thomée [33] (similar techniques are used for hyperbolic equations in [13]). This type of convergence analysis relies crucially on estimates given by Bramble and Hilbert [4] and Thomée [33]. The analysis of schemes for retarded potential integrals is much more complicated than those for PDEs, and much of this section is devoted to formulating the problem in such a way so as to use these estimates.

Throughout this section, C will denote a generic constant that can depend upon the mesh ratio, σ, T, and the norm exponents m and r but is independent of u, a, and h.

5.1. Hypotheses and definitions. We make the following assumptions on the problem and numerical solution.

HYPOTHESES. *Suppose that*

(H1) *the incident field* $a \in H_*^{5+\beta}(0,T;H^{6+\beta}(\mathbb{R}^2))$*;*

(H2) *numerical scheme S1Tβ for* (2.1) *is stable at the mesh ratio* $\rho = \Delta t/h \in (0,\infty)$*, and the mesh ratio remains fixed as* Δt *and* h *go to zero.*

As in section 3, the approximate solution corresponding to S1Tβ for $\beta \in \{0,1\}$ is denoted by $U(\boldsymbol{x},t)$. We explicitly need to make assumption (H2) because stability for these schemes has only been verified numerically and not proved rigorously. Note that it follows from (H1) and Lemma 2.3 that $u \in H_*^{4+\beta}(0,T;H^{5+\beta}(\mathbb{R}^2))$.

We now define convergence for an RPIE scheme, and in the subsequent lemma we show what quantities need to be bounded in order to prove that the schemes converge.

DEFINITION 5.1. *A scheme for the RPIE* (2.1) *is convergent on* $(0,T)$ *if the difference between the exact and approximate solutions* $\|u(\cdot,t)-U(\cdot,t)\|_h \to 0$ *as* $h \to 0$ *whenever* $t < T$.

LEMMA 5.2. *For RPIE* (2.1) *with* $a(\boldsymbol{x},t)$ *satisfying* (H1), *schemes S1T*β *for* $\beta \in \{0,1\}$ *satisfy*

$$\|u(\cdot,t^{n*})-U(\cdot,t^{n*})\|_h \le Ch^r \|a\|_{H_*^2(0,T;H^{r+2}(\mathbb{R}^2))} + \|\varepsilon_n\|_{\mathcal{F}_h}$$

for $1 < r \le 4+\beta$, *where* $t^{n*} = t^{n-(1-\beta)/2}$ *and* ε_n *satisfies the convolution equation*

$$\sum_{m=1}^{n} q_{n-m}(\boldsymbol{\omega})\, \varepsilon_m(\boldsymbol{\omega}) = \mathcal{E}_n(\boldsymbol{\omega}) \tag{5.1}$$

with

$$\mathcal{E}_n(\boldsymbol{\omega}) = 2\pi \int_0^{t^n} J_0(\omega(t^n - R))\, \widehat{u}(\boldsymbol{\omega},R)\, dR - \sum_{m=1}^{n} q_{n-m}(\boldsymbol{\omega})\, \widehat{u}(\boldsymbol{\omega}, t^{m-(1-\beta)/2}) \tag{5.2}$$

and the q_m *given by* (4.5). *Hence they are convergent if* $\|\varepsilon_n\|_{\mathcal{F}_h} \to 0$ *as* $h \to 0$.

Proof. It follows from the discrete Parseval identity (2.5) and the triangle inequality that

$$\|u(\cdot,t^{n*})-U(\cdot,t^{n*})\|_h \le \|\tilde{u}(\cdot,t^{n*})-\widehat{u}(\cdot,t^{n*})\|_{\mathcal{F}_h} + \|\tilde{U}^n - \widehat{u}(\cdot,t^{n*})\|_{\mathcal{F}_h}.$$

The first term on the right-hand side above can be bounded using Proposition 2.1. When $r > 1$ this gives

$$\begin{aligned} \|\widehat{u}(\cdot,t)-\tilde{u}(\cdot,t)\|_{\mathcal{F}_h} &\le Ch^r \|u(\cdot,t)\|_{H^r(\mathbb{R}^2)} \\ &\le Ch^r \|u\|_{H_*^1(0,T;H^r(\mathbb{R}^2))} \\ &\le Ch^r \|a\|_{H_*^2(0,T;H^{r+1}(\mathbb{R}^2))} \end{aligned} \tag{5.3}$$

from Lemma 2.3.

We now examine the second term. Comparing the Fourier transformed RPIE (2.13) at $t = t^n$ with the DFT of the numerical scheme (4.1) gives

$$\sum_{m=1}^{n} q_{n-m}(\boldsymbol{\omega})\, (\tilde{U}^m(\boldsymbol{\omega}) - \widehat{u}(\boldsymbol{\omega},t^{m*})) = \tilde{a}^n(\boldsymbol{\omega}) - \widehat{a}^n(\boldsymbol{\omega}) + \mathcal{E}_n(\boldsymbol{\omega}).$$

Setting

$$\beta_m(\boldsymbol{\omega}) = \tilde{U}^m(\boldsymbol{\omega}) - \widehat{u}(\boldsymbol{\omega},t^{m*}) - \varepsilon_m(\boldsymbol{\omega}),$$

it follows from the definition (5.1) of ε_m that

$$\sum_{m=1}^{n} q_{n-m}(\boldsymbol{\omega})\, \beta_m(\boldsymbol{\omega}) = \tilde{a}^n(\boldsymbol{\omega}) - \widehat{a}^n(\boldsymbol{\omega}). \tag{5.4}$$

The triangle inequality gives

$$\|\tilde{U}^n - \widehat{u}(\cdot, t^{n*})\|_{\mathcal{F}_h} \le \|\varepsilon_n\|_{\mathcal{F}_h} + \|\beta_n\|_{\mathcal{F}_h}$$

and so it remains only to show that $\|\beta_n\|_{\mathcal{F}_h} \to 0$ as $h \to 0$.

Inverting the convolution sum (5.4) using the formula (4.3) gives

$$\beta_n = q_0^{-1} \sum_{m=1}^{n} p_{n+1-m} \left(\tilde{a}^m - \widehat{a}^m\right),$$

where the p_m are defined by (4.4). The scheme is stable by hypothesis (H2), which means that the p_n are bounded, and hence it follows from the triangle inequality and the lower bound on q_0 given in Lemma 4.2 that

$$\|\beta_n\|_{\mathcal{F}_h} \le C\, h^{-1} \sum_{m=1}^{n} \|\tilde{a}^m - \widehat{a}^m\|_{\mathcal{F}_h}$$

for some constant C. Hypothesis (H1) and Proposition 2.1 together give

$$\|\tilde{a}(\cdot,t) - \widehat{a}(\cdot,t)\|_{\mathcal{F}_h} \le C\, h^{r+2}\, \|a(\cdot,t)\|_{H^{r+2}(\mathbb{R}^2)} \le C\, h^{r+2}\, \|a\|_{H^1_*(0,T;H^{r+2}(\mathbb{R}^2))}$$

when $t < T$, for any $r > -1$. Thus

$$\|\beta_n\|_{\mathcal{F}_h} \le C\, h^{-1} \sum_{m=1}^{n} \|\tilde{a}^m - \widehat{a}^m\|_{\mathcal{F}_h} \le C\, h^{r}\, \|a\|_{H^1_*(0,T;H^{r+2}(\mathbb{R}^2))}$$

since $n \le T/(\rho h)$ (where ρ is the mesh ratio). Combining this with inequality (5.3) completes the proof. ∎

The rest of this section is devoted to deriving two different bounds on ε_n; the first bound is valid for all $\boldsymbol{\omega}$ in S_h and the second when $h\omega$ is small. These bounds are then combined to show that $\|\varepsilon_n\|_{\mathcal{F}_h} = O(h^2)$ as $h \to 0$, and hence that we can use the previous lemma with $r = 2$ to prove second order convergence for the schemes S1Tβ.

5.2. Bound on $\boldsymbol{\varepsilon_n}$ for all $\boldsymbol{\omega \in S_h}$. Here we combine a bound on the size of the error term $\mathcal{E}_n(\boldsymbol{\omega})$ defined in (5.2) with the stability hypothesis (H2) in order to bound ε_n.

LEMMA 5.3. *Under hypotheses* (H1) *and* (H2), *there exists* ζ *with* $(1+\omega)^2\zeta(\omega) \in L^2(\mathbb{R}^2)$ *such that*

$$|\varepsilon_n(\boldsymbol{\omega})| \le \zeta(\boldsymbol{\omega}) \quad \textit{when } t^n < T. \tag{5.5}$$

Proof. Using (4.10) to replace the q_m terms in (5.2) gives

$$\frac{\mathcal{E}_n}{2\pi} = \int_0^{t^n} J_0(\omega(t^n - R))\, \widehat{u}(\boldsymbol{\omega}, R)\, dR - \Delta t \sum_{m=1}^{n} \left[J_0(\omega t^{n-m*}) - \alpha_{n-m}(\boldsymbol{\omega})\right] \widehat{u}(\boldsymbol{\omega}, t^{m*}).$$

This is the error in the midpoint (resp., trapezoidal) rule approximation of the integral when $\beta = 0$ (resp., 1), with additional terms involving the α's. It follows from standard results for these quadrature rules that if $t^n < T$ then

$$\left|\frac{\mathcal{E}_n}{2\pi}\right| \le Ch^2 \left|\frac{\partial^2}{\partial R^2} J_0(\omega(t^n - R))\, \widehat{u}(\boldsymbol{\omega}, R)\right|_{R=\mu} + \Delta t \sum_{m=1}^{n} |\alpha_{n-m}(\boldsymbol{\omega})\, \widehat{u}(\boldsymbol{\omega}, {}^{m*})|$$

for some $\mu \in (0, t^n)$. Hence, using the bound (4.11) on the size of the $|\alpha_m|$, and the fact that $J_0(z)$, $J_0'(z)$, and $J_0''(z)$ are all bounded it follows that

$$|\mathcal{E}_n| \le Ch^2(\omega^2|\widehat{u}(\boldsymbol{\omega},\mu)| + \omega|\widehat{u}^{(1)}(\boldsymbol{\omega},\mu)| + |\widehat{u}^{(2)}(\boldsymbol{\omega},\mu)|)$$

and the pointwise bound from Lemma 2.4 then gives

$$|\mathcal{E}_n| \le Ch^2(1+\omega)^3\|\widehat{a}(\boldsymbol{\omega},\cdot)\|_{H^4(0,T)}. \tag{5.6}$$

Now inverting the convolution sum (5.1) and using an identical argument to Lemma 5.2, we get

$$|\varepsilon_n| \le C\,h^{-1}\sum_{m=1}^{n}|\mathcal{E}_m| \le C\,(1+\omega)^3\|\widehat{a}(\boldsymbol{\omega},\cdot)\|_{H^4(0,T)} \equiv \zeta(\boldsymbol{\omega})$$

for $t_n \le T$. Hypothesis (H1) guarantees that $(1+\omega)^2\zeta(\boldsymbol{\omega}) \in L^2(\mathbb{R}^2)$ as required. ∎

5.3. Bound on $\boldsymbol{\varepsilon_n}$ for small $\boldsymbol{h\omega}$. This is the most technical part of the convergence proof. We need to get an $O(h^2)$ bound on $\|\varepsilon_n\|$ when $h\omega$ is sufficiently small. Taking the Z transform (2.10) of the convolution sum (5.1) gives

$$Zq(\boldsymbol{\omega},s)\,Z\varepsilon(\boldsymbol{\omega},s) = Z\mathcal{E}(\boldsymbol{\omega},s)$$

and hence

$$|Z\varepsilon(\boldsymbol{\omega},s)| = |Zq(\boldsymbol{\omega},s)|^{-1}|Z\mathcal{E}(\boldsymbol{\omega},s)|$$

(for $Zq \ne 0$) where $s = \sigma + i\eta$ and $\eta \in [-\pi/\Delta t, \pi/\Delta t]$. Ideally, we would obtain upper bounds on $1/|Zq|$ and $|Z\mathcal{E}|$, use them to bound $|Z\varepsilon|$, and use the inverse Z transform to bound $|\varepsilon_n|$. Unfortunately this is not straightforward, but we can make progress by a less direct route. We first use (4.10) and Lemma 4.4 to obtain the following information on the Z transform of the q_m.

LEMMA 5.4. *We can write $q_m = q_m^a + q_m^b$ for all $0 \le m\Delta t \le T$, where*

$$|q_m^b| \le C\Delta t(h\omega)^2$$

and the sequence q_m^a is defined through its Z transform

$$Zq^a(\boldsymbol{\omega},s) = 2\pi\begin{cases} e^{s\Delta t/2}\left(\dfrac{1}{\sqrt{s^2+\omega^2}} - \dfrac{1}{s} + \dfrac{\Delta t}{e^{s\Delta t/2}-e^{-s\Delta t/2}}\right), & \beta = 0, \\ \dfrac{1}{\sqrt{s^2+\omega^2}} - \dfrac{1}{s} + \dfrac{\Delta t}{2}\left(\dfrac{e^{s\Delta t}+1}{e^{s\Delta t}-1}\right), & \beta = 1. \end{cases}$$

Proof. The two cases are very similar so we just consider $\beta = 0$. From Lemma 4.4 we have $q_m = 2\pi\Delta t J_0(\omega t^{m+1/2}) - 2\pi\Delta t\alpha_m$ with $|\alpha_m| \le C(h\omega)^2$. We write the Bessel function term as $J_0(\omega t) = f(t) + 1$, where $f(t) \equiv J_0(\omega t) - 1$ and take its Z transform to get

$$\sum_{m=0}^{\infty} J_0(\omega t^{m+1/2})\,e^{-sm\Delta t} = \sum_{m=0}^{\infty} f(t^{m+1/2})\,e^{-sm\Delta t} + \frac{1}{1-e^{-s\Delta t}}\,. \tag{5.7}$$

(The reason for working with f rather than directly with $J_0(\omega t)$ is that $f(0) = 0$.) Splitting the sum $\sum_{m=0}^{\infty} f_{m/2}\, e^{-sm\Delta t/2}$ into odd and even terms and rearranging give

$$\sum_{m=0}^{\infty} f(\omega t^{m+1/2})\, e^{-sm\Delta t} = e^{s\Delta t/2} \left(\sum_{m=0}^{\infty} f(\omega t^{m/2})\, e^{-sm\Delta t/2} - \sum_{m=0}^{\infty} f(\omega t^{m})\, e^{-sm\Delta t} \right).$$

The Laplace Poisson sum formula (2.12) with spacing Δt is

$$\Delta t \sum_{m=0}^{\infty} f(\omega t^m) e^{-sm\Delta t} = \frac{1}{\sqrt{s^2+\omega^2}} - \frac{1}{s} + \sum_{l\neq 0} \theta_l(s,\omega,\Delta t),$$

where

$$\theta_l(s,\omega,\Delta t) = \frac{1}{\sqrt{s_l^2+\omega^2}} - \frac{1}{s_l} \quad \text{for} \quad s_l = s + i\,\frac{2\pi l}{\Delta t}.$$

Substituting this and the similar Poisson sum formula with spacing $\Delta t/2$ into the above identity for f gives

$$\Delta t \sum_{m=0}^{\infty} f(\omega t^{m+1/2})\, e^{-sm\Delta t} = e^{s\Delta t/2} \left(\frac{1}{\sqrt{s^2+\omega^2}} - \frac{1}{s} + \Theta \right),$$

where

$$\Theta = \sum_{l\neq 0} \{2\theta_l(s,\omega,\Delta t/2) - \theta_l(s,\omega,\Delta t)\}.$$

It then follows from (5.7) that

$$2\pi\Delta t \sum_{m=0}^{\infty} J_0(\omega t^{m+1/2}) e^{-sm\Delta t} = Zq^a(\boldsymbol{\omega}, s) + Z\kappa, \tag{5.8}$$

where $\{\kappa_m\}$ is the inverse Z transform of $2\pi e^{s\Delta t/2}\Theta$.

It can be shown that if $hw < 1/(\rho\sqrt{2})$ and h is sufficiently small, then $|\Theta| \leq C\Delta t\,(h\omega)^2$. Hence it follows from the inverse transform formula (2.11) that

$$|\kappa_n| \leq \Delta t\, e^{\sigma(n+1/2)\Delta t} \int_{-\pi/\Delta t}^{\pi/\Delta t} |\Theta| d\eta \leq C e^{T\sigma} \Delta t (h\omega)^2$$

for $n\Delta t \leq T$. The result then follows upon comparing (5.8) with (4.10) and using the bound on $|\alpha_m|$ given in Lemma 4.4. ∎

We next obtain upper bounds on $1/|Zq^a|$.

LEMMA 5.5. *If $\omega\Delta t \leq \pi/\sqrt{2}$ and Δt is small enough, then the Z transforms defined in the previous lemma satisfy*

$$\frac{1}{|Zq^a|} \leq \begin{cases} \frac{2}{\pi}\,|\sqrt{s^2+\omega^2}|\,, & \beta = 0, \\ (2\pi\sigma)^{-1}|s^2+\omega^2|\,, & \beta = 1, \end{cases}$$

where $s = \sigma + i\eta$ and $\eta \in [-\pi/\Delta t, \pi/\Delta t]$.

Proof. The two cases work quite differently, and a great deal of algebraic manipulation (the details are omitted) is required to obtain the results.

Case $\boldsymbol{\beta = 0}$. Set

$$P = \frac{1}{\sqrt{s^2+\omega^2}} \quad \text{and} \quad Q = \frac{\Delta t}{(e^{s\Delta t/2} - e^{-s\Delta t/2})} - \frac{1}{s}\,.$$

Then

$$|Zq^a| = 2\pi\, e^{\sigma\Delta t/2}\ \{|P+Q|\} \geq 2\pi\, e^{\sigma\Delta t/2}\ \{|P| - |Q|\}.$$

It can be shown that $|Q|$ is monotonic increasing in $\eta\Delta t$ for $\eta\Delta t \in [0,\pi]$, and hence

$$|Q| \leq |Q|_{\eta\Delta t=\pi} = \frac{\Delta t}{\pi}\sqrt{1-\pi+\pi^2/4} + O(\Delta t^2).$$

So if Δt is sufficiently small, then $|Q| \leq 3\Delta t/(5\pi)$. It can also be shown that $|P| \geq 4\Delta t/(5\pi)$ if Δt is sufficiently small and $\omega\Delta t \leq \pi/\sqrt{2}$. Hence under these conditions we have $|Q| \leq 3|P|/4$ and so

$$|Zq^a| \geq \pi\, e^{\sigma\Delta t/2}\, |P|/2 \geq \pi|P|/2,$$

and the result follows from the definition of P.

Case $\boldsymbol{\beta = 1}$. We use P as above and define

$$R = \frac{\Delta t}{2}\frac{(e^{s\Delta t}+1)}{(e^{s\Delta t}-1)} - \frac{1}{s}\,.$$

Then

$$|Zq^a| = 2\pi\,|P+R| \geq 2\pi\,\Re(P+R) = 2\pi\ \{\Re(P) + \Re(R)\}.$$

It can be shown that

$$\Re(P) \geq \frac{\sigma}{|s^2+\omega^2|} > 0 \quad \text{and} \quad \Re(R) \geq 0,$$

from which the result follows immediately. ∎

We split the error from (5.1) into two parts, $\varepsilon_n = \varepsilon_n^a + \varepsilon_n^b$, satisfying

$$\sum_{m=0}^{n} q^a_{n-m}(\boldsymbol{\omega})\varepsilon^a_m(\boldsymbol{\omega}) = \mathcal{E}_n(\boldsymbol{\omega}) \quad \text{and} \quad \sum_{m=0}^{n} q_{n-m}(\boldsymbol{\omega})\varepsilon^b_m(\boldsymbol{\omega}) = -\sum_{m=0}^{n} q^b_{n-m}(\boldsymbol{\omega})\varepsilon^a_m(\boldsymbol{\omega}),$$

where $q_m = q^a_m + q^b_m$ as defined in Lemma 5.4, and we have taken all sums to start from $m=0$ rather than $m=1$ for ease of manipulation (the $m=0$ terms are zero by causality). We first bound $|\varepsilon^b_m|$ in terms of $|\varepsilon^a_m|$, so that the problem reduces to finding a bound on $|\varepsilon^a_m|$. Inverting the second convolution sum gives

$$\varepsilon^b_n = \frac{-1}{q_0}\sum_{m=0}^{n} p_{n-m}\sum_{k=0}^{m} q^b_{m-k}\varepsilon^a_m,$$

where $|p_m| \leq C$ by the stability hypothesis (H2), and $q_0 \geq C\Delta t$ from Lemma 4.2. If $n\Delta t \leq T$, then it follows from Lemma 5.4 that

$$|\varepsilon^b_n| \leq \frac{C}{\Delta t}\sum_{m=0}^{n}\sum_{k=0}^{m}|q^b_{m-k}|\,|\varepsilon^a_m| \leq \frac{CT^2h^2\omega^2}{\Delta t^2}\max_{m\leq n}|\varepsilon^a_m| \leq C\omega^2\max_{m\leq n}|\varepsilon^a_m|\,. \tag{5.9}$$

It remains to bound $|\varepsilon_n^a(\boldsymbol{\omega})|$. To do this we embed the time-discrete convolution $\sum_{m=0}^n q_m^a \varepsilon_{n-m}^a = \mathcal{E}_n$ into a time-continuous problem

$$\sum_{m=0}^{\infty} q_m^a(\boldsymbol{\omega})\varepsilon^a(\boldsymbol{\omega}, t-t_m) = \mathcal{E}(\boldsymbol{\omega}, t), \tag{5.10}$$

where $\mathcal{E}(\boldsymbol{\omega}, t)$ and $\varepsilon^a(\boldsymbol{\omega}, t)$ interpolate $\mathcal{E}_n$ and ε_n^a at time levels $t = t_n$. The aim is to obtain a bound on $\|\varepsilon^a(\boldsymbol{\omega}, \cdot)\|_{H^1(\mathbb{R}^+)}$ and hence on the point values $|\varepsilon_n^a(\boldsymbol{\omega})| = |\varepsilon^a(\boldsymbol{\omega}, t_n)|$ via (2.19). We generalize the formula (5.2) for $\mathcal{E}_n$ to obtain the interpolant

$$\mathcal{E}(\boldsymbol{\omega}, t) = 2\pi\int_0^{\infty} J_0(\omega R)\, \widehat{u}(\boldsymbol{\omega}, t-R)dR - \sum_{m=0}^{\infty} q_m(\boldsymbol{\omega})\, \widehat{u}(\boldsymbol{\omega}, t - t^{m+(1-\beta)/2}), \tag{5.11}$$

and note that it follows from causality of u that $\mathcal{E}(\boldsymbol{\omega}, t_n) = \mathcal{E}_n(\boldsymbol{\omega})$.

We bound $\|\varepsilon^a(\boldsymbol{\omega}, \cdot)\|_{H^1(\mathbb{R}^+)}$ via the Laplace transform of the time-continuous problem (5.10):

$$\bar{\varepsilon^a}(\boldsymbol{\omega}, s)\, Zq^a(\boldsymbol{\omega}, s) = \bar{\mathcal{E}}(\boldsymbol{\omega}, s).$$

This implies

$$|\bar{\varepsilon^a}(\boldsymbol{\omega}, s)| \le |Zq^a(\boldsymbol{\omega}, s)|^{-1}\, |\bar{\mathcal{E}}(\boldsymbol{\omega}, s)|,$$

with the upper bound on $|Zq^a|^{-1}$ given in Lemma 5.5. Using this bound, multiplying by $1 + |s|$ and applying the equivalence inequality (2.9), then gives

$$\|e^{-\sigma t}\varepsilon^a(\boldsymbol{\omega}, t)\|_{H^1(\mathbb{R}^+)} \le \begin{cases} C(1+\omega)\|e^{-\sigma t}\mathcal{E}(\boldsymbol{\omega}, t)\|_{H^2(\mathbb{R}^+)}, & \beta = 0 \\ C(1+\omega)^2\|e^{-\sigma t}\mathcal{E}(\boldsymbol{\omega}, t)\|_{H^3(\mathbb{R}^+)}, & \beta = 1. \end{cases} \tag{5.12}$$

The pointwise result

$$|\varepsilon^a(\boldsymbol{\omega}, t)| \le \frac{e^{\sigma T}}{\sqrt{2\pi}}\, \|e^{-\sigma t}\varepsilon^a(\boldsymbol{\omega}, t)\|_{H^1(\mathbb{R}^+)} \tag{5.13}$$

for $t \in (0, T)$, then follows from (2.19), and the next lemma provides the crucial $O(h^2)$ term that leads to the second order convergence result.

LEMMA 5.6. *The error term $\mathcal{E}(\boldsymbol{\omega}, t)$ defined by* (5.11) *satisfies*

$$\|e^{-\sigma t}\mathcal{E}(\boldsymbol{\omega}, t)\|_{H^m(0,T)} \le Ch^2(1+\omega)^3\|\widehat{a}(\boldsymbol{\omega}, \cdot)\|_{H^{m+3}(0,T)}$$

for $1 \le m \le 2 + \beta$.

Proof. The Laplace transform of (5.11) is

$$\bar{\mathcal{E}}(\boldsymbol{\omega}, s) = E(\boldsymbol{\omega}, s)\, \bar{\widehat{u}}(\boldsymbol{\omega}, s), \tag{5.14}$$

where

$$E(\boldsymbol{\omega}, s) \stackrel{\text{def}}{=} \left(\frac{2\pi}{\sqrt{\omega^2 + s^2}} - Zq(\boldsymbol{\omega}, s)e^{s\Delta t(\beta-1)/2}\right).$$

Multiplying by $(1 + |s|)^m$, square integrating over $\mathbb{R}$, and using (2.9) give

$$\|e^{-\sigma t}\, \mathcal{E}(\boldsymbol{\omega}, t)\|^2_{H^m(\mathbb{R}^+)} \le C\int_{-\infty}^{\infty} (1+|s|)^{2m}|E(\boldsymbol{\omega}, s)|^2|\bar{\widehat{u}}(\boldsymbol{\omega}, s)|^2 d\eta,$$

where $s = \sigma + i\eta$. We obtain two different bounds for $|E(\boldsymbol{\omega}, s)|$, valid for "high" and "low" values of $|\eta|$.

When $|\eta\Delta t| > \pi$, the triangle inequality implies that

$$|E(\boldsymbol{\omega}, s)| \leq \frac{2\pi}{|\sqrt{\omega^2 + s^2}|} + e^{(\beta-1)\sigma\Delta t/2}\,|Zq(\boldsymbol{\omega}, s)|$$

and we consider each term separately. If $|\eta\Delta t| > \pi$, then

$$\frac{1}{|\omega^2 + s^2|^2} \leq \frac{\Delta t^2}{2\sigma^2\pi^2} \leq C$$

when Δt is small. The second term can also be bounded by a constant: by definition

$$|Zq(\boldsymbol{\omega}, s)| \leq \sum_{n=0}^{\infty} |q_n\, e^{-sn\Delta t}| \leq C\Delta t \sum_{n=0}^{\infty} e^{-\sigma n\Delta t}$$

since (4.10) and (4.11) imply that each $|q_n| < C\Delta t$. Hence

$$|Zq(\boldsymbol{\omega}, s)| \leq \frac{C\Delta t}{1 - e^{-\sigma\Delta t}} \leq C$$

if Δt is sufficiently small.

Thus we have shown that if $|\eta\Delta t| > \pi$ then $|E(\boldsymbol{\omega}, s)| \leq C$. In this region $|s| > |\eta| > \pi/\Delta t$, and so $|s|\Delta t/\pi > 1$, which means that

$$|E(\boldsymbol{\omega}, s)| \leq C < C\,(|s|\Delta t/\pi)^2 = C\,|s|^2\,h^2$$

since $\Delta t/h$ is fixed.

When $|\eta\Delta t| \leq \pi$ we use Lemma 5.4 and consider the cases $\beta = 0$ and $\beta = 1$ separately. Define

$$E_0 = \frac{1}{s} - \frac{\Delta t}{e^{s\Delta t/2} - e^{-s\Delta t/2}}\,, \quad E_1 = \frac{1}{s} - \frac{\Delta t}{2}\left(\frac{e^{s\Delta t} + 1}{e^{s\Delta t} - 1}\right)$$

so that $E = E_\beta - Zq^b\, e^{(\beta-1)s\Delta t/2}$. Lemma 5.4 implies that $|q_n^b| \leq C\Delta t(h\omega)^2$ in either case, and so it follows from an identical argument to that used above to bound $|Zq|$ that $|Zq^b\, e^{(\beta-1)s\Delta t/2}| \leq e^{(\beta-1)\sigma\Delta t/2}\, C\,(h\omega)^2 \leq C\,(h\omega)^2$ if Δt is sufficiently small. It can be shown (again by considerable algebraic manipulation) that $|E_\beta| \leq \Delta t^2\,|s|$ for $\beta = 0,\ 1$ when $|\eta\Delta t| \leq \pi$. Hence if $|\eta\Delta t| \leq \pi$ and Δt is sufficiently small, we get $|E| \leq C\,\left(\Delta t^2\,|s| + h^2\omega^2\right)$.

We thus have the bound

$$|E(\boldsymbol{\omega}, s)| \leq Ch^2(1+\omega)^2(1+|s|)^2 \quad \forall \eta \in \mathbb{R}\,.$$

Inserting this into the integral in (5.14) gives

$$\|e^{-\sigma t}\mathcal{E}(\boldsymbol{\omega}, t)\|_{H^m(\mathbb{R}^+)} \leq Ch^2(1+\omega)^2\|e^{-\sigma t}\widehat{u}(\boldsymbol{\omega}, t)\|_{H^{m+2}(\mathbb{R}^+)}$$

and the result follows from Lemma 2.4. ∎

We now use this result to bound $|\varepsilon_n|$: (5.9) implies that

$$|\varepsilon_n(\boldsymbol{\omega})| \leq C(1+\omega)^2 \max_{n \leq T/\Delta t} |\varepsilon_n^a(\boldsymbol{\omega})|$$

for $n\Delta t \leq T$, and using bounds (5.12) and (5.13) gives

$$|\varepsilon_n(\boldsymbol{\omega})| \leq Ch^2(1+\omega)^{6+\beta}\|\widehat{a}(\boldsymbol{\omega}, \cdot)\|_{H^{5+\beta}(0,T)} \equiv h^2\zeta_\beta(\boldsymbol{\omega}) \tag{5.15}$$

for $\beta = 0, 1$. Hypothesis (H1) guarantees that $\zeta_\beta(\boldsymbol{\omega}) \in L_2(\mathbb{R}^2)$, which completes the small $h\omega$ bound calculation.

5.4. A bound for $\|\varepsilon_n\|_{\mathcal{F}_h}$. We split the range of integration of the Fourier norm $\|\varepsilon_n\|_{\mathcal{F}_h}$ into a "low" frequency section $\boldsymbol{\omega} \in L_h \equiv \{\boldsymbol{\omega} : h\omega < \gamma\}$, where inequality (5.15) is used (where the constant γ is chosen to be less than $1/(\rho\sqrt{2})$ so that all the small $h\omega$ bounds hold), and a "high" frequency section $\boldsymbol{\omega} \in S_h \backslash L_h$, where inequality (5.5) is used. The result is

$$\|\varepsilon_n\|_{\mathcal{F}_h} \le Ch^2 \left(\int_{L_h} |\zeta_\beta(\boldsymbol{\omega})|^2 \, d\boldsymbol{\omega} \right)^{1/2} + C \left(\int_{S_h \backslash L_h} |\zeta(\boldsymbol{\omega})|^2 \, d\boldsymbol{\omega} \right)^{1/2},$$

where ζ was introduced in section 5.2. The integral over low frequencies satisfies

$$\int_{L_h} |\zeta_\beta(\boldsymbol{\omega})|^2 \, d\boldsymbol{\omega} \le \int_{\mathbb{R}^2} |\zeta_\beta(\boldsymbol{\omega})|^2 \, d\boldsymbol{\omega} \le \|a\|^2_{H_*^{5+\beta}(0,T;H^{6+\beta}(\mathbb{R}^2))}.$$

Following the arguments used by Thomeé [33], the high frequency integral satisfies

$$\int_{S_h \backslash L_h} |\zeta(\boldsymbol{\omega})|^2 \, d\boldsymbol{\omega} \le \int_{S_h \backslash L_h} \left| \left(\frac{\omega h}{\gamma} \right)^2 \zeta(\boldsymbol{\omega}) \right|^2 d\boldsymbol{\omega} \le Ch^4,$$

since $\omega^2 \zeta(\boldsymbol{\omega}) \in L^2(\mathbb{R}^2)$ by Lemma 5.3.

Combining the low and high frequency bounds above and using Lemma 5.2 with $r = 2$ yield the final result.

THEOREM 5.7. *Under hypotheses* (H1) *and* (H2) *for* $\beta = 0, 1$, *the global error for schemes* $S1T\beta$ *satisfies the bound*

$$\|u(\cdot, t^{n-(1-\beta)/2}) - U(\cdot, t^{n-(1-\beta)/2})\|_h \le Ch^2$$

as $h \to 0$ *whenever* $t^n \le T$, *where* C *is a constant.*

6. Conclusions. We have presented two new schemes for the RPIE (2.1) that appear stable over a wide range of mesh ratio values, and are hence likely to be useful and reliable in practice. We have also given what we believe is the first rigorous convergence proof with reasonable, checkable hypotheses that RPIE collocation schemes converge at the optimal $O(h^2)$ rate one would expect from the underlying approximation methods. This is a great improvement on our earlier work [11], where we obtained proof of covergence at the rate $O(1/|\ln h|)$ for all but extremely smooth incident fields, whose spatial Fourier transforms decay faster than $e^{-\gamma_0 \omega}$ for a constant γ_0.

This improved result is mostly due to a change in approach to the error analysis for low spatial frequencies (section 5.3) from a Volterra integral equation analysis in the style of [17, 23, 24], to an approach using Z and Laplace transforms in the style of Lubich [27]. We believe that our new smoothness requirements may be relaxed further by more refined or alternative methods of proof, and we conjecture that this convergence rate will be achieved for a wider class of excitations.

Acknowledgments. We are indebted to the anonymous referee of the first version of this manuscript who very kindly pointed out how we could use Z transforms to improve our convergence rate from $O(1/|\ln h|)$ to $O(h^2)$. We are also grateful to B. P. Rynne for many helpful discussions.

This work was completed when we visited the Isaac Newton Institute for Mathematical Sciences in Cambridge, UK, as participants of the Computational Challenges in PDEs programme.

REFERENCES

[1] K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Springer-Verlag, New York, 2001.

[2] A. Bachelot and A. Pujols, *Time dependent integral equations for the Maxwell system*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 639–644.

[3] A. Bamberger and T. Ha-Duong, *Formulation variationnelle espace–temps pour le calcul par potentiel retardé de la diffraction d'une onde acoustique* (i), Math. Methods Appl. Sci., 8 (1986), pp. 405–435.

[4] J. H. Bramble and S. R. Hilbert, *Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation*, SIAM J. Numer. Anal., 7 (1970), pp. 112–124.

[5] O. Bruno and L. Kunyansky, *A fast, high-order algorithm for the solution of surface scattering problems: basic implementation, tests and applications*, J. Comput. Phys., 169 (2001), pp. 80–110.

[6] P. J. Davies, *Numerical stability and convergence of approximations of retarded potential integral equations*, SIAM J. Numer. Anal., 31 (1994), pp. 856–875.

[7] P. J. Davies, *Stability of time-marching numerical schemes for the electric field integral equation*, J. Electromag. Waves Appl., 8 (1994), pp. 85–114.

[8] P. J. Davies, *A stability analysis of a time marching scheme for the general surface electric field integral equation*, Appl. Numer. Math., 27 (1998), pp. 33–57.

[9] P. J. Davies and D. B. Duncan, *Accuracy and Convergence of Time Marching Schemes for RPIEs*, Technical report, University of Dundee, Dundee, Scotland, 1995.

[10] P. J. Davies and D. B. Duncan, *Averaging techniques for time-marching schemes for retarded potential integral equations*, Appl. Numer. Math., 23 (1997), pp. 291–310.

[11] P. J. Davies and D. B. Duncan, *Numerical stability of collocation schemes for time domain boundary integral equations*, in Computational Electromagnetics: Proceedings of the GAMM Workshop, Kiel, 2001, C. Carstensen, S. A. Funken, W. Hackbusch, R. H. W. Hoppe, and P. Monk, eds., Springer-Verlag, New York, 2003, pp. 51–67.

[12] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed., Academic Press, Orlando, 1984.

[13] D. B. Duncan and D. F. Griffiths, *The study of a Petrov-Galerkin method for first-order hyperbolic equations*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 147–166.

[14] D. B. Duncan and M. A. M. Lynch, *Jacobi iteration in implicit difference schemes for the wave equation*, SIAM J. Numer. Anal., 28 (1991), pp. 1661–1679.

[15] A. A. Ergin, B. Shanker, and E. Michielssen, *The plane–wave time–domain algorithm for the fast analysis of transient wave phenomena*, IEEE Ant. Prop. Magazine, 41 (1999), pp. 39–52.

[16] A. A. Ergin, B. Shanker, and E. Michielssen, *Fast analysis of transient acoustic wave scattering from rigid bodies using the multilevel plane wave time domain algorithm*, J. Acoust. Soc. Am., 107 (2000), pp. 1168–1178.

[17] C. J. Gladwin and R. Jeltsch, *Stability of quadrature rule methods for first kind Volterra integral equations*, BIT, 14 (1974), pp. 144–151.

[18] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 5th ed., Academic Press, Boston, 1994.

[19] T. Ha-Duong, *On the transient acoustic scattering by a flat object*, Japan J. Appl. Math., 7 (1990), pp. 489–513.

[20] T. Ha-Duong, *On retarded potential boundary integral equations and their discretisation*, in Topics in Computational Wave Propagation: Direct and Inverse Problems, M. Ainsworth, P. J. Davies, D. B. Duncan, P. A. Martin, and B. P. Rynne, eds., Springer-Verlag, New York, 2003, pp. 301–336.

[21] D. S. Jones, *The Theory of Electromagnetism*, MacMillan Co., New York, 1964.

[22] D. S. Jones, *Methods in Electromagnetic Wave Propagation*, 2nd ed., Clarendon Press, Oxford University Press, New York, 1994.

[23] J. G. Jones, *On the numerical solution of convolution integral equations and systems of such equations*, Math. Comp., 15 (1961), pp. 131–142.

[24] P. Linz, *Analytical and Numerical Methods for Volterra Equations*, SIAM, Philadelphia, 1985.

[25] J. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications,* Vol. I–II, Springer-Verlag, 1972.

[26] M. Lu, J. Wang, A. A. Ergin, and E. Michielssen, *Fast evaluation of two–dimensional transient wave fields*, J. Comput. Phys., 158 (2000), pp. 161–185.

[27] C. Lubich, *On the multistep time discretization of linear initial–boundary value problems and their boundary integral equations*, Numer., Math., 67 (1994), pp. 365–389.
[28] S. M. Rao, D. R. Wilton, and A. W. Glisson, *Electromagnetic scattering by surfaces of arbitrary shape*, IEEE Trans. Antennas and Propagation, 30 (1982), pp. 409–418.
[29] B. P. Rynne, *Time domain scattering from arbitrary surfaces using the electric field equation*, J. Electromagn. Waves Appl., 5 (1991), pp. 93–112.
[30] B. P. Rynne, *The well–posedness of the electric field integral equation for transient scattering from a perfectly conducting body*, Math. Methods Appl. Sci., 22 (1999), pp. 619–631.
[31] B. P. Rynne and P. D. Smith, *Stability of time marching algorithms for the electric field equation*, J. Electromagn. Waves Appl., 4 (1990), pp. 1181–1205.
[32] J. Steinig, *The real zeros of Struve's function*, SIAM J. Math. Anal., 1 (1970), pp. 365–375.
[33] V. Thomée, *Convergence estimates for semi-discrete Galerkin methods for initial-value problems*, in Numerische, inbesondere approximationstheoretische Behandlung von Funktionalgleichungen, Lecture Notes in Math. 333, A. Dold and B. Eckmann, eds., Springer-Verlag, Berlin, 1973, pp. 243–262.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1189–1207

AN ABSOLUTELY STABLE PRESSURE-POISSON STABILIZED FINITE ELEMENT METHOD FOR THE STOKES EQUATIONS*

PAVEL BOCHEV† AND MAX GUNZBURGER‡

Abstract. The pressure-Poisson stabilized Galerkin method for the Stokes equation requires the choice of a positive parameter. Existing theoretical predictions for the range of parameter values that yield stable discretizations seem to be very pessimistic when compared to the computational evidence. Motivated by this wide gap, we first examine a continuous prototype for this class of schemes. We show that the prototype is absolutely stable; i.e., it is stable for all parameter values, and is optimally accurate. We then define a new, practical variant of the well-known pressure-Poisson stabilized scheme. We prove that the new method is absolutely stable just like its continuous prototype and that it achieves optimal convergence rates with respect to the same mesh-independent norms. The new method differs from the standard pressure-Poisson stabilized method in several important aspects. First, its definition does not degrade to a penalty formulation for the lowest order nodal spaces. Second, the method is absolutely stable with respect to the natural norm for the problem, while the standard pressure-Poisson stabilized method is stable with respect to a mesh-dependent norm.

Key words. stabilized finite element methods, mixed methods, Stokes problem

AMS subject classifications. 76D05, 76D07, 65F10, 65F30

DOI. 10.1137/S0036142903416547

1. Introduction.

The stable and accurate finite element solution of the Stokes problem requires pairs of velocity and pressure spaces that satisfy the inf-sup (or LBB) compatibility condition; see, e.g., [7, 15, 16]. In the past two decades, the formulation of finite element methods that either circumvent or ameliorate this restrictive condition has attracted significant attention. Examples include augmented Lagrangian methods [12], least-squares finite element methods [4], and a group of methods collectively known as *consistently stabilized* Galerkin methods; see [1, 3, 8, 11, 13, 14, 17, 18, 19]. In what follows, we will refer to the members of the latter group as the *standard stabilized methods*.

In this paper, we develop and analyze a new stabilized formulation that can be related to one of the standard methods originally proposed in [18] and widely known as the *pressure-Poisson stabilized Galerkin method*. To demonstrate the connection between the new and the standard methods, we introduce the notion of continuous stabilized prototypes. Continuous prototypes are idealized finite element methods that are not necessarily practical. Their role is to provide a template that reveals the proper functional settings and guides the development of practical schemes. In

*Received by the editors October 21, 2002; accepted for publication (in revised form) September 10, 2003; published electronically September 18, 2004.

http://www.siam.org/journals/sinum/42-3/41654.html

†Computational Mathematics and Algorithms Department, Sandia National Laboratories, Albuquerque, NM 87185-1110 (pbboche@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC-94AL85000.

‡School of Computational Science and Information Technology, Florida State University, Tallahassee, FL 32306-4120 (gunzburg@csit.fsu.edu). The research of this author was supported in part by CSRI, Sandia National Laboratories under contract 18407.

addition, prototypes serve as a gauge to measure the deviation of practical methods from the idealized mathematical setting. All practical methods associated with a particular prototype form a class of methods. Here, we will derive the prototypes that engender the three most commonly used stabilized methods for the Stokes problem. For reasons that will be explained later, we call the three classes the GLS, SGLS, and RGLS method classes.

Consistently stabilized methods contain a positive parameter that must be set to define the method. It is well known that standard stabilized methods can be divided into those that are conditionally stable and those that are absolutely stable, i.e., those that are stable only for a set of restricted values of the parameter and those that are stable for all values of the parameter, respectively. According to previous theoretical analyses, the standard Galerkin least-squares [17] and pressure-Poisson [18] methods fall into the first category while the method of [11] is an example of an absolutely stable method. Stability classifications of stabilized methods are based on sufficient (weak or strong) coercivity conditions for the corresponding forms. Thus, in principle, they represent the worst case scenario and, in practice, there may be a gap between the theoretically predicted stability range of a method and the stability range observed in computational implementations. For the GLS method this gap is very narrow if it exists at all; see [13] or [2]. In other words, for this method, the stability range predicted by existing theory agrees with great accuracy with its practical stability range.

The main focus of this paper will be on the SGLS class which contains the standard pressure-Poisson stabilized method. Our interest in this class is not incidental. In [2], we reported an unusually large discrepancy between the well-known theoretical stability analysis of [8] and the actual, computationally observed stability range of the standard pressure-Poisson Galerkin method. In fact, what was observed computationally indicates that this method is actually absolutely stable. In this paper, we show that there are indeed grounds for such a stability pattern. Most notably, we prove that the continuous SGLS prototype is absolutely stable. Then we define a new discrete member of this class which also turns out to be absolutely stable.

Our new method differs from the standard pressure-Poisson Galerkin formulation in several important aspects. First, its definition does not degrade to a penalty formulation for the lowest-order nodal spaces. Second, we show that our method is absolutely stable with respect to the natural norm on $\mathbf{H}^1(\Omega) \times L_0^2(\Omega)$, while stability of the standard method is with respect to a mesh-dependent norm. Last, while the new method is not fully consistent, it is weakly inconsistent in the sense that finite element approximations converge to all smooth solutions at the best possible rate.

Our analysis suggests that the new, implementable SGLS method is a potentially strong contender in the field of stabilized formulations for the Stokes problem. The absolute stability makes it an attractive alternative to GLS methods that, both theoretically and practically, are known to be only conditionally stable. Compared with the absolutely stable RGLS methods, the new formulation avoids the appearance of local biharmonic terms that in principle should lead to better conditioned matrices. This conjecture is supported by our studies in [2] which suggest that Krylov subspace solvers generally tend to perform better for members of the SGLS family of stabilized methods. Nevertheless, further numerical studies will be needed to reach a definitive conclusion about the practical performance of our new method. These will be reported in a forthcoming paper.

We have organized the paper as follows. In section 2, we summarize notations

and quote technical results that are used throughout the paper. Section 3 develops the notion of continuous stabilized prototypes starting from a penalized Lagrangian formulation of the Stokes problem. Sections 4 and 5 are the core of this paper. Their focus is on the SGLS class of stabilized methods. In section 4, we consider the continuous prototype of this class and show that it is absolutely stable. Then, in section 5, we proceed to define a new discrete member of the SGLS class and establish its absolute stability and optimal convergence. In section 6, we conclude the paper with several remarks concerning implementation of the new method.

2. Quotation of results. Let Ω denote a bounded region in $\mathbb{R}^n$, $n = 2, 3$, with a Lipschitz continuous boundary $\Gamma = \partial\Omega$. For $p > 0$, $H^p(\Omega)$ denotes a Sobolev space of order p with norm and inner product denoted by $\|\cdot\|_p$ and $(\cdot,\cdot)_p$, respectively. When $p = 0$ we use the standard notation $L^2(\Omega)$. The symbol $|\cdot|_k$, $0 \le k \le p$, denotes the kth seminorm on $H^p(\Omega)$. We recall the subspace $L^2_0(\Omega)$ of all square integrable functions with vanishing mean and the subspace $H^1_0(\Omega)$ of all $H^1(\Omega)$ functions with vanishing trace. The Poincaré's inequality

$$C_P\|\phi\|_0 \le \|\nabla\phi\|_0 \quad \forall\, \phi \in H^p(\Omega) \cap H^1_0(\Omega) \tag{2.1}$$

implies that the seminorm $|\phi|_1 = \|\nabla\phi\|_0$ is an equivalent norm on $H^1_0(\Omega)$. Vector analogues of the Sobolev spaces along with vector-valued functions are denoted by upper and lower case bold face font, respectively, e.g., $\mathbf{H}^1(\Omega)$, $\mathbf{L}^2(\Omega)$, and $\mathbf{u}$. For vectors in Euclidean spaces, we use vector notation, e.g., $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$. Matrices are denoted by block letters, e.g., $\mathbb{A}$ and $\mathbb{B}$.

$\mathbf{V}^h$ and S^h will denote a pair of finite element subspaces of $\mathbf{H}^1_0(\Omega)$ and $L^2_0(\Omega)$, respectively. We assume that these spaces are defined with respect to the *same* regular triangulation $\mathcal{T}_h$ of the domain Ω into finite elements $\mathcal{K}$, where h denotes some measure of the grid size. For example, $\mathcal{K}$ can be hexahedrons or tetrahedrons in three dimensions or triangles or quadrilaterals in two dimensions. We will use C to denote a generic constant that is independent of h but whose value may change from place to place. Let $r > 0$ and $s > 0$ be two integers. It is further assumed that for every $\mathbf{u} \in \mathbf{H}^{r+1}(\Omega)$ and $p \in H^{s+1}(\Omega)$, there exist functions $\mathbf{u}^h_I \in \mathbf{V}^h$ and $p^h_I \in S^h$ such that

$$\|\mathbf{u} - \mathbf{u}^h_I\|_0 + h\|\mathbf{u} - \mathbf{u}^h_I\|_1 \le Ch^{r+1}\|\mathbf{u}\|_{r+1} \tag{2.2}$$

and

$$\|p - p^h_I\|_0 + h\|p - p^h_I\|_1 \le Ch^{s+1}\|p\|_{s+1}, \tag{2.3}$$

respectively. We recall the inverse inequalities

$$\|\mathbf{u}^h\|_1 \le C_I h^{-1}\|\mathbf{u}^h\|_0 \qquad \text{and} \qquad \|p^h\|_1 \le C_I h^{-1}\|p^h\|_0 \tag{2.4}$$

that hold for finite element spaces on regular triangulations; see [10] or [15].

2.1. Negative norm and inner product. Let $\mathbf{H}^{-1}(\Omega)$ denote the dual of $\mathbf{H}^1_0(\Omega)$. Using the equivalence of $|\cdot|_1$ and $\|\cdot\|_1$ on $\mathbf{H}^1_0(\Omega)$, we equip $\mathbf{H}^{-1}(\Omega)$ with the norm

$$\|\mathbf{f}\|_{-1} = \sup_{\boldsymbol{\phi}\in\mathbf{H}^1_0(\Omega)} \frac{(\mathbf{f}, \boldsymbol{\phi})_0}{|\boldsymbol{\phi}|_1} \quad \forall\, \mathbf{f} \in \mathbf{H}^{-1}(\Omega). \tag{2.5}$$

The following representation results hold (cf. [5, 6]).

LEMMA 2.1. *For all* $\mathbf{f} \in \mathbf{H}^{-1}(\Omega)$*, we have*

$$\|\mathbf{f}\|_{-1}^2 = (\mathbf{S}\mathbf{f}, \mathbf{f})_0,$$

where $\mathbf{S} : \mathbf{H}^{-1}(\Omega) \mapsto \mathbf{H}_0^1(\Omega)$ *is the solution operator for the vector Poisson equation*

$$-\triangle \mathbf{u} = \mathbf{f} \quad \textit{in } \Omega \qquad \textit{and} \qquad \mathbf{u} = 0 \quad \textit{on } \Gamma,$$

i.e., $\mathbf{u} = \mathbf{S}\mathbf{f}$ *if and only if*

$$(\nabla \mathbf{u}, \nabla \mathbf{v})_0 = (\mathbf{f}, \mathbf{v})_0 \quad \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega).$$

If $(\cdot,\cdot)_{-1}$ *is the inner product associated with* $\|\cdot\|_{-1}$*, then*

$$(\mathbf{f}, \mathbf{g})_{-1} = (\mathbf{S}\mathbf{f}, \mathbf{g})_0 = (\mathbf{f}, \mathbf{S}\mathbf{g})_0 \quad \forall\, \mathbf{f}, \mathbf{g} \in \mathbf{H}^{-1}(\Omega). \tag{2.6}$$

Using (2.6), it is not difficult to show that

$$(-\triangle \mathbf{u}, \mathbf{v})_{-1} = (\mathbf{u}, \mathbf{v})_0 \quad \forall\, \mathbf{u} \in \mathbf{H}_0^1(\Omega),\ \mathbf{v} \in \mathbf{H}^{-1}(\Omega). \tag{2.7}$$

We also recall the well-known result (cf. [15, p. 20]) that for any connected Ω there exists a $C_N > 0$ such that

$$C_N \|p\|_0 \le \|\nabla p\|_{-1} \quad \forall\, p \in L_0^2(\Omega). \tag{2.8}$$

3. Stabilization of mixed methods for the Stokes problem. We consider the Stokes equations

$$\begin{aligned} -\triangle \mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} \quad \text{on } \Gamma. \end{aligned} \tag{3.1}$$

A weak formulation of the Stokes problem is to seek $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ such that

$$A(\mathbf{u}, \mathbf{v}) + B(\mathbf{v}, p) = F(\mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega), \tag{3.2}$$

$$B(\mathbf{u}, q) = 0 \quad \forall\, q \in L_0^2(\Omega), \tag{3.3}$$

where $A(\cdot,\cdot)$, $B(\cdot,\cdot)$, and $F(\cdot)$ are defined by

$$A(\mathbf{u}, \mathbf{v}) = \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v}\, d\Omega, \quad B(\mathbf{v}, p) = -\int_\Omega p \nabla \cdot \mathbf{v}\, d\Omega, \quad \text{and} \quad F(\mathbf{v}) = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\Omega,$$

respectively. We recall that (3.2)–(3.3) is the optimality system for the saddle-point $(\mathbf{u}, p)$ of the Lagrangian functional

$$L(\mathbf{v}, q) = \frac{1}{2} A(\mathbf{v}, \mathbf{v}) - F(\mathbf{v}) + B(\mathbf{v}, q). \tag{3.4}$$

Therefore, the pressure p is the Lagrange multiplier that is introduced into (3.4) to enforce the (weak) incompressibility constraint (3.3). The restriction of (3.2)–(3.3) to a pair of finite element subspaces $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$ and $S^h \subset L_0^2(\Omega)$ yields the *Galerkin mixed method*: seek $(\mathbf{u}^h, p^h) \in \mathbf{V}^h \times S^h$ such that

$$A(\mathbf{u}^h, \mathbf{v}^h) + B(\mathbf{v}^h, p^h) = F(\mathbf{v}^h) \quad \forall\, \mathbf{v}^h \in \mathbf{V}^h, \tag{3.5}$$

$$B(\mathbf{u}^h, q^h) = 0 \quad \forall\, q^h \in S^h. \tag{3.6}$$

For continuous pressure approximations and for velocity fields that vanish on the boundary, $B(\cdot,\cdot)$ can be replaced by the equivalent bilinear form

$$B^*(\mathbf{v}, p) = \int_\Omega \mathbf{v} \cdot \nabla p \, d\Omega.$$

It is easy to see that (3.5)–(3.6) is equivalent to the symmetric, indefinite linear algebraic system

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^T \\ \mathbb{B} & 0 \end{pmatrix} \begin{pmatrix} \vec{\mathbf{u}} \\ \vec{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{0}} \end{pmatrix}, \tag{3.7}$$

where the elements of $\vec{\mathbf{u}}$ and $\vec{\mathbf{p}}$ are the coefficients in the representation in terms of bases of the finite element pair $(\mathbf{u}^h, p^h)$; the matrices $\mathbb{A}$ and $\mathbb{B}$ are deduced in the usual manner, using the bases for $\mathbf{V}^h$ and S^h, from the bilinear forms $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ (or $B^*(\cdot,\cdot)$), respectively.

The problems (3.5)–(3.6) and (3.7) are equivalent representations of the optimality system for the saddle-point $(\mathbf{u}^h, p^h)$ of (3.4) out of $\mathbf{V}^h \times S^h$; i.e., they represent a discrete saddle-point problem. As a result, they lead to stable and accurate approximations of $(\mathbf{u}, p)$ if and only if the pair $(\mathbf{V}^h, S^h)$ satisfies the following conditions: first, the inf-sup condition (see [7, 15, 16]) *there exists $C > 0$, independent of h, such that*

$$\sup_{\mathbf{v}^h \in \mathbf{V}^h} \frac{B(\mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_1} \geq C \|q^h\|_0 \quad \forall\, q^h \in S^h,$$

and second, $A(\cdot,\cdot)$ is coercive on $\mathbf{Z}^h \times \mathbf{Z}^h$, where $\mathbf{Z}^h = \{\mathbf{v}^h \in \mathbf{V}^h \mid B(q^h, \mathbf{v}^h) = 0 \; \forall\, q \in S^h\}$ is the subspace of discretely solenoidal functions belonging to $\mathbf{V}^h$. Examples of unstable pairs include all equal order interpolation spaces defined with respect to the same triangulation of Ω into finite elements, as well as such combinations as the bilinear-constant pair; see [15, 16].

3.1. Continuous stabilized prototypes. In the literature, the term *finite element stabilization* is commonly applied to describe the application of various regularization techniques either to (3.4) or directly to (3.5)–(3.6) in order to circumvent the inf-sup condition. Stabilization leads to finite element methods that allow for an unrestricted choice of velocity and pressure spaces, including the choice of equal order interpolation. *Consistent stabilization* is one of the most popular types of regularization because it avoids penalty errors and can, in principle, be extended to achieve an arbitrarily high order of accuracy. Typically, consistently stabilized methods are defined at the discrete level and employ mesh-dependent norms and inner products. In this section, we formulate continuous prototypes for these methods. The prototypes represent idealized variational problems that can be used to derive practical finite element schemes. The origin of the continuous prototypes can be best understood by considering first the regularization of (3.4) by penalty. The relevant penalized Lagrangian functional is

$$L(\mathbf{v}, q) = \frac{1}{2} A(\mathbf{v}, \mathbf{v}) - F(\mathbf{v}) + B(\mathbf{v}, q) - \delta \|q\|_0^2. \tag{3.8}$$

The optimality system for (3.8) is to seek $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ such that

$$A(\mathbf{u}, \mathbf{v}) + B(\mathbf{v}, p) = F(\mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega), \tag{3.9}$$

$$B(\mathbf{u}, q) - \delta M(p, q) = 0 \quad \forall\, q \in L_0^2(\Omega), \tag{3.10}$$

where $M(p, q) = (p, q)_0$. Thus, the effect emanating from the penalty term in (3.8) is to relax the constraint in (3.3). In terms of algebraic problems, this means that instead of the indefinite problem (3.7), now finite element discretization yields a linear system of the form

$$\begin{pmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{B}^T & -\delta\mathbb{M} \end{pmatrix} \begin{pmatrix} \vec{\mathbf{u}} \\ \vec{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{0}} \end{pmatrix}, \tag{3.11}$$

having a "definite" coefficient matrix.[1] As a result, one can show that a finite element method based on (3.8) is stable for any conforming choice of $\mathbf{V}^h$ and S^h. The trouble with (3.8) is the penalty error that limits the order of approximation to $O(\sqrt{\delta})$, regardless of the interpolation order of the pair $(\mathbf{V}^h, S^h)$.

The idea of consistent stabilization is to modify (3.2) and (3.3) to a problem like (3.9) and (3.10) but without incurring a penalty error. This requires a term that will generate the desired stabilizing contribution but will vanish on all sufficiently smooth exact solutions. To construct such a term, note that thanks to (2.8)

$$C_P \|p\|_0 \le \|\nabla p\|_{-1} \le C \|p\|_0,$$

i.e., $\|\nabla p\|_{-1}$ is an equivalent norm on $L_0^2(\Omega)$. As a result, $\|\nabla p\|_{-1}^2$ will have the same stabilization effect as $\|p\|_0^2$. However, unlike the latter, $\|\nabla p\|_{-1}^2$ can be included via the residual of (3.1) and so, when added to (3.2)–(3.3), the term

$$\delta(-\triangle \mathbf{u} + \nabla p - \mathbf{f}, -\alpha \triangle \mathbf{v} + \nabla q)_{-1}$$

will generate the appropriate stabilizing contribution but without the penalty error. This leads to a family of *continuous stabilized prototypes*: seek $(\mathbf{u}^h, p^h) \in \mathbf{V}^h \times S^h$ such that

$$Q_\alpha^\beta(\mathbf{u}^h, p^h; \mathbf{v}^h, q^h) = F_\alpha^\beta(\mathbf{v}^h, q^h) \tag{3.12}$$

for all $(\mathbf{v}^h, q^h) \in \mathbf{V}^h \times S^h$, where

$$\begin{aligned} Q_\alpha^\beta(\mathbf{u}, p; \mathbf{v}, q) = {} & A(\mathbf{u}, \mathbf{v}) + B(\mathbf{v}, p) + \beta B(\mathbf{u}, q) \\ & -\delta(-\triangle \mathbf{u} + \nabla p, -\alpha \triangle \mathbf{v} + \beta \nabla q)_{-1} \end{aligned} \tag{3.13}$$

and

$$F_\alpha^\beta(\mathbf{v}, q) = F(\mathbf{v}) - \delta(\mathbf{f}, -\alpha \triangle \mathbf{v} + \beta \nabla q)_{-1} \tag{3.14}$$

[1] The matrix in (3.11) is definite in the sense that

$$\begin{pmatrix} \mathbb{A} & \mathbb{B} \\ -\mathbb{B}^T & +\delta\mathbb{M} \end{pmatrix},$$

which is obtained from the coefficient matrix in (3.11) by multiplying the lower block of equations by -1, is real, positive definite.

are a bilinear form $[\mathbf{H}_0^1(\Omega)\times L^2(\Omega)]^2 \mapsto \mathbb{R}$ and a linear functional $\mathbf{H}_0^1(\Omega)\times L^2(\Omega)\mapsto\mathbb{R}$ parametrized by α, β, and δ. In (3.13) and (3.14), α and β take on the values $\{-1,0,1\}$ and $\{-1,1\}$, respectively, and δ is a positive, real valued parameter. A method is called *absolutely stable* if the form Q_β^α is weakly or strongly coercive for all values of δ. If this is true only for selected values of δ, the method is called *conditionally stable*. In what follows, we will work exclusively with continuous pressure approximations, in which case we can write

$$\begin{aligned} Q_\alpha^\beta(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h) \equiv{}& A(\mathbf{u}^h,\mathbf{v}^h)+B^*(\mathbf{v}^h,p^h)+\beta B^*(\mathbf{u}^h,q^h)\\ &-\delta(-\triangle\mathbf{u}^h+\nabla p^h,-\alpha\triangle\mathbf{v}^h+\beta\nabla q^h)_{-1}. \end{aligned} \tag{3.15}$$

We call (3.12) *prototypes* because the $\mathbf{H}^{-1}(\Omega)$ inner product is not computable so that (3.13) or (3.15) and (3.14) cannot be used directly in a finite element method. However, if the $\mathbf{H}^{-1}(\Omega)$ inner product appearing in (3.13) or (3.15) and (3.14) is replaced by a discrete approximation, each prototype will give rise to a practical method. All methods that can be associated with a particular prototype by virtue of such a substitution form the stabilized class generated by this prototype.

Remark 3.1. While the stabilized problem (3.12) is a modification of an equation that represents an optimality system, it is not necessarily itself an optimality system of some modified Lagrangian. Many of the methods defined by (3.12) can only be derived as modifications of (3.5) and (3.6); i.e., they cannot be formulated starting from a modification of (3.4) and then deriving the associated optimality system.

Remark 3.2. If $\mathbf{u}$ is approximated by piecewise linear or bilinear finite element functions, the second order derivative terms in (3.13) vanish and the prototypes (3.12) reduce to a penalized formulation in which the Lagrangian functional (3.4) is penalized by $-\delta\|\nabla q\|_{-1}^2$.

Introducing the bilinear forms

$$D(\mathbf{u},\mathbf{v})=\delta(-\triangle\mathbf{u},-\triangle\mathbf{v})_{-1},\qquad C(\mathbf{v},q)=\delta(\triangle\mathbf{v},\nabla q)_{-1},$$

and

$$K(p,q)=\delta(\nabla p,\nabla q)_{-1}$$

defined on $\mathbf{H}_0^1(\Omega)\times\mathbf{H}_0^1(\Omega)$, $\mathbf{H}_0^1(\Omega)\times L^2(\Omega)$, and $L^2(\Omega)\times L^2(\Omega)$, respectively, we can write (3.15) in the form

$$\begin{aligned} Q_\alpha^\beta(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h) = {}& A(\mathbf{u}^h,\mathbf{v}^h)+B^*(\mathbf{v}^h,p^h)+\beta B^*(\mathbf{u}^h,q^h)\\ &-\alpha D(\mathbf{u}^h,\mathbf{v}^h)+\alpha C(\mathbf{v}^h,p^h)+\beta C(\mathbf{u}^h,q^h)-\beta K(p^h,q^h). \end{aligned}$$

It is then easy to see that the discrete system (3.12) is equivalent to a family of linear algebraic systems of the form

$$\begin{pmatrix} \mathbb{A}-\alpha\mathbb{D} & \mathbb{B}+\alpha\mathbb{C}\\ \beta(\mathbb{B}+\mathbb{C})^T & -\beta\mathbb{K}\end{pmatrix}\begin{pmatrix}\vec{\mathbf{u}}\\ \vec{\mathbf{p}}\end{pmatrix}=\begin{pmatrix}\vec{\mathbf{f}}_1\\ \vec{\mathbf{f}}_2\end{pmatrix}, \tag{3.16}$$

where the matrices $\mathbb{C}$, $\mathbb{D}$, and $\mathbb{K}$ are respectively deduced in the usual manner from the bilinear forms $C(\cdot,\cdot)$, $D(\cdot,\cdot)$, and $K(\cdot,\cdot)$.

Choosing different α and β gives rise to different bilinear forms in (3.13) and to different matrices in (3.16). It is easy to see that the choices $\{\alpha,\beta\}$ and $\{\alpha,-\beta\}$ define variational problems that can be derived from one another by simply changing

the pressure test function in (3.12) from q^h to $-q^h$. Likewise, the linear system (3.16) generated by the choice $\{\alpha, -\beta\}$ can be derived from that for the choice $\{\alpha, \beta\}$ by simply scaling the second row of blocks by -1. Therefore, the linear systems produced by the two choices $\{\alpha, \beta\}$ and $\{\alpha, -\beta\}$ are equivalent in the sense that they have exactly the same solution. [2] We will call these variational problems, along with their associated bilinear forms and linear algebraic systems, *complementary*. The choice of α determines the class of complementary forms while the two forms within each class are generated by selecting β equal to either 1 or -1.

For consistency with the established terminology, we call the prototype corresponding to $\alpha = 1$ *Galerkin least-squares*, or GLS. Since taking $\alpha = 0$ "simplifies" the weighting function, we call this class of methods *simplified Galerkin least-squares*, or SGLS. Finally, choosing $\alpha = -1$ "reflects" the sign of the second order term and so we refer to this prototype as *reflected Galerkin least-squares*, or RGLS.

The *standard members of the GLS, SGLS, and RGLS classes of methods* are obtained when the $\mathbf{H}^{-1}(\Omega)$ inner product appearing in (3.15) and (3.14) is approximated by a weighted $\mathbf{L}^2$ inner product in the following manner:

$$
\begin{aligned}
Q^{\beta}_{\alpha,h}(\mathbf{u}^h, p^h; \mathbf{v}^h, q^h) &= A(\mathbf{u}^h, \mathbf{v}^h) + B^*(\mathbf{v}^h, p^h) + \beta B^*(\mathbf{u}^h, q^h) \\
&\quad - \sum_{\mathcal{K}\in\mathcal{T}_h} \delta h^2_{\mathcal{K}} (-\triangle \mathbf{u}^h + \nabla p^h, -\alpha \triangle \mathbf{v}^h + \beta \nabla q^h)_{0,\mathcal{K}}
\end{aligned} \tag{3.17}
$$

and

$$
F^{\beta}_{\alpha,h}(\mathbf{v}^h, q^h) = F(\mathbf{v}^h) - \sum_{\mathcal{K}\in\mathcal{T}_h} \delta h^2_{\mathcal{K}} (\mathbf{f}, -\alpha \triangle \mathbf{v}^h + \beta \nabla q^h)_{0,\mathcal{K}}, \tag{3.18}
$$

respectively. When $\alpha = 1$ and $\beta = 1$, we recover from (3.17) and (3.18) the original GLS method of [17]. For $\alpha = 0$ and $\beta = -1$, they give the original pressure-Poisson stabilized mixed method of [18]. The case $\alpha = -1$ and $\beta = 1$ gives the method of [11].

The weighted L^2 norm is not a particularly accurate approximation of the negative norm. Its main defect is that

$$
C_1(h\|\mathbf{u}^h\|_0) \le \|\mathbf{u}^h\|_{-1} \le C_2 h^{-1}(h\|\mathbf{u}^h\|_0).
$$

This equivalence relation, including the factor h^{-1} in the upper bound, is sharp, and means that (3.17) is stable with respect to a mesh-dependent norm that is not uniformly (in h) equivalent to the norm on $\mathbf{H}^1(\Omega) \times L^2(\Omega)$. A more sophisticated but also more complicated approximation is to use a discrete equivalent proposed in [6] in the context of least-squares finite element methods. For stabilized methods based on this norm, we refer to [9].

Analyses of the standard GLS and RGLS methods in [17] and [11], respectively, classify the first one as a conditionally stable scheme and the second one as an absolutely stable scheme. This means that for $\alpha = 1$, the choice of δ in (3.17) and (3.18) is restricted to some finite interval $0 < \delta_0 \le \delta \le \delta_{\max}$, while for $\alpha = -1$, the form in (3.17) is stable for any positive δ. In both cases, theoretical classifications agree well with the practical stability of the respective finite element methods; see [13] and

[2] Although the choices $\{\alpha, \beta\}$ and $\{\alpha, -\beta\}$ yield the same solution, the algebraic properties of the corresponding coefficient matrices can be vastly different. As a result, the performance of iterative solution techniques can also be vastly different; cf. [2].

[2]. However, this is not so for the standard SGLS method. The formal analysis of [8] classified this method as conditionally stable, with a stability range estimate very close to that of the standard GLS method. In practice, after extensive numerical experiments, we found that the standard SGLS behaves much more like the absolutely stabilized RGLS method; see [2]. This unexpected practical stability prompted us to reexamine the SGLS class starting from its continuous prototype. Thus, for the remainder of this paper, our focus will be on SGLS methods.

4. Continuous SGLS. In this section, we show that the continuous SGLS prototype

$$Q_0^{\pm}(\mathbf{u},p;\mathbf{v},q) = A(\mathbf{u},\mathbf{v}) + B^*(\mathbf{v},p) \pm B^*(\mathbf{u},q) - \delta(-\triangle\mathbf{u} + \nabla p, \pm\nabla q)_{-1}$$

is absolutely stable.

THEOREM 4.1. *Let $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$ and $S^h \subset L_0^2(\Omega) \cap H^1(\Omega)$. Then $Q_0^-(\cdot;\cdot)$ is coercive for $0 < \delta < 4$ and $Q_0^{\pm}(\cdot;\cdot)$ are weakly coercive for any $\delta \geq 4$; i.e., there exists $C > 0$, independent of h, such that*

$$Q_0^-(\mathbf{u}^h,p^h;\mathbf{u}^h,p^h) \geq C\left(\|\mathbf{u}^h\|_1^2 + \|p^h\|_0^2\right) \quad \forall\, 0 < \delta < 4$$

and

$$\left.\begin{aligned} \sup_{(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h} \frac{Q_0^{\pm}(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h)}{\|\mathbf{u}^h\|_1 + \|p^h\|_0} &\geq C(\|\mathbf{u}^h\|_1 + \|p^h\|_0) \\ \sup_{(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h} \frac{Q_0^{\pm}(\mathbf{v}^h,q^h;\mathbf{u}^h,p^h)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0} &> 0 \end{aligned}\right\} \quad \forall \delta \geq 4$$

for any $(\mathbf{u}^h,p^h) \in \mathbf{V}^h \times S^h$.

Proof. Since complementary forms can be obtained from one another by changing the sign of the pressure test functions, it suffices to carry out the proofs for only one of the forms. Here, we choose to work with the minus form Q_0^-. Using (2.7), the stabilizing term in Q_0^- simplifies to

$$\delta(-\triangle\mathbf{u}^h + \nabla p^h, \nabla q^h)_{-1} = \delta((\mathbf{u}^h, \nabla q^h)_0 + (\nabla p^h, \nabla q^h)_{-1}).$$

As a result,

$$Q_0^-(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h) = A(\mathbf{u}^h,\mathbf{v}^h) + (\nabla p^h,\mathbf{v}^h)_0 + (\delta-1)(\nabla q^h,\mathbf{u}^h)_0 + \delta(\nabla p^h,\nabla q^h)_{-1}.$$

To prove strong the coercivity result, let δ be a number between 0 and 4 and consider $Q_0^-(\mathbf{u}^h,p^h;\mathbf{u}^h,p^h)$. Using Cauchy's inequality and the ϵ inequality,

$$\begin{aligned} Q_0^-(\mathbf{u}^h,p^h;\mathbf{u}^h,p^h) &= A(\mathbf{u}^h,\mathbf{u}^h) + \delta(\nabla p^h,\mathbf{u}^h)_0 + \delta(\nabla p^h,\nabla p^h)_{-1} \\ &\geq |\mathbf{u}^h|_1^2 + \delta\|\nabla p^h\|_{-1}^2 - \delta\|\nabla p^h\|_{-1}|\mathbf{u}^h|_1 \\ &\geq \left(1 - \frac{\delta}{2\epsilon}\right)|\mathbf{u}^h|_1^2 + \delta\left(1 - \frac{\epsilon}{2}\right)\|\nabla p^h\|_{-1}^2. \end{aligned}$$

To ensure coercivity, both coefficients above must be positive. Therefore, δ and ϵ must satisfy the inequalities

$$0 < \delta < 2\epsilon \quad \text{and} \quad \epsilon < 2.$$

This is always possible when $0 < \delta < 4$. Since $p^h \in L_0^2(\Omega)$ and $\mathbf{u}^h \in \mathbf{H}_0^1(\Omega)$, the final bound

$$Q_0^-(\mathbf{u}^h, p^h; \mathbf{u}^h, p^h) \geq C(\delta, C_P, C_N)\left(\|\mathbf{u}^h\|_1^2 + \|p\|_0^2\right)$$

follows from (2.8) and (2.1).

To show that Q_0^- is weakly coercive for $\delta \geq 4$, let $(\widetilde{\mathbf{v}}^h, \widetilde{q}^h) = (\mathbf{u}^h, \gamma p^h)$ for some positive γ. Then

$$Q_0^-(\mathbf{u}^h, p^h; \widetilde{\mathbf{v}}^h, \widetilde{q}^h) = |\mathbf{u}^h|_1^2 + \gamma\delta\|\nabla p^h\|_{-1}^2 + (1 + \gamma(\delta - 1))(\nabla p^h, \mathbf{u}^h)_0.$$

Letting $\gamma = 1/(\delta - 1)$, the Cauchy and ϵ inequalities further give

$$\begin{aligned} Q_0^-(\mathbf{u}^h, p^h; \widetilde{\mathbf{v}}^h, \widetilde{q}^h) &\geq |\mathbf{u}^h|_1^2 + \frac{\delta}{\delta - 1}\|\nabla p^h\|_{-1}^2 - 2\|\nabla p^h\|_{-1}|\mathbf{u}^h|_1 \\ &\geq (1 - \epsilon)|\mathbf{u}^h|_1^2 + \left(\frac{\delta}{\delta - 1} - \frac{1}{\epsilon}\right)\|\nabla p^h\|_{-1}^2. \end{aligned}$$

Since $\delta \geq 4$, we can always choose a positive ϵ such that

$$\frac{\delta - 1}{\delta} < \epsilon < 1.$$

This makes both coefficients in the lower bound positive and we can conclude that there exists $C(\delta, C_P, C_N)$, independent of h, such that

$$Q_0^-(\mathbf{u}^h, p^h; \widetilde{\mathbf{v}}^h, \widetilde{q}^h) \geq C(\delta, C_P, C_N)\left(\|\mathbf{u}^h\|_1^2 + \|p^h\|_0^2\right).$$

To complete the proof of the first weak coercivity condition, we note that $\|\widetilde{\mathbf{v}}^h\|_1 + \|\widetilde{q}^h\|_0 = \|\mathbf{u}^h\|_1 + \frac{1}{\delta - 1}\|p^h\|_0$ so that the last inequality can be recast into

$$Q_0^-(\mathbf{u}^h, p^h; \widetilde{\mathbf{v}}^h, \widetilde{q}^h) \geq C(\delta, C_P, C_N)\left(\|\mathbf{u}^h\|_1 + \|p^h\|_0\right)\left(\|\widetilde{\mathbf{v}}^h\|_1 + \|\widetilde{q}^h\|_0\right).$$

To prove the second weak coercivity condition, we choose $\mathbf{v}^h = -\mathbf{S}(\nabla p^h)$ and $q^h \equiv p^h$. Using Lemma 2.1

$$A(-\mathbf{S}(\nabla p^h), \mathbf{u}^h) = -(\nabla p^h, \mathbf{u}^h) \qquad \text{and} \qquad -\triangle(-\mathbf{S}(\nabla p^h)) = -\nabla p^h.$$

It is now easy to see that

$$Q_0^-(-\mathbf{S}(\nabla p^h), p^h; \mathbf{u}^h, p^h) = (\mathbf{S}(\nabla p^h), \nabla p^h)_0 = \|\nabla p^h\|_{-1}^2 > 0,$$

where the last identity follows again from Lemma 2.1. ∎

It is a straightforward matter to demonstrate that $Q_0^{\pm}$ is continuous. Then standard finite element arguments can be used to show that the method is optimally accurate.

THEOREM 4.2. *Let* $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \cap \mathbf{H}^{r+1}(\Omega) \times L_0^2(\Omega) \cap H^{s+1}(\Omega)$ *denote a solution of the Stokes problem and let* $(\mathbf{u}^h, p^h)$ *solve* (3.12) *for* $\alpha = 0$. *Then there exists a constant* $C > 0$ *independent of* h *such that*

$$\|\mathbf{u} - \mathbf{u}^h\|_1 + \|p - p^h\|_0 \leq C(h^r\|\mathbf{u}\|_{r+1} + h^{s+1}\|p\|_{s+1}).$$

We note for future reference that the stability and error estimates of the SGLS prototype are given in terms of the natural mesh-independent norm of $\mathbf{H}^1(\Omega) \times L^2(\Omega)$.

5. Discrete SGLS. While the continuous SGLS prototype is not a practical method, its analysis hints at a possibility that members of the SGLS family of methods may have far better stability properties than previously thought. In this section we will define a new member of this family that not only is practical but also inherits the absolute stability of its continuous prototype in terms of the same mesh-independent norms. In addition, the new method is also optimally accurate and converges at the same rate as the continuous prototype. To formulate and analyze the new method, we will make use of several discrete operators along with their relevant properties. These are reviewed next.

5.1. Discrete operators. Given a finite element subspace $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$, we define the *discrete Laplace operator* $-\triangle^h$ as the mapping $-\triangle^h : \mathbf{H}_0^1(\Omega) \mapsto \mathbf{V}^h$ such that $-\triangle^h \mathbf{u} = \mathbf{z}^h$ if and only if

$$(\mathbf{z}^h, \mathbf{v}^h)_0 = (\nabla \mathbf{u}, \nabla \mathbf{v}^h)_0 \quad \forall\, \mathbf{v}^h \in \mathbf{V}^h. \tag{5.1}$$

The *discrete inverse Laplace operator* $\mathbf{S}^h$ is the mapping $\mathbf{S}^h : \mathbf{H}^{-1}(\Omega) \mapsto \mathbf{V}^h$ such that $\mathbf{S}^h \mathbf{u} = \mathbf{z}^h$ if and only if

$$(\nabla \mathbf{z}^h, \nabla \mathbf{v}^h)_0 = (\mathbf{u}, \mathbf{v}^h)_0 \quad \forall\, \mathbf{v}^h \in \mathbf{V}^h. \tag{5.2}$$

The last operator that we will need is the L^2 *projection operator onto* $\mathbf{V}^h$. This operator is the mapping $\mathbf{Q}^h : \mathbf{L}^2(\Omega) \mapsto \mathbf{V}^h$ such that $\mathbf{Q}^h \mathbf{u} = \mathbf{z}^h$ if and only if

$$(\mathbf{z}^h, \mathbf{v}^h)_0 = (\mathbf{u}, \mathbf{v}^h)_0 \quad \forall\, \mathbf{v}^h \in \mathbf{V}^h. \tag{5.3}$$

If the supremum in (2.5) is restricted to the subspace $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$, we obtain the *discrete negative seminorm*

$$\|\mathbf{f}\|_{-h} = \sup_{\boldsymbol{\phi}^h \in \mathbf{V}^h} \frac{(\mathbf{f}, \boldsymbol{\phi}^h)_0}{|\boldsymbol{\phi}^h|_1} \quad \forall\, \mathbf{f} \in \mathbf{H}^{-1}(\Omega). \tag{5.4}$$

The next theorem summarizes the properties of the discrete operators and norms that are relevant to our analysis; for part 3, note that

$$\|(\mathbf{I} - \mathbf{Q}^h)\mathbf{u}\|_{-k} = \sup_{\boldsymbol{\phi} \in \mathbf{H}_0^k(\Omega)} \frac{((\mathbf{I} - \mathbf{Q}^h)\mathbf{u}, \boldsymbol{\phi})_0}{\|\boldsymbol{\phi}\|_k},$$

where $\mathbf{H}_0^k(\Omega) \equiv \mathbf{H}^k(\Omega) \cap \mathbf{H}_0^1(\Omega)$.

THEOREM 5.1. 1. *For any* $\mathbf{f}, \mathbf{g} \in \mathbf{H}^{-1}(\Omega)$, *define* $(\mathbf{f}, \mathbf{g})_{-h} = (\mathbf{S}^h \mathbf{f}, \mathbf{g})_0 = (\mathbf{f}, \mathbf{S}^h \mathbf{g})_0$. *Then*

$$\|\mathbf{f}\|_{-h}^2 = (\mathbf{f}, \mathbf{f})_{-h}. \tag{5.5}$$

2. *For any* $\mathbf{u} \in \mathbf{L}^2(\Omega)$

$$\|\mathbf{Q}^h \mathbf{u}\|_0 \le C_I h^{-1} \|\mathbf{u}\|_{-h}, \tag{5.6}$$

$$\|\mathbf{u}\|_{-1}^2 \le C \left(h^2 \|\mathbf{u}\|_0^2 + \|\mathbf{u}\|_{-h}^2 \right), \tag{5.7}$$

$$-\triangle^h \cdot \mathbf{S}^h \mathbf{u} = \mathbf{Q}^h \mathbf{u}. \tag{5.8}$$

3. *For any* $\mathbf{u} \in \mathbf{L}^2(\Omega)$ *and* $0 < k \leq r+1$

$$\|(\mathbf{I}-\mathbf{Q}^h)\mathbf{u}\|_{-k} \leq Ch^k\|\mathbf{u}\|_0. \tag{5.9}$$

Proof. For the proof of the characterization (5.5) and the lower equivalence bound (5.7), we refer to [5] or [6]. Here, we will only demonstrate the proofs for the inverse inequality (5.6), the identity (5.8), and the duality estimate (5.9).

Let $\mathbf{u} \in \mathbf{L}^2(\Omega)$. Using the definition (5.3) of $\mathbf{Q}^h$ in (5.4),

$$\|\mathbf{u}\|_{-h} = \sup_{\boldsymbol{\phi}^h \in \mathbf{V}^h} \frac{(\mathbf{u},\boldsymbol{\phi}^h)_0}{|\boldsymbol{\phi}^h|_1} = \sup_{\boldsymbol{\phi}^h \in \mathbf{V}^h} \frac{(\mathbf{Q}^h\mathbf{u},\boldsymbol{\phi}^h)_0}{|\boldsymbol{\phi}^h|_1} \geq \frac{(\mathbf{Q}^h\mathbf{u},\mathbf{Q}^h\mathbf{u})_0}{|\mathbf{Q}^h\mathbf{u}|_1}.$$

Using the first inequality in (2.4) for $\mathbf{Q}^h\mathbf{u}$ gives that

$$|\mathbf{Q}^h\mathbf{u}|_1 \leq C_I h^{-1}\|\mathbf{Q}^h\mathbf{u}\|_0.$$

As a result,

$$\|\mathbf{u}\|_{-h} \geq \frac{\|\mathbf{Q}^h\mathbf{u}\|_0^2}{|\mathbf{Q}^h\mathbf{u}|_1} \geq \frac{h\|\mathbf{Q}^h\mathbf{u}\|_0^2}{C_I\|\mathbf{Q}^h\mathbf{u}\|_0} = hC_I^{-1}\|\mathbf{Q}^h\mathbf{u}\|_0,$$

which proves (5.6). A straightforward application of (5.1)–(5.3) shows that

$$(-\triangle^h \mathbf{S}^h\mathbf{u},\mathbf{v}^h) = (\nabla(\mathbf{S}^h\mathbf{u}),\nabla\mathbf{v}^h) = (\mathbf{u},\mathbf{v}^h),$$

which proves (5.8). To prove (5.9), we use the definition (5.3) of $\mathbf{Q}^h$ and Cauchy's inequality to show that

$$((\mathbf{I}-\mathbf{Q}^h)\mathbf{u},\boldsymbol{\phi})_0 = (\mathbf{u},(\mathbf{I}-\mathbf{Q}^h)\boldsymbol{\phi})_0 \leq \|\mathbf{u}\|_0\|(\mathbf{I}-\mathbf{Q}^h)\boldsymbol{\phi}\|_0$$

and then use (2.2) to obtain

$$\|(\mathbf{I}-\mathbf{Q}^h)\boldsymbol{\phi}\|_0 \leq Ch^k\|\boldsymbol{\phi}\|_k.$$

Combining these bounds shows that

$$\|(\mathbf{I}-\mathbf{Q}^h)\mathbf{u}\|_{-k} \leq \sup_{\boldsymbol{\phi}\in\mathbf{H}_0^k(\Omega)} \frac{h^k C\|\mathbf{u}\|_0\|\boldsymbol{\phi}\|_k}{\|\boldsymbol{\phi}\|_k} = Ch^k\|\mathbf{u}\|_0. \qquad \square$$

5.2. An absolutely stable discrete SGLS method. We introduce the bilinear form

$$\begin{aligned} Q_{0,h}^{\pm}(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h) = {} & A(\mathbf{u}^h,\mathbf{v}^h) + B^*(\mathbf{v}^h,p^h) \pm B^*(\mathbf{u}^h,q^h) \\ & -\delta h^2(-\triangle^h\mathbf{u}^h + \nabla p^h, \pm\nabla q^h)_0 \end{aligned} \tag{5.10}$$

and the linear functional

$$F_{0,h}^{\pm}(\mathbf{v}^h,q^h) = F(\mathbf{v}^h) - \delta h^2(\mathbf{f},\pm\nabla q^h)_0.$$

The new member of the SGLS family of methods is to seek $(\mathbf{u}^h,p^h) \in \mathbf{V}^h \times S^h$ such that

$$Q_{0,h}^{\pm}(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h) = F_{0,h}^{\pm}(\mathbf{v}^h,q^h) \quad \forall\,(\mathbf{v}^h,q^h) \in \mathbf{V}^h \times S^h. \tag{5.11}$$

Before we continue with the stability and error analysis of the new method, let us point out that thanks to definition (5.1)

$$A(\mathbf{u}^h, \mathbf{v}^h) \equiv (\nabla \mathbf{u}^h, \nabla \mathbf{v}^h)_0 = (-\triangle^h \mathbf{u}^h, \mathbf{v}^h)_0.$$

As a result,

$$\begin{aligned} Q^{\pm}_{0,h}(\mathbf{u}^h, p^h; \mathbf{v}^h, q^h) &= (-\triangle^h \mathbf{u}^h + \nabla p^h, \mathbf{v}^h)_0 \pm B^*(\mathbf{u}^h, q^h) \\ &\quad -\delta h^2 \left(-\triangle^h \mathbf{u}^h + \nabla p^h, \pm \nabla q^h\right)_0 \\ &= \left(-\triangle^h \mathbf{u}^h + \nabla p^h, \mathbf{v}^h \mp \delta h^2 \nabla q^h\right)_0 \pm B^*(\mathbf{u}^h, q^h) \end{aligned} \tag{5.12}$$

is an equivalent representation of (5.10) and

$$(-\triangle^h \mathbf{u}^h + \nabla p^h, \mathbf{v}^h \mp \delta h^2 \nabla q^h)_0 \pm B^*(\mathbf{u}^h, q^h) = (\mathbf{f}, \mathbf{v}^h \mp \delta h^2 \nabla q^h)_0 \tag{5.13}$$

is an equivalent form of (5.11). Problem (5.13) leads to an interesting interpretation for the new method: it can be viewed as a Petrov–Galerkin-like scheme obtained by modification of the velocity weight function to $\mathbf{v}^h \mp \delta h^2 \nabla q^h$.

Because we have replaced $-\triangle \mathbf{u}^h$ with $-\triangle^h \mathbf{u}^h$ in the definition of the method, the term

$$(-\triangle^h \mathbf{u} + \nabla p - \mathbf{f}, \mp \delta h^2 \nabla q^h)_0 \neq 0;$$

i.e., the new method, is not, strictly speaking, a consistent formulation. However, as we will see in the next lemma, the inconsistency is very weak. In particular, we will prove that it does not degrade the optimal convergence rate of the method.

LEMMA 5.2. *Let* $(\mathbf{u}, p) \in \mathbf{H}^1_0(\Omega) \cap \mathbf{H}^{r+1}(\Omega) \times L^2_0(\Omega) \cap H^{s+1}(\Omega)$ *denote a solution of the Stokes problem and let* $(\mathbf{u}^h, p^h)$ *be a solution of* (5.11). *Then*

$$\begin{aligned} Q^{\pm}_{0,h}(\mathbf{u} - \mathbf{u}^h, p - p^h; \mathbf{v}^h, q^h) &= \delta h^2 (-\triangle \mathbf{u}, (\mathbf{Q}^h - \mathbf{I}) \nabla q^h)_0 \\ &\leq \delta C h^r \|\mathbf{u}\|_{r+1} \|q^h\|_0 \end{aligned} \tag{5.14}$$

for all $(\mathbf{v}^h, q^h) \in \mathbf{V}^h \times S^h$.

Proof. Consider the minus form. It is easy to see that

$$\begin{aligned} Q^{-}_{0,h}(\mathbf{u} - \mathbf{u}^h, p - p^h; \mathbf{v}^h, q^h) &= \delta h^2 (-\triangle^h \mathbf{u} + \nabla p - \mathbf{f}, \nabla q^h)_0 \\ &= \delta h^2 (-(\triangle^h - \triangle) \mathbf{u}, \nabla q^h)_0. \end{aligned}$$

From the fact that $-\triangle^h \mathbf{u} \in \mathbf{V}^h$, the definition (5.3) of the L^2 projection, and the definition (5.1) of $-\triangle^h$, it follows that

$$(-\triangle^h \mathbf{u}, \nabla q^h)_0 = (-\triangle^h \mathbf{u}, \mathbf{Q}^h \nabla q^h)_0 = (\nabla \mathbf{u}, \nabla \mathbf{Q}^h \nabla q^h)_0 = (-\triangle \mathbf{u}, \mathbf{Q}^h \nabla q^h)_0$$

and so

$$(-(\triangle^h - \triangle) \mathbf{u}, \nabla q^h)_0 = (-\triangle \mathbf{u}, (\mathbf{Q}^h - \mathbf{I}) \nabla q^h)_0$$

so that the equality in (5.14) is proved. Next, with the help of (5.9) and the inverse inequality (2.4), we have

$$\begin{aligned} (-\triangle \mathbf{u}, (\mathbf{Q}^h - \mathbf{I}) \nabla q^h)_0 &\leq \|\triangle \mathbf{u}\|_{r-1} \, \|(\mathbf{Q}^h - \mathbf{I}) \nabla q^h\|_{1-r} \\ &\leq C h^{r-1} \|\mathbf{u}\|_{r+1} \|\nabla q^h\|_0 \\ &\leq C h^{r-2} \|\mathbf{u}\|_{r+1} \|q^h\|_0, \end{aligned}$$

from which the inequality in (5.14) follows. ∎

5.3. Stability and convergence. The main results of this section are to show that the method (5.11) is absolutely stable and that finite element solutions of (5.11) converge at optimal rates. We begin by establishing the absolute stability of the method, i.e., that the bilinear form (5.10) is weakly coercive for all values of the parameter δ. The proof relies upon a technical result presented in the next lemma.

LEMMA 5.3. *For any* $q^h \in S^h$

$$\|\nabla q^h\|_{-1}^2 \le C\left(h^2\|(\mathbf{I}-\mathbf{Q}^h)\nabla q^h\|_0^2 + \|\nabla q^h\|_{-h}^2\right). \tag{5.15}$$

Proof. Since we restrict attention to continuous pressure approximations, $\nabla q^h \in L^2(\Omega)$. Therefore, (5.7) from Theorem 5.1 implies that

$$\|\nabla q^h\|_{-1}^2 \le C\left(h^2\|\nabla q^h\|_0^2 + \|\nabla q^h\|_{-h}^2\right).$$

Adding and subtracting $\mathbf{Q}^h\nabla q^h$ to the first term and using the triangle inequality give the upper bound

$$\|\nabla q^h\|_{-1}^2 \le C\left(h^2\|(\mathbf{I}-\mathbf{Q}^h)\nabla q^h\|_0^2 + h^2\|\mathbf{Q}^h\nabla q^h\|_0^2 + \|\nabla q^h\|_{-h}^2\right).$$

The lemma follows by using the inverse inequality (5.6) to bound $h^2\|\mathbf{Q}^h\nabla q^h\|_0^2$ by $C_I\|\nabla q^h\|_{-h}^2$. ∎

THEOREM 5.4. *Assume that* $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$ *and* $S^h \subset L_0^2(\Omega)\cap H^1(\Omega)$. *Then, for any* $\delta > 0$, *there exists a positive constant* $C(\delta)$, *independent of* h, *such that*

$$\begin{aligned}
&\sup_{(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h} \frac{Q_{0,h}^{\pm}(\mathbf{u}^h,p^h;\mathbf{v}^h,q^h)}{\|\mathbf{v}^h\|_1+\|q^h\|_0} \ge C(\delta)(\|\mathbf{u}^h\|_1+\|p^h\|_0),\\
&\sup_{(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h} \frac{Q_{0,h}^{\pm}(\mathbf{v}^h,q^h;\mathbf{u}^h,p^h)}{\|\mathbf{v}^h\|_1+\|q^h\|_0} > 0
\end{aligned} \tag{5.16}$$

for all $(\mathbf{u}^h,p^h)\in\mathbf{V}^h\times S^h$.

Proof. We recall that the complementary plus and minus forms define equivalent problems, and so it suffices to carry the proof for just one of the forms. Here, we choose again to work with the minus form. Given a positive δ, we will construct a test function $(\widetilde{\mathbf{v}}^h,\widetilde{q}^h)$ such that

$$Q_{0,h}^{-}(\mathbf{u}^h,p^h;\widetilde{\mathbf{v}}^h,\widetilde{q}^h) \ge C(\|\mathbf{u}^h\|_1+\|p^h\|_0)\left(\|\widetilde{\mathbf{v}}^h\|_1+\|\widetilde{q}^h\|_0\right).$$

To find such a function, note that definition (5.2) implies the identity

$$(\nabla\mathbf{u}^h,\nabla\mathbf{S}^h(\nabla q^h))_0 = (\mathbf{u}^h,\nabla q^h)_0.$$

Thus, if $q^h\in S^h$ is arbitrary and $\mathbf{v}_1^h = \mathbf{S}^h(\nabla q^h)$,

$$Q_{0,h}^{-}(\mathbf{u}^h,p^h;\mathbf{v}_1^h,q^h) = (\nabla p^h,\mathbf{S}^h\nabla q^h)_0 + \delta h^2(-\triangle^h\mathbf{u}^h+\nabla p^h,\nabla q^h)_0.$$

Adding and subtracting $\mathbf{Q}^h\nabla p^h$ from the last term give

$$\begin{aligned}
Q_{0,h}^{-}(\mathbf{u}^h,p^h;\mathbf{v}_1^h,q^h) = (\nabla p^h,\mathbf{S}^h\nabla q^h)_0 &+ \delta h^2((\mathbf{I}-\mathbf{Q}^h)\nabla p^h,\nabla q^h)_0\\
&+\delta h^2(-\triangle^h\mathbf{u}^h+\mathbf{Q}^h\nabla p^h,\nabla q^h)_0
\end{aligned}$$

while the orthogonality

$$((\mathbf{I}-\mathbf{Q}^h)\nabla p^h,\mathbf{Q}^h\nabla q^h) = 0$$

and the fact that $-\triangle^h \mathbf{u}^h + \mathbf{Q}^h \nabla p^h \in \mathbf{V}^h$ allow us to rewrite the last identity as

$$\begin{aligned} Q^-_{0,h}(\mathbf{u}^h, p^h; \mathbf{v}_1^h, q^h) &= (\nabla p^h, \mathbf{S}^h \nabla q^h)_0 + \delta h^2((\mathbf{I}-\mathbf{Q}^h)\nabla p^h, (\mathbf{I}-\mathbf{Q}^h)\nabla q^h)_0 \\ &\quad + \delta h^2 \left(-\triangle^h \mathbf{u}^h + \mathbf{Q}^h \nabla p^h, \mathbf{Q}^h \nabla q^h\right)_0. \end{aligned} \tag{5.17}$$

Next, (5.12) implies that

$$Q^-_{0,h}(\mathbf{u}^h, p^h; \mathbf{v}^h, 0) = (-\triangle^h \mathbf{u}^h + \mathbf{Q}^h \nabla p^h, \mathbf{v}^h)_0.$$

Choosing $\mathbf{v}_2^h = -\delta h^2 \mathbf{Q}^h \nabla q^h$ then gives the identity

$$Q^-_{0,h}(\mathbf{u}^h, p^h; \mathbf{v}_2^h, 0) = -\delta h^2 (-\triangle^h \mathbf{u}^h + \mathbf{Q}^h \nabla p^h, \mathbf{Q}^h \nabla q^h)_0. \tag{5.18}$$

Therefore, if $q^h = p^h$, (5.17), (5.18), and (5.15) together with the discrete negative norm characterization in (5.5) imply that

$$\begin{aligned} Q^-_{0,h}(\mathbf{u}^h, p^h; \mathbf{v}_1^h + \mathbf{v}_2^h, p^h) &= (\nabla p^h, \mathbf{S}^h \nabla p^h)_0 + \delta h^2((\mathbf{I}-\mathbf{Q}^h)\nabla p^h, (\mathbf{I}-\mathbf{Q}^h)\nabla p^h)_0 \\ &= \|\nabla p^h\|^2_{-h} + \delta h^2 \|(\mathbf{I}-\mathbf{Q}^h)\nabla p^h\|_0^2 \geq C(\delta)\|\nabla p^h\|^2_{-1}. \end{aligned}$$

Since $p^h \in L^2_0(\Omega)$, the last inequality in combination with (2.8) gives a bound in terms of L^2 pressure norm:

$$Q^-_{0,h}(\mathbf{u}^h, p^h; \mathbf{v}_1^h + \mathbf{v}_2^h, p^h) \geq C_1(\delta)\|p^h\|_0^2. \tag{5.19}$$

To complete the proof of the first weak coercivity condition, note that

$$\begin{aligned} Q^-_{0,h}(\mathbf{u}^h, p^h; \mathbf{u}^h, 0) &= |\mathbf{u}^h|_1^2 + (\nabla p^h, \mathbf{u}^h)_0 = |\mathbf{u}^h|_1^2 - (p^h, \nabla \cdot \mathbf{u}^h)_0 \\ &\geq C_P^2 \|\mathbf{u}^h\|_1^2 - \sqrt{n}\|p^h\|_0 \|\mathbf{u}^h\|_1 \geq \frac{C_P^2}{2}\|\mathbf{u}^h\|_1^2 - \frac{n}{2C_P^2}\|p^h\|_0^2. \end{aligned}$$

Therefore, letting $\mathbf{v}_3^h = n^{-1} C_1(\delta) C_P^2 \mathbf{u}^h$ gives

$$Q^-_{0,h}(\mathbf{u}^h, p^h; \mathbf{v}_3^h, 0) \geq \frac{C_1(\delta) C_P^4}{2n}\|\mathbf{u}^h\|_1^2 - \frac{C_1(\delta)}{2}\|p^h\|_0^2,$$

where $C_1(\delta)$ is the constant from (5.19). As a result,

$$Q^-_{0,h}(\mathbf{u}^h, p^h; \mathbf{v}_1^h + \mathbf{v}_2^h + \mathbf{v}_3^h, p^h) \geq \frac{C_1(\delta) C_P^4}{2n}\|\mathbf{u}^h\|_1^2 + \frac{C_1(\delta)}{2}\|p^h\|_0^2 \tag{5.20}$$

and the association

$$(\widetilde{\mathbf{v}}^h, \widetilde{q}^h) = (\mathbf{v}_1^h + \mathbf{v}_2^h + \mathbf{v}_3^h, p^h)$$

will fit our purpose if we can show that $\|\widetilde{\mathbf{v}}^h\|_1 + \|\widetilde{q}^h\|_0$ is bounded by $\|\mathbf{u}^h\|_1 + \|p^h\|_0$. Using Poincaré's inequality (2.1), we have

$$\begin{aligned} \|\widetilde{\mathbf{v}}^h\|_1 &\leq C\|\nabla \widetilde{\mathbf{v}}^h\|_0 \\ &\leq C(\|\nabla \mathbf{v}_1^h\|_0 + \|\nabla \mathbf{v}_2^h\|_0 + \|\nabla \mathbf{v}_3^h\|_0) \\ &\leq C(\|\nabla(\mathbf{S}^h \nabla p^h)\|_0 + \delta h^2 \|\nabla(\mathbf{Q}^h \nabla p^h)\|_0 + \|\nabla \mathbf{u}^h\|_0). \end{aligned}$$

To estimate the first term, we use the definition of $\mathbf{S}^h$ and Poincaré's inequality to find that

$$\begin{aligned}\|\nabla(\mathbf{S}^h\nabla p^h)\|_0^2 &= (\nabla\mathbf{S}^h\nabla p^h, \nabla\mathbf{S}^h\nabla p^h)_0\\ &= (\nabla p^h, \mathbf{S}^h\nabla p^h)_0 = -(p^h, \nabla\cdot\mathbf{S}^h\nabla p^h)_0\\ &\le \sqrt{n}\|p^h\|_0\|\mathbf{S}^h\nabla p^h\|_1 \le C\|p^h\|_0\|\nabla\mathbf{S}^h\nabla p^h\|_0\end{aligned}$$

and, as a result,

$$\|\nabla(\mathbf{S}^h\nabla p^h)\|_0 \le C\|p^h\|_0.$$

For the second term, application of the inverse inequality (2.4) twice and the fact that $\mathbf{Q}^h$ is bounded gives

$$\delta h^2\|\nabla(\mathbf{Q}^h\nabla p^h)\|_0 \le \delta h C_I\|\mathbf{Q}^h\nabla p^h\|_0 \le \delta h C_I\|\nabla p^h\|_0 \le \delta C_I^2\|p^h\|_0.$$

Combining all bounds shows that

$$\|\widetilde{\mathbf{v}}^h\|_1 \le C(\|p^h\|_0 + \|\mathbf{u}^h\|_1)$$

and so we can rewrite (5.20) as

$$Q_{0,h}^-(\mathbf{u}^h, p^h; \widetilde{\mathbf{v}}^h, \widetilde{q}^h) \ge C(\|\mathbf{u}^h\|_1 + \|p^h\|_0)(\|\widetilde{\mathbf{v}}^h\|_1 + \|\widetilde{q}^h\|_0),$$

which proves the first part of (5.16). To prove the second weak coercivity condition we proceed as in the proof of Theorem 4.1 and set $\mathbf{v}^h = -\mathbf{S}^h\nabla p^h$ and $q^h \equiv p^h$. Using definitions (5.1)–(5.3) and Lemma 5.3, we find that

$$\begin{aligned}Q_{0,h}^-(-\mathbf{S}^h\nabla p^h, p^h; \mathbf{u}^h, p^h) &= (\mathbf{S}^h\nabla p^h, \nabla p^h) + \delta h^2((\mathbf{I}-\mathbf{Q}^h)\nabla p^h, \nabla p^h)\\ &= \|\nabla p^h\|_{-h}^2 + \delta h^2\|(\mathbf{I}-\mathbf{Q}^h)\nabla p^h\|_0^2 \ge C(\delta)\|\nabla p^h\|_0^2 > 0. \qquad \square\end{aligned}$$

This theorem shows that the new discrete method is stable with respect to the same norms as its continuous prototype, i.e., the natural norm on $\mathbf{H}^1(\Omega)\times L^2(\Omega)$. This valuable feature of the new method distinguishes it from the standard discrete SGLS of [18], which is stable with respect to a mesh-dependent norm.

Let us now consider the convergence of finite element solutions. The next theorem shows that the new method yields the same convergence rates as its continuous prototype with respect to the same mesh-independent norms.

THEOREM 5.5. *Let* $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega)\cap\mathbf{H}^{r+1}(\Omega)\times L_0^2(\Omega)\cap H^{s+1}(\Omega)$ *denote a solution of the Stokes problem and let* $(\mathbf{u}^h, p^h)$ *solve* (5.11). *Then*

$$\|\mathbf{u}-\mathbf{u}^h\|_1 + \|p-p^h\|_0 \le C(h^r\|\mathbf{u}\|_{r+1} + h^{s+1}\|p\|_{s+1}). \tag{5.21}$$

Proof. We begin by splitting the error into discrete and approximation theoretic parts:

$$\|\mathbf{u}-\mathbf{u}^h\|_1 + \|p-p^h\|_0 \le (\|\mathbf{u}_I^h-\mathbf{u}^h\|_1 + \|p_I^h-p^h\|_0) + (\|\mathbf{u}-\mathbf{u}_I^h\|_1 + \|p-p_I^h\|_0).$$

Since the interpolation error is of optimal order, to prove the theorem it suffices to

estimate the discrete error. Using (5.16),

$$C(\delta)(\|\mathbf{u}_I^h - \mathbf{u}^h\|_1 + \|p_I^h - p^h\|_0) \le \sup_{(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h} \frac{Q_{0,h}^{\pm}(\mathbf{u}_I^h - \mathbf{u}^h, p_I^h - p^h; \mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0}$$

$$\le \sup_{(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h} \frac{Q_{0,h}^{\pm}(\mathbf{u} - \mathbf{u}^h, p - p^h; \mathbf{v}^h, q^h) + Q_{0,h}^{\pm}(\mathbf{u}_I^h - \mathbf{u}, p_I^h - p; \mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0}$$

$$\le \sup_{(\mathbf{v}^h,q^h)\in\mathbf{V}^h\times S^h} \frac{Q_{0,h}^{\pm}(\mathbf{u} - \mathbf{u}^h, p - p^h; \mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_1 + \|q^h\|_0} + C(\|\mathbf{u} - \mathbf{u}_I^h\|_1 + \|p - p_I^h\|_0)$$

$$\le \sup_{q^h\in S^h} \frac{\delta h^2(-\triangle\mathbf{u}, (\mathbf{Q}^h - \mathbf{I})\nabla q^h)_0}{\|q^h\|_0} + C(h^r\|\mathbf{u}\|_{r+1} + h^{s+1}\|p\|_{s+1}),$$

where to obtain the last bound we have used (5.14) in Lemma 5.2 and (2.2) and (2.3). From (5.14), it easily follows that

$$\sup_{q^h\in S^h} \frac{\delta h^2(-\triangle\mathbf{u}, (\mathbf{Q}^h - \mathbf{I})\nabla q^h)_0}{\|q^h\|_0} \le Ch^r\|\mathbf{u}\|_{r+1}.$$

This means that the discrete error is of optimal order, i.e.,

$$\|\mathbf{u}_I^h - \mathbf{u}^h\|_1 + \|p_I^h - p^h\|_0 \le C(h^r\|\mathbf{u}\|_{r+1} + h^{s+1}\|p\|_{s+1})$$

and since the interpolation error is of the same order, (5.21) immediately follows. ∎

6. Concluding remarks. Using the notion of continuous prototypes, we formulated a new absolutely stable method for the Stokes problem. The new method is a close relative of the standard pressure-Poisson stabilized method of [18] in the sense that they share the same continuous prototype.

However, the two methods differ in several important aspects. The new formulation is weakly inconsistent in the sense that, although being strictly speaking not consistent, it still leads to optimal error estimates for all C^0 finite element subspaces, including the lowest order piecewise linear case. In contrast, the standard method not only is not consistent for piecewise linear approximations (because the Laplace operator annihilates the linear velocity field in (3.17) and (3.18)) but also results in errors that do not vanish with vanishing grid sizes; i.e., there remains an error proportional to the parameter δ. Furthermore, the new method is stable with respect to the norm on $\mathbf{H}^1(\Omega) \times L^2(\Omega)$, while the standard method is stable with respect to a mesh-dependent norm that is not *equivalent* to the norm on $\mathbf{H}^1(\Omega) \times L^2(\Omega)$.

Implementation of the new method requires evaluation of the discrete operator $-\triangle^h$. Given a finite element function $\mathbf{u}^h \in \mathbf{V}^h$, the coefficients $\vec{\mathbf{z}}$ of $\mathbf{z}^h = -\triangle^h\mathbf{u}^h$ can be determined from definition (5.1) by solving the linear system

$$\mathbb{M}\,\vec{\mathbf{z}} = \vec{\mathbf{r}}.$$

$\mathbb{M}$ is a mass matrix that can be assembled in the usual manner and $\vec{\mathbf{r}}$ is a vector with components

$$\vec{\mathbf{r}}_i = (\nabla\mathbf{u}^h, \nabla\boldsymbol{\phi}_i^h)_0,$$

where $\{\boldsymbol{\phi}_k^h\}_{k=1}^N$ is a nodal basis for $\mathbf{V}^h$. In practical computations, $\mathbb{M}$ can be replaced by a lumped mass matrix or local projection.

While computation of $-\triangle^h$ may seem as an additional overhead compared to the implementation of the standard method, it is well worth the effort thanks to the improved accuracy, especially when *piecewise linear* finite elements are used, and the guaranteed absolute, mesh-independent stability of the new method. It should be mentioned that essentially the same auxiliary problem, involving the inversion of a mass matrix, arises in standard stabilized methods with improved consistency; see [19]. These methods aim to restore the loss of consistency caused by the annihilation of all second order derivatives in the element residual when piecewise linear elements are used. The idea of [19] is to apply an L^2 projection to the first derivative of the finite element solution *before* the application of the second derivative so as to avoid its annihilation. Specialized to our context, this method can be viewed as providing an alternative definition for the discrete Laplace operator. Instead of the operator $-\triangle^h : \mathbf{H}_0^1(\Omega) \mapsto \mathbf{V}^h$ used in our method, they use the operator $-\triangle_A^h : \mathbf{V}^h \mapsto \mathbf{L}^2(\Omega)$ defined by

$$-\triangle_A^h = -\nabla \cdot (\mathbf{Q}^h \nabla \mathbf{u}^h). \tag{6.1}$$

Let us conclude by noting that an important open question that remains to be answered is whether or not the absolute stability of the continuous SGLS prototype is inherited by other members of this class. It seems particularly worthwhile to exploit extensions of our analysis to an SGLS method defined using the alternative discrete operator (6.1) and to the original pressure-Poisson method of [18] which, as we recall, behaves numerically just like an absolutely stable formulation. Extending our results to discontinuous pressure spaces would also be valuable.

REFERENCES

[1] C. Baiocchi and F. Brezzi, *Stabilization of unstable numerical methods*, in Proceedings of Problemi attuali dell' analisi e della fisica matematica, Taormina, Rome, 1992, pp. 59–63.

[2] T. Barth, P. Bochev, M. Gunzburger, and J. Shadid, *A taxonomy of consistently stabilized finite element methods for the Stokes problem*, SIAM J. Sci. Comput., 25 (2004), pp. 1585–1607.

[3] M. Behr, L. Franca, and T. Tezduyar, *Stabilized finite element methods for the velocity-pressure-stress formulation of incompressible flows,* Comput. Methods Appl. Mech. Engrg., 104 (1993), pp. 31–48.

[4] P. Bochev and M. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[5] J. Bramble, R. Lazarov, and J. Pasciak, *A Least Squares Approach Based on a Discrete Minus One Inner Product for First Order Systems*, Technical Report 94-32, Math. Sci. Institute, Cornell University, Ithaca, NY, 1994.

[6] J. Bramble and J. Pasciak, *Least-squares methods for Stokes equations based on a discrete minus one inner product*, J. Comput. Appl. Math., 74 (1996), pp. 155–173.

[7] F. Brezzi, *On existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers,* RAIRO Modél. Math. Anal. Numér., 21 (1974), pp. 129–151.

[8] F. Brezzi and J. Douglas, *Stabilized mixed methods for the Stokes problem,* Numer. Math., 53 (1988), pp. 225–235.

[9] Z. Cai and J. Douglas, *Stabilized finite element methods with fast iterative solution algorithms for the Stokes problem*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 115–129.

[10] P. Ciarlet, *Finite Element Methods for Elliptic Problems,* North-Holland, Amsterdam, 1978, reprinted as Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[11] J. Douglas and J. Wang, *An absolutely stabilized finite element method for the Stokes problem,* Math. Comp., 52 (1989), pp. 495–508.

[12] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to Numerical Solution of Boundary Value Problems*, Stud. Math. Appl. 15, North-Holland, Amsterdam, 1983.

[13] L. Franca, S. Frey, and T. Hughes, *Stabilized finite element methods:* I. *Application to the advective-diffusive model,* Comput. Methods Appl. Mech. Engrg., 95 (1992), pp. 253–276.

[14] L. Franca and R. Stenberg, *Error analysis of some Galerkin least-squares methods for the elasticity equations,* SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.

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

[16] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows,* Academic, Boston, 1989.

[17] T. Hughes and L. Franca, *A new finite element formulation for computational fluid dynamics:* VII. *The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity pressure spaces*, Comput. Methods Appl. Mech. Engrg., 65 (1987), pp. 85–96.

[18] T. Hughes, L. Franca, and M. Balestra, *A new finite element formulation for computational fluid dynamics:* V. *Circumventing the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations,* Comput. Methods Appl. Mech. Engrg., 59 (1986), pp. 85–99.

[19] K. Jansen, S. Collis, C. Whiting, and F. Shakib, *A better consistency for low-order stabilized finite element methods*, Comput. Methods Appl. Mech. Engrg., 174 (1999), pp. 153–170.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1208–1225

ANALYSIS OF THE BREZZI–PITKÄRANTA STABILIZED GALERKIN SCHEME FOR CREEPING FLOWS OF BINGHAM FLUIDS*

J.-C. LATCHÉ† AND D. VOLA†

Abstract. In this paper we propose and analyze a finite element scheme for a class of variational nonlinear and nondifferentiable mixed inequalities including balance equations governing incompressible creeping flows of Bingham fluids. For numerical efficiency reasons, equal-order piecewise linear approximations are used for both velocity and pressure, and the numerical scheme is stabilized by a Brezzi–Pitkäranta perturbation term. We obtain error estimates of the same order as for stable discretizations, namely $h^{1/2}$ for velocity and pressure solutions in $\left[\mathbf{H}^2(\Omega)\right]^d$ and $\mathbf{H}^1(\Omega)$, respectively. A decomposition-coordination algorithm to solve the discrete nonlinear algebraic system is presented, together with its convergence properties. Finally, numerical tests are performed. The solution of the problem under consideration presents particular regularity properties that are shown to permit convergence order improvement to $h|\log(h)|^{1/2}$. This estimate is confirmed by numerical results.

Key words. variational inequality, finite element method, Brezzi–Pitkäranta stabilization, error bound, Bingham fluid, creeping flow, decomposition- coordination method, augmented Lagrangian, algorithm convergence

AMS subject classifications. 65N12, 65N15, 65N30, 76D07, 76A05

DOI. 10.1137/S0036142903424386

1. Introduction.

Balance equations governing Bingham fluid flows take the form of nonlinear and nondifferentiable variational inequality problems, and their numerical solution is still a challenging task.

Steady one-directional flow in a pipe was the first of this type of flows to be the subject of an in-depth study. In the early 1980s, Fortin and Glowinski developed the so-called decomposition-coordination method for this particular situation; see [8], [10] and references therein. The principle of this numerical method is to isolate the nonlinear and nondifferentiable terms in the variational problem by introducing an auxiliary variable, the strain rate tensor, to enforce the consistency of this new variable and the velocity field by a duality method, and to solve the mixed problem by Uzawa algorithm variants. If the approximation space for the strain rate tensor is chosen so that no interelement continuity is required, it appears that the global nonlinear minimization problem degenerates in the Uzawa algorithm into a family of element-related subproblems which then can be solved efficiently, and even explicitly for a piecewise constant approximation. This aspect represents the main interest of the decomposition-coordination technique. In addition, the use of the constitutive relation is localized to the subproblems, which allows effortless changes. This numerical method was first used by Begis [10, Chapter VI] in the stream-function-vorticity formulation framework and then applied by several authors to pipe flows; see [13], [17].

The extension of this method to multidimensional incompressible flows was tested by Fortin, Côté, and Tanguy [7] and Roquet and Saramito [16]. To our knowledge,

*Received by the editors March 14, 2003; accepted for publication (in revised form) September 11, 2003; published electronically September 18, 2004.
http://www.siam.org/journals/sinum/42-3/42438.html
†Direction de Pr'evention des Accidents Majeurs (DPAM), Institut de Radioprotection et de Sûreté Nucléaire (IRSN), BP3-13115 St. Paul-lez-Durance CEDEX, France (jean-claude.latche@irsn.fr, didier.vola@irsn.fr).

the first analysis of this scheme is due to Han and Reddy [11], who dealt with the same mathematical problem for elastoplasticity applications (see [3] for a formulation for Bingham fluid flows and in the frame of the decomposition-coordination method). Convergence is proven and the error is found to be bounded by the square root of the so-called interpolation error, the latter depending on the discretization spaces and solution regularity. This analysis is based on the usual discrete Babuska–Brezzi stability condition for velocity/pressure approximation. Unfortunately, in our applications to Bingham fluid flows this limitation appeared to be somewhat restrictive, because it leads us to employ high degree approximations for the velocity that are not well suited to the poor regularity which can be expected from the solution. Moreover, the construction of strain rate approximation spaces which preserve the scheme accuracy and efficiency for high degree velocity approximations did not appear straightforward.

This explains the attractiveness of using piecewise linear equal-order approximation for the velocity and the pressure. This finite element is known to be cost-effective, and the discretization of the strain rate by piecewise constants would allow us to match the coherence constraint between velocity and strain rate perfectly. Of course, a stabilizing procedure must be applied with this approximation. The usual derivation of consistently stabilized schemes (e.g., [12], [9]) requires making the numerical residuals explicit which, in turn, needs a "strong differential formulation" of the problem. The latter can be obtained by regularizing the constitutive law. A numerical method built following these lines is presented in [14]. Conversely, if one chooses to deal with the problem without regularization, the natural choice then seems to be to implement a Brezzi–Pitkäranta stabilization [4]. The main purpose of this paper is to analyze such a scheme, combining ideas from [11] and from the analysis of stabilized schemes for Newtonian flows [9].

In section 2, we present the error analysis of a stabilized approximation of an abstract mixed nonlinear variational inequality. The application of this result to Bingham creeping flows problems is performed in section 3. Section 4 is devoted to the extension of the decomposition-coordination method to the stabilized scheme and the presentation of the solution algorithm and of its convergence properties. In section 5, we present numerical tests. As the solution of the considered problem presents extra regularity properties, the analysis of sections 2 and 3 yields an improved error bound, which is established.

Throughout this paper, Ω stands for an open and bounded subset of $\mathbb{R}^d$, $d \leq 3$, with Lipschitz domain boundary $\partial\Omega$, and we use standard notation for Sobolev spaces $\mathbf{L}^2(\Omega)$, $\mathbf{H}^1(\Omega)$, $\mathbf{H}_0^1(\Omega)$, $\mathbf{H}^2(\Omega)$, (see [1]). $\|.\|_0$ and $\|.\|_1$ stand for the norms of $\mathbf{L}^2(\Omega)$ and $\mathbf{H}^1(\Omega)$, and $|.|_1$ and $|.|_2$ are the usual seminorms of $\mathbf{H}^1(\Omega)$ and $\mathbf{H}^2(\Omega)$, respectively.

2. An abstract convergence result. The aim of this section is to analyze a numerical scheme to solve the following abstract variational inequality:

(2.1)

Find u and p in $\left[\mathbf{H}_0^1(\Omega)\right]^d$ and $\mathbf{L}^2(\Omega)$, respectively, such that

$$\left|\begin{array}{ll} a(u, v-u) + j(v) - j(u) + b(v-u, p) \geq \langle f, v-u\rangle & \forall v \in \left[\mathbf{H}_0^1(\Omega)\right]^d, \\ b(u, q) = \langle g, q\rangle & \forall q \in \mathbf{L}^2(\Omega), \end{array}\right.$$

where $a(.,.)$ and $b(.,.)$ are two continuous bilinear forms defined on $\left[\mathbf{H}_0^1(\Omega)\right]^d \times \left[\mathbf{H}_0^1(\Omega)\right]^d$ and $\left[\mathbf{H}_0^1(\Omega)\right]^d \times \mathbf{L}^2(\Omega)$, respectively, and $j(.)$ is a real convex Lipschitz-continuous function defined on $\left[\mathbf{H}_0^1(\Omega)\right]^d$. f and g belong to the dual space of

$\left[\mathbf{H}_0^1(\Omega)\right]^d$ and $\mathbf{L}^2(\Omega)$, respectively, and $\langle .,. \rangle$ stands for the duality product. Finally, we suppose that the bilinear form $a(.,.)$ is coercive over $\left[\mathbf{H}_0^1(\Omega)\right]^d$, and we note $|a|$, δ, $|b|$, and, β four positive real usual constants such that

$$\begin{aligned}
&a(u,v) \le |a|\ \|u\|_1\ \|v\|_1 && \forall\ u,v \in \left[\mathbf{H}_0^1(\Omega)\right]^d,\\
&a(u,u) \ge \delta\ \|u\|_1^2 && \forall\ u \in \left[\mathbf{H}_0^1(\Omega)\right]^d,\\
&b(v,q) \le |b|\ \|v\|_1\ \|q\|_0 && \forall\ v \in \left[\mathbf{H}_0^1(\Omega)\right]^d,\ \forall q \in \mathbf{L}^2(\Omega),\\
&j(u) - j(v) \le \beta\ \|u-v\|_1 && \forall\ u,v \in \left[\mathbf{H}_0^1(\Omega)\right]^d.
\end{aligned}$$

We suppose, in addition, that the Babuska–Brezzi condition is satisfied:

$$\exists c > 0 \text{ such that } \quad \forall q \in \mathbf{L}^2(\Omega), \quad \sup_{v \in \left[\mathbf{H}_0^1(\Omega)\right]^d} \frac{b(v,q)}{\|v\|_1} \ge c\|q\|_0.$$

Under the preceding assumptions, the existence of solutions and the uniqueness of u are proven in [11].

Let $\mathbf{U}_h$ and $\mathbf{Q}_h$ be two finite element spaces such that $\mathbf{U}_h \subset \left[\mathbf{H}_0^1(\Omega)\right]^d$, $\mathbf{Q}_h \subset \mathbf{H}^1(\Omega)$, and the following interpolation results are satisfied:

$$\exists c \text{ independent of the discretization step, } h, \text{ such that } \forall\ u \in \left[\mathbf{H}^2(\Omega)\right]^d,$$
$$\exists\ r_h u \in \mathbf{U}_h \text{ satisfying } \quad \left| \begin{array}{l} \|u - r_h u\|_1 \le ch\ |u|_2, \\ \|u - r_h u\|_0 \le ch^2\ |u|_2, \end{array} \right.$$

$$\exists c \text{ independent of } h \text{ such that } \forall p \in \mathbf{H}^1(\Omega),$$
$$\exists\ r_h p \in \mathbf{Q}_h \text{ satisfying } \left| \begin{array}{l} \|p - r_h p\|_0 \le ch\ |p|_1, \\ \|p - r_h p\|_1 \le c\ |p|_1, \end{array} \right.$$

where h stands for the discretization step, with the standard definition.

The numerical scheme considered here reads as follows:

(2.2)
Find $u_h \in \mathbf{U}_h$ and $p_h \in \mathbf{Q}_h$ such that

$$\left| \begin{array}{ll} a(u_h, v_h - u_h) + j(v_h) - j(u_h) + b(v_h - u_h, p_h) \ge \langle f, v_h - u_h \rangle & \forall v_h \in \mathbf{V}_h, \\ -b(u_h, q_h) + \alpha\ c_h(p_h, q_h) = -\langle g, q_h \rangle & \forall q_h \in \mathbf{Q}_h, \end{array} \right.$$

where α is a positive parameter and $c_h(.,.)$ is a mesh-dependent bilinear form such that the following assumptions hold:

(H1) $c_h(p,q)$ is defined for any couple of functions $p,\ q \in \mathbf{H}^1(\Omega)$.

(H2) $[.]_h$ defined by $[q_h]_h^2 = c_h(q_h, q_h)$ is a mesh-dependent norm.

(H3) $\forall p_h,\ q_h \in \mathbf{Q}_h, \quad c_h(p_h, q_h) \le [p_h]_h\ [q_h]_h$.

(H4) $\exists \gamma$, a positive constant independent of h, and $k > 0$ such that
$$\forall v_h \in \mathbf{U}_h, \forall q_h \in \mathbf{Q}_h, \quad b(v_h, q_h) \le \gamma\ \frac{1}{h^k}\ \|v_h\|_0\ [q_h]_h$$
(in further practical applications, $k = \frac{1}{2}$ or 1).

(H5) $\exists c$, a positive constant independent of h, such that
$$\forall q \in \mathbf{H}^1(\Omega), \quad [q]_h \le ch^k\ \|q\|_1.$$

We define as $B(.,.)$ and $B_h(.,.)$, the bilinear forms defined on $(\mathbf{V}_h \times \mathbf{Q}_h) \times (\mathbf{V}_h \times \mathbf{Q}_h)$, by

$$B(u_h, p_h; v_h, q_h) = a(u_h, v_h) + b(v_h, p_h) - b(u_h, q_h),$$
$$B_h(u_h, p_h; v_h, q_h) = B(u_h, p_h; v_h, q_h) + \alpha\ c_h(p_h, q_h).$$

The following stability lemma holds.

LEMMA 2.1. *For any $v_h \in \mathbf{U}_h$ and $q_h \in \mathbf{Q}_h$, we have*

$$B_h(u_h, p_h; u_h, p_h) \geq \delta \|u_h\|_1^2 + \alpha [p_h]_h^2.$$

Proof. This result is a straightforward consequence of the coercivity of the bilinear form $a(.,.)$ in $\left[\mathbf{H}_0^1(\Omega)\right]^d$, combined with the fact that we use a conforming discretization ($\mathbf{U}_h \subset \left[\mathbf{H}_0^1(\Omega)\right]^d$). ∎

COROLLARY 2.2. *There exists a unique solution to problem* (2.2).

Proof. Let $\mathbf{X}_h$ be the product space $\mathbf{V}_h \times \mathbf{Q}_h$, provided with the norms inherited from $\left[\mathbf{H}_0^1(\Omega)\right]^d$ and $\mathbf{L}^2(\Omega)$, respectively. Let $J(.)$ be the convex Lipschitz-continuous functional defined from $\mathbf{X}_h$ to $\mathbb{R}$ by

$$J(v_h, q_h) = j(v_h)$$

and F be the element of the dual of $\mathbf{X}_h$ such that

$$\langle F; (v_h, q_h)\rangle = \langle f, v_h\rangle - \langle g, q_h\rangle.$$

With these notations, problem (2.2) reads as follows.

Find $(u_h, p_h) \in \mathbf{X}_h$such that,$\forall (v_h, q_h) \in \mathbf{X}_h$,

$$B_h(u_h, p_h; (v_h, q_h) - (u_h, p_h)) + J(v_h, q_h) - J(u_h, p_h) \geq \langle F; (v_h, q_h) - (u_h, p_h)\rangle.$$

As on finite dimensional spaces all norms are equivalent, the preceding lemma shows the coercivity of the bilinear form $B_h(.,.)$, and the existence and uniqueness of the solution follow by standard optimization results [8]. ∎

We are now in position to prove the following convergence result.

THEOREM 2.3. *Let u and p be a solution of problem* (2.1) *and u_h and p_h be the solution of* (2.2). *We suppose, in addition, that $u \in \left[\mathbf{H}^2(\Omega)\right]^d$ and $p \in \mathbf{H}^1(\Omega)$. Then the following error bound holds:*

$$\|u - u_h\|_1 \leq c\ h^{\frac{1}{2}}\ (|u|_2 + |p|_1).$$

Proof. Let v_h and q_h be generic elements of $\mathbf{V}_h$ and $\mathbf{Q}_h$, respectively. We suppose that $p \in \mathbf{H}^1(\Omega)$, and consequently, $[p]_h$ is well defined. By the triangular inequality,

$$\begin{aligned} &\delta\ \|u - u_h\|_1^2 + \alpha\ [p - p_h]_h^2 \\ &\qquad \leq \delta\ (\|u - v_h\|_1 + \|u_h - v_h\|_1)^2 + \alpha\ ([p - q_h]_h + [p_h - q_h]_h)^2 \\ &\qquad \leq 2\ [\ \delta \|u - v_h\|_1^2 + \alpha [p - q_h]_h^2 + \underbrace{\delta \|u_h - v_h\|_1^2 + \alpha [p_h - q_h]_h^2}_{\text{(i)}}\ ]. \end{aligned} \tag{2.3}$$

By the stability Lemma 2.1, we obtain the following bound of the last term of this relation:

$$\begin{aligned} \text{(i)} \quad &\leq B_h(u_h - v_h, p_h - q_h; u_h - v_h, p_h - q_h) \\ &= \underbrace{B_h(u_h, p_h; u_h - v_h, p_h - q_h)}_{\text{(ii)}} - B_h(v_h, q_h; u_h - v_h, p_h - q_h). \end{aligned} \tag{2.4}$$

Developing term (ii) of this inequality yields

$$\text{(ii)} = \underbrace{a(u_h, u_h - v_h) + b(u_h - v_h, p_h)}_{\text{(iii)}} - \underbrace{b(u_h, p_h - q_h) + \alpha\ c_h(p_h, p_h - q_h)}_{\text{(iv)}}. \tag{2.5}$$

By the second relation of (2.2) then the second one of (2.1), we get for (iv) the following expression:

$$\text{(iv)} = -\langle \mathrm{g}, \mathrm{p_h} - \mathrm{q_h} \rangle = -\mathrm{b}(\mathrm{u}, \mathrm{p_h} - \mathrm{q_h}). \tag{2.6}$$

Taking $v = u_h$ and then $v = 2u - v_h$ as test functions in the first relation of (2.1) yields

$$\begin{aligned} a(u, u_h - u) + j(u_h) - j(u) + b(u_h - u, p) &\geq \langle f, u_h - u \rangle, \\ a(u, u - v_h) + j(2u - v_h) - j(u) + b(u - v_h, p) &\geq \langle f, u - v_h \rangle. \end{aligned}$$

Summing up these inequalities, we obtain

$$\langle f, u_h - v_h \rangle \leq a(u, u_h - v_h) + j(u_h) - j(u) + j(2u - v_h) - j(u) + b(u_h - v_h, p)$$

and, by the first relation of (2.2),

$$\begin{aligned} \text{(iii)} \quad &\leq \langle f, u_h - v_h \rangle - j(u_h) + j(v_h) \\ &\leq a(u, u_h - v_h) + j(v_h) + j(2u - v_h) - 2j(u) + b(u_h - v_h, p). \end{aligned} \tag{2.7}$$

Using (2.6) and (2.7) in (2.5), we get

$$\text{(ii)} \leq a(u, u_h - v_h) + b(u_h - v_h, p) - b(u, p_h - q_h) + j(v_h) + j(2u - v_h) - 2j(u).$$

And, finally, using this estimate for (ii) in (2.4) and developing the term $B_h(v_h, q_h; u_h - v_h, p_h - q_h)$, we obtain

(2.8)

$$\begin{aligned} \text{(i)} \quad &\leq \quad a(u, u_h - v_h) + b(u_h - v_h, p) - b(u, p_h - q_h) + j(v_h) + j(2u - v_h) - 2j(u) \\ &\quad - [a(v_h, u_h - v_h) + b(u_h - v_h, q_h) - b(v_h, p_h - q_h) + \alpha\ c_h(q_h, p_h - q_h)] \\ &= \quad \underbrace{a(u - v_h, u_h - v_h)}_{\text{(v)}} + \underbrace{b(u_h - v_h, p - q_h)}_{\text{(vi)}} - \underbrace{b(u - v_h, p_h - q_h)}_{\text{(vii)}} \\ &\quad + \underbrace{j(v_h) + j(2u - v_h) - 2j(u)}_{\text{(viii)}} + \underbrace{\alpha\ c_h(p - q_h, p_h - q_h)}_{\text{(ix)}} - \underbrace{\alpha\ c_h(p, p_h - q_h)}_{\text{(x)}}. \end{aligned}$$

Tracing back the origin of the last term (x) of the preceding relation, it may be checked that it appears because $B_h(u_h, p_h; u_h - v_h, p_h - q_h)$ is used as a discrete counterpart of $B(u, p; u_h - v_h, p_h - q_h)$, which cannot be identified, under regularity assumptions on u and p, to $B_h(u, p; u_h - v_h, p_h - q_h)$, as would be the case for a consistently stabilized scheme. It thus can be viewed as the consistency error of the scheme.

The next step consists of bounding each of the terms of the relation (2.8):

$$
\begin{aligned}
\text{(v)} \quad &\leq |a|\ \|u-v_h\|_1\ \|u_h-v_h\|_1 \leq \frac{|a|^2}{2\delta c_1}\ \|u-v_h\|_1^2 + \frac{c_1\delta}{2}\ \|u_h-v_h\|_1^2,\\
\text{(vi)} \quad &\leq |b|\ \|u_h-v_h\|_1\ \|p-q_h\|_0 \leq \frac{|b|^2}{2\delta c_2}\ \|p-q_h\|_0^2 + \frac{c_2\delta}{2}\ \|u_h-v_h\|_1^2,\\
\text{(vii)} \quad &\leq \frac{\gamma}{h^k}\ \|u-v_h\|_0\ [p_h-q_h]_h \qquad \text{(by assumption (H4))}\\
&\leq \frac{\gamma^2}{2\alpha c_3 h^{2k}}\ \|u-v_h\|_0^2 + \frac{c_3\alpha}{2}\ [p_h-q_h]_h^2,\\
\text{(viii)} \quad &= (j(2u-v_h)-j(u))-(j(u)-j(v_h))\\
&\leq 2\beta\|u-v_h\|_1 \qquad \text{(by the Lipschitz continuity of } j(.)),\\
\text{(ix)} \quad &\leq \alpha\ [q_h-p]_h\ [p_h-q_h]_h \qquad \text{(by assumption (H3))}\\
&\leq \frac{\alpha}{2c_4}\ [q_h-p]_h^2 + \frac{\alpha c_4}{2}\ [p_h-q_h]_h^2,\\
\text{(x)} \quad &\leq \frac{\alpha}{2c_5}\ [p]_h^2 + \frac{\alpha c_5}{2}\ [p_h-q_h]_h^2 \qquad \text{(idem).}
\end{aligned}
$$

Choosing $c_1 = c_2 = 1/2$, $c_3 = c_4 = c_5 = 1/3$, substituting these expressions for (v)–(x) in 2.8, and combining terms on the left-hand side, we get

$$
\begin{aligned}
&\frac{1}{2}\ \left[\delta\ \|u_h-v_h\|_1^2 + \alpha[p_h-q_h]_h^2\right]\\
&\quad\leq \frac{|a|^2}{\delta}\ \|u-v_h\|_1^2 + \frac{3\gamma^2}{2\alpha h^{2k}}\ \|u-v_h\|_0^2 + 2\beta\ \|u-v_h\|_1 + \frac{|b|^2}{\delta}\ \|p-q_h\|_0^2\\
&\qquad + \frac{3\alpha}{2}\ [p-q_h]_h^2 + \frac{3\alpha}{2}\ [p]_h^2.
\end{aligned}
$$

Combined with the initial triangular inequality (2.3), this last relation yields

$$
\begin{aligned}
&\delta\ \|u-u_h\|_1^2 + \alpha[p-p_h]_h^2\\
&\leq \inf_{v_h\in\mathbf{V}_h}\left[\left(\frac{4|a|^2}{\delta}+2\delta\right)\ \|u-v_h\|_1^2 + \frac{6\gamma^2}{\alpha h^{2k}}\ \|u-v_h\|_0^2 + 8\beta\ \|u-v_h\|_1\right]\\
&+ \inf_{q_h\in\mathbf{Q}_h}\left[\frac{4|b|^2}{\delta}\ \|p-q_h\|_0^2 + 7\alpha\ [p-q_h]_h^2\right] + 6\alpha\ [p]_h^2.
\end{aligned}
\tag{2.9}
$$

Associated with the assumed approximation properties of the spaces $\mathbf{V}_h$ and $\mathbf{Q}_h$ and hypothesis (H5), this last inequality completes the proof. □

Remark. Due to the fact that the seminorm $[.]_h$ behaves like $h^k\ \|.\|_1$, the presence of the term $[p-q_h]_h$ on the left-hand side of inequality (2.9) does not provide any convergence result for the approximation of p. This was expected: p is not even guaranteed to be unique. However, we obtain the following weaker result: as soon as $k \leq \frac{1}{2}$, the solution of the numerical scheme p_h remains bounded independently of the discretization step.

The sharpest result for the convergence of u_h toward the solution u is obtained with $k \geq 1$. In this case, the leading-order term on the right-hand side of inequality 2.9 is $\beta\ \|u-v_h\|_1$ and stems from the nonlinearity of the problem.

3. A numerical scheme for Bingham fluid creeping flows. The so-called Bingham fluids belong to the category of fluids exhibiting a yield stress, i.e., which behave as a solid in regions of the flow where the shear stress magnitude falls under a threshold value. Their constitutive law reads as

$$\begin{cases} \text{if } \|\tau\| \leq \tau_{\text{YS}}, \quad \dot{\varepsilon}(v) = 0, \\ \text{if } \|\tau\| \geq \tau_{\text{YS}}, \quad \tau = \left(\dfrac{\tau_{\text{YS}}}{\|\dot{\varepsilon}\|} + 2\mu \right) \dot{\varepsilon}(v), \end{cases}$$

where v stands for the fluid velocity, $\dot{\varepsilon}(.)$ for the usual strain rate tensor, τ for the shear stress tensor, μ for the dynamic viscosity, τ_{YS} for the fluid yield stress, and $\|.\|$ is the Euclidean norm in $\mathbb{R}^{d\times d}$.

The creeping flow of a Bingham fluid in a domain Ω of $\mathbb{R}^d$ with adherence conditions at the boundaries is governed by a set of balance equations that admit a variational formulation of general form (2.1) [6, Chapter 6], with the following specific expressions for each bilinear and linear form of the problem,

$$a(u,v) = \int_\Omega 2\mu \ \dot{\varepsilon}(u) : \dot{\varepsilon}(v), \qquad b(v,p) = -\int_\Omega p \ \nabla \cdot u, \qquad \langle f, v\rangle = \int_\Omega f.v,$$

and for the Lipschitz-continuous functional,

$$j(v) = \int_\Omega \tau_{\text{YS}} \ \|\dot{\varepsilon}(v)\|,$$

where f stands for the volume forces. The assumptions of continuity and coercivity of $a(.,.)$, continuity of $b(.,.)$, and the Babuska–Brezzi condition are standard results of theoretical computational fluid dynamics. The Lipschitz–continuity of $j(.)$ is a consequence of the Lipschitz–continuity of the Euclidean norm.

We suppose a given family of triangulations of the domain, and we choose as approximation space the standard Lagrange linear continuous finite element space for both the velocity and the pressure. For a particular triangulation T_h of n d-simplexes $(K_i)_{i\leq n}$, the finite element subspaces are given by

$$\mathbf{U}_h = \{v_h \in [C^1(\bar{\Omega})]^d, \quad v_h|_{K_i} \in [\mathbb{P}_1(K_i)]^d, \quad 1 \leq i \leq n, \quad \text{and } v_h = 0 \text{ on } \partial\Omega\},$$
$$\mathbf{Q}_h = \{q_h \in C^1(\bar{\Omega}), \quad q_h|_{K_i} \in \mathbb{P}_1(K_i), \quad 1 \leq i \leq n\},$$

where $\mathbb{P}_1(K_i)$ stands for the space of degree ≤ 1 polynomials over the polyhedron K_i.

These finite element spaces verify the regularity and approximation properties used in the preceding section, provided that the meshing is regular (i.e., the ratio between the diameter of the largest inscribed ball and the mesh diameter is bounded away from zero for each element of the family of triangulations [5]).

The numerical scheme considered here uses the following stabilization bilinear form:

$$c_h(p_h, q_h) = c_{\text{b}} \sum_{i=1}^{n} h_{K_i}^2 \int_{K_i} \nabla p_h . \nabla q_h.$$

With this definition, the assumptions (H1), (H2), (H3), and (H5) are easily checked (with the usual restriction to zero mean value functions for the pressure space). The following lemma means that the hypothesis (H4) stands, under an additional constraint for the family of triangulations.

LEMMA 3.1. *We assume that the family of triangulations is quasi-uniform, i.e., that the ratio of the smallest element diameter to the largest one is bounded away from zero. Then there exists a positive constant c independent of h such that, for any* $v_h \in \mathbf{U}_h$ *and* $q_h \in \mathbf{Q}_h$*, we have*

$$|b(v_h, q_h)| \leq c\ \frac{1}{h}\ \|v_h\|_0\ [q_h]_h.$$

Proof. From one part, we have

$$[q_h]_h^2 = c_{\mathrm{b}}\ \sum_{i=1}^{n}\ h_{K_i}^2\ \int_{K_i} \nabla\, q_h.\nabla\, q_h \quad \geq c_{\mathrm{b}}\ \left(\min_{0\leq i\leq n} h_{K_i}\right)^2\ \sum_{i=1}^{n}\ \int_{K_i} \nabla\, q_h.\nabla\, q_h$$
$$\geq c_{\mathrm{b}}\ \left(\frac{\min_{0\leq i\leq n} h_{K_i}}{h}\right)^2\ h^2\ |q_h|_1^2.$$

The first factor of this last expression is bounded away from zero if the family of triangulations is quasi-uniform.

On the other hand, as $\mathbf{Q}_h \subset \mathbf{H}^1(\Omega)$ and $\mathbf{V}_h \subset \left[\mathbf{H}_0^1(\Omega)\right]^d$, the following integration by parts is valid:

$$b(v_h, q_h) = -\int_\Omega q_h \nabla \cdot v_h = \int_\Omega \nabla\, q_h.v_h$$

and, by the Cauchy–Schwarz inequality,

$$b(v_h, q_h) \leq \|v_h\|_0 |q_h|_1.$$

The result is obtained by combining both inequalities. ∎

As a consequence, we obtain the following error estimate for the numerical scheme under consideration.

THEOREM 3.2. *Let u be the solution of the problem. Let* u_h *be the generic element of a family of approximate solutions obtained with the present scheme using a family of regular and quasi-uniform triangulations. If we assume that* $u \in \left[\mathbf{H}^2(\Omega)\right]^d$*, the following error bound holds:*

$$\|u - u_h\|_1 \leq c\ h^{1/2}\ |u|_2,$$

where the positive constant c is independent of h.

Remark. The extension of this analysis to problems with nonhomogeneous Dirichlet boundary conditions can be performed as for the Newtonian Stokes problem, without any additional difficulty (e.g., [15]).

4. Practical implementation: The decomposition-coordination method. The object of this section is first to adapt the decomposition-coordination method of Fortin and Glowinski [8] to the numerical scheme under consideration, then to describe an algorithm for solving the discrete problem and to analyze its convergence. The difficulty of the first task lies in the fact that, due to the presence of the stabilization term, the standard theory does not apply to the problem under consideration, and the equivalence (in a sense to be defined) between the final discrete system and the initial one has to be proven.

Our starting point is the approximate problem:

Find $u_h \in \mathbf{U}_h$ and $p_h \in \mathbf{Q}_h$ such that

$$
\left|
\begin{aligned}
&\int_\Omega 2\mu\ \dot{\varepsilon}(u_h) : \dot{\varepsilon}(v_h - u_h) - \int_\Omega p_h\ \nabla \cdot (v_h - u_h) \\
&\qquad + \int_\Omega \tau_{\text{YS}}\ \|\dot{\varepsilon}(v_h)\| - \int_\Omega \tau_{\text{YS}}\ \|\dot{\varepsilon}(u_h)\| \geq \int_\Omega f.(v_h - u_h) \qquad \forall v_h \in \mathbf{U}_h, \\
&\int_\Omega p_h\ \nabla \cdot u_h + c_{\text{b}} \sum_{i=1}^{n} h_{K_i}^2 \int_{K_i} \nabla p_h . \nabla q_h = 0 \qquad \forall q_h \in \mathbf{Q}_h.
\end{aligned}
\right. \tag{4.1}
$$

Let $\mathbf{W}_h$ be the following finite element space:

$$\mathbf{W}_h = \{z_h \in [\mathbf{L}^2(\Omega)]^{d\times d}, \quad z_h|_{K_i} \in [\mathbb{P}_0(K_i)]^{d\times d}, \quad 1 \leq i \leq n\},$$

where $\mathbb{P}_0(K_i)$ stands for the space of constant functions over polyhedron K_i.

We introduce the following discrete variational problem:

Find $u_h \in \mathbf{U}_h,\ p_h \in \mathbf{Q}_h,\ w_h \in \mathbf{W}_h,$ and $s_h \in \mathbf{W}_h$ such that

$$
\left|
\begin{aligned}
&\int_\Omega 2\mu\ w_h : (z_h - w_h) - \int_\Omega (z_h - w_h) : s_h \\
&\qquad + \int_\Omega \tau_{\text{YS}}\ \|z_h\| - \int_\Omega \tau_{\text{YS}}\ \|w_h\| \\
&\qquad + \sum_{i=1}^{n} r_i \int_{K_i} [w_h - \dot{\varepsilon}(u_h)] : [z_h - w_h] \geq 0 \qquad \forall z_h \in \mathbf{W}_h, \\
&\sum_{i=1}^{n} r_i \int_{K_i} [\dot{\varepsilon}(u_h) - w_h] : \dot{\varepsilon}(v_h) - \int_\Omega p_h\ \nabla \cdot v_h \\
&\qquad + \int_\Omega \dot{\varepsilon}(v_h) : s_h = \int_\Omega f.v_h \qquad \forall v_h \in \mathbf{U}_h, \\
&\int_\Omega q_h\ \nabla \cdot u_h + c_{\text{b}} \sum_{i=1}^{n} h_{K_i}^2 \int_{K_i} \nabla p_h . \nabla q_h = 0 \qquad \forall q_h \in \mathbf{Q}_h, \\
&\int_\Omega (\dot{\varepsilon}(u_h) - w_h) : t_h = 0 \qquad \forall t_h \in \mathbf{W}_h.
\end{aligned}
\right. \tag{4.2}
$$

The family of parameters r_i is chosen such that $0 < r_i$ for each $1 \geq i \geq n$.

We then have the following result.

PROPOSITION 4.1. *Problem* (4.2) *has at least one solution; in addition, the first three components u_h, w_h, p_h are the same for any solution of the problem.*

Moreover, if u_h and p_h are (part of) one solution of (4.2)*, u_h and p_h are solutions of* (4.1).

Proof. Let $A(.,.)$ be the bilinear form defined over $(\mathbf{U}_h \times \mathbf{W}_h \times \mathbf{Q}_h) \times (\mathbf{U}_h \times \mathbf{W}_h \times \mathbf{Q}_h)$, $B(.,.)$ be the bilinear form defined over $(\mathbf{U}_h \times \mathbf{W}_h \times \mathbf{Q}_h) \times \mathbf{W}_h$, and $J(.)$ and $F(.)$ be, respectively, the form and the nonlinear convex functional defined over

$\mathbf{U}_h \times \mathbf{W}_h \times \mathbf{Q}_h$ as follows:

$$\begin{aligned} A((u_h,w_h,p_h);(v_h,z_h,q_h)) &= \int_\Omega 2\mu\ w_h : z_h + \sum_{i=1}^{n} r_i \int_{K_i} [w_h - \dot{\varepsilon}(u_h)] : z_h \\ &+ \sum_{i=1}^{n} r_i \int_{K_i} [\dot{\varepsilon}(u_h) - w_h] : \dot{\varepsilon}(v_h) - \int_\Omega p_h\ \nabla\cdot v_h + \int_\Omega q_h\ \nabla\cdot u_h \\ &+ c_{\mathrm{b}} \sum_{i=1}^{n} h_{K_i}^2 \int_{K_i} \nabla p_h . \nabla q_h, \end{aligned}$$

$$B((v_h,z_h,q_h);t_h) = \int_\Omega (\dot{\varepsilon}(u_h) - w_h) : t_h,$$

$$J(v_h,z_h,q_h) = j(z_h),$$

$$F(v_h,z_h,q_h) = \int_\Omega f.v_h.$$

With these notations, problem (4.2) can be recast under the form of the following nonlinear mixed variational inequality:

Find $(u_h,w_h,p_h) \in (\mathbf{U}_h \times \mathbf{W}_h \times \mathbf{Q}_h)$ and $s_h \in \mathbf{W}_h$ such that

$$\left|\begin{array}{ll} A((u_h,w_h,p_h);(v_h,z_h,q_h)-(u_h,w_h,p_h)) + J(v_h,z_h,q_h) - J(u_h,w_h,p_h) & \\ \qquad + B((v_h,z_h,q_h);s_h) \geq F((v_h,z_h,q_h)-(u_h,w_h,p_h)) & \\ \qquad \forall (v_h,z_h,q_h) \in (\mathbf{U}_h \times \mathbf{W}_h \times \mathbf{Q}_h), & \\ B((u_h,w_h,p_h);t_h) = 0 & \forall t_h \in \mathbf{W}_h. \end{array}\right.$$

We find the following result in [11]: assuming that $A(.,.)$ is coercive, the LBB condition holds for $B(.,.)$, and $J(.)$ is Lipschitz-continuous and convex, this variational problem has solutions; moreover, the primal component of the solution (u_h,w_h,p_h) is unique.

Checking the LBB condition is straightforward here. We are going to prove the coercivity of the bilinear form $A(.,.)$:

$$\begin{aligned} A((u_h,w_h,p_h);(u_h,w_h,p_h)) &= \int_\Omega 2\mu\ w_h : w_h + c_{\mathrm{b}} \sum_{i=1}^{n} h_{K_i}^2 \int_{K_i} \nabla p_h . \nabla p_h \\ &+ \sum_{i=1}^{n} r_i \int_{K_i} [w_h - \dot{\varepsilon}(u_h)] : [w_h - \dot{\varepsilon}(u_h)]. \end{aligned}$$

Developing and using the Cauchy–Schwarz inequality yields, for any strictly positive constant α,

$$\begin{aligned} &A((u_h,w_h,p_h);(u_h,w_h,p_h)) \\ &\quad = 2\mu\|w_h\|_0^2 + c_{\mathrm{b}}[p_h]_h^2 + \sum_{i=1}^{n} r_i \left[\|w_h\|_{0,\mathrm{K}}^2 + \|\dot{\varepsilon}(u_h)\|_{0,\mathrm{K}}^2 + \int_{K_i} w_h : \dot{\varepsilon}(u_h) \right] \\ &\quad \geq 2\mu\|w_h\|_0^2 + c_{\mathrm{b}}[p_h]_h^2 + \sum_{i=1}^{n} r_i \left[(1-\alpha)\ \|w_h\|_{0,\mathrm{K}}^2 + \left(1 - \frac{1}{\alpha}\right)\ \|\dot{\varepsilon}(u_h)\|_{0,\mathrm{K}}^2 \right]. \end{aligned}$$

In finite dimensional spaces, all norms are equivalent; the coercivity of $A(.,.)$ then follows by choosing $\alpha > 1$ such that $(\alpha - 1)r_i \leq \mu$ for all $1 \leq i \leq n$.

Due to the choice of the discrete spaces, the constraint $\dot{\varepsilon}(u_h) = w_h$ holds exactly. As a consequence, the first equation of (4.1) is recovered by summing up the first two equations of (4.2) and choosing $z_h = \dot{\varepsilon}(v_h)$ as a test function. As the second equation of (4.1) remains valid, the solution of (4.1) and the first and third components of the solution of (4.2) are the same. ∎

Replacing the last third relations with their algebraic counterpart, problem (4.2) reads as follows:

(i) Find $w_h \in \mathbf{W}_h$ such that

$$\int_\Omega 2\mu\ w_h : (z_h - w_h) - \int_\Omega (z_h - w_h) : s_h + \int_\Omega \tau_{\text{YS}}\ \|z_h\| - \int_\Omega \tau_{\text{YS}}\ \|w_h\| + \sum_{i=1}^{n} r_i \int_{K_i} [w_h - \dot{\varepsilon}(u_h)] : [z_h - w_h] \geq 0 \qquad \forall z_h \in \mathbf{W}_h,$$

(ii) $\mathrm{A}\,\mathbf{u}_h - \mathrm{D}\,\mathbf{w}_h - \mathrm{B}^{\,t}\mathbf{p}_h + \mathrm{D}\,\mathbf{s}_h = \mathbf{f}$,

(iii) $\mathrm{B}\,\mathbf{u}_h + \mathrm{C}\,\mathbf{p}_h = 0$,

(iv) $\mathrm{D}\,\mathbf{u}_h - \mathrm{E}_w\mathbf{w}_h = 0$,

where, in the last three relations, the expressions typed in boldface stand for usual finite element vectors of degrees of freedom and the discrete operators are obtained by using the standard finite element process. With the particular discrete spaces used here, a suitable choice of the parameters r_i leads to $\mathrm{A} = r_1 \mathrm{D}\,\mathrm{E}_w^{-1}\mathrm{D}$, where r_1 is a positive augmentation parameter and E_w is the strain rates mass matrix.

Finally, we simulate an augmentation relative to the divergence constraint by premultiplying the equation (iii) by $r_2 \mathrm{B}^{\,t}\mathrm{E}_p^{-1}$, with E_p the lumped pressure mass matrix, and adding the obtained relation to equation (ii). The final system reads as follows:

(4.3)

(i) Find $w_h \in \mathbf{W}_h$ such that

$$\int_\Omega 2\mu\ w_h : (z_h - w_h) - \int_\Omega (z_h - w_h) : s_h + \int_\Omega \tau_{\text{YS}}\ \|z_h\| - \int_\Omega \tau_{\text{YS}}\ \|w_h\| + \sum_{i=1}^{n} r_i \int_{K_i} [w_h - \dot{\varepsilon}(u_h)] : [z_h - w_h] \geq 0 \qquad \forall z_h \in \mathbf{W}_h,$$

(ii) $(r_1 \mathrm{D}\,\mathrm{E}_w^{-1}\mathrm{D} + r_2 \mathrm{B}^{\,t}\mathrm{E}_p^{-1}\mathrm{B})\mathbf{u}_h - \mathrm{C}\,\mathbf{w}_h + (r_2 \mathrm{B}^{\,t}\mathrm{E}_p^{-1}\mathrm{C} - \mathrm{B}^{\,t})\mathbf{p}_h + \mathrm{D}\,\mathbf{s}_h = \mathbf{f}$,

(iii) $\mathrm{B}\,\mathbf{u}_h + \mathrm{C}\,\mathbf{p}_h = 0$,

(iv) $\mathrm{D}\,\mathbf{u}_h - \mathrm{E}_w\mathbf{w}_h = 0$.

This last algebraic manipulation does not change the properties of the system.

Thanks to the particular structure of the approximation space $\mathbf{W}_h$, the variational inequality (i) of (4.3) degenerates to a family of scalar minimization problems that

admit an explicit solution for $\mathbf{w}_h$ as a function of the other unknowns $\mathbf{u}_h$ and $\mathbf{s}_h$:

$$\text{For } 1 \le i \le n$$

$$\left| \begin{array}{ll} \text{if } \|\sigma_i\| < \tau_{\text{YS}}, & \mathbf{w}_{hi} = 0, \\ \text{if } \|\sigma_i\| < \tau_{\text{YS}}, & \mathbf{w}_{hi} = \dfrac{1 - \tau_{\text{YS}}/\|\sigma_i\|}{2(\mu + r_1/2)}\ \sigma_i, \end{array} \right. \tag{4.4}$$

$$\text{where } \sigma_i = \mathbf{s}_{hi} + \frac{r_1}{meas(K_i)} \int_{K_i} \dot{\varepsilon}(\mathbf{u}_h).$$

The nonlinear system is solved by an Uzawa-like algorithm that reads as follows:

$\mathbf{u}_h^{n-1}, \mathbf{w}_h^{n-1}, \mathbf{p}_h^{n-1}, \mathbf{s}_h^{n-1}$ being known,

(1) compute $\mathbf{u}_h^n$ as a function of $\mathbf{w}_h^{n-1}, \mathbf{p}_h^{n-1}, \mathbf{s}_h^{n-1}$ by (ii),

(2) compute $\mathbf{w}_h^n$ as a function of $\mathbf{u}_h^n, \mathbf{s}_h^{n-1}$ by (4.4),

(3) compute $\mathbf{p}_h^n$ and $\mathbf{s}_h^n$ by

$$\left| \begin{array}{l} \left(\dfrac{1}{\rho_2}\mathrm{E}_p + \mathrm{C}\right)\ \mathbf{p}_h^n = \dfrac{1}{\rho_2}\mathrm{E}_p\ \mathbf{p}_h^{n-1} - \mathrm{B}\,\mathbf{u}_h^n, \\ \dfrac{1}{\rho_1}\mathrm{E}_w\ \mathbf{s}_h^n = \dfrac{1}{\rho_1}\mathrm{E}_w\ \mathbf{s}_h^{n-1} - (\mathrm{D}\,\mathbf{u}_h^n - \mathrm{E}_w \mathbf{w}_h^n). \end{array} \right. \tag{4.5}$$

We have the following convergence results.

THEOREM 4.2. *Algorithm* (4.5) *converges to a solution of the system* (4.3)*, provided the following condition holds:*

$$\rho_1 < \frac{1+\sqrt{5}}{2}\ r_1 \qquad \textit{and} \qquad \rho_2 < 2r_2.$$

Proof. The proof of this theorem is obtained by minor modifications of the convergence study of the standard algorithm (i.e., without regularization term) that can be found in [3]. This development closely follows the seminal work of Fortin and Glowinski [8], [10]. □

5. Numerical experiment. In this section we are interested in the numerical validation of the proven error estimates against a particular problem that admits an explicit solution. We will see that, under particular regularity assumptions verified by the problem being considered, the convergence proof of section 2 can yield an improvement of the convergence rate to $h|\log(h)|^{1/2}$; this error bound is then confirmed by numerical experiments.

5.1. Position of the test problem. To our knowledge, no problem admitting an analytic solution and set on a general multidimensional polygonal domain is found in the literature. As an alternative, we consider an axisymmetrical problem, treated in the following as a fully bidimensional one: the tangential flow of a Bingham fluid in a viscosimeter made of two coaxial cylinders. The inner cylinder of radius $r_{inn} = 0.5$ is kept fixed, whereas a constant angular velocity $W = 1$ is imposed on the outer cylinder (radius $r_{out} = 1$). Finally, the fluid is assumed to stick to the apparatus boundaries, and we recall here the problem solution [2, section 4.5]. As the plasticity

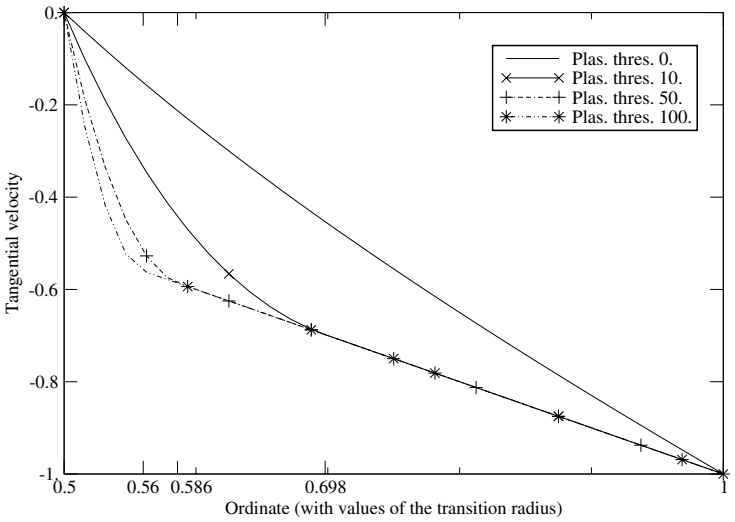


FIG. 5.1. *Tangential velocity at* $(x = 0, y \in [0.5; 1])$*, analytical solution.*

threshold increases, a rigid zone appears near the outer cylinder. The transition radius r_{tra} between the "flow" region and the "rigid" zone obeys the nonlinear equation

$$(2r_{tra})^2 - 2\ln(2r_{tra}) - 2\sqrt{2}\frac{\mu}{\tau_{\text{YS}}} = 1, \tag{5.1}$$

and the following tangential velocity and pressure are solution to the problem:

$$\begin{aligned}
&\text{if } r \geq r_{tra}, && v_\theta(r) = r;\\
&\text{if } r < r_{tra}, && v_\theta(r) = r\left[1 + \frac{\sqrt{2}\tau}{2\mu}\left(\frac{1}{2} - \frac{1}{2}\left(\frac{r_{tra}}{r}\right)^2 + ln\left(\frac{r_{tra}}{r}\right)\right)\right];\\
&r_{inn} \leq r \leq r_{out}, && p = 0.
\end{aligned}$$

This analytical solution has been plotted for various plasticity threshold values in Figure 5.1.

5.2. An improved error estimate for the particular problem at hand. We begin with a technical lemma.

LEMMA 5.1. *Let* u *and* v *be two nonzero vectors of* $\mathbb{R}^d$*. The following relation holds:*

$$\|v\| - \|u\| = \frac{u}{\|u\|}.(v-u) + \frac{\|v-u\|^2 - \left[\frac{u}{\|u\|}.(v-u)\right]^2}{\|v\| + \frac{u}{\|u\|}.v}.$$

The solution u of the problem under consideration is continuous and belongs to the Sobolev space $[\mathbf{W}^{2,\infty}(\Omega)]^d$ of fields whose components' first and second derivatives are essentially bounded. As a consequence, the Lagrangian interpolate of u, $r_h u$ is well defined and the following estimates hold:

$$\|u - r_h u\|_{1,\infty} \leq c\ h\ |u|_{2,\infty}, \tag{5.2}$$

where $c_{\rm a}$ is a positive constant independent of the discretization step h.

For the particular problem at hand, we will prove the following improved error estimate.

PROPOSITION 5.2. *For a sufficiently small discretization step h, the numerical solution of the problem is such that*

$$\|u - u_h\|_1 \leq c\ \tau_{\rm YS}^{1/2}\ h\ |\log(h)|^{1/2},$$

where c is a positive constant dependent on the solution u but independent of h.

Proof. To obtain the preceding estimate, it is sufficient to prove that, for h small enough, the following relation holds:

$$j(2u - r_h u) + j(r_h u) - 2j(u) \leq c\ \tau_{\rm YS}\ h^2\ |\log(h)|.$$

By definition, the left-hand member of this inequality reads as

$$(j(2u - r_h u) - j(u)) - (j(u) - j(r_h u)) = \tau_{\rm YS} \int_\Omega \|\dot\varepsilon(2u - r_h u)\| + \|\dot\varepsilon(r_h u)\| - 2\ \|\dot\varepsilon(u)\|.$$

We split this integral into three parts:

$$\begin{aligned} & 2\pi \underbrace{\int_{r_{inn}}^{r_{tra}-\gamma h} [\ \|\dot\varepsilon(2u - r_h u)\| + \|\dot\varepsilon(r_h u)\| - 2\ \|\dot\varepsilon(u)\|\ ]\ r\ dr}_{\rm (i)} \\ + \quad & 2\pi \underbrace{\int_{r_{tra}-\gamma h}^{r_{tra}+h} [\ \|\dot\varepsilon(2u - r_h u)\| + \|\dot\varepsilon(r_h u)\| - 2\ \|\dot\varepsilon(u)\|\ ]\ r\ dr}_{\rm (ii)} \\ + \quad & 2\pi \underbrace{\int_{r_{tra}+h}^{r_{out}} [\ \|\dot\varepsilon(2u - r_h u)\| + \|\dot\varepsilon(r_h u)\| - 2\ \|\dot\varepsilon(u)\|\ ]\ r\ dr}_{\rm (iii)}, \end{aligned}$$

where γ is a positive parameter independent of h to be chosen later.

The three vertices of each element that intersects the integration domain of (iii) lie in the rigid zone. As the solution in this zone is linear, u and $r_h u$ are equal and $\dot\varepsilon(r_h u)$ also vanishes with $\dot\varepsilon(u)$. As a consequence, (iii) $= 0$.

By the triangular inequality for the norm in $\mathbb{R}^{d\times d}$,

$$\begin{aligned} & \|\dot\varepsilon(2u - r_h u)\| + \|\dot\varepsilon(r_h u)\| - 2\ \|\dot\varepsilon(u)\| \\ & \qquad = (\|\dot\varepsilon(2u - r_h u)\| - \|\dot\varepsilon(u)\|) - (\|\dot\varepsilon(u)\| - \|\dot\varepsilon(r_h u)\|) \\ & \qquad \leq \|\dot\varepsilon(2u - r_h u) - \dot\varepsilon(u)\| + \|\dot\varepsilon(u) - \dot\varepsilon(r_h u)\| \\ & \qquad = 2\ \|\dot\varepsilon(u) - \dot\varepsilon(r_h u)\|. \end{aligned}$$

As a consequence of the approximation inequality (5.2), the following pointwise estimate holds:

$$\|\dot\varepsilon(u) - \dot\varepsilon(r_h u)\|_{0,\infty} \leq c_{\rm a}\ h\ |u|_{2,\infty}$$

and

$$\text{(ii)} \leq c\ h|u|_{2,\infty} \int_{r_{tra}-\gamma h}^{r_{tra}+h} r dr \leq c\ |u|_{2,\infty}\ (r_{tra} + h)\ (\gamma + 1)\ h^2.$$

For $r \leq r_{tra}$, the solution u is such that

$$\|\dot{\varepsilon}(u)\| \geq c_{\mathrm{e}}\ (r_{tra} - r) \qquad \text{with} \quad c_{\mathrm{e}} = \max\left(1, \frac{\sqrt{2}\ \tau_{\mathrm{YS}}}{\mu}\right)\ \frac{r_{tra} + r_{inn}}{r_{tra}^2}. \tag{5.3}$$

As a consequence of the approximation inequality (5.2) and the triangular inequality for the Euclidean norm of $\mathbb{R}^{d\times d}$, for a value of the parameter γ such that $c_{\mathrm{e}}\gamma > c_{\mathrm{a}}\ |u|_{2,\infty}$,

$$\left.\begin{array}{l} \|\dot{\varepsilon}(r_h u)\| > 0 \\ \|\dot{\varepsilon}(2u - r_h u)\| > 0 \end{array}\right| \quad \forall r \in [r_{inn}, r_{tra}].$$

We choose the following value for γ:

$$\gamma = 2\ \frac{c_{\mathrm{a}}\ |u|_{2,\infty}}{c_{\mathrm{e}}}. \tag{5.4}$$

This choice is possible for a value of h such that $\gamma h < r_{tra} - r_{inn}$. Lemma 5.1 then applies, and

$$\begin{aligned} \text{(i)} \quad &= \int_{r_{inn}}^{r_{tra}-\gamma h} [(\|\dot{\varepsilon}(2u - r_h u)\| - \|\dot{\varepsilon}(u)\|) - (\|\dot{\varepsilon}(u)\| - \|\dot{\varepsilon}(r_h u)\|)]\ r\ dr \\ &= \int_{r_{inn}}^{r_{tra}-\gamma h} \left[\left(\frac{\dot{\varepsilon}(u)}{\|\dot{\varepsilon}(u)\|} : \dot{\varepsilon}(u - r_h u) + \frac{N(u, r_h u)}{D_1(u, r_h u)}\right)\right. \\ &\qquad \left. - \left(\frac{\dot{\varepsilon}(u)}{\|\dot{\varepsilon}(u)\|} : \dot{\varepsilon}(u - r_h u) + \frac{N(u, r_h u)}{D_2(u, r_h u)}\right)\right]\ r\ dr, \end{aligned}$$

where

$$N(u, r_h u) = \|\dot{\varepsilon}(u) - \dot{\varepsilon}(r_h u)\|^2 + \left[\frac{\dot{\varepsilon}(u)}{\|\dot{\varepsilon}(u)\|} : (\dot{\varepsilon}(u) - \dot{\varepsilon}(r_h u))\right]^2,$$

$$D_1(u, r_h u) = \|\dot{\varepsilon}(2u) - \dot{\varepsilon}(r_h u)\| + \left[\frac{\dot{\varepsilon}(u)}{\|\dot{\varepsilon}(u)\|} : [\dot{\varepsilon}(2u) - \dot{\varepsilon}(r_h u)\right],$$

$$D_2(u, r_h u) = \|\dot{\varepsilon}(r_h u)\| + \frac{\dot{\varepsilon}(u)}{\|\dot{\varepsilon}(u)\|} : \dot{\varepsilon}(r_h u).$$

By the approximation inequality (5.2), we then get

$$|N(u, r_h u)| \leq c\ h^2\ |u|_{2,\infty} \qquad \forall r \in [r_{inn}, r_{tra} - \gamma h]$$

and, by the triangular inequality for the Euclidean norm of $\mathbb{R}^{d\times d}$,

$$|D_1(u, r_h u) - 2\ \|\dot{\varepsilon}(u)\|| \leq 2\ \|\dot{\varepsilon}(u) - \dot{\varepsilon}(r_h u)\|.$$

Inequality (5.3) then implies, with the particular choice for γ given by (5.4),

$$|D_1(u, r_h u)| < 2c_{\mathrm{e}}\ \left(\left(r_{tra} - \frac{\gamma}{2}\ h\right) - r\right) \qquad \forall r \in [r_{inn}, r_{tra} - \gamma h].$$

Consequently,

$$\begin{aligned} \int_{r_{inn}}^{r_{tra}-\gamma h} \frac{N(u, r_h u)}{D_1(u, r_h u)}\ r\ dr &\leq c\ h^2\ |u|_{2,\infty} \int_{r_{inn}}^{r_{tra}-\gamma h} \frac{1}{(r_{tra} - \frac{\gamma}{2}\ h) - r}\ r\ dr \\ &\leq c\ h^2\ \log\left(\frac{\gamma}{2}\ h\right). \end{aligned}$$

A similar estimate of the second part of (i) follows, applying the same arguments. □

Remark. The regularity of the solution of the problem under consideration is far beyond what can be expected in the general case. However, arguments similar to those employed in the preceding proof may extend to a wide range of practical situations, where, in particular, regularity of the rigid zones boundaries, often suggested a posteriori by numerical results, can be conjectured.

5.3. Numerical tests. The whole domain has been meshed with triangles; see, for example, a coarse mesh in Figure 5.2. Each mesh is used to build a finer one by cutting each of its right triangles into four right triangles of equal size. In Figure 5.3, we present the velocity cuts for various values of the meshing parameter. h_0 corresponds to a coarser mesh than the mesh presented in Figure 5.2, as it has only two layers of triangles in the radial direction.

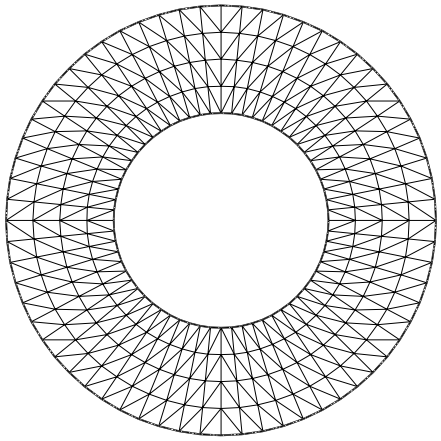


FIG. 5.2. *Coarse mesh.*

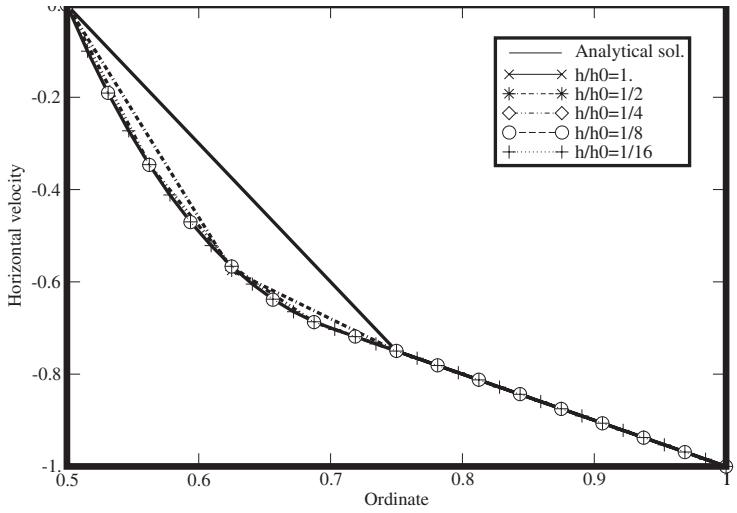


FIG. 5.3. *Velocity cuts for various meshing parameters,* $\tau_{YS} = 10Mpa$.

The $\mathcal{H}_1$ estimates for the velocity are drawn on Figure 5.4 as a function of the meshing parameter for various values of the plasticity threshold. One can observe that in this specific case the estimate is of order $O(h)$, thus confirming the preceding analysis.

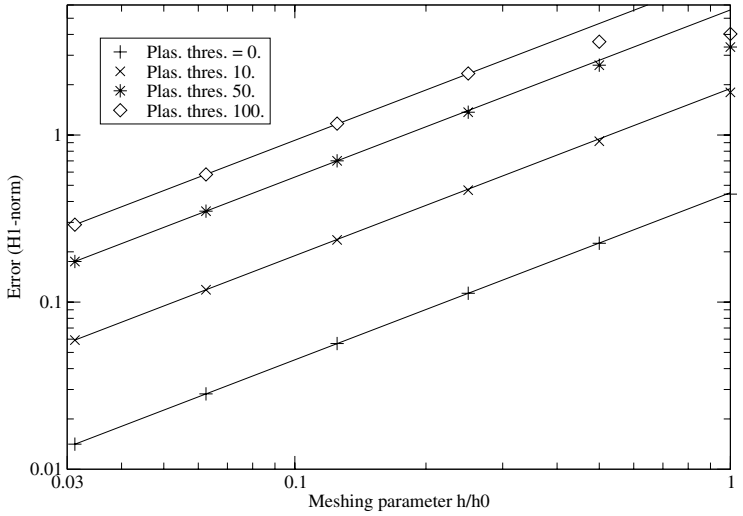


FIG. 5.4. *Error bound versus meshing parameter for various plasticity thresholds,* H1-norm.

6. Conclusion. We have proposed and analyzed in this paper a stabilized finite element scheme for the computation of incompressible creeping flows of Bingham fluids, using equal-order piecewise linear approximations for both velocity and pressure. This numerical scheme has shown several advantages due to the low degree of the velocity approximation: its convergence rate is the same as equivalent stable schemes, with an improved efficiency; in addition, the decomposition-coordination method can be used with a perfect matching of the consistency constraint between the auxiliary variable, namely the strain rate tensor, and the velocity field.

This numerical method can be extended straightforwardly to noncreeping Bingham flows using a characteristic-Galerkin strategy. Some results obtained in this way have been published in [18].

Acknowledgments. This work was performed at the French Institut de Radioprotection et Sûreté Nucléaire (IRSN), in the framework of reactor safety modeling activities. Developments were implemented in C++ as an application of PELICANS, an object-oriented platform developed by our team to provide general frameworks and software components for the implementation of PDE solvers.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] R. B. BIRD, R. C. ARMSTRONG, AND O. HASSAGER, *Dynamics of Polymeric Liquids, Vol.* 1*, Fluid Mechanics*, 2nd ed., Wiley-Interscience, New York, 1987.
[3] L. BOSCARDIN, *Méthodes de Lagrangien Augmenté pour la Résolution des Equations de Navier-Stokes dans le Cas d'Ecoulements de Fluide de Bingham*, Ph.D. thesis, Université de Franche-Comté, 1999.

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

[5] P. G. Ciarlet, *Handbook of Numerical Analysis Volume* II : *Finite Elements Methods—Basic Error Estimates for Elliptic Problems*, North–Holland, Amsterdam, 1991.

[6] G. Duvaut and J. L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin and New York, 1976.

[7] A. Fortin, D. Côté, and P. A. Tanguy, *On the imposition of friction boundary conditions for the numerical simulation of Bingham fluid flows*, Comput. Methods Appl. Mech. Engrg., 88 (1991), pp. 97–109.

[8] M. Fortin and R. Glowinski, *Méthodes de Lagrangien Augmenté*, Dunod, Paris, 1982.

[9] L. P. Franca and R. Stenberg, *Error analysis of some Galerkin least squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.

[10] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[11] W. Han and B. D. Reddy, *On the finite element method for mixed variational inequalities arising in elastoplasticity*, SIAM J. Numer. Anal., 32 (1995), pp. 1778–1807.

[12] T. J. R. Hughes, L. P. Franca, and M. Balestra, *A new finite element formulation for computational fluid dynamics:* V. *Circumventing the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations*, Comput. Meth. Appl. Mech. Engrg., 59 (1986), pp. 85–99.

[13] R. R. Huigol and M. P. Panizza, *On the determination of the plug flow region in Bingham fluids through the application of variational inequalities*, J. Non-Newtonian Fluid Mech., 58 (1995), pp. 207–217.

[14] D. Perić and S. Slijepčević, *Computational modelling of viscoplastic fluids based on a stabilized finite element method*, Engrg. Comput., 18 (2001), pp. 577–591.

[15] O. Pironneau, *Finite Element Methods for Fluids*, John Wiley and Sons / Masson, New York and Paris, 1989.

[16] N. Roquet and P. Saramito, *An adaptive finite element method for Bingham fluid flows around a cylinder*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 3317–3341.

[17] N. Roquet, P. Saramito, and R. Michel, *An adaptive finite element method for viscoplastic fluid flows in pipes*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 5391–5412.

[18] D. Vola, L. Boscardin, and J. Latché, *Laminar unsteady flows of Bingham fluids: A numerical strategy and some benchmark results*, J. Comput. Phys., 187 (2003), pp. 441–456.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1226–1260

ON THE CONSTRUCTION AND ANALYSIS OF HIGH ORDER LOCALLY CONSERVATIVE FINITE VOLUME-TYPE METHODS FOR ONE-DIMENSIONAL ELLIPTIC PROBLEMS*

MICHAEL PLEXOUSAKIS† AND GEORGIOS E. ZOURARIS‡

Abstract. Locally conservative, finite volume-type methods based on continuous piecewise polynomial functions of degree $r \geq 2$ are introduced and analyzed in the context of indefinite elliptic problems in one space dimension. The new methods extend and generalize the classical finite volume method based on piecewise linear functions. We derive a priori error estimates in the L^2, H^1, and L^∞ norm and discuss superconvergence effects for the error and its derivative. Explicit, residual-based a posteriori error bounds in the L^2 and energy norm are also derived. We compute the experimental order of convergence and show the results of an adaptive algorithm based on the a posteriori error estimates.

Key words. elliptic problems, finite volume methods, locally conservative methods, a priori error estimates, superconvergence, a posteriori error estimates, adaptive computations

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/S0036142902406302

1. Introduction.

1.1. Locally conservative and finite volume methods. Let $\Omega = (x_L, x_R)$ be a bounded interval and consider the following two-point boundary value problem: seek $u : \overline{\Omega} \to \mathbb{R}$ satisfying

$$Lu \equiv -(a(x)u')' + \beta(x)u' + \gamma(x)u = f(x) \quad \forall\, x \in \Omega, \tag{1.1a}$$

$$u(x_L) = u'(x_R) = 0, \tag{1.1b}$$

where a, β, γ, f are smooth, real-valued functions defined on $\overline{\Omega}$, and

$$0 < a_* \leq a(x) \quad \forall\, x \in \overline{\Omega}. \tag{1.1c}$$

We assume that (1.1) has a unique solution which is sufficiently smooth. In what follows, we shall operate under the assumptions that $f \in H^m(\Omega)$, $a \in C^{m+1}(\overline{\Omega})$, and $\beta, \gamma \in C^m(\overline{\Omega})$ for some $m \in \mathbb{N}$, which ensure that $u \in H^{m+2}(\Omega)$ (see, e.g., [1], [22]). The assumption (1.1c) ensures the strict ellipticity of L in the sense of [1] or [22].

Integrating (1.1a) over a subinterval $\omega = (\omega_L, \omega_R)$ of Ω we obtain

$$\mathcal{F}(u, \omega) \equiv -\big[(au')(\omega_R^-) - (au')(\omega_L^+)\big] + \int_\omega (\beta u' + \gamma u)\, dx - \int_\omega f\, dx = 0, \tag{1.2a}$$

which, when $\omega_R = x_R$, is written as

$$\mathcal{G}(u, \omega) \equiv (au')(\omega_L^+) + \int_\omega (\beta u' + \gamma u)\, dx - \int_\omega f\, dx = 0, \tag{1.2b}$$

*Received by the editors April 24, 2002; accepted for publication (in revised form) September 26, 2003; published electronically October 28, 2004. This research was supported by European Union TMR grant ERBFMRX-CT98-0234 (Viscosity Solutions and their Applications).
http://www.siam.org/journals/sinum/42-3/40630.html
†Department of Applied Mathematics, University of Crete, GR-714 09 Heraklion, Crete, Greece (plex@tem.uoc.gr).
‡Department of Mathematics, University of the Aegean, GR-832 00 Karlovassi, Samos, Greece (zouraris@aegean.gr).

due to the Neumann boundary condition (1.1b). When $\beta = \gamma = 0$ the relations (1.2a)–(1.2b) express a local *conservation* property for the solution of (1.1a)–(1.1b). The physical interpretation of this property depends on the situation where the equation is used as a model. For example, considering (1.1) as the steady state of a heat flow problem in a rod Ω, the property (1.2a) expresses the conservation of the thermal energy in a part ω of the rod.

A *locally conservative finite volume-type method* (or simply, *locally conservative method*) approximates the solution u of (1.1a)–(1.1b) by a function u_h, in a given finite element space X_h, that satisfies (1.2a) or (1.2b) on the subintervals of a covering $\{\Omega_j\}_{j=1}^{M_h}$ of Ω, where M_h depends on the dimension of X_h. We shall say that a locally conservative method for (1.1) is a *finite volume method* if the approximation $u_h \in X_h$ is entirely determined by the fulfillment of (1.2a) and/or (1.2b) on the subintervals $\{\Omega_j\}_{j=1}^{M_h}$. In the context of finite volume methods, the subintervals $\{\Omega_j\}_{j=1}^{M_h}$ are known as *control volumes* and we shall use the same terminology even in the case of locally conservative methods. If u_h satisfies (1.2a) (resp., (1.2b)) on a boundary subinterval $\omega = (\omega_L, x_R)$, then we shall say that ω is a boundary control volume of type I (resp., type II). A consequence of the above definition is that a locally conservative method requires less computational effort to assemble the matrix of the resulting linear system and the right-hand side, compared to the standard finite element method. In particular, the stiffness matrix (i.e., the part of the system matrix related to a) in the finite volume method does not require numerical quadrature.

The finite volume method has wide applicability in the approximation of solutions of hyperbolic equations of conservation laws (see, e.g., [30], [23], [26] and the references therein). These equations contain a term in space-divergence form and therefore a local conservation property, analogous to (1.2a), holds. The interest in extending the finite volume method to elliptic problems arises from the need to treat a second-order, regularizing viscosity term (e.g., a Laplacian), which is also written in space-divergence form. However, the finite volume method has also had early applications in the approximation of solutions of elliptic equations (see, e.g., [32]). In the bibliography, the proposed locally conservative methods for the approximation of solutions of elliptic problems are, mainly, of the finite volume type, and X_h consists of either piecewise constant functions (see, e.g., [21]), piecewise linear functions (see, e.g., [5], [24], [11], [10]), or piecewise quadratic functions [31]. The piecewise constant finite volume approach has a well developed theoretical background and wide applicability, and for some applications (e.g., digital image processing) is probably the most natural (see, e.g., [21], [34]). Variants of the finite volume method on piecewise linear functions, where (1.2a) is not satisfied, have also been proposed (see, e.g., [8], [24], [14]). Similar in spirit are the generalized finite difference schemes analyzed in, e.g., [29] or [39], which are based on Steklov averaging operators. Recently, the locally discontinuous Galerkin method (LDG) has been introduced (see, e.g., [12], [4]) with the aim of constructing approximations of solutions of elliptic problems in one or two space dimensions that have local conservation properties. The main idea in these methods is to rewrite the elliptic problem as a first-order system for the unknowns u and u' (or ∇u). Then the method constructs approximations $u_h \in X_h$ and $q_h \in X_h$ (or $q_h \in (X_h)^2$) of u and u' (or ∇u) in a finite element space X_h consisting of discontinuous functions. From the corresponding variational formulation follows that q_h has a local conservation property resembling (1.2a). In contrast, in the locally conservative methods investigated in this paper, the finite element space X_h consists of continuous functions and an approximation $u_h \in X_h$ is sought which, along with its derivative, satisfies the balance equation (1.2a).

As the references cited above indicate, a great deal of attention has been given to the construction and analysis of finite volume methods for elliptic problems in two space dimensions. The main tool in that analysis is the early idea (cf. [33]) to connect (1.2a) with a variational formulation similar to that of the finite element method (cf., e.g., [5], [24], [15], [14], [20]). In the piecewise linear case, and for self-adjoint or non-self-adjoint elliptic operators in two space dimensions, optimal first-order H^1 error estimates have been obtained by several authors under the regularity assumption $u \in H^2$ (cf., e.g., [5], [24], [25], [13], [35], [20]). Also, the authors of [9], [10], and [11] prove some second-order error estimates in a discrete H^1 norm. An optimal L^2 second-order error estimate for the Poisson equation is obtained by Hackbusch in [24] by assuming $u \in H^2$ and $f \in H^1$. Later, Jianguo and Shitong in [25], and Ewing, T. Lin, and Y. Lin in [20], show by examples in one and two dimensions that, in general, it is not possible to obtain optimal second-order L^2 convergence by assuming $u \in H^2$ and $f \in L^2$. Optimal order L^2 error estimates for general self-adjoint elliptic problems have been obtained by Chatzipantelidis in [13], for a nonconforming finite volume method, and by Chou and Li in [15] for a conforming one. The result in [13] relies on the assumption $u \in H^2$ and $f \in H^1$, while in [15] it is assumed that $u \in H^3$; here the second assumption is stronger because, in general, the regularity of the solution of an elliptic problem in the two-dimensional case also depends on the regularity of the boundary of the domain and not only on the regularity of the data. Analogous results are obtained in [27] for the Stokes problem and in [18], [19] for one-dimensional integro-differential equations. Recently in [20], Ewing, T. Lin, and Y. Lin show that if $f \in H^\alpha$ and $u \in H^{1+\alpha}$ for $\alpha \in (0, 1]$, then the order of the L^2 convergence of the finite volume method considered in [15] is equal to 2α. For the piecewise quadratic case, a finite volume method proposed by Liebau in [31] attains optimal, second-order convergence in the H^1 norm assuming $u \in H^3$. To the best of our knowledge, this is the only finite volume method based on piecewise polynomial functions of degree greater than one. The absence of a general theory for finite volume methods based on piecewise polynomial approximation spaces, analogous to that of the finite element method (cf., e.g., [7], [16]), is apparent.

An interesting step forward could be to find a systematic way of deriving finite volume methods for elliptic problems, based on finite element spaces X_h consisting of functions which are piecewise polynomials of degree greater than one. The main difficulty in achieving this goal arises from the fact that the definition of a finite volume method requires (except from the finite element space X_h) a set of control volumes with cardinal number proportional to the dimension of X_h. Hence, the control volume quest seems to be a very complicated procedure when the degree of the piecewise polynomial functions of X_h increases. It is worthwhile to notice that the choice of the control volumes may influence, apart from the well-posedness, the order of convergence of the obtained finite volume method. This phenomenon has been observed in the piecewise linear case where only special families of control volumes lead to a finite volume method which has the same order of convergence with that of the corresponding finite element method (cf. [24], [13], [15]). In the recent work of Liebau [31] the finite volume method has been extended on quadratics over a triangular mesh, but still we do not know if a finite volume method based on quadratics can attain optimal order of convergence in the L^2 norm.

1.2. Description of the results of the paper. In this paper we derive general classes of *new* optimal order, locally conservative, finite volume-type methods for the problem (1.1a)–(1.1b), based on continuous piecewise polynomial spaces of degree

$r \geq 2$. For these methods, we provide a general error estimation theory in the H^1, L^2, and L^∞ norms, analogous to that of the finite element method. Also, we discuss some pointwise error estimates for $r \geq 2$ and some a posteriori error estimates for $r = 2$.

Even though the "battlefield" of the current research in the area of the finite volume methods is the two-dimensional case, we restrict ourselves to the one-dimensional case so as to fix the ideas and gain insight into the structure of such methods. This choice came after preliminary work on the two-dimensional case. In particular, we arrived at the conclusion that, using integration rules, we may construct finite volume methods based on quadratic finite elements over a triangular mesh which are different from that of [31] and have optimal order of convergence in the H^1 norm. However, the existence of such a finite volume method with optimal order of convergence in the L^2 norm seems to be unsure. Then, we thought that we might find an answer to this question if we knew how to construct systematically finite volume methods, based on high order finite element spaces, in the one-dimensional case where the finite element geometry is simpler. The extension of the work at hand to the two-dimensional case, which is the our next step, is not straightforward because of the greater variety of the geometry of the relevant finite element spaces. However, the ideas and methods developed here may be used in two-dimensional problems when the discretization of the computational domain is achieved by rectangular elements (cf. Remark 2.2).

Before presenting a summary of the present paper, we list the basic results of our work:

- For every $r \geq 2$, we construct *locally conservative* methods for (1.1a)–(1.1b) where part of the degrees of freedoms are determined by finite element equations and the rest by equations concerning local conservativity in the mean of (1.2) (cf. Propositions 3.1–3.2). The interesting fact is that these methods have optimal order of convergence in the H^1, L^2, and L^∞ norm assuming the same regularity as in the finite element method.
- If $r = 2, 4$, or 6, we construct *finite volume* methods for (1.1a)–(1.1b) which have optimal order of convergence in the H^1, L^2, and L^∞ norm (cf. Proposition 3.7). Their control volumes are related to a dual mesh based on the roots of the well-known Legendre polynomial with degree r. Our opinion is that the technique used in deriving these finite volume methods works only for all *even* r but we are not able to provide a general proof for this fact.
- For every $r \geq 3$ we construct a general family of *finite volume* methods for (1.1a)–(1.1b) which have optimal order of convergence in the H^1 norm (cf. Proposition 3.6). Their control volumes are related to a dual mesh based on $r - 2$ arbitrary internal nodes of $[0, 1]$.

Let us present briefly the contents of the work at hand. Here, we extend and modify the framework proposed in [14] for the analysis of finite volume methods for elliptic problems in two space dimensions based on piecewise linear finite element spaces, in order to construct locally conservative methods for problem (1.1a)–(1.1b). Hence, we consider a general class of numerical methods based on continuous piecewise polynomial spaces which have a common variational formulation of the form

$$B_h(u_h, \chi) = (f, \Lambda_h \chi) \quad \text{for } \chi \in X_h, \tag{1.3}$$

where Λ_h is a linear operator defined on piecewise $\mathbb{P}^r$ functions into a finite-dimensional space consisting of piecewise polynomial functions, and B_h is a bilinear form that depends on Λ_h. The exact description of B_h is given in (2.9). We assume throughout

that the operator Λ_h satisfies (2.7), a *local* L^2 approximation property of order $\alpha \geq 1$, i.e.,

$$\| \Lambda_h v - v \|_{L^2(x_1,x_2)} \leq C\,(x_2 - x_1)^s \, \|v\|_{H^s(x_1,x_2)}, \quad s = 1, \ldots, \alpha, \tag{1.4}$$

and (2.8), a *local* error orthogonality property on $\mathbb{P}^{r-2+\sigma}$ with $\sigma = 0$ or 1, i.e.,

$$(\Lambda_h v - v, q)_{L^2(x_1,x_2)} = 0 \quad \forall q \in \mathbb{P}^{r-2+\sigma}(x_1, x_2), \tag{1.5}$$

where x_1, x_2 are consecutive nodes of a partition of Ω.

In section 3, we construct two large families of methods based on specific operators Λ_h satisfying(1.4)–(1.5), and yielding conservative equations of the form (1.2a), i.e., producing locally conservative methods. For the first family of the methods (see section 3.1) the operator Λ_h is defined as a local L^2-projection from $\mathbb{P}^r$ to $\mathbb{P}^{r-1}$ or $\mathbb{P}^{r-2}$, while for the second family of methods (see section 3.2) the operator Λ_h is defined via a special projection (see Proposition 3.5) from $\mathbb{P}^r$ to a space of piecewise constant functions. Both families of methods are infinite, with respect to the degree r of the underlying polynomial space. The methods of section 3.1 are *locally conservative* with $\alpha \geq 2$ and $\sigma = 0$ (see Proposition 3.1) or $\sigma = 1$ (see Proposition 3.2), while those of section 3.2 are of the *finite volume* type or *locally conservative* with $\alpha = 1$ and $\sigma = 0$ (see Proposition 3.6) or $\sigma = 1$ (see Proposition 3.7, Remark 3.5). In particular, for the methods of Proposition 3.1 and Proposition 3.6 we show that the discrete approximation u_h belongs to $C^1(\overline{\Omega})$ and satisfies the Neumann boundary condition at x_R. We note that the generality of the variational formulation (1.3) allows us to include as particular members, finite volume methods based on piecewise quadratic functions, where $\alpha = 1$, σ takes only the value 0, and the control volumes are related to the nodes of a quadrature rule (e.g., Simpson, Radau). The latter methods were introduced in [36] but cannot be generalized to $r \geq 3$ so that (2.8) is satisfied, and this is the reason we keep them out of our presentation. Hence, the connection of a finite volume method to a quadrature rule has a limited usefulness when $r \geq 2$.

The main result of the a priori error analysis in the L^2 and H^1 norms, for the general methods (1.3) introduced in section 2, is contained in Theorem 4.6 of section 4. There it is shown that convergence in the H^1 norm is always of optimal order r, while the optimal rate of convergence $r + 1$ in the L^2 norm is attained when $\alpha \geq 2$ or $\sigma = 1$. The latter result holds for the methods of Proposition 3.1 with $r \geq 3$, where $\alpha = r - 1$, the methods of Proposition 3.2 where $\alpha = r = 2m$, the methods of Remark 3.5 where $\alpha = 1$ and $\sigma = 1$, and finally for the *new* finite volume methods derived in Proposition 3.7 where also $\alpha = 1$ and $\sigma = 1$. The exact solution u is assumed to be in $H^{s_0}(\Omega)$, where for the H^1 estimate $s_0 = r + 1$ and for the L^2 estimate $s_0 = r + 1$ if $\alpha \geq 2$ and $s_0 = r + 1 + \sigma$ if $\alpha = 1$.

The a priori error analysis in L^∞ is carried out in section 5. We show that the methods with optimal order of L^2 convergence also attain optimal order of convergence in the L^∞ norm. Theorem 5.5 covers the case $\alpha \geq 2$ under the additional condition that Λ_h preserves piecewise linear functions which holds for the optimal order methods of section 3.1 (cf. Proposition 3.1, Proposition 3.2), while Theorem 5.7 covers the case $\alpha = 1$ and $\sigma = 1$ (cf. Proposition 3.7, Remark 3.5).

In section 6 we prove a pointwise error estimate of order $r+\alpha+\sigma-1$ at the nodes of the finite element partition (cf. Proposition 6.1). This is a superconvergence result when $\alpha \geq 3$ and $\sigma = 0$, or $\alpha \geq 2$ and $\sigma = 1$, which is the case of the locally conservative methods in Proposition 3.1 for $r \geq 4$ ($\sigma = 0$, $\alpha = r - 1$) and in Proposition 3.2 ($\sigma = 1$, $\alpha = r = 2m$). In addition, if a locally conservative approximation is C^1 at

the boundary points of a set of nonoverlapping control volumes and the boundary control volume is of type II, then we prove a general error estimate for the derivative of the error at these points (cf. Proposition 6.2, Remark 6.4). This result holds for the methods of Propositions 3.1, 3.6, and 3.7 and Remark 3.5.

The a posteriori error analysis for the case of quadratic finite element spaces is elaborated in section 7. We develop a residual-based a posteriori error analysis by extending the corresponding framework of the finite element method. In section 8 we verify numerically the convergence rates of some of the methods of section 3 and compare with the finite element method. In addition, we use the a posteriori error bounds of section 7 to construct an adaptive algorithm and test its performance by applying it to an appropriate test-problem.

2. Preliminaries and the formulation of the numerical method.

2.1. Notation and preliminaries. For $I \subset \mathbb{R}$ an open and bounded interval, $s \in \mathbb{N}_0$ and $1 \le p \le \infty$, we let $W^{s,p}(I)$ denote the Sobolev space of functions having generalized derivatives up to order s in the space $L^p(I)$. The norm of $v \in W^{s,p}(I)$ will be denoted by $\|v\|_{s,p,I} = (\sum_{j=0}^{s} \int_I |D^j v|^p \, dx)^{\frac{1}{p}}$ for $1 \le p < \infty$ and $\|v\|_{s,\infty,I} = \max_{0\le j\le s} (\operatorname{ess\,sup}_I |D^j v|)$ for $p = \infty$. We shall write $H^s(I) = W^{s,2}(I)$ and omit the index 2 from the symbol of its norm, i.e., $\|\cdot\|_{s,I} = \|\cdot\|_{s,2,I}$. $H_0^1(I)$ will denote the subspace of $H^1(I)$ consisting of functions which vanish at the endpoints of I in the sense of trace. The inner product and norm of $L^2(I)$ will be denoted by $(\cdot,\cdot)_I$ and $\|\cdot\|_{0,I}$, respectively, and the norm of $L^\infty(I)$ by $|\cdot|_{\infty,I}$. We shall also omit the index I from the norm and inner product symbols when $I = \Omega$. $C_B^s(I)$ will denote the space of $C^s(I)$-functions which, along with their classical derivatives up to order s, are continuously extensible to $\overline{I}$. Finally, we denote the restriction of L on $H^2(I)$ by L_I, so that $L \equiv L_\Omega$, and the characteristic function of I by $\mathcal{X}_I$.

The length of the interval I will be denoted by h_I. Further, if $I = (y_L, y_R)$ and $v \in L^2(I)$ is such that there exists $\delta > 0$ for which $v|_{(y_L, y_L+\delta)} \in C_B^0(y_L, y_L + \delta)$ and $v|_{(y_R-\delta, y_R)} \in C_B^0(y_R - \delta, y_R)$, we shall write $[\![v]\!]_{\partial I} = v(y_R - 0) - v(y_L + 0)$. Here, and in what follows, $v(x \pm 0) = \lim_{\varepsilon\to 0^+} v(x \pm \varepsilon)$. We also let $\mathcal{H}(I) = \{v \in H^1(I) : v(y_L) = 0\}$.

For $h \in (0,1)$, let $\mathcal{P}_h$ denote a (nonuniform) partition of Ω with J_h+1 nodes, $x_L = x_0^h < x_1^h < \cdots < x_{J_h}^h = x_R$, such that $\max_{1\le j\le J_h}(x_j^h - x_{j-1}^h) \le C\,h$ for some positive constant C independent of h. We stress that this weak mesh assumption is sufficient for the error estimates of sections 4–7, and thus, a stronger mesh assumption such as quasi uniformity is not required. We shall write $I_j^h = (x_{j-1}^h, x_j^h)$ for $j = 1, \ldots, J_h$ and $x_{j+z}^h = x_j^h + z h_{I_{j+1}^h}$ for $j = 0, \ldots, J_h - 1$ and $z \in [0,1]$; furthermore, we let $x_{j-z}^h = x_{j-1+(1-z)}^h$ for $j = 1, \ldots, J_h$ and $z \in [0,1]$. Also, for $j = 1, \ldots, J_h$, we define $\xi_j^h : I_j^h \to [0,1]$ by $\xi_j^h(x) := (x - x_{j-1}^h)/h_{I_j^h}$ for $x \in I_j^h$.

For $m \in \mathbb{N}_0$, we denote by $\mathbb{P}^m(I)$ the space of polynomials of degree less than or equal to m, restricted on I, and let $\mathbb{P}_h^m = \{v \in L^2(\Omega) : v|_I \in \mathbb{P}^m(I)\ \forall\, I \in \mathcal{P}_h\}$. Letting $\mathcal{L}_m \in \mathbb{P}^m(-1,1)$ be the well-known Legendre polynomial of degree m corresponding to the weight function $w \equiv 1$ (cf., e.g., [38]), we denote by W_m its shift to $(0,1)$, i.e., $W_m \in \mathbb{P}^m(0,1)$ and $W_m(x) = \mathcal{L}_m(2x-1)$ for $x \in (0,1)$. Then W_m has m discrete roots in $(0,1)$ and satisfies $W_m(0) = (-1)^m$ and $W_m(1) = 1$. We shall also make use of the space $\mathcal{C}_h^m = \{v \in L^2(\Omega) : v|_I \in C_B^m(I)\ \ \forall\, I \in \mathcal{P}_h\}$. In addition, we introduce the space $H_h^m = \{v \in L^2(\Omega) : v|_I \in H^m(I)\ \ \forall\, I \in \mathcal{P}_h\}$ and equip it with the mesh-dependent norm $\|v\|_{m,h} = \{\sum_{I\in\mathcal{P}_h} \|v\|_{m,I}^2\}^{\frac{1}{2}}$ for $v \in H_h^m$.

We shall seek approximations of the solution u of (1.1a)–(1.1b) in the space S_h^r, $r \geq 2$, of continuous functions which vanish at $x_{_L}$ and reduce to polynomials of degree less than or equal to r on each $I \in \mathcal{P}_h$, i.e., $S_h^r = \mathbb{P}_h^r \cap \mathcal{H}(\Omega)$. Since the polynomial spaces, restricted to I, have finite dimension, all the norms are equivalent but the constants depend on h_I. In particular, the following *local inverse property* holds (see, e.g., [7, section 4.5]): there exists a constant $C = C(r)$, independent of h, such that for $0 \leq m \leq \ell \leq r$

$$\|\chi\|_{\ell,I} \leq C\, h_I^{m-\ell} \|\chi\|_{m,I} \quad \forall\, I \in \mathcal{P}_h, \quad \forall\, \chi \in \mathbb{P}_h^r. \tag{2.1}$$

For $v \in \mathcal{C}_h^0$ and $m \in \mathbb{N}$, we denote by $\mathcal{I}_h^m v$ the Lagrange interpolant of v, i.e., the unique function of $\mathbb{P}_h^m$ which, for $j = 0, \ldots, J_h - 1$, satisfies $(\mathcal{I}_h^m v - v)(x^h_{j+\frac{k}{m}}) = 0$, $k = 1, \ldots, m-1$, $(\mathcal{I}_h^m v - v)(x_j^h + 0) = 0$, and $(\mathcal{I}_h^m v - v)(x_{j+1}^h - 0) = 0$. It is well known (cf., e.g., [7, section 4.4]) that $\mathcal{I}_h^m$ possesses the following approximation properties:

$$\sum_{s=0}^{m} h_I^s \, \|\mathcal{I}_h^m v - v\|_{s,I} \leq \widetilde{C}_m \, h_I^{m+1} \, \|v\|_{m+1,I} \quad \forall\, v \in H^{m+1}(I), \quad \forall\, I \in \mathcal{P}_h, \tag{2.2}$$

$$|\mathcal{I}_h^m v - v|_{\infty,I} \leq \widetilde{C}_{\infty,m} \, h_I^{m+1} \, \|v\|_{m+1,\infty,I} \quad \forall\, v \in W^{m+1,\infty}(I), \quad \forall\, I \in \mathcal{P}_h, \tag{2.3}$$

$$\|\mathcal{I}_h^1 v - v\|_{0,I} \leq h_I \, \|v\|_{1,I} \quad \forall\, v \in H^1(I), \quad \forall\, I \in \mathcal{P}_h. \tag{2.4}$$

We shall also make use of the L^2-projection operator Π_h on $\mathbb{P}_h^0$, defined by

$$(\Pi_h v, q) = (v, q) \quad \forall\, q \in \mathbb{P}_h^0, \quad \forall\, v \in L^2(\Omega).$$

It is easy to show that $\Pi_h v|_I = \frac{1}{h_I} \int_I v \, dx \;\forall\, I \in \mathcal{P}_h$, $\forall\, v \in L^2(\Omega)$, and

$$\|v - \Pi_h v\|_{0,I} \leq h_I \, \|v'\|_{0,I} \quad \forall\, I \in \mathcal{P}_h, \quad \forall\, v \in H_h^1, \tag{2.5}$$

$$|v - \Pi_h v|_{\infty,I} \leq h_I \, |v'|_{\infty,I} \quad \forall\, I \in \mathcal{P}_h, \quad \forall\, v \in \mathcal{C}_h^1. \tag{2.6}$$

2.2. A variational formulation of the method: The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$. In this section we formulate a family of numerical methods for the approximation of the solution of (1.1a)–(1.1b) from the finite element space S_h^r, $r \geq 2$, which are generalizations of the standard Galerkin method. As we shall see later in section 3, we can define *finite volume* and general *locally conservative* methods as particular members of this family. A similar approach has been introduced in [14] for the analysis of finite volume element methods based on piecewise linear functions and applied to two-dimensional linear elliptic problems.

The basic ingredient in the definition of our numerical methods is a linear operator $\Lambda_h \colon \mathbb{P}_h^r \to L^2(\Omega)$, satisfying the following stability- and consistency-like assumptions:

$$\exists\, \alpha \in \mathbb{N} : \; \|\Lambda_h v - v\|_{0,I} \leq \widehat{C}_{r,s} \, h_I^s \, \|v\|_{s,I} \quad \forall\, I \in \mathcal{P}_h, \quad s = 1, \ldots, \alpha, \; \forall\, v \in \mathbb{P}_h^r, \tag{2.7}$$

$$\exists\, \sigma \in \{0, 1\} : \; (\Lambda_h v - v, q)_I = 0 \quad \forall\, q \in \mathbb{P}^{r-2+\sigma}(I), \quad \forall\, I \in \mathcal{P}_h, \; \forall\, v \in \mathbb{P}_h^r. \tag{2.8}$$

A discrete variational formulation of (1.1a)–(1.1b) is then defined as follows: for $h \in (0, 1)$ we seek $u_h \in S_h^r$ such that

$$B_h(u_h, \chi) = (f, \Lambda_h \chi) \quad \forall\, \chi \in S_h^r, \tag{2.9}$$

where the bilinear form $B_h : H_h^2 \times \mathbb{P}_h^r \to \mathbb{R}$ is defined as

$$B_h(v, \chi) = \sum_{I \in \mathcal{P}_h} \left\{ [\![a v' \chi]\!]_{\partial I} + (L_I v, \Lambda_h \chi)_I \right\} \quad \forall\, v \in H_h^2, \quad \forall\, \chi \in \mathbb{P}_h^r.$$

That the method (2.9) is well defined is discussed later, in Proposition 4.4. However, the consistency of the method (2.9) is straightforward. Indeed, the solution $u \in H^2(\Omega)$ of (1.1) satisfies $B_h(u,\chi) = (f, \Lambda_h\chi)$ for $\chi \in S_h^r$, since $u', \chi \in H^1(\Omega) \subset C(\overline{\Omega})$. Thus we arrive at

$$B_h(u - u_h, \chi) = 0 \quad \forall\, \chi \in S_h^r, \tag{2.10}$$

which is analogous to the orthogonality property of the finite element method. Note also that in the standard Galerkin finite element method we seek $\widetilde{u}_h \in S_h^r$ such that $B(\widetilde{u}_h, \chi) = (f, \chi)\ \forall\, \chi \in S_h^r$, where

$$B(v,\chi) = \sum_{I \in \mathcal{P}_h} \left\{ (av', \chi')_I + (\beta v', \chi)_I + (\gamma v, \chi)_I \right\} \quad \forall\, v, \chi \in H_h^1.$$

Using integration by parts we obtain the fundamental, for the error estimation, identity

$$B_h(v,\chi) = B(v,\chi) + \sum_{I \in \mathcal{P}_h} (L_I v, \Lambda_h\chi - \chi)_I \quad \forall\, v \in H_h^2, \quad \forall\, \chi \in \mathbb{P}_h^r, \tag{2.11}$$

relating the bilinear forms of the finite element and the locally conservative methods. Thus, the standard Galerkin finite element method and (2.9) coincide when we choose $\Lambda_h\chi = \chi$, for $\chi \in \mathbb{P}_h^r$.

REMARK 2.1. *Let us consider the two-point boundary value problem*

$$-(\widehat{a}(x)\,\widehat{u}' - \widehat{\beta}(x)\,\widehat{u})' + \widehat{\gamma}(x)\,\widehat{u} = \widehat{f}(x)\ \forall\, x \in \Omega, \quad \widehat{u}(x_L) = \widehat{u}'(x_R) = 0, \tag{2.12}$$

where $\inf_\Omega \widehat{a} > 0$, $\widehat{a}, \widehat{\beta} \in C^1(\overline{\Omega})$, $\widehat{\gamma} \in C^0(\overline{\Omega})$, *and* $\widehat{f} \in L^2(\Omega)$. *We can construct an approximation* $\widehat{u}_h \in S_h^r$ *of the solution* $\widehat{u}$ *by writing* (2.12) *in the equivalent form* (1.1a) *with* $a = \widehat{a}$, $\beta = \widehat{\beta}$, $\gamma = \widehat{\gamma} + \widehat{\beta}'$ *and applying method* (2.9). *Next, let us also assume that* (2.9) *is locally conservative with control volumes* $\omega \in \mathcal{V}$. *Then* $-[\![\widehat{a}\,\widehat{u}_h']\!]_{\partial\omega} + \int_\omega (\widehat{\beta}\,\widehat{u}_h' + \gamma\,\widehat{u}_h)\,dx = \int_\omega \widehat{f}\,dx$ *for* $\omega \in \mathcal{V}$. *Since* $\widehat{\beta}\,\widehat{u}_h \in H^1(\Omega)$ *we obtain* $-[\![\widehat{a}\,\widehat{u}_h' - \widehat{\beta}\,\widehat{u}_h]\!]_{\partial\omega} + \int_\omega \widehat{\gamma}\,\widehat{u}_h\,dx = \int_\omega \widehat{f}\,dx$ *for* $\omega \in \mathcal{V}$. *If* ω *is a boundary control volume of type* II *we have* $(\widehat{a}\,\widehat{u}_h')(\omega_L^+) + [\![\widehat{\beta}\,\widehat{u}_h]\!]_{\partial\omega} + \int_\omega \widehat{\gamma}\,\widehat{u}_h\,dx = \int_\omega \widehat{f}\,dx$, *where* ω_L *is the left endpoint of* ω. *Hence, if the method* (2.9) *is locally conservative for the problem* (1.1a)–(1.1b), *then we can use it to derive locally conservative approximations of problem* (2.12), *which is the conservative form of* (1.1a)–(1.1b).

REMARK 2.2. *In a two-dimensional setting, the derivation of a high order, locally conservative method requires that the term* $[\![av'\chi]\!]_{\partial I}$ *in the definition of* $B_h(v,\chi)$ *be replaced by* $\int_\tau A\nabla v \cdot n\, Q_h\chi\,dS$, *where* A *is the diffusion matrix of the elliptic problem,* τ *is an element of a partition of the domain* Ω, *and* Q_h *is some appropriate operator with properties analogous to those of* Λ_h (*cf.* [14] *for finite volume methods based on piecewise linear functions*). *In the case of triangular elements it is not obvious how to construct an operator* Q_h *yielding locally conservative methods with optimal order convergence properties. In the case where the discretization of the computational domain is achieved by rectangular elements, it seems possible to construct such an operator* Q_h *using the techniques of the present work.*

3. Examples of locally conservative methods. In this section, for $r \geq 2$, we provide examples of operators $\Lambda_h : \mathbb{P}_h^r \to L^2(\Omega)$ satisfying the assumptions (2.7)–(2.8) and yielding locally conservative methods based on S_h^r. We note that for $r = 2$ we can construct locally conservative and finite volume methods using *quadrature rules* exact for polynomials of degree at least two (cf. [36] for details). These methods, however, cannot be generalized to $r \geq 3$ so that (2.8) is satisfied.

3.1. Methods based on local L^2 projections. We present here some locally conservative methods on S_h^r based on a local L^2 projection of $\mathbb{P}^r$ onto $\mathbb{P}^{r-2}$ for $r \geq 2$ (cf. Proposition 3.1), or a local L^2 projection of $\mathbb{P}^r$ onto $\mathbb{P}^{r-1}$ for even $r \geq 2$ (cf. Proposition 3.2). The basic characteristic of these methods is that (2.7) holds with $\alpha \geq 2$, and thus, as we shall show later in sections 4 and 5, we may obtain optimal order of convergence in the H^1, L^2, and L^∞ norm, assuming for u the same regularity as in the standard finite element method. These methods can be considered as a bridge between the finite element and the finite volume method.

PROPOSITION 3.1. *Let $r \geq 2$ and $\Lambda_h : \mathbb{P}_h^r \to \mathbb{P}_h^{r-2}$ be defined by*

$$(\Lambda_h p - p, q) = 0 \quad \forall\, q \in \mathbb{P}_h^{r-2}, \quad \forall\, p \in \mathbb{P}_h^r.$$

Then (2.7) holds with $\alpha = r-1$ and (2.8) is satisfied with $\sigma = 0$. Also, the method (2.9) is a locally conservative method on S_h^r with overlapping control volumes $\{I_j^h\}_{j=1}^{J_h}$ and $\{(x_{j-1}^h, x_{j+1}^h)\}_{j=1}^{J_h-1}$, where the interval $I_{J_h}^h$ is a boundary control volume of both types I *and* II. *Moreover, the corresponding approximation u_h belongs to $S_h^r \cap C^1(\overline{\Omega})$ and satisfies a homogeneous Neumann boundary condition at x_R.*

Proof. The definition of Λ_h immediately implies (2.8) for $\sigma = 0$. Let $I \in \mathcal{P}_h$ and $v \in \mathbb{P}^r(I)$. Since $\|\Lambda_h v - v\|_{0,I} = \inf_{\chi \in \mathbb{P}^{r-2}(I)} \|\chi - v\|_{0,I}$, using (2.2), (2.5), and (2.4) we conclude that (2.7) holds with $\alpha = r-1$.

Let θ_1, $\theta_2 \in \{0,1\}$ and $\theta = (\theta_1, \theta_2)$. We can find a polynomial $p_\theta \in \mathbb{P}^r(0,1)$ such that $p_\theta(0) = \theta_1$, $p(1) = \theta_2$ and whose L^2-projection on $\mathbb{P}^{r-2}(0,1)$ is the constant 1. Indeed, if p_θ is such a polynomial, then $p_\theta(x) = w_\theta(x)\left(\sum_{j=1}^{r-1} a_j^\theta\, x^{j-1}\right)$, where $w_\theta(x) = |x-\theta_1|\,|x-\theta_2|$. Requiring $\int_0^1 x^{\ell-1}\, p_\theta(x)\, dx = \frac{1}{\ell}$, $\ell = 1, \ldots, r-1$, we get $\sum_{j=1}^{r-1} a_j^\theta\, A_{\ell j}^\theta = \frac{1}{\ell}$ for $\ell = 1, \ldots, r-1$, where $A_{\ell j}^\theta = \int_0^1 w_\theta(x)\, x^{\ell-1} x^{j-1}\, dx$. A^θ is invertible being the Gram matrix of the linearly independent vectors $\{\sqrt{w_\theta(x)}\, x^{k-1}\}_{k=1}^{r-1}$ of $L^2(0,1)$. Thus, p_θ is fully determined.

Set now $\overline{p} = p_{(0,0)}$ and $\widetilde{p} = p_{(0,1)}$, and consider the linearly independent functions $\{\overline{\phi}_j, \widetilde{\phi}_j\}_{j=1}^{J_h}$ of S_h^r defined by $\overline{\phi}_j = (\overline{p} \circ \xi_j^h)\, \mathcal{X}_{I_j^h}$ for $j = 1, \ldots, J_h$, $\widetilde{\phi}_j = (\widetilde{p} \circ \xi_j^h)\, \mathcal{X}_{I_j^h} + (\widetilde{p} \circ (1 - \xi_{j+1}^h))\, \mathcal{X}_{I_{j+1}^h}$ for $j = 1, \ldots, J_h - 1$, and $\widetilde{\phi}_{J_h} = (\widetilde{p} \circ \xi_{J_h}^h)\, \mathcal{X}_{I_{J_h}^h}$. Then we have $\Lambda_h \overline{\phi}_j = \mathcal{X}_{I_j^h}$ for $j = 1, \ldots, J_h$, $\Lambda_h \widetilde{\phi}_j = \mathcal{X}_{I_j^h \cup I_{j+1}^h}$ for $j = 1, \ldots, J_h - 1$, and $\Lambda_h \widetilde{\phi}_{J_h} = \mathcal{X}_{I_{J_h}^h}$. Setting $\chi = \overline{\phi}_j$ in (2.9) we obtain

$$\int_{I_j^h} L_{I_j^h} u_h\, dx = \int_{I_j^h} f\, dx, \quad j = 1, \ldots, J_h. \tag{3.1}$$

With $\chi = \widetilde{\phi}_j$ in (2.9) we get

$$(au_h')(x_j^h - 0) - (au_h')(x_j^h + 0) + \sum_{\ell \in \{j, j+1\}} \int_{I_\ell^h} \left(L_{I_\ell^h} u_h - f\right) dx = 0 \tag{3.2}$$

for $j = 1, \ldots, J_h - 1$ and

$$(au_h')(x_{J_h}^h - 0) + \int_{I_{J_h}^h} \left(L_{I_{J_h}^h}\, u_h - f\right) dx = 0. \tag{3.3}$$

Using integration by parts, we arrive at

$$\mathcal{F}(u_h, (x_{j-1}^h, x_{j+1}^h)) = 0, \quad j = 1, \ldots, J_h - 1, \quad \mathcal{G}(u_h, I_{J_h}^h) = 0, \tag{3.4}$$

$$\mathcal{F}(u_h, I_j^h) = 0, \quad j = 1, \ldots, J_h. \tag{3.5}$$

The relations in (3.4) and (3.5) show that the method (2.9) is a locally conservative method with control volumes $\{I_j^h\}_{j=1}^{J_h}$ and $\{(x_{j-1}^h, x_{j+1}^h)\}_{j=1}^{J_h-1}$. The interval $I_{J_h}^h$ is a boundary control volume of type I due to (3.5) and of type II due to the second equation in (3.4). Also, (3.2), (3.3), and (3.1) imply $u_h'(x_R - 0) = 0$ and $u_h'(x_j^h - 0) = u_h'(x_j^h + 0)$ for $j = 1, \ldots, J_h - 1$, which completes the proof. ∎

REMARK 3.1. *We note that the method of Proposition* 3.1 *with* $r = 2$ *is a finite volume method.*

PROPOSITION 3.2. *Let* $m \in \mathbb{N}$, $r = 2m$, *and* $\Lambda_h : \mathbb{P}_h^r \to \mathbb{P}_h^{r-1}$ *be defined by*

$$(\Lambda_h p - p, q) = 0 \quad \forall\, q \in \mathbb{P}_h^{r-1}, \quad \forall\, p \in \mathbb{P}_h^r. \tag{3.6}$$

Then (2.7) *holds with* $\alpha = r$ *and* (2.8) *is satisfied with* $\sigma = 1$. *Also, the method* (2.9) *is a locally conservative method on* S_h^r *with control volumes* $\{I_j^h\}_{j=1}^{J_h}$, *where the interval* $I_{J_h}^h$ *is a boundary control volume of type* I.

Proof. The definition of Λ_h immediately yields that (2.8) holds for $\sigma = 1$. Let $I \in \mathcal{P}_h$ and $v \in \mathbb{P}^r(I)$. Since $\|\Lambda_h v - v\|_{0,I} \le \|\mathcal{I}_h^{r-1} v - v\|_{0,I}$, the use of (2.2) and (2.4), implies that (2.7) holds with $\alpha = r$.

We define $\widehat{p} \in \mathbb{P}^r(0,1)$ by $\widehat{p}(x) = 1 - W_r(x)$ for $x \in (0,1)$ (cf. section 2.1). Then we have $\widehat{p}(0) = 0$, $\widehat{p}(1) = 0$ and $\int_0^1 (\widehat{p}(x) - 1)\, \widehat{q}(x) dx = 0$ $\forall \widehat{q} \in \mathbb{P}^{r-1}(0,1)$, because of the orthogonality property of the Legendre polynomials, i.e., $\int_{-1}^1 \mathcal{L}_r(x)\, q(x) dx = 0$ $\forall q \in \mathbb{P}^{r-1}(-1,1)$. Consider the linearly independent functions $\{\widehat{\phi}_j\}_{j=1}^{J_h} \subset S_h^r$ defined by $\widehat{\phi}_j = (\widehat{p} \circ \xi_j^h) \mathcal{X}_{I_j^h}$, $j = 1, \ldots, J_h$. By construction we have that $\Lambda_h \widehat{\phi}_j = \mathcal{X}_{I_j^h}$, $j = 1, \ldots, J_h$. Setting $\chi = \widehat{\phi}_j$ in (2.9) we arrive at $\mathcal{F}(u_h, I_j^h) = 0$, $j = 1, \ldots, J_h$, which completes the proof of the proposition. ∎

REMARK 3.2. *Let* $m \in \mathbb{N}$, $r = 2m$, Λ_h *be defined by* (3.6) *and* $\widehat{B} = \{\widehat{\phi}_j\}_{j=1}^{J_h}$ *be the set of linearly independent functions of* S_h^r *defined in the proof of Proposition* 3.2. *If* $\widetilde{B}$ *is a basis of* S_h^{r-1}, *then the set* $\widetilde{B} \cup \widehat{B}$ *is a basis of* S_h^r *because* $\mathrm{card}(\widetilde{B}) = (r-1)J_h$ *and* $\widehat{\phi}_j|_{I_j^h} \in \mathbb{P}^r(I_j^h) \backslash \mathbb{P}^{r-1}(I_j^h)$ *for* $j = 1, \ldots, J_h$. *Since* $\Lambda_h \widetilde{\phi} = \widetilde{\phi}$, *for* $\widetilde{\phi} \in \widetilde{B}$, *the rest degrees of freedom of the approximation* u_h *are specified by the finite element-type equations*: $B(u_h, \widetilde{\phi}) = (f, \widetilde{\phi})$ *for* $\widetilde{\phi} \in \widetilde{B}$.

REMARK 3.3. *Let* $m \in \mathbb{N}$, $r = 2m + 1$ *and* Λ_h *be defined by* (3.6). *Proceeding as in the proof of Proposition* 3.1, *we conclude that there is a unique polynomial* $p_* \in \mathbb{P}^r(0,1)$ *such that* $p_*(0) = 0$ *and whose* L^2 *projection on* $\mathbb{P}^{r-1}(0,1)$ *is equal to* 1, *i.e.,* $\int_0^1 (p_*(x) - 1)\, q(x) dx = 0$ $\forall q \in \mathbb{P}^{r-1}(0,1)$. *Setting* $q = p_*' \in \mathbb{P}^{r-1}(0,1)$ *we obtain* $p_*(1)(p_*(1) - 2) = 0$, *so that either* $p_*(1) = 0$ *or* $p_*(1) = 2$. *Since* r *is odd, it is easily seen that* $p_* = W_r + 1$, *which yields* $p_*(1) = 2$. *This means that it is not possible to have* $p_*(1) = 0$ *or* $p_*(1) = 1$, *and hence we cannot employ the technique used in Propositions* 3.1 *and* 3.2 *to show that the method is locally conservative.*

3.2. Methods based on a special projection of $\mathcal{C}_h^0$ onto piecewise constants. In this subsection we define a special projection of the piecewise continuous functions of $\mathcal{C}_h^0$ onto a space consisting of piecewise constant functions and use it to derive finite volume methods on S_h^r, for $r \ge 2$. For these finite volume methods, (2.7) holds with $\alpha = 1$ and (2.8) is satisfied with $\sigma = 0$ or 1. The convergence analysis of sections 4 and 5 shows that the order of convergence is optimal in the H^1 norm when $\sigma = 0$ and in the H^1, L^2, and L^∞ norm when $\sigma = 1$.

Let $s \ge 2$ and $\varrho = \{\varrho_j\}_{j=0}^s \subset \mathbb{R}$ be the nodes of a partition of $[0,1]$, i.e., $\varrho_0 = 0$, $\varrho_s = 1$, and $\varrho_{j-1} < \varrho_j$ for $j = 1, \ldots, s$. The next lemma shows that there is a unique

polynomial of degree less than or equal to $s-1$ with prescribed integral over each subinterval of the partition.

LEMMA 3.3. *Let $s \geq 2$ and $\varrho = \{\varrho_j\}_{j=0}^s$ be the nodes of a partition of $[0,1]$. Then, for given $E = \{\varepsilon_j\}_{j=1}^s \subset \mathbb{R}$ there exists a unique $p_E \in \mathbb{P}^{s-1}[0,1]$ such that*

$$\int_{\varrho_{j-1}}^{\varrho_j} p_E(x)\, dx = \varepsilon_j, \quad j = 1, \ldots, s. \tag{3.7}$$

Proof. Observing that every $p \in \mathbb{P}^{s-1}[0,1]$ has a unique representation as $p(x) = \sum_{i=1}^s i\, a_i\, x^{i-1}$, it is easily seen that (3.7) is equivalent to $\sum_{i=1}^s a_i\, (\varrho_j^i - \varrho_{j-1}^i) = \varepsilon_j$, $j = 1, \ldots, s$. Since $\varrho_0 = 0$, the last relation and a simple induction argument yields that (3.7) is, finally, equivalent to $\sum_{i=1}^s a_i\, \hat{A}_{ji} = \frac{\varepsilon_j + \varepsilon_{j-1}}{\varrho_j}$ $j = 1, \ldots, s$, where $\hat{A}_{ji} = (\varrho_j)^{i-1}$ and $\varepsilon_0 = 0$. Since $\hat{A} \in \mathbb{R}^{s\times s}$ is a Vandermonde matrix, the coefficients $\{a_i\}_{i=1}^s$ are uniquely determined and this proves the assertion of the lemma. ∎

The result of Lemma 3.3 allows us to construct a basis of $\mathbb{P}^{s-1}$ for $s \geq 2$ that is useful for our purposes.

LEMMA 3.4. *Let $s \geq 2$ and $\varrho = \{\varrho_j\}_{j=0}^s$ be the nodes of a partition of $[0,1]$. The polynomials $\{z_\ell^{\varrho,s-1}\}_{\ell=1}^s \subset \mathbb{P}^{s-1}[0,1]$ satisfying*

$$\int_{\varrho_{j-1}}^{\varrho_j} z_\ell^{\varrho,s-1}(x)\, dx = \delta_{\ell j}, \quad j, \ell = 1, \ldots, s, \tag{3.8}$$

form a basis of $\mathbb{P}^{s-1}[0,1]$.

Proof. Lemma 3.3 ensures that $\{z_\ell^{\varrho,s-1}\}_{\ell=1}^s$ are well-defined elements of $\mathbb{P}^{s-1}[0,1]$. To show that they are linearly independent, we assume that there exist real numbers $\{\lambda_j\}_{j=1}^s$ such that $\sum_{j=1}^s \lambda_j\, z_j^{\varrho,s-1} = 0$. Then, for $j_0 = 1, \ldots, s$, integrate the last relation over $[\varrho_{j_0-1}, \varrho_{j_0}]$ and use (3.8) to obtain $\sum_{j=1}^s \lambda_j\, \delta_{jj_0} = 0$, or $\lambda_{j_0} = 0$. ∎

REMARK 3.4. *Let $\{w_\ell^{\varrho,s}\}_{\ell=1}^s \subset \mathbb{P}^s(0,1)$ be such that $w_\ell^{\varrho,s}(\varrho_j) = 0$ for $j = 0, \ldots, \ell-1$ and $w_\ell^{\varrho,s}(\varrho_j) = 1$ for $j = \ell, \ldots, s$. Then we have $z_\ell^{\varrho,s-1} = (w_\ell^{\varrho,s})'$ for $\ell = 1, \ldots, s$.*

We define now a linear operator that maps the piecewise continuous functions $\mathcal{C}_h^0$ onto piecewise constant functions and has properties analogous to (2.7) and (2.8).

PROPOSITION 3.5. *Let $s \geq 2$, $\varrho = \{\varrho_j\}_{j=0}^s$ be the nodes of a partition of $[0,1]$, $\{z_\ell^{\varrho,s-1}\}_{\ell=1}^s$ be the basis of $\mathbb{P}^{s-1}[0,1]$ described in Lemma 3.4 and $\widetilde{\Lambda}_h^{\varrho,s} : \mathcal{C}_h^0 \to L^2(\Omega)$ be a linear operator defined by*

$$\widetilde{\Lambda}_h^{\varrho,s} v|_{I_j^h} = \frac{1}{h_{I_j^h}} \sum_{\ell=1}^s \big(v, (z_\ell^{\varrho,s-1} \circ \xi_j^h)\big)_{I_j^h}\, \mathcal{X}_{I_{j,\ell}^{\varrho,h}}, \quad j = 1, \ldots, J_h, \tag{3.9}$$

where $I_{j,\ell}^{\varrho,h} = (x_{j-1+\varrho_{\ell-1}}^h, x_{j-1+\varrho_\ell}^h)$. Then, it holds that

$$\big(\widetilde{\Lambda}_h^{\varrho,s} v - v, q\big)_I = 0 \quad \forall\, q \in \mathbb{P}^{s-1}(I), \quad \forall\, I \in \mathcal{P}_h, \quad \forall\, v \in \mathcal{C}_h^0, \tag{3.10}$$

$$\big\|\widetilde{\Lambda}_h^{\varrho,s} \tilde{v} - \tilde{v}\big\|_{0,I} \leq\ C\, h_I\, \|\tilde{v}\|_{1,I} \quad \forall\, I \in \mathcal{P}_h, \quad \forall\, \tilde{v} \in H_h^1. \tag{3.11}$$

Proof. Let $v \in \mathcal{C}_h^0$, $\tilde{v} \in H_h^1$, and $j \in \{1, \ldots, J_h\}$. Using (3.8) and (3.9), we obtain

$$\begin{aligned}\big(\widetilde{\Lambda}_h^{\varrho,s} v, (z_i^{\varrho,s-1} \circ \xi_j^h)\big)_{I_j^h} &= \sum_{\ell=1}^s \big(v, (z_\ell^{\varrho,s-1} \circ \xi_j^h)\big)_{I_j^h} \int_{\varrho_{\ell-1}}^{\varrho_\ell} z_i^{\varrho,s-1}(x)\, dx \\ &= \big(v, (z_i^{\varrho,s-1} \circ \xi_j^h)\big)_{I_j^h}, \quad i = 1, \ldots, s,\end{aligned}$$

which proves (3.10) since $\{z_i^{\varrho,s-1} \circ \xi_j^h\}_{i=1}^s$ is a basis of $\mathbb{P}^{s-1}(I_j^h)$. By (3.8) and (3.9) we have

$$\int_{I_j^h} |\widetilde{\Lambda}_h^{\varrho,s}\tilde{v} - \tilde{v}|^2\, dx = \sum_{\ell=1}^{s} \int_{I_{j,\ell}^{\varrho,h}} \left| \int_0^1 \left(\tilde{v}(x_{j-1}^h + z h_{I_j^h}) - \tilde{v}(x)\right) z_\ell^{\varrho,s-1}(z)\, dz \right|^2 dx$$
$$\leq \sum_{\ell=1}^{s} |I_{j,\ell}^{\varrho,h}| \left(\int_{I_j^h} |\widetilde{v}'|\, dx \right)^2 \left(\int_0^1 |z_\ell^{\varrho,s-1}|\, dz \right)^2.$$

Using Cauchy–Schwarz's inequality and the equality $|I_{j,\ell}^{\varrho,h}| = (\varrho_\ell - \varrho_{\ell-1}) h_{I_j^h}$ we obtain

$$\int_{I_j^h} |\widetilde{\Lambda}_h^{\varrho,s}\tilde{v} - \tilde{v}|^2\, dx \leq\ h_{I_j^h}^2\ \|\tilde{v}\|_{1,I_j^h}^2\ \left\{ \sum_{\ell=1}^{s} (\varrho_\ell - \varrho_{\ell-1})\ \left(\int_0^1 |z_\ell^{\varrho,s-1}|\, dz \right)^2 \right\},$$

which yields (3.11). ∎

Next, we show that it is possible to construct arbitrarily high order finite volume methods based on the projection on piecewise constants introduced in Proposition 3.5.

PROPOSITION 3.6. *Let $r \geq 3$, $\varrho = \{\varrho_i\}_{i=0}^{r-1}$ be the nodes of a partition of $[0,1]$, and*

$$\Lambda_h = \widetilde{\Lambda}_h^{\varrho,r-1}|_{\mathbb{P}_h^r}. \tag{3.12}$$

Then Λ_h satisfies (2.7) with $\alpha = 1$ and (2.8) with $\sigma = 0$. Also, the method (2.9) is a finite volume method with control volumes $\left\{\{(x_{j-1+\varrho_{i-1}}^h, x_{j-1+\varrho_i}^h)\}_{i=1}^{r-1}\right\}_{j=1}^{J_h}$ and $\{(x_{j-1+\varrho_{r-2}}^h, x_{j+\varrho_1}^h)\}_{j=1}^{J_h-1}$, where the interval $(x_{J_h-1+\varrho_{r-2}}^h, x_{J_h}^h)$ is a boundary control volume of both types I *and* II. *Moreover, the corresponding approximation u_h belongs to $S_h^r \cap C^1(\overline{\Omega})$ and satisfies a homogeneous Neumann boundary condition at x_R.*

Proof. Proposition 3.5, for $s = r-1$, yields that (2.7) holds with $\alpha = 1$ and (2.8) is satisfied with $\sigma = 0$. We shall show that the method (2.9) (under the choice (3.12) for Λ_h) is a finite volume method by constructing an appropriate basis of S_h^r.

Let $\theta_1,\ \theta_2 \in \{0,1\}$, and $\theta = (\theta_1, \theta_2)$. We consider the problem of finding a polynomial $p_\theta \in \mathbb{P}^r[0,1]$ such that $p_\theta(0) = \theta_1$, $p_\theta(1) = \theta_2$ and $\int_0^1 p_\theta(x)\ z_\ell^{\varrho,r-2}(x)\, dx = \varepsilon_\ell^\theta$ for $\ell = 1, \ldots, r-1$, where $\{\varepsilon_\ell^\theta\}_{\ell=1}^{r-1}$ are given real numbers. Using Lemma 3.4, the first two conditions for p_θ are equivalent to $p_\theta = w_\theta\, (\sum_{i=1}^{r-1} a_i^\theta\ z_i^{\varrho,r-2})$, where $w_\theta(x) = |x-\theta_1|\, |x-\theta_2|$. Hence, the conditions for p_θ are equivalent to $\sum_{i=1}^{r-1} a_i^\theta\ A_{\ell i}^\theta = \varepsilon_\ell^\theta$ for $\ell = 1, \ldots, r-1$, where $A_{\ell i}^\theta = \int_0^1 w_\theta(x)\, z_\ell^{\varrho,r-2}(x)\, z_i^{\varrho,r-2}(x)\, dx$. Since A^θ is invertible being the Gram matrix corresponding to the linearly independent vectors $\left\{\sqrt{w^\theta}\ z_k^{\varrho,r-2}\right\}_{k=1}^{r-1}$ of $L^2(0,1)$, there exists a unique $p_\theta \in \mathbb{P}^r[0,1]$ with the aforementioned properties.

For $i = 1, \ldots, r-1$, we denote by $\widetilde{p}_i$ the polynomial $p_{(0,0)}$ when $\varepsilon_\ell^{(0,0)} = \delta_{i\ell}$ for $\ell = 1, \ldots, r-1$. By $\overline{p}_*$ we denote the polynomial $p_{(1,0)}$, when $\varepsilon_1^{(1,0)} = 1$ and $\varepsilon_\ell^{(1,0)} = 0$ for $\ell = 2, \ldots, r-1$. In addition, we shall denote by $\widehat{p}_*$ the polynomial $p_{(0,1)}$, when $\varepsilon_\ell^{(0,1)} = 0$ for $\ell = 1, \ldots, r-2$, and $\varepsilon_{r-1}^{(0,1)} = 1$.

Now, we define $\Psi = \left\{\{\psi_{i,j}\}_{j=1}^{J_h}\right\}_{i=1}^{r} \subset S_h^r$ by $\psi_{i,j} = (\widetilde{p}_i \circ \xi_j^h)\ \mathcal{X}_{I_j^h}$ for $i = 1, \ldots, r-1$ and $j = 1, \ldots, J_h$, $\psi_{r,j} = (\widehat{p}_* \circ \xi_j^h)\ \mathcal{X}_{I_j^h} + (\overline{p}_* \circ \xi_{j+1}^h)\ \mathcal{X}_{I_{j+1}^h}$ for $j = 1, \ldots, J_h - 1$, and $\psi_{r,J_h} = (\widehat{p}_* \circ \xi_{J_h}^h)\ \mathcal{X}_{I_{J_h}^h}$. The definition of Λ_h, $\widehat{p}_*$, $\overline{p}_*$, and $\{\widetilde{p}_i\}_{i=1}^{r-1}$ yields $\Lambda_h \psi_{i,j} = \mathcal{X}_{I_{j,i}^{\varrho,h}}$ for $i = 1, \ldots, r-1$ and $j = 1, \ldots, J_h$, $\Lambda_h \psi_{r,j} = \mathcal{X}_{I_{j,r-1}^{\varrho,h} \cup I_{j+1,1}^{\varrho,h}}$ for $j = 1, \ldots, J_h - 1$, and $\Lambda_h \psi_{r,J_h} = \mathcal{X}_{I_{J_h,r-1}^{\varrho,h}}$. To prove that the elements of Ψ are linearly independent we

assume that there exist $\big\{\{\lambda_{i,j}\}_{j=1}^{J_h}\big\}_{i=1}^{r} \subset \mathbb{R}$ such that $\psi \equiv \sum_{i=1}^{r}\sum_{j=1}^{J_h} \lambda_{i,j}\,\psi_{i,j} = 0$. Evaluating ψ at $\{x_j^h\}_{j=1}^{J_h}$, we obtain $\lambda_{r,j} = 0$ for $j = 1,\dots,J_h$. Since $\Lambda_h\psi = 0$, we conclude that $\sum_{i=1}^{r-1}\sum_{j=1}^{J_h} \lambda_{i,j}\mathcal{X}_{I_{j,i}^{\varrho,h}} = 0$, which obviously yields $\lambda_{i,j} = 0$ for $i = 1,\dots,r-1$ and $j = 1,\dots,J_h$. Since $\dim(S_h^r) = rJ_h$, Ψ is a basis of S_h^r.

Setting $\chi = \psi_{i,j}$ in (2.9), we obtain

$$\int_{I_{j,i}^{\varrho,h}} L_{I_j^h} u_h \, dx = \int_{I_{j,i}^{\varrho,h}} f \, dx, \quad i = 1,\dots,r-1, \quad j = 1,\dots,J_h, \tag{3.13}$$

$$\begin{aligned}(au_h')(x_j^h - 0) - (au_h')(x_j^h + 0) + \int_{I_{j,r-1}^{\varrho,h}} (L_{I_j^h} u_h - f)\, dx \\ + \int_{I_{j+1,1}^{\varrho,h}} (L_{I_{j+1}^h} u_h - f)\, dx = 0, \quad j = 1,\dots,J_h - 1,\end{aligned} \tag{3.14}$$

$$(au_h')(x_{J_h}^h - 0) + \int_{I_{J_h,r-1}^{\varrho,h}} (L_{I_{J_h}^h} u_h - f)\, dx = 0. \tag{3.15}$$

Integration by parts, (3.14), and (3.15) yield

$$\mathcal{F}\big(u_h, (x_{j-1+\varrho_{r-2}}^h, x_{j+\varrho_1}^h)\big) = 0,\ j = 1,\dots,J_h - 1, \quad \mathcal{G}\big(u_h, I_{J_h,r-1}^{\varrho,h}\big) = 0. \tag{3.16}$$

The relations (3.13) and (3.16) show that the method (2.9) with Λ_h given by (3.12) is a finite volume method, the control volumes of which are $\{(x_{j-1+\varrho_{r-2}}^h, x_{j+\varrho_1}^h)\}_{j=1}^{J_h-1}$ and $\big\{\{I_{j,i}^{\varrho,h}\}_{j=1}^{J_h}\big\}_{i=1}^{r-1}$. We note that the interval $I_{J_h,r-1}^{\varrho,h}$ is a boundary control volume both of types I and II. Also, we combine (3.13), (3.14), and (3.15) to get $u_h'(x_R - 0) = 0$ and $u_h'(x_j^h - 0) = u_h'(x_j^h + 0)$ for $j = 1,\dots,J_h - 1$, which completes the proof. ∎

We close this subsection by discussing the possibility of constructing finite volume methods of the form (2.9) with $\alpha = 1$, $\sigma = 1$.

PROPOSITION 3.7. *Let $m \in \mathbb{N}$, $r = 2m$, $\varrho = \{\varrho_i\}_{i=0}^{r}$ be the nodes of a partition of $[0,1]$, and*

$$\Lambda_h = \widetilde{\Lambda}_h^{\varrho,r}|_{\mathbb{P}_h^r}. \tag{3.17}$$

Then Λ_h satisfies (2.7) *with $\alpha = 1$ and* (2.8) *with $\sigma = 1$. Also, the method* (2.9) *is a locally conservative method with control volumes $\{I_j^h\}_{j=1}^{J_h}$, where $I_{J_h}^h$ is a boundary control volume of type* I. *Moreover, if $r \in \{2,4,6\}$ and*

$$\{\varrho_j\}_{j=1}^{r-1} \subset \{z \in (0,1) : W_r(z) = 0\}, \tag{3.18}$$

then the method (2.9) *is a finite volume method with rJ_h overlapping control volumes:*

$$\{I_j^h\}_{j=1}^{J_h}, \quad \big\{\{(x_{j-1+\varrho_{r-i}}^h,\ x_{j+\varrho_i}^h)\}_{i=1}^{r-1}\big\}_{j=1}^{J_h-1}, \quad \{(x_{J_h-1+\varrho_{r-i}}^h,\ x_{J_h}^h)\}_{i=1}^{r-1}, \tag{3.19}$$

where the latter intervals are boundary control volumes of type II *and $I_{J_h}^h$ is a boundary control volume of type* I.

Proof. Proposition 3.5, for $s = r$, yields (2.7) with $\alpha = 1$ and (2.8) is satisfied with $\sigma = 1$. Proceeding as in the proof of Proposition 3.6, we conclude that for $\theta_0 \in \{0,1\}$ there is only one polynomial $\overline{p}_{\theta_0} \in \mathbb{P}^r(0,1)$ such that $\overline{p}_{\theta_0}(\theta_0) = 0$ and

$\int_0^1 \overline{p}_{\theta_0}(x)\ z_\ell^{\varrho,r-1}(x)\ dx = \overline{\varepsilon}_\ell$, $\ell = 1,\dots,r$, where $\{\overline{\varepsilon}_\ell\}_{\ell=1}^r$ are given real numbers. For $i = 1,\dots,r-1$, we set $\widetilde{p}_i = \overline{p}_0$ when $\overline{\varepsilon}_\ell = 0$ for $\ell = 1,\dots,r-i$ and $\overline{\varepsilon}_\ell = 1$ for $\ell = r-i+1,\dots,r$, and $\widetilde{q}_i = \overline{p}_1$ when $\overline{\varepsilon}_\ell = 1$ for $\ell = 1,\dots,i$ and $\overline{\varepsilon}_\ell = 0$ for $\ell = i+1,\dots,r$. We also set $\widetilde{p}_r = \overline{p}_0$ when $\overline{\varepsilon}_\ell = 1$ for $\ell = 1,\dots,r$.

Since r is even, using the orthogonality property of the Legendre polynomials and (3.8) we conclude that $\widetilde{p}_r = 1 - W_r$, so we have, in addition, that $\widetilde{p}_r(1) = 0$. Now we consider the linearly independent functions $\{\psi_j^r\}_{j=1}^{J_h} \subset S_h^r$ defined by $\psi_j^r = (\widetilde{p}_r \circ \xi_j^h)\ \mathcal{X}_{I_j^h}$ for $j = 1,\dots,J_h$. By construction we get $\Lambda_h\psi_j^r = \mathcal{X}_{I_j^h}$ for $j = 1,\dots,J_h$. For $\chi = \psi_j^r$, (2.9) yields (3.5), so the method is locally conservative with control volumes $\{I_j^h\}_{j=1}^{J_h}$, and the interval $I_{J_h}^h$ is a boundary control volume of type I.

Let $r \in \{2,4,6\}$ and $i \in \{1,\dots,r-1\}$. A long but straightforward calculation reveals that

$$\widetilde{p}_i(1) = 1 - W_r(\varrho_{r-i}) \quad \text{and} \quad \widetilde{q}_i(0) = 1 - W_r(\varrho_i). \tag{3.20}$$

Since (3.18) holds we finally obtain that $\widetilde{p}_i(1) = \widetilde{q}_i(0) = 1$. Now, we define $\{\psi_j^i\}_{j=1}^{J_h} \subset S_h^r$ by $\psi_j^i = (\widetilde{p}_i \circ \xi_j^h)\ \mathcal{X}_{I_j^h} + (\widetilde{q}_i \circ \xi_{j+1}^h)\ \mathcal{X}_{I_{j+1}^h}$ for $j = 1,\dots,J_h-1$ and $\psi_{J_h}^i = (\widetilde{p}_i \circ \xi_{J_h}^h)\ \mathcal{X}_{I_{J_h}^h}$. Then, we have $\Lambda_h\psi_j^i = \mathcal{X}_{(x^h_{j-1+\varrho_{r-i}},\, x^h_{j+\varrho_i})}$ for $j = 1,\dots,J_h-1$, and $\Lambda_h\psi_{J_h}^i = \mathcal{X}_{(x^h_{J_h-1+\varrho_{r-i}},\, x^h_{J_h})}$. Setting $\chi = \psi_j^i$ in (2.9) we get

$$\begin{gathered}\mathcal{F}\big(u_h,(x^h_{j-1+\varrho_{r-i}},x^h_{j+\varrho_i})\big) = 0, \quad j = 1,\dots,J_h-1,\\ \mathcal{G}\big(u_h,(x^h_{J_h-1+\varrho_{r-i}},x^h_{J_h})\big) = 0.\end{gathered} \tag{3.21}$$

The last step in our proof is to show that the set $\Psi = \big\{\{\psi_j^i\}_{i=1}^r\big\}_{j=1}^{J_h} \subset S_h^r$ consists of linearly independent functions. Let $\big\{\{\lambda_j^i\}_{i=1}^r\big\}_{j=1}^{J_h} \subset \mathbb{R}$ such that $\psi \equiv \sum_{i=1}^r \sum_{j=1}^{J_h} \lambda_j^i\ \psi_j^i = 0$. Evaluating ψ at $\{x_j^h\}_{j=1}^{J_h}$ we get

$$\sum_{i=1}^{r-1} \lambda_j^i = 0, \quad j = 1,\dots,J_h. \tag{3.22}$$

Since $\Lambda_h\psi\big|_{I_{1,1}^{\varrho,h}} = \lambda_1^r$ and $\Lambda_h\psi\big|_{I_{j,1}^{\varrho,h}} = \lambda_j^r + \sum_{i=1}^{r-1}\lambda_{j-1}^i$ for $j = 2,\dots,J_h$, using that $\Lambda_h\psi = 0$ and the result (3.22), we obtain

$$\lambda_j^r = 0, \quad j = 1,\dots,J_h. \tag{3.23}$$

Using (3.23), we have

$$\Lambda_h\psi\big|_{I_{1,i+1}^{\varrho,h}} = \sum_{m=r-i}^{r-1} \lambda_1^m \quad \text{and} \quad \Lambda_h\psi\big|_{I_{j,i+1}^{\varrho,h}} = \sum_{m=r-i}^{r-1} \lambda_j^m + \sum_{m'=i+1}^{r-1} \lambda_{j-1}^{m'} \tag{3.24}$$

for $i = 1,\dots,r-1$ and $j = 2,\dots,J_h$. Since $\Lambda_h\psi = 0$, applying an induction argument on (3.24) we conclude that

$$\lambda_j^i = 0, \quad i = 1,\dots,r-1, \quad j = 1,\dots,J_h. \tag{3.25}$$

The relations (3.25) and (3.23) complete the independence proof. Hence, Ψ is a basis of S_h^r. Finally, (3.5) and (3.21) yield that (2.9) (under the choice (3.17) for Λ_h) is a finite volume method with the control volumes defined in (3.19). ∎

REMARK 3.5. *Let $r \in \{3,5\}$ and Λ_h be defined by* (3.17)*. Then, by Proposition* 3.5*, we have $\alpha = 1$ and $\sigma = 1$. Since r is odd, we have $\widetilde{p}_r = 1 + W_r$ and hence $\widetilde{p}_r(1) = 2 \notin \{0,1\}$. Moreover, we get*

$$\widetilde{p}_i(1) = 1 - W_r(\varrho_{r-i}) \quad \text{and} \quad \widetilde{q}_i(0) = 1 + W_r(\varrho_i), \quad i = 1, \dots, r-1. \tag{3.26}$$

Assuming that (3.18) *holds,* (3.26) *yields $\widetilde{p}_i(1) = \widetilde{q}_i(0) = 1$ for $i = 1, \dots, r-1$. Therefore, proceeding as in the proof of Proposition* 3.7 *we conclude that in the case above the method* (2.9) *is a locally conservative method with control volumes $\big\{\{(x^h_{j-1+\varrho_{r-i}},\, x^h_{j+\varrho_i})\}_{i=1}^{r-1}\big\}_{j=1}^{J_h-1}$ and $\{(x^h_{J_h-1+\varrho_{r-i}},\, x^h_{J_h})\}_{i=1}^{r-1}$, which are boundary control volumes of type* II*.*

REMARK 3.6. *It is our opinion that* (3.20) *is true for all even r and that* (3.26) *is true for all odd r, but we are not able to provide a general proof of this fact.*

4. A priori estimates in the L^2 and H^1 norm. Our aim in this section is to derive a priori error estimates, in the usual Sobolev norms, for the method (2.9) under the assumptions (2.7) and (2.8). In particular, in Theorem 4.6 we shall show that (i) in the H^1 norm the order of convergence is always optimal and (ii) in the L^2 norm the order of convergence is optimal when $\alpha \geq 2$, or $\alpha = 1$ and $\sigma = 1$.

The following lemmata will be used throughout the remainder of this paper.

LEMMA 4.1. *Let $r \geq 2$, $j \in \{0, \sigma\}$, and assume that Λ_h satisfies* (2.7) *and* (2.8)*. Then $\forall\, h \in (0,1)$, $v \in H_h^{r+1+j}$ and $p \in \mathbb{P}_h^r$, we have*

$$\big|B_h(v - \mathcal{I}_h^\ell v, p) - B(v - \mathcal{I}_h^\ell v, p)\big| \leq C\, h^{s+\ell-1+j}\, \|v\|_{\ell+1+j,h}\, \|p\|_{s,h}$$

for $s = 1, \dots, \alpha$ and $\ell = 1, \dots, r$.

Proof. Let $h \in (0,1)$, $v \in H_h^{r+1+j}$, $p \in \mathbb{P}_h^r$, $s \in \{1, \dots, \alpha\}$ and $\ell \in \{1, \dots, r\}$. Also, for simplicity we set $g = v - \mathcal{I}_h^\ell v$. From (2.11) we have

$$B_h(g,p) - B(g,p) = G_{_A} + G_{_B}, \tag{4.1}$$

where $G_{_A} = -\sum_{I \in \mathcal{P}_h} (ag'', \Lambda_h p - p)_I$ and $G_{_B} = \sum_{I \in \mathcal{P}_h} \big((\beta - a')g' + \gamma g, \Lambda_h p - p\big)_I$. Using (2.8), (2.6), (2.2), and (2.7) it follows that

$$\begin{aligned}
|G_{_A}| &= \Bigg| \sum_{I \in \mathcal{P}_h} \int_I g''\, (a - \Pi_h a)\, (\Lambda_h p - p)\, dx \\
&\qquad - \sum_{I \in \mathcal{P}_h} \int_I \Pi_h a\, (v'' - \mathcal{I}_h^{\ell-2+j} v'')\, (\Lambda_h p - p)\, dx \Bigg| \\
&\leq C \sum_{I \in \mathcal{P}_h} h_I^s \left(h_I\, \|g''\|_{0,I} + h_I^{\ell-1+j}\, \|v''\|_{\ell-1+j,I}\right) \|p\|_{s,I} \\
&\leq C \sum_{I \in \mathcal{P}_h} h_I^s \left(h_I^\ell\, \|v\|_{\ell+1,I} + h_I^{\ell-1+j}\, \|v\|_{\ell+1+j,I}\right) \|p\|_{s,I} \\
&\leq C \sum_{I \in \mathcal{P}_h} h_I^{s+\ell-1+j}\, \|v\|_{\ell+1+j,I}\, \|p\|_{s,I} \\
&\leq C\, h^{s+\ell-1+j}\, \|v\|_{\ell+1+j,h}\, \|p\|_{s,h}.
\end{aligned}$$

To estimate G_B we use (2.7) and (2.2) to obtain

$$\begin{aligned}|G_B| &\le C \sum_{I\in\mathcal{P}_h} h_I^s \, \|g\|_{1,I} \, \|p\|_{s,I} \\ &\le C \sum_{I\in\mathcal{P}_h} h_I^{\ell+s} \, \|v\|_{\ell+1,I} \, \|p\|_{s,I} \\ &\le C \, h^{\ell+s} \, \|v\|_{\ell+1,h} \, \|p\|_{s,h}.\end{aligned}$$

The lemma now follows from (4.1) and the estimates for G_A and G_B above. ∎

LEMMA 4.2. *Let $r \ge 2$ and assume that Λ_h satisfies* (2.7) *and* (2.8). *Then $\forall\, h \in (0,1)$ we have*

$$\big|B_h(\phi,p) - B(\phi,p)\big| \le C \, h^{s+\sigma} \, \|\phi\|_{1,h} \, \|p\|_{s,h}, \quad s = 1,\dots,\alpha, \quad \forall\, \phi, p \in \mathbb{P}_h^r.$$

Proof. Let $h \in (0,1)$, $s \in \{1,\dots,\alpha\}$ and $\phi, p \in \mathbb{P}_h^r$. From (2.11) we have

$$B_h(\phi,p) - B(\phi,p) = E_A + E_B + E_C, \tag{4.2}$$

where $E_A = \sum_{I\in\mathcal{P}_h} \big((\beta - a')\phi', \Lambda_h p - p\big)_I$, $E_B = \sum_{I\in\mathcal{P}_h} \big(\gamma\phi, \Lambda_h p - p\big)_I$ and $E_C = -\sum_{I\in\mathcal{P}_h} (a\phi'', \Lambda_h p - p)_I$. If $\sigma = 0$ in (2.8), then using (2.7) it follows that

$$\begin{aligned}|E_A| + |E_B| &\le C \sum_{I\in\mathcal{P}_h} h_I^s \, \|\phi\|_{1,I} \, \|p\|_{s,I} \\ &\le C \, h^s \, \|\phi\|_{1,h} \, \|p\|_{s,h}.\end{aligned}$$

If $\sigma = 1$ in (2.8), then using (2.8), (2.6), (2.5), and (2.7), it follows that

$$\begin{aligned}|E_A| &= \left| \sum_{I\in\mathcal{P}_h} \int_I \phi' \, \big((\beta - a') - \Pi_h(\beta - a')\big) \, (\Lambda_h p - p) \, dx \right| \\ &\le C \sum_{I\in\mathcal{P}_h} h_I^{s+1} \, \|\phi\|_{1,I} \, \|p\|_{s,I} \\ &\le C \, h^{s+1} \, \|\phi\|_{1,h} \, \|p\|_{s,h}\end{aligned}$$

and

$$\begin{aligned}|E_B| &= \left| \sum_{I\in\mathcal{P}_h} \int_I \phi \, \big(\gamma - \Pi_h\gamma\big) \, (\Lambda_h p - p) \, dx + \sum_{I\in\mathcal{P}_h} \int_I \Pi_h\gamma \, (\phi - \Pi_h\phi) \, (\Lambda_h p - p) \, dx \right| \\ &\le C \sum_{I\in\mathcal{P}_h} h_I^{s+1} \, \|\phi\|_{1,I} \, \|p\|_{s,I} \\ &\le C \, h^{s+1} \, \|\phi\|_{1,h} \, \|p\|_{s,h}.\end{aligned}$$

To estimate E_C we use (2.8), (2.2) or (2.3), (2.1), and (2.7) to obtain

$$\begin{aligned}|E_C| &= \left| \sum_{I\in\mathcal{P}_h} \int_I \phi''(a - \mathcal{I}_h^* a)(\Lambda_h p - p) \, dx \right| \\ &\le C \sum_{I\in\mathcal{P}_h} h_I^{\sigma+1} \, \|\phi''\|_{0,I} \, \|\Lambda_h p - p\|_{0,I} \\ &\le C \sum_{I\in\mathcal{P}_h} h_I^{\sigma} \, \|\phi\|_{1,I} \, \|\Lambda_h p - p\|_{0,I} \\ &\le C \, h^{s+\sigma} \, \|\phi\|_{1,h} \, \|p\|_{s,h},\end{aligned}$$

where $\mathcal{I}_h^* = \Pi_h$ when $\sigma = 0$, and $\mathcal{I}_h^* = \mathcal{I}_h^1$ when $\sigma = 1$. The lemma now follows from (4.2) and the estimates for E_A, E_B, and E_C above. ∎

Next we show that B_h, like B, satisfies a Gårding-type inequality.

LEMMA 4.3. *Let $r \geq 2$ and assume that Λ_h satisfies (2.7) and (2.8). Then, there exists $\widetilde{h}_0 \in (0,1)$ and constants $C_G \geq 0$, $C_E > 0$, such that*

$$B_h(\phi,\phi) + C_G \|\phi\|_0^2 \geq C_E \|\phi\|_1^2 \quad \forall\, \phi \in S_h^r, \quad \forall\, h \in (0, \widetilde{h}_0). \tag{4.3}$$

In particular, if $\beta = \gamma = 0$, then $C_G = 0$.

Proof. Let $h \in (0,1)$ and $\phi \in S_h^r$. It is well known that there exist constants $C_\Gamma \geq 0$ and $C_\Delta > 0$, independent of ϕ, such that

$$B(\phi,\phi) + C_\Gamma \|\phi\|_0^2 \geq C_\Delta \|\phi\|_1^2. \tag{4.4}$$

Lemma 4.2 now gives $\big|B_h(\phi,\phi) - B(\phi,\phi)\big| \leq C_Z\, h^{\sigma+1} \|\phi\|_1^2$, which along with (4.4) yields

$$B_h(\phi,\phi) + C_\Gamma \|\phi\|_0^2 \geq (C_\Delta - h^{\sigma+1}\, C_Z)\, \|\phi\|_1^2.$$

If $C_Z = 0$, then (4.3) holds with $\widetilde{h}_0 = 1$, $C_G = C_\Gamma$ $C_E = C_\Delta$. If $C_Z > 0$, then (4.3) holds, for example, with $\widetilde{h}_0 = \min\{1, (\frac{C_\Delta}{2C_Z})^{\frac{1}{\sigma+1}}\}$, $C_G = C_\Gamma$, $C_E = \frac{C_\Delta}{2}$. If $\beta = \gamma = 0$, then $C_\Gamma = 0$ from the Poincaré–Friedrichs inequality. Consequently, $C_G = 0$. ∎

REMARK 4.1. *For later use we note that if $\beta = \gamma = 0$, then*

$$B_h(\phi,\phi) + C_Z\, h^{\sigma+1}\, \|\phi\|_1^2 \geq a_*\, \|\phi'\|_0^2 \quad \forall\, \phi \in \mathbb{P}_h^r, \quad \forall\, h \in (0,1).$$

As in the finite element analysis (cf. [40]), we relate the approximation error in the L^2 norm to that in the H^1 norm.

PROPOSITION 4.4. *Let $r \geq 2$, $m \in \{0, \sigma\}$, and assume that Λ_h satisfies (2.7) and (2.8). If $u \in H^{r+1+m}(\Omega) \cap \mathcal{H}(\Omega)$, then there exists $\widehat{h}_0 \in (0, \widetilde{h}_0)$ such that for $h \in (0, \widehat{h}_0)$ the method (2.9) is well defined and*

$$\|u - u_h\|_0 \leq C\, h\, \|u - u_h\|_1 + C\, h^{r+m+\alpha^*-1}\, \|u\|_{r+1+m}, \tag{4.5}$$

where $\alpha^* = \min\{\alpha, 2\}$ and $\widetilde{h}_0$ is the constant specified in Lemma 4.3.

Proof. Let $h \in (0, \widetilde{h}_0)$ and $w \in H^2(\Omega)$ be the solution of the dual problem

$$L^* w \equiv -(a(x)w')' - \beta(x)w' + (\gamma(x) - \beta'(x))w = \psi(x) \quad \forall\, x \in \Omega, \tag{4.6a}$$

$$w(x_L) = 0, \quad a(x_R)w'(x_R) + \beta(x_R)w(x_R) = 0, \tag{4.6b}$$

with $\psi \in L^2(\Omega)$, [28]. We let T^* denote the solution operator of (4.6), so that $w = T^*\psi$. Under the assumptions on the data of (1.1) stated in section 1.1, $w \in H^2(\Omega)$ and elliptic regularity yields

$$\|w\|_2 \leq C_R\, \|\psi\|_0 \tag{4.7}$$

for some constant C_R which depends only on the domain Ω and the coefficients of L^*. Taking the $L^2(\Omega)$-inner product of (4.6a) with $\psi \in \mathcal{H}(\Omega)$ and using (2.2) we have

$$\begin{aligned} \|\psi\|_0^2 &= B(\psi, w - \mathcal{I}_h^1 w) + B(\psi, \mathcal{I}_h^1 w) \\ &\leq C\, h\, \|\psi\|_1 \|w\|_2 + B(\psi, \mathcal{I}_h^1 w), \end{aligned}$$

which, along with (4.7), yields

$$\|\psi\|_0^2 \leq C\, h\, \|\psi\|_1 \|\psi\|_0 + B(\psi, \mathcal{I}_h^1 w). \tag{4.8}$$

We shall first show that the method (2.9) is well defined. Obviously, (2.9) is equivalent to a linear system of algebraic equations with matrix M with $M_{ij} = B_h(\phi_j, \phi_i)$, $i, j = 1, \ldots, N_h$, where $N_h = \dim(S_h^r)$ and $\{\phi_j\}_{j=1}^{N_h}$ is a basis of S_h^r. If M were not invertible there would exist a nonzero $\phi_* \in S_h^r$ such that

$$B_h(\phi_*, \chi) = 0 \quad \forall\, \chi \in S_h^r. \tag{4.9}$$

Then, from (4.3) we have

$$\|\phi_*\|_1 \le C\, \|\phi_*\|_0. \tag{4.10}$$

Choosing $\psi = \phi_*$ in (4.6) and using (4.8) and (4.10) it follows that

$$(1 - Ch)\|\phi_*\|_0^2 \le B(\phi_*, \mathcal{I}_h^1 w), \tag{4.11}$$

where $w = T^*\phi_*$. Also, (4.9) yields $B(\phi_*, \mathcal{I}_h^1 w) = B(\phi_*, \mathcal{I}_h^1 w) - B_h(\phi_*, \mathcal{I}_h^1 w)$. Thus, from Lemma 4.2 with $s = 1$, (2.2), and (4.7) we have

$$\begin{aligned} B(\phi_*, \mathcal{I}_h^1 w) &\le C\, h^{1+\sigma}\, \|\phi_*\|_1\, \|\mathcal{I}_h^1 w\|_1 \\ &\le C\, h^{1+\sigma}\, \|\phi_*\|_1\, \|w\|_2 \\ &\le C\, h^{1+\sigma}\, \|\phi_*\|_1\, \|\phi_*\|_0, \end{aligned}$$

which along with (4.10) yields

$$B(\phi_*, \mathcal{I}_h^1 w) \le C\, h^{1+\sigma}\, \|\phi_*\|_0^2. \tag{4.12}$$

From (4.11) and (4.12) we now have $(1 - C_* h)\|\phi_*\|_0^2 \le 0$. We set $\widehat{h}_0 = \widetilde{h}_0$ when $C_* = 0$ and $\widehat{h}_0 = \min\{\widetilde{h}_0, \frac{1}{2C_*}\}$ otherwise. If $h \in (0, \widehat{h}_0)$, then $\phi_* = 0$, which is a contradiction; consequently, M is invertible and (2.9) is well defined.

To prove (4.5), let $h \in (0, \widehat{h}_0)$ and write $e = u_h - u = (u_h - \mathcal{I}_h^r u) + (\mathcal{I}_h^r u - u) = \theta_h + \eta$. From (2.10) and (4.6a) with $\psi = e$, so that $w = T^* e$, we have

$$\begin{aligned} B(e, \mathcal{I}_h^1 w) &= B(e, \mathcal{I}_h^1 w) - B_h(e, \mathcal{I}_h^1 w) \\ &= D_A + D_B, \end{aligned} \tag{4.13}$$

where $D_A = B(\eta, \mathcal{I}_h^1 w) - B_h(\eta, \mathcal{I}_h^1 w)$ and $D_B = B(\theta_h, \mathcal{I}_h^1 w) - B_h(\theta_h, \mathcal{I}_h^1 w)$. Using Lemma 4.1 (with $s = \alpha^*$, $\ell = r$, and $j = m$), (2.2), and (4.7) it follows that

$$\begin{aligned} |D_A| &\le C\, h^{r+m+\alpha^*-1}\, \|u\|_{r+1+m}\, \|\mathcal{I}_h^1 w\|_1 \\ &\le C\, h^{r+m+\alpha^*-1}\, \|u\|_{r+1+m}\, \|w\|_2 \\ &\le C\, h^{r+m+\alpha^*-1}\, \|u\|_{r+1+m}\, \|e\|_0. \end{aligned} \tag{4.14}$$

To estimate D_B we note that Lemma 4.2 with $s = \alpha^*$ yields

$$|D_B| \le C\, h^{\alpha^*+\sigma}\, \|\theta_h\|_1\, \|\mathcal{I}_h^1 w\|_1,$$

so again using (2.2) and (4.7) we obtain

$$|D_B| \le C\, h^{\alpha^*+\sigma}\, (h^r \|u\|_{r+1} + \|e\|_1)\, \|e\|_0. \tag{4.15}$$

The estimate (4.5) follows by combining (4.8) with $\psi = e$, (4.13), (4.14), and (4.15). ∎

PROPOSITION 4.5. *Let $r \geq 2$, $\widehat{h}_0$ be the constant specified in* Proposition 4.4*, and assume that Λ_h satisfies* (2.7) *and* (2.8)*. If $u \in H^{r+1}(\Omega) \cap \mathcal{H}(\Omega)$, then*

$$\|u_h - \mathcal{I}_h^r u\|_1^2 \leq C\, h^r\, \|u\|_{r+1}\, \|u_h - \mathcal{I}_h^r u\|_1 + \tfrac{C_G}{C_E}\, \|u_h - \mathcal{I}_h^r u\|_0^2 \quad \forall\, h \in (0, \widehat{h}_0). \tag{4.16}$$

Proof. Let $h \in (0, \widehat{h}_0)$ and, as before, $\theta_h = u_h - \mathcal{I}_h^r u$, $\eta = \mathcal{I}_h^r u - u$. Using (2.10) and Lemma 4.1 with $\ell = r$, $s = 1$, and $j = 0$, we obtain

$$\begin{aligned} B_h(\theta_h, \theta_h) &= -B_h(\eta, \theta_h) \\ &= \big[B(\eta, \theta_h) - B_h(\eta, \theta_h)\big] - B(\eta, \theta_h) \\ &\leq C\, \left(h^r\, \|u\|_{r+1} + \|\eta\|_1\right)\, \|\theta_h\|_1. \end{aligned}$$

The interpolation estimate (2.2) and the above relation imply

$$B_h(\theta_h, \theta_h) \leq C\, h^r\, \|u\|_{r+1}\, \|\theta_h\|_1. \tag{4.17}$$

The proposition now follows by combining (4.3) and (4.17). ∎

We are now ready to prove the main error estimate of this section.

THEOREM 4.6. *Let $r \geq 2$, $\widehat{h}_0$ be the constant specified in Proposition* 4.4*, and assume that Λ_h satisfies* (2.7) *and* (2.8)*. Also, let $\overline{m}_\alpha = 0$ when $\alpha \geq 2$, and $\overline{m}_\alpha = \sigma$ when $\alpha = 1$. If $u \in H^{r+1+\overline{m}_\alpha}(\Omega) \cap \mathcal{H}(\Omega)$, then there exists $h_0 \in (0, \widehat{h}_0)$ such that $\forall\, h \in (0, h_0)$,*

$$\|u - u_h\|_1 \leq C\, h^r \|u\|_{r+1}, \tag{4.18}$$

$$\|u - u_h\|_0 \leq C\, h^{r+\min\{\overline{m}_\alpha + \alpha - 1, 1\}}\, \|u\|_{r+1+\overline{m}_\alpha}. \tag{4.19}$$

Proof. Let $h \in (0, \widehat{h}_0)$ and $e = u - u_h$. Combining (4.16), (4.5), and (2.2) we obtain $\|e\|_1^2 \leq C_1\, h^{2r}\, \|u\|_{r+1}^2 + C_2\, h^2\, \|e\|_1^2 + \frac{1}{2}\|e\|_1^2$, or equivalently

$$\left(\frac{1}{2} - C_2\, h^2\right)\, \|e\|_1^2 \leq C_1\, h^{2r}\, \|u\|_{r+1}^2.$$

The estimate (4.18) follows from the above relation for h sufficiently small. The estimate (4.19) follows from (4.18) and (4.5). ∎

5. A priori estimates in the L^∞ norm. This section is devoted to the derivation of a priori error estimates in the L^∞ norm for the method (2.9), under the assumptions (2.7) and (2.8). Since we work in one space dimension, the L^∞ norm is dominated by the H^1 norm. Hence, when $\alpha = 1$ and $\sigma = 0$, using Theorem 4.6 we obtain the error estimate $|u - u_h|_\infty \leq C\, h^r\, \|u\|_{r+1}$, the suboptimal order of which has been observed numerically for the method of Proposition 3.1 with $r = 2$ (see Table 8.4 in section 8). For this reason, in the rest of this section we shall examine the following two cases: (i) $\alpha \geq 2$ and Λ_h is identity on $\mathbb{P}_h^1$, and (ii) $\alpha = 1$ and $\sigma = 1$, where, according to Theorem 4.6, the order of convergence in the L^2 norm is optimal.

5.1. Maximum norm estimates when $\alpha \geq 2$. In this subsection we shall assume that (2.7) holds with $\alpha \geq 2$. Moreover, we shall make the hypothesis that the operator $\Lambda_h : \mathbb{P}_h^r \to L^2(\Omega)$ has the property

$$\Lambda_h p = p \quad \forall\, p \in \mathbb{P}_h^1, \tag{5.1}$$

which is obviously satisfied for the methods of Proposition 3.1 for $r \geq 3$, and the methods of Proposition 3.2. In the analysis below, we follow the steps of the L^∞ estimate of [44] for the standard Galerkin finite element method.

For an open interval $I \subset \Omega$ and $v \in H^2(I)$ we let $\mathcal{L}_I v = -(av')'$, where a is the diffusion coefficient of the problem (1.1). For $h \in (0,1)$ and $r \geq 2$, we define the bilinear form $\widetilde{B}_h : H_h^2 \times \mathbb{P}_h^r \to \mathbb{R}$ by

$$\widetilde{B}_h(v,\chi) = \sum_{I \in \mathcal{P}_h} \left\{ [\![a v' \chi]\!]_{\partial I} + (\mathcal{L}_I v, \Lambda_h \chi)_I \right\} \quad \forall\, v \in H_h^2, \quad \forall\, \chi \in \mathbb{P}_h^r. \tag{5.2}$$

From Lemma 4.3 there exists $\overline{h}_0 \in (0,1)$ such that

$$\widetilde{B}_h(\phi,\phi) \geq C_E^* \, \|\phi\|_1^2 \quad \forall\, \phi \in S_h^r, \quad \forall\, h \in (0, \overline{h}_0). \tag{5.3}$$

For $h \in (0,\overline{h}_0)$, we introduce a Ritz projection operator $R_h^r : H_h^2 \to S_h^r$ by

$$\widetilde{B}_h(R_h^r v - v, \chi) = 0 \quad \forall\, \chi \in S_h^r, \quad \forall\, v \in H_h^2, \tag{5.4}$$

which in view of (5.3) is well defined. Moreover, when Λ_h satisfies (2.8) and (2.7) with $\alpha \geq 2$, the orthogonality property (5.4) and the analysis of section 4 yield

$$\|R_h^r v - v\|_0 + h \, \|R_h^r v - v\|_1 \leq C \, h^{r+1} \, \|v\|_{r+1} \tag{5.5}$$

$\forall\, h \in (0,\overline{h}_0)$ and $v \in H^{r+1}(\Omega) \cap \mathcal{H}(\Omega)$.

The first step towards obtaining a maximum norm error estimate is the estimation of the difference $R_h^r u - u$ at the nodes of the partition $\mathcal{P}_h$.

LEMMA 5.1. *Let $r \geq 2$ and assume that Λ_h satisfies* (2.8), (5.1), *and* (2.7) *with $\alpha \geq 2$. If $u \in H^{r+1}(\Omega) \cap \mathcal{H}(\Omega)$, then we have*

$$\max_{1 \leq j \leq J_h} \left| (R_h^r u - u)(x_j^h) \right| \leq C \, h^{r+1} \, \|u\|_{r+1} \quad \forall\, h \in (0,\overline{h}_0). \tag{5.6}$$

Proof. Let $h \in (0,\overline{h}_0)$. For $j \in \{1,\ldots,J_h\}$, we define the function

$$g_{x_j^h}(z) = \begin{cases} z - x_{\scriptscriptstyle L}, & z \leq x_j^h, \\ x_j^h - x_{\scriptscriptstyle L}, & z \geq x_j^h, \end{cases} \quad \forall\, z \in \overline{\Omega}, \tag{5.7}$$

which clearly belongs to $S_h^r \cap \mathbb{P}_h^1$. So with $\zeta = R_h^r u - u$, using the definition of R_h^r and (5.1), we obtain

$$\begin{aligned} 0 = \widetilde{B}_h(\zeta, g_{x_j^h}) &= \sum_{I \in \mathcal{P}_h} \left\{ [\![a \zeta' g_{x_j^h}]\!]_{\partial I} + (\mathcal{L}_I \zeta, g_{x_j^h})_I \right\} \\ &= (a\zeta', g'_{x_j^h}) \\ &= (a\zeta)(x_j^h) - \int_{x_{\scriptscriptstyle L}}^{x_j^h} a' \, \zeta \, dx, \quad j = 1,\ldots,J_h, \end{aligned}$$

and consequently, $\max_{1 \leq j \leq J_h} \left| \zeta(x_j^h) \right| \leq \frac{\|a'\|_0}{a_*} \, \|\zeta\|_0$. The proof of the lemma now follows from the latter relation and (5.5). ∎

Our next step is the estimation of the error $u_h - u$ at the right endpoint $x_{\scriptscriptstyle R}$ of Ω.

LEMMA 5.2. *Let $r \geq 2$, h_0 be the constant specified in* Theorem 4.6, *and assume that Λ_h satisfies* (2.8) *and* (2.7) *with $\alpha \geq 2$. If $u \in H^{r+1}(\Omega) \cap \mathcal{H}(\Omega)$, then*

$$\left| (u - u_h)(x_{\scriptscriptstyle R}) \right| \leq C \, h^{r+1} \, \|u\|_{r+1} \quad \forall\, h \in (0, h_0).$$

Proof. Consider the dual elliptic problem

$$-(a\ w')' - \beta\ w' + (\gamma_0 + \gamma - \beta')\ w = 0 \quad \text{on}\,\Omega, \tag{5.8a}$$

$$w(x_{_L}) = 0, \quad a(x_{_R})\ w'(x_{_R}) + \beta(x_{_R})\ w(x_{_R}) = 1, \tag{5.8b}$$

where γ_0 is a positive real number, sufficiently large to ensure the existence of an $H^2(\Omega)$-solution of the problem (cf. [1]). Note that this problem, in contrast with the dual problem (4.6) when $\psi = u_h - u$, is independent of h. Now taking the inner product of (5.8a) with the error, $e = u_h - u$, and using integration by parts we obtain

$$\begin{aligned} e(x_{_R}) &= B(e, w) + \gamma_0\ (e, w) \\ &= B(e, w - \mathcal{I}_h^1 w) + B(e, \mathcal{I}_h^1 w) + \gamma_0\ (e, w). \end{aligned} \tag{5.9}$$

First, using Theorem 4.6 and (2.2) we have

$$\big|B(e, w - \mathcal{I}_h^1 w)\big| \le\ C\ h^{r+1}\ \|u\|_{r+1}\ \|w\|_2, \tag{5.10}$$

$$\big|\gamma_0\ (e, w)\big| \le\ C\ h^{r+1}\ \|u\|_{r+1}\ \|w\|_0. \tag{5.11}$$

The term $B(e, \mathcal{I}_h^1 w)$ may be estimated as the corresponding term in (4.13) using (2.10) and Lemmas 4.1 and 4.2, so that

$$\big|B(e, \mathcal{I}_h^1 w)\big| \le\ C\ \big(h^{r+1}\ \|u\|_{r+1} + h\ \|u_h - \mathcal{I}_h^r u\|_1\big)\ \|\mathcal{I}_h^1 w\|_1.$$

Using again Theorem 4.6 and (2.2) in the relation above we obtain

$$\big|B(e, \mathcal{I}_h^1 w)\big| \le C\ h^{r+1}\ \|u\|_{r+1}\ \|w\|_2. \tag{5.12}$$

The lemma now follows by combining (5.10), (5.11), and (5.12) with (5.9). ∎

Next we present an H^1-superconvergent estimate for the difference $R_h^r u - u_h$.

PROPOSITION 5.3. *Let $r \ge 2$. Assume that Λ_h satisfies* (2.8) *and* (2.7) *with $\alpha \ge 2$. If u belongs to $H^{r+1}(\Omega) \cap \mathcal{H}(\Omega)$, then, for h sufficiently small, we have*

$$\|R_h^r u - u_h\|_1 \le\ C\ h^{r+1}\ \|u\|_{r+1}. \tag{5.13}$$

Proof. Let $e = u_h - u$ and $\nu_h = R_h^r u - u_h$. From (5.4) and (2.10) it follows that

$$\widetilde{B}_h(\nu_h, \nu_h) = D_1 + D_2, \tag{5.14}$$

where $D_1 = -(\beta e' + \gamma e, \Lambda_h \nu_h - \nu_h)$ and $D_2 = -(\beta e' + \gamma e, \nu_h)$. From (2.7) and Theorem 4.6 we have

$$\begin{aligned} \big|D_1\big| &\le\ C\ \sum_{I \in \mathcal{P}_h} h_I\ \|e\|_{1,I}\ \|\nu_h\|_{1,I} \\ &\le\ C\ h^{r+1}\ \|u\|_{r+1}\ \|\nu_h\|_1. \end{aligned} \tag{5.15}$$

Using integration by parts, Lemma 5.2, and the Sobolev inequality, we obtain

$$\begin{aligned} \big|D_2\big| &\le\ C\ \|e\|_0\ \|\nu_h\|_0 + \big|(\beta\ e\ \nu_h)(x_{_R}) - (e, \beta'\ \nu_h + \beta\ \nu_h')\big| \\ &\le\ C\big(\ \|e\|_0\ \|\nu_h\|_1 + |e(x_{_R})|\ |\nu_h(x_{_R})|\ \big) \\ &\le\ C\ h^{r+1}\ \|u\|_{r+1}\ \|\nu_h\|_1. \end{aligned} \tag{5.16}$$

The claim of the proposition now follows by (5.14), (5.15), (5.16), and (5.3). ∎

For $h \in (0,1)$ and $I \in \mathcal{P}_h$, we define the local bilinear forms $\widetilde{B}_I : H^2(I) \times \mathbb{P}^r(I) \to \mathbb{R}$ and $B_I : H^1(I) \times H^1(I) \to \mathbb{R}$ by

$$\widetilde{B}_I(v,p) = [\![av'p]\!]_{\partial I} + (\mathcal{L}_I v, \Lambda_h p)_I \quad \forall\, v \in H^2(I), \quad \forall\, p \in \mathbb{P}^r(I),$$
$$B_I(v,p) = (av', p')_I \quad \forall\, v, p \in H^1(I).$$

Following the proof of Lemma 4.2 we have

$$\left|\widetilde{B}_I(\phi,p) - B_I(\phi,p)\right| \le C\ h_I\ \|\phi\|_{1,I}\ \|p\|_{1,I} \quad \forall\, \phi, p \in \mathbb{P}^r(I).$$

Moreover, proceeding as in the proof of Lemma 4.3 we conclude that there exists $h_\infty \in (0,1)$ such that

$$\widetilde{B}_I(\phi,\phi) \ge C\ \|\phi\|^2_{1,I} \quad \forall\, \phi \in \mathbb{P}^r_0(I), \quad \forall\, h \in (0, h_\infty), \tag{5.17}$$

where $\mathbb{P}^r_0(I)$ is the subset of $\mathbb{P}^r(I)$ consisting of functions which vanish at the endpoints of I. Next, for $h \in (0, h_\infty)$ we define the operator $Y_I : H^1_0(I) \to \mathbb{P}^r_0(I)$ by

$$\widetilde{B}_I(Y_I v, p) = \widetilde{B}_I(v,p) \quad \forall\, p \in \mathbb{P}^r_0(I), \quad \forall\, v \in H^1_0(I).$$

In view of (5.17), the operator Y_I is well defined. We shall also make use of the operator $\mu_I : H^1(I) \to \mathbb{P}^r(I)$, defined by $\mu_I v := Y_I(v - \mathcal{I}^1_h v) + \mathcal{I}^1_h v \ \ \forall\, v \in H^1(I)$. The next proposition establishes an approximation property of μ_I in the L^∞ norm.

PROPOSITION 5.4. *Let $r \ge 2$ and assume that Λ_h satisfies* (2.7) *and* (2.8). *If $v \in W^{r+1,\infty}(\Omega)$, then we have*

$$|\mu_I v - v|_{\infty,I} \le\ C\ h_I^{r+1}\ \|v\|_{r+1,\infty} \quad \forall\, I \in \mathcal{P}_h, \quad \forall\, h \in (0, h_\infty). \tag{5.18}$$

Proof. Let $h \in (0, h_\infty)$. It is enough to prove (5.18) for $v \in C^{r+1}(\overline{\Omega})$. Indeed, if $I = (x_A, x_B) \in \mathcal{P}_h$, Taylor's theorem implies that

$$v(x) = \sum_{\ell=0}^{r} \frac{(x - x_A)^\ell}{\ell!}\ v^{(\ell)}(x_A) + \Phi(x) \quad \forall\, x \in I,$$

where $\Phi(x) = \frac{1}{r!} \int_{x_A}^x (x-z)^r\ v^{(r+1)}(z)\ dz \ \ \forall\, x \in I$. From the fact that $\mu_I p = p$, for $p \in \mathbb{P}^r(I)$, it follows that

$$\begin{aligned} |\mu_I v - v|_{\infty,I} &= |\mu_I \Phi - \Phi|_{\infty,I} \\ &\le |\mu_I \Phi - \mathcal{I}^1_h \Phi|_{\infty,I} + |\mathcal{I}^1_h \Phi - \Phi|_{\infty,I}. \end{aligned} \tag{5.19}$$

Since $E := \mu_I \Phi - \mathcal{I}^1_h \Phi \in \mathbb{P}^r_0(I)$, the definition of μ_I implies that

$$\begin{aligned} \widetilde{B}_I(E,E) &= \widetilde{B}_I(Y_I(\Phi - \mathcal{I}^1_h \Phi), E) \\ &= \widetilde{B}_I(\Phi - \mathcal{I}^1_h \Phi, E) \\ &= D_1 + D_2 + D_3, \end{aligned} \tag{5.20}$$

where $D_1 = \big(a(\Phi - \mathcal{I}^1_h\Phi)'', \Lambda_h E - E\big)_I$, $D_2 = \big(a'(\Phi - \mathcal{I}^1_h\Phi)', \Lambda_h E - E\big)_I$, $D_3 = \big(a(\Phi - \mathcal{I}^1_h\Phi)', E'\big)_I$. Using (2.7) and the fact that $\|\Phi''\|_{0,I} \le \frac{1}{(r-2)!}\ h_I^{r-\frac{1}{2}}\ \|v\|_{r+1,\infty}$, we obtain

$$|D_1| \le C\ h_I^{r+\frac{1}{2}}\ \|v\|_{r+1,\infty}\ \|E\|_{1,I}.$$

The identity $(\mathcal{I}_h^1\Phi)' = \Pi_h\Phi'$, the estimate for $\|\Phi''\|_{0,I}$ above, and (2.7), (2.5) yield

$$\begin{aligned}|D_2| &\le C\ h_I\ \|\Phi' - \Pi_h\Phi'\|_{0,I}\ \|E\|_{1,I}\\ &\le C\ h_I^2\ \|\Phi''\|_{0,I}\ \|E\|_{1,I}\\ &\le C\ h_I^{r+\frac{3}{2}}\ \|v\|_{r+1,\infty}\ \|E\|_{1,I},\end{aligned}$$

and similarly,

$$|D_3| \le C\ h_I^{r+\frac{1}{2}}\ \|v\|_{r+1,\infty}\ \|E\|_{1,I}.$$

Using (5.17) in (5.20) and the estimates for D_1, D_2, D_3 above we conclude that

$$\begin{aligned}|E|_{\infty,I} &\le C\ \sqrt{h_I}\ \|E\|_{1,I}\\ &\le C\ h^{r+1}\ \|v\|_{r+1,\infty}.\end{aligned} \tag{5.21}$$

Finally, (5.19), (5.21), and the estimate

$$\begin{aligned}|\mathcal{I}_h^1\Phi - \Phi|_{\infty,I} &\le |\Phi|_{\infty,I}\\ &\le C\ h_I^{r+1}\ \|v\|_{r+1,\infty},\end{aligned}$$

complete the proof of the proposition. ∎

We are now ready to prove the main result of this section.

THEOREM 5.5. *Let $r \ge 2$ and assume that Λ_h satisfies* (5.1), (2.8), *and* (2.7) *with $\alpha \ge 2$. If $u \in W^{r+1,\infty}(\Omega) \cap \mathcal{H}(\Omega)$, then, for h sufficiently small, we have*

$$|u - u_h|_\infty \le\ C\ h^{r+1}\ \|u\|_{r+1,\infty}. \tag{5.22}$$

Proof. Let $I \in \mathcal{P}_h$. Since $\alpha \ge 2$, in view of Propositions 5.3 and 5.4, we have

$$\begin{aligned}|e|_{\infty,I} &\le |u_h - R_h^r u|_{\infty,I} + |R_h^r u - \mu_I u|_{\infty,I} + |\mu_I u - u|_{\infty,I}\\ &\le C\ h^{r+1}\ \|u\|_{r+1,\infty} + |R_h^r u - \mu_I u|_{\infty,I}.\end{aligned} \tag{5.23}$$

Hence, it remains for us to estimate $\xi_I := R_h^r u|_I - \mu_I u \in \mathbb{P}^r(I)$. From the definitions of the operators R_h^r and μ_I we obtain $\widetilde{B}_I(\xi_I, p) = 0 \quad \forall\ p \in \mathbb{P}_0^r(I)$. Setting $p = \xi_I - \mathcal{I}_h^1\xi_I$ in the latter relation and using Remark 4.1 and (5.1), we obtain

$$\begin{aligned}a_*\ \|\xi_I'\|_{0,I}^2 &\le \widetilde{B}_I(\xi_I, \xi_I) + C\ h\ \|\xi_I\|_{1,I}^2\\ &= B_I(\xi_I, \mathcal{I}_h^1\xi_I) + C\ h\ \|\xi_I\|_{1,I}^2\\ &\le C\ h_I^{-\frac{1}{2}} \max_{1\le j\le J_h} \left|(R_h^r u - u)(x_j^h)\right|\ \|\xi_I'\|_{0,I} + C\ h\ (\|\xi_I\|_{0,I}^2 + \|\xi_I'\|_{0,I}^2),\end{aligned} \tag{5.24}$$

where we have used the fact that ξ_I and $R_h^r u - u$ coincide at the nodes of the partition. Also, using (5.5) and (5.18) we have

$$\|\xi_I\|_{0,I} \le C\ h^{r+1}\|u\|_{r+1,\infty}. \tag{5.25}$$

Combining (5.24), (5.25), and Lemma 5.1 we conclude that

$$\|\xi_I'\|_{0,I} \le C\ h^{r+\frac{1}{2}}\ \|u\|_{r+1,\infty} \tag{5.26}$$

for h sufficiently small. Since $|\xi_I|_{\infty,I} \le \max_{1\le j\le J_h} |\xi_I(x_j^h)| + \sqrt{h_I}\ \|\xi_I'\|_{0,I}$, from (5.26) and Lemma 5.1 we obtain

$$|\xi_I|_{\infty,I} \le C\ h^{r+1}\ \|u\|_{r+1,\infty}. \tag{5.27}$$

The theorem now follows by combining (5.27) and (5.23). ∎

5.2. Maximum norm estimates when $\alpha = 1$ and $\sigma = 1$. Here, we shall assume that $\alpha = 1$ and $\sigma = 1$, which covers the methods in Proposition 3.7 and Remark 3.5. Also, the assumption (5.1) for Λ_h, which was fundamental in the analysis of the previous subsection, is not needed now.

For $r \geq 2$ and $h \in (0,1)$, we define the bilinear form $\overline{B} : H_h^1 \times H_h^1 \to \mathbb{R}$ by $\overline{B}(v,\chi) = (a\, v', \chi')$ for $v, \chi \in H_h^1$ and introduce the standard elliptic projection operator $\overline{R}_h^r : H^1(\Omega) \to S_h^r$ by

$$\overline{B}(\overline{R}_h^r v - v, \chi) = 0 \quad \forall\, \chi \in S_h^r, \quad \forall\, v \in H^1(\Omega). \tag{5.28}$$

Using (1.1c) and the Poincaré–Friedrichs inequality we obtain $\overline{B}(\tilde{v},\tilde{v}) \geq C_{E,*}\, \|\tilde{v}\|_1^2$ for $\tilde{v} \in \mathcal{H}(\Omega)$, where $C_{E,*}$ is a positive constant. Thus, the projection $\overline{R}_h^r$ is well defined. Also, it is well known (cf. [41]) that $\overline{R}_h^r$ has the following approximation property:

$$\|\overline{R}_h^r v - v\|_0 + h\, \|\overline{R}_h^r v - v\|_1 \leq\ C\, h^{r+1}\, \|v\|_{r+1} \tag{5.29}$$

$\forall\, h \in (0,1)$ and $v \in H^{r+1}(\Omega) \cap \mathcal{H}(\Omega)$.

We start by estimating the difference $\overline{R}_h^r u - u$ at the nodes of the partition $\mathcal{P}_h$.

LEMMA 5.6. *Let $r \geq 2$. If $u \in H^{r+1}(\Omega) \cap \mathcal{H}(\Omega)$, then we have*

$$\max_{1 \leq j \leq J_h} \left|(\overline{R}_h^r u - u)(x_j^h)\right| \leq\ C\, h^{r+1}\, \|u\|_{r+1} \quad \forall\, h \in (0,1). \tag{5.30}$$

Proof. Let $h \in (0,1)$ and $\overline{\zeta} = \overline{R}_h^r u - u$. Also, we consider the functions $\{g_{x_j^h}\}_{j=1}^{J_h} \subset S_h^r$ defined in (5.7). Using (5.28) we obtain $(a\, \overline{\zeta}', g'_{x_j^h}) = 0$ for $j = 1, \ldots, J_h$. Proceeding now as in the proof of Lemma 5.1, we get $\max_{1 \leq j \leq J_h} |\overline{\zeta}(x_j^h)| \leq \frac{\|a'\|_0}{a_*}\, \|\overline{\zeta}\|_0$. The estimate (5.30) follows by combining the last estimate and (5.29). ∎

The next theorem is the maximum norm error estimate when $\alpha = 1$ and $\sigma = 1$.

THEOREM 5.7. *Let $r \geq 2$ and assume that Λ_h satisfies (2.7) with $\alpha = 1$ and (2.8) with $\sigma = 1$. If $u \in H^{r+2}(\Omega) \cap \mathcal{H}(\Omega)$, then, for h sufficiently small, we have*

$$|u - u_h|_\infty \leq\ C\, h^{r+1}\, \|u\|_{r+2}. \tag{5.31}$$

Proof. Let $h \in (0, h_0)$, $\overline{\nu}_h = \overline{R}_h^r u - u_h \in S_h^r$, $\overline{\zeta} = \overline{R}_h^r u - u$, and $\overline{e} = u - u_h$, where h_0 is the constant specified in Theorem 4.6. Thus, we have $\overline{\nu}_h = \overline{\zeta} + \overline{e}$. Now using (2.10) and (5.28), we obtain

$$B_h(\overline{\nu}_h, \overline{\nu}_h) = D_A + D_B + D_C, \tag{5.32}$$

where $D_A = B_h(\overline{R}_h^r u - \mathcal{I}_h^r u, \overline{\nu}_h) - B(\overline{R}_h^r u - \mathcal{I}_h^r u, \overline{\nu}_h)$, $D_B = B_h(\mathcal{I}_h^r u - u, \overline{\nu}_h) - B(\mathcal{I}_h^r u - u, \overline{\nu}_h)$, and $D_C = (\beta\, \overline{\zeta}' + \gamma\, \overline{\zeta}, \overline{\nu}_h)$. Using Lemma 4.2, (5.29), and (2.2), we have

$$\begin{aligned} \left|D_A\right| &\leq C\, h^2\, \|\overline{R}_h^r u - \mathcal{I}_h^r u\|_1\, \|\overline{\nu}_h\|_1 \\ &\leq C\, h^{r+2}\, \|u\|_{r+1}\, \|\overline{\nu}_h\|_1. \end{aligned} \tag{5.33}$$

Lemma 4.1 directly yields

$$\left|D_B\right| \leq\ C\, h^{r+1}\, \|u\|_{r+2}\, \|\overline{\nu}_h\|_1. \tag{5.34}$$

Using integration by parts, the Sobolev inequality, (5.29), and Lemma 5.6, we get

$$\begin{aligned} \left|D_C\right| &\le C\,\|\overline{\zeta}\|_0\,\|\overline{\nu}_h\|_0 + \left|(\beta\,\overline{\zeta}\,\overline{\nu}_h)(x_R) - (\overline{\zeta},\beta'\,\overline{\nu}_h + \beta\,\overline{\nu}_h')\right| \\ &\le C\,\left(\,\|\overline{\zeta}\|_0 + |\overline{\zeta}(x_R)|\,\right)\,\|\overline{\nu}_h\|_1 \\ &\le C\,h^{r+1}\,\|u\|_{r+1}\,\|\overline{\nu}_h\|_1. \end{aligned} \tag{5.35}$$

Combining (5.32), (5.33), (5.34), (5.35), and Lemma 4.3, we have

$$\|\overline{\nu}_h\|_1^2 \le\ C\,h^{r+1}\,\|u\|_{r+2}\,\|\overline{\nu}_h\|_1 + C\,\|\overline{\nu}_h\|_0^2.$$

The inequality above, along with (5.29) and Theorem 4.6, yields

$$\|\overline{\nu}_h\|_1 \le\ C\,h^{r+1}\,\|u\|_{r+2}, \tag{5.36}$$

which is an H^1-superconvergent estimate analogous to that of Proposition 5.3.

When $\Lambda_h p = p$ for $p \in \mathbb{P}_h^r$ in (5.2), then from (5.4) and (5.28) we have $R_h^r u = \overline{R}_h^r u$. Thus the analysis of the previous subsection (cf. Theorem 5.5) yields

$$|\overline{\zeta}|_\infty \le\ C\,h^{r+1}\,\|u\|_{r+1,\infty}. \tag{5.37}$$

The estimate (5.31) follows, combining (5.36) and (5.37). ∎

6. Some pointwise estimates. In the first proposition of this section we present a pointwise estimate for the error $u_h - u$ at the nodes of the partition, generalizing the well-known superconvergence results from the finite element case (see, e.g., [43] and the references therein).

PROPOSITION 6.1. *Let $r \ge 2$, h_0 be the constant defined in* Theorem 4.6*, and assume that Λ_h satisfies* (2.7) *and* (2.8)*. If $u \in H^{r+1+\sigma}(\Omega) \cap \mathcal{H}(\Omega)$, then we have*

$$\max_{1\le j\le J_h} \left|(u-u_h)(x_j^h)\right| \le C\,h^{r+\alpha+\sigma-1}\,\|u\|_{r+1+\sigma} \quad \forall\, h \in (0,h_0). \tag{6.1}$$

Proof. Let $h \in (0,h_0)$. For ξ a node of the partition $\mathcal{P}_h$, denote by $G_\xi \in H^1(\Omega)$ the Green's function of problem (1.1a)–(1.1b) (cf., e.g., [37]), which satisfies

$$B(v,G_\xi) = v(\xi) \quad \forall\, v \in H^1(\Omega). \tag{6.2}$$

In addition, we have $G_\xi \in H^{r+1}(x_L,\xi) \cap H^{r+1}(\xi,x_R)$ and

$$\|G_\xi\|_{r+1,(x_L,\xi)} + \|G_\xi\|_{r+1,(\xi,x_R)} \le \overline{C} \tag{6.3}$$

for some constant $\overline{C}$ depending only on Ω and the coefficients of L. Hence, from (2.2) we obtain

$$\|G_\xi - \mathcal{I}_h^r G_\xi\|_1 \le C\,h^r. \tag{6.4}$$

With $v = e = u_h - u$, (6.2) yields the error representation

$$e(\xi) = B(e, G_\xi - \mathcal{I}_h^r G_\xi) + B(e, \mathcal{I}_h^r G_\xi). \tag{6.5}$$

Using Theorem 4.6 and (6.4) we conclude that

$$\left|B(e, G_\xi - \mathcal{I}_h^r G_\xi)\right| \le C\,h^{2r}\,\|u\|_{r+1}. \tag{6.6}$$

To estimate $B(e,\mathcal{I}_h^r G_\xi)$ we use the orthogonality property (2.10) to write it as

$$\begin{aligned} B(e,\mathcal{I}_h^r G_\xi) &= B(e,\mathcal{I}_h^r G_\xi) - B_h(e,\mathcal{I}_h^r G_\xi) \\ &= E_A + E_B, \end{aligned} \tag{6.7}$$

where $E_A = B(\eta,\mathcal{I}_h^r G_\xi) - B_h(\eta,\mathcal{I}_h^r G_\xi)$, $E_B = B(\theta_h,\mathcal{I}_h^r G_\xi) - B_h(\theta_h,\mathcal{I}_h^r G_\xi)$, and $\theta_h = u_h - \mathcal{I}_h^r u$, $\eta = \mathcal{I}_h^r u - u$. First, from Lemma 4.1 (with $\ell = r$, $s = \alpha$, and $j = \sigma$), (2.2), and the bound (6.3), we have

$$\begin{aligned} \big|E_A\big| &\le C\, h^{\alpha+r-1+\sigma}\, \|u\|_{r+1+\sigma}\, \|\mathcal{I}_h^r G_\xi\|_{\alpha,I} \\ &\le C\, h^{\alpha+r-1+\sigma}\, \|u\|_{r+1+\sigma}. \end{aligned} \tag{6.8}$$

To estimate E_B we note that from Lemma 4.2 (with $s = \alpha$), Theorem 4.6, (2.2), and (6.3), we have

$$\begin{aligned} \big|E_B\big| &\le C\, h^{\alpha+\sigma}\, \|\theta_h\|_1\, \|\mathcal{I}_h^r G_\xi\|_{\alpha,h} \\ &\le C\, h^{\alpha+r+\sigma}\, \|u\|_{r+1}. \end{aligned} \tag{6.9}$$

The estimate (6.1) now follows by combining (6.5)–(6.9). ∎

REMARK 6.1. *The result of* Proposition 6.1 *is a superconvergence estimate for the error e, when the order of convergence $r+\alpha+\sigma-1$ is greater than the order of convergence in the L^∞ norm. This happens when $\alpha+\sigma-1 \ge 2$, which means $\alpha \ge 3$ when $\sigma = 0$ and $\alpha \ge 2$ when $\sigma = 1$. For $r \ge 4$ and for the locally conservative methods of Proposition* 3.1*, the order of convergence in* (6.1) *is equal to $2(r-1)$, since $\sigma = 0$ and $\alpha = r-1$. For $r = 2m$ and for the methods of Proposition* 3.2*, the order of convergence in* (6.1) *is equal to $2r$ since $\alpha = r$ and $\sigma = 1$ (cf. Table* 8.5 *in section* 8 *for $r = 2$), which is the same as that of the finite element method. For $r \in \{2,4,6\}$ and for the finite volume methods in Proposition* 3.7 *we do not have a superconvergence result. Indeed the order of convergence in* (6.1) *is equal to $r+1$ (since $\alpha = 1$ and $\sigma = 1$) and has been observed numerically (cf. Table* 8.6 *in section* 8 *for $r = 2$).*

The next proposition presents a pointwise bound for the derivative of the error $u_h - u$. This is interesting because it is related to the characteristic property (1.2a) of the locally conservative methods and the Neumann boundary condition in (1.1b).

PROPOSITION 6.2. *Let $r \ge 2$. Assume that $u_h \in S_h^r$ satisfies*

$$\mathcal{F}\big(u_h,(\omega_{j-1}^h,\omega_j^h)\big) = 0, \quad j = 1,\dots,M_h-1, \tag{6.10a}$$

$$\mathcal{G}\big(u_h,(\omega_{M_h-1}^h,\omega_{M_h}^h)\big) = 0, \quad u_h'(\omega_j^h-0) = u_h'(\omega_j^h+0), \quad j = 1,\dots,M_h-1, \tag{6.10b}$$

where $\{(\omega_{j-1}^h,\omega_j^h)\}_{j=1}^{M_h} \subset \Omega$ are open intervals with $x_L \le \omega_0^h$, $\omega_{M_h}^h = x_R$ and $\omega_{j-1}^h < \omega_j^h$, $j = 1,\dots,M_h$, i.e., $(\omega_{M_h-1}^h,\omega_{M_h}^h)$ is a boundary control volume of type II. *Then we have*

$$\max_{0\le m\le M_h-1} \big|(u-u_h)'(\omega_m^h)\big| \le C\, \left\{ \max_{0\le m\le M_h} \big|(u-u_h)(\omega_m^h)\big| + \|u-u_h\|_0 \right\}. \tag{6.11}$$

Proof. Let $e = u_h - u$. From (6.10) and (1.2) we obtain

$$a(\omega_j^h)e'(\omega_j^h) - a(\omega_{j-1}^h)e'(\omega_{j-1}^h) = \int_{\omega_{j-1}^h}^{\omega_j^h} (\beta e' + \gamma e)\, dx, \quad j = 1,\dots,M_h-1,$$

$$-a(\omega_{M_h-1}^h)\, e'(\omega_{M_h-1}^h) = \int_{\omega_{M_h-1}^h}^{x_R} (\beta e' + \gamma e)\, dx.$$

Adding the above equalities with respect to j we arrive at

$$a(\omega_m^h)\, e'(\omega_m^h) = -\int_{\omega_m^h}^{x_R} (\beta\, e' + \gamma\, e)\, dx, \quad m = 0, \ldots, M_h - 1.$$

Using integration by parts we obtain

$$a(\omega_m^h) e'(\omega_m^h) = \beta(\omega_m^h)\, e(\omega_m^h) - \beta(x_R)\, e(x_R) + \int_{\omega_m^h}^{x_R} (\beta' - \gamma)\, e\, dx \tag{6.12}$$

for $m = 0, \ldots, M_h - 1$, which yields (6.11). ∎

REMARK 6.2. *A final error estimate follows by combining* (6.11) *with the estimates in Theorems* 4.6, 5.5, *and* 5.7. *We also note that when* $\beta = \gamma = 0$ *then* $u_h'(\omega_j^h) = u'(\omega_j^h)$, *for* $j = 0, \ldots, M_h - 1$, *which is not the case for the finite element method.*

REMARK 6.3. *The methods in Proposition* 3.2 *satisfy the assumptions* (6.10a) *with* $M_h = J_h$ *and* $\omega_j^h = x_j^h$, $j = 0, \ldots, J_h$, *but not* (6.10b). *The result* (6.11) *also applies to the methods of Propositions* 3.6–3.7 *and Remark* 3.5.

REMARK 6.4. *For the locally conservative methods of Proposition* 3.1 *with* $r \geq 4$ *we have* $M_h = J_h$ *and* $\omega_j^h = x_j^h$ *for* $j = 0, \ldots, J_h$. *From* (6.12) *and Remark* 6.1 *we obtain*

$$\max_{0 \leq m \leq J_h - 1} |e'(x_m^h)| \leq\ C\ \left\{ h^{2(r-1)}\ \|u\|_{r+1} + \max_{0 \leq m \leq J_h - 1} \left| \int_{x_m^h}^{x_R} (\beta' - \gamma) e\, dx \right| \right\}. \tag{6.13}$$

Let $\xi \in [x_L, x_R)$ *and* $g = \mathcal{X}_{(\xi, x_R)}$. *For* $\delta \in (0,1)$ *there exists* $g_\delta \in C^{r-3}(\Omega)$ *such that*

$$\int_\Omega |g_\delta - g|\, dx \leq C\, \delta \qquad \text{and} \qquad \|g_\delta\|_{r-3} \leq C\, \delta^{-(r-3)+\frac{1}{2}}, \tag{6.14}$$

where the constant C *is independent of* ξ *and* δ. *Using* g *and* g_δ, *we introduce the following splitting:*

$$\int_\xi^{x_R} (\beta' - \gamma)\, e\, dx = \int_{x_L}^{x_R} (\beta' - \gamma)\, e\, (g - g_\delta)\, dx + \int_{x_L}^{x_R} (\beta' - \gamma)\, e\, g_\delta\, dx.$$

From Theorem 5.5 *and* (6.14), *we have*

$$\left| \int_{x_L}^{x_R} (\beta' - \gamma)\, e\, (g - g_\delta)\, dx \right| \leq C\, \delta\, h^{r+1}\, \|u\|_{r+1,\infty}. \tag{6.15}$$

For $w_\delta = T^*\big((\beta' - \gamma) g_\delta\big)$, *where* T^* *is the solution operator of* (4.6), *we obtain the identity*

$$\int_{x_L}^{x_R} e\, (\beta' - \gamma)\, g_\delta\, dx = B(e, w_\delta - \mathcal{I}_h^{r-2} w_\delta) + B(e, \mathcal{I}_h^{r-2} w_\delta).$$

From the definition of Λ_h *in Proposition* 3.1, (2.10), *and* (2.11), *we conclude that* $B(e, \mathcal{I}_h^{r-2} w_\delta) = 0$. *Hence, using* (2.2), (4.18), *the fact that* $\|w_\delta\|_{r-1} \leq C\, \|g_\delta\|_{r-3}$, *and* (6.14), *it follows that*

$$\begin{aligned} \left| \int_{x_L}^{x_R} e\, (\beta' - \gamma)\, g_\delta\, dx \right| &\leq C\, h^{2(r-1)}\, \|g_\delta\|_{r-3}\, \|u\|_{r+1} \\ &\leq C\, h^{2(r-1)}\, \delta^{-(r-3)+\frac{1}{2}}\, \|u\|_{r+1}. \end{aligned} \tag{6.16}$$

With $\delta = h^\nu$, *where* $\nu = \frac{2(r-3)}{2(r-3)+1}$, *the estimates* (6.13), (6.15), *and* (6.16) *yield* $\max_{0 \leq m \leq J_h - 1} \big|e'(x_m^h)\big| \leq\ C\, h^{r+1+\nu}\, \|u\|_{r+1,\infty}$.

7. A posteriori estimates. In this section we derive explicit residual-based a posteriori error estimates in the energy and L^2 norm for the methods of section 2 with $r = 2$, following the approach of [17] for the standard Galerkin finite element method. Our estimates are based on the general formulation (2.9) and make use of the orthogonality properties (2.8) and (2.10). We refer to [2] and [6] for a posteriori error estimates in the L^2 norm for finite volume methods based on piecewise linear polynomial spaces and applied to linear and nonlinear elliptic problems in two space dimensions.

PROPOSITION 7.1. *Let $r = 2$ and assume that Λ_h satisfies* (2.7) *and* (2.8). *For $h \in (0, \widehat{h}_0)$, we define the residual $R_E^h \in L^2(\Omega)$ of* (2.9) *by $R_E^h|_I = f - L_I u_h$ for $I \in \mathcal{P}_h$, where $\widehat{h}_0$ is the constant specified in Proposition* 4.4. *If $\sigma = 0$ in* (2.8), *then there exists a constant $C^P_{\alpha^*,0}$, independent of the solution u and the right-hand side function f, such that*

$$\|u - u_h\|_0 \le C^P_{\alpha^*,0} \left\{ \sum_{I\in\mathcal{P}_h} h_I^{2\alpha^*} \lambda^2_{\alpha^*}(h_I) \|R_E^h\|^2_{0,I} \right\}^{\frac{1}{2}} \quad \forall\, h \in (0, \widehat{h}_0), \tag{7.1}$$

where $\alpha^* = \min\{2, \alpha\}$, $\lambda_{\alpha^*}(h_I) = 1 + \widehat{C}_{2,1}(1 + h_I)$ when $\alpha^* = 1$, and $\lambda_{\alpha^*}(h_I) = \widetilde{C}_1 + \widehat{C}_{2,2}(1 + h_I)$ when $\alpha^* = 2$. If $\alpha = 1$ in (2.7) *and $\sigma = 1$ in* (2.8), *then there exists a constant $C^P_{1,1}$, independent of the solution u and the right-hand side function f, such that*

$$\|u - u_h\|_0 \le C^P_{1,1} \left\{ \sum_{I\in\mathcal{P}_h} h_I^2 \left(\widetilde{C}_1\, h_I \|R_E^h\|_{0,I} + \widehat{C}_{2,1}\,(1 + h_I) \|R_E^h - \mathcal{I}_h^1 R_E^h\|_{0,I} \right)^2 \right\}^{\frac{1}{2}} \tag{7.2}$$

$\forall\, h \in (0, \widehat{h}_0)$. *In the estimates above, $\widehat{C}_{2,1}$ and $\widehat{C}_{2,2}$ are the constants in* (2.7), *and $\widetilde{C}_1$ is the constant in* (2.2).

Proof. Let $h \in (0, \widehat{h}_0)$, $e = u - u_h$, and $w = T^* e \in H^2(\Omega) \cap \mathcal{H}(\Omega)$. Using integration by parts and (2.9) we have

$$\begin{aligned}
\|e\|_0^2 &= B(e, w) \\
&= (f, w) - B(u_h, w) \\
&= \sum_{I\in\mathcal{P}_h} (R_E^h, w - \Lambda_h \mathcal{I}_h^1 w)_I \\
&= \sum_{I\in\mathcal{P}_h} \left[(R_E^h, w - \mathcal{I}_h^1 w)_I + (R_E^h, \mathcal{I}_h^1 w - \Lambda_h \mathcal{I}_h^1 w)_I \right].
\end{aligned}$$

When $\alpha = 1$ and $\sigma = 0$, using the identity $(\mathcal{I}_h^1 w)' = \Pi_h w'$, (2.4), and (2.7), we have

$$\begin{aligned}
\|e\|_0^2 &\le \sum_{I\in\mathcal{P}_h} h_I \|R_E^h\|_{0,I} \left(\|w\|_{1,I} + \widehat{C}_{2,1} \|\mathcal{I}_h^1 w\|_{1,I} \right) \\
&\le \sum_{I\in\mathcal{P}_h} h_I \lambda_1(h_I) \|R_E^h\|_{0,I} \|w\|_{1,I} \\
&\le C_Q \left\{ \sum_{I\in\mathcal{P}_h} h_I^2 \lambda_1^2(h_I) \|R_E^h\|^2_{0,I} \right\}^{\frac{1}{2}} \|e\|_0,
\end{aligned}$$

where $C_{_Q} = \sup\big\{\|T^*g\|_1 : \ g \in L^2(\Omega) \ \text{ with } \ \|g\|_0 = 1\big\}$. When $\alpha \geq 2$ and $\sigma = 0$, using (2.2), (2.4), and (2.7), we have

$$\begin{aligned}\|e\|_0^2 &\leq \sum_{I\in\mathcal{P}_h} h_I^2 \, \|R_{_E}^h\|_{0,I} \, (\widetilde{C}_1 \, \|w\|_{2,I} + \widehat{C}_{2,2} \, \|\mathcal{I}_h^1 w\|_{2,I}) \\ &\leq \sum_{I\in\mathcal{P}_h} h_I^2 \, \lambda_2(h_I) \, \|R_{_E}^h\|_{0,I} \, \|w\|_{2,I} \\ &\leq C_{_R} \, \left\{ \sum_{I\in\mathcal{P}_h} h_I^4 \, \lambda_2^2(h_I) \, \|R_{_E}^h\|_{0,I}^2 \right\}^{\frac{1}{2}} \, \|e\|_0,\end{aligned}$$

where $C_{_R}$ is the constant of the elliptic regularity estimate (4.7). The estimates above yield (7.1) with $C_{1,0}^{_P} = C_{_Q}$ and $C_{2,0}^{_P} = C_{_R}$. When $\alpha = 1$ and $\sigma = 1$, we use the orthogonality property (2.8) and the estimates (2.2), (2.7), (2.4) to get

$$\|e\|_0^2 \leq \sum_{I\in\mathcal{P}_h} h_I \, \big(\widetilde{C}_1 \, h_I \, \|R_{_E}^h\|_{0,I} + \widehat{C}_{2,1} \, (1+h_I) \, \|R_{_E}^h - \mathcal{I}_h^1 R_{_E}^h\|_{0,I}\big) \, \|w\|_{2,I}$$

which along with (4.7) yields (7.2) with $C_{1,1}^{_P} = C_{_R}$. □

Remark 7.1. *When* (5.1) *holds, we have* $\Lambda_h \mathcal{I}_h^1 w = \mathcal{I}_h^1 w$, *so, instead of* (7.1) *we obtain* $\|u - u_h\|_0 \leq \widetilde{C}_1 \, C_{_R} \, \{\sum_{I\in\mathcal{P}_h} h_I^2 \, \|R_{_E}^h\|_{0,I}^2\}^{1/2}$ *for* $h \in (0, \widehat{h}_0)$, *which has the form of a finite element a posteriori error estimator.*

Remark 7.2. *If we assume that* $C_{_\Gamma} = 0$ *in* (4.4), *then we also have* $C_{_G} = 0$ *in* (4.3). *Proceeding as in Proposition* 7.1, *we conclude that*

$$B(e,e) = \sum_{I\in\mathcal{P}_h} (R_{_E}^h, e - \Lambda_h \mathcal{I}_h^1 e)_I \leq \ \Big\{ \sum_{I\in\mathcal{P}_h} h_I^2 \, \lambda_1^2(h_I) \, \|R_{_E}^h\|_{0,I}^2 \Big\}^{\frac{1}{2}} \, \|e\|_1.$$

Using (4.4) *we obtain* $\big[B(e,e)\big]^{1/2} \leq (C_{_\Delta})^{-1/2} \, \big\{\sum_{I\in\mathcal{P}_h} h_I^2 \, \lambda_1^2(h_I) \, \|R_{_E}^h\|_{0,I}^2\big\}^{\frac{1}{2}}$, *which is an a posteriori error bound in the energy norm.*

Remark 7.3. *We note that in the a posteriori error bounds of Proposition* 7.1 *and Remarks* 7.1 *and* 7.2 *the jump of the derivative at the interval boundaries is absent. This is due to the use of the interpolation operator* $\mathcal{I}_h^1$ *and this also applies in the finite element case in one space dimension* (*see, e.g.,* [17]).

Remark 7.4. *Assume that* $f \in H^1(\Omega)$ *or* $f \in H^2(\Omega)$. *Observing that*

$$B(u - u_h, v) = \sum_{I\in\mathcal{P}_h} \big\{(R_{_E}^h, v)_I - [\![a u_h' v]\!]_{\partial I}\big\} \quad \forall \, v \in H^1(\Omega),$$

and moving along the lines of [42] *or section* 3.4 *in* [3], *we can show that*

$$\|R_{_E}^h\|_{0,I} \leq C \, \big\{\|R_{_E}^h - \mathcal{I}_h^1 R_{_E}^h\|_{0,I} + \tfrac{1}{h_I} \, \|e\|_{1,I}\big\} \quad \forall \, I \in \mathcal{P}_h.$$

Since $L_I u_h \in C_{_B}^2(I)$, *using* (2.4), (2.2), *and* (2.1), *we obtain*

$$\begin{aligned}\|R_{_E}^h - \mathcal{I}_h^1 R_{_E}^h\|_{0,I} &\leq C \, \big(\|f - \mathcal{I}_h^1 f\|_{0,I} + h_I^2 \, \|L_I u_h\|_{2,I}\big) \\ &\leq C \, \big(h_I^s \, \|f\|_{s,I} + h_I^2 \, \|u_h\|_{2,I}\big) \\ &\leq C \, h_I^s \, \big(\|f\|_{s,I} + \|u_h\|_{1,I}\big), \quad s = 1, 2, \quad \forall \, I \in \mathcal{P}_h.\end{aligned}$$

We conclude from the two relations above that

$$\|R_E^h\|_{0,I} \le C\left\{h_I^s\,(\|f\|_{1,I}+\|u_h\|_{1,I})+\tfrac{1}{h_I}\,\|e\|_{1,I}\right\}\quad\forall\, I\in\mathcal{P}_h,\quad s=1,2. \tag{7.3}$$

The estimates above yield

$$\left\{\sum_{I\in\mathcal{P}_h} h_I^{2\overline{\alpha}}\,\|R_E^h\|_{0,I}^2\right\}^{\frac12} \le C\,\{h^{\overline{\alpha}+1}\,(\|f\|_1+\|u_h\|_1)+h^{\overline{\alpha}-1}\,\|e\|_1\}\quad\forall\,\overline{\alpha}\in\mathbb{N},$$

$$\left\{\sum_{I\in\mathcal{P}_h} h_I^2\,\|R_E^h-\mathcal{I}_h^1 R_E^h\|_{0,I}^2\right\}^{\frac12} \le C\,h^3\,(\|f\|_2+\|u_h\|_1).$$

Hence, the a posteriori error estimators of Proposition 7.1 *and Remarks* 7.1 *and* 7.2 *and the corresponding approximation errors are of the same order, and we expect them to be useful in the construction of an adaptive algorithm. In addition, using* (7.3) *we obtain the following lower bound of the* H^1 *error (cf. Theorem* 3.7 *in* [3]*):*

$$\sum_{I\in\mathcal{P}_h} h_I^2\|R_E^h\|_{0,I}^2 \le C\,\left\{\sum_{I\in\mathcal{P}_h} h_I^{2s+2}\,(\|f\|_{s,I}+\|u_h\|_{1,I})^2+\|e\|_1^2\right\},\quad s=1,2.$$

For the purpose of describing later in section 8 an algorithm for the control of the L^2 approximation error, we shall write our error estimators in the form

$$\mathcal{E}_h = C_{_M}\,\left\{\sum_{I\in\mathcal{P}_h}\eta_I^2\right\}^{\frac12}, \tag{7.4}$$

where η_I is a nonnegative, computable error indicator for the interval $I\in\mathcal{P}_h$. Indeed, for the error estimator (7.1) of Proposition 7.1 we have

$$C_{_M}=C^{P}_{\alpha^*,0},\quad \eta_I = h_I^{\alpha^*}\lambda_{\alpha^*}(h_I)\,\|R_E^h\|_{0,I}, \tag{7.5}$$

while for the error estimator (7.2)

$$C_{_M}=C^{P}_{1,1},\quad \eta_I = h_I\,\big(\widetilde{C}_1\,h_I\,\|R_E^h\|_{0,I}+\widehat{C}_{2,1}\,(1+h_I)\,\|R_E^h-\mathcal{I}_h^1R_E^h\|_{0,I}\big). \tag{7.6}$$

8. Numerical experiments. In this section we present the results of numerical experiments performed with some of the methods of section 3, based on piecewise quadratic functions. We shall refer to the method of Proposition 3.2 with $r=2$ as Method 3.1 and to the method of Proposition 3.6 with $r=2$ and $\varrho_1=\frac12-\frac{\sqrt3}{6}$ as Method 3.2. All numerical schemes were implemented in a `FORTRAN` program using double precision arithmetic. The resulting linear systems were solved using the `LINPACK` subroutines `DGBFA` and `DGBSL`. All runs were performed on a Sun UltraSparc 5 running SunOS 5.6, using the native version of the `FORTRAN` compiler.

8.1. Experimental order of convergence. Our first task is to verify numerically the convergence rate of Methods 3.1 and 3.2 and compare them with the standard Galerkin finite element method based on S_h^2. To do this we consider the problem

$$-\big((2+\cos(\tfrac{\pi}{2}x))\,u'\big)'+u'+u=f(x)\quad\forall\,x\in(0,1),\quad u(0)=u'(1)=0, \tag{8.1}$$

TABLE 8.1
Rates of convergence of Method 3.1 *for the problem* (8.1).

J_h	$\|u-u_h\|_0$	Rate	$\|u-u_h\|_1$	Rate	$\|u-u_h\|_\infty$	Rate
20	1.97047(-6)		2.55364(-4)		3.84016(-6)	
40	2.46264(-7)	3.0003	6.38364(-5)	2.0001	4.79942(-7)	3.0002
80	3.07815(-8)	3.0000	1.59588(-5)	2.0000	5.99904(-8)	3.0001
160	3.84765(-9)	3.0000	3.98969(-6)	2.0000	7.49872(-9)	3.0000

TABLE 8.2
Rates of convergence of Method 3.2 *for the problem* (8.1).

J_h	$\|u-u_h\|_0$	Rate	$\|u-u_h\|_1$	Rate	$\|u-u_h\|_\infty$	Rate
20	3.19946(-6)		2.55434(-4)		5.22213(-6)	
40	4.04199(-7)	2.9847	6.38407(-5)	2.0004	6.55057(-7)	2.9949
80	5.07905(-8)	2.9924	1.59591(-5)	2.0000	8.20274(-8)	2.9974
160	6.36727(-9)	2.9958	3.98971(-6)	2.0000	1.02662(-8)	2.9982

TABLE 8.3
Rates of convergence of the finite element method on S_h^2 *for the problem* (8.1).

J_h	$\|u-u_h\|_0$	Rate	$\|u-u_h\|_1$	Rate	$\|u-u_h\|_\infty$	Rate
20	1.96957(-6)		2.55318(-4)		3.84727(-6)	
40	2.46236(-7)	2.9998	6.38336(-5)	1.9999	4.80545(-7)	3.0011
80	3.07808(-8)	2.9999	1.59586(-5)	2.0000	6.00330(-8)	3.0008
160	3.84762(-9)	3.0000	3.98968(-6)	2.0000	7.50155(-9)	3.0005

where f is chosen so that the problem admits the C^∞–solution $u(x) = \sin(\frac{\pi}{2}x)$. This problem was solved numerically on a uniform grid consisting of $J_h = 20$, 40, 80, or 160 intervals. The L^2 and H^1 norms of the error $e = u - u_h$ were computed using an eight-point Gauss quadrature rule, and the L^∞ norm of the error was estimated by a finite sampling at the abscissae of the aforementioned quadrature rule. The results of these computations are summarized in Tables 8.1–8.2 and clearly confirm the convergence estimates of sections 4 and 5. The experimental rate of convergence and the discretization error for the standard Galerkin finite element method based on S_h^2 are shown in Table 8.3. The close agreement between the discretization errors of Method 3.1 and the finite element method is easily explained by the fact that u_h satisfies the finite element equations in addition to the finite volume equations (cf. Remark 3.2). Moreover, it is worth noting that the H^1 norm of the discretization error is approximately the same for all three methods.

For the sake of completeness we performed similar experiments with the first member of the infinite family of the methods constructed in Proposition 3.1 (cf. Remark 3.1) and listed the results in Table 8.4. Recall that for this method (referred to in Table 8.4 as the "Π_h-method") we have $\Lambda_h = \Pi_h$ and $\alpha = 1$, $\sigma = 0$. The experimental rates of convergence clearly agree with the rates predicted in Theorem 4.6.

To confirm the results of Proposition 6.1 (see also Remark 6.1) we determined the rate of convergence of the quantities $|e(1/4)|$, $|e(4/5)|$, and $|e(1)|$ for Methods 3.1–3.2, applied to the test problem (8.1). The results are shown in Tables 8.5–8.6.

8.2. Adaptive computations. As our next task we undertake the development and testing of an adaptive algorithm based on Methods 3.1 and 3.2, which uses the

TABLE 8.4
Rates of convergence of the Π_h-method *for the problem* (8.1).

J_h	$\|u-u_h\|_0$	Rate	$\|u-u_h\|_1$	Rate	$\|u-u_h\|_\infty$	Rate
20	2.82391(-4)		5.12046(-4)		4.03168(-4)	
40	7.06080(-5)	1.9998	1.28019(-4)	1.9999	1.00779(-4)	2.0002
80	1.76526(-5)	2.0000	3.20053(-5)	2.0000	2.51941(-5)	2.0000
160	4.41320(-6)	2.0000	8.00136(-6)	2.0000	6.29848(-6)	2.0000

TABLE 8.5
Rates of convergence of $|e(1/4)|$, $|e(4/5)|$, *and* $|e(1)|$ (*Method* 3.1).

J_h	$\|e(1/4)\|$	Rate	$\|e(4/5)\|$	Rate	$\|e(1)\|$	Rate
20	9.64734(-9)		2.73913(-8)		3.17698(-8)	
40	6.03302(-10)	3.9992	1.71231(-9)	3.9997	1.98573(-9)	3.9999
80	3.77473(-11)	3.9984	1.07111(-10)	3.9987	1.24206(-10)	3.9989
160	2.33780(-12)	4.0130	6.62781(-12)	4.0144	7.69784(-12)	4.0121

TABLE 8.6
Rates of convergence of $|e(1/4)|$, $|e(4/5)|$, *and* $|e(1)|$ (*Method* 3.2).

J_h	$\|e(1/4)\|$	Rate	$\|e(4/5)\|$	Rate	$\|e(1)\|$	Rate
20	3.00789(-7)		3.87146(-6)		5.28432(-6)	
40	3.75175(-8)	3.0031	4.82915(-7)	3.0030	6.58883(-7)	3.0036
80	4.68532(-8)	3.0013	6.03067(-8)	3.0014	8.22646(-8)	3.0017
160	5.86658(-10)	2.9976	7.53832(-9)	3.0000	1.02810(-8)	3.0003

a posteriori error bounds derived in section 7 to control the L^2 norm of the error. The goal is to compute an approximation u_h of the exact solution u of (1.1a)–(1.1b) such that $\|u-u_h\|_0 \le \text{TOL}$, for a given tolerance $\text{TOL} > 0$. In what follows, for $m \in \mathbb{N}_0$, we shall denote by $\mathcal{P}^{(m)}$ a partition of Ω with $J^{(m)}$ intervals and by $S^{(m)} \subset \mathcal{H}(\Omega)$ the finite element space consisting of functions which vanish at x_L and reduce to polynomials of degree less than or equal to two on each interval I of the partition $\mathcal{P}^{(m)}$. The adaptive algorithm starts with an initial partition $\mathcal{P}^{(0)}$ and an initial approximation $u^{(0)} \in S^{(0)}$ of u. Then it computes successive approximations $u^{(k)} \in S^{(k)}$, $k \ge 1$, where $S^{(k-1)} \subset S^{(k)}$, by means of the following iterative procedure.

Step 1. Given an approximate solution $u^{(k-1)} \in S^{(k-1)}$, compute the error indicators $\eta_I^{(k-1)}$ for $I \in \mathcal{P}^{(k-1)}$ and the corresponding error estimator $\mathcal{E}_h^{(k-1)}$ from (7.4).

Step 2. If $\mathcal{E}_h^{(k-1)} \le \text{TOL}$ stop. Otherwise, construct a new partition $\mathcal{P}^{(k)}$ of Ω by bisecting the intervals I of $\mathcal{P}^{(k-1)}$ for which $\eta_I^{(k-1)} > \frac{\text{TOL}}{C_M\sqrt{J^{(k-1)}}}$.

Step 3. Compute a new approximate solution $u^{(k)} \in S^{(k)}$, increment k, and go to Step 1.

We shall apply this adaptive algorithm on the test problem

$$-((x+\epsilon)\,u')' = 1 \quad \forall\, x \in (0,1), \quad u(0) = u'(1) = 0, \tag{8.2}$$

where $\epsilon > 0$ is a parameter. The exact solution of (8.2) is $u(x) = (1+\epsilon)\,\ln\big(1+\frac{x}{\epsilon}\big) - x$. We note that, for small ϵ, u and u' change rapidly near $x = 0$ and thus a locally fine grid is required to approximate u accurately. The constant C_M needed in the adaptive algorithm depends on the constants C_R, $\widetilde{C}_1$, and $C_{1,1}^P$ of the a posteriori

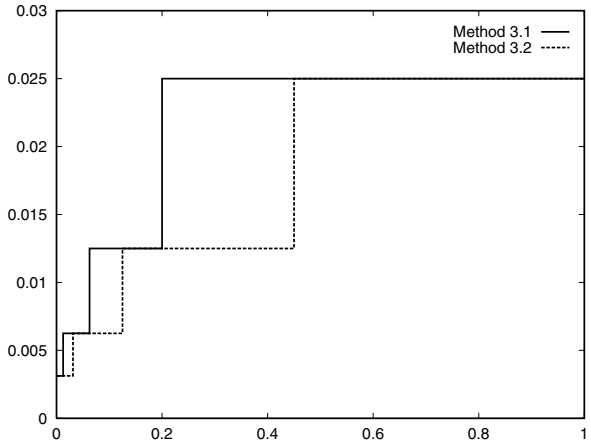


FIG. 8.1. *The mesh-size function for problem* (8.2) *with* $\epsilon = 0.01$ *and* TOL $= 0.0001$.

error estimates; cf. (7.4), (7.5), (7.6), and Remark 7.1. For Method 3.1 we have $C_M = \widetilde{C}_1\, C_R = \frac{1}{8}\, C_R$ (cf. [17]), while for Method 3.2 we have $C_M = C_R$. The constant $\widehat{C}_{2,1}$, needed in the computation of the error indicators for Method 3.2, may be estimated as in the proof of Proposition 3.5, and we have $\widehat{C}_{2,1} = \frac{17}{8} - \varrho_1$. To find upper bounds for the constants C_Q and C_R we observe that if $g \in L^2(\Omega)$ and $w = T^* g$, then for (8.2) we have $w'(x) = \frac{1}{x+\epsilon} \int_x^1 g(s)\, ds$ for $x \in [0,1]$. It follows that $C_Q \le \widetilde{C}_Q = \sqrt{A}$ and $C_R \le \widetilde{C}_R = \{(\frac{1}{\epsilon} + B)^2 + \frac{3}{2}\widetilde{C}_Q\}^{\frac{1}{2}}$, where $A = \int_0^1 \frac{1-x}{(x+\epsilon)^2}\, dx$ and $B = \int_0^1 \frac{1-x}{(x+\epsilon)^4}\, dx$. In the numerical experiments we used the upper bounds $\widetilde{C}_Q$ and $\widetilde{C}_R$ instead of C_Q and C_R, respectively. Figure 8.1 shows the local mesh-size function (i.e., the piecewise constant function whose restriction on each interval of the partition equals the length of the interval) after our adaptive algorithm terminated. We used $\epsilon = 0.01$, TOL $= 0.0001$, and an initial uniform partition of $[0,1]$ with 20 intervals. The final number of subintervals and the L^2 norm of the exact error for Method 3.1 were 47 and 2.33017(-5), respectively, and for Method 3.2 they were 55 and 3.19060(-5), respectively. The corresponding errors on uniform grids with the same number of intervals, 0.78595(-3) and 0.634621(-3), are both greater than TOL. We also note that for both methods the length of the smallest interval is the same, but for Method 3.2 the size of the finest grid region is approximately twice as large as that of Method 3.1. This is in good agreement with the discretization errors of Tables 8.1–8.2.

Acknowledgments. The authors wish to thank the referees for their valuable suggestions and constructive comments and gratefully acknowledge the support and warm hospitality of the Department of Numerical Analysis and Computer Science (NADA), Royal Institute of Technology (KTH), Stockholm, Sweden, and of the Centre de Recherche en Mathématiques de la Décision (CEREMADE), UMR CNRS 7534, Université de Paris IX-Dauphine, Paris, France.

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, D. Van Nostrand Co., Inc., Princeton, NJ, 1965.

[2] A. Agouzal and F. Oudin, *A posteriori error estimator for finite volume methods*, Appl. Math. Comput., 110 (2000), pp. 239–250.
[3] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Comput. Methods Appl. Mech. Engrg., 142 (1997), pp. 1–88.
[4] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 1749–1779.
[5] R. E. Bank and D. J. Rose, *Some error estimates for the box method*, SIAM J. Numer. Anal., 24 (1987), pp. 777–787.
[6] A. Bergam and Z. Mghazli, *Estimateurs a posteriori d'un schéma de volumes finis pour un problème non linéaire*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 475–478.
[7] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer–Verlag, New York, 1994.
[8] B. Brighi, M. Chipot, and E. Gut, *Finite differences on triangular grids*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 567–579.
[9] Z. Q. Cai, *On the finite volume element method*, Numer. Math., 58 (1991), pp. 713–735.
[10] Z. Q. Cai, J. Mandel, and S. McCormick, *The finite volume element method for diffusion equations on general triangulations*, SIAM J. Numer. Anal., 28 (1991), pp. 392–402.
[11] Z. Q. Cai and S. McCormick, *On the accuracy of the finite volume element method for diffusion equations on composite grids*, SIAM J. Numer. Anal., 27 (1990), pp. 636–655.
[12] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.
[13] P. Chatzipantelidis, *A finite volume method based on the Crouzeix–Raviart element for elliptic PDE's in two dimensions*, Numer. Math., 82 (1999), pp. 409–432.
[14] P. Chatzipantelidis, *Finite volume methods for elliptic pde's: A new approach*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 307–324.
[15] S.-H. Chou and Q. Li, *Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach*, Math. Comp., 69 (2000), pp. 103–120.
[16] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[17] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, in Acta Numerica 1995, Cambridge University Press, Cambridge, UK, pp. 105–158.
[18] R. E. Ewing, R. D. Lazarov, and Y. Lin, *Finite volume element approximation of nonlocal reactive flows in porous media*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 285–311.
[19] R. E. Ewing, R. D. Lazarov, and Y. Lin, *Finite volume element approximations of non-local in time one-dimensional flows in porous media*, Computing, 64 (2000), pp. 157–182.
[20] R. E. Ewing, T. Lin, and Y. Lin, *On the accuracy of the finite volume element method based on piecewise linear polynomials*, SIAM J. Numer. Anal., 39 (2002), pp. 1865–1888.
[21] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 713–1020.
[22] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer–Verlag, Berlin, 1983.
[23] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer–Verlag, New York, 1996.
[24] W. Hackbusch, *On first and second order box schemes*, Computing, 41 (1989), pp. 277–296.
[25] H. Jianguo and X. Shitong, *On the finite volume element method for general self–adjoint elliptic problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1762–1774.
[26] D. Kröner, *Numerical Schemes for Conservation Laws*, Adv. Numer. Math., John Wiley and Sons, Chichester, B. G. Teubner, Stuttgart, 1997.
[27] S. Kang and D. Y. Kwak, *Error estimate in L^2 of a covolume method for the generalized Stokes problem*, in Proceedings of the 8th KAIST Math Workshop on Finite Element Methods, 1997, pp. 121–139.
[28] O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics*, Appl. Math. Sci. 49, Springer–Verlag, New York, 1985.
[29] R. D. Lazarov, V. L. Makarov, and W. Weinelt, *On the convergence of difference schemes for the approximation of solutions u from W^m, ($m > 0.5$) of elliptic equations with mixed derivatives*, Numer. Math., 44 (1984), pp. 223–232.
[30] R. J. Leveque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge Text Appl. Math., Cambridge University Press, Cambridge, UK, 2002.

[31] F. Liebau, *The finite volume element method with quadratic basis functions*, Computing, 57 (1996), pp. 281–299.
[32] R. H. Macneal, *An asymmetrical finite difference network*, Quart. Math. Appl., 11 (1953), pp. 295–310.
[33] G. Marchouk and V. Agochkov, *Introduction aux méthodes des élément finis*, Éditions Mir, Moscow, 1985.
[34] K. Mikula and N. Ramarosy, *Semi-implicit finite volume scheme for solving nonlinear diffusion equations in image processing*, Numer. Math., 89 (2001), pp. 1473–1501.
[35] I. D. Mishev, *Finite volume methods for non-definite problems*, Numer. Math., 83 (1999), pp. 161–175.
[36] M. Plexousakis and G. E. Zouraris, *High Order Finite Volume Element Approximations of One Dimensional Indefinite Elliptic Problems*, Technical Report TRITA–NA–0108, NADA, KTH, Stockholm, Sweden, 2001.
[37] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Texts Appl. Math. 13, Springer–Verlag, New York, 1993.
[38] T. J. Rivlin, *An Introduction to the Approximation of Functions*, Dover, New York, 1981.
[39] A. A. Samarskii, R. D. Lazarov, and V. L. Makarov, *Difference Schemes for Differential Equations Having Generalized Solutions*, Visshaya Shkola Publ., Moscow, 1987 (in Russian).
[40] A. H. Schatz, *An observation concerning Ritz–Galerkin methods with indefinite bilinear form*, Math. Comp., 28 (1974), pp. 959–962.
[41] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Ser. Comput. Math. 25, Springer–Verlag, Berlin, 1997.
[42] R. Verfürth, *A posteriori error estimation and adaptive mesh–refinement techniques*, J. Comput. Appl. Math., 50 (1994), pp. 67–83.
[43] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Lecture Notes in Math. 1605, Springer–Verlag, Berlin, 1995.
[44] M. F. Wheeler, *An optimal L_∞ error estimate for Galerkin approximations to solutions of two-point boundary value problems*, SIAM J. Numer. Anal., 10 (1973), pp. 914–917.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1261–1291

CONVERGENCE ANALYSIS OF A MULTIGRID METHOD FOR A CONVECTION-DOMINATED MODEL PROBLEM*

MAXIM A. OLSHANSKII† AND ARNOLD REUSKEN‡

Abstract. The paper presents a convergence analysis of a multigrid solver for a system of linear algebraic equations resulting from the discretization of a convection-diffusion problem using a finite element method. We consider piecewise linear finite elements in combination with a streamline diffusion stabilization. We analyze a multigrid method that is based on canonical intergrid transfer operators, a "direct discretization" approach for the coarse-grid operators and a smoother of line-Jacobi type. A robust (diffusion and h-independent) bound for the contraction number of the two-grid method and the multigrid W-cycle are proved for a special class of convection-diffusion problems, namely with Neumann conditions on the outflow boundary, Dirichlet conditions on the rest of the boundary, and a flow direction that is constant and aligned with gridlines. Our convergence analysis is based on modified smoothing and approximation properties. The arithmetic complexity of one multigrid iteration is optimal up to a logarithmic term.

AMS subject classifications. 65F10, 65N22, 65N30, 65N55

Key words. multigrid, streamline diffusion, convection-diffusion

DOI. 10.1137/S0036142902418679

1. Introduction.

Concerning the theoretical analysis of multigrid methods, different fields of application have to be distinguished. For linear self-adjoint elliptic boundary value problems the convergence theory is well developed (cf. [5, 9, 35, 36]). In other areas the state of the art is (far) less advanced. For example, for convection-diffusion problems the development of a multigrid convergence analysis is still in its infancy. In this paper we present a convergence analysis of a multilevel method for a special class of two-dimensional convection-diffusion problems.

An interesting class of problems for the analysis of multigrid convergence is given by

$$\begin{cases} -\varepsilon\Delta u + b\cdot\nabla u = f & \text{in } \Omega = (0,1)^2, \\ u = g & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

with $\varepsilon > 0$ and $b = (\cos\phi, \sin\phi)$, $\phi \in [0, 2\pi)$. The application of a discretization method results in a large sparse linear system which depends on a mesh size parameter h_k. For a discussion of discretization methods for this problem we refer to [28, 1, 2] and the references therein. Note that in the discrete problem we have three interesting parameters: h_k (mesh size), ε (convection-diffusion ratio), and ϕ (flow direction). For the approximate solution of this type of problems *robust* multigrid methods have been developed which are efficient solvers for a large range of relevant values for the parameters h_k, ε, ϕ. To obtain good robustness properties the components in the multigrid method have to be chosen in a special way because, in general, the

*Received by the editors November 26, 2002; accepted for publication (in revised form) October 28, 2003; published electronically October 28, 2004.
http://www.siam.org/journals/sinum/42-3/41867.html

†Department of Mechanics and Mathematics, Moscow M. V. Lomonosov University, Moscow 119899, Russia (Maxim.Olshanskii@mtu-net.ru). The research of this author was partially supported by Russian Foundation for Basic Research grants 02-01-06715 and 03-01-06460 linked to project 02-01-00592.

‡Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany (reusken@igpm.rwth-aachen.de).

"standard" multigrid approach used for a diffusion problem does not yield satisfactory results when applied to a convection-dominated problem. To improve robustness several modifications have been proposed in the literature, such as "robust" smoothers, matrix-dependent prolongations, and restrictions and semicoarsening techniques. For an explanation of these methods we refer to [9, 33, 4, 13, 14, 18, 19, 37]. These modifications are based on heuristic arguments and empirical studies and rigorous convergence analysis proving robustness is still missing for most of these modifications.

Related to the theoretical analysis of multigrid applied to convection-diffusion problems we note the following. In the literature one finds convergence analyses of multigrid methods for nonsymmetric elliptic boundary value problems which are based on perturbation arguments [6, 9, 17, 32]. If these analyses are applied to the problem in (1.1) the constants in the estimates depend on ε and the results are not satisfactory for the case $\varepsilon \ll 1$. In [11, 25] multigrid convergence for a one-dimensional convection-diffusion problem is analyzed. These analyses, however, are restricted to the one-dimensional case. In [23, 26] convection-diffusion equations as in (1.1) with *periodic* boundary conditions are considered. A Fourier analysis is applied to analyze the convergence of two- or multigrid methods. In [23] the problem (1.1) with periodic boundary conditions and $\phi = 0$ is studied. For the discretization the streamline diffusion finite element method on a uniform grid is used. A bound for the contraction number of a multigrid V-cycle with point Jacobi smoother is proved which is uniform in ε and h_k provided $\varepsilon \sim h_k$ is satisfied. Note that due to the fact that a point Jacobi smoother is used one can not expect robustness of this method for $h_k \gg \varepsilon \downarrow 0$. In [26] a two-grid method for solving a first order upwinding finite difference discretization of the problem (1.1) with periodic boundary conditions is analyzed, and it is proved that the two-grid contraction number is bounded by a constant smaller than one which does not depend on any of the parameters ε, h_k, ϕ. In [3] the application of the hierarchical basis multigrid method to a finite element discretization of problems as in (1.1) is studied. The analysis there shows how the convergence rate depends on ε and on the flow direction, but the estimates are not uniform with respect to the mesh size parameter h_k. In [27] the convergence of a multigrid method applied to a standard finite difference discretization of the problem (1.1) with $\phi = 0$ is analyzed. This method is based on *semicoarsening* and a matrix-dependent prolongation and restriction. It is proved that the multigrid W-cycle has a contraction number smaller than one independent of h_k and ε. The analysis in [27] is based on linear algebra arguments only and is not applicable in a finite element setting. Moreover, the case with *standard coarsening*, which will be treated in the present paper, is not covered by the analysis in [27].

In the present paper we consider the convection-diffusion problem

$$\begin{aligned} -\varepsilon \Delta u + u_x &= f \quad \text{in} \quad \Omega := (0,1)^2, \\ \frac{\partial u}{\partial x} &= 0 \quad \text{on} \quad \Gamma_E := \{\, (x,y) \in \overline{\Omega} \mid x = 1 \,\}, \\ u &= 0 \quad \text{on} \quad \partial\Omega \setminus \Gamma_E. \end{aligned} \tag{1.2}$$

In this problem we have Neumann boundary conditions on the outflow boundary and Dirichlet boundary conditions on the remaining part of the boundary. Hence, the solution may have parabolic layers but exponential boundary layers at the outflow boundary do not occur. For this case an a priori regularity estimate of the form $\|u\|_{H^2} \leq c\,\varepsilon^{-1}\|f\|_{L^2}$ holds, whereas for the case with an exponential boundary layer one only has $\|u\|_{H^2} \leq c\,\varepsilon^{-\frac{3}{2}}\|f\|_{L^2}$. Due to the Dirichlet boundary conditions a Fourier analysis is not applicable.

For the discretization we use conforming linear finite elements. As far as we know there is no multigrid convergence analysis for convection-dominated problems known in the literature that can be applied in a finite element setting with nonperiodic boundary conditions and yields robustness for the parameter range $0 \leq \varepsilon \leq h_k \leq 1$. In this paper we present an analysis which partly fills this gap. We use the streamline diffusion finite element method (SDFEM). The SDFEM ensures a higher order of accuracy than a first order upwind finite difference method (cf. [28, 38]). In SDFEM a mesh-dependent anisotropic diffusion, which acts only in the streamline direction, is added to the discrete problem. Such anisotropy is important for the high order of convergence of this method and also plays a crucial role in our convergence analysis of the multigrid method. In this paper we only treat the case of a uniform triangulation which is taken such that the streamlines are aligned with gridlines. Whether our analysis can be generalized to the situation of an unstructured triangulation is an open question.

We briefly discuss the different components of the multigrid solver.

- For the *prolongation and restriction* we use the canonical intergrid transfer operators that are induced by the nesting of the finite element spaces.
- The hierarchy of *coarse grid discretization operators* is constructed by applying the SDFEM on each grid level. Note that due to the level-dependent stabilization term we have level-dependent bilinear forms and the Galerkin property $A_{k-1} = r_k A_k p_k$ does not hold.
- Related to the *smoother* we note the following. First we emphasize that due to a certain crosswind smearing effect in the finite element discretization the x-line Jacobi or Gauss–Seidel methods do *not* yield robust smoothers (i.e., they do not result in a direct solver in the limit case $\varepsilon = 0$; cf. [9]). This is explained in more detail in Remark 6.1 in section 6. In the present paper we use a smoother of x-line-Jacobi type.

These components are combined in a standard W-cycle algorithm.

The convergence analysis of the multigrid method is based on the framework of the smoothing- and approximation property as introduced by Hackbusch [9, 10]. However, the splitting of the two-grid iteration matrix that we use in our analysis is not the standard one. This splitting is given in (6.8). It turns out to be essential to keep the preconditioner corresponding to the smoother (W_k in (6.8)) as part of the approximation property. Moreover, in the analysis we have to distinguish between residuals which after presmoothing are zero close to the inflow boundary and those that are nonzero. This is done by using a cut-off operator (Φ_k in (6.8)). The main reason for this distinction is the following. As is usually done in the analysis of the approximation property we use finite element error bounds combined with regularity results. In the derivation of a L^2 bound for the finite element discretization error we use a duality argument. However, the formal dual problem has poor regularity properties, since the inflow boundary of the original problem is the outflow boundary of the dual problem. Thus Dirichlet outflow boundary conditions would appear and we obtain poor estimates due to the poor regularity. To avoid this, we consider a dual problem with Neumann outflow and Dirichlet inflow conditions. To be able to deal with the inconsistency caused by these "wrong" boundary conditions we assume the input residuals for the coarse grid correction to be zero near the inflow boundary. Numerical experiments from section 11 related to the approximation property show that such analysis is sharp.

In our estimates there are terms that grow logarithmically if the mesh size parameter h_k tends to zero. To compensate this the number of presmoothings has to

be taken level dependent. This then results in a two-grid method with a contraction number $\|T_k\|_{A^TA} \leq c < 1$ and a complexity $\mathcal{O}(N_k(\ln N_k)^4)$, with $N_k = h_k^{-2}$. Using standard arguments we obtain a similar convergence result for the multigrid W-cycle.

The remainder of this paper is organized as follows. In section 2 we give the weak formulation of the problem (1.2) and describe the SDFEM. In section 3 some useful properties of the stiffness matrix are derived. In section 4 we prove some a priori estimates for the continuous and the discrete solution. In section 5 we derive quantitative results concerning the upstream influence of a right-hand side on the solution. These results are needed in the proof of the modified approximation property. *Section 6 contains the main results of this paper*. In this section we describe the multigrid algorithm and present the convergence analysis. In sections 7–10 we give proofs of some important results that are used in the analysis in section 6. In section 11 we present results of a few numerical experiments.

2. The continuous problem and its discretization. For the weak formulation of the problem (1.2) we use the $L^2(\Omega)$ scalar product which is denoted by $(\cdot,\cdot)$. For the corresponding norm we use the notation $\|\cdot\|$. With the Sobolev space $\mathbf{V} := \{\, v \in H^1(\Omega) \mid v = 0 \ \text{ on } \ \partial\Omega \setminus \Gamma_E \,\}$ the weak formulation is as follows: find $u \in \mathbf{V}$ such that

$$a(u,v) := \varepsilon(u_x, v_x) + \varepsilon(u_y, v_y) + (u_x, v) = (f, v) \ \text{ for all } \ v \in \mathbf{V}. \tag{2.1}$$

From the Lax–Milgram lemma it follows that a unique solution of this problem exists. For the discretization we use linear finite elements on a uniform triangulation. For this we use a mesh size $h_k := 2^{-k}$ and grid points $x_{i,j} = (ih_k, jh_k)$, $0 \leq i,j \leq h_k^{-1}$. A uniform triangulation is obtained by inserting diagonals that are oriented from southwest to northeast. Let $\mathbb{V}_k \subset \mathbf{V}$ be the space of continuous functions that are piecewise linear on this triangulation and have zero values on $\partial\Omega \setminus \Gamma_E$. For the discretization of (2.1) we consider the SDFEM: find $u_k \in \mathbb{V}_k$ satisfying

$$(\varepsilon + \delta_k h_k)((u_k)_x, v_x) + \varepsilon((u_k)_y, v_y) + ((u_k)_x, v) = (f, v + \delta_k h_k v_x) \ \text{ for all } \ v \in \mathbb{V}_k \tag{2.2}$$

with

$$\delta_k = \begin{cases} \bar{\delta} & \text{if } \ \frac{h_k}{2\varepsilon} \geq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

The stabilization parameter $\bar{\delta}$ is a given constant of order 1. For an analysis of the SDFEM we refer to [28, 15]. In this paper we assume

$$\bar{\delta} \in \left[\frac{1}{3}, 1\right]. \tag{2.4}$$

The value $\frac{1}{3}$ for the lower bound is important for our analysis. The choice of 1 for the upper bound is made for technical reasons and this value is rather arbitrary. The finite element formulation (2.2) gives rise to the (stabilized) bilinear form

$$a_k(u,v) := (\varepsilon + \delta_k h_k)(u_x, v_x) + \varepsilon(u_y, v_y) + (u_x, v), \quad u, v \in \mathbf{V}. \tag{2.5}$$

Note the following relation for the bilinear form $a_k(\cdot,\cdot)$:

$$a_k(v,v) = \varepsilon\|v_y\|^2 + (\varepsilon + \delta_k h_k)\|v_x\|^2 + \frac{1}{2}\int_{\Gamma_E} v^2\, dy \quad \text{for } \ v \in \mathbf{V}. \tag{2.6}$$

The main topic of this paper is a convergence analysis of a multigrid solver for the algebraic system of equations that corresponds to (2.2). In this convergence analysis the particular form of the right-hand side in (2.2), which is essential for consistency in the SDFEM, does not play a role. Therefore for an arbitrary $f \in \mathrm{L}^2(\Omega)$ we will consider the problems

$$u \in \mathbf{V} \quad \text{such that} \quad a_k(u,v) = (f,v) \quad \text{for all} \quad v \in \mathbf{V}, \tag{2.7}$$

$$u_k \in \mathbb{V}_k \quad \text{such that} \quad a_k(u_k,v_k) = (f,v_k) \quad \text{for all} \quad v_k \in \mathbb{V}_k. \tag{2.8}$$

Note that u and u_k depend on the stabilization term in the bilinear form and that these solutions differ from those in (2.1) and (2.2).

3. Representation of the stiffness matrix. We now derive a representation of the stiffness matrix corresponding to the bilinear form $a_k(\cdot,\cdot)$ that will be used in the analysis below. The standard nodal basis in $\mathbb{V}_k$ is denoted by $\{\phi_\ell\}_{1\le\ell\le N_k}$ with N_k the dimension of the finite element space, $N_k := h_k^{-1}(h_k^{-1}-1)$. Define the isomorphism:

$$P_k : X_k := \mathbb{R}^{N_k} \to \mathbb{V}_k, \quad P_k x = \sum_{i=1}^{N_k} x_i \phi_i.$$

On X_k we use a scaled Euclidean scalar product $\langle x,y\rangle_k = h_k^2 \sum_{i=1}^{N_k} x_i y_i$ and corresponding norm denoted by $\|\cdot\|$ (note that this notation is also used to denote the $\mathrm{L}^2(\Omega)$ norm). The adjoint $P_k^* : \mathbb{V}_k \to X_k$ satisfies $(P_k x, v) = \langle x, P_k^* v\rangle_k$ for all $x \in X_k,\ v \in \mathbb{V}_k$. The following norm equivalence holds:

$$C^{-1}\|x\| \le \|P_k x\| \le C\|x\| \quad \text{for all} \quad x \in \mathrm{X}_k, \tag{3.1}$$

with a constant C independent of k. The stiffness matrix A_k on level k is defined by

$$\langle A_k x, y\rangle_k = a_k(P_k x, P_k y) \quad \text{for all} \quad x,y \in \mathrm{X}_k. \tag{3.2}$$

In an interior grid point the discrete problem has the stencil

$$\frac{1}{h_k^2}\begin{bmatrix} 0 & -\varepsilon & 0 \\ -\varepsilon_k & 2(\varepsilon_k+\varepsilon) & -\varepsilon_k \\ 0 & -\varepsilon & 0 \end{bmatrix} + \frac{1}{h_k}\begin{bmatrix} 0 & -\frac{1}{6} & \frac{1}{6} \\ -\frac{1}{3} & 0 & \frac{1}{3} \\ -\frac{1}{6} & \frac{1}{6} & 0 \end{bmatrix}, \qquad \varepsilon_k := \varepsilon + \delta_k h_k\ . \tag{3.3}$$

For a matrix representation of the discrete operator we first introduce some notation and auxiliary matrices. Let $n_k := h_k^{-1}$ and

$$\hat{D}_x := \frac{1}{h_k}\mathrm{tridiag}(-1,1,0) \in \mathbb{R}^{n_k\times n_k}\ ,$$

$$\hat{A}_x := \hat{D}_x^T\hat{D}_x = \frac{1}{h_k^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{n_k\times n_k}\ ,$$

$$\hat{A}_y := \frac{1}{h_k^2}\mathrm{tridiag}(-1,2,-1) \in \mathbb{R}^{(n_k-1)\times(n_k-1)}\ ,$$

$$\hat{J} := \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & \frac{1}{2} \end{pmatrix} \in \mathbb{R}^{n_k\times n_k}, \quad \hat{T} := \mathrm{tridiag}(0,0,1) \in \mathbb{R}^{n_k\times n_k}\ .$$

Furthermore, let I_m be the $m \times m$ identity matrix. We finally introduce the following $N_k \times N_k$ matrices

$$D_x := I_{n_k-1} \otimes \hat{D}_x \ , \quad A_x := I_{n_k-1} \otimes \hat{A}_x = D_x^T D_x \ , \quad A_y := \hat{A}_y \otimes \hat{J}$$

and the $N_k \times N_k$ blocktridiagonal matrix

$$B := \text{blocktridiag}(I_{n_k}, 4I_{n_k}, \hat{T}) \ .$$

Using all this notation we consider the following representation for the stiffness matrix A_k in (3.2):

$$A_k = \left(\varepsilon + \left(\delta_k - \frac{1}{3}\right) h_k\right) A_x + \varepsilon A_y + \frac{1}{6} B D_x. \tag{3.4}$$

The latter decomposition can be written in stencil notation as

$$\frac{\bar{\varepsilon}_k}{h_k^2} \begin{bmatrix} 0 & 0 & 0 \\ -1 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix} + \frac{\varepsilon}{h_k^2} \begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{bmatrix} + \frac{1}{6h_k} \begin{bmatrix} 0 & -1 & 1 \\ -4 & 4 & 0 \\ -1 & 1 & 0 \end{bmatrix} \tag{3.5}$$

with $\bar{\varepsilon}_k = \varepsilon + (\delta_k - \frac{1}{3})h_k > 0$.

Some properties of the matrices used in the decomposition (3.4) are collected in the following lemma.

For $B, C \in \mathbb{R}^{n \times n}$ we write $B \geq C$ iff $x^T B x \geq x^T C x$ for all $x \in \mathbb{R}^n$.

LEMMA 3.1. *The following inequalities hold:*

$$A_x D_x^{-1} \geq 0, \tag{3.6}$$

$$A_y D_x^{-1} \geq 0, \tag{3.7}$$

$$B \geq 2I, \tag{3.8}$$

$$A_k D_x^{-1} \geq \frac{1}{3} I, \tag{3.9}$$

$$\|D_x A_k^{-1}\| \leq 3. \tag{3.10}$$

Proof. To check (3.6) observe $A_x D_x^{-1} = D_x^T D_x D_x^{-1} = D_x^T$. Now note that $D_x^T + D_x$ is symmetric positive definite.

To prove (3.7) it suffices to show that $D_x^T A_y \geq 0$ holds. We have

$$K := D_x^T A_y = (I_{n_k-1} \otimes \hat{D}_x^T)(\hat{A}_y \otimes \hat{J}) = \hat{A}_y \otimes \tilde{D}_x^T \ ,$$

with the matrix

$$\tilde{D}_x^T = \frac{1}{h_k} \begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -\frac{1}{2} \\ & & & \frac{1}{2} \end{pmatrix}.$$

Hence in the matrix $K + K^T = \hat{A}_y \otimes (\tilde{D}_x^T + \tilde{D}_x)$ both factors $\hat{A}_y$ and $\tilde{D}_x^T + \tilde{D}_x$ are symmetric positive definite. From this the result follows.

To prove (3.8) we define $R := B - 4I$ and note that $\|R\|^2 \leq \|R\|_\infty \|R\|_1 \leq 4$. Using this we get

$$\langle Bx, x\rangle_k = 4\|x\|^2 + \langle Rx, x\rangle_k \geq 4\|x\|^2 - \|R\|\|x\|^2 \geq 2\|x\|^2$$

which proves the desired result. Inequality (3.9) follows immediately from the representation of A_k in (3.4) and inequalities (3.6)–(3.8). From the result in (3.9) it follows that $D_x^T A_k \geq \frac{1}{3} D_x^T D_x$. This implies $\|D_x x\|^2 \leq 3\langle A_k x, D_x x\rangle_k \leq 3\|A_k x\|\|D_x x\|$ for all $x \in \mathrm{X}_k$ and thus estimate (3.10) is also proved. ∎

4. A priori estimates. In this paper we study the convergence of a multigrid method for solving the system of equations

$$A_k x_k = b, \tag{4.1}$$

with A_k the stiffness matrix from the previous section. As already noted in the introduction, our analysis relies on smoothing and approximation properties. For establishing a suitable approximation property we will use regularity results and a priori estimates for solutions of the continuous and the discrete problems. Such results are collected in this section. In the remainder of the paper we restrict ourselves to the convection-dominated case.

Assumption 4.1. We consider only values of k and ε such that $\varepsilon \leq \frac{1}{2} h_k$.

If instead of the factor $\frac{1}{2}$ in this assumption we take another constant C, our analysis can still be applied but some technical modifications are needed (to distinguish between $\delta_k = \bar{\delta}$ and $\delta_k = 0$) which make the presentation less transparent.

We consider this convection-dominated case to be the most interesting one. Many results that will be presented also hold for the case of an arbitrary positive ε but the proofs for the diffusion-dominated case often differ from those for the convection-dominated case. In view of the presentation we decided to treat only the convection-dominated case. Note that then

$$\delta_k = \bar{\delta} \in \left[\frac{1}{3}, 1\right] \quad \text{and} \quad \frac{1}{3} h_k \leq \varepsilon_k = \varepsilon + \bar{\delta} h_k \leq \frac{3}{2} h_k. \tag{4.2}$$

For the inflow boundary we use the notation $\Gamma_W := \{(x, y) \in \overline{\Omega} \mid x = 0\}$. For the continuous solution u the following a priori estimates hold.

THEOREM 4.1. *For $f \in L_2(\Omega)$ let u be the solution of* (2.7). *There is a constant c independent of k and ε such that*

$$\|u\| + \|u_x\| \leq c\|f\|, \tag{4.3}$$

$$\sqrt{\varepsilon}\|u_y\| \leq c\|f\|, \tag{4.4}$$

$$h_k\|u_{xx}\| + \sqrt{\varepsilon h_k}\|u_{xy}\| + \varepsilon\|u_{yy}\| \leq c\|f\|, \tag{4.5}$$

$$\int_{\Gamma_E} u^2\, dy + h_k \int_{\Gamma_W} u_x^2\, dy + \varepsilon \int_{\Gamma_E} u_y^2\, dy \leq c\|f\|^2. \tag{4.6}$$

Proof. Since $f \in L_2(\Omega)$, the regularity theory from [8] ensures that the solution u of (2.7) belongs to $\mathrm{H}^2(\Omega)$. Hence we can consider the strong formulation of (2.7),

$$-\varepsilon u_{yy} - \varepsilon_k u_{xx} + u_x = f, \tag{4.7}$$

with boundary conditions as in (1.2). Now we multiply (4.7) with u_x and integrate by parts. Taking boundary conditions into account, we get the following terms:

$$\begin{aligned}
-\varepsilon(u_{yy},u_x) &= \frac{\varepsilon}{2}((u_y^2)_x,1) = \frac{\varepsilon}{2}\int_{\Gamma_E} u_y^2\,dy,\\
-\varepsilon_k(u_{xx},u_x) &= -\frac{\varepsilon_k}{2}((u_x^2)_x,1) = \frac{\varepsilon_k}{2}\int_{\Gamma_W} u_x^2\,dy \geq c\,h_k\int_{\Gamma_W} u_x^2\,dy \quad \text{(we use (4.2))},\\
(u_x,u_x) &= \|u_x\|^2 \geq \|u\|^2,\\
(f,u_x) &\leq \frac{1}{2}\|f\|^2 + \frac{1}{2}\|u_x\|^2.
\end{aligned}$$

From these relations the results (4.3) and (4.6), except the bound for $\int_{\Gamma_E} u^2\,dy$, easily follow. Next we multiply (4.7) with u and integrate by parts to obtain

$$\varepsilon\|u_y\|^2 + \varepsilon_k\|u_x\|^2 + \frac{1}{2}\int_{\Gamma_E} u^2\,dy = (f,u) \leq \|f\|\|u\| \leq c\,\|f\|^2 \quad \text{(we use (4.3))}.$$

Estimate (4.4) and the remainder of (4.6) now follow. To prove (4.5) we introduce $F = f - u_x$. Due to (4.3) we have $\|F\| \leq c\,\|f\|$. Moreover $-\varepsilon u_{yy} - \varepsilon_k u_{xx} = F$ holds. If we square both sides of this equality and integrate over Ω we obtain

$$\varepsilon^2\|u_{yy}\|^2 + 2\varepsilon\varepsilon_k(u_{yy},u_{xx}) + \varepsilon_k^2\|u_{xx}\|^2 = \|F\|^2 \leq c\,\|f\|^2. \tag{4.8}$$

Further note that for any sufficiently smooth function v, satisfying the boundary conditions in (1.2), the relations

$$v_{xx}(x,0) = v_{xx}(x,1) = 0,\ x \in (0,1), \quad v_y(0,y) = v_{xy}(1,y) = 0,\ y \in (0,1),$$

hold, and thus

$$(v_{yy},v_{xx}) = -(v_y,v_{xxy}) = (v_{xy},v_{xy}).$$

Using a standard density argument we conclude that for the solution $u \in \mathrm{H}^2(\Omega)$ of (2.7) the relation $(u_{yy},u_{xx}) = (u_{xy},u_{xy})$ holds. Now (4.8) gives

$$\varepsilon^2\|u_{yy}\|^2 + 2\varepsilon\varepsilon_k\|u_{xy}\|^2 + \varepsilon_k^2\|u_{xx}\|^2 \leq c\,\|f\|^2.$$

In combination with (4.2) this yields (4.5). ∎

The next lemma states that the x-derivative of the *discrete* solution is also uniformly bounded if the right-hand side is from $\mathbb{V}_k$.

LEMMA 4.2. *For $f_k \in \mathbb{V}_k$ let $u_k \in \mathbb{V}_k$ be a solution to* (2.8)*; then*

$$\|(u_k)_x\| \leq c\,\|f_k\|. \tag{4.9}$$

Proof. The result in (4.9) follows from the estimate (3.10) in Lemma 3.1. To show this we need some technical considerations.

First we show how the size of the x-derivative of a finite element function $v \in \mathbb{V}_k$ can be determined from its corresponding coefficient vector $P_k^{-1}v \in X_k$. Let $\mathcal{I}$ be the index set $\{(i,j) \mid 0 \leq i \leq n_k - 1,\ \ 1 \leq j \leq n_k - 1\}$. For $(i,j) \in \mathcal{I}$ let $T^l_{(i,j)}$ and $T^u_{(i,j)}$ be the two triangles in the triangulation which have the line between the grid

points $x_{i,j}$ and $x_{i+1,j}$ as a common edge. Let $v \in \mathbb{V}_k$ be given. For $1 \le j \le n_k - 1$ we introduce the vector $\mathbf{v}_j = (v(x_{1,j}), \dots, v(x_{n_k,j}))^T$. We then obtain

$$\begin{aligned}\|v_x\|^2 &= \sum_{(i,j)\in\mathcal{I}} \left(\int_{T^l_{(i,j)}} v_x^2 \, dxdy + \int_{T^u_{(i,j)}} v_x^2 \, dxdy \right) \\ &= \sum_{(i,j)\in\mathcal{I}} \left(\frac{v(x_{i+1,j}) - v(x_{i,j})}{h_k} \right)^2 h_k^2 = h_k^2 \sum_{1\le j\le n_k-1} \big(D_x\mathbf{v}_j\big)^T (D_x\mathbf{v}_j) \\ &= h_k^2 \big(D_x P_k^{-1} v\big)^T \big(D_x P_k^{-1} v\big) = \|D_x P_k^{-1} v\|^2.\end{aligned}$$

Therefore

$$\|v_x\| = \|D_x P_k^{-1} v\| \quad \text{for any} \quad v \in \mathbb{V}_k. \tag{4.10}$$

For the discrete solution of (2.8) with $f = f_k$ we have the representation $u_k = P_k A_k^{-1} P_k^* f_k$. Now from (3.10) and (4.10) it follows that

$$\|(u_k)_x\| = \|D_x A_k^{-1} P_k^* f_k\| \le 3 \, \|P_k^* f_k\| \le c \, \|f_k\|$$

with a constant c independent of k and ε. ∎

The next lemma gives some bounds on the difference between discrete and continuous solutions

LEMMA 4.3. *Define the error $e_k = u - u_k$, where u and u_k are solutions of the problems* (2.7) *and* (2.8) *with right-hand side $f = f_k \in \mathbb{V}_k$. Then the following estimates hold:*

$$\|(e_k)_x\| \le c\|f_k\| \tag{4.11}$$

$$\varepsilon\|(e_k)_y\|^2 + \frac{1}{2}\int_{\Gamma_E} e_k^2 \, dy \le \, c\frac{h_k^2}{\varepsilon}\|f_k\|^2. \tag{4.12}$$

Proof. Estimate (4.11) directly follows from (4.3) and (4.9) by a triangle inequality. The proof of (4.12) is based on standard arguments: the Galerkin orthogonality, approximation properties of $\mathbb{V}_k$, and a priori estimates from (4.5). Indeed

$$\begin{aligned}&\varepsilon\|(e_k)_y\|^2 + (\varepsilon + \bar\delta h_k)\|(e_k)_x\|^2 + \frac{1}{2}\int_{\Gamma_E} e_k^2 \, dy = a_k(e_k, e_k) = \inf_{v_k\in\mathbb{V}_k} a_k(e_k, u - v_k) \\ &\quad\le \inf_{v_k\in\mathbb{V}_k} \big(\varepsilon\|(e_k)_y\|\|(u-v_k)_y\| + (\varepsilon + \bar\delta h_k)\|(e_k)_x\|\|(u-v_k)_x\| + \|(e_k)_x\|\|u - v_k\|\big) \\ &\quad\le c\,(\varepsilon\, h_k\|(e_k)_y\|\|u\|_{H^2} + h_k^2\|(e_k)_x\|\|u\|_{H^2}) \\ &\quad\le c\left(h_k\|(e_k)_y\|\|f_k\| + \frac{h_k^2}{\varepsilon}\|f_k\|^2\right) \le \frac{\varepsilon}{2}\|(e_k)_y\|^2 + c\frac{h_k^2}{\varepsilon}\|f_k\|^2.\end{aligned}$$

The estimate (4.12) follows. ∎

5. Upstream influence of the streamline diffusion method. Consider the continuous problem (2.7). The goal of this section is to estimate the upstream influence of the right-hand side function f on the solution u. The same will be done for the corresponding discrete problem. In the literature, results of such type are known for the problem with Dirichlet boundary conditions and typically formulated

in the form of estimates on the (discrete) Greens function (see, e.g., [31, 20, 16]). A typical result is that the value of the solution at a point x is essentially determined by the values of the right-hand side in a "small" strip that contains x. This strip has a crosswind width of size $O(\varepsilon^*|\ln h|)$, where $\varepsilon^* = \max\{\varepsilon, h^{\frac{3}{2}}\}$, and in the streamline direction it ranges from the inflow boundary to a $\mathcal{O}(h|\ln h|)$ upstream distance from x. In our analysis we need precise quantitative results for the case with Neumann outflow boundary conditions. In the literature we did not find such results. Hence we present proofs of the results that are needed for the multigrid convergence analysis further on. Our analysis uses the known technique of cut-off functions (e.g., [7, 16]), it avoids the use of an adjoint problem and is based on the following lemma.

LEMMA 5.1. *For $\varepsilon_k = \varepsilon + \bar{\delta} h_k$ assume a function $\phi \in H^1_\infty(0,1)$, such that $0 \le -\varepsilon_k \phi_x \le \phi$. Denote by $\|\cdot\|_\phi$ a semi-norm induced by the scalar product $(\phi\cdot,\cdot)$. Then the solution u of (2.7) satisfies*

$$\|u_x\|_\phi \le 2\|f\|_\phi, \tag{5.1}$$

$$\varepsilon_k \ \phi(0) \int_{\Gamma_W} u_x^2 \, dy \le \|f\|_\phi^2, \tag{5.2}$$

$$\frac{1}{4}\|u\|^2_{-\phi_x} + \varepsilon \|u_y\|_\phi^2 \le (\phi\, f, u). \tag{5.3}$$

Proof. We consider the strong formulation (4.7) and multiply it with ϕu_x and integrate by parts. We then get the following terms:

$$\begin{aligned}
-\varepsilon(u_{yy}, \phi u_x) &= \frac{\varepsilon}{2}\|u_y\|^2_{-\phi_x} + \frac{\varepsilon}{2}\phi(1)\int_{\Gamma_E} u_y^2 \, dy \ge 0,\\
-\varepsilon_k(u_{xx}, \phi u_x) &= -\frac{\varepsilon_k}{2}\|u_x\|^2_{-\phi_x} + \frac{\varepsilon_k}{2}\phi(0)\int_{\Gamma_W} u_x^2 \, dy \ge -\frac{1}{2}\|u_x\|_\phi^2 + \frac{\varepsilon_k}{2}\phi(0)\int_{\Gamma_W} u_x^2 \, dy,\\
(u_x, \phi u_x) &= \|u_x\|_\phi^2,\\
(f, \phi u_x) &\le \|f\|_\phi \|u_x\|_\phi \le \|f\|_\phi^2 + \frac{1}{4}\|u_x\|_\phi^2.
\end{aligned}$$

Now (5.1) and (5.2) immediately follow. To obtain the estimate (5.3) we multiply (4.7) with $\phi\, u$ and integrate by parts. We get the following terms:

$$\begin{aligned}
-\varepsilon(u_{yy}, \phi\, u) &= \varepsilon\|u_y\|_\phi^2,\\
-\varepsilon_k(u_{xx}, \phi\, u) &= \varepsilon_k\|u_x\|_\phi^2 + \varepsilon_k(u_x, \phi_x\, u)\\
&\ge \varepsilon_k\|u_x\|_\phi^2 - \varepsilon_k^2\|u_x\|^2_{-\phi_x} - \frac{1}{4}\|u\|^2_{-\phi_x} \ge -\frac{1}{4}\|u\|^2_{-\phi_x},\\
(u_x, \phi\, u) &= \frac{1}{2}\|u\|^2_{-\phi_x} + \frac{\phi(1)}{2}\int_{\Gamma_E} u^2 \, dy.
\end{aligned}$$

Thus (5.3) follows. ∎

For arbitrary $\xi \in [0,1]$ consider the function

$$\phi_\xi(x) = \begin{cases} 1 & \text{for } x \in [0,\xi],\\ \exp\left(-\frac{x-\xi}{\varepsilon_k}\right) & \text{for } x \in (\xi, 1]. \end{cases}$$

For any ξ the function $\phi_\xi(x)$ satisfies the assumptions of Lemma 5.1. For $0 < \xi < \eta < 1$ we define the domains

$$\Omega_\xi = \{(x,y) \in \Omega : x < \xi\}, \quad \Omega_\eta = \{(x,y) \in \Omega : x > \eta\}.$$

Direct application of Lemma 5.1 with $\phi = \phi_\xi$ gives the following corollary.

COROLLARY 5.2. *Consider $f \in L_2(\Omega)$ such that* $\mathrm{supp}(f) \in \Omega_\eta$ *and let u be the corresponding solution of problem* (2.7). *Assume $\eta - \xi \geq 2\,\varepsilon_k\, p\,|\ln h_k|,\ p > 0$. Then we have*

$$\|u_x\|_{L_2(\Omega_\xi)} \leq h_k^p \|f\|, \tag{5.4}$$

$$\varepsilon_k \int_{\Gamma_W} u_x^2\, dy \leq h_k^{2p} \|f\|^2, \tag{5.5}$$

$$\sqrt{\varepsilon}\|u_y\|_{L_2(\Omega_\xi)} \leq \sqrt{\varepsilon_k}\, h_k^p \|f\|. \tag{5.6}$$

Proof. The estimate $\|f\|_\phi^2 = (\phi f, f)_{\Omega_\eta} \leq \phi(\eta)\|f\|_{\Omega_\eta}^2 = h_k^{2p}\|f\|^2$ and (5.1), (5.2) imply the results (5.4) and (5.5). We also have

$$\begin{aligned}(\phi f, u) = (\phi f, u)_{\Omega_\eta} &\leq \varepsilon_k \|f\|_\phi^2 + \frac{1}{4\varepsilon_k}(\phi u, u)_{\Omega_\eta} = \varepsilon_k \|f\|_\phi^2 + \frac{1}{4}(-\phi_x u, u)_{\Omega_\eta} \\ &\leq \varepsilon_k \|f\|_\phi^2 + \frac{1}{4}\|u\|_{-\phi_x}^2.\end{aligned}$$

Together with (5.3) this yields (5.6). ∎

We need an analogue of estimate (5.1) for the finite element solution u_k of (2.8). To this end consider a vector $\phi = (\phi_0, \ldots, \phi_{n_k})$, such that $\phi_i > 0$ for all i and

$$0 \leq -\varepsilon_k \frac{\phi_i - \phi_{i-1}}{h_k} \leq c_0 \phi_i, \quad i = 1, \ldots, n_k, \tag{5.7}$$

with a constant $c_0 \in (0, \frac{4}{9})$ and $\varepsilon_k = \varepsilon + \bar{\delta} h_k$.

Define $\hat{\Phi}_k := \mathrm{diag}(\phi_i)_{1 \leq i \leq n_k}$, $\Phi_k := I_{n_k - 1} \otimes \hat{\Phi}_k$ with ϕ_i satisfying (5.7). Let $\langle \cdot, \cdot \rangle_\Phi = \langle \Phi_k \cdot, \cdot \rangle_k$.

LEMMA 5.3. *There exists a constant $c > 0$ independent of k and ε such that*

$$\langle A_k x, D_x x \rangle_\Phi \geq c\, \|D_x x\|_\Phi^2 \quad \textit{for all} \ \ x \in X_k.$$

Proof. We use similar arguments as in the proof of (3.10). We use the representation (3.4) of the stiffness matrix: $A_k = \bar{\varepsilon}_k A_x + \varepsilon A_y + \frac{1}{6} B D_x$. Note that

$$D_x^T \Phi_k A_y = (I_{n_k - 1} \otimes \hat{D}_x^T)(I_{n_k - 1} \otimes \hat{\Phi}_k)(\hat{A}_y \otimes \hat{J}) = \hat{A}_y \otimes \hat{D}_x^T \hat{\Phi}_k \hat{J}.$$

The matrix $\hat{A}_y$ is symmetric positive definite. Using $\phi_i \leq \phi_{i-1}$ and a Gershgorin theorem it follows that $\hat{D}_x^T \hat{\Phi}_k \hat{J} + \hat{J} \hat{\Phi}_k \hat{D}_x$ is symmetric positive definite, too. Hence, $D_x^T \Phi_k A_y \geq 0$ holds, i.e.,

$$\langle A_y x, D_x x \rangle_\Phi \geq 0 \quad \text{for all} \ \ x \in X_k. \tag{5.8}$$

From the assumption on ϕ it follows that $\phi_{i-1} \leq (1 + \frac{c_0 h_k}{\varepsilon_k})\phi_i$ for all i. Using this and the relation

$$\frac{1}{2}(\hat{\Phi}_k^{\frac{1}{2}} \hat{D}_x^T \hat{\Phi}_k^{-\frac{1}{2}} + \hat{\Phi}_k^{-\frac{1}{2}} \hat{D}_x \hat{\Phi}_k^{\frac{1}{2}}) = \frac{1}{2h_k} \mathrm{tridiag}\left(\sqrt{\frac{\phi_{i-1}}{\phi_i}}\,,\, 2\,,\, \sqrt{\frac{\phi_i}{\phi_{i+1}}} \right)$$

it follows that

$$\Phi_k^{\frac{1}{2}} D_x^T \Phi_k^{-\frac{1}{2}} \geq \frac{1}{2h_k}\left(2 - 2\sqrt{1+\frac{c_0 h_k}{\varepsilon_k}}\right) I \geq -\frac{c_0}{2\varepsilon_k} I \geq -\frac{c_0}{2\bar{\varepsilon}_k} I$$

holds. And thus

$$\bar{\varepsilon}_k \langle A_x x, D_x x\rangle_\Phi = \bar{\varepsilon}_k \langle \Phi_k D_x^T D_x x, D_x x\rangle \geq -\frac{1}{2} c_0 \langle D_x x, D_x x\rangle_\Phi \quad \text{for all } \ x \in X_k. \tag{5.9}$$

We decompose B as $B = 4I - R$. A simple computation yields

$$\|\Phi_k^{\frac{1}{2}} R \Phi_k^{-\frac{1}{2}}\|_1 \leq 1 + \sqrt{1+\frac{c_0 h_k}{\varepsilon_k}} \leq 1 + \sqrt{1+3c_0} \leq 2 + \frac{3}{2} c_0.$$

Similarly we get $\|\Phi_k^{\frac{1}{2}} R \Phi_k^{-\frac{1}{2}}\|_\infty \leq 2 + \frac{3}{2} c_0$ and thus $\|\Phi_k^{\frac{1}{2}} R \Phi_k^{-\frac{1}{2}}\| \leq 2 + \frac{3}{2} c_0$. Hence

$$\Phi_k^{\frac{1}{2}} B \Phi_k^{-\frac{1}{2}} \geq \left(4 - \left(2 + \frac{3}{2} c_0\right)\right) I = \left(2 - \frac{3}{2} c_0\right) I$$

and thus

$$\frac{1}{6}\langle B D_x x, D_x x\rangle_\Phi \geq \left(\frac{1}{3} - \frac{1}{4} c_0\right) \langle D_x x, D_x x\rangle_\Phi \quad \text{for all } \ x \in X_k. \tag{5.10}$$

Combination of the results in (5.8), (5.9), and (5.10) yields

$$\langle A_k x, D_x x\rangle_\Phi \geq \left(\frac{1}{3} - \frac{3}{4} c_0\right) \langle D_x x, D_x x\rangle_\Phi \geq c \langle D_x x, D_x x\rangle_\Phi \quad \text{for all } \ x \in X_k$$

with a constant $c > 0$ (use that $c_0 \in (0, \frac{4}{9})$). ∎

LEMMA 5.4. *For $f = f_k \in \mathbb{V}_k$ let u_k be the solution of the problem* (2.8)*. Then*

$$\sum_{i=1}^{n_k} \sum_{j=1}^{n_k - 1} h_k^2 \phi_i \left(\frac{u_{i,j} - u_{i-1,j}}{h_k}\right)^2 \leq C \sum_{i=1}^{n_k} \sum_{j=1}^{n_k - 1} h_k^2 \phi_i (M_k \hat{f})_{i,j}^2 \tag{5.11}$$

holds. Here u_{ij} is the nodal value of u_k at the grid point $x_{i,j}$, $\hat{f}$ is the vector of nodal values of f_k, M_k is the mass matrix, and ϕ_i satisfies (5.7)*.*

Proof. Let $\hat{u}_k = P_k^{-1} u_k \in X_k$ be the vector of nodal values of u_k; then

$$A_k \hat{u}_k = M_k \hat{f} =: \hat{b}_k. \tag{5.12}$$

The diagonal matrices Φ_k and $\hat{\Phi}_k$ are as in Lemma 5.3. The statement of the lemma is equivalent to $\langle \Phi_k D_x \hat{u}_k, D_x \hat{u}_k\rangle_k \leq c \langle \Phi_k \hat{b}_k, \hat{b}_k\rangle_k$, with a constant c that is independent of $\hat{b}_k$. This is the same as

$$\|D_x A_k^{-1}\|_\Phi \leq c. \tag{5.13}$$

Note that (5.13) is a generalization of the result in (3.10). From Lemma 5.3 we obtain

$$\|D_x x\|_{\hat{\Phi}}^2 < \frac{1}{c}\langle A_k x, D_x x\rangle_{\hat{\Phi}} \leq \frac{1}{c}\|A_k x\|_\Phi \|D_x x\|_\Phi \quad \text{for all } \ x \in X_k;$$

thus $\|D_x x\|_\Phi \leq \tilde{c}\|A_k x\|_\Phi$ for all x. Hence we have proved the result in (5.13). ∎

For the discrete case we consider

$$\phi_i^\xi = \begin{cases} 1 & \text{for } ih_k \in [0,\xi], \\ \exp\left(-\frac{ih_k-\xi}{4h_k}\right) & \text{for } ih_k > \xi. \end{cases} \tag{5.14}$$

It is straightforward to check that $-(\phi_i^\xi - \phi_{i-1}^\xi) = (\exp(\frac{1}{4}) - 1)\,\phi_i^\xi$ if $ih_k > \xi$. Therefore, using $\varepsilon_k \le \frac{3}{2}h_k$,

$$0 \le -\varepsilon_k \frac{\phi_i^\xi - \phi_{i-1}^\xi}{h_k} \le \frac{3}{2}\left(\exp\left(\frac{1}{4}\right) - 1\right)\phi_i^\xi, \quad i = 1, 2, \ldots . \tag{5.15}$$

For any ξ the vector ϕ_i^ξ, $1 \le i \le n_k$, satisfies the condition (5.7) with $c_0 = \frac{3}{2}(\exp(\frac{1}{4}) - 1)$. This constant is less than $\frac{4}{9}$. As a consequence of Lemma 5.4 we obtain discrete versions of the results in Corollary 5.2.

COROLLARY 5.5. *Consider $f_k \in \mathbb{V}_k$ such that $\operatorname{supp}(f_k) \in \Omega_\eta$ and let u_k be a the corresponding solution of problem* (2.8). *Assume $\eta - \xi \ge 8\,h_k\,p\,|\ln h_k|$, $p > 0$; then*

$$\|(u_k)_x\|_{L_2(\Omega_\xi)} \le c\,h_k^p \|f_k\|, \tag{5.16}$$

$$\|(u_k)_y\|_{L_2(\Omega_\xi)} \le c\,\xi\,h_k^{p-1} \|f_k\|. \tag{5.17}$$

Proof. Estimate (5.16) is a consequence of (5.11). Indeed, observe the following inequalities:

$$\begin{aligned} \|(u_k)_x\|_{L_2(\Omega_\xi)} &\le c \sum_{i:\, ih \le \xi} \sum_{j=1}^{n_k-1} h_k^2 \left(\frac{u_{i,j} - u_{i-1,j}}{h_k}\right)^2 \\ &= c \sum_{i:\, ih \le \xi} \sum_{j=1}^{n_k-1} h_k^2 \phi_i \left(\frac{u_{i,j} - u_{i-1,j}}{h_k}\right)^2 \le c \sum_{i=1}^{n_k} \sum_{j=1}^{n_k-1} h_k^2 \phi_i (M_k \hat{f})_{i,j}^2 \\ &\le c \left(\max_{ih \ge \eta} \phi_i\right) \sum_{i=1}^{n_k} \sum_{j=1}^{n_k-1} h_k^2 (M_h \hat{f})_{i,j}^2 \le c \left(\max_{ih \ge \eta} \phi_i\right) \|f_k\|^2 \le c\,h_k^{2p} \|f_k\|^2. \end{aligned}$$

Estimate (5.17) follows from an inverse inequality, the Friedrichs inequality, and (5.16):

$$\|(u_k)_y\|_{L_2(\Omega_\xi)} \le c\,h_k^{-1} \|u_k\|_{L_2(\Omega_\xi)} \le c\,\xi\,h_k^{-1} \|(u_k)_x\|_{L_2(\Omega_\xi)} \le c\,\xi h_k^{p-1} \|f\|. \qquad \square$$

COROLLARY 5.6. *Consider $f_k \in \mathbb{V}_k$ such that $\operatorname{supp}(f_k) \in \Omega_\eta$. Let u and u_k be the solutions* (2.7) *and* (2.8), *respectively. Assume $\eta - \xi \ge 8\,h_k\,p\,|\ln h_k|$, $p > 0$. Then for $e_k = u - u_k$ we have*

$$\begin{aligned} &\|(e_k)_x\|_{L_2(\Omega_\xi)} \le c\,h_k^p \|f_k\|, \\ &\|(e_k)_y\|_{L_2(\Omega_\xi)} \le c \max\left\{\sqrt{\frac{\varepsilon_k}{\varepsilon}};\ \frac{\xi}{h_k}\right\} h_k^p \|f_k\|. \end{aligned}$$

Proof. The proof is made by direct superposition of estimates in Corollaries 5.2 and 5.5. $\square$

The result in Corollary 5.6 shows that the H^1-norm of errors close to the inflow boundary can be made arbitrarily small if the right-hand side is zero on a sufficiently

large subdomain $(\Omega \setminus \Omega_\eta)$ that is adjacent to this inflow boundary. In the proof of the approximation property in section 10 we will need these estimates for the case $\xi = h_k$ and $p = \frac{1}{2}$. Hence we take $\eta = 4h_k|\ln h_k| + h_k$. Note that for the results in the previous corollaries to be applicable we need right-hand side functions f_k which are zero in $\Omega \setminus \Omega_\eta$. For technical reasons we take Ω_η such that the right boundary of the domain $\Omega \setminus \Omega_\eta$ coincides with a grid line. We use $|\ln h_k| = k \ln 2$ and thus $4h_k|\ln h_k| + h_k \leq (3k+1)h_k$ and introduce the following auxiliary domains for each grid level:

$$\Omega_k^{in} := \{\, (x,y) \in \Omega \mid x < (3k+1)h_k \,\}. \tag{5.18}$$

As a direct consequence of the previous corollary we then obtain the following.

COROLLARY 5.7. *Consider $f_k \in \mathbb{V}_k$ such that f_k is zero on the subdomain Ω_k^{in}. Let u and u_k be the solutions of* (2.7) *and* (2.8)*, respectively. Then for $e_k = u - u_k$ we have*

$$\|(e_k)_x\|_{L_2(\Omega_{h_k})} \leq c\, h_k^{\frac{1}{2}} \|f_k\|, \tag{5.19}$$

$$\|(e_k)_y\|_{L_2(\Omega_{h_k})} \leq c\, \frac{h_k}{\sqrt{\varepsilon}} \|f_k\|. \tag{5.20}$$

6. Multigrid method and convergence analysis. In this section we describe the multigrid method for solving a problem of the form $A_k x = \hat{b}$ with the stiffness matrix A_k from section 2 and present a convergence analysis.

For the prolongation and restriction in the multigrid algorithm we use the canonical choice:

$$p_k : \mathrm{X}_{k-1} \to \mathrm{X}_k, \quad p_k = P_k^{-1} P_{k-1}, \quad r_k = \frac{1}{4} p_k^T. \tag{6.1}$$

Let $W_k : \mathrm{X}_k \to \mathrm{X}_k$ be a nonsingular matrix. We consider a smoother of the form

$$x^{\mathrm{new}} = \mathcal{S}_k(x^{\mathrm{old}}, \hat{b}) = x^{\mathrm{old}} - \omega_k W_k^{-1}(A_k x^{\mathrm{old}} - \hat{b}) \quad \text{for } x^{\mathrm{old}}, \hat{b} \in \mathrm{X}_k, \tag{6.2}$$

with corresponding iteration matrix denoted by

$$S_k = I - \omega_k W_k^{-1} A_k. \tag{6.3}$$

The preconditioner W_k we use is of line-Jacobi type:

$$W_k = \frac{4\varepsilon}{h_k^2} I + D_x \ . \tag{6.4}$$

Note that W_k is a blockdiagonal matrix with diagonal blocks that are $n_k \times n_k$ bidiagonal matrices. A suitable choice for the parameter ω_k follows from the analysis below.

Remark 6.1. In the literature it is often recommended to apply a so-called *robust smoother* for solving singularly perturbed elliptic problem using multigrid. Such a smoother should have the property that it becomes a direct solver if the singular perturbation parameter tends to zero (cf. [9], chapter 10). In the formulation (6.2) one then must have a splitting such that $A_k - W_k = \mathcal{O}(\varepsilon)$ (the constant in $\mathcal{O}$ may depend on k). Such robust smoothers are well known for some anisotropic problems. For anisotropic problems in which the anisotropy is *aligned with the gridlines* one

can use a line (Jacobi or Gauss–Seidel) method or an ILU factorization as a robust smoother. Theoretical analyses of these methods can be found in [29, 30, 34].

If the convection-diffusion problem (1.2) is discretized using standard finite differences it is easy to see that an appropriate line solver yields a robust smoother. However, in the finite element setting such line methods do *not yield a robust smoother*. This is clear from the stencil in (3.3). For $\varepsilon \to 0$ the diffusion part yields an x-line difference operator which can be represented exactly by an x-line smoother, but in the convection stencil the $[0 \ -\frac{1}{6} \ \frac{1}{6}]$ and $[-\frac{1}{6} \ \frac{1}{6} \ 0]$ parts of the difference operator are not captured by such a smoother. It is not clear to us how for the finite element discretization, with a stencil as in (3.3), a robust smoother can be constructed.

In multigrid analyses for reaction-diffusion or anisotropic diffusion problems one usually observes a ε^{-1} dependence in the standard approximation property that is then compensated by an ε factor from the smoothing property (cf. [21, 22, 29, 30, 34]). However, we cannot apply a similar technique, due to the fact that for our problem class a robust smoother is not available. Instead, we use another splitting of the iteration matrix of the two-grid method, leading to modified (ε-independent) smoothing and approximation properties. ∎

We consider a standard multigrid method with pre- and postsmoothers of the form as in (6.2), (6.4). In the analysis we will need different damping parameters for the pre- and postsmoother. Thus we introduce

$$S_{k,pr} := I - \omega_{k,pr} W_k^{-1} A_k, \quad S_{k,po} := I - \omega_{k,po} W_k^{-1} A_k.$$

We also define the transformed iteration matrices

$$\tilde{S}_{k,pr} := A_k S_{k,pr} A_k^{-1}, \quad \tilde{S}_{k,po} := A_k S_{k,po} A_k^{-1}.$$

We will analyze a standard two-grid method with iteration matrix

$$T_k = S_{k,po}^{\nu_k} \big(I - p_k A_{k-1}^{-1} r_k A_k\big) S_{k,pr}^{\mu_k}. \tag{6.5}$$

For the corresponding multigrid W-cycle the iteration matrix (cf. [10]) is given by

$$M_0^{\mathrm{mgm}} := 0, \quad M_k^{\mathrm{mgm}} = T_k + S_{k,po}^{\nu_k} p_k (M_{k-1}^{\mathrm{mgm}})^2 A_{k-1}^{-1} r_k A_k S_{k,pr}^{\mu_k}, \quad k > 1. \tag{6.6}$$

In the convergence analysis of this method the auxiliary inflow domain Ω_k^{in} defined in (5.18) plays a crucial role. As in the analysis of the upstream influence in section 5 we will use a cut-off function in the x-direction. We define diagonal matrices $\hat{\Phi}_k$, Φ_k as follows:

$$\xi := (3k+1)h_k, \quad \hat{\Phi}_k := \mathrm{diag}(\phi_1^\xi, \dots, \phi_{n_k}^\xi), \quad \Phi_k := I_{n_k - 1} \otimes \hat{\Phi}_k; \tag{6.7}$$

here ϕ_i^ξ is the cut-off function defined in (5.14) with $\xi = (3k+1)h_k$. For notational simplicity we drop the superscript ξ in ϕ_i^ξ in the remainder. Note that the diagonal matrix Φ_k is positive definite.

For any symmetric positive definite matrix $C \in \mathbb{R}^{m \times m}$ we define

$$\langle x, y \rangle_C := x^T C y, \quad \|x\|_C^2 := \langle x, x \rangle_C, \quad \|B\|_C := \|C^{\frac{1}{2}} B C^{-\frac{1}{2}}\|$$

with $x, y \in \mathbb{R}^m$, $B \in \mathbb{R}^{m \times m}$. Note that if $C = E^T E$ for some nonsingular matrix E then $\|B\|_C = \|EBE^{-1}\|$.

The convergence analysis is based on the following splitting, with $A := A_k$:

$$\begin{aligned}
\|T_k\|_{A^TA} &= \|S_{k,po}^{\nu_k}(I - p_k A_{k-1}^{-1} r_k A_k) S_{k,pr}^{\mu_k}\|_{A^TA} \\
&= \|S_{k,po}^{\nu_k}(A_k^{-1} - p_k A_{k-1}^{-1} r_k)\big((I - \Phi_k^{\frac{1}{2}}) + \Phi_k^{\frac{1}{2}}\big) A_k S_{k,pr}^{\mu_k}\|_{A^TA} \\
&\le \|S_{k,po}^{\nu_k}(A_k^{-1} - p_k A_{k-1}^{-1} r_k)(I - \Phi_k^{\frac{1}{2}}) A_k S_{k,pr}^{\mu_k}\|_{A^TA} \\
&\quad + \|S_{k,po}^{\nu_k}(A_k^{-1} - p_k A_{k-1}^{-1} r_k)\Phi_k^{\frac{1}{2}} A_k S_{k,pr}^{\mu_k}\|_{A^TA} \\
&\le \|\tilde{S}_{k,po}^{\nu_k} A_k W_k^{-1}\| \|W_k(A_k^{-1} - p_k A_{k-1}^{-1} r_k)(I - \Phi_k^{\frac{1}{2}})\| \|\tilde{S}_{k,pr}^{\mu_k}\| \\
&\quad + \|\tilde{S}_{k,po}^{\nu_k}\| \|I - A_k p_k A_{k-1}^{-1} r_k\| \|\Phi_k^{\frac{1}{2}} \tilde{S}_{k,pr}^{\mu_k}\|.
\end{aligned} \tag{6.8}$$

Remark 6.2. Note that the splitting in (6.8) differs from the usual splitting that is used in the theory based on the smoothing and approximation property introduced by Hackbusch (cf. [10]). In this theory the approximation property of the form $\|A_k^{-1} - p_k A_{k-1}^{-1} r_k\| \le C_A\, g(h_k, \varepsilon)$ is combined with a smoothing property of the form $\|A_k S_{k,po}^{\mu_k}\| \le \eta(\mu_k)\, g(h_k, \varepsilon)^{-1}$ with some $\eta(\mu_k)$ such that $\eta(\mu_k) \to 0,\ \mu_k \to \infty$ uniformly with respect to h_k and ε. In numerical experiments we observed that bounds of this type are not likely to be valid. Due to the fact that the smoother is not an exact solver for $\varepsilon \downarrow 0$ (cf. Remark 6.1), it is essential to have the preconditioner W_k as part of the approximation property. Furthermore, it turns out that for obtaining an appropriate bound for $\|W_k(A_k^{-1} - p_k A_{k-1}^{-1} r_k) f_k\|$ the right-hand side function f_k must vanish near the inflow boundary. We illustrate this by numerical experiments in section 11. This motivates the introduction of the "cut-off" matrix Φ_k in the decomposition.

We now formulate the main results on which the convergence analysis will be based. The proofs of these results will be given further on.

THEOREM 6.1. *The following holds:*

$$W_k A_k^{-1} \ge \frac{1}{8} I \quad \text{for} \quad k = 1, 2, \ldots . \tag{6.9}$$

Proof. The proof is given in section 7. ∎

LEMMA 6.2. *From* (6.9) *it follows that*

$$\|I - \omega A_k W_k^{-1}\| \le 1 \quad \text{for all} \ \ \omega \in \left[0, \frac{1}{4}\right].$$

Proof. The proof is elementary. ∎

Assumption 6.1. In the postsmoother $S_{k,po}$ we take $\omega_{k,po} := \frac{1}{8}$.

We note that the analysis below applies for any fixed $\omega_{k,po} \in (0, \frac{1}{8}]$. We obtain the following smoothing property.

COROLLARY 6.1. *There exists a constant c_1 independent of k and ε such that*

$$\|\tilde{S}_{k,po}^{\nu_k} A_k W_k^{-1}\| \le \frac{c_1}{\sqrt{\nu_k}}. \tag{6.10}$$

Proof. Follows from Lemma 6.2 and Theorem 10.6.8 in [10] (or results in [12, 24]). The result holds with $c_1 = \frac{32}{\sqrt{2\pi}}$. ∎

We now turn to the presmoother.

THEOREM 6.3. *There exist constants $d_1 > 0, d_2 > 0$ independent of k and ε such that*

$$\left\| \Phi_k^{\frac{1}{2}} \Big(I - \frac{d_1}{k^2} A_k W_k^{-1} \Big) \Phi_k^{-\frac{1}{2}} \right\| \leq 1 - \frac{d_2}{k^4}. \tag{6.11}$$

Proof. The proof is given in section 8. ∎

Assumption 6.2. In the presmoother $S_{k,pr}$ we take $\omega_{k,pr} := \min\{ \frac{1}{4}, \frac{d_1}{k^2} \}$.

Remark 6.3. The result in (6.11) can be written as $\|I - \frac{d_1}{k^2} A_k W_k^{-1}\|_{\Phi_k} \leq 1 - \frac{d_2}{k^4}$. Hence, we have a contraction result in the almost degenerated norm $\|\cdot\|_{\Phi_k}$. This norm, however, coincides with the Euclidean one for the vectors that have a support only in Ω_k^{in}. Hence the result in (6.11) indicates that the presmoother is a fast solver *near the inflow boundary* (cf. section 11).

Concerning the approximation property the following result holds.

THEOREM 6.4. *There exists a constant c_2 independent of k and ε such that*

$$\|W_k(A_k^{-1} - p_k A_{k-1}^{-1} r_k)(I - \Phi_k^{\frac{1}{2}})\| \leq c_2 \quad \textit{for} \quad k = 2, 3, \ldots . \tag{6.12}$$

Proof. The proof is given in section 10. ∎

Finally, we present two results related to stability of the coarse-grid correction. It is well known that for the canonical restriction operator the inequality

$$\|r_k\| \leq c_r$$

holds with a constant c_r independent of k. The second stability result is the following.

THEOREM 6.5. *There exists a constant c_3 independent of k and ε such that*

$$\|A_k p_k A_{k-1}^{-1}\| \leq c_3 \quad \textit{for} \quad k = 2, 3, \ldots . \tag{6.13}$$

Proof. The proof is given in section 9. ∎

We now obtain a two-grid convergence result.

THEOREM 6.6. *For the two-grid method we then have*

$$\|T_k\|_{A^T A} \leq \frac{c_1 c_2}{\sqrt{\nu_k}} + (1 + c_r c_3) \left(1 - \frac{d_2}{k^4} \right)^{\mu_k}.$$

Proof. The proof is based on results from (6.9), (6.11), (6.12), and (6.13). We use the splitting in (6.8). From the Assumptions 6.1 and 6.2 and Lemma 6.2 it follows that $\|\tilde{S}_{k,pr}\| \leq 1$ and $\|\tilde{S}_{k,po}\| \leq 1$. From Assumption 6.2, Theorem 6.3, and $\|\Phi_k\| \leq 1$ we obtain

$$\|\Phi_k^{\frac{1}{2}} \tilde{S}_{k,pr}^{\mu_k}\| \leq \|(\Phi_k^{\frac{1}{2}} \tilde{S}_{k,pr} \Phi_k^{-\frac{1}{2}})^{\mu_k}\| \|\Phi_k^{\frac{1}{2}}\| \leq \left(1 - \frac{d_2}{k^4} \right)^{\mu_k}$$

Combine these bounds with the results in Corollary 6.1 and Theorems 6.4 and 6.5. ∎

Using the two-grid result of Theorem 6.6 we derive a multigrid W-cycle convergence result based on standard arguments.

THEOREM 6.7. *In addition to the assumptions of Theorem* 6.6 *we assume that the number of smoothing steps on every grid level is sufficiently large:*

$$\nu_k \geq c_{po}, \quad \mu_k \geq c_{pr}\, k^4$$

with suitable constants c_{po}, c_{pr}. Then for the contraction number of the multigrid W-cycle the inequality

$$\|M_k^{\mathrm{mgm}}\|_{A^TA} \leq \xi^* \tag{6.14}$$

holds, with a constant $\xi^* < 1$ independent of k and ε.

Proof. Define $\xi_k := \|M_k^{\mathrm{mgm}}\|_{A_k^TA_k}$. Using the recursion relation (6.6) for M_k^{mgm} it follows that

$$\begin{aligned}\xi_k &\leq \|T_k\|_{A_k^TA_k} + \|\tilde{S}_{k,po}\|^{\nu_k}\|A_kp_kA_{k-1}^{-1}\|\xi_{k-1}^2\|r_k\|\|\tilde{S}_{k,pr}\|^{\mu_k} \\ &\leq \|T_k\|_{A_k^TA_k} + c_3c_r\xi_{k-1}^2.\end{aligned}$$

Now use the two-grid bound given in Theorem 6.6 and a fixed point argument. ∎

Remark 6.4. We briefly discuss the arithmetic work needed in one W-cycle iteration. The arithmetic work for a matrix vector multiplication on level k is of order $\mathcal{O}(N_k) = \mathcal{O}(n_k^2)$. The work needed in one smoothing iteration is of order $\mathcal{O}(N_k)$. The number of smoothings behaves like $\nu_k + \mu_k \sim k^4$. Using a standard recursive argument it follows that for a multigrid W-cycle iteration the arithmetic complexity is of the order $N_k(\ln N_k)^4$. Hence this multigrid method has suboptimal complexity.

7. Proof of Theorem 6.1. We recall the representation of the stiffness matrix in (3.4)

$$A_k = \left(\varepsilon + \left(\bar{\delta} - \frac{1}{3}\right)h_k\right)A_x + \varepsilon A_y + \frac{1}{6}BD_x.$$

We will need the following lemma:

LEMMA 7.1. *The inequality $BD_x \geq 0$ holds.*

Proof. The matrix $\frac{1}{6}BD_x - \frac{1}{3}h_kA_x$ is the stiffness matrix corresponding to the bilinear form $(u,v) \to \int_\Omega u_xv\,dxdy$. For any $x \in X_k$ we get

$$\frac{1}{6}\langle BD_xx,x\rangle_k - \frac{1}{3}\langle h_kA_xx,x\rangle_k = \int_\Omega (P_kx)_x(P_kx)\,dxdy = \frac{1}{2}\int_{\Gamma_E}(P_kx)^2\,dxdy \geq 0.$$

Since the matrix A_x is symmetric positive definite the result now follows. ∎

We now consider the preconditioner $W_k = \frac{4\varepsilon}{h_k^2}I + D_x$, as in (6.4).

THEOREM 7.2 (=Theorem 6.1). *The inequality $W_kA_k^{-1} \geq \frac{1}{8}I$ holds.*

Proof. First note that

$$h_k\hat{D}_x\hat{D}_x^T = \hat{D}_x + \hat{D}_x^T - \frac{1}{h_k}(1,0,\ldots,0)^T(1,0,\ldots,0) \leq \hat{D}_x + \hat{D}_x^T$$

and thus $h_k\hat{D}_x^T\hat{D}_x\hat{D}_x^T\hat{D}_x \leq \hat{D}_x^T(\hat{D}_x + \hat{D}_x^T)\hat{D}_x$ holds. Using $\hat{A}_x = \hat{D}_x^T\hat{D}_x$ this results in $h_k\hat{A}_x^2 \leq 2\hat{D}_x^T\hat{A}_x$ and thus

$$\frac{1}{2}h_kA_x^2 \leq D_x^TA_x. \tag{7.1}$$

Note that the following inequality holds for any $a,b,c \in \mathbb{R}$ and $\sigma_1,\sigma_2,\sigma_3 > 0$:

$$(a+b+c)^2 \leq (1+\sigma_2+\sigma_3^{-1})a^2 + (1+\sigma_3+\sigma_1^{-1})b^2 + (1+\sigma_1+\sigma_2^{-1})c^2.$$

We apply this inequality with $\sigma_2 = 2, \sigma_1 = \sigma_3 = 1$. Also using $\|A_y\| \leq 4h_k^{-2}$ and $\|B\| \leq 6$ we get for any $x \in \mathrm{X}_k$

$$\begin{aligned}\|A_k x\|^2 &\leq 4\varepsilon^2 \|A_y x\|^2 + 3\bar{\varepsilon}_k^2 \|A_x x\|^2 + \frac{5}{2}\left\|\frac{1}{6} B D_x x\right\|^2 \\ &\leq 16\left(\frac{\varepsilon}{h_k}\right)^2 \langle A_y x, x\rangle_k + 3\bar{\varepsilon}_k^2 \|A_x x\|^2 + \frac{5}{2}\|D_x x\|^2.\end{aligned} \tag{7.2}$$

We recall that $\bar{\varepsilon}_k = \varepsilon_k - \bar{\delta} h_k \leq \frac{7}{6} h_k$. Now apply the result (7.1) and the estimates in Lemmas 3.1 and Lemma 7.1 to obtain

$$\begin{aligned}\langle W_k x, A_k x\rangle_k &= \left\langle \frac{4\varepsilon}{h_k^2} x + D_x x\,,\, \varepsilon A_y x + \bar{\varepsilon}_k A_x x + \frac{1}{6} B D_x x \right\rangle_k \\ &\geq 4\left(\frac{\varepsilon}{h_k}\right)^2 \langle A_y x, x\rangle_k + \bar{\varepsilon}_k \langle D_x x, A_x x\rangle_k + \left\langle D_x x, \frac{1}{6} B D_x x\right\rangle_k \\ &\geq 4\left(\frac{\varepsilon}{h_k}\right)^2 \langle A_y x, x\rangle_k + \frac{3}{7}\bar{\varepsilon}_k^2 \|A_x x\|^2 + \frac{1}{3}\|D_x x\|^2 \\ &= \frac{1}{8}\left(32\left(\frac{\varepsilon}{h_k}\right)^2 \langle A_y x, x\rangle_k + \frac{24}{7}\bar{\varepsilon}_k^2 \|A_x x\|^2 + \frac{8}{3}\|D_x x\|^2\right) \\ &\geq \frac{1}{8}\left(16\left(\frac{\varepsilon}{h_k}\right)^2 \langle A_y x, x\rangle_k + 3\bar{\varepsilon}_k^2 \|A_x x\|^2 + \frac{5}{2}\|D_x x\|^2\right).\end{aligned}$$

Combination of this with the inequality in (7.2) proves the theorem. ∎

8. Proof of Theorem 6.3. We start with an elementary known result on the convergence of basic iterative methods.

LEMMA 8.1. *Assume $C, A, W \in \mathbb{R}^{n\times n}$ with C symmetric positive definite. If there are constants $c_0 > 0$, c_1 such that*

$$c_0 \langle Ay, Ay\rangle_C \leq \langle Wy, Wy\rangle_C \leq c_1 \langle Wy, Ay\rangle_C \quad \text{for all } y \in \mathbb{R}^n \tag{8.1}$$

then for arbitrary $d \in [0,1]$ we have

$$\|I - \alpha \frac{c_0}{c_1} A W^{-1}\|_C \leq \sqrt{1 - d\frac{c_0}{c_1^2}} \quad \text{if} \quad 1 - \sqrt{1-d} \leq \alpha \leq 1 + \sqrt{1-d}.$$

Proof. Let $D := AW^{-1}$. From (8.1) we get

$$\langle Dy, y\rangle_C \geq c_1^{-1} \langle y, y\rangle_C\,, \quad \langle Dy, Dy\rangle_C \leq c_0^{-1} \langle y, y\rangle_C \quad \text{for all } y.$$

Note that

$$\begin{aligned}\left\|\left(I - \alpha \frac{c_0}{c_1} A W^{-1}\right) y\right\|_C^2 &= \langle y, y\rangle_C - 2\alpha \frac{c_0}{c_1}\langle Dy, y\rangle_C + \alpha^2 \frac{c_0^2}{c_1^2}\langle Dy, Dy\rangle_C \\ &\leq \left(1 - 2\alpha \frac{c_0}{c_1^2} + \alpha^2 \frac{c_0}{c_1^2}\right)\|y\|_C^2 = \left(1 - (2\alpha - \alpha^2)\frac{c_0}{c_1^2}\right)\|y\|_C^2\end{aligned}$$

and $2\alpha - \alpha^2 \geq d$ if $1 - \sqrt{1-d} \leq \alpha \leq 1 + \sqrt{1-d}$. ∎

Below we use the scalar product $\langle \cdot,\cdot\rangle_\Phi := \langle \Phi_k \cdot,\cdot\rangle_k$ with Φ_k defined in (6.7). We recall the result proved in Lemma 5.3,

$$\langle A_k x, D_x x\rangle_\Phi \geq c\,\|D_x x\|_\Phi^2 \quad \text{for all} \;\; x \in X_k \tag{8.2}$$

with $c > 0$ independent of k and of ε.

We introduce the diagonal projection matrix $J_k := I_{n_k-1} \otimes \hat{J}_k$ with $\hat{J}_k$ the $n_k \times n_k$ diagonal matrix with $(\hat{J}_k)_{i,i} = 1$ if $(\hat{\Phi}_k)_{i,i} = 1$ and $(\hat{J}_k)_{i,i} = 0$ otherwise.

LEMMA 8.2. *There exists a constant $c > 0$ independent of k and ε such that*

$$\|W_k x\|_\Phi^2 \leq ck^2 \left(\frac{\varepsilon}{h_k^3} \|(I - J_k)x\|_\Phi^2 + \|D_x x\|_\Phi^2 \right) \quad \textit{for all} \;\; x \in X_k.$$

Proof. Note that

$$\begin{aligned} \|J_k x\|_\Phi &= \|J_k D_x^{-1} J_k D_x x\|_\Phi \leq \|J_k D_x^{-1} J_k\|_\Phi \|D_x x\|_\Phi \\ &= \|J_k D_x^{-1} J_k\| \|D_x x\|_\Phi \leq (3k+1) h_k \|D_x x\|_\Phi. \end{aligned}$$

And thus, using $\varepsilon \leq \frac{1}{2} h_k$ we get

$$\begin{aligned} \|W_k x\|_\Phi &= \left\| \frac{4\varepsilon}{h_k^2} x + D_x x \right\|_\Phi \leq \frac{4\varepsilon}{h_k^2} \|(I - J_k)x\|_\Phi + \frac{4\varepsilon}{h_k^2} \|J_k x\|_\Phi + \|D_x x\|_\Phi \\ &\leq \frac{4\varepsilon}{h_k^2} \|(I - J_k)x\|_\Phi + ck \|D_x x\|_\Phi \leq ck \left(\frac{4\varepsilon}{h_k^2} \|(I - J_k)x\|_\Phi + \|D_x x\|_\Phi \right). \end{aligned}$$

Squaring this result and using $(\frac{\varepsilon}{h_k^2})^2 \leq \frac{1}{2} \frac{\varepsilon}{h_k^3}$ completes the proof. ∎

We define $\hat{\Phi}_x := \frac{1}{h_k} \mathrm{diag}(\phi_i - \phi_{i+1})_{1 \leq i \leq n_k}$ with $\phi_i = \phi_i^\xi$ as in (6.7). Consider the diagonal matrix $\Phi_x := I_{n_k-1} \otimes \hat{\Phi}_x$. Note that $\Phi_x \geq 0$.

LEMMA 8.3. *The following estimate holds:*

$$\langle A_k x, x\rangle_\Phi \geq \frac{1}{30} \|\Phi_x^{\frac{1}{2}} x\|^2 \quad \textit{for all} \;\; x \in X_k.$$

Proof. Recall

$$A_k = \bar{\varepsilon}_k A_x + \varepsilon A_y + \frac{1}{6} B D_x. \tag{8.3}$$

Note that

$$\Phi_k A_y = (I_{n_k-1} \otimes \hat{\Phi}_k)(\hat{A}_y \otimes \hat{J}) = \hat{A}_y \otimes \hat{\Phi}_k \hat{J} \geq 0. \tag{8.4}$$

We consider the term $\bar{\varepsilon}_k \Phi_k A_x = \bar{\varepsilon}_k (I_{n_k-1} \otimes \hat{\Phi}_k \hat{A}_x)$. Note that $\hat{\Phi}_k \hat{A}_x = \hat{\Phi}_k \hat{D}_x^T \hat{D}_x$. A simple computation yields $\hat{\Phi}_k \hat{D}_x^T - \hat{D}_x^T \hat{\Phi}_k = -\hat{\Phi}_x \hat{T}$, with $\hat{T} := \mathrm{tridiag}(0,0,1)$, and thus

$$\bar{\varepsilon}_k \hat{\Phi}_k \hat{A}_x = \bar{\varepsilon}_k \hat{D}_x^T \hat{\Phi}_k \hat{D}_x - \bar{\varepsilon}_k \hat{\Phi}_x \hat{T} \hat{D}_x. \tag{8.5}$$

From the Cauchy–Schwarz inequality it follows that

$$\bar{\varepsilon}_k \langle \hat{\Phi}_x \hat{T} \hat{D}_x y, y\rangle \leq \bar{\varepsilon}_k^2 \frac{9}{4} \|\hat{\Phi}_x^{\frac{1}{2}} \hat{T} \hat{D}_x y\|^2 + \frac{1}{9} \|\hat{\Phi}_x^{\frac{1}{2}} y\|^2 \quad \text{for all} \;\; y \in \mathbb{R}^{n_k}. \tag{8.6}$$

Using the property (5.15) we get

$$\hat{T}^T\hat{\Phi}_x\hat{T} \le \bar{\varepsilon}_k^{-1}c_0\hat{\Phi}_k. \tag{8.7}$$

Combination of the results in (8.5), (8.6), (8.7) and using $c_0 \le \frac{4}{9}$ yields

$$\begin{aligned}\bar{\varepsilon}_k\langle\hat{\Phi}_k\hat{A}_x y, y\rangle &\ge \bar{\varepsilon}_k\|\hat{D}_x y\|^2_{\hat{\Phi}_k} - \bar{\varepsilon}_k\frac{9}{4}c_0\|\hat{D}_x y\|^2_{\hat{\Phi}_k} - \frac{1}{9}\|\hat{\Phi}_x^{\frac{1}{2}}y\|^2\\ &\ge -\frac{1}{9}\|\hat{\Phi}_x^{\frac{1}{2}}y\|^2 \quad \text{for all} \quad y\in\mathbb{R}^{n_k}.\end{aligned}$$

And thus

$$\bar{\varepsilon}_k\Phi_k A_x \ge -\frac{1}{9}\Phi_x \tag{8.8}$$

holds. Finally we consider the term $\frac{1}{6}\langle BD_x x, x\rangle_\Phi$. First we note

$$BD_x = \text{blocktridiag}(\hat{D}_x, 4\hat{D}_x, \hat{S}_x), \quad \hat{S}_x := \frac{1}{h_k}\begin{pmatrix}-1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1\\ & & & 0\end{pmatrix} \in \mathbb{R}^{n_k\times n_k}$$

and thus $K := \frac{1}{6}\Phi_k BD_x = \frac{1}{6}\text{blocktridiag}(\hat{\Phi}_k\hat{D}_x, 4\hat{\Phi}_k\hat{D}_x, \hat{\Phi}_k\hat{S}_x)$. Hence

$$\frac{1}{2}(K+K^T) = \frac{1}{12}\text{blocktridiag}\big(\hat{\Phi}_k\hat{D}_x + \hat{S}_x^T\hat{\Phi}_k, 4(\hat{\Phi}_k\hat{D}_x + \hat{D}_x^T\hat{\Phi}_k), \hat{\Phi}_k\hat{S}_x + \hat{D}_x^T\hat{\Phi}_k\big).$$

A simple computation yields

$$\hat{\Phi}_k\hat{D}_x + \hat{D}_x^T\hat{\Phi}_k = \hat{\Phi}_x + \frac{1}{h_k}\text{tridiag}(-\phi_i, \phi_i+\phi_{i+1}, -\phi_{i+1})_{1\le i\le n_k} =: \hat{\Phi}_x + R \tag{8.9}$$

and $\hat{\Phi}_k\hat{S}_x + \hat{D}_x^T\hat{\Phi}_k = \hat{\Phi}_x\hat{T} + \frac{1}{h_k}\phi_n e_n e_n^T$, with $n := n_k$ and e_n the nth basis vector in $\mathbb{R}^n$. Thus we obtain

$$\begin{aligned}\frac{1}{2}(K+K^T) &= \frac{1}{12}\text{blocktridiag}\big(\hat{T}^T\hat{\Phi}_x, 4\hat{\Phi}_x, \hat{\Phi}_x\hat{T}\big)\\ &\quad + \frac{1}{12}\text{blocktridiag}\left(\frac{1}{h_k}\phi_n e_n e_n^T, 4R, \frac{1}{h_k}\phi_n e_n e_n^T\right)\\ &\ge \frac{1}{12}\text{blocktridiag}\big(\hat{T}^T\hat{\Phi}_x, 4\hat{\Phi}_x, \hat{\Phi}_x\hat{T}\big).\end{aligned}$$

By $\hat{\Phi}_x^{-1}$ (Φ_x^{-1}) we denote the pseudoinverse of $\hat{\Phi}_x$ (Φ_x). We then have

$$\frac{1}{2}\Phi_x^{-\frac{1}{2}}(K+K^T)\Phi_x^{-\frac{1}{2}} \ge \frac{1}{12}\text{blocktridiag}\big(\hat{\Phi}_x^{-\frac{1}{2}}\hat{T}^T\hat{\Phi}_x^{\frac{1}{2}}, 4I, \hat{\Phi}_x^{\frac{1}{2}}\hat{T}\hat{\Phi}_x^{-\frac{1}{2}}\big).$$

Note that

$$\|\hat{\Phi}_x^{-\frac{1}{2}}\hat{T}^T\hat{\Phi}_x^{\frac{1}{2}}\|_\infty = \|\hat{\Phi}_x^{\frac{1}{2}}\hat{T}\hat{\Phi}_x^{-\frac{1}{2}}\|_\infty = \max_{i\ge 3k+2}\left(\frac{\phi_{i-1}-\phi_i}{\phi_i-\phi_{i+1}}\right)^{\frac{1}{2}} = e^{\frac{1}{8}}.$$

And thus we get $\frac{1}{2}\Phi_x^{-\frac{1}{2}}(K+K^T)\Phi_x^{-\frac{1}{2}} \ge \frac{1}{12}(4-2e^{\frac{1}{8}})I$. Hence

$$\frac{1}{6}\Phi_k BD_x = K \ge \frac{1}{6}(2-e^{\frac{1}{8}})\Phi_x. \tag{8.10}$$

Combination of the results in (8.3), (8.4), (8.8), and (8.10) yields

$$\Phi_k A_k \geq \left(-\frac{1}{9} + \frac{1}{6}(2 - e^{\frac{1}{8}}) \right) \Phi_x > \frac{1}{30}\Phi_x. \qquad \square$$

Using the previous two lemmas we can show a result as in the second inequality in (8.1).

THEOREM 8.4. *There exists a constant c_1 independent of k and ε such that*

$$\langle W_k x, W_k x\rangle_\Phi \leq c_1 k^2 \langle W_k x, A_k x\rangle_\Phi \quad \textit{for all } \; x \in X_k.$$

Proof. From Lemma 8.3 and (8.2) we get

$$\begin{aligned} \langle W_k x, A_k x\rangle_\Phi &= \frac{4\varepsilon}{h_k^2}\langle x, A_k x\rangle_\Phi + \langle D_x x, A_k x\rangle_\Phi \\ &\geq c\left(\frac{\varepsilon}{h_k^2}\langle \Phi_x x, x\rangle_k + \|D_x x\|_\Phi^2 \right) \end{aligned} \tag{8.11}$$

with $c > 0$ independent of k and ε. Using $\phi_i - \phi_{i+1} = (1 - e^{-\frac{1}{4}})\phi_i \geq \frac{1}{5}\phi_i$ for $i \geq 3k+1$ we get

$$\langle \Phi_x x, x\rangle_k \geq \frac{1}{5} h_k^{-1} \langle (I - J_k)\Phi_k x, x\rangle_k = \frac{1}{5} h_k^{-1} \|(I - J_k)x\|_\Phi^2. \tag{8.12}$$

From (8.11) and (8.12) we obtain

$$\langle W_k x, A_k x\rangle_\Phi \geq c\left(\frac{\varepsilon}{h_k^3}\|(I - J_k)x\|_\Phi^2 + \|D_x x\|_\Phi^2 \right)$$

Now combine this with the result in Lemma 8.2. $\square$

We now consider the first inequality in (8.1).

THEOREM 8.5. *There exists a constant $c_0 > 0$ independent of k and ε such that*

$$c_0 \langle A_k x, A_k x\rangle_\Phi \leq \langle W_k x, W_k x\rangle_\Phi \quad \textit{for all } \; x \in X_k.$$

Proof. The constants c that appear in the proof are all strictly positive and independent of k and ε. First note that $\|A_k x\|_\Phi \leq \bar{\varepsilon}_k \|A_x x\|_\Phi + \varepsilon \|A_y x\|_\Phi + \frac{1}{6}\|B D_x x\|_\Phi$. We have

$$\|A_y\|_\Phi = \|(I_{n_k - 1} \otimes \hat{\Phi}_k^{\frac{1}{2}})(\hat{A}_y \otimes \hat{J})(I_{n_k - 1} \otimes \hat{\Phi}_k^{-\frac{1}{2}})\| = \|\hat{A}_y \otimes \hat{J}\| \leq \frac{4}{h_k^2}.$$

Note that $|\phi_i \phi_{i+1}^{-1}| \leq e^{\frac{1}{4}}$ and thus $\|\hat{\Phi}_k^{\frac{1}{2}} \hat{D}_x^T \hat{\Phi}_k^{-\frac{1}{2}}\| \leq c h_k^{-1}$ holds. From this it follows that $\|D_x^T\|_\Phi \leq c h_k^{-1}$ holds. With a similar argument we get $\|B\|_\Phi \leq c$. Thus we obtain, using $\bar{\varepsilon}_k \leq \frac{3}{2} h_k$,

$$\begin{aligned} \|A_k x\|_\Phi &\leq \bar{\varepsilon}_k \|D_x^T\|_\Phi \|D_x x\|_\Phi + \frac{4\varepsilon}{h_k^2}\|x\|_\Phi + c\|D_x x\|_\Phi \\ &\leq c\left(\frac{\varepsilon}{h_k^2}\|x\|_\Phi + \|D_x x\|_\Phi \right). \end{aligned} \tag{8.13}$$

From (8.9) it follows that $\langle D_x x, x\rangle_\Phi \geq 0$ holds. Using this we get

$$\begin{aligned}\|W_k x\|_\Phi^2 &= \frac{16\varepsilon^2}{h_k^4}\|x\|_\Phi^2 + \frac{16\varepsilon}{h_k^2}\langle D_x x, x\rangle_\Phi + \|D_x x\|_\Phi^2 \\ &\geq c\left(\frac{\varepsilon^2}{h_k^4}\|x\|_\Phi^2 + \|D_x x\|_\Phi^2\right).\end{aligned} \tag{8.14}$$

Now combine (8.13) with (8.14). ∎

Combination of the results of Theorems 8.4 and 8.5 with the second result in Lemma 8.1 shows that Theorem 6.3 holds.

9. Proof of Theorem 6.5. Let $\mathrm{g}_{k-1} \in X_{k-1}$ be given and define $g_{k-1} := (P_{k-1}^*)^{-1}\mathrm{g}_{k-1} \in \mathbb{V}_{k-1}$. Let $u_{k-1} \in \mathbb{V}_{k-1}$ be such that

$$a_{k-1}(u_{k-1}, v_{k-1}) = (g_{k-1}, v_{k-1}) \quad \text{for all} \quad v_{k-1} \in \mathbb{V}_{k-1}.$$

Then $A_{k-1}^{-1}\mathrm{g}_{k-1} = P_{k-1}^{-1}u_{k-1}$ holds. The corresponding continuous solution $u \in \mathbf{V}$ satisfies $a_{k-1}(u, v) = (g_{k-1}, v)$ for all $v \in \mathbf{V}$. Now note that

$$\begin{aligned}\|A_k p_k A_{k-1}^{-1}\mathrm{g}_{k-1}\| &= \max_{y\in X_k}\frac{\langle A_k p_k P_{k-1}^{-1}u_{k-1}, y\rangle_k}{\|y\|} \leq c\max_{v_k\in\mathbb{V}_k}\frac{a_k(u_{k-1}, v_k)}{\|v_k\|} \\ &\leq c\max_{v_k\in\mathbb{V}_k}\frac{a_{k-1}(u_{k-1}, v_k)}{\|v_k\|} + c\max_{v_k\in\mathbb{V}_k}\frac{a_k(u_{k-1}, v_k) - a_{k-1}(u_{k-1}, v_k)}{\|v_k\|}.\end{aligned} \tag{9.1}$$

Define $e_{k-1} := u - u_{k-1}$. For the first term in (9.1) we get, using the results of Lemma 4.3,

$$\begin{aligned}a_{k-1}(u_{k-1}, v_k) &\leq |a_{k-1}(e_{k-1}, v_k)| + |a_{k-1}(u, v_k)| \\ &\leq ch_k\|(e_{k-1})_x\|\|(v_k)_x\| + \varepsilon\|(e_{k-1})_y\|\|(v_k)_y\| + \|(e_{k-1})_x\|\|v_k\| + |(g_{k-1}, v_k)| \\ &\leq c\left(\|(e_{k-1})_x\| + \frac{\varepsilon}{h_k}\|(e_{k-1})_y\|\right)\|v_k\| + \|g_{k-1}\|\|v_k\| \\ &\leq c\|g_{k-1}\|\|v_k\| \leq c\|\mathrm{g}_{k-1}\|\|v_k\|.\end{aligned} \tag{9.2}$$

For the second term in (9.1) we have, using Lemma 4.2,

$$\begin{aligned}|a_k(u_{k-1}, v_k) - a_{k-1}(u_{k-1}, v_k)| &= \bar{\delta}h_k|((u_{k-1})_x, (v_k)_x)| \\ &\leq c\|(u_{k-1})_x\|\|v_k\| \\ &\leq c\|g_{k-1}\|\|v_k\| \leq c\|\mathrm{g}_{k-1}\|\|v_k\|.\end{aligned} \tag{9.3}$$

Combination of the results in (9.1), (9.2), and (9.3) yields $\|A_k p_k A_{k-1}^{-1}\mathrm{g}_{k-1}\| \leq c\|\mathrm{g}_{k-1}\|$ and thus the result in Theorem 6.5 holds. ∎

10. Proof of Theorem 6.4. We briefly comment on the idea of the proof. As usual to prove an estimate for the error in the L^2-norm we use a duality argument. However, the formal dual problem has poor regularity properties, since in this dual problem Γ_E is the "inflow" boundary and Γ_W is the "outflow" boundary. Thus Dirichlet outflow boundary conditions would appear and we obtain poor estimates due to the poor regularity. To avoid this, we consider a dual problem with Neumann outflow and Dirichlet inflow conditions. To be able to deal with the inconsistency caused by these "wrong" boundary conditions we *assume* the right-hand side is zero

near the boundary Γ_W. In order to satisfy this assumption we use the cut-off operator with matrix Φ_k.

A further problem we have to deal with is the fact that due to the level dependent stabilization term we have to treat k-dependent bilinear forms.

We introduce the space

$$\mathbb{V}_k^0 := \{ v_k \in \mathbb{V}_k \mid v_k(x) = 0 \quad \text{for all} \quad x \in \Omega_k^{in} \}.$$

Let $\hat{b}_k \in X_k$ be given. In view of Theorem 6.4 we must prove an estimate $\|W_k(A_k^{-1} - p_k A_{k-1}^{-1} r_k)(I - \Phi_k)\hat{b}_k\| \leq c\|\hat{b}_k\|$ with a constant c that is independent of k, ε, and $\hat{b}_k$. Note that $(P_k^*)^{-1}(I - \Phi_k^{\frac{1}{2}})\hat{b}_k =: f_k \in \mathbb{V}_k^0$ holds. For this $f_k \in \mathbb{V}_k^0$ we define corresponding discrete solutions and continuous solutions as follows:

$$\begin{aligned}
u_k \in \mathbb{V}_k : &\quad a_k(u_k, v_k) = (f_k, v_k) && \text{for all} \quad v_k \in \mathbb{V}_k, \\
u \in \mathbf{V} : &\quad a_k(u, v) = (f_k, v) && \text{for all} \quad v \in \mathbf{V}, \\
u_{k-1} \in \mathbb{V}_{k-1} : &\quad a_{k-1}(u_{k-1}, v_{k-1}) = (f_k, v_{k-1}) && \text{for all} \quad v_{k-1} \in \mathbb{V}_{k-1}, \\
\tilde{u} \in \mathbf{V} : &\quad a_{k-1}(\tilde{u}, v) = (f_k, v) && \text{for all} \quad v \in \mathbf{V}.
\end{aligned} \tag{10.1}$$

In the proof of Lemma 4.2 we showed that $\|v_x\| = \|D_x P_k^{-1} v\|$ holds for all $v \in \mathbb{V}_k$. We use that $W_k = \frac{4\varepsilon}{h_k^2} I + D_x$ and obtain

$$\begin{aligned}
\|W_k(A_k^{-1} - p_k A_{k-1}^{-1} r_k)(I - \Phi_k^{\frac{1}{2}})\hat{b}_k\| &\leq \frac{4\varepsilon}{h_k^2} \|(A_k^{-1} - p_k A_{k-1}^{-1} r_k)(I - \Phi_k^{\frac{1}{2}})\hat{b}_k\| \\
&\quad + \|D_x A_k^{-1}(I - \Phi_k^{\frac{1}{2}})\hat{b}_k\| + \|D_x p_k A_{k-1}^{-1} r_k (I - \Phi_k^{\frac{1}{2}})\hat{b}_k\| \\
&\leq c \left(\frac{\varepsilon}{h_k^2} \|u_k - u_{k-1}\| + \|(u_k)_x\| + \|(u_{k-1})_x\| \right) \\
&\leq c \left(\frac{\varepsilon}{h_k^2} \big(\|u - u_k\| + \|\tilde{u} - u_{k-1}\| + \|u - \tilde{u}\| \big) + \|(u_k)_x\| + \|(u_{k-1})_x\| \right).
\end{aligned} \tag{10.2}$$

From Lemma 4.2 we get

$$\|(u_k)_x\| + \|(u_{k-1})_x\| \leq c\|f_k\|. \tag{10.3}$$

From the result in Theorem 10.1 below it follows that

$$\|u_k - u\| + \|u_{k-1} - \tilde{u}\| \leq c \, \frac{h_k^2}{\varepsilon} \, \|f_k\|. \tag{10.4}$$

Finally, from Theorem 10.4 we have

$$\|u - \tilde{u}\| \leq c \, h_k \|f_k\|. \tag{10.5}$$

If we insert the results (10.3),(10.4), and (10.5) in (10.2) we get

$$\|W_k(A_k^{-1} - p_k A_{k-1}^{-1} r_k)(I - \Phi_k^{\frac{1}{2}})\hat{b}_k\| \leq c\|f_k\| \leq c\|(P_k^*)^{-1}\| \|I - \Phi_k^{\frac{1}{2}}\| \|\hat{b}_k\| \leq c\|\hat{b}_k\|$$

and thus the result of Theorem 6.4 is proved. It remains to prove the results in Theorems 10.1 and 10.4.

THEOREM 10.1. *For $f_k \in \mathbb{V}_k^0$ let u and u_k be as defined in* (10.1). *Then*

$$\|u - u_k\| \leq c \, \frac{h_k^2}{\varepsilon} \, \|f_k\| \tag{10.6}$$

holds.

Proof. Define $e_k := u - u_k$. Let $w \in H^2(\Omega)$ be such that

$$-\varepsilon w_{yy} - \varepsilon_k w_{xx} - w_x = e_k \tag{10.7}$$

with

$$w_x = 0 \text{ on } \Gamma_W \ , \quad w = 0 \text{ on } \Gamma \setminus \Gamma_W. \tag{10.8}$$

Note that for this problem Γ_E is the "inflow" boundary and Γ_W is the "outflow" boundary. We multiply (10.7) with e_k and integrate by parts to get

$$\begin{aligned}\|e_k\|^2 &= \varepsilon((e_k)_y, w_y) + \varepsilon_k((e_k)_x, w_x) - \varepsilon_k \int_{\Gamma_E} w_x e_k \, dy + ((e_k)_x, w) \\ &= a_k(e_k, w) - \varepsilon_k \int_{\Gamma_E} w_x e_k \, dy.\end{aligned}$$

We use (4.6) with w and e_k instead of u and f, respectively, and (4.12) to estimate

$$\left|\varepsilon_k \int_{\Gamma_E} w_x e_k \, dy\right| \le \varepsilon_k^{\frac{1}{2}} \left(\varepsilon_k \int_{\Gamma_E} w_x^2 \, dy\right)^{\frac{1}{2}} \left(\int_{\Gamma_E} e_k^2 \, dy\right)^{\frac{1}{2}} \le c\, h_k^{\frac{1}{2}} \|e_k\| \frac{h_k}{\sqrt{\varepsilon}} \|f_k\|. \tag{10.9}$$

From this estimate and the Galerkin orthogonality for the error it follows that for any $v_k \in \mathbb{V}_k$

$$\begin{aligned}\|e_k\|^2 \le{}& \varepsilon\, ((e_k)_y, (w - v_k)_y) + \varepsilon_k\, ((e_k)_x, (w - v_k)_x) \\ &+ ((e_k)_x, w - v_k) + c\, \|e_k\| \frac{h_k^{\frac{3}{2}}}{\sqrt{\varepsilon}} \|f_k\|.\end{aligned} \tag{10.10}$$

Let $\Omega_h := \Omega_{h_k}$ be as defined in (5), i.e., Ω_h is the set of triangles with at least one vertex on Γ_W. In the remainder of the domain, $\omega = \Omega \backslash \Omega_h$, we take v_k as a nodal interpolant to w and we put $v_k = 0$ on Γ_W to ensure $v_k \in \mathbb{V}_k$. Note that v_k is a proper interpolant of w everywhere in Ω except in Ω_h. Therefore we will estimate scalar products in (10.10) over ω and Ω_h, separately. We continue (10.10) with

$$\begin{aligned}\|e_k\|^2 &\le c\, \varepsilon\, h_k \|(e_k)_y\|_\omega \|w\|_{H^2(\omega)} + c\, \varepsilon_k\, h_k \|(e_k)_x\|_\omega \|w\|_{H^2(\omega)} \\ &\quad + c\, h_k^2 \|(e_k)_x\|_\omega \|w\|_{H^2(\omega)} + c\, \|e_k\| \frac{h_k^{\frac{3}{2}}}{\sqrt{\varepsilon}} \|f_k\| + \mathrm{I}_{\Omega_h} \\ &\le c\, h_k^2 \|f_k\| \frac{1}{\varepsilon} \|e_k\| + \mathrm{I}_{\Omega_h}.\end{aligned} \tag{10.11}$$

The term I_{Ω_h} collects integrals over Ω_h:

$$\mathrm{I}_{\Omega_h} = \varepsilon\, ((e_k)_y, (w - v_k)_y)_{\Omega_h} + \varepsilon_k\, ((e_k)_x, (w - v_k)_x)_{\Omega_h} + ((e_k)_x, w - v_k)_{\Omega_h}\,.$$

To estimate I_{Ω_h} we use Corollary 5.7 and the following auxiliary estimate for the interpolant $v_k \in \mathbb{V}_k$ of w, with $\omega_h = \{(x, y) \in \Omega : x \in (h_k, 2h_k)\}$:

$$\begin{aligned}\|v_k\|_{\Omega_h} &\le c\|v_k\|_{\omega_h} \le c(\|w\|_{\omega_h} + \|v_k - w\|_\omega) \\ &= c\left(\left(\int_0^1 \int_{h_k}^{2h_k} \left[w(0, y) + \int_0^x w_\eta(\eta, y)\, d\eta \right]^2 dx\, dy \right)^{\frac{1}{2}} + \|v_k - w\|_\omega \right) \\ &\le c\left(h_k^{\frac{1}{2}} \left(\int_{\Gamma_W} w^2 \, dy \right)^{\frac{1}{2}} + h_k \|w_x\| + h_k^2 \|w\|_{H^2(\omega)} \right) \le c \left(h_k^{\frac{1}{2}} + \frac{h_k^2}{\varepsilon} \right) \|e_k\|.\end{aligned}$$

We proceed estimating terms from I_{Ω_h}, where we use the previous result:

$$\begin{aligned}
\varepsilon\,((e_k)_y,(w-v_k)_y)_{\Omega_h} &\le \varepsilon\|(e_k)_y\|_{\Omega_h}\,(\|w_y\|+\|(v_k)_y\|_{\Omega_h})\\
&\le c\,\varepsilon^{\frac12}h_k\|f_k\|\,(\varepsilon^{-\frac12}\|e_k\|+h_k^{-1}\|v_k\|_{\Omega_h})\\
&\le c\,\varepsilon^{\frac12}h_k\|f_k\|\left(\varepsilon^{-\frac12}+h_k^{-\frac12}+\frac{h_k}{\varepsilon}\right)\|e_k\|\le c\left(h_k+\frac{h_k^2}{\sqrt{\varepsilon}}\right)\|f_k\|\|e_k\|,\\
\varepsilon_k\,((e_k)_x,(w-v_k)_x)_{\Omega_h} &\le \varepsilon_k\|(e_k)_x\|_{\Omega_h}\,(\|w_x\|+\|(v_k)_x\|_{\Omega_h})\\
&\le c\,h_k^{\frac12}\varepsilon_k\|f_k\|\,(\|e_k\|+h_k^{-1}\|v_k\|_{\Omega_h})\le c\left(h_k+\frac{h_k^{\frac52}}{\varepsilon}\right)\|f_k\|\|e_k\|,\\
((e_k)_x,w-v_k)_{\Omega_h} &\le \|(e_k)_x\|_{\Omega_h}(\|w\|_{\Omega_h}+\|v_k\|_{\Omega_h})\\
&\le c\,h_k^{\frac12}\|f_k\|\left(h_k^{\frac12}\left(\int_{\Gamma_W}w^2\,dy\right)^{\frac12}+h_k\|w_x\|_{\Omega_h}+\|v_k\|_{\Omega_h}\right)\\
&\le c\left(h_k+\frac{h_k^{\frac52}}{\varepsilon}\right)\|f_k\|\|e_k\|.
\end{aligned}$$

Inserting these estimates into (10.11) and using $\varepsilon\le\frac12 h_k$ we obtain

$$\|e_k\|^2\le c\,\frac{h_k^2}{\varepsilon}\|f_k\|\|e_k\|+c\left(h_k+\frac{h_k^2}{\sqrt{\varepsilon}}+\frac{h_k^{\frac52}}{\varepsilon}\right)\|f_k\|\|e_k\|\le c\,\frac{h_k^2}{\varepsilon}\|f_k\|\|e_k\|.$$

and thus the theorem is proved. ∎

For the proof of Theorem 10.4 we first formulate two lemmas.

LEMMA 10.2. *Consider a function $g\in\mathrm{H}^1(\Omega)$. The solution of*

$$-\varepsilon_k u_{xx}-\varepsilon u_{yy}+u_x=g_x \tag{10.12}$$

with boundary conditions as in (1.2) *satisfies*

$$\int_{\Gamma_E}u^2\,dy\le c\left(h_k^{-1}\,\|g\|^2+\int_{\Gamma_E}g^2\,dy+h_k\,\|g_x\|^2\right). \tag{10.13}$$

Proof. We multiply (10.12) with u and integrate by parts to get

$$\varepsilon_k\|u_x\|^2+\varepsilon\|u_y\|^2+\frac12\int_{\Gamma_E}u^2\,dy=-(g,u_x)+\int_{\Gamma_E}g\,u\,dy. \tag{10.14}$$

For the right-hand side in (10.14) we have

$$|(g,u_x)|\le\|g\|\|u_x\|\le c\,\|g\|\|g_x\|\le c\,\left(h_k^{-1}\,\|g\|^2+h_k\|g_x\|^2\right)$$

and

$$\int_{\Gamma_E}g\,u\,dy\le\int_{\Gamma_E}g^2\,dy+\frac14\int_{\Gamma_E}u^2\,dy.$$

Combining these estimates and (10.14) the lemma is proved. ∎

LEMMA 10.3. *Assume $g \in H^1$ and $g|_{\Gamma_E} = 0$, let u be the corresponding solution of* (10.12). *Then the following holds:*

$$\|u\| \le c \left(\|g\| + h_k \|g_x\| + \left(\int_{\Gamma_W} g^2 \, dy \right)^{\frac{1}{2}} + h_k \left(\int_{\Gamma_W} u_x^2 \, dy \right)^{\frac{1}{2}} \right). \tag{10.15}$$

(Note that the standard a priori estimates would give only $\|u\| \le c \|g_x\|$.)

Proof. Consider the auxiliary function $v(x,y) := \int_0^x u(\xi, y) \, d\xi$. It satisfies

$$-\varepsilon_k v_{xx} - \varepsilon v_{yy} + v_x = g + \varepsilon_k u_{in} + g_{in}, \tag{10.16}$$

with $u_{in}(x,y) = u_x(0,y)$ and $g_{in} = g(0,y)$. The corresponding boundary conditions are

$$v_x = u(1,y) \text{ on } \Gamma_E \,, \qquad v = 0 \text{ on } \partial\Omega \setminus \Gamma_E. \tag{10.17}$$

Then the estimate (10.15) is equivalent to

$$\|v_x\| \le c \left(\|g\| + h_k \|g_x\| + \left(\int_{\Gamma_W} g^2 \, dy \right)^{\frac{1}{2}} + h_k \left(\int_{\Gamma_W} u_x^2 \, dy \right)^{\frac{1}{2}} \right). \tag{10.18}$$

The estimate (10.18) is proved by the following arguments. We multiply (10.16) with v_x and integrate by parts to obtain

$$\begin{aligned} &\|v_x\|^2 + \frac{\varepsilon}{2} \int_{\Gamma_E} (v_y)^2 \, dy + \frac{\varepsilon_k}{2} \int_{\Gamma_W} (v_x)^2 \, dy \\ &\qquad = (g, v_x) + \varepsilon_k (u_{in}, v_x) + (g_{in}, v_x) + \frac{\varepsilon_k}{2} \int_{\Gamma_E} (v_x)^2 \, dy. \end{aligned} \tag{10.19}$$

Since $g|_{\Gamma_E} = 0$ the estimate (10.13) yields

$$\int_{\Gamma_E} (v_x)^2 \, dy = \int_{\Gamma_E} u^2 \, dy \le c \left(h_k^{-1} \|g\|^2 + h_k \|g_x\|^2 \right). \tag{10.20}$$

Now (10.18) follows from (10.19) by applying the Cauchy inequality and estimate (10.20). ∎

Using these lemmas we can prove the final result we need.

THEOREM 10.4. *For $f \in \mathbb{V}_k^0$ let u and $\tilde{u}$ be the continuous solutions defined in* (10.1). *Then the following holds:*

$$\|u - \tilde{u}\| \le c \, h_k \|f_k\|. \tag{10.21}$$

Proof. The difference $e := u - \tilde{u}$ solves the equation

$$-\varepsilon_k e_{xx} - \varepsilon e_{yy} + e_x = g_x, \tag{10.22}$$

with $g = -\bar{\delta} h_k \tilde{u}_x$ and boundary conditions as in (1.2). Now the result of Lemma 10.3 can be applied. We obtain

$$\begin{aligned} \|e\| &\le c \left(\|g\| + h_k \|g_x\| + \left(\int_{\Gamma_W} g^2 \, dy \right)^{\frac{1}{2}} + h_k \left(\int_{\Gamma_W} e_x^2 \, dy \right)^{\frac{1}{2}} \right) \\ &\le c \, h_k \left(\|\tilde{u}_x\| + h_k \|\tilde{u}_{xx}\| + \left(\int_{\Gamma_W} u_x^2 \, dy \right)^{\frac{1}{2}} + \left(\int_{\Gamma_W} \tilde{u}_x^2 \, dy \right)^{\frac{1}{2}} \right). \end{aligned}$$

To estimate the norms $\|\tilde{u}_x\|$ and $\|\tilde{u}_{xx}\|$ we use a priori bounds from Theorem 4.1. Further we use the fact that $f_k = 0$ in Ω_k^{in}. Due to the choice of Ω_k^{in} (cf. (5.18)) we can apply Corollary 5.2 with $\xi = h_k$, $\eta = \varepsilon_k |\ln h_k| + h_k$, and $p = \frac{1}{2}$. Using (5.5) and $\varepsilon_k \geq \frac{1}{3} h_k$ we get $\int_{\Gamma_W} u_x^2 \, dy \leq c \, \|f_k\|^2$. The same estimate holds for $\int_{\Gamma_W} \tilde{u}_x^2 \, dy$. Thus we obtain $\|e\| \leq c \, h_k \|f_k\|$. ∎

11. Numerical experiments. In this section we present results of a few numerical experiments to illustrate that in a certain sense our analysis is sharp. In particular it will be shown that the nonstandard splitting in (6.8) which forms the basis of our convergence analysis reflects some important phenomena.

In the experiments we use the following parameters. For $\bar{\delta}$ in (2.4) we take $\bar{\delta} = \frac{1}{2}$. The pre- and postsmoother are as in (6.2), (6.4) with $\omega_k = 1$. We take a random right-hand side vector and a starting vector equal to zero. For the stopping criterion we take a reduction of the relative residual by a factor 10^9 . Thus in the tables below convergence is measured in the norm $\|\cdot\|_{A^T A}$. We use the notation $Pe_h := \frac{h}{2\varepsilon}$.

First we present results for a standard V-cycle with $\mu_k = \nu_k = 2$. In Table 11.1 we give the number of iterations needed to satisfy the stopping criterion and (between brackets) the average residual reduction per iteration. These results clearly show robustness of the multigrid solver. For a W-cycle we also observed robust results.

TABLE 11.1
Multigrid convergence: V-cycle with $\nu_k = \nu_k = 2$.

	h			
Pe_h	1/8	1/32	1/128	1/512
1	**8**(0.06)	**10**(0.12)	**11**(0.13)	**11**(0.13)
10	**7**(0.04)	**8**(0.07)	**8**(0.07)	**8**(0.07)
1e+3	**8**(0.05)	**11**(0.14)	**11**(0.14)	**11**(0.14)
1e+5	**7**(0.04)	**11**(0.14)	**11**(0.14)	**11**(0.14)

Number of iterations and average reduction factor.

If we consider only the smoother and do not use a coarse grid correction, then for $\varepsilon \approx h$ this method has an h-dependent convergence rate. This is illustrated in Table 11.2.

We consider the standard splitting in the convergence analysis based on the smoothing and approximation property. For $\varepsilon = h^2$ some results are presented in Table 11.3. The estimates that are given in this table result from the computation of

$$\frac{\|(A_h^{-1} - p A_{2h}^{-1} r)\hat{f}\|}{\|\hat{f}\|} \quad \text{and} \quad \frac{\|(A_h S_h^2)\hat{f}\|}{\|\hat{f}\|}$$

TABLE 11.2
h-dependence of convergence of the smoothing iterations.

	h			
Pe_h	1/8	1/32	1/128	1/512
1	**119**(0.83)	**244**(0.91)	**533**(0.94)	**1495**(0.986)
10	**26**(0.44)	**51**(0.61)	**66**(0.72)	**173**(0.88)

Number of iterations and average reduction factor.

with $\hat{f} \in \mathbb{V}_h$ a discrete point source in the grid point $(\frac{1}{2}, \frac{1}{2})$. These results indicate $\mathcal{O}(h^{-1})$ behavior for the smoothing property (as expected) and $\mathcal{O}(\sqrt{h})$ behavior for the approximation property. Hence this splitting is not satisfactory for proving a robustness result.

TABLE 11.3
Standard splitting for approximation and smoothing properties.

	h			
Estimates for	1/8	1/32	1/128	1/512
$\|A_h^{-1} - pA_{2h}^{-1}r\|$	8.4e-2	5.0e-2	2.7e-2	1.4e-2
$\|A_h S_h^2\|$	1.25	4.48	17.7	70.8

The proof of the modified approximation property is based on the result in Theorem 10.1. In that theorem a $\frac{h_k^2}{\varepsilon}$ bound is proved *provided the right-hand side function f_k is zero close to the inflow boundary.* We performed an experiment with a function f_k which has values equal to one in all grid points $(h_k, jh_k),\ j = 1, \ldots, n_k$, and zero elsewhere. Results are given in Table 11.4. We observe an $h_k^{-\frac{1}{2}}$ effect. This justifies the splitting using the cut-off operator Φ_k.

TABLE 11.4
Approximation property if f_k has support near inflow.

	h			
Pe_h	1/8	1/32	1/128	1/512
1	0.31	0.60	1.23	2.53
10	0.07	0.17	0.23	0.46

Values of $\frac{\varepsilon}{h^2}\|(A_h^{-1} - pA_{2h}^{-1}r)f\|/\|f\|$.

Finally we performed a numerical experiment related to the result in Theorem 6.3. For the smoother we computed residual reduction factors in the almost degenerated norm $\|\Phi_k^{\frac{1}{2}} \cdot\|$ with $\Phi_k := I_{n_{k-1}} \otimes \mathrm{diag}(\phi)$ and

$$\phi_i = \begin{cases} 1 & \text{for } 1 \le i < 5, \\ \exp(4-i) & \text{for } 5 \le i \le n_k. \end{cases}$$

For the relaxation parameter ω in the smoother we take the value $\omega = 1.2$. The results in Table 11.5 show h-independent and "fast" convergence of the smoother in this norm.

TABLE 11.5
Residual reduction of the smoother in the $\|\Phi^{\frac{1}{2}} \cdot\|$-norm.

	h			
Pe_h	1/8	1/32	1/128	1/512
1	**93**(0.8)	**131**(0.85)	**133**(0.85)	**133**(0.85)
10	**23**(0.40)	**28**(0.47)	**28**(0.47)	**28**(0.47)

Number of iterations and average reduction factor.

Acknowledgment. The authors thank the referees for valuable comments which lead to a significant improvement of the paper.

REFERENCES

[1] O. Axelsson and W. Layton, *Defect-correction methods for convection-dominated convection-diffusion problems*, RAIRO Model. Math. Anal. Numer., 24 (1990), pp. 423–455.

[2] O. Axelsson and M. Nikolova, *Adaptive refinement for convection-diffusion problems based on a defect-correction technique and finite difference method*, Computing, 58 (1997), pp. 1–30.

[3] R. E. Bank and M. Benbourenane, *The hierarchical basis multigrid method for convection-diffusion equations*, Numer. Math., 61 (1992), pp. 7–37.

[4] J. Bey and G. Wittum, *Downwind numbering: Robust multigrid for convection-diffusion problems*, Appl. Numer. Math., 23 (1997), pp. 177–192.

[5] J. H. Bramble, *Multigrid Methods*, Longman, Harlow, UK, 1993.

[6] J. H. Bramble, J. E. Pasciak, and J. Xu, *The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems*, Math. Comp., 51 (1988), pp. 389–414.

[7] K. Eriksson and C. Johnson, *Adaptive streamline diffusion finite element methods for stationary convection-diffusion problems*, Math. Comp., 60 (1993), pp. 167–188, S1–S2.

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

[9] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, 1985.

[10] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, New York, 1994.

[11] W. Hackbusch, *Multigrid convergence for a singular perturbation problem*, Linear Algebra Appl., 58 (1984), pp. 125–145.

[12] W. Hackbusch, *A note on Reusken's Lemma*, Computing, 55 (1995), pp. 181–189.

[13] W. Hackbusch and T. Probst, *Downwind Gauss-Seidel smoothing for convection dominated problems*, Numer. Linear Algebra Appl., 4 (1997), pp. 85–102.

[14] K. Johannsen, *Robust Smoothers for Convection-Diffusion Problems*, Preprint IWR, University of Heidelberg, Heidelberg, Germany, 1999.

[15] C. Johnson, A. H. Schatz, and L. B. Wahlbin, *Crosswind smear and pointwise errors in streamline diffusion finite element methods*, Math. Comp., 49 (1987), pp. 25–38.

[16] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comp. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[17] J. Mandel, *Multigrid convergence for nonsymmetric, indefinite variational problems and one smoothing step*, Appl. Math. Comput., 19 (1986), pp. 201–216.

[18] W. Mulder, *A new multigrid approach to convection problems*, J. Comput. Phys., 83 (1989), pp. 303–323.

[19] N. H. Naik and J. van Rosedale, *The improved robustness of multigrid elliptic solvers based on multiple semicoarsened grids*, SIAM J. Numer. Anal., 30 (1993), pp. 215–229.

[20] K. Niijima, *Pointwise error estimates for a streamline diffusion finite element scheme*, Numer. Math., 56 (1990), pp. 707–719.

[21] M. A. Olshanskii and A. Reusken, *On the convergence of a multigrid method for linear reaction-diffusion problem*, Computing, 65 (2000), pp. 193–202.

[22] M. A. Olshanskii and A. Reusken, *Navier-Stokes equations in rotation form: A robust multigrid solver for the velocity problem*, SIAM J. Sci. Comput., 23 (2002), pp. 1683–1706.

[23] I. Persson, K. Samuelsson, and A. Szepessy, *On the convergence of multigrid methods for flow problems*, Electron. Trans. Numer. Anal., 8 (1999), pp. 46–87.

[24] A. Reusken, *On maximum norm convergence of multigrid methods for two-point boundary value problems*, SIAM J. Numer. Anal., 29 (1992), pp. 1569–1578.

[25] A. Reusken, *Multigrid with matrix-dependent transfer operators for a singular perturbation problem*, Computing, 50 (1994), pp. 199–211.

[26] A. Reusken, *Fourier analysis of a robust multigrid method for convection-diffusion equations*, Numer. Math., 71 (1995), pp. 365–397.

[27] A. Reusken, *Convergence analysis of a multigrid method for convection-diffusion equations*, Numer. Math., 91 (2002), pp. 323–349.

[28] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.

[29] R. P. Stevenson, *New estimates of the contraction number of V-cycle multi-grid with applications to anisotropic equations*, in Incomplete Decompositions, Proceedings of the Eighth GAMM Seminar, W. Hackbusch and G. Wittum, eds., Notes Numer. Fluid Mech., 41 Vieweg, Braunschweig, Germany, (1993), pp. 159–167.
[30] R. P. Stevenson, *Robustness of multi-grid applied to anisotropic equations on convex domains and on domains with re-entrant corners*, Numer. Math., 66 (1993), pp. 373–398.
[31] L. B. Wahlbin, *Local behavior in finite element methods*, in Handbook of Numerical Analysis, Vol. II, Finite Element Methods, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 353–522.
[32] J. Wang, *Convergence analysis of multigrid algorithms for nonselfadjoint and indefinite elliptic problems*, SIAM J. Numer. Anal., 30 (1993), pp. 275–285.
[33] P. Wesseling, *An Introduction to Multigrid Methods*, Wiley, Chichester, UK, 1992.
[34] G. Wittum, *On the robustness of ILU smoothing*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 699–717.
[35] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[36] H. Yserentant, *Old and new convergence proofs for multigrid methods*, Acta Numer., (1993), pp. 285–326.
[37] P. M. de Zeeuw, *Matrix-dependent prolongations and restrictions in a blackbox multigrid solver*, J. Comput. Appl. Math., 33 (1990), pp. 1–27.
[38] G. Zhou, *How accurate is the streamline diffusion finite element method*, Math. Comp., 66 (1997), pp. 31–44.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1292–1323

DIFFERENCE APPROXIMATIONS OF THE NEUMANN PROBLEM FOR THE SECOND ORDER WAVE EQUATION*

HEINZ-OTTO KREISS†, N. ANDERS PETERSSON‡, AND JACOB YSTRÖM§

Abstract. Stability theory and numerical experiments are presented for a finite difference method that directly discretizes the Neumann problem for the second order wave equation. Complex geometries are discretized using a Cartesian embedded boundary technique. Both second and third order accurate approximations of the boundary conditions are presented. Away from the boundary, the basic second order method can be corrected to achieve fourth order spatial accuracy. To integrate in time, we present both a second order and a fourth order accurate explicit method. The stability of the method is ensured by adding a small fourth order dissipation operator, locally modified near the boundary to allow its application at all grid points inside the computational domain. Numerical experiments demonstrate the accuracy and long-time stability of the proposed method.

Key words. wave equation, stability, accuracy, embedded boundary

AMS subject classifications. 65M06, 65M12

DOI. 10.1137/S003614290342827X

1. Introduction.

There are many methods to solve the wave equation numerically. Methods based on variational principles [1] have the advantage that the energy is conserved, but they are not as efficient as difference methods. On the other hand, difference methods are prone to instabilities. To avoid these one often has to add dissipative terms, and the energy is not conserved. Luckily, the instabilities are often weak and caused by high frequency waves which are not accurately represented anyway. Therefore, one constructs the dissipation in such a way that it acts mainly only on these frequencies. We feel that the fixation on energy conservation often goes too far. Large phase-errors can destroy the solution as well.

In this paper we continue the development of numerical methods that directly discretize the second order wave equation without first rewriting it as a system of first order equations. In particular, we want to discuss the kind of instabilities that can arise and how to control them. Since we treated the Dirichlet problem in [9], we consider here only the Neumann problem

$$\begin{aligned} u_{tt} &= \Delta u + F(\boldsymbol{x},t), \quad \boldsymbol{x}\in\Omega,\ t>0,\\ \frac{\partial u}{\partial n}(\boldsymbol{x},t) &= f(\boldsymbol{x},t), \quad \boldsymbol{x}\in\Gamma,\ t>0,\\ u(\boldsymbol{x},0) = u_0(\boldsymbol{x}), &\quad u_t(\boldsymbol{x},0) = u_1(\boldsymbol{x}), \quad \boldsymbol{x}\in\Omega, \end{aligned} \tag{1.1}$$

where Ω is a bounded one- or two-dimensional domain with boundary Γ.

*Received by the editors May 15, 2003; accepted for publication (in revised form) January 20, 2004; published electronically October 28, 2004. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

http://www.siam.org/journals/sinum/42-3/42827.html

†Department of Mathematics, University of California, Los Angeles, CA 90024 (Kreiss@math.ucla.edu).

‡Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551 (andersp@llnl.gov).

§Department of Numerical Analysis and Computing Science, Royal Institute of Technology, S-100 44 Stockholm, Sweden (yxan@nada.kth.se).

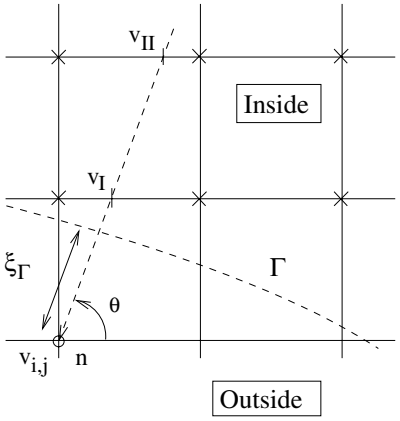


FIG. 1. *The points used for discretizing the Neumann boundary condition.*

We will discretize (1.1) on a Cartesian embedded boundary grid. The embedded boundary technique for discretizing partial differential equations dates back to the first order technique by Weller and Shortley [15] and the higher order generalizations of Collatz [3]. More recently, several embedded boundary methods have been presented for various types of partial differential equations. For example, Pember et al. [12] used a Cartesian grid method for solving the time-dependent equations of gas dynamics. Zhang and LeVeque [16] solved the acoustic wave equation with discontinuous coefficients written as a first order system. They derived special difference stencils that satisfy the jump conditions at the interior interfaces, where the coefficients are discontinuous. A staggered grid method was used by Ditkowski, Dridi, and Hesthaven [4] for solving Maxwell's equations on a Cartesian grid. The methods described in these papers all solve first order systems (in time). For Poisson's equation with Dirichlet boundary conditions, Johansen and Colella [6] derived an embedded boundary technique based on the finite volume method combined with multigrid.

We proceed by presenting the highlights of our proposed method. The domain Ω is covered by a Cartesian grid with step size h where the grid points are located at $\boldsymbol{x}_{i,j} = (x_i, y_j)^T = (ih, jh)^T$, and the boundary Γ is allowed to cut through the grid in an arbitrary manner; see Figure 1. Let $t_n = nk$, $k = 0, 1, 2, \ldots$, denote the time-discretization with step size k, and let $v_{i,j}^n$ be the difference approximation of $u(x_i, y_j, t_n)$. A second order accurate approximation of the Laplacian of u is given by

$$\Delta_h v_{i,j}^n =: \frac{1}{h^2}(v_{i+1,j}^n + v_{i-1,j}^n + v_{i,j+1}^n + v_{i,j-1}^n - 4v_{i,j}^n). \tag{1.2}$$

To be able to evaluate $\Delta_h v_{i,j}^n$ at all grid points inside Ω, we use ghost points just outside the domain. Consider the case in Figure 1 where the grid point $\boldsymbol{x}_{i,j}$ is outside of Ω, but $\boldsymbol{x}_{i,j+1}$ is inside. To aid in the approximation of the Neumann boundary condition, we construct a third order accurate interpolant between three points along the normal: $(0, v_{i,j}^n)$, (ξ_I, v_I^n), (ξ_{II}, v_{II}^n). Here ξ_I and $\xi_{II} = 2\xi_I$ are the distances between $\boldsymbol{x}_{i,j}$, along the normal going through that point, and the horizontal grid lines y_{j+1} and y_{j+2}, respectively. After differentiating the interpolant, we get a second order accurate approximation of the (outward) normal derivative

$$D_n^{(2)} v_{i,j}^n =: g_0 v_{i,j}^n + g_I v_I^n + g_{II} v_{II}^n = \frac{\partial v}{\partial n}(\boldsymbol{x}_{i,j}^\Gamma, t_n) + O(h^2), \tag{1.3}$$

where $\boldsymbol{x}_{i,j}^{\Gamma}$ is the intersection point between the boundary and the normal going through $\boldsymbol{x}_{i,j}$. The coefficients g_j are given by

$$g_0 = \frac{3\xi_I - 2\xi_\Gamma}{2\xi_I^2}, \quad g_I = \frac{2\xi_\Gamma - 2\xi_I}{\xi_I^2}, \quad g_{II} = \frac{\xi_I - 2\xi_\Gamma}{2\xi_I^2}, \tag{1.4}$$

where ξ_Γ is the distance between $\boldsymbol{x}_{i,j}$ and the boundary. Since the coefficients $g_j = \mathcal{O}(1/h)$, we need to use third order accurate approximations for v_I^n and v_{II}^n. Here we use Lagrangian interpolation along the grid lines y_{j+1} and y_{j+2}:

$$\begin{aligned} v_I^n &= c_0 v_{i,j+1}^n + c_1 v_{i+1,j+1}^n + c_2 v_{i+2,j+1}^n, \\ v_{II}^n &= c_3 v_{i,j+2}^n + c_4 v_{i+1,j+2}^n + c_5 v_{i+2,j+2}^n. \end{aligned}$$

The resulting formula for $D_n^{(2)} v_{i,j}$ holds when the angle θ between the x-axis and the normal satisfies $\pi/4 \le \theta \le \pi/2$. When $0 \le \theta \le \pi/4$, the horizontal interpolations to obtain v_I and v_{II} are replaced by corresponding interpolations in the vertical direction. The expressions in the remaining three quadrants are simply obtained by reflections in index space, leading to a total of eight different cases to treat all possible directions of the boundary.

The second order boundary condition formula results in an overall second order scheme, but since the boundary condition is discretized using one-sided differences, the truncation error will be larger at the boundary than in the interior, where a centered scheme is used. We can easily modify the above technique to construct a third order accurate formula $D_n^{(3)} v_{i,j}$ to make the coefficient in front of the leading second order truncation error term smaller. In this case, three interior values v_I, v_{II}, and v_{III} are interpolated using fourth order Lagrangian interpolation along three contiguous grid lines. Hence this stencil involves 12 interior points. The third order boundary condition formula works nicely for well-resolved geometries where there are enough interior points. For less resolved geometries, or for very thin regions where two parts of the boundary are close to each other, we will use the second order boundary condition formula.

All ghost point values in (1.2) can be eliminated using formulas of the type (1.3). The discrete approximation of the Laplacian of u (for functions subject to the boundary condition $\partial u/\partial n = f(\boldsymbol{x}^{\Gamma}, t)$) can then be written in matrix form:

$$\Delta u = A\boldsymbol{v} + \boldsymbol{b}(t) + O(h^2). \tag{1.5}$$

Here the array $\boldsymbol{v}$ contains the solution at all grid points inside Ω, and $\boldsymbol{b}(t)$ is the discrete counterpart of the boundary forcing $f(\boldsymbol{x}_\Gamma, t)$.

Because of the discretized form of the Neumann boundary condition, the matrix A will not be symmetric. As a result, the basic scheme proposed in [9],

$$\frac{\boldsymbol{v}^{n+1} - 2\boldsymbol{v}^n + \boldsymbol{v}^{n-1}}{k^2} = A\boldsymbol{v}^n + \boldsymbol{b}(t_n) + \boldsymbol{F}(t_n),$$

suffers from a weak instability (here $\boldsymbol{F}(t_n)$ is the discretized version of the internal forcing $F(\boldsymbol{x}, t_n)$). The definition of a weak instability will be given in section 4. To understand the loss of stability, we analyze a number of model problems. We start with the one-dimensional half-plane (section 2) and strip (section 3) problems, proving that the difference approximation is stable in these cases, without damping. The two-dimensional case is analyzed in sections 4–6, where we show that the tangential

derivatives that occur in the truncation error of the boundary condition can lead to instabilities, both for the half-plane and strip problems. We also show that our scheme can be stabilized by a small fourth order artificial dissipation of the type $h^3\Delta^2\boldsymbol{v}_t$. However, a centered finite difference stencil such as $\Delta_h^2\boldsymbol{v}_t$ is wider than the discretized Laplacian, so it is not possible to use this damping term all the way up to the boundary (without adding extra numerical boundary conditions). Instead, we suggest using the discrete operator $h^3A^T(A(\boldsymbol{v}^n-\boldsymbol{v}^{n-1})/k)$ which *can* be applied all the way up to the boundary. Away from the boundary, it is equivalent to $\Delta_h^2(\boldsymbol{v}^n-\boldsymbol{v}^{n-1})/k$. For the general case with inhomogeneous boundary conditions and internal forcing, the proposed scheme becomes

$$\frac{\boldsymbol{v}^{n+1}-2\boldsymbol{v}^n+\boldsymbol{v}^{n-1}}{k^2} = A\boldsymbol{v}^n+\boldsymbol{b}(t_n)+\boldsymbol{F}(t_n) - \alpha h^3A^T\left(A(\boldsymbol{v}^n-\boldsymbol{v}^{n-1})/k+\frac{d\boldsymbol{b}}{dt}(t_n)\right). \tag{1.6}$$

We note that the sparse structure of A can be used to efficiently evaluate both $A\boldsymbol{v}$ and $A^T\boldsymbol{v}$, without the need to store the matrix explicitly; see Appendix A.

In section 7.1, we will demonstrate that this discretization does not suffer from the "small cell" stiffness problem that commonly is encountered when the finite volume method is used on a Cartesian grid with an embedded boundary; cf. [2]. We will also show that the damping term inflicts an $\mathcal{O}(h^2)$ perturbation of the undamped scheme (section 7.2), and by numerical experiments in section 8 we will demonstrate that it suffices to take α very small (of the order $\mathcal{O}(10^{-3})$). Hence, the resulting numerical solution will be second order accurate, and the scheme is well suited for long-time calculations where it is important to keep damping to a minimum. In section 7.3, we also present correction terms that optionally can be added to make the scheme fourth order accurate in time and space (away from the boundary). A number of numerical examples are presented in section 8 to assess the accuracy and long-time stability of the method with and without fourth order corrections, both for smooth boundaries and in the presence of corners. The proposed method is finally used for a resonance analysis of wave propagation in a harbor.

2. The one-dimensional half-plane problem. We start with the half-plane problem

$$\begin{gathered} u_{tt}=u_{xx},\quad 0\le x<\infty,\ t\ge 0,\\ u(x,0)=f(x),\end{gathered} \tag{2.1}$$

with boundary conditions

$$u_x(0,t)=0,\qquad \lim_{x\to\infty}u(x,t)=0. \tag{2.2}$$

Let $x_\nu=\nu h$, $h>0$, denote the grid points, $v(x_\nu,t)$ be a grid function, and $D_+v(x_\nu,t)=(v(x_{\nu+1},t)-v(x_\nu,t))/h$ represent the usual forward difference operator. We want to solve (2.1), (2.2) by the simplest central difference approximation

$$\begin{gathered} v_{tt}(x_\nu,t)=D_+D_-v(x_\nu,t),\quad \nu=1,2,\dots,\\ v(x_\nu,0)=f(x_\nu),\end{gathered} \tag{2.3}$$

with boundary conditions

$$D_+v(0,t)+\alpha hD_+^2v(0,t)+\beta h^2D_+^3v(0,t)=0,\qquad \lim_{x_\nu\to\infty}v(x_\nu,t)=0. \tag{2.4}$$

If we set $\alpha = \beta = 0$ or $\alpha = -\frac{1}{2},\ \beta = 0$, we obtain a first order or second order accurate approximation, respectively. In these cases we can prove stability by energy estimates; see [9]. If $\alpha = -\frac{1}{2}$ and $\beta = \frac{1}{3}$, we obtain the third order accurate approximation

$$D_+ v(0,t) - \frac{1}{2} h D_+^2 v(0,t) + \frac{1}{3} h^2 D_+^3 v(0,t) = 0. \tag{2.5}$$

In this case, we do not know how to prove stability by energy estimates. Instead, we will use mode analysis.

For simplicity, we keep time continuous. In actual calculations we use the method of lines. In [10] we have shown that the stability of the semidiscrete approximation implies the stability of the totally discretized method for most standard methods of lines.

By stability we mean here that there are no exponentially growing solutions. Therefore, a test for stability is that (2.3), (2.4) has no solutions of type

$$v(x_\nu, t) = e^{st} \varphi(x_\nu), \quad |\varphi(x_\nu)| \leq \text{const} \tag{2.6}$$

for $\operatorname{Re} s > 0$, satisfying the boundary condition (2.4). Introducing (2.6) into (2.3) gives us

$$h^2 s^2 \varphi(x_\nu) = h^2 D_+ D_- \varphi(x_\nu) = \varphi(x_\nu + h) - 2\varphi(x_\nu) + \varphi(x_\nu - h). \tag{2.7}$$

Since (2.7) is a difference equation with constant coefficients, its general solution is of the form

$$\varphi(x_\nu) = \sigma_1 \kappa_1^\nu + \sigma_2 \kappa_2^\nu, \tag{2.8}$$

where κ_1, κ_2 are solutions of the characteristic equation

$$(\kappa - 1)^2 - h^2 s^2 \kappa = 0. \tag{2.9}$$

We have $\kappa_2 = \kappa_1^{-1}$, and we simplify the notation by removing the index of the roots and set $\kappa_1 = \kappa$, $\kappa_2 = \kappa^{-1}$.

LEMMA 2.1. *For $|hs| \ll 1$, the roots of* (2.9) *are of the form*

$$\begin{aligned} \kappa &= 1 - hs + \frac{h^2 s^2}{2} + \mathcal{O}(h^3 s^3) = e^{-hs(1+\mathcal{O}(h^2 s^2))}, \\ \kappa^{-1} &= 1 + hs + \frac{h^2 s^2}{2} + \mathcal{O}(h^3 s^3) = e^{hs(1+\mathcal{O}(h^2 s^2))}. \end{aligned} \tag{2.10}$$

Also, for $\operatorname{Re} s > 0$, (2.9) *has no root with $|\kappa| = 1$ and exactly one root κ with $|\kappa| < 1$.*

Proof. Equation (2.10) follows by asymptotic expansion of the roots. (It is not surprising: The corresponding solutions of (2.1) are e^{-sx}, e^{sx}, and (2.3) is second order accurate.)

Assume that (2.9) has a solution

$$|\kappa| = 1, \quad \text{i.e.,} \quad \kappa = e^{i\tau},\ \tau \text{ real},$$

for some s with $\operatorname{Re} s > 0$. Then (2.9) becomes

$$-4 \sin^2(\tau/2) = h^2 s^2.$$

Therefore, $\operatorname{Re} s = 0$, which is a contradiction.

For $hs \to \infty$, $s > 0$ real, the solutions of (2.9) satisfy

$$\lim_{hs\to\infty} \kappa = 0, \qquad \lim_{hs\to\infty} \kappa^{-1} = \infty.$$

Since the roots are smooth functions of s and $|\kappa| \neq 1$ for $\operatorname{Re} s > 0$, we always have $|\kappa| < 1$, $|\kappa^{-1}| > 1$. This proves the lemma. □

The lemma shows that the solution can only stay bounded in space if $\sigma_2 = 0$, so

$$\varphi(x_\nu) = \sigma_1 \kappa^\nu, \quad |\kappa| < 1. \tag{2.11}$$

Introducing (2.11) into the boundary condition (2.4) gives us

$$(\kappa - 1)(1 + \alpha(\kappa - 1) + \beta(\kappa - 1)^2) = 0. \tag{2.12}$$

The cubic equation (2.12) has three roots $\kappa = \kappa_j$, $j = 1, 2, 3$, which lead to possible solutions of (2.6). We obtain the corresponding s from the characteristic equation (2.9), i.e.,

$$hs = \pm\sqrt{\frac{(\kappa - 1)^2}{\kappa}} = \pm(\kappa^{1/2} - \kappa^{-1/2}). \tag{2.13}$$

The first root, $\kappa_1 = 1$, does not generate a growing solution. In fact, any root with $|\kappa| = 1$ has this property, since inserting $\kappa = e^{i\xi}$ into (2.13) yields

$$hs = \pm 2i \sin\frac{\xi}{2}, \quad \text{i.e.,} \quad \operatorname{Re} s = 0. \tag{2.14}$$

Roots with $|\kappa| > 1$ are not permissible because $\varphi(x_\nu) = \sigma\kappa^\nu$ becomes unbounded as $\nu \to \infty$ and violates the boundary condition (2.4). However, solutions of the type $\kappa = e^{i\xi - \eta}$, ξ, η real, $\eta > 0$, correspond to

$$h\operatorname{Re} s = \pm(e^{-\eta/2} - e^{\eta/2}) \cos\frac{\xi}{2},$$

which grows rapidly in time if $\xi \neq \pi + 2n\pi$, $n = 0, 1, 2, \ldots$. These solutions decay rapidly away from the boundary, and we therefore denote these solutions as *boundary layer instabilities*.

Often one tries to stabilize numerical methods by adding a dissipative term to the difference equation. Instead of (2.3), we then consider

$$v_{tt} = D_+ D_- v + \sigma h D_+ D_- v_t.$$

For boundary layer instabilities, this does not work. If the boundary layer is oscillatory, then one can stabilize the method, but the amount of necessary dissipation is, in general, too large for accuracy reasons. Therefore, the only useful boundary condition approximations are those where $|\kappa_2| > 1$, $|\kappa_3| > 1$. While this condition is violated for general coefficients α, β, it is easy to see that the third order approximation (2.5) satisfies the requirement. That approximation has $\alpha = -\frac{1}{2}$, $\beta = \frac{1}{3}$, and (2.12) has the solutions

$$\kappa_1 = 1, \quad \kappa_{2,3} = \frac{7}{4} \pm i\sqrt{3 - \frac{9}{16}}, \quad |\kappa_{2,3}| = \frac{\sqrt{88}}{4} > \frac{9}{4}. \tag{2.15}$$

3. The one-dimensional strip problem. We consider now the wave equation (2.1) for $0 \le x \le 1$, $t \ge 0$. As boundary conditions we use

$$u_x(0,t) = 0, \qquad u(1,t) = 0. \tag{3.1}$$

We approximate the continuous problem by

$$\begin{aligned} v_{tt}(x_\nu,t) = D_+D_-v(x_\nu,t), \quad \nu = 1,2,\dots,N-1, \quad Nh = 1, \\ v(x_\nu,0) = f(x_\nu), \end{aligned} \tag{3.2}$$

with boundary conditions

$$\begin{aligned} L_hv =: D_+v(0,t) - \frac{1}{2}hD_+^2v(0,t) + \frac{1}{3}h^2D_+^3v(0,t) = 0, \\ v(1,t) = 0. \end{aligned} \tag{3.3}$$

For the analytic problem (2.1), (3.1) there is an energy estimate. Also, we can represent the solution by an eigenfunction expansion

$$u(x,t) = \sum_{j=0}^{\infty} e^{\lambda_j t}\psi_j(x).$$

The eigenvalues λ_j are purely imaginary and are solutions of the eigenvalue problem

$$\lambda^2\psi = \psi_{xx}, \qquad \psi_x(0) = \psi(1) = 0. \tag{3.4}$$

Again we want to investigate whether (3.2), (3.3) has exponentially growing solutions. We make the ansatz (2.6) and obtain

$$\begin{aligned} h^2s^2\varphi(x_\nu) = h^2D_+D_-\varphi(x_\nu), \\ L_h\varphi = 0, \quad \varphi(1) = 0, \end{aligned} \tag{3.5}$$

and start our discussion with the case that $|sh| \ll 1$. The discretized eigenvalue problem (3.5) is an approximation of the continuous problem (3.4), and since the difference stencil is compact, solutions of (3.5) with $|sh| \ll 1$ are close to solutions of the continuous problem; see Kreiss [7]. The question is whether the eigenvalues also are purely imaginary.

The general solution of (3.5) is

$$\varphi(x_\nu) = \sigma_1\kappa_1^\nu + \sigma_2\kappa_2^\nu, \tag{3.6}$$

where κ_j, $j = 1,2$, are the solutions of the characteristic equation (2.9)

$$\kappa^2 - (2 + h^2s^2)\kappa + 1 = 0.$$

Therefore, $\kappa_1\kappa_2 = 1$, i.e., $\kappa_2 = \kappa_1^{-1}$. By (2.10) we can write

$$\kappa_1 = e^{h\tilde{s}}, \quad \kappa_2 = e^{-h\tilde{s}}, \quad \tilde{s} =: s(1 + \mathcal{O}(h^2s^2)),$$

and

$$\varphi(x) = \sigma_1e^{\tilde{s}x} + \sigma_2e^{-\tilde{s}x}, \quad x = x_\nu,\ \nu = 0,1,2,\dots,N, \quad Nh = 1.$$

For smooth functions $w(x)$,

$$L_h w = w_x(0) + \gamma_4 h^3 w_{xxxx}(0) + \gamma_5 h^4 w_{xxxxx}(0) + \mathcal{O}(h^5).$$

Therefore, introducing (3.6) into the boundary conditions gives us

$$\sigma_1(1 + \tilde{\gamma}_4 h^3 \tilde{s}^3 + \tilde{\gamma}_5 h^4 \tilde{s}^4) - \sigma_2(1 - \tilde{\gamma}_4 h^3 \tilde{s}^3 + \tilde{\gamma}_5 h^4 \tilde{s}^4) = 0, \tag{3.7}$$

$$\sigma_1 e^{\tilde{s}} + \sigma_2 e^{-\tilde{s}} = 0. \tag{3.8}$$

Here

$$\tilde{\gamma}_4 = \gamma_4 + \gamma_{41}\tilde{s}^2 h^2 + \cdots, \quad \tilde{\gamma}_5 = \gamma_5 + \gamma_{51}\tilde{s}^2 h^2 + \cdots$$

account for the higher order terms.

There is a nontrivial solution of (3.7), (3.8) if and only if

$$\frac{\sigma_2}{\sigma_1} = -e^{2\tilde{s}} = \frac{1 + \tilde{\gamma}_5 h^4 \tilde{s}^4 + \tilde{\gamma}_4 h^3 \tilde{s}^3}{1 + \tilde{\gamma}_5 h^4 \tilde{s}^4 - \tilde{\gamma}_4 h^3 \tilde{s}^3}.$$

If $|sh| \ll 1$, then the eigenvalues of (3.5) converge to the eigenvalues $\lambda_n = i(\frac{\pi}{2} + n\pi)$ of (3.4) where $|h\lambda_n| \ll 1$. Thus we make the ansatz

$$h\tilde{s} = h\lambda_n + ih\tau = ih\mu_n + ih\tau, \quad \mu_n = \frac{\pi}{2} + n\pi,$$

and obtain

$$e^{2i\tau} = \frac{1 + \tilde{\gamma}_5((\mu_n + \tau)h)^4 - i\tilde{\gamma}_4((\mu_n + \tau)h)^3}{1 + \tilde{\gamma}_5((\mu_n + \tau)h)^4 + i\tilde{\gamma}_4((\mu_n + \tau)h)^3} =: S. \tag{3.9}$$

Since $\tilde{\gamma}_4$, $\tilde{\gamma}_5$ are real and bounded and μ_n is real, an asymptotic expansion of $|S|$ in h yields

$$|S| = 1 + \mathcal{O}(|\tau|^3 h^3) = e^{2i\tau}.$$

Hence, τ must be real-valued; that is, $\tilde{s}$ must be purely imaginary. By the above expansion, it follows that there is a unique solution close to λ_n with

$$i\tau = i\tilde{\gamma}_4(\mu_n h)^3 + \mathcal{O}((\mu_n h)^4), \ \tau \text{ real}.$$

Thus the eigenvalues of the discrete problem are purely imaginary, provided $|sh| \ll 1$.

Now we consider the case that $|sh| \geq \tilde{\delta} > 0$. The characteristic equation (2.9) implies

$$|\kappa - 1|^2 = |sh|^2|\kappa| \geq \tilde{\delta}^2|\kappa|.$$

Hence, when $|\kappa| \geq 1/2$,

$$|\kappa - 1| \geq \tilde{\delta}/\sqrt{2} = \delta > 0.$$

Furthermore, when $|\kappa| \leq 1/2$, the triangle inequality gives $|1 - \kappa| \geq 1 - |\kappa| \geq 1/2$. Thus κ cannot be arbitrarily close to 1 when $|sh| \geq \tilde{\delta} > 0$.

In the following, we use the representation

$$\varphi(x_\nu) = \sigma_1 \kappa^{+\nu} + \sigma_2 \kappa^{-\nu}, \quad |\kappa| \geq 1.$$

The discrete eigenvalue problem (3.5) has a nontrivial solution if and only if

$$\sigma_1 P_h(\kappa - 1) + \sigma_2 P_h(\kappa^{-1} - 1) = 0, \qquad \sigma_1 \kappa^N + \sigma_2 \kappa^{-N} = 0,$$

i.e.,

$$\kappa^N P_h(\kappa^{-1} - 1) - \kappa^{-N} P_h(\kappa - 1) = 0, \tag{3.10}$$

has a nontrivial solution. Here

$$P_h(y) = y - \frac{1}{2}y^2 + \frac{1}{3}y^3 \equiv y(y - y_2)(y - y_3), \quad y_{2,3} = \frac{3}{4} \pm i\sqrt{3 - \frac{9}{16}}. \tag{3.11}$$

LEMMA 3.1. *There is a constant $C > 0$ such that* (3.10) *has no solution for*

$$|\kappa| \geq e^{Ch}.$$

Proof. Assume that $|\kappa| = e^{Ch}$. By (3.11), the zeros of $P_h(y)$ are $y = 0$ and $y = y_{2,3}$ with $\operatorname{Re} y_{2,3} = 3/4$. For $|\kappa - 1| \geq \delta$ and $|\kappa| \geq 1$, κ^{-1} is inside the unit circle but bounded away from 1. Therefore, $\kappa^{-1} - 1$ is inside a unit circle centered at -1 but bounded away from zero. There are no zeros of P_h in this region, and we have

$$\min_{|\kappa|\geq 1,\, |\kappa - 1| \geq \delta} |P_h(\kappa^{-1} - 1)| \geq d > 0.$$

Since $|\kappa|^N = e^C$, and $\kappa^{-3} P_h(\kappa - 1) \leq \text{const}$ for $|\kappa| \geq 1$,

$$|\kappa^{-N} P_h(\kappa - 1)| = |\kappa^{-N+3}|\, |\kappa^{-3} P_h(\kappa - 1)| \leq \text{const}\ e^{-C},$$
$$|\kappa^N P_h(\kappa^{-1} - 1)| \geq d e^C.$$

Hence, (3.10) has no solution if C is sufficiently large, and the lemma follows. ∎

We can now prove the following theorem.

THEOREM 3.2. *For sufficiently small h, all eigenvalues s of* (3.5) *are purely imaginary and the discrete problem* (3.2), (3.3) *is stable.*

Proof. We have already shown that all eigenvalues with $|sh| \ll 1$ are purely imaginary. For $|sh| \geq \tilde{\delta}$, the eigenvalue problem (3.5) has a solution if and only if (3.10) has a solution with $|\kappa| \geq 1$ and $|\kappa - 1| \geq \delta$. We can write (3.10) in the form

$$Q(\kappa) =: \frac{P_h(\kappa^{-1} - 1)}{P_h(\kappa - 1)} = \kappa^{-2N}. \tag{3.12}$$

By noting that $\overline{P_h(y)} = P_h(\overline{y})$ and that $\overline{e^{i\xi} - 1} = e^{-i\xi} - 1$, it is easy to see that

$$|Q(e^{i\xi})| = 1, \quad |(e^{i\xi})^{-2N}| = 1.$$

Now consider $\kappa = e^{i\xi + \eta h}$, $0 \leq \eta \leq C$. Then,

$$|Q(e^{i\xi+\eta h})| = 1 + \mathcal{O}(\eta h), \quad \text{but } |(e^{i\xi+\eta h})^{-2N}| = e^{-2\eta}. \tag{3.13}$$

For sufficiently small h, (3.12) can only have solutions for $\eta = 0$ since only the left-hand side of (3.13) scales with h. Lemma 3.1 tells us that (3.10) has no solution for $|\kappa| \geq e^{Ch}$, and we conclude that all solutions of (3.10) must have $|\kappa| = 1$. By solving the characteristic equation (2.9) for s and setting $\kappa = e^{i\xi}$, we get (2.14) which shows that all eigenvalues are purely imaginary.

Since we can represent the solution of the discrete problem (3.2), (3.3) in an eigenfunction expansion where all eigenvalues are purely imaginary, there can be no exponentially growing solutions, and we conclude that the discrete problem is stable. ∎

4. A continuous two-dimensional model problem. We now start our discussion of two-dimensional problems. The results from the one-dimensional model seem to indicate that we need only to avoid boundary instabilities. However, there are also highly oscillatory instabilities which can be controlled by small amounts of dissipation. As will be demonstrated in section 5, our embedded boundary approximation of the Neumann condition in general two-dimensional domains introduces truncation errors in both the tangential and normal directions. To illustrate the type of instabilities that the tangential terms can give, we study the solutions of the wave equation with perturbed Neumann conditions. We start with the half-plane problem

$$\begin{aligned} u_{tt} &= u_{xx} + u_{yy}, \quad 0 \le x < \infty,\ -\infty < y < \infty,\ t \ge 0, \\ u_x(0, y, t) &= \epsilon u_y(0, y, t), \end{aligned} \tag{4.1}$$

where ϵ is a real parameter. It turns out the size of ϵ is of minor importance, and we will for simplicity consider the case $\epsilon = 1$. Corresponding to section 2, the problem is unstable if we can find exponentially growing solutions of the type

$$u = e^{st + i\omega y}\varphi(x), \quad \operatorname{Re} s > 0,\ |\varphi(x)| \le \text{const},\ \omega \text{ real}. \tag{4.2}$$

Introducing (4.2) into (4.1) gives us the eigenvalue problem for s:

$$\begin{aligned} \varphi_{xx} &= (s^2 + \omega^2)\varphi, \\ \varphi_x(0) &= i\omega\varphi(0), \quad |\varphi(x)| \le \text{const}. \end{aligned} \tag{4.3}$$

Since (4.3) is a differential equation with constant coefficients, its general solution is

$$\varphi(x) = \sigma_1 e^{\lambda x} + \sigma_2 e^{-\lambda x}, \quad \lambda = \sqrt{s^2 + \omega^2},\ \operatorname{Re}\lambda \ge 0. \tag{4.4}$$

Clearly, $\operatorname{Re}\lambda > 0$ for $\operatorname{Re} s > 0$. Therefore, $|\varphi(x)| \le \text{const}$ if and only if $\sigma_1 = 0$, i.e.,

$$\varphi(x) = \sigma_2 e^{-\lambda x}, \quad \operatorname{Re}\lambda > 0. \tag{4.5}$$

Introducing (4.5) into the boundary condition gives us

$$-\lambda = i\omega.$$

Since $\operatorname{Re}\lambda > 0$, there are no solutions of type (4.2). However, let $s = i\sqrt{2}\,\omega + \eta$, $\eta > 0$. Solving (4.4) for λ gives

$$\lim_{\eta \to 0} \lambda = i|\omega|.$$

Thus, for $\omega < 0$, there is a solution of type (4.2) but with $\operatorname{Re} s = 0$,

$$u = e^{i\sqrt{2}\,\omega t - i|\omega|(x+y)}. \tag{4.6}$$

There is no exponential growth in time and, for large ω, the solutions are highly oscillatory in space. Furthermore, there is no decay in the x-direction. Hence, $s = i\sqrt{2}\omega$ is a generalized eigenvalue (see [5] for a definition) which forecasts instabilities for the corresponding problem on a bounded domain.

To demonstrate these instabilities, we next consider the strip problem,

$$\begin{aligned} u_{tt} &= u_{xx} + u_{yy}, \quad 0 \le x \le 1,\ -\infty < y < \infty,\ t \ge 0, \\ u_x(0,y,t) &= u_y(0,y,t), \qquad u_x(1,t) = 0. \end{aligned} \tag{4.7}$$

Again, we construct solutions of the type (4.2). Instead of (4.3), we now obtain the eigenvalue problem

$$\begin{aligned} \varphi_{xx} &= (s^2 + \omega^2)\varphi, \\ \varphi_x(0) &= i\omega\varphi(0), \quad \varphi_x(1) = 0. \end{aligned} \tag{4.8}$$

Introducing the general solution (4.4) into the boundary conditions shows that (4.8) has a solution if

$$\frac{\lambda - i\omega}{\lambda + i\omega} = e^{2\lambda}. \tag{4.9}$$

THEOREM 4.1. *The strip problem* (4.7) *is unstable. For large* $|\omega|$ *there are solutions of the type* (4.2) *with*

$$\operatorname{Re} s \approx \frac{1}{\sqrt{8}} \log(2|\omega|), \quad \textit{i.e.}, \quad e^{(\operatorname{Re} s)t} = (2|\omega|)^{t/\sqrt{8}}. \tag{4.10}$$

Proof. Let

$$\lambda = \lambda_r + i\lambda_i,$$

and assume that

$$\lambda_i = -\omega, \quad |\omega| \gg 1.$$

By (4.9),

$$\frac{\lambda_r - 2i\omega}{\lambda_r} = e^{2\lambda_r} e^{-2i\omega}. \tag{4.11}$$

Take $|\omega|$ large and $\arg\omega$ such that the arguments of the left- and right-hand sides of (4.11) match. The modulus matches if

$$\frac{\lambda_r^2 + 4\omega^2}{\lambda_r^2} = e^{4\lambda_r},$$

i.e., to the highest order in ω,

$$\lambda_r \approx \frac{1}{2} \log(2|\omega|).$$

Thus,

$$s = \pm\sqrt{-\omega^2 + \lambda^2} \approx \pm\sqrt{-2\omega^2 - i\omega\, \log(2|\omega|)},$$

and (4.10) follows. ∎

The above example shows that the stability of the left and right half-plane problems is not enough to ensure stability for the strip problem. The reason is that the generalized eigenfunctions (4.6) do not decay in space but are reflected back and forth between the boundaries at $x = 0, 1$, respectively. Every time they hit the left boundary they are amplified. These are highly oscillatory instabilities, and we will see that they can easily be controlled by small amounts of dissipation. This example also illustrates that nondissipative difference methods of our type are prone to weak instabilities, i.e., instabilities that grow only algebraically in time (see (4.10)). Note that a weak instability also occurs if the tangential derivative in (4.1) is replaced by a higher order, odd, tangential derivative.

To demonstrate a strong instability, we study the half-plane problem where the boundary condition in (4.1) is replaced by

$$u_x = \beta u_{yy}, \quad x = 0, \tag{4.12}$$

where β is a constant. As before, we look for solutions of the type (4.2), and using the same arguments as above, we know the solution must have the form (4.5). Inserting this ansatz into the boundary condition (4.12) gives

$$-\lambda = -\beta\omega^2. \tag{4.13}$$

Since $\operatorname{Re}\lambda > 0$, there are no solutions with $\operatorname{Re} s > 0$ when $\beta < 0$. Next we investigate if there are any generalized eigenvalues. Setting $s = i\tau$ yields $\lambda = \sqrt{-\tau^2 + \omega^2}$, so $-\lambda$ is either real and negative or purely imaginary. When $\beta < 0$, the right-hand side of (4.13) is always real and positive, and we conclude that there are not any generalized eigenvalues either. Hence, the case $\beta < 0$ is stable.

When $\beta > 0$ and ω is large, (4.13) is solved by

$$s \approx \beta\omega^2,$$

and inserting (4.13) into (4.5) gives

$$u = e^{\beta\omega^2 t - \beta\omega^2 x + i\omega y}, \quad \omega \text{ large}, \ \beta > 0.$$

Hence, these solutions have a thin boundary layer in space and grow exponentially in time. This is a strong instability. As we shall see in section 5, this type of instability can only be controlled by dissipation when the coefficient β is small. Perturbing the Neumann condition by a higher order, even, tangential derivative results in the same behavior; i.e., the stability depends on the sign of the coefficient.

5. The discrete half-plane problem in two dimensions. We consider next the two-dimensional half-plane problem for

$$u_{tt} = u_{xx} + u_{yy}, \quad 0 \le x < \infty, \ -\infty < y < \infty, \ t \ge 0, \tag{5.1}$$

with the boundary condition

$$u_x(0, y, t) = 0, \quad |u(x, y, t)| \le \text{const}, \tag{5.2}$$

and approximate it by

$$v_{tt} = (D_{+x}D_{-x} + D_{+y}D_{-y})v, \tag{5.3}$$

with the third order accurate boundary condition (2.5)

$$L_h v(0,y,t) = 0, \quad |v(x,y,t)| \le \text{const.} \tag{5.4}$$

Here v is a discrete function varying on a grid

$$\{x_\nu = \nu h,\ y_\mu = \mu h\}, \quad \nu = 0,1,2,\dots,\ \mu = 0,\pm 1,\pm 2,\dots.$$

We Fourier-transform the difference equation with respect to y and obtain

$$\begin{aligned} &\hat{v}_{tt} = \left(D_{+x}D_{-x} - \frac{4}{h^2}\sin^2(\omega h/2)\right)\hat{v}, \\ &L_h\hat{v}(0,\omega,t) = 0, \quad |\hat{v}(x,\omega,t)| \le \text{const.} \end{aligned} \tag{5.5}$$

Thus, we obtain a one-dimensional problem for every fixed ω and can apply mode analysis as before. Then

$$\hat{v}(x_\nu,t) = e^{st}\varphi(x_\nu), \quad \operatorname{Re} s > 0,$$

is a solution of (5.5) if there are solutions $\varphi(x_\nu)$ of the eigenvalue problem

$$\left(s^2 + \frac{4}{h^2}\sin^2(\omega h/2)\right)\varphi = D_{+x}D_{-x}\varphi, \tag{5.6}$$

$$L_h\varphi = 0, \quad |\varphi(x)| \le \text{const}, \tag{5.7}$$

with $\operatorname{Re} s > 0$. The eigenvalue problem (5.6), (5.7) is of the same type as for the one-dimensional half-plane problem in section 2. In particular, the general solution has the form (2.8), where κ now is a solution of the two-dimensional characteristic equation

$$(\kappa - 1)^2 - (s^2h^2 + 4\sin^2(\omega h/2))\kappa = 0. \tag{5.8}$$

It is straightforward to show that this characteristic equation has the same essential properties as in the one-dimensional case. To be precise, we have the following lemma.

LEMMA 5.1. *For $|hs| \ll 1$ and $|h\omega| \ll 1$, the roots of* (5.8) *are of the form*

$$\begin{aligned} \kappa &= 1 - h\lambda + \frac{h^2\lambda^2}{2} + \mathcal{O}(h^3\lambda^3) = e^{-h\lambda(1+\mathcal{O}(h^2\lambda^2))}, \\ \kappa^{-1} &= 1 + h\lambda + \frac{h^2\lambda^2}{2} + \mathcal{O}(h^3\lambda^3) = e^{h\lambda(1+\mathcal{O}(h^2\lambda^2))}, \end{aligned}$$

where $\lambda = \sqrt{s^2+\omega^2}$, $\operatorname{Re}\lambda > 0$, for $\operatorname{Re} s > 0$. Also, for each fixed ω and for $\operatorname{Re} s > 0$, (5.8) *has no root with $|\kappa| = 1$ and exactly one root κ with $|\kappa| < 1$.*

Proof. The proof follows by straightforward generalization of Lemma 2.1. ∎

Since the boundary conditions are the same as in the one-dimensional case, we can use the same arguments as in section 2 to show that there are no solutions of (5.6), (5.7) for $\operatorname{Re} s > 0$, which implies that there are no exponentially growing solutions of (5.5). Note that for the Neumann boundary condition approximation (5.4), the boundary normal is aligned with the x-direction.

Our goal is to construct stable difference approximations for general domains. In this case the boundary condition for the differential equation is

$$\partial u/\partial n = 0, \quad \partial/\partial n: \text{ derivative normal to the boundary}, \tag{5.9}$$

and in general the normal is not aligned with the mesh. In the following, we will study the continuous boundary conditions perturbed by the leading truncation error terms. In the literature this technique is often used for the Cauchy or spatially periodic problems, and the truncation terms appear only in the differential equation. The obtained equation is often called the "modified equation"; see, for example, [14] or [11]. Here we use the technique to analyze the influence of truncation errors in the boundary conditions. The modified equation is a more accurate description of the discretized problem than the continuous problem. Or rephrased, the numerical solution approximates the modified equation to a higher order of accuracy than the continuous problem. However, the modified equation can only model low and intermediate frequencies in the discrete solution, and we rely on the dissipation to control the highest frequencies.

In section 4 we have discussed half-plane and strip problems. The reason is this: For analytic initial boundary value problems where there are no direct energy estimates, the study of wellposedness can be reduced to the study of half-plane and strip problems. This is done in the following way. In the neighborhood of every boundary point P with tangent T_g we use a locally smooth map to transform the curved boundary locally onto T_g. Then we study the half-plane problem with T_g as the boundary. After freezing the variable coefficients we can solve the problem by Fourier–Laplace transform. If for all these half-plane problems there are no eigenvalues or generalized eigenvalues s with $\operatorname{Re} s \geq 0$, then the original problem is well posed; see Kreiss and Lorenz [8].

We shall now apply this technique to analyze the stability of the discrete problem. Let the angle θ between the outward normal and the x-axis be defined as in Figure 1. We consider the differential equation on the half-plane $\boldsymbol{n} \cdot \boldsymbol{x} \leq 0$, i.e.,

$$x \cos\theta + y \sin\theta \geq 0. \tag{5.10}$$

To be able to calculate the truncation error of the discrete boundary condition we assume that the solution u of the differential equation is smooth and decays rapidly to zero for $x^2 + y^2 \to \infty$, in the half-plane (5.10). Also, we extend it smoothly beyond the boundary such that the extended u decays rapidly to zero for $x^2 + y^2 \to \infty$ in the whole plane.

The truncation error in the third order Neumann boundary condition satisfies ($\pi/4 \leq \theta \leq \pi/2$)

$$D_n^{(3)} u(x_i, y_j) = \frac{\partial u}{\partial n}(\boldsymbol{x}_{i,j}^{\Gamma}) + C_1 h^4 \frac{\partial^5 u}{\partial n^5}(\boldsymbol{x}_{i,j}^{\Gamma}) + C_2 h^3 \frac{\partial^4 u}{\partial n^4}(\boldsymbol{x}_{i,j}^{\Gamma}) + h^4 R_1 + h^3 R_2 + O(h^5).$$

Here,

$$R_1 = \sum_{\nu=1}^{3} C_{1\nu} \frac{\partial^5 u(\tilde{x}_\nu, y_{j+\nu})}{\partial x^5}, \quad R_2 = \sum_{\nu=1}^{3} C_{2\nu} \frac{\partial^4 u(\tilde{x}_\nu, y_{j+\nu})}{\partial x^4}.$$

The terms in R_1, R_2 originate from interpolation errors in v_I, v_{II}, and v_{III}, respectively.

Derivatives with respect to x and y can be related to normal ($\partial/\partial n$) and tangential ($\partial/\partial \sigma$) derivatives. We have

$$\frac{\partial}{\partial x} = -\sin\theta \frac{\partial}{\partial \sigma} - \cos\theta \frac{\partial}{\partial n}, \qquad \frac{\partial}{\partial y} = \cos\theta \frac{\partial}{\partial \sigma} - \sin\theta \frac{\partial}{\partial n}.$$

We can also use Taylor expansions to express derivatives at $(\tilde{x}_\nu, y_{j+\nu})$ in terms of derivatives at the boundary point x_{ij}^Γ. After some calculations we obtain

$$D_n^{(3)} u(x_i, y_j) = (1+R)\frac{\partial u}{\partial n} - \left(h^4 \beta_1 \frac{\partial^5 u}{\partial \sigma^5} + h^3 \beta_2 \frac{\partial^4 u}{\partial \sigma^4} \right) + \mathcal{O}\left(h^5 \frac{\partial^6 u}{\partial n^{6-j} \partial \sigma^j} \right).$$

Here R is an operator of the form

$$R = \sum_{p+q \geq 3} \beta_{pq} h^{p+q} \frac{\partial^{p+q}}{\partial n^p \partial \sigma^q}.$$

We can write the half-plane problem for the differential equation in the form

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial n^2} + \frac{\partial^2 u}{\partial \sigma^2}, \quad n \geq 0, \ -\infty < \sigma < \infty,$$
$$\frac{\partial u}{\partial n} = 0 \quad \text{for } n = 0.$$

After Fourier-transforming with respect to σ and Laplace-transforming with respect to t, we obtain

$$\frac{\partial^2 \hat{u}}{\partial n^2} = (s^2 + \omega^2)\hat{u}.$$

Thus,

$$\hat{u} = e^{-\sqrt{s^2+\omega^2}\, n} u_0(s, \omega), \quad \frac{\partial \hat{u}}{\partial n} = -\sqrt{s^2+\omega^2}\hat{u},$$

and the Fourier–Laplace transform of $\partial/\partial n$ is $-\sqrt{s^2+\omega^2}$. After freezing the coefficients, we Fourier–Laplace-transform the truncation error and obtain

$$\hat{D}_n^{(3)} = -(1+\hat{R})\sqrt{s^2+\omega^2} - (i\beta_1 h^4 \omega^5 + \beta_2 h^3 \omega^4) + \mathcal{O}((|\omega| + |s|)^6 h^5).$$

Here,

$$\hat{R} = \sum_{p+q \geq 3} \beta_{pq} (-\sqrt{(hs)^2 + (h\omega)^2})^p (ih\omega)^q = \mathcal{O}((|hs| + |h\omega|)^3).$$

For $|hs| + |h\omega|$ sufficiently small, $\|\hat{R}\| \leq 1/2$, and we can write

$$\begin{aligned} \hat{D}_n^{(3)} &= (1+\hat{R}) \left(-\sqrt{s^2+\omega^2} - \frac{i\beta_1 h^4 \omega^5 + \beta_2 h^3 \omega^4}{1+\hat{R}} \right) + \mathcal{O}((|\omega| + |s|)^6 h^5) \\ &= (1+\hat{R})(-\sqrt{s^2+\omega^2} - (i\beta_1 h^4 \omega^5 + \beta_2 h^3 \omega^4)) + \mathcal{O}((|\omega| + |s|)^6 h^5). \end{aligned}$$

By neglecting the $\mathcal{O}(h^5)$ term and transforming back to physical space, the boundary condition $D_n^{(3)} u = 0$ corresponds to

$$(1+R)\left(\frac{\partial u}{\partial n} - \beta_1 h^4 \frac{\partial^5 u}{\partial \sigma^5} - \beta_2 h^3 \frac{\partial^4 u}{\partial \sigma^4} \right) = 0.$$

By assumption, $\|R\|$ is small and the boundary condition can only be satisfied if the term following $(1+R)$ is zero. After changing spatial variables, $n \to x$ and $\sigma \to y$,

we arrive at the modified equation model corresponding to the discrete half-plane problem (5.3)–(5.4) in domains where the normal is not aligned with the mesh:

$$u_{tt} = u_{xx} + u_{yy} - \alpha h^3 u_{tyyyy}, \quad \alpha \geq 0,\ x \geq 0,\ -\infty < y < \infty, \tag{5.11}$$

$$u_x = \beta_1 h^4 u_{yyyyyy} + \beta_2 h^3 u_{yyyy}, \quad x = 0,\ |u| \leq \text{const}. \tag{5.12}$$

We have added a dissipation term to the differential equation because we shall need it later. Note that we have only added dissipation in the tangential direction, to avoid having to add any extra boundary conditions.

After Fourier-transforming in y and Laplace-transforming in t, we obtain

$$\hat{u}_{xx} = (s^2 + \omega^2)\hat{u} + sa\hat{u}, \quad a = \alpha h^3 \omega^4,\ \operatorname{Re} s > 0, \tag{5.13}$$

with boundary conditions

$$\hat{u}_x(0) = b\hat{u}(0), \quad |\hat{u}| \leq \text{const},\ b = i\beta_1 \omega^5 h^4 + \beta_2 \omega^4 h^3. \tag{5.14}$$

As $|\omega|$ gets larger, the Fourier symbol of the second divided difference, $-\frac{4}{h^2}\sin^2(\frac{\omega h}{2})$, deviates more and more from the Fourier symbol of a second derivative, $-\omega^2$. In particular, for the highest frequency on the mesh ($\omega h = \pi$), the symbol of the second divided difference is $-4/h^2$, while the symbol of the second derivative is $-\pi^2/h^2$. Hence, the highest frequencies on the mesh are not accurately modeled by the modified equation. We therefore restrict the following analysis to $|\omega h| \leq 1$.

The general solution of (5.13) is given by

$$\hat{u} = \sigma_1 e^{\lambda x} + \sigma_2 e^{-\lambda x}, \tag{5.15}$$

where λ now satisfies

$$\lambda = \sqrt{s^2 + \omega^2 + sa}, \quad \operatorname{Re}\lambda \geq 0. \tag{5.16}$$

Since $\operatorname{Re}\lambda > 0$ for $\operatorname{Re} s > 0$, the boundary conditions are satisfied if and only if

$$\sigma_1 = 0, \qquad \lambda = -b = -(i\beta_1 \omega^5 h^4 + \beta_2 \omega^4 h^3). \tag{5.17}$$

There are two possibilities:

1. If $\beta_2 \geq 0$, then there are no solutions of (5.17) with $\operatorname{Re} s > 0$ since $\operatorname{Re}\lambda > 0$, but $\operatorname{Re}(-b) \leq 0$. Furthermore, when the dissipation coefficient $\alpha > 0$, $\operatorname{Re}\lambda > 0$ also for $\operatorname{Re} s = 0$, $\omega \neq 0$. Hence, there are no generalized eigenvalues when $\alpha > 0$, and we conclude that the half-plane problem is stable.

2. If $\beta_2 < 0$, then the problem can be unstable. We want to show that if $|\beta_1|$ and $|\beta_2|$ are small, the problem can be stabilized by a small $\alpha > 0$.

THEOREM 5.2. *If $\beta_2 < 0$, $|\beta_2| \ll 1$, $|\beta_1| \ll 1$, and $\alpha \geq K|\beta_1\beta_2|$, $K = \text{const}$, the modified half-plane problem* (5.11)–(5.12) *is stable; i.e., the eigenvalue problem* (5.13)–(5.14) *has no solutions with* $\operatorname{Re} s > 0$ *and no generalized eigenvalues* $\operatorname{Re} s = 0$ *for* $\omega \neq 0$.

Proof. Introducing (5.17) into (5.16) gives

$$s^2 + sa + \omega^2 - b^2 = 0.$$

Since $|\omega h| \leq 1$, $|a| \leq |\alpha||\omega|$. If we assume $0 \leq \alpha \leq 1$, we have $|a| \leq |\omega|$ and $\omega^2 - a^2/4 \geq 3\omega^2/4$. Therefore,

$$s = -\frac{a}{2} \pm i\sqrt{\left(\omega^2 - \frac{a^2}{4}\right)}\sqrt{1 - \frac{b^2}{\omega^2 - a^2/4}}. \tag{5.18}$$

By assumption $|b^2| \ll \omega^2$. We can therefore expand the square root and conclude that

$$s = -\frac{a}{2} \pm \left(i\sqrt{\left(\omega^2 - \frac{a^2}{4}\right)} - \frac{ib^2}{2\sqrt{\omega^2 - a^2/4}} + \cdots \right).$$

Hence,

$$\operatorname{Re} s \approx -\frac{\alpha}{2} h^3 \omega^4 \pm \frac{\operatorname{Im} b^2}{2\sqrt{\omega^2 - a^2/4}} = -\frac{\alpha}{2} h^3 \omega^4 \pm \frac{\beta_1 \beta_2 (\omega h)^7 \omega^2}{\sqrt{\omega^2 - a^2/4}} < 0$$

for $\alpha \geq 4|\beta_1 \beta_2|/\sqrt{3}$. ∎

We have numerically computed the truncation error coefficients β_1 and β_2 for our boundary condition approximation. To conserve space, we will only report the result of these computations here. For all possible directions of the boundary normal and all permissible distances between the ghost point and the boundary, we found that $-0.065 < \beta_2 < 0.015$ and $-0.063 < \beta_1 < 0.063$. It is critical that $|\beta_2|$ is small since the case $\beta_2 < 0$, $|\beta_2| = \mathcal{O}(1)$ cannot be stabilized by adding a dissipative term to the differential equation. In earlier versions of our numerical code we added a tangential smoothing operator to the boundary condition approximation. In terms of the modified problem this means that $\beta_2 > 0$. The dissipation operator proposed in section 1 seems to be so efficient that this extra smoothing operator is not needed.

6. The two-dimensional strip problem. Here we generalize the modified equation approach to study the stability of solutions on a bounded domain,

$$u_{tt} = u_{xx} + u_{yy} - \alpha h^3 u_{tyyyy}, \quad \alpha \geq 0,\ 0 \leq x \leq 1,\ -\infty < y < \infty, \tag{6.1}$$

$$u_x = \beta_1 h^4 u_{yyyyy} + \beta_2 h^3 u_{yyyy}, \quad x = 0,\ u_x = 0,\ x = 1. \tag{6.2}$$

Remark. In reality the boundary condition at $x = 1$ also contains truncation order terms, but the results are the same.

After Fourier- and Laplace-transforming the problem, (6.1)–(6.2) becomes

$$s^2 \hat{u} = \hat{u}_{xx} - (\omega^2 + as)\hat{u}, \quad a = \alpha h^3 \omega^4, \tag{6.3}$$

$$\hat{u}_x(0) = b\hat{u}(0), \quad \hat{u}_x(1) = 0, \quad b = i\beta_1 \omega^5 h^4 + \beta_2 \omega^4 h^3. \tag{6.4}$$

The general solution of (6.3) now has the form

$$\hat{u} = \sigma_1 e^{\lambda x} + \sigma_2 e^{-\lambda x}, \tag{6.5}$$

where λ is the solution of (5.16), i.e., $\operatorname{Re} \lambda > 0$ for $\operatorname{Re} s > 0$. Introducing (6.5) into (6.4) shows that there is a nontrivial solution if and only if

$$\frac{\lambda - b}{\lambda + b} = e^{2\lambda}. \tag{6.6}$$

We have already studied the corresponding half-plane problem and shown that for $\beta_2 \geq 0$, there are no eigenvalues s, with $\operatorname{Re} s > 0$, and that there are no generalized eigenvalues when $\alpha > 0$. For $\beta_2 < 0$, Theorem 5.2 shows when the half-plane problem is stable. Hence, it can be expected that the strip problem also is stable. In Appendix B, we perform a detailed calculation to verify the stability of the strip problem. From this calculation, we can also read off the order of magnitude of the dissipation coefficient α that is necessary for stability. The results are summarized in the following theorem.

THEOREM 6.1. *If the half-plane problem* (5.11)–(5.12) *is stable, the modified strip problem* (6.1)–(6.2) *is stable for* $\alpha > 0$, $\alpha = \mathcal{O}(h^{3/4})$.

7. General two-dimensional domains. In this section, we will add some details to our proposed scheme that were left out of the general description in section 1.

7.1. Near boundary behavior of the discretized Laplacian. The discretized Neumann boundary condition (1.3) can be used to eliminate all ghost point values in the discretized Laplacian (1.2). Referring to the case shown in Figure 1, we get at the point $(i, j+1)$,

$$\begin{aligned}\Delta_h v^n_{i,j+1} = {} & \frac{1}{h^2}(v^n_{i+1,j+1} + v^n_{i-1,j+1} + v^n_{i,j+2} - 4v^n_{i,j+1}) \\ & - \frac{g_I}{h^2 g_0}(c_0 v^n_{i,j+1} + c_1 v^n_{i+1,j+1} + c_2 v^n_{i+2,j+1}) \\ & - \frac{g_{II}}{h^2 g_0}(c_3 v^n_{i,j+2} + c_4 v^n_{i+1,j+2} + c_5 v^n_{i+2,j+2}) + \frac{f(\boldsymbol{x}^\Gamma_{i,j}, t_n)}{h^2 g_0},\end{aligned} \tag{7.1}$$

assuming that (i, j) is the only nearest neighbor of $(i, j+1)$ that is outside of Ω. If additional points are outside, other formulas of the type (1.3) would be used to eliminate those points as well. The coefficients g_0, g_I, g_{II} are given by (1.4). Since $0 \le \xi_\Gamma \le \xi_I$ and $h \le \xi_I \le \sqrt{2}h$, the denominator g_0 satisfies

$$\frac{1}{2\sqrt{2}h} \le \frac{1}{2\xi_I} \le |g_0| \le \frac{3}{2\xi_I} \le \frac{3}{2h}.$$

Because the coefficient g_0 in (7.1) is bounded away from zero, we conclude that this discretization of the Laplacian does *not* suffer from the "small cell" stiffness problem.

7.2. Accuracy of the damped scheme. For simplicity, let the grid function $\boldsymbol{v}$ satisfy the semidiscrete problem, where time is left continuous,

$$\boldsymbol{v}_{tt} = A\boldsymbol{v} + \boldsymbol{b} + \boldsymbol{F} - \alpha h^3 A^T(A\boldsymbol{v}_t + \boldsymbol{b}_t).$$

Let the error in the discrete solution be $\boldsymbol{e} = u - \boldsymbol{v}$, where u is the solution of the continuous problem (1.1) evaluated on the grid. We have

$$\begin{aligned}\boldsymbol{e}_{tt} &= \Delta u - A\boldsymbol{v} - \boldsymbol{b} + \alpha h^3 A^T(A(\boldsymbol{v}_t + u_t - u_t) + \boldsymbol{b}_t) \\ &= \Delta u - Au - \boldsymbol{b} + A\boldsymbol{e} - \alpha h^3 A^T A\boldsymbol{e}_t + \alpha h^3 A^T(Au_t + \boldsymbol{b}_t).\end{aligned}$$

We split the error according to $\boldsymbol{e} = \boldsymbol{e}^I + \boldsymbol{e}^{II}$ and let $\boldsymbol{e}^I$ satisfy

$$A\boldsymbol{e}^I = -\alpha h^3 A^T(Au_t + \boldsymbol{b}_t). \tag{7.2}$$

Now, $Au_t + \boldsymbol{b}_t$ is a second order accurate approximation of Δu_t evaluated on the grid. Furthermore, away from the boundary, $A^T\Delta u_t$ is a second order approximation of $\Delta^2 u_t$, but near the boundary $A^T\Delta u_t = \mathcal{O}(\Delta u_t/h^2)$. Hence the right-hand side of (7.2) is $\mathcal{O}(h)$ near the boundary but $\mathcal{O}(h^3)$ in the interior. Due to the smoothing properties of the elliptic operator A (see Figure 2 and Table 1 for a numerical example), we gain one order of magnitude when solving for $\boldsymbol{e}^I$, resulting in

$$\boldsymbol{e}^I = \mathcal{O}(h^2).$$

Since the right-hand side of (7.2) is smooth in time, we also have $\boldsymbol{e}^I_{tt} = \mathcal{O}(h^2)$ and $A\boldsymbol{e}^I_t = \mathcal{O}(h)$. The equation for $\boldsymbol{e}^{II}$ is

$$\boldsymbol{e}^{II}_{tt} = A\boldsymbol{e}^{II} - \alpha h^3 A^T A\boldsymbol{e}^{II}_t - \boldsymbol{e}^I_{tt} - \alpha h^3 A^T A\boldsymbol{e}^I_t + \Delta u - (Au + \boldsymbol{b}).$$

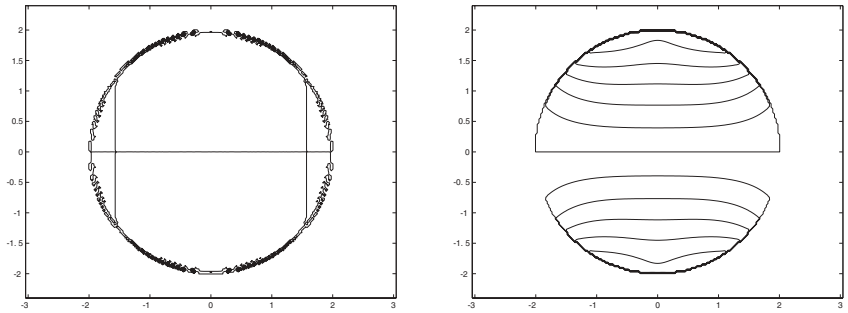


FIG. 2. *Numerical test of the smoothing properties of* $A\boldsymbol{e}^I = h^3 A^T \boldsymbol{u}$. *The left figure shows a contour plot of the right-hand side* $h^3 A^T \boldsymbol{u}$, *and the right figure shows the solution* $\boldsymbol{e}^I$ *for* $N = 171$; *see Table* 1 *for quantitative information. In this case, the computational domain was a circle with unit radius and the test function was* $\boldsymbol{u}_{i,j} = \cos(x_i)\sin(y_j)$. *The problem was solved using the conjugated gradient algorithm.*

TABLE 1

Smoothing properties of the operator A investigated by solving $A\boldsymbol{e}^I = h^3 A^T \boldsymbol{u}$ *for different grid sizes for the case shown in Figure* 2. *Clearly,* $\boldsymbol{e}^I = \mathcal{O}(h^2)$ *while* $h^3 A^T \boldsymbol{u} = \mathcal{O}(h)$.

N	$\|\boldsymbol{e}^I\|_\infty$	$\|h^3 A^T \boldsymbol{u}\|_\infty$	h
171	3.43×10^{-4}	4.36×10^{-2}	2.82×10^{-2}
341	9.21×10^{-5}	2.20×10^{-2}	1.41×10^{-2}
681	2.40×10^{-5}	1.20×10^{-2}	7.06×10^{-3}

Because $Au + \boldsymbol{b}$ is a second order accurate approximation of Δu and $h^3 A^T A \boldsymbol{e}_t^I = \mathcal{O}(h^2)$, all forcing terms are of the order $\mathcal{O}(h^2)$. Hence

$$\boldsymbol{e}^{II} = \mathcal{O}(h^2),$$

which shows that the damped scheme is second order accurate.

7.3. Fourth order corrections. To reduce the phase-error away from the boundary, we can optionally add a fourth order correction term,

$$\Delta_{h,4} v_{i,j}^n = -\frac{h^2}{12}(D_+^x D_-^x \gamma_{i,j} D_+^x D_-^x + D_+^y D_-^y \gamma_{i,j} D_+^y D_-^y) v_{i,j}^n,$$

to our second order accurate approximation of the Laplacian. Clearly, this stencil is too wide to be evaluated all the way up to the boundary, so the grid function $\gamma_{i,j}$ must be identically zero in a band near the boundary. Away from the boundary we want $\gamma_{i,j} \equiv 1$ to make the correction term cancel the second order truncation error in $\Delta_h v_{i,j}$. To aid in the construction of $\gamma_{i,j}$, we initially compute a smoothed distance function $d_{i,j} \geq 0$ using the technique described in [13]. The value of the distance function at a grid point approximately equals the distance between that grid point and the nearest boundary. Hence, the distance function is zero on the boundary and increases monotonically away from the boundary, making it straightforward to construct a smooth $\gamma_{i,j}$ that is zero near the boundary ($d_{i,j} \leq \epsilon_1$) and one away from the boundary ($d_{i,j} \geq \epsilon_2$). In all numerical examples presented below, we used $\epsilon_1 = 3h$ and $\epsilon_2 = 13h$. The resulting scheme can be written in semidiscrete form as

$$\boldsymbol{v}_{tt} = A\boldsymbol{v} + B\boldsymbol{v} + \boldsymbol{F} + \boldsymbol{b} - \alpha h^3 A^T \left(A\boldsymbol{v}_t + \frac{d\boldsymbol{b}}{dt}(t_n) \right), \tag{7.3}$$

where B represents the fourth order correction term. The symmetry of B and the smoothness of the distance function of $\gamma_{i,j}$ seem to give stability. A heuristic argument for this is that B cannot generate any boundary layer instability, since this type of instability decays rapidly away from the boundary and in this region $\Delta_{h,4} v_{i,j}^n$ is arbitrarily small. And the other type of instability discussed above, highly oscillatory in the whole domain, is effectively stabilized by our damping term. The smoothness of the distance function implies that no new spurious solutions are generated. The smoothness furthermore guarantees accuracy of order two in the transition region. Hence, the resulting scheme will only be second order accurate. The main benefit of the fourth order spatial correction will be a reduced phase-error away from the boundary. For this reason we will call the resulting scheme the "internally fourth order" method.

We can also improve the basic second order time-integration method by using a fourth order accurate Taylor series method. Consider the second order system of ordinary differential equations

$$\boldsymbol{w}_{tt} = C\boldsymbol{w} + \boldsymbol{F},$$

where C is a symmetric negative semidefinite matrix. A fourth order time-discretization is given by

$$\frac{\boldsymbol{w}^{n+1} - 2\boldsymbol{w}^n + \boldsymbol{w}^{n-1}}{k^2} = C\boldsymbol{w}^n + \boldsymbol{F}^n + \frac{k^2}{12}(C(C\boldsymbol{w}^n + \boldsymbol{F}^n) + \boldsymbol{F}_{tt}^n), \tag{7.4}$$

and it is stable for

$$\max_j (-\lambda_j) k^2 < 12,$$

where λ_j are the real-valued nonpositive eigenvalues of C. The scheme (7.4) can be formulated in predictor-corrector form,

$$\tilde{\boldsymbol{w}}^{n+1} = 2\boldsymbol{w}^n - \boldsymbol{w}^{n-1} + k^2 C\boldsymbol{w}^n + k^2 \boldsymbol{F}^n, \tag{7.5}$$

$$\boldsymbol{w}^{n+1} = \tilde{\boldsymbol{w}}^{n+1} + \frac{k^2}{12}(C(\tilde{\boldsymbol{w}}^{n+1} - 2\boldsymbol{w}^n + \boldsymbol{w}^{n-1}) + k^2 \boldsymbol{F}_{tt}^n). \tag{7.6}$$

Hence, the predictor step (7.5) is simply the second order time-integration scheme presented above. The discrete damping term is added to the predictor-corrector scheme in the same way as in (1.6). For the spatially fourth order method, we take $C = A + B$; otherwise $C = A$. We note that the corrector step (7.6) needs only a second order accurate approximation of $\boldsymbol{w}_{tt}$. Hence, from an accuracy standpoint we can omit the correction term and always take $C = A$ in this step. Numerical experiments (see section 8) indicate that the resulting scheme is stable.

We start the time-integration at $n = 0$. For the fourth order time-discretization, we take $v_{i,j}^0 = u_0(x_i, y_j)$ and need to use a fifth order accurate approximation of $u(x_i, y_j, -k)$ for $v_{i,j}^{-1}$. This is achieved by using the differential equation to

approximate higher order time derivatives,

$$
\begin{aligned}
v_{i,j}^{-1} = {} & u_0(x_i,y_j) - ku_1(x_i,y_j) + \frac{k^2}{2}(D_+^xD_-^x + D_+^yD_-^y)u_0(x_i,y_j) + \frac{k^2}{2}F(\boldsymbol{x}_{i,j},0) \\
& -\frac{k^3}{6}(D_+^xD_-^x + D_+^yD_-^y)u_1(x_i,y_j) - \frac{k^3}{6}F_t(\boldsymbol{x}_{i,j},0) \\
& -\frac{k^2h^2}{24}((D_+^xD_-^x)^2 - (D_+^yD_-^y)^2)u_0(x_i,y_j) \\
& +\frac{k^4}{24}(D_+^xD_-^x + D_+^yD_-^y)^2u_0(x_i,y_j) + \frac{k^4}{24}F_{tt}(\boldsymbol{x}_{i,j},0).
\end{aligned} \tag{7.7}
$$

Note that the last three lines can be omitted for the second order time-discretization.

8. Numerical examples. In this section we numerically solve (1.1) with the schemes described above. For the cases where an analytical solution is known, we use this solution to initialize the computation at time levels $t = -k$ and $t = 0$. For the cases where an analytical solution is not known we use the initialization (7.7).

We will denote the CFL-number by $\mathrm{CFL} \equiv k/h$. Note that for a two-dimensional periodic domain, our second order time-integration scheme (1.6) is stable for CFL $\leq 1/\sqrt{2} \approx 0.71$, while the fourth order predictor-corrector scheme (7.5), (7.6) is stable for $\mathrm{CFL} \leq \sqrt{3/2} \approx 1.22$. Also note that all errors are measured in max-norm.

In all examples presented below, the fourth order predictor-corrector time-integrator (7.5), (7.6) is used together with the internally fourth order spatial correction. The second order scheme (1.6) is always used together with the second order spatial discretization. Unless otherwise noted, the Neumann boundary condition is discretized using the third order accurate formula to reduce the constant in the second order truncation error, as was mentioned in the introduction.

To evaluate the accuracy of the method, the forcing function is chosen such that the exact solution is the trigonometric traveling wave:

$$
u(x,y,t) = \sin(\omega(x-t))\sin(\omega y), \quad \omega = 4\pi. \tag{8.1}
$$

The domain Ω is taken to be an ellipse centered at the origin with semiaxes $x_s = 1$ and $y_s = 0.75$. The Cartesian grid covers the rectangle $-1.1 \leq x \leq 1.1$, $-0.85 \leq y \leq 0.85$. In Table 2, we present a grid refinement study for the second order scheme (1.6) and the internally fourth order predictor-corrector scheme (7.5), (7.6). The fourth order correction applies only in the interior of the domain, and the second order errors near the boundary clearly dominate the total error. Hence, in this case, there is no apparent benefit of using the internally fourth order method. The time step can be taken twice as large, but this gain is balanced by having to evaluate the Laplacian twice instead of once per time step. Also note that the influence of the damping term is so small that it changes only the last digit in the error in one of these runs.

To more clearly illustrate the benefits of using a fourth order correction away from the boundary, we select the forcing function F and boundary data f such that the exact solution is a spatially localized, outwardly traveling wave,

$$
u(x,y,t) = \phi(\sqrt{x^2+y^2} - t), \qquad \phi(\xi) = \frac{1}{2}\left(1 + \tanh\frac{\xi-\xi_0}{\epsilon}\right)\left(1 - \tanh\frac{\xi-\xi_1}{\epsilon}\right). \tag{8.2}
$$

Note that such waves are exact solutions to the unforced wave equation in one and three space dimensions, but not in the two-dimensional case. The domain Ω is taken

TABLE 2

Grid refinement study showing the errors in the computed solutions when the exact solution is the trigonometric function (8.1). *Here, CFL* $= 0.5$ *for the second order scheme and CFL* $= 1.0$ *for the internally fourth order predictor-corrector scheme. The grid size* $N = 101$ *corresponds to* $h = 2.4 \times 10^{-2}$ *and* $N = 201$ *corresponds to* $h = 1.2 \times 10^{-2}$. *The first line corresponds to the undamped case,* $\alpha = 0$, *and the second line shows the damped case with* $\alpha = 0.001$.

		Second order scheme			Predictor-corrector scheme		
t	α	$N = 101$	$N = 201$	ratio	$N = 101$	$N = 201$	ratio
2.0	0.0	8.75e-02	2.10e-02	4.17	10.7e-02	2.17e-02	4.93
2.0	0.001	8.77e-02	2.10e-02	4.18	10.7e-02	2.17e-02	4.93

TABLE 3

Grid refinement study showing the errors in the computed solutions when the exact solution is the outwardly traveling wave function (8.2). *Here, CFL* $= 0.5$ *for the second order scheme and CFL* $= 1.0$ *for the predictor-corrector scheme. The grid size* $N = 201$ *corresponds to* $h = 1.8 \times 10^{-2}$ *and* $N = 401$ *corresponds to* $h = 9.0 \times 10^{-3}$. *In all cases, the damping coefficient was* $\alpha = 10^{-3}$.

	Second order scheme			Predictor-corrector scheme		
t	$N = 201$	$N = 401$	ratio	$N = 201$	$N = 401$	ratio
0.5	3.29e-2	8.63e-3	3.8	1.23e-3	8.78e-5	14.0
0.75	4.59e-2	1.23e-2	3.7	1.73e-3	1.26e-4	13.7
1.0	1.05e-1	3.12e-2	3.4	2.71e-2	3.23e-3	8.4
1.25	5.89e-2	1.73e-2	3.4	1.76e-2	2.53e-3	6.9

to be the circle, $|r| \leq 1.5$, and the Cartesian grid covers the square $-1.6 \leq x \leq 1.6$, $-1.6 \leq y \leq 1.6$. The parameters in ϕ are taken to be

$$\xi_0 = 0.3, \quad \xi_1 = 0.5, \quad \epsilon = 0.07.$$

The wave reaches the boundary at $t \approx 0.8$. In Table 3 we see that for the internally fourth order method, the error is at least one order of magnitude smaller and the convergence rate is much higher before the wave hits the boundary. No such distinction can be made for the second order method, where the errors grow more gradually in time. Furthermore, the errors in the internally fourth order method are substantially smaller than those of the second order method, especially before the wave hits the boundary.

We proceed by investigating the long-time stability properties of the method. We take the domain to be the same ellipse used above and take the forcing functions such that the exact solution is the trigonometric traveling wave (8.1). In Figure 3, we show the error in the solution as a function of time for different values of α and for different grid sizes. We conclude that it is sufficient to take $\alpha = 2 \times 10^{-3}$ for both the second order and the predictor-corrector scheme. Note that these computations integrated the solution for long times. In particular, the second order scheme on the finer grid ($N = 401$) required 66,666 time steps to reach $t = 200$. Also note that there is no long-time increase in the error, which indicates that the damping is very mild.

We next study the homogeneous problem

$$F(\boldsymbol{x}, t) \equiv 0, \quad f(\boldsymbol{x}, t) \equiv 0,$$

in a domain bounded by an ellipse centered at the origin, with semiaxes $x_s = 2.0$ and $y_s = 2.54$. The Cartesian grid covers the square $-2.1 \leq x \leq 2.1$, $-2.64 \leq y \leq 2.64$.

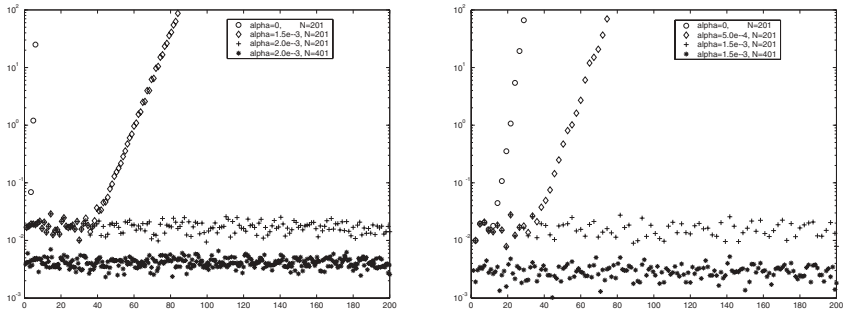


FIG. 3. *The max-norm of the error in the solution as a function of time. The second order scheme* (1.6) *was run at CFL* $= 0.5$ *(left), and the predictor-corrector method* (7.5), (7.6) *was run at CFL* $= 1.0$ *(right). Note that to stabilize the solution, the damping coefficient had to be slightly larger for the second order scheme* (2×10^{-3}) *than for the predictor-corrector method* (1.5×10^{-3}).

We take initial data to be

$$u_0(x, y) = \phi(\sqrt{x^2 + (y - y_F)^2}),$$

where $\phi(\xi)$ is given by (8.2). The upper focal point is located at $y_F = \sqrt{y_s^2 - x_s^2} \approx 1.56$ and

$$u_1(\boldsymbol{x}) = -\phi'(\sqrt{x^2 + (y - y_F)^2}).$$

The parameters in $\phi(\xi)$ are

$$\xi_0 = 0.2, \quad \xi_1 = 0.4, \quad \epsilon = 0.035.$$

Note that the initial data is chosen such that the wave is essentially traveling radially outwardly from the focal point $(0, y_F)$. By making a ray-tracing argument, we see that a high frequency wave should reflect the boundary and refocus at the other focal point $(0, -y_F)$. This was verified for the Dirichlet problem in [9] (Figure 6). For the Neumann boundary condition, we should get a similar behavior, except that the solution should have the opposite phase compared to the Dirichlet case. This is confirmed in Figure 4, where we show a well-resolved calculation using the predictor-corrector scheme with $N = 801$ and $CFL = 1.0$. It is interesting to use this calculation as a yard-stick to compare the quality of the solutions from both schemes at a lower resolution, $N = 401$; see Figures 5 and 6. Observe the more pronounced over- and undershoots for the second order method in comparison to the predictor-corrector method, indicating that the phase-error dominates at the time of comparison. In all these calculations, the damping coefficient was $\alpha = 0.001$.

While all theory and all numerical experiments up to this point have been presented for the third order accurate discretization of the boundary conditions, our practical experience with the second order boundary condition stencil is at least as good. The advantage of the second order stencil is that it uses fewer internal points, which becomes important for thin or marginally resolved geometries. However, near true corners, the second order boundary condition needs to be modified to avoid using grid points where the solution is undefined; see Figure 7. To avoid this problem, all grid points are first scanned in a preprocessing step to detect interior points within $\sqrt{2}h$ of corners. All such points that also have at least two exterior nearest neighbors

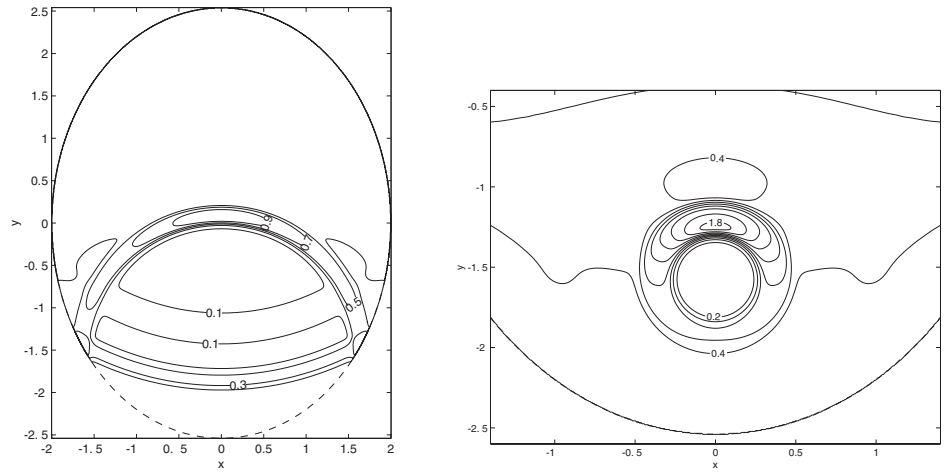


FIG. 4. *Contours of the bouncing wave solution to the Neumann problem. Here a reference solution is produced with the predictor-corrector scheme,* $CFL = 1.0, N = 801$, $t = 3.12$ *(left), and* $t = 4.41$ *(right). The dashed line is the boundary and the contour spacing is* 0.2.

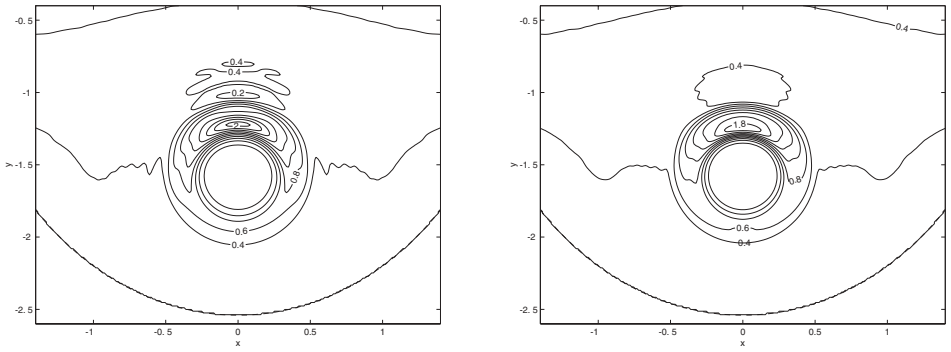


FIG. 5. *Contours of the bouncing wave solution to the Neumann problem. The second order scheme is used with* $CFL = 0.5$ *(left) and the predictor-corrector scheme is used with* $CFL = 1.0$ *(right). Here* $N = 401$, $t = 4.41$, *and the contour spacing is* 0.2.

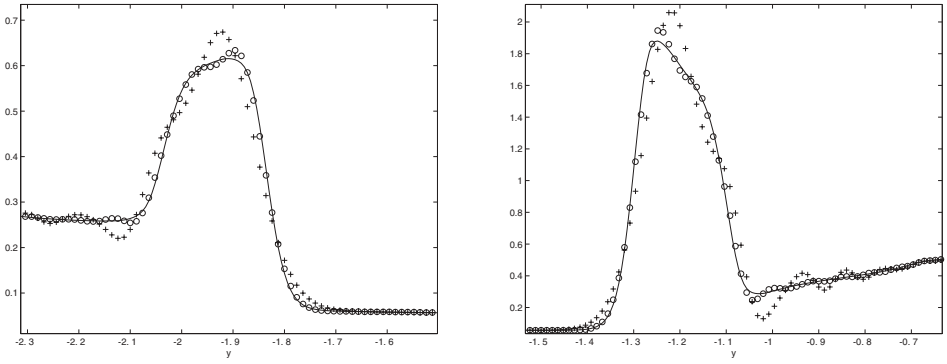


FIG. 6. *Comparison of the bouncing wave solution for the Neumann problem at* $t = 4.41$ *along the line* $x = 0$ *centered around* $y = -2.0$ *(left) and* $y = -1.2$ *(right). The reference solution is for* $N = 801$ *(solid), the second order scheme is for* $N = 401$, $CFL = 0.5$(“+”), *and the predictor-corrector scheme is for* $N = 401$, $CFL = 1.0$(“o”).

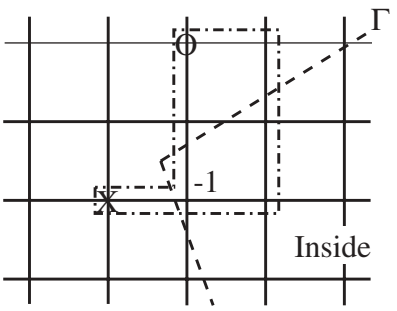


FIG. 7. *The standard second order boundary condition stencil (outlined with a dash-dotted line) for the ghost point at "X" involves the point "O", where the solution is undefined due to the corner. In this case, the boundary stencil at "X" is reduced to a divided difference between the solution at "−1" and "X."*

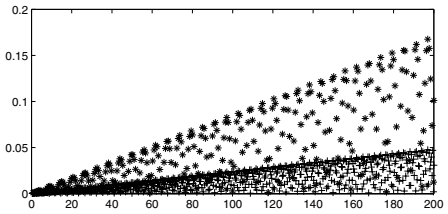


FIG. 8. *The max-norm of the error for a rotated square domain, as a function of time. The "∗" correspond to the grid size $h = 1.417 \times 10^{-2}$, and the "+" represent the grid size $h = 7.087 \times 10^{-3}$. The damping coefficient was $\alpha = 2 \times 10^{-3}$.*

get marked with a "−1." The boundary condition stencil at ghost points neighboring a "−1" point is then modified to be a divided difference between the ghost point and the "−1" point; i.e., the direction of the normal is locally changed to be either vertical or horizontal. As a consequence, no undefined points are involved in the boundary stencil near the corner, and the resulting contribution to the discretized Laplace operator (the matrix A) will be locally symmetric.

While the modified boundary condition approximation will be at most first order accurate near each corner, it is not clear what impact that truncation error has on the accuracy of the solution. We are also interested in the long-time stability of the resulting scheme. To investigate these issues, we take the domain to be a square with side length 2, rotated 10 degrees relative to the grid directions. In rotated coordinates $\tilde{x} = x\cos(\theta) + y\sin(\theta)$, $\tilde{y} = -x\sin(\theta) + y\cos(\theta)$, $\theta = 10\pi/180$, an exact solution of the homogeneous wave equation can be constructed using Fourier expansion. Here we take

$$u(\tilde{x}, \tilde{y}, t) = \sin\left(\frac{\pi\tilde{x}}{2}\right)\sin\left(\frac{3\pi\tilde{y}}{2}\right)\cos(\omega t), \quad \omega = \frac{\pi\sqrt{10}}{2},$$

which satisfies homogeneous Neumann conditions along $\tilde{x} = \pm 1$ and $\tilde{y} = \pm 1$, respectively. The errors in the computed solutions on two grid sizes are reported in Figure 8, indicating that the solution is almost second order accurate despite the corners. However, for reasons not currently understood, the errors accumulate and seem to grow linearly in time.

In our last numerical example, we use the numerical method to compute the

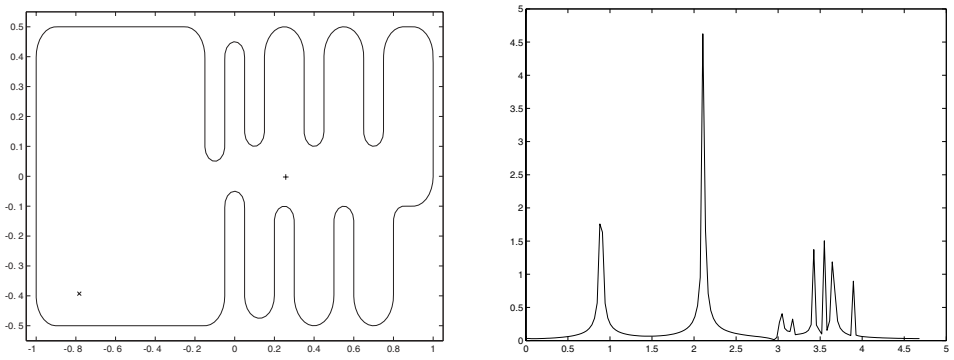


FIG. 9. *The geometry for the harbor model (left). The computational grid covered* $-1.05 \le x \le 1.05$, $-0.55 \le y \le 0.55$ *and had* 801×401 *grid points corresponding to the grid size* $h = 2.6 \times 10^{-3}$. *The forcing is located at the "*x*" and the signal is recorded at the "+." The right figure shows the lowest modes in the discrete Fourier transform of the recorded signal, as a function of the frequency. The spikes indicate eigenfrequencies.*

eigenfrequencies and eigenmodes of the domain shown in Figure 9. Since the wave equation models the propagation of small amplitude water waves, we may think of this geometry as representing a simple harbor. Even though the grid is rather fine, the wide stencil used by the third order boundary condition couples the solution at some ghost points near the ends of the convex fingers protruding into the domain. By coupling we mean that at least one of the interior points in one boundary condition stencil is also a ghost point. Satisfying the boundary conditions at all ghost points would then require an iteration over the ghost point values. To avoid this iterative procedure, we will instead use the second order boundary condition, which uses fewer interior points in its stencil. For this case, the solution does not get coupled at any ghost points. To estimate the eigenfrequencies, we apply a forcing to two consecutive points in space,

$$F(x_{i,j}, y_{i,j}, t) = \begin{cases} Ke^{-(t-t_0)^2/\varepsilon_1^2}, & i = I_1,\ j = J_1, \\ -Ke^{-(t-t_0)^2/\varepsilon_1^2}, & i = I_1 + 1,\ j = J_1, \\ 0, & \text{otherwise.} \end{cases}$$

Here $K = 10^5$, $\varepsilon_1 = 0.07$, $t_0 = 1.0$. We choose this forcing since it is likely to have a component along each eigenmode, except the constant mode corresponding to the zero eigenvalue, which is present due to the Neumann boundary condition. We start the computation from rest and integrate up to time $T = 200$. During the computation, the solution is recorded at another point (I_r, J_r). This signal is then Fourier-transformed in time, after which the eigenfrequencies of the domain appear as spikes in the spectrum; see Figure 9. Note that the frequency resolution is limited by $2\pi/T$, so a longer computation leads to a more accurate estimate of the eigenfrequencies. Also note that the eigenvalues of

$$\begin{aligned} \Delta u &= \lambda u \quad \text{in } \Omega, \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on } \Gamma \end{aligned}$$

are related to the eigenfrequencies ω through $\lambda = -\omega^2$.

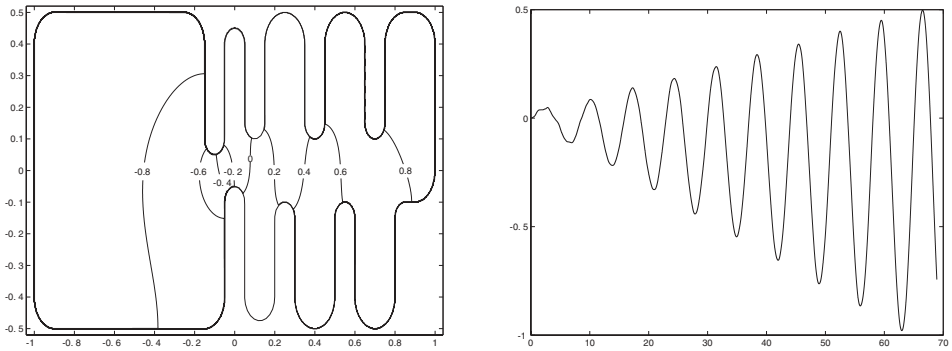


FIG. 10. *A contour plot of the solution at $t = 62.4$ approximating the eigenmode corresponding to the eigenfrequency $\omega_r = 0.90$ (left). Here, the contour levels are equally spaced between -0.8 and 0.8. The right figure shows the time-history of the solution at the point $(x, y) = (-0.6526, -0.1976)$. Because of resonance, the amplitude grows linearly in time.*

To compute the corresponding eigenmode, we perform a second computation, where the forcing is taken to be

$$F(x, y, t) = \sin(\omega_r t)\gamma'(x - x_0)\gamma(y), \quad \gamma(\xi) = e^{-\xi^2/\varepsilon_2^2}, \quad x_0 = -0.6,\ \varepsilon_2 = 0.2.$$

The frequency of the time-harmonic forcing is chosen to obtain resonance. In this computation, we take $\omega_r = 0.90$, which is the approximate location of the first spike in the spectrum; see Figure 9. Due to resonance, the solution will be more and more dominated by the corresponding eigenmode as time increases, assuming that the forcing is not orthogonal to that mode. The resulting eigenmode is shown in Figure 10 together with the time-history of the solution in one point, which demonstrates the expected linear growth in amplitude.

9. Conclusions. We have presented stability theory and numerical examples for a Cartesian embedded boundary scheme which directly discretizes the second order wave equation subject to Neumann boundary conditions, without rewriting the problem as a system of first order equations. Since the discrete approximation of the Laplacian subject to the Neumann boundary condition leads to a matrix A that is not symmetric, the stability theory developed in [9] does not directly apply. Indeed, numerical experiments in two-dimensional domains indicate that the basic undamped scheme is unstable. In the one-dimensional case, we prove that the semidiscrete scheme is stable, thus indicating that the instability is due to two-dimensional effects. In two dimensions, tangential derivatives are present in the truncation error of the boundary condition, when the boundary is not aligned with the mesh. A two-dimensional stability theory is presented that first is used to show the destabilizing effect of perturbing a Neumann boundary condition by tangential derivatives. The stability theory also predicts that a small fourth order dissipative term $h^3\Delta^2 u_t$ can control the destabilizing effects of high order tangential derivatives. The discrete stabilization term $h^3 A^T A(u^n - u^{n-1})/k$ is proposed for the practical computation. This term can be evaluated all the way up to the boundary so no extra numerical boundary conditions are necessary. After discretization in space, the system of second order ordinary differential equations is integrated in time using a second or fourth order explicit method. Improved spatial accuracy can be achieved away from the boundary by adding a fourth order spatial correction term. Our numerical examples indicate

that the resulting scheme is second order accurate measured in max-norm and that the time step can be chosen independently of small grid cells near the boundary. Numerical experiments also show that the amount of dissipation needed to stabilize the scheme is very small and, for smooth boundaries, long-time computations do not show any accumulation of the error. A simple modification of the scheme in the vicinity of corners is proposed, but more work is needed to fully understand its implications.

Work is underway to generalize the proposed method to Maxwell's equations written as a system of second order wave equations, which requires more complicated boundary conditions to be satisfied. Further work is also planned to extend the method to three space dimensions.

Appendix A. Computing $\boldsymbol{A^T u}$. Using standard notation for an $N \times N$ matrix $\mathcal{A}$ and vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, the most straightforward way of computing $\boldsymbol{v} = \mathcal{A}^T \boldsymbol{u}$ might be

$$v_i = \sum_{j=1}^{N} \mathcal{A}_{j,i} u_j.$$

However, when the matrix is sparse, it is inefficient to store all matrix elements explicitly. If we let $\boldsymbol{a}_i^T$ denote the ith row of $\mathcal{A}$, we can write $\mathcal{A}$ in row form,

$$\mathcal{A} = \begin{pmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{a}_N^T \end{pmatrix}, \quad \mathcal{A}^T = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N), \tag{A.1}$$

and $\boldsymbol{v} = \mathcal{A}^T \boldsymbol{u} = \sum_{j=1}^{N} \boldsymbol{a}_j u_j$. Hence, another way of computing $\mathcal{A}^T \boldsymbol{u}$ is by accumulating the contributions from each column of $\mathcal{A}^T$, i.e., each row of $\mathcal{A}$:

1. $\boldsymbol{v} = 0$;
2. for $j = 1, 2, \ldots, N$ do $\boldsymbol{v} += \boldsymbol{a}_j u_j$.

Here the operator $+=$ means evaluate the right-hand side and add the result to the left-hand side (as it is defined in the "C" programming language).

Next consider the particular form of the matrix $\mathcal{A} = A$ that arises in our embedded boundary discretization. Away from the boundary, A is defined by (1.2). Near the boundary, outside points in the stencil get eliminated using the discretized Neumann boundary condition, resulting in a stencil of the type (7.1). In general, each row of A will only have a few nonzero entries. To simplify the notation, we define $u(k,l) =: u_{k,l}$. Each row of the matrix can then be written in sparse form as

$$A\boldsymbol{u}|_{i,j} =: \sum_{k=1}^{NZ_{i,j}} a_{i,j}^{(k)} u(I_{i,j}^{(k)}, J_{i,j}^{(k)}), \tag{A.2}$$

where $NZ_{i,j}$ is the number of nonzero entries for the row corresponding to grid point (i,j), and $(I_{i,j}^{(k)}, J_{i,j}^{(k)})$ is the grid point index of the kth contribution to $A\boldsymbol{u}$ in that row.

Equation (A.2) represents the matrix A in a sparse row form corresponding to (A.1). The operation $\boldsymbol{v} = A^T \boldsymbol{u}$ can therefore be computed using the above accumulation algorithm,

1. $\boldsymbol{v} = 0$;
2. for all grid points (i,j) inside Ω do
   for $k = 1, 2, \ldots, NZ_{i,j}$ do

$$v(I_{i,j}^{(k)}, J_{i,j}^{(k)}) += a_{i,j}^{(k)} u(i,j).$$

We note that it is only necessary to form the sparse representation of A at interior points where some neighbors are outside Ω. If all neighbors of (i,j) are interior, the "for k"-loop in the second step in the accumulation algorithm can be replaced by

$$v(i+1,j) += \frac{u(i,j)}{h^2}, \quad v(i-1,j) += \frac{u(i,j)}{h^2}, \quad v(i,j) += -\frac{4u(i,j)}{h^2},$$
$$v(i,j-1) += \frac{u(i,j)}{h^2}, \quad v(i,j+1) += \frac{u(i,j)}{h^2}.$$

Appendix B. Proof of Theorem 6.1. The proof is divided into three cases: $|\lambda| \gg |b|$, $|\lambda| \ll 1$, and $|\lambda| \leq C|b|$.

Case 1, $|\lambda| \gg |b|$. We have $(\lambda - b)/(\lambda + b) \sim 1$. To make the modulus of the right-hand side of (6.6) be close to one,

$$\lambda = Ni\pi + \tilde{\lambda}, \quad |\tilde{\lambda}| \ll 1, \ N \geq 1 \text{ integer}.$$

(Note that $N = 0$; i.e., $|\lambda| \ll 1$ is treated in Case 2 below.) To first approximation in $\tilde{\lambda}$,

$$\frac{\tilde{\lambda} + iN\pi - b}{\tilde{\lambda} + iN\pi + b} = 1 + 2\tilde{\lambda}.$$

Therefore,

$$\tilde{\lambda} + iN\pi - b = \tilde{\lambda} + iN\pi + b + 2\tilde{\lambda}^2 + 2iN\pi\tilde{\lambda} + 2b\tilde{\lambda}$$

or

$$\tilde{\lambda}^2 + (iN\pi + b)\tilde{\lambda} + b = 0.$$

Since $|\lambda| \gg |b|$, $|b| \ll N\pi$, and we can expand the roots of $\tilde{\lambda}$ in the small parameter $\epsilon = b/N\pi$, $|\epsilon| \ll 1$,

$$\begin{aligned}\tilde{\lambda} &= -\frac{iN\pi + b}{2} \pm i\frac{N\pi}{2}\sqrt{1 - \frac{4b}{N\pi}\left(\frac{i}{2} - \frac{1}{N\pi} + \frac{b}{4N\pi}\right)} \\ &= -\frac{iN\pi + b}{2} \pm \left(\frac{iN\pi + b}{2} + \frac{ib}{N\pi} + \mathcal{O}(\epsilon^2)\right).\end{aligned}$$

Only the plus sign gives $|\tilde{\lambda}| \ll 1$, and we have

$$\lambda = iN\pi + \tilde{\lambda} \approx iN\pi + \frac{ib}{N\pi}.$$

By solving the characteristic equation (5.16) for s and inserting the above expression for λ,

$$s = -\frac{a}{2} \pm \sqrt{\frac{a^2}{4} - \omega^2 - N^2\pi^2 - 2b - \frac{b^2}{(N\pi)^2}}. \tag{B.1}$$

We assume that

$$0 \le \alpha \le 1. \tag{B.2}$$

Since $|\omega h| \le 1$, we have that

$$|a| \le |\omega|, \tag{B.3}$$

so $N^2\pi^2 + \omega^2 - a^2/4$ is real and positive. Because $|b|/N\pi \ll 1$, we can expand the roots of (B.1),

$$\begin{aligned} s &= -\frac{a}{2} \pm i\sqrt{\left(N^2\pi^2 + \omega^2 - \frac{a^2}{4}\right)}\sqrt{1 + \frac{2b}{N^2\pi^2 + \omega^2 - a^2/4}\left(1 + \frac{b}{2N^2\pi^2}\right)} \\ &= -\frac{a}{2} \pm \left(i\sqrt{\left(N^2\pi^2 + \omega^2 - \frac{a^2}{4}\right)} + \frac{ib}{\sqrt{N^2\pi^2 + \omega^2 - a^2/4}} + \cdots\right). \end{aligned}$$

We have $ib = -\beta_1\omega^5h^4 + i\beta_2\omega^4h^3$, and $N^2\pi^2 + \omega^2 - a^2/4 \ge 3\omega^2/4$, so

$$\operatorname{Re} s \approx -\frac{1}{2}\alpha h^3\omega^4 \mp \frac{\beta_1\omega^5h^4}{\sqrt{N^2\pi^2 + \omega^2 - a^2/4}} \le -\frac{1}{2}\alpha h^3\omega^4 + \frac{2|\beta_1|\omega^4h^4}{\sqrt{3}}. \tag{B.4}$$

Therefore, $\operatorname{Re} s < 0$ for $\alpha \ge 4|\beta_1|h/\sqrt{3}$, and we conclude that there can be no exponentially growing solutions with $|\lambda| \gg |b|$, when α exceeds that value.

Case 2, $|\lambda| \ll 1$. If $|\lambda| \ll 1$, we can replace (6.6) by

$$\frac{\lambda - b}{\lambda + b} = 1 + 2\lambda,$$

i.e.,

$$\lambda^2 = -b + \mathcal{O}(b^{3/2}). \tag{B.5}$$

Then (5.16) gives us

$$s = -\frac{a}{2} \pm \sqrt{\frac{a^2}{4} - \omega^2 + \lambda^2} \approx -\frac{a}{2} \pm \sqrt{\frac{a^2}{4} - \omega^2 - i\beta_1\omega^5h^4 - \beta_2\omega^4h^3}, \tag{B.6}$$

and by making the same expansion as above we obtain

$$\operatorname{Re} s \approx -\frac{1}{2}\alpha h^3\omega^4 \pm \frac{|\beta_1|\omega^4h^4}{\sqrt{3}}.$$

Hence, in this case, $\operatorname{Re} s < 0$ for $\alpha > 2|\beta_1|h/\sqrt{3}$, and there can be no exponentially growing solutions with $|\lambda| \ll 1$ when α satisfies that inequality. Note that (B.5) implies that $|b| \ll 1$ when $|\lambda| \ll 1$.

Case 3, $|\lambda| \le C|b|$. From Case 2 above, we know that $|b| \ll 1$ when $|\lambda| \ll 1$. We can therefore assume that $|b| \ge \delta_1 > 0$. Since $|b| = \omega^4h^3\sqrt{\beta_2^2 + \beta_1^2\omega^2h^2}$ and $|\omega h| \le 1$,

$$c_1h^{-3/4} \le |\omega| \le h^{-1}. \tag{B.7}$$

Let us define a complex number ρ such that

$$\lambda + b = \rho b, \tag{B.8}$$

that is, $|\rho - 1| \leq C$. There are two possibilities:

$$\text{(a)} \quad |\rho| \geq \delta > 0, \quad \text{(b)} \quad |\rho| \leq \epsilon \ll 1.$$

For possibility (a), we start by deriving a bound for $\operatorname{Re}\lambda$. Let $\lambda_r = \operatorname{Re}\lambda$ and $\lambda_i = \operatorname{Im}\lambda$. From (6.6)

$$e^{2\lambda_r} = \left|\frac{\lambda + b - 2b}{\lambda + b}\right| \leq 1 + \left|\frac{2b}{\lambda + b}\right| = 1 + \frac{2}{|\rho|} \leq 1 + \frac{2}{\delta},$$

and therefore

$$\lambda_r \leq \frac{1}{2}\log\left(1 + \frac{2}{\delta}\right) = c_2. \tag{B.9}$$

By solving the characteristic equation (5.16) for s, we have

$$s = -\frac{a}{2} \pm \sqrt{\frac{a^2}{4} - \omega^2 - \lambda_i^2 + 2i\lambda_i\lambda_r + \lambda_r^2}.$$

Since (B.7) bounds $|\omega|$ from below, $\lambda_r^2 \ll \omega^2$. Hence, we can neglect this term, expand the roots of s as before, and use (B.9) to get

$$\operatorname{Re} s \leq -\frac{1}{2}\alpha h^3\omega^4 + \frac{c_2|\lambda_i|}{\sqrt{\frac{3}{4}\omega^2 + \lambda_i^2}}.$$

Let $\rho_r = \operatorname{Re}\rho$ and $\rho_i = \operatorname{Im}\rho$. The relation (B.8) gives $\lambda_i = \xi\omega^4h^3$, where the real-valued coefficient $\xi = \rho_i\beta_2 + (\rho_r - 1)\beta_1\omega h$. Clearly, for all $|\rho| \geq \delta$ and $|\omega h| \leq 1$,

$$\frac{|\lambda_i|}{\sqrt{\frac{3}{4}\omega^2 + \lambda_i^2}} = \frac{|\xi||\omega h|^3}{\sqrt{\frac{3}{4} + \xi^2\omega^6h^6}} \leq c_3|\omega h|^3.$$

Since $|\omega|$ is bounded from below by (B.7), $\operatorname{Re} s < 0$ for $\alpha > 2c_2c_3c_1^{-1}h^{3/4}$.

For possibility (b), we exploit that $\operatorname{Re}\lambda > 0$ for $\operatorname{Re} s > 0$. We have

$$\lambda = -b(1 - \rho), \quad |\rho| \ll 1. \tag{B.10}$$

When $\rho = 0$, this case reverts to (5.17) and the half-plane problem. For $\beta_2 > 0$, $\operatorname{Re}(-b) = -\beta_2\omega^4h^3 < 0$, but $\operatorname{Re}\lambda > 0$ for $\operatorname{Re} s > 0$, which is contradicted by (B.10). Hence there are no solutions with $\operatorname{Re} s > 0$ when $\beta_2 > 0$. When $\beta_2 < 0$, $\beta_1 \ll 1$, and $|\beta_2| \ll 1$, Theorem 5.2 applies and the problem can be stabilized by a small amount of dissipation $\alpha \geq K|\beta_1\beta_2|$.

For the perturbed case, $|\rho| = \epsilon$, $\epsilon \ll 1$, a simple computation yields

$$\operatorname{Re}(-b + \rho b) = (-1 + \rho_r)\beta_2\omega^4h^3 - \rho_i\beta_1\omega^5h^4,$$

and $\operatorname{Re}(-b + \rho b) < 0$ if $(-1 + \rho_r)\beta_2 + |\rho_i\beta_1| < 0$, i.e.,

$$\beta_2 > \frac{|\rho_i|}{1 - \rho_r}|\beta_1| \approx \epsilon|\beta_1|. \tag{B.11}$$

Hence, when $|\rho| \leq \epsilon$, there cannot be any solutions with $\operatorname{Re} s > 0$ when (B.11) is satisfied. For $\beta_2 < \epsilon|\beta_1|$, we can apply the same expansion as in Theorem 5.2 for the half-plane problem. Since λ is perturbed by $\rho b = \mathcal{O}(\epsilon)$, the roots of s can only be perturbed by $\mathcal{O}(\epsilon)$, and the amount of dissipation necessary to stabilize the problem remains essentially the same.

This concludes the proof of Theorem 6.1.

REFERENCES

[1] G. C. Cohen, *Higher-Order Numerical Methods for Transient Wave Equations*, Springer-Verlag, New York, 2002.

[2] P. Colella, *Volume-of-fluid methods for partial differential equations*, in Godunov Methods: Theory and Applications, E. F. Toro, ed., Kluwer Academic/Plenum Publishers, New York, 2001, pp. 161–177.

[3] L. Collatz, *The Numerical Treatment of Differential Equations*, 3rd ed., Springer-Verlag, New York, 1960.

[4] A. Ditkowski, K. Dridi, and J. S. Hesthaven, *Convergent Cartesian grid methods for Maxwell's equations in complex geometries*, J. Comput. Phys., 170 (2001), pp. 39–80.

[5] B. Gustafsson, H.-O. Kreiss, and J. Oliger, *Time Dependent Problems and Difference Methods*, Wiley-Interscience, New York, 1995.

[6] H. Johansen and P. Colella, *A Cartesian grid embedded boundary method for Poisson's equation on irregular domains*, J. Comput. Phys., 147 (1998), pp. 60–85.

[7] H.-O. Kreiss, *Difference approximations for boundary and eigenvalue problems for ordinary differential equations*, Math. Comp., 26 (1972), pp. 605–624.

[8] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press, New York, 1989.

[9] H.-O. Kreiss, N. A. Petersson, and J. Ystrom, *Difference approximations for the second order wave equation*, SIAM J. Numer. Anal., 40 (2002), pp. 1940–1967.

[10] H.-O. Kreiss and L. Wu, *On the stability definition of difference approximations for the initial boundary value problem*, Appl. Numer. Math., 12 (1993), pp. 213–227.

[11] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, 1992, pp. 117–118.

[12] R. B. Pember, J. B. Bell, P. Collella, W. Y. Crutchfield, and M. Welcome, *An adaptive Cartesian grid method for unsteady compressible flow in irregular regions*, J. Comput. Phys., 120 (1995), pp. 278–304.

[13] M. Sussman, P. Smereka, and S. Osher, *A level set approach for computing solutions to incompressible two-phase flow*, J. Comput. Phys., 114 (1994), pp. 146–159.

[14] R. F. Warming and B. J. Hyett, *The modified equation approach to the stability and accuracy analysis of finite-difference methods*, J. Comput. Phys., 14 (1974), pp. 159–179.

[15] R. Weller and G. H. Shortley, *Calculation of stresses within the boundary of photoelastic models*, J. Appl. Mech., 6 (1939), pp. A-71–A-78.

[16] C. Zhang and R. LeVeque, *The immersed interface method for acoustic wave equations with discontinuous coefficients*, Wave Motion, 25 (1997), pp. 237–263.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1324–1341

A FINITE ELEMENT APPROXIMATION OF A VARIATIONAL INEQUALITY FORMULATION OF BEAN'S MODEL FOR SUPERCONDUCTIVITY*

C. M. ELLIOTT†, D. KAY†, AND V. STYLES†

Abstract. We introduce a finite element approximation of a variational formulation of Bean's model for the physical configuration of an infinitely long cylindrical superconductor subject to a transverse magnetic field. We prove an error between the exact solution and the approximate solution for the current density and the magnetic field in appropriate norms of order $h^{1/2} + \Delta t$. Numerical simulations for a variety of applied magnetic fields are also presented.

Key words. finite elements, variational inequalities, superconductors

AMS subject classifications. 49J40, 74S05, 82D55

DOI. 10.1137/S0036142902412324

1. Introduction. In this paper we consider the numerical approximation of an evolutionary variational inequality arising from a critical state model for a type-II superconductor. The physical setting is that of an infinitely long cylinder of type-II superconducting material subject to an applied transverse magnetic field. We take the cylindrical superconductor to occupy the region $D = \Omega \times \mathbb{R}$, where $\Omega \subset \mathbb{R}^2$, a bounded, simply connected domain in $\mathbb{R}^2$, is the cross section of the superconductor. The physical vector fields that are relevant are the current density $\mathbf{J} = (0, 0, J(\underline{x}, t))$, which is parallel to the axis of the cylinder, and the magnetic field $\mathbf{H} = (\underline{H}(\underline{x}, t), 0)$, which is orthogonal to the cylinder's axis, for $\underline{x} \in \mathbb{R}^2$. The well-known Bean critical state model can be formulated as an evolutionary variational inequality for $J(\underline{x}, t)$ of the form (see [10]):

(**P**) Find $J(\cdot, t) \in K$ for a.e. $t \geq 0$ such that $J(\cdot, 0) = J_0 \in K$ and

$$\left(\frac{\partial GJ}{\partial t}, \eta - J\right) \geq (f, \eta - J) \quad \forall\, \eta \in K. \tag{1.1}$$

Here $(\cdot, \cdot)$ denotes the standard L^2 inner product over Ω,

$$\mathcal{V} := \left\{\eta \in L^2_{\mathrm{loc}}(\mathbb{R}^2) : \nabla\eta \in L^2(\mathbb{R}^2),\ (\eta, 1) = 0\right\},$$

$$K = \left\{\eta \in \mathcal{V} : \eta = 0 \ \text{ on } \mathbb{R}^2/\overline{\Omega}, |\eta| \leq J_c, \ \ (\eta, 1) = 0\right\}$$

and $G : \mathcal{V}' \to \mathcal{V}$ is the "inverse Laplacian" operator defined by the solution to the following variational problem:

Given $v \in \mathcal{V}'$, find $Gv \in \mathcal{V}$ such that

$$(\nabla Gv, \nabla\eta)_{\mathbb{R}^2} = \langle v, \eta\rangle \quad \forall\, \eta \in \mathcal{V} \tag{1.2}$$

*Received by the editors July 30, 2002; accepted for publication (in revised form) June 19, 2003; published electronically October 28, 2004.

http://www.siam.org/journals/sinum/42-3/41232.html

†Centre for Mathematical Analysis and its Applications, University of Sussex, Falmer, Brighton BN1 9QH, UK (c.m.elliott@sussex.ac.uk, d.a.kay@sussex.ac.uk, v.styles@sussex.ac.uk). The research of the second author was supported by an NUF-NAL 00 grant. The research of the third author was supported by a Leverhulme 2000 Fellowship.

with $\langle\cdot,\cdot\rangle$ denoting the duality pairing between $\mathcal{V}'$ and $\mathcal{V}$. For $v \in \mathcal{F} \subset \mathcal{V}'$ we have

$$\langle v, \eta\rangle = (v, \eta) \quad \forall\ \eta \in \mathcal{V},$$

where

$$\mathcal{F} := \Big\{\eta \in \mathcal{V}' : \eta \in L^2_{\rm loc}(\mathbb{R}^2) :\ \eta = 0 \text{ on } \mathbb{R}^2/\overline{\Omega}\Big\}.$$

Setting

$$\mathcal{F}_0 := \Big\{\eta \in \mathcal{F} : (\eta, 1) = 0\Big\},$$

we have the following for all $v \in \mathcal{F}_0$:

$$-\Delta Gv = v \quad \text{in } \mathbb{R}^2, \qquad \int_\Omega Gv\ d\underline{x} = 0, \qquad \text{and} \qquad \nabla Gv \sim 0 \quad \text{at } \infty, \tag{1.3}$$

and Gv is unique.

Throughout the remaining sections we assume that

$$f \in L^2(0,T; H^2(\Omega)), \quad f_t \in L^2(0,T; H^1(\Omega)). \tag{1.4}$$

It follows from the classical theory of evolutionary variational inequalities that (**P**) has a unique solution; see [10, 5].

2. Derivation of the model and reduction to a bounded domain.

2.1. Derivation of the model. We suppose that all field variables depend only on t and $\underline{x} \in \mathbb{R}^2$, and that there is a prescribed, time dependent, smooth magnetic field $\mathbf{H}^a = (\underline{H}^a(\underline{x}, t), 0)$ applied at infinity and a prescribed, bounded current density $\mathbf{J}^a = J^a(\underline{x}, t)\mathbf{e}_3$, exterior to the superconductor, such that the compatibility condition

$$\mathbf{J}^a - \operatorname{curl} \mathbf{H}^a \to 0 \quad \text{as } |\underline{x}| \to \infty$$

is satisfied. Then Maxwell's equations, neglecting displacement current, are

$$\begin{aligned} \frac{\partial \mathbf{H}}{\partial t} + \operatorname{curl} \mathbf{E} &= \mathbf{0} \qquad \text{in } \mathbb{R}^2, \\ \operatorname{curl} \mathbf{H} &= \mathbf{J} \qquad \text{in } \mathbb{R}^2, \\ \nabla \cdot \mathbf{H} &= 0 \qquad \text{in } \mathbb{R}^2, \end{aligned}$$

where $\mathbf{E}$ is the electric field; see [10]. Note we have taken the magnetic permeability equal to 1 for simplicity.

The critical state model assumes the following nonlinear Ohm's law in the superconductor,

$$\mathbf{E} \ = \ \rho\mathbf{J} \qquad \text{in } \Omega$$

with

$$|\mathbf{J}| \ \le \ J_c \qquad \text{in } \Omega,$$

and the effective resistivity ρ achieves the constraint on $|\mathbf{J}|$ by the relation $\rho \in \beta(|\mathbf{J}|)$, where β is a multivalued map given by the graph

$$\beta(r) = \begin{cases} (-\infty, 0] & \text{if} \quad r = -J_c, \\ 0 & \text{if} \quad |r| < J_c, \\ [0, \infty) & \text{if} \quad r = J_c. \end{cases}$$

We assume that exterior to the superconductor the current is prescribed so

$$\mathbf{J} = \mathbf{J}^a \qquad \text{in } \mathbb{R}^2/\overline{\Omega}.$$

To complete this set of equations we require initial and boundary conditions for the magnetic field given, respectively, by

$$\mathbf{H}(\underline{x}, 0) \;=\; \mathbf{H}_0(\underline{x})$$

and

$$\mathbf{H} \;\to\; \mathbf{H}^a \quad \text{as } |\underline{x}| \to \infty.$$

On the boundary of the superconductor, $\partial\Omega$, we have that

$$[\mathbf{H}_\tau] = [\mathbf{H}_\nu] = 0,$$

where $[\mathbf{H}_\tau]$ and $[\mathbf{H}_\nu]$ denote the jumps in the tangential and normal components, respectively, of $\mathbf{H}$ across $\partial\Omega$.

In order to consider homogeneous boundary conditions at infinity, it is convenient to introduce a current density $\mathbf{J}^e$ defined by

$$\mathbf{J}^e \;= \begin{cases} \mathbf{0} & \text{in } \Omega, \\ \mathbf{J}^a & \text{in } \mathbb{R}^2/\overline{\Omega}. \end{cases}$$

Associated with $\mathbf{J}^e$ is the magnetic field $\mathbf{H}^e$ such that

$$\begin{aligned} \operatorname{curl} \mathbf{H}^e &= \mathbf{J}^e \quad \text{in } \mathbb{R}^2, \\ \nabla \cdot \mathbf{H}^e &= 0 \quad \text{in } \mathbb{R}^2, \\ \mathbf{H}^e &\to \mathbf{H}^a \quad \text{as } |\underline{x}| \to \infty. \end{aligned}$$

Finally, we use the shift

$$\hat{\mathbf{J}} = \mathbf{J} - \mathbf{J}^e \quad \text{and} \quad \hat{\mathbf{H}} = \mathbf{H} - \mathbf{H}^e$$

to give the problem

$$\frac{\partial \hat{\mathbf{H}}}{\partial t} + \operatorname{curl}(\rho \hat{\mathbf{J}}) = -\frac{\partial \mathbf{H}^e}{\partial t} \qquad \text{in } \Omega, \tag{2.1}$$

$$\operatorname{curl} \hat{\mathbf{H}} = \hat{\mathbf{J}} \qquad \text{in } \mathbb{R}^2, \tag{2.2}$$

$$\nabla \cdot \hat{\mathbf{H}} = 0 \qquad \text{in } \mathbb{R}^2, \tag{2.3}$$

$$|\hat{\mathbf{J}}| \le J_c \qquad \text{in } \Omega \tag{2.4}$$

together with the boundary condition

$$\hat{\mathbf{H}} \to \mathbf{0} \text{ as } |\underline{x}| \to \infty.$$

Note that interpreting (2.2), (2.3) in conservation form yields the compatibility boundary conditions

$$\left[\hat{\mathbf{H}}_\nu\right] = \left[\hat{\mathbf{H}}_\tau\right] = 0 \quad \text{on } \partial\Omega.$$

It follows by the assumption $\mathbf{J}^a = J^a(\underline{x}, t)\mathbf{e}_3$ and the definitions of $\mathbf{J}$ and $\mathbf{J}^e$ that $\hat{\mathbf{J}} = (0, 0, J)$, where $J \in K$. From this last set of equations and using the assumption that $\hat{\mathbf{H}}$ lies in the (x_1, x_2) plane, we see that there exists a scalar potential $q(\underline{x}, t)$, $\underline{x} \in \mathbb{R}^2$, for $\hat{\mathbf{H}}$ such that $\hat{\mathbf{H}} = (\nabla^\perp q, 0)$.

Furthermore, q satisfies

$$-\Delta q = J \quad \text{in } \mathbb{R}^2 \tag{2.5a}$$

and

$$|\nabla^\perp q| \to 0 \quad \text{as } |\underline{x}| \to \infty. \tag{2.5b}$$

Imposing the condition

$$\int_\Omega q d\underline{x} = 0, \tag{2.5c}$$

the problem (2.5a)–(2.5c) is known to have a unique solution, which we denote by

$$q = GJ.$$

Similarly, there exists a scalar potential q^e for $\mathbf{H}^e$, unique up to a constant function in time, such that

$$\mathbf{H}^e = (\nabla^\perp q^e, 0), \quad \nabla^\perp q^e \to \underline{H}^a \quad \text{as } |\underline{x}| \to \infty.$$

We may rewrite (2.1) in the form

$$\nabla^\perp \left(\frac{\partial q}{\partial t} + \rho J \right) = -\nabla^\perp \frac{\partial q^e}{\partial t}$$
$$\Rightarrow \nabla^\perp \left(\frac{\partial GJ}{\partial t} + \rho J \right) = -\nabla^\perp \frac{\partial q^e}{\partial t}.$$

Hence, fixing q^e, we obtain

$$\frac{\partial GJ}{\partial t} + \rho J - \lambda(t) = -\frac{\partial q^e}{\partial t} := f,$$

where λ is an arbitrary function of time.

Multiplying the above equation by $\eta - J$ for $\eta \in K$, integrating over Ω, and using the fact that $(1, \eta - J) = 0$, we have

$$\left(\frac{\partial GJ}{\partial t}, \eta - J \right) = (f, \eta - J) - (\rho J, \eta - J).$$

Since $\rho(r) \in \beta(|J|)$ and $|\eta| \le J_c$, we have

$$(\rho J, \eta - J) \le 0.$$

Hence, we obtain problem (**P**).

The above formulation of Bean's model is the basis of the numerical algorithm proposed by Prigozhin in [9, 11] using an explicit formula for the integral operator G. The discretization is then based upon piecewise constant finite elements. This approach leads to a dense matrix. In the following we use the finite element method to approximate G but never form the matrix associated with this finite element approximation. Whenever G is required we use an elliptic solve. In this paper an error bound is proved and an iterative method is proposed for the resulting discrete variational inequality. For an engineering application of (**P**), see [2, 3].

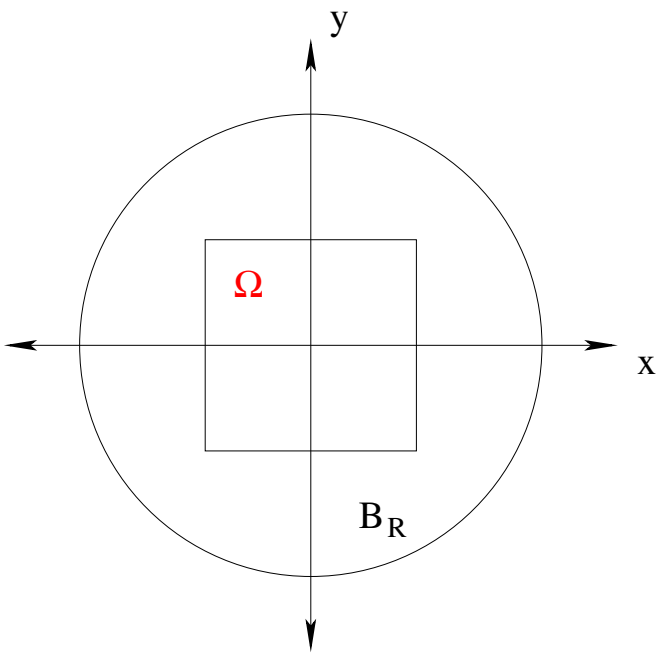


FIG. 2.1. *Reduction in the domain of the problem.*

2.2. Reduction to a bounded domain. From a computational viewpoint, discretizing the whole of $\mathbb{R}^2$ in order to find the operator G is not practical. A natural approach is to restrict the problem to a large bounded region B_R containing Ω and to write an exact boundary condition for Gv on ∂B_R.

Consider the situation where Ω is embedded in a large circle B_R of radius R; see Figure 2.1.

We consider a Dirichlet-to-Neumann mapping which relies on the harmonic property of Gv outside B_R and the boundedness of $\nabla^\perp Gv$ in $L^2(\mathbb{R}^2)$. This method of truncating a problem defined on an infinite domain to one defined on a finite domain is described in [6]. An overview is given here.

For $w \in H^{1/2}(\partial B_R)$ let z solve

$$-\Delta z = 0 \quad \text{in} \quad \mathbb{R}^2 \backslash \overline{B_R}, \tag{2.6}$$

$$z = w \quad \text{on} \quad \partial B_R, \tag{2.7}$$

$$\nabla z \in L^2(\mathbb{R}^2 \backslash \overline{B_R}). \tag{2.8}$$

It follows that we have a Fourier expansion

$$z(r,\theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos(k\theta) + b_k \sin(k\theta)) R^k r^{-k},$$

where a_k, b_k are the Fourier coefficients for $w = w(\theta)$ on ∂B_R.

Differentiating with respect to r and letting $r \to R$ gives

$$\frac{\partial z}{\partial r}(R,\theta) = -\sum_{k=1}^{\infty} \frac{k}{R}(a_k \cos(k\theta) + b_k \sin(k\theta)). \tag{2.9}$$

Since

$$a_k = -\frac{1}{k\pi}\int_0^{2\pi} \frac{\partial w}{\partial \varphi}\sin(k\varphi)d\varphi \quad \text{and} \quad b_k = \frac{1}{k\pi}\int_0^{2\pi} \frac{\partial w}{\partial \varphi}\cos(k\varphi)d\varphi,$$

substituting into (2.9) gives the relation

$$\left.\frac{\partial z}{\partial r}\right|_{\partial B_R}(\theta) = \mathcal{B}(w)(\theta) := -\sum_{k=1}^{\infty}\frac{1}{R\pi}\int_0^{2\pi}\frac{\partial w}{\partial \varphi}\sin(k(\varphi-\theta))d\varphi. \tag{2.10}$$

Let z be a solution of (2.6)–(2.8) for w being the trace of Gv on ∂B_R. Taking $\mathcal{B}(\cdot)$ to be defined as above, it follows that Gv solves the following Neumann problem defined on B_R:

$$-\Delta Gv = v \quad \text{in } B_R, \frac{\partial Gv}{\partial \nu} = \mathcal{B}(Gv) \quad \text{on } \partial B_R. \tag{2.11}$$

Multiplying (2.11) by a test function $\eta \in H^1(B_R)$, integrating over B_R, and then integrating by parts yield the equivalent variational problem:

For $v \in \mathcal{F}_0$, find $Gv \in H^1(B_R)$ such that

$$(Gv, 1) = 0, \quad a(Gv,\eta) + b(Gv,\eta) = (v,\eta) \quad \forall\ \eta \in H^1(B_R), \tag{2.12}$$

where for $\xi, \eta \in H^1(B_R)$,

$$a(\xi,\eta) := \int_{B_R}\nabla\xi\cdot\nabla\eta\ d\underline{x} \quad \text{and} \quad b(\xi,\eta) := \int_{\partial B_R}\mathcal{B}(\xi)\eta\ dS.$$

The existence of a unique solution Gv to this variational problem is easily proved.

We define

$$A(\xi,\eta) := a(\xi,\eta) + b(\xi,\eta) \quad \forall\ \xi,\eta \in H^1(B_R) \tag{2.13}$$

together with the seminorm and norm

$$|\eta|_A^2 := A(\eta,\eta) \quad \forall\ \eta \in H^1(B_R), \quad ||\eta||_{A^{-1}}^2 := |G\eta|_A^2 \quad \forall\ \eta \in \mathcal{F}_0. \tag{2.14}$$

Henceforth we define the L^2 norm and the H^1 norm and seminorm over X respectively by

$$||\eta||_{0,X}^2 = \int_X |\eta|^2 d\underline{x}, \quad ||\eta||_{1,X}^2 = \int_X (|\eta|^2 + |\nabla\eta|^2)d\underline{x} \quad \text{and} \quad |\eta|_{1,X}^2 = \int_X |\nabla\eta|^2 d\underline{x}.$$

From [6] we have that A is continuous with respect to the H^1 norm; that is, for all $\xi, \eta \in H^1(B_R)$

$$|A(\xi,\eta)| \ \le\ C||\xi||_{1,B_R}||\eta||_{1,B_R}. \tag{2.15}$$

Using (2.12)–(2.15), we have the following useful result:

$$(\xi,\eta) = A(G\xi,\eta) \le |G\xi|_A|\eta|_A \le C||\xi||_{A^{-1}}||\eta||_{1,B_R} \quad \forall\ \eta \in H^1(B_R),\ \xi \in \mathcal{F}_0. \tag{2.16}$$

3. Finite element approximation. In this section we consider a finite element approximation of (**P**) under the following assumptions on the partitioning:

(A) Let Ω be a polygon and let T_h^1 be a quasi-uniform partitioning of Ω into disjoint open simplices κ with $h_\kappa := \operatorname{diam}(\kappa)$ and $h := \max_{\kappa\in T_h^1} h_\kappa$, so that $\overline{\Omega} = \cup_{\kappa\in T_h^1}\overline{\kappa}$.

(B) Let T_h^2 be a partitioning of B_R into disjoint open elements $\kappa \in T_h^2$ such that
- $\cup_{\kappa\in T_h^2}\overline{\kappa} = \overline{B}_R$,
- either $\kappa \cap \Omega$ is empty or $\kappa \in T_h^1$,
- if $\overline{\kappa} \cap \partial B_R = \emptyset$, or a point, then κ is a simplex; otherwise, κ is a three-sided element with a curved edge on ∂B_R.

Associated with T_h^1 is the finite element space of continuous piecewise linear functions on Ω such that

$$S_h^1 = \left\{\chi \in C(\overline{\Omega}) : \chi|_\kappa \quad \text{is linear } \forall\ \kappa \in T_h^1\right\} \subset H^1(\Omega).$$

Similarly associated with T_h^2 is the finite element space of continuous functions on B_R such that

$$S_h^2 = \left\{\chi \in C(\overline{B_R}) : \chi|_\kappa \quad \text{is linear } \ \forall\ \kappa \in T_h^2\right\} \subset H^1(B_R).$$

The discrete inner product $(\cdot,\cdot)^h$ is defined by numerical integration in the following way.

Associated with each node $\underline{x}_i$, $i = 1, 2, \dots, M$, of S_h^1 we have a lumped mass matrix value $\mathcal{M}_i > 0$. We now introduce a discrete semi-inner product on $L^2(\Omega)$, defined by

$$(\eta_1, \eta_2)^h := \int_\Omega \Pi^h(\eta_1\eta_2)d\underline{x} = \sum_{i=1}^{M} \mathcal{M}_i(\eta_1\eta_2)(\underline{x}_i), \tag{3.1}$$

where $\Pi^h : C(\overline{\Omega}) \to S_h^1$ is the standard linear interpolation operator.

We introduce the $L^2(\Omega)$ projection operator $Q^h : L^2(\Omega) \to S_h^1$ such that

$$\left(Q^h\eta, \chi\right)^h = (\eta, \chi) \quad \forall\ \chi \in S_h^1. \tag{3.2}$$

Similar to (2.12) we introduce the operator $G^h : \mathcal{F}_0 \to \mathcal{V}_h := \{v_h \in S_h^2 : (v_h, 1) = 0\}$ such that

$$A(G^h\xi, \chi) = (Q^h\eta, \chi)^h \quad \forall\ \xi \in \mathcal{F}_0,\ \chi \in S_h^2, \tag{3.3}$$

and we define the norm

$$||\eta||^2_{A^{-h}} := \left|G^h\eta\right|^2_A \quad \forall \eta \in \mathcal{F}_0.$$

It follows from (3.2) and (3.3) similarly to (2.16) that

$$(\xi, \chi) \ \le\ C||\xi||_{A^{-h}}||\chi||_{1,B_R} \quad \forall\ \chi \in S_h^2,\ \xi \in \mathcal{F}_0. \tag{3.4}$$

From [6] we have the following useful results:

$$\left\|(G - G^h)\eta\right\|_{0,B_R} \ \le\ C_R h^2 \left\|\eta\right\|_{0,\Omega} \qquad \forall\ \eta \in \mathcal{F}_0, \tag{3.5}$$

$$\left|(G-G^h)\eta\right|_{1,B_R} \;\le\; C_R h\,\|\eta\|_{0,\Omega} \qquad \forall\ \eta\in\mathcal{F}_0, \tag{3.6}$$

$$\left|G^h\chi\right|_A \;\le\; |G\chi|_A \qquad \forall\ \chi\in S_h^1, \tag{3.7}$$

and using (3.6) it follows that

$$|G\chi|_A \le C\left|G^h\chi\right|_A \qquad \forall\ \chi\ \in S_h^1. \tag{3.8}$$

Lastly from (2.16) and an inverse inequality we have the following for all $\chi\in\mathcal{F}_0\cap S_h^2$:

$$\|\chi\|_{0,\Omega}^2 \;\le C\ |G\chi|_A||\chi||_{1,\Omega} \;\le\; Ch^{-1}|G\chi|_A\,\|\chi\|_{0,\Omega}$$

$$\Rightarrow \|\chi\|_{0,\Omega}\le Ch^{-1}|G\chi|_A. \tag{3.9}$$

LEMMA 3.1. *We have*

$$\left|G(\eta-Q^h\eta)\right|_A \le Ch\,\|\eta\|_{0,\Omega} \qquad \forall\ \eta\in\mathcal{F}_0. \tag{3.10}$$

Proof. Using (2.12), (3.2), (3.3), (3.5), Hölder's inequality, and the well-known estimate

$$\left|(\xi,\eta)-(\xi,\eta)^h\right| \le Ch^2|\xi|_{1,\Omega}|\eta|_{1,\Omega}\le Ch|\xi|_{1,\Omega}\,\|\eta\|_{0,\Omega} \quad \forall\ \xi,\eta\in S_h^1, \tag{3.11}$$

we have the following for all $\eta\in\mathcal{F}_0$:

$$\begin{aligned}
\left|G(\eta-Q^h\eta)\right|_A^2 &= A(G(\eta-Q^h\eta),G(\eta-Q^h\eta))\\
&= (G(\eta-Q^h\eta),\eta-Q^h\eta)\\
&= ((G-G^h)(\eta-Q^h\eta),\eta-Q^h\eta)\\
&\quad +(G^h(\eta-Q^h\eta),Q^h\eta)^h-(G^h(\eta-Q^h\eta),Q^h\eta)\\
&\le \left\|(G-G^h)(\eta-Q^h\eta)\right\|_{0,\Omega}\left\|\eta-Q^h\eta\right\|_{0,\Omega}\\
&\quad +Ch\left|G^h(\eta-Q^h\eta)\right|_{1,\Omega}\left\|Q^h\eta\right\|_{0,\Omega}\\
&\le Ch^2\left\|\eta-Q^h\eta\right\|_{0,\Omega}^2+Ch\left|G^h(\eta-Q^h\eta)\right|_A\left\|Q^h\eta\right\|_{0,\Omega}.
\end{aligned}$$

The result follows by noting (3.7) and using Young's inequality. ∎

Finally we introduce a finite element approximation of (**P**):

(**P**$_h$) Find $J_h\in K_h$ such that $J_h(\cdot,0)=Q^hJ_0$ and

$$\left(\frac{\partial}{\partial t}G^hJ_h,\chi-J_h\right)\ge(f,\chi-J_h) \qquad \forall\ \chi\in K_h, \tag{3.12}$$

where

$$K_h:=\Big\{\chi\in S_h^1:|\chi|\le J_c,\ \ (\chi,1)=0\Big\}.$$

Remark 3.1. Let the assumptions (A) hold. Then there exists a unique solution J_h to (**P**$_h$) such that

$$||J_h||_{L^\infty(0,T;L^\infty(\Omega))}+\left\|\frac{\partial}{\partial t}GJ_h\right\|_{L^\infty(0,T;A)}\le C. \tag{3.13}$$

LEMMA 3.2. *The unique solutions of* $(\mathbf{P}_h)$ *and* $(\mathbf{P}_h)$ *satisfy*

$$||J - J_h||^2_{L^\infty(0,T;A^{-1})} \leq Ch. \tag{3.14}$$

Proof. Since $J_h \in K$ using (1.1), (2.16), and (3.12) we have

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|J - J_h\|^2_{A^{-1}} &= \left(\frac{\partial}{\partial t}G(J - J_h), J - J_h\right) \\
&\leq (f, J - J_h) - \left(\frac{\partial}{\partial t}GJ_h, J - J_h\right) \\
&= (f, J - J_h) - \left(\frac{\partial}{\partial t}G^hJ_h, Q^hJ - J_h\right) - \left(\frac{\partial}{\partial t}G^hJ_h, J - Q^hJ\right) \\
&\quad -\left(\frac{\partial}{\partial t}(G - G^h)J_h, J - J_h\right) \\
&\leq (f, J - J_h) - \left(f, Q^hJ - J_h\right) - \left(\frac{\partial}{\partial t}G^hJ_h, J - Q^hJ\right) \\
&\quad -\left(\frac{\partial}{\partial t}(G - G^h)J_h, J - J_h\right) \\
&= \left(f - \frac{\partial}{\partial t}G^hJ_h, J - Q^hJ\right) - \left(\frac{\partial}{\partial t}(G - G^h)J_h, J - J_h\right) \\
&\leq \left|f - \frac{\partial}{\partial t}G^hJ_h\right|_A \|J - Q^hJ\|_{A^{-1}} + \left\|\frac{\partial}{\partial t}(G - G^h)J_h\right\|_{0,\Omega} \|J - J_h\|_{0,\Omega}.
\end{aligned}$$

Using the above inequality together with (1.4), (3.9), (3.5), (3.10), and (3.13) yields

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|J - J_h\|^2_{A^{-1}} &\leq Ch + Ch^2\left\|\frac{\partial}{\partial t}J_h\right\|_{0,\Omega}\|J - J_h\|_{0,\Omega} \\
&\leq Ch + Ch\left|\frac{\partial}{\partial t}GJ_h\right|_A.
\end{aligned}$$

Integrating from 0 to t and using (3.13) gives the required result. ∎

Remark 3.2. This is a suboptimal error bound because of the error term $(f - \frac{\partial}{\partial t}G^hJ_h, J - P^hJ)$, arising due to the variational inequality, which only gives $\mathcal{O}(h)$ because of the lack of H^1 regularity of J.

4. Fully discrete model. In this section we consider a fully discrete discretization of $(\mathbf{P})$. Setting $N\Delta t = T$ and $t_n : n\Delta t$ for $n = 0 \to N$ and for any $\chi_h \in S^1_h$, $n = 0, 1, \ldots,$ we set

$$\delta_t\chi^n = \frac{\chi^n - \chi^{n-1}}{\Delta t}.$$

We consider the following fully discrete discretization of $(\mathbf{P})$:

$(\mathbf{P}_{h,\Delta t})$ For $n = 1 \to N$, find $J^n_h \in K_h$ such that $J^0_h = Q^hJ_0$ and

$$\left(G^h(\delta_tJ^n_h), \chi - J^n_h\right) \geq (f^n, \chi - J^n_h) \quad \forall\ \chi\ \in K_h, \tag{4.1}$$

where $f^n := f(\cdot, t_n)$.

LEMMA 4.1. *Let the assumptions* (A) *hold. Then for* $n = 1 \to N$ *there exists a unique solution* J_h^n *to* $(\mathbf{P}_{h,\Delta t})$ *such that*

$$\max_{n=1\to N} ||\delta_t J_h^n||_{A^{-h}}^2 \le C. \tag{4.2}$$

Proof. Existence and uniqueness for (4.1) are standard. Setting $\chi = J_h^{n-1}$ in (4.1), dividing by Δt, and noting (1.4) and (3.4) gives

$$\left(G^h \delta_t J_h^n, \delta_t J_h^n\right) \le (f^n, \delta_t J_h^n)$$

$$\Rightarrow ||\delta_t J_h^n||_{A^{-h}}^2 \le ||f^n||_{1,B_R} ||\delta_t J_h^n||_{A^{-h}}$$

which together with (1.4) yields (4.2). ∎

Before we derive an error bound on the solutions of $(\mathbf{P}_h)$ and $(\mathbf{P}_{h,\Delta t})$ we introduce some useful notation. For $n \ge 1$ we set

$$J_{h,\Delta t}(t) := \frac{t - t_{n-1}}{\Delta t} J_h^n + \frac{t_n - t}{\Delta t} J_h^{n-1}, \quad f_{\Delta t}(t) := \frac{t - t_{n-1}}{\Delta t} f^n + \frac{t_n - t}{\Delta t} f^{n-1} \quad \forall t \in [t_{n-1}, t_n], \tag{4.3}$$

and

$$\hat{J}_{h,\Delta t}(t) := J_h^n, \quad \hat{f}_{\Delta t}(t) := f^n \quad \forall\, t \in (t_{n-1}, t_n]. \tag{4.4}$$

From (4.3) and (4.4) it follows that for a.e. $t \in (0, T)$,

$$J_{h,\Delta t} - \hat{J}_{h,\Delta t} = -(t_n - t)\frac{\partial}{\partial t} J_{h,\Delta t}, \quad f_{\Delta t} - \hat{f}_{\Delta t} = -(t_n - t)\frac{\partial}{\partial t} f_{\Delta t}. \tag{4.5}$$

We also introduce for $t \in (0, T)$,

$$\mathcal{R}(t) := \left(\hat{f}_{\Delta t} - \frac{\partial}{\partial t} G^h J_{h,\Delta t}, \hat{J}_{h,\Delta t} - J_{h,\Delta t}\right)$$

$$= (t_n - t)\left(\hat{f}_{\Delta t} - \frac{\partial}{\partial t} G^h J_{h,\Delta t}, \frac{\partial}{\partial t} J_{h,\Delta t}\right), \quad t \in (t_{n-1}, t_n], \tag{4.6}$$

and for $t \in (0, T]$,

$$\mathcal{D}(t) := \mathcal{D}^n := -\left(G^h(\delta_t J_h^n), \delta_t J_h^n\right) + \left(G^h(\delta_t J_h^{n-1}), \delta_t J_h^n\right), \quad t \in (t_{n-1}, t_n], \tag{4.7}$$

with J_h^{-1} satisfying (4.1) and

$$\left|\left|\frac{J^0 - J^{-1}}{\Delta t}\right|\right|_{A^{-h}}^2 = \left(G^h(\delta_t J_h^0), \delta_t J_h^0\right) \le C. \tag{4.8}$$

LEMMA 4.2. *For a.e.* $t \in (0, T)$ *we have that*

$$\mathcal{R}(t) \le (t_n - t)\left[\mathcal{D}(t) + \Delta t\left(\frac{\partial}{\partial t} f_{h,\Delta t}, \frac{\partial}{\partial t} J_{h,\Delta t}\right)\right], \quad t \in (t_{n-1}, t_n], \tag{4.9}$$

and

$$\int_0^T \mathcal{R}(t)dt \le C(\Delta t)^2. \tag{4.10}$$

Proof. Setting $\chi = J_h^n$ in (4.1) for $n = n-1$ and using the definitions of $\mathcal{D}(t)$ and $\mathcal{R}(t)$, we have

$$\begin{aligned}\mathcal{R}(t) &= -(t_n - t)\left(\frac{\partial}{\partial t}G^h J_{h,\Delta t}, \frac{\partial}{\partial t}J_{h,\Delta t}\right) + (t_n - t)\left(\hat{f}_{\Delta t}(t) - \hat{f}_{\Delta t}(t-\Delta t), \frac{\partial}{\partial t}J_{h,\Delta t}\right)\\ &\quad +(t_n - t)\left(\hat{f}_{\Delta t}(t-\Delta t), \frac{\partial}{\partial t}J_{h,\Delta t}\right)\\ &\le (t_n - t)\mathcal{D}^n + (t_n - t)\left(\hat{f}_{\Delta t}(t) - \hat{f}_{\Delta t}(t-\Delta t), \frac{\partial}{\partial t}J_{h,\Delta t}\right),\end{aligned}$$

and (4.9) follows by using (4.3). We now integrate (4.9) from 0 to t and use (1.4), (3.4), and (4.2) to obtain

$$\int_0^T \mathcal{R}(t)dt = \sum_{n=1}^N \mathcal{D}^n \int_{t_{n-1}}^{t_n} (t_n - t)dt + \int_0^T \Delta t\left(\frac{\partial}{\partial t}f_{\Delta t}, \frac{\partial}{\partial t}J_{h,\Delta t}\right)(t_n - t)dt$$

$$\le \sum_{n=1}^N \frac{(\Delta t)^2}{2}\mathcal{D}^n + (\Delta t)^2 \int_0^T \left|\left|\frac{\partial}{\partial t}f_{\Delta t}\right|\right|_{1,B_R}\left|\left|\frac{\partial}{\partial t}J_{h,\Delta t}\right|\right|_{A^{-h}} dt \tag{4.11}$$

$$\le \sum_{n=1}^N \frac{(\Delta t)^2}{2}\mathcal{D}^n + C(\Delta t)^2. \tag{4.12}$$

To bound the first term on the right-hand side we use the identity

$$2\big(G^h(a-b), a\big) = \big(G^h a, a\big) - \big(G^h b, b\big) + \big(G^h(a-b), a-b\big)$$

to obtain

$$2\mathcal{D}^n \le \big(G^h(\delta_t J_h^{n-1}), \delta_t J_h^{n-1}\big) - \big(G^h(\delta_t J_h^n), \delta_t J_h^n\big).$$

Summing the above inequality from $n = 1 \to N$ and using (3.4) and (4.8), we have

$$\begin{aligned}2\sum_{n=1}^N \mathcal{D}^n &\le \big(G^h(\delta_t J_h^0), \delta_t J_h^0\big) - \big(G^h(\delta_t J_h^N), \delta_t J_h^N\big)\\ &\le \big(G^h(\delta_t J_h^0), \delta_t J_h^0\big) \le C.\end{aligned} \tag{4.13}$$

Using (4.13) in (4.12), we conclude (4.10). ∎

LEMMA 4.3. *The unique solutions of* $(\mathbf{P}_h)$ *and* $(\mathbf{P}_{h,\Delta t})$ *satisfy*

$$||J_h - J_{h,\Delta t}||_{L^\infty(0,T;A^{-1})} \le C\Delta t. \tag{4.14}$$

Proof. Setting $\chi = J_h$ in (4.1) and $\chi = J_{h,\Delta t}$ in (3.12) and adding the resulting inequalities gives

$$\left(\frac{\partial}{\partial t}G^h(J_{h,\Delta t} - J_h), J_{h,\Delta t} - J_h\right) \le \left(\hat{f}_{\Delta t} - f, J_{h,\Delta t} - J_h\right) + \left(\frac{\partial}{\partial t}G^h J_{h,\Delta t} - \hat{f}_{\Delta t}, J_{h,\Delta t} - \hat{J}_{h,\Delta t}\right).$$

Noting (3.4) and (4.6), we obtain

$$\frac{1}{2}\frac{\partial}{\partial t}||J_{h,\Delta t} - J_h||^2_{A^{-h}} \le \left|\left|\hat{f}_{\Delta t} - f\right|\right|_{1,B_R} ||J_{h,\Delta t} - J_h||_{A^{-h}} + \mathcal{R}.$$

From Lemma 3.6 in [8] we conclude that

$$\max_{t\in[0,T]} ||J_{h,\Delta t} - J_h||_{A^{-h}} \le \left(||J_{h,\Delta t}(0) - J_h(0)||^2_{A^{-h}} + \int_0^T \mathcal{R}(t)dt\right)^{1/2} + \int_0^T \left|\left|\hat{f}_{\Delta t} - f\right|\right|_{1,B_R} dt,$$

and noting (1.4), (3.8), (4.4), and (4.10) yields the required result. ∎

Finally we have our main result.

THEOREM 4.4. *Let the assumptions* (A) *hold. Then the unique solutions* $\{J_h^n\}_{n=0}^N$ *to* $(\mathbf{P}_{h,\Delta t})$ *and* J *to* $(\mathbf{P})$ *satisfy*

$$||J - J_{h,\Delta t}||_{L^\infty(0,T;A^{-1})} \le C(T)(h^{1/2} + \Delta t).$$

Proof. The desired result follows directly from (3.14) and (4.14). ∎

Recalling that $\underline{\hat{H}} = \nabla^\perp GJ$ and setting $\underline{\hat{H}}_h^n = \underline{\hat{H}}_h(t_n) := \nabla^\perp G^h J_h(t_n)$ and $\underline{\hat{H}}_{h,\Delta t}$ as in (4.3), we conclude the following.

COROLLARY 4.1. *The error between the magnetic field* $\underline{H}$ *and its approximation* $\underline{H}_{h,\Delta t}$ *is*

$$\left\|\underline{H} - \underline{H}_{h,\Delta t}\right\|_{L^\infty(0,T;L^2(B_R))} \le C(T)(h^{1/2} + \Delta t).$$

5. Algorithm for solving $(\widehat{\mathbf{P}}_{\mathbf{h},\Delta\mathbf{t}})$. In the numerical simulations presented in section 6 we solve the following approximation of $(\widehat{\mathbf{P}}_{h,\Delta t})$:

$(\widehat{\mathbf{P}}_{h,\Delta t})$ For $n = 1 \to N$, find $J_h^n \in K_h$ such that $J_h^0 = Q^h J_0$ and

$$\left(\hat{G}^h\left(\delta_t \widehat{J}_h^n\right), \chi - \widehat{J}_h^n\right)^h \ge \left(f^n, \chi - \widehat{J}_h^n\right)^h \quad \forall\, \chi \in K_h, \tag{5.1}$$

where $f^n := f(\cdot, t_n)$ and the operator $\hat{G}^h : \mathcal{V}_h \to \mathcal{V}_h$ is such that

$$A(\hat{G}^h\xi, \chi) = (\xi, \chi)^h \quad \forall\, \xi \in \mathcal{V}_h, \chi \in S_h^2.$$

Below we give an algorithm for solving $(\widehat{\mathbf{P}}_{h,\Delta t})$. See [5] for an account of iterative methods for solving discrete variational inequalities.

Reformulating $(\widehat{\mathbf{P}}_{h,\Delta t})$ gives the following problem:

Given $J_h^0 = P^h J_0$, for $n = 1 \to N$, find $J_h^n \in K_h$ and $\lambda^n \in \mathbb{R}$ such that $|J_h^n| \le J_c$, $(J_h^n, 1)^h = 0$, and

$$\left(\hat{G}^h \widehat{J}_h^n, \chi - \widehat{J}_h^n\right)^h \ge \left(\Delta t f^n + \lambda^n + \hat{G}^h \widehat{J}_h^{n-1}, \chi - \widehat{J}_h^n\right)^h \quad \forall\, \chi \in S_h^1 \text{ such that } |\chi| \le J_c.$$

Setting $\Lambda_h^n := \Delta t f^n + \hat{G}^h \widehat{J}_h^{n-1}$, the above problem is equivalent to the following problem:

Find $\widehat{J}_h^n \in S_h^1$ such that $(J_h^n, 1)^h = 0$ and

$$\hat{G}^h \widehat{J}_h^n - \Lambda_h^n - \lambda^n + \beta_h^n = 0$$

$$\Leftrightarrow \frac{1}{\mu}\widehat{J}_h^n + \hat{G}^h \widehat{J}_h^n - \lambda^n + \beta_h^n = \Lambda_h^n + \frac{1}{\mu}\widehat{J}_h^n, \tag{5.2}$$

where $\beta_h^n(\underline{x}_i) \in \beta(J_h^n(\underline{x}_i))$.

We solve (5.2) iteratively using a splitting algorithm of Lions and Mercier [7]. Let $\widehat{J}_h^0$ be given; for fixed μ we construct $J_h^{n,k+1}$, $\beta_h^{n,k+1}$, and $\lambda^{n,k+1}$ iteratively by solving for $k \ge 0$:

$$\frac{1}{\mu}\widehat{J}_h^{n,k+1/2} + \hat{G}^h \widehat{J}_h^{n,k+1/2} = \Lambda_h^n + \frac{1}{\mu}\widehat{J}_h^{n,k} - \beta_h^{n,k} + \lambda^{n,k} := \tilde{\Lambda}_h^{n,k}, \tag{5.3}$$

$$\frac{1}{\mu}\widehat{J}_h^{n,k+1} - \lambda^{n,k+1} + \beta_h^{n,k+1} = \Lambda_h^n + \frac{1}{\mu}\widehat{J}_h^{n,k+1/2} - \hat{G}^h \widehat{J}_h^{n,k+1/2} := F^{n,k+1/2}, \tag{5.4}$$

$$(J_h^{n,k+1}, 1)^h = 0,$$

where $\beta_h^{n,k+1}(\underline{x}_i) \in \beta(J_h^{n,k+1}(\underline{x}_i))$. To solve (5.3) we use (3.3) to rewrite

$$\frac{1}{\mu}\left(\bar{J}_h^{k+1/2}, \chi\right)^h + \left(\hat{G}^h \bar{J}_h^{k+1/2}, \chi\right)^h = \left(\tilde{\Lambda}_h^{n,k}, \chi\right)^h$$

as

$$\frac{1}{\mu}A(\hat{G}^h \bar{J}_h^{k+1/2}, \chi) + \left(\hat{G}^h \bar{J}_h^{k+1/2}, \chi\right)^h = \left(\tilde{\Lambda}_h^{n,k}, \chi\right)^h, \tag{5.5}$$

where $\bar{J}_h^{k+1/2} = \widehat{J}_h^{k+1/2} - f^n$.

At the ith node we may rewrite (5.4) using the projection

$$J_i^{n,k+1} = P(\mu(F_i^{n,k+1/2} + \lambda^{n,k+1})), \tag{5.6}$$

where

$$P(r) = \begin{cases} J_c & \text{if} \quad r \ge J_c, \\ r & \text{if} \quad |r| < J_c, \\ -J_c & \text{if} \quad r \le -J_c. \end{cases}$$

Noting that $(J_h^{n,k+1}, 1)^h = 0$, $\lambda^{n,k+1}$ solves the equation

$$g(\lambda) = \sum_i \mathcal{M}_i P(\mu(F_i^{n,k+1/2} + \lambda)) = 0. \tag{5.7}$$

To obtain the solution at the $(k+1)$th time step we proceed as follows:

Step 1. Solve (5.5) to obtain $\hat{G}^h \bar{J}_h^{k+1/2}$.

Step 2. Set $\hat{G}^h J_h^{n,k+1/2} = \hat{G}^h \bar{J}_h^{k+1/2} + f^n$.

Step 3. Use (5.3) to obtain $\widehat{J}_h^{n,k+1/2}$.

Step 4. Solve (5.6) and (5.7) to obtain $\widehat{J}_h^{n,k+1}$.

Step 5. Use (5.4) to obtain $\beta_h^{n,k+1}$.

Step 6. If $|\widehat{J}_h^{n,k+1} - \widehat{J}_h^{n,k}| \leq \text{tol}$, then set $\widehat{J}_h^n = \widehat{J}_h^{n,k+1}$; else set $\widehat{J}_h^{n,k} = \widehat{J}_h^{n,k+1/2}$ and go to Step 1.

The above procedure is relatively cheap apart from Step 1, which involves the solution of a large sparse matrix problem,

$$A\mathbf{x} = \mathbf{f}, \quad A \in \mathbb{R}^{N \times N}.$$

In general N is required to be large, so that interfaces between critical current and noncritical current can be captured.

Since the matrix A remains fixed throughout time, we could calculate the inverse, or an LU decomposition, of A at the beginning. Due to the nonlocal boundary condition the LU decomposition of this matrix produces $O(N^{3/2})$ entries, and thus, for large problems this is not practical.

Since Step 1 is part of an iteration, we need not solve this problem exactly. In the following section ten or fewer preconditioned GMRES iterations (see [12]) are used with an ILU decomposition used as a preconditioner. This allows large problems to be solved and accurate solutions to be obtained.

Note that Step 4 is well defined for $J_h^{n,k+1}$. It is easily seen that the function g is continuous and monotone piecewise linear which takes negative values for sufficiently negative λ and positive values for sufficiently positive λ, and hence (5.7) has a solution. Furthermore it has only a nonunique solution when $g(\lambda) = 0$ in an interval and in such an interval we observe that $P(F_i^{n,k+1/2} + \lambda)$ is constant for each i; hence the solution of (5.6) is unique. A solution of (5.7) can be found by efficiently by using the bisection method.

In [9, 11] Prigozhin solves the discrete variational inequality associated with the full matrix approximation of G using a projected SOR algorithm. We avoid doing this by using the splitting algorithm defined above in which it is not necessary to form the solution operator G explicitly but its action is calculated by the use of an elliptic solve. That is, (5.3) is implemented using elliptic solve (5.5). The constraint condition is then handled by (5.4), which is easily solved by the projection (5.6) and the Lagrange multiplier equation (5.7).

In practice we do not actually compute $G^h J_h$. Instead we approximate it by replacing the nonlocal boundary inner product $b(\cdot,\cdot)$ with a truncated version $b_M(\cdot,\cdot)$, where

$$b_M(\xi,\eta) = \int_{\partial\Omega} \mathcal{B}_M(\xi)\eta dS$$

with

$$\mathcal{B}_M(w)(\theta) := \int_{\partial\Omega} \sum_{k=1}^{M} \frac{1}{R\pi} \int_0^{2\pi} \frac{\partial w}{\partial \varphi} \sin\left(k(\varphi - \theta)\right) d\varphi.$$

Error analysis for this approximation can be found in [6].

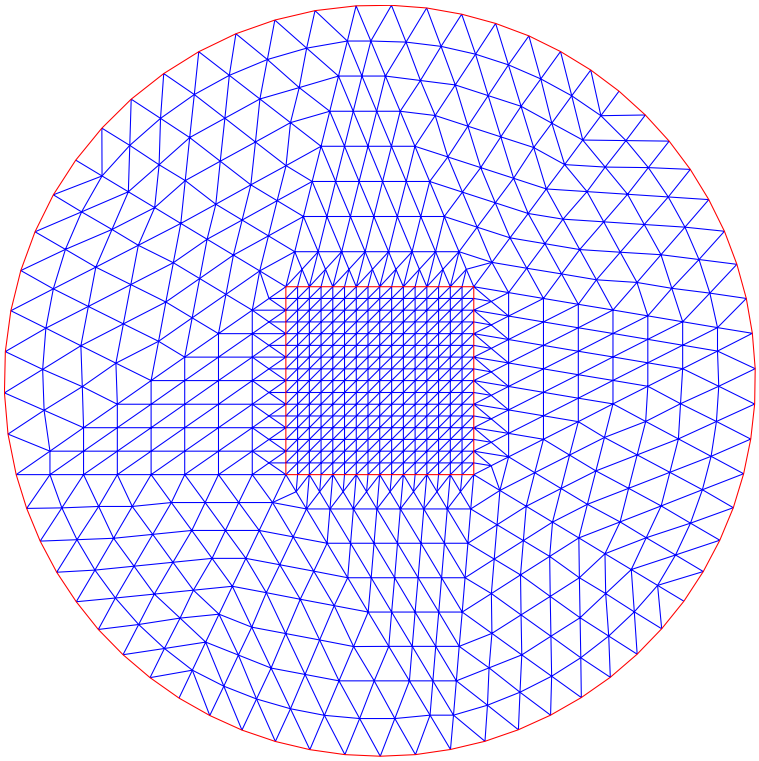


FIG. 6.1. *A typical mesh used for numerical simulations.*

6. Numerical results. In this section we present three sets of computational simulations. All results are calculated on domains of the form seen in Figure 6.1, where the superconductor is located in the square region $(-0.5, 0.5) \times (-0.5, 0.5)$. For all simulations the critical current density is taken to be $J_c = 1$ and the truncated sum for the nonlocal boundary inner product has $M = 5$.

In the first set (Figure 6.2) we take an applied magnetic field

$$\mathbf{H}^a = (0, \min\{t, H_{max}\}, 0)^T$$

for four values of H_{max}. For each value of H_{max} we display steady state solutions of the current density J_h. We see that while the applied magnetic field is increasing, the region in which the current takes critical values also increases.

In the second set of results (Figure 6.3) we apply an oscillating magnetic field of the form

$$\mathbf{H}^a = \left(0, 0.14 \sin \frac{\pi t}{2}, 0\right)^T, \tag{6.1}$$

and we display plots of the current density J_h at times $t = 1, 1.5, 2$, and 2.5.

In Table 6.1 we display the calculated error

$$\left\| \tilde{J}(\cdot, t^*) - J_{h,\Delta t}(\cdot, t^*) \right\|_{A^{-1}}$$

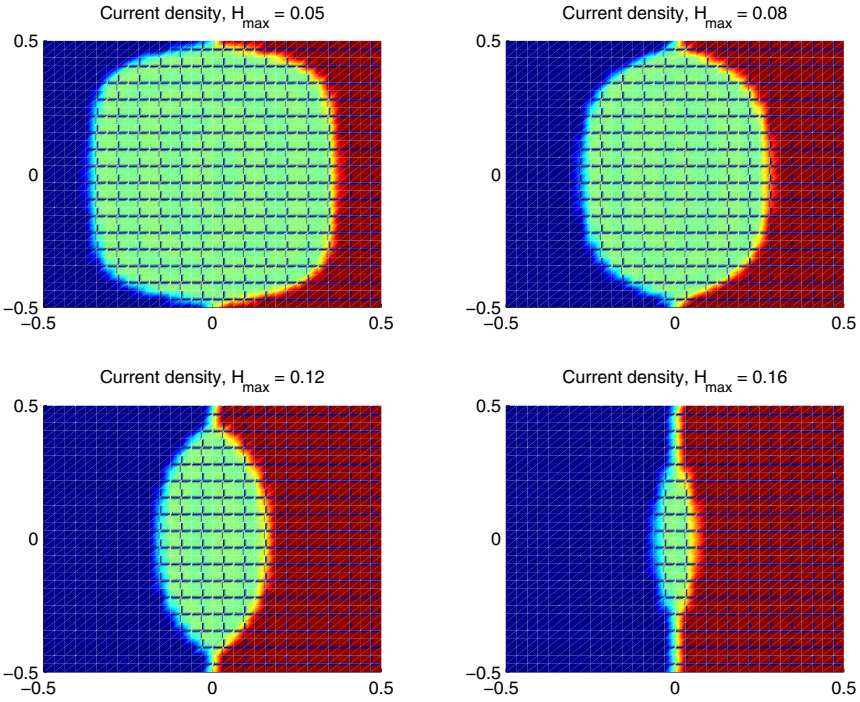


FIG. 6.2. *Steady state solutions:* $\mathbf{H}^a = (0, \min\{t, H_{max}\}, 0)^T$.

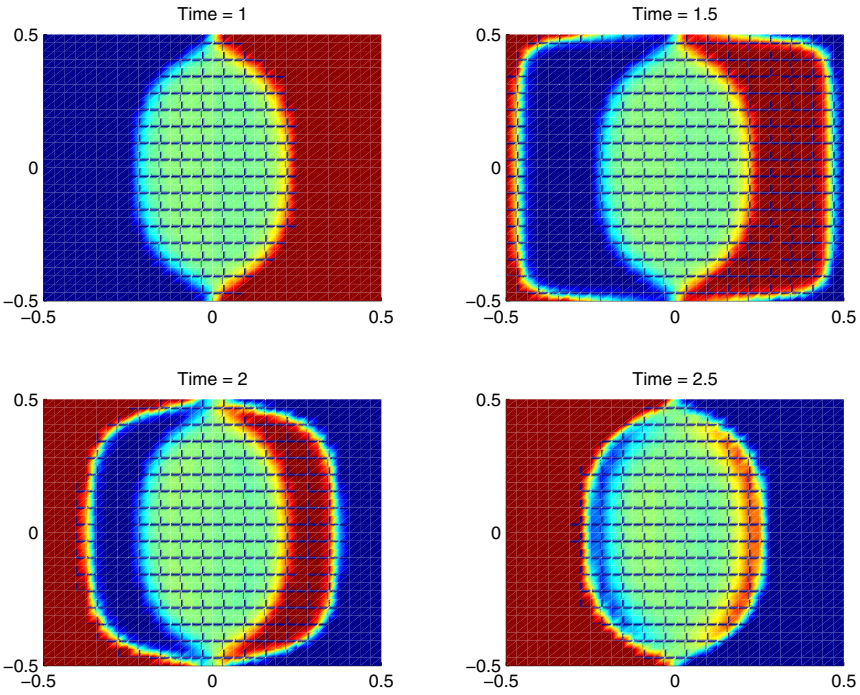


FIG. 6.3. *Current density for oscillating problem:* $\mathbf{H}^a = (0, 0.14 \sin \frac{\pi t}{2}, 0)^T$.

TABLE 6.1
Estimated errors for varying times and meshes.

	$t^* = 1.0$	$t^* = 2.0$	$t^* = 3.0$
$h = 1/8,\ \Delta t = 1/16$	0.0236	0.0255	0.0236
$h = 1/16,\ \Delta t = 1/64$	0.0126	0.0130	0.0126
$h = 1/32,\ \Delta t = 1/256$	0.0063	0.0068	0.0063
$h = 1/64,\ \Delta t = 1/1024$	0.0030	0.0037	0.0030

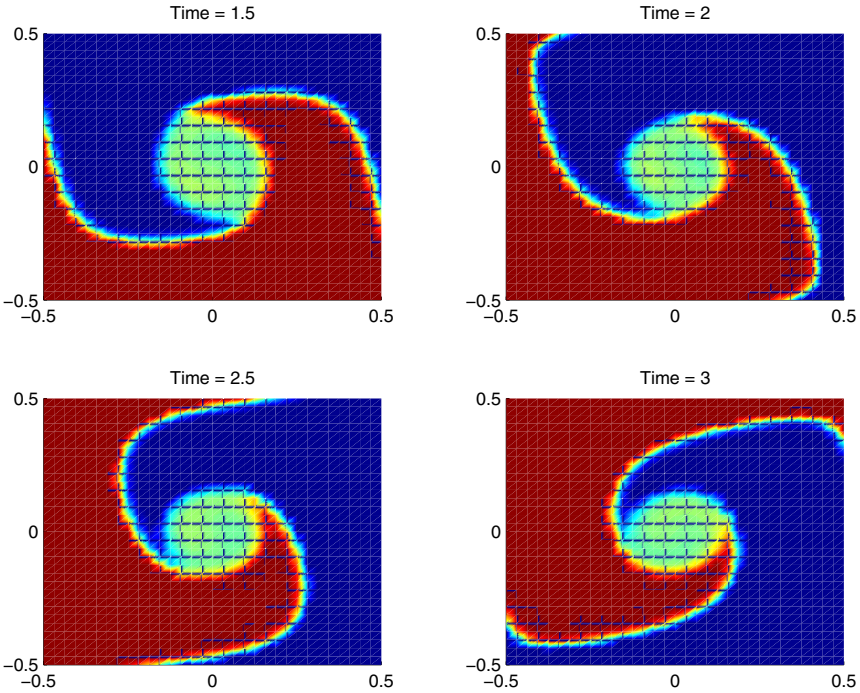


FIG. 6.4. *Current density for rotating problem:* $\mathbf{H}^a = \min\{t, 0.14\}(\sin\frac{\pi t}{2}, \cos\frac{\pi t}{2}, 0)^T$.

for the oscillating magnetic field (6.1). Here $\tilde{J}$ is the solution of $(\widehat{\mathbf{P}}_{h,\Delta t})$ obtained using a fine mesh ($h = 1/256$) and small time step ($\Delta t = 0.001$). These results are consistent with an error of $\mathcal{O}(h)$.

Finally, in Figure 6.4 we take a rotating applied magnetic field of the form

$$\mathbf{H}^a = \min\{t, 0.14\}\left(\sin\frac{\pi t}{2}, \cos\frac{\pi t}{2}, 0\right)^T,$$

and we display plots of the current density J_h at times $t = 1.5, 2, 2.5$, and 3.

REFERENCES

[1] J. F. BLOWEY AND C. M. ELLIOTT, *The Cahn-Hilliard gradient theory for phase separation with nonsmooth free energy Part* II: *Numerical analysis*, European J. Appl. Math., 3 (1992), pp. 147–179.

[2] G. BARNES, M. MCCULLOCH, AND D. DEW-HUGHES, *Computer modelling of type* II *superconductors in applications*, Supercond. Sci. Technol., 12 (1999), pp. 518–522.

[3] G. Barnes, M. McCulloch, and D. Dew-Hughes, *Finite difference modelling of bulk high temperature superconducting cylindrical hysteresis machines*, Supercond. Sci. Technol., 13 (2000), pp. 229–236.

[4] S. J. Chapman, *A mean-field model of superconducting vortices in three dimensions*, SIAM J. Appl. Math., 55 (1995), pp. 1259–1274.

[5] R. Glowinski, J. L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1976.

[6] H. Han and X. Wu, *Approximation of infinite boundary condition and its application to finite element methods*, J. Comput. Math., 3 (1985), pp. 179–192.

[7] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.

[8] R. H. Nochetto, G. Savaré, and C. Verdi, *A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations*, Comm. Pure Appl. Math., 53 (2000), pp. 525–589.

[9] L. Prigozhin, *Analysis of critical state problems in type-*II *superconductivity*, IEEE Trans. on Appl. Supercond., 7 (1997), pp. 3866–3873.

[10] L. Prigozhin, *On the Bean critical state model in superconductivity*, European J. Appl. Math., 7 (1996), pp. 237–247.

[11] L. Prigozhin, *The Bean model in superconductivity: Variational formulation and numerical solution*, J. Comput. Phys., 129 (1996), pp. 190–200.

[12] Y. Saad and M. H. Schultz, *A generalized minimal residual algorithm for solving nonsymmetric linear systems*, J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1342–1356

MULTILEVEL SOLVERS FOR A FINITE ELEMENT DISCRETIZATION OF A DEGENERATE PROBLEM*

SVEN BEUCHLER†

Abstract. In this paper, finite element discretizations of the degenerate operator $-\omega^2(y)u_{xx} - \omega^2(x)u_{yy} = g$ in the unit square are investigated, where the weight function satisfies $\omega(\xi) > 0$ for $\xi \in (0,1]$ and is monotonically increasing. We propose two multilevel methods in order to solve the resulting system of linear algebraic equations. The first method is a multigrid algorithm with line smoother. A proof of the smoothing property is given. The second method is a Bramble–Pasciak–Xu-like preconditioner with line smoother which we call multiple tridiagonal scaling Bramble–Pasciak–Xu preconditioner.

Key words. multigrid, preconditioning

AMS subject classifications. 65N55, 65N22, 65N30, 65F30

DOI. 10.1137/S0036142903428414

1. Introduction.

In this paper, we consider the following problem: Find $u \in H^1_{0,\omega}(\Omega)$ such that

$$a(u,v) := \int_\Omega (\omega(y))^2 u_x v_x + (\omega(x))^2 u_y v_y = \int_\Omega gv =: \langle g, v\rangle \quad \forall v \in H^1_{0,\omega}(\Omega), \tag{1.1}$$

where $H^1_{0,\omega}(\Omega) = \{u \in L^2(\Omega), \omega(x)u_y, \omega(y)u_x \in L^2(\Omega), u \mid_{\partial\Omega} = 0\}$. The domain $\Omega = (0,1)^2$ is the unit square.

Assumption 1.1. The weight function $\omega(\xi) \in L^\infty((0,1))$ is assumed to be monotonically increasing and satisfies $\omega(\xi) > 0$ for $\xi \in (0,1]$.

Remark 1.1. If $\omega(0) = 0$, then the differential operator in (1.1) is not uniformly elliptic in the Sobolev space $H^1_0(\Omega)$, and an estimate of the type

$$a(u,u) \geq \gamma \parallel u \parallel^2_{H^1(\Omega)} \quad \forall u \in H^1_0(\Omega) \tag{1.2}$$

with a constant $\gamma > 0$ is not satisfied.

The integrand on the left-hand side in (1.1) is of the type $(\nabla u)^T D(x,y)\nabla v$ with the diffusion tensor

$$D(x,y) = \begin{bmatrix} \omega^2(y) & 0 \\ 0 & \omega^2(x) \end{bmatrix}. \tag{1.3}$$

Therefore, the matrix D is symmetric and positive definite for all $(x,y) \in \Omega$ but not uniformly positive definite if $\omega(0) = 0$. Moreover, the matrix D is bounded for each $(x,y) \in \Omega$. Such problems are called degenerate problems. In the past, degenerate problems have been considered relatively rarely. One reason is the unphysical behavior of the PDE which is quite unusual in technical applications. One work focusing on this type of PDE is the book of Kufner and Sändig [17]. Nowadays, problems of this type have become more and more popular because there are stochastic PDEs which

*Received by the editors May 23, 2003; accepted for publication (in revised form) March 16, 2004; published electronically December 1, 2004. This work was supported by the DFG Sonderforschungsbereich 393.

http://www.siam.org/journals/sinum/42-3/42841.html

†Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, A-4040 Linz, Austria (sven.beuchler@oeaw.ac.at).

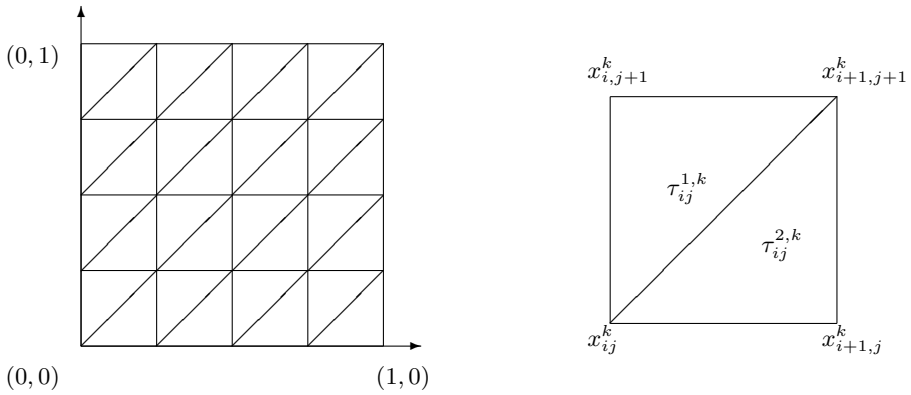


FIG. 1.1. *Mesh for the finite element method (left). Notation within a macroelement $\mathcal{E}^k_{ij}$ (right).*

have a similar structure. An example of a degenerate stochastic PDE is the Black–Scholes PDE in [21]. Moreover, the solver related to the problem of the subdomains embedded in a domain decomposition preconditioner for the p-version of the finite element method can be interpreted as an h-version fem-discretization matrix of (1.1) in the case of the weight function $\omega(\xi) = \xi$. We refer to [6] and [7] for more details.

We discretize problem (1.1) by finite elements. For this purpose, some notation is introduced. Let k be the level of approximation and $n = 2^k$. Let $x^k_{ij} = (\frac{i}{n}, \frac{j}{n})$, where $i, j = 0, \ldots, n$. The domain Ω is divided into congruent, isosceles, right-angled triangles $\tau^{s,k}_{ij}$, where $0 \le i, j < n$ and $s = 1, 2$ (see Figure 1.1). The triangle $\tau^{1,k}_{ij}$ has the three vertices $x^k_{ij}, x^k_{i+1,j+1}$, and $x^k_{i,j+1}$, $\tau^{2,k}_{ij}$ has the three vertices $x^k_{ij}, x^k_{i+1,j+1}$, and $x^k_{i+1,j}$ (see Figure 1.1). Furthermore, let $\mathcal{E}^k_{ij} = \overline{\tau^{1,k}_{ij} \cup \tau^{2,k}_{ij}}$ be the macroelement $\left[\frac{i}{n}, \frac{i+1}{n}\right] \times \left[\frac{j}{n}, \frac{j+1}{n}\right]$. Piecewise linear finite elements are used on the mesh $T_k = \{\tau^{s,k}_{ij}\}^{n-1,n-1,2}_{i=0,j=0,s=1}$. The subspace of piecewise linear functions ϕ^k_{ij} with

$$\phi^k_{ij} \in H^1_0(\Omega),\ \phi^k_{ij}\,|_{\tau^{s,k}_{lm}} \in \mathbb{P}_1(\tau^{s,k}_{lm})$$

is denoted by $\mathbb{V}_k$, where $\mathbb{P}_1$ is the space of polynomials of degree ≤ 1. A basis of $\mathbb{V}_k$ is the system of the usual hat-functions $\{\phi^k_{ij}\}^{n-1}_{i,j=1}$ uniquely defined by

$$\phi^k_{ij}(x^k_{lm}) = \delta_{il}\delta_{jm}$$

and $\phi^k_{ij} \in \mathbb{V}_k$, where δ_{il} is the Kronecker delta. Now, we can formulate the discretized problem.

Find $u^k \in \mathbb{V}_k$ such that

$$a(u^k, v^k) = \langle g, v^k \rangle \quad \forall v^k \in \mathbb{V}_k \tag{1.4}$$

holds. Problem (1.4) is equivalent to solving the system of linear algebraic equations

$$K_{\omega,k}\underline{u}_k = \underline{g}_k, \tag{1.5}$$

where $K_{\omega,k} = \left[a(\phi^k_{ij}, \phi^k_{lm})\right]^{n-1}_{i,j,l,m=1}$, $\underline{u}_k = [u_{ij}]^{n-1}_{i,j=1}$, and $\underline{g}_k = \left[\langle g, \phi^k_{lm}\rangle\right]^{n-1}_{l,m=1}$. The index ω denotes the weight function ω. Then, $u^k = \sum^{n-1}_{i,j=1} u_{ij}\phi^k_{ij}$ is the solution of

(1.4). In this paper, we will derive fast solution methods for (1.5). Because of the right-angled triangles $\tau_{ij}^{s,k}$, and the diagonal matrix $D(x,y)$ in (1.3), the matrix $K_{\omega,k}$ is a sparse matrix with 5-point stencil structure and $\mathcal{O}(n^2)$ nonzero matrix entries. Therefore, it is important to find a method which solves (1.5) in $\mathcal{O}(n^2)$ arithmetical operations. Using direct methods, an additional memory requirement is necessary. Moreover, the arithmetical cost is at least $\mathcal{O}(n^3)$ (see [13] and [14]).

Using iterative methods, no additional memory requirement is necessary in order to save the matrix $K_{\omega,k}$.

However, efficient preconditioners are needed. For systems of finite element equations arising from the discretization of boundary value problems as, for example, $-u_{xx} - u_{yy} = f$, efficient solution techniques have been developed in the last two decades. Examples for such solvers are the preconditioned conjugate gradient (PCG) method with Bramble–Pasciak–Xu preconditioners (see [11]), or hierarchical basis preconditioners (see [26]), and multigrid methods (see [15] and [16]).

However, the differential operator in (1.1) is not spectrally equivalent to the Laplacian. It is an elliptic, but not uniformly elliptic, differential operator (cf. (1.2)). In a certain way, this differential operator can be interpreted as an operator with local anisotropies, where the range of anisotropy ε goes to zero, if the discretization parameter h tends to zero.

A typical anisotropic model problem considered in the literature (see [15]) is

$$-\frac{\partial^2 u}{\partial x^2} - \varepsilon \frac{\partial^2 u}{\partial y^2} = f, \quad \varepsilon \text{ small.}$$

One iterative method with a rate of convergence independent of the choice of ε is the multigrid algorithm with a line Gauss–Seidel (GS) smoother (cf. [16]). Bramble and Zhang [12] considered multigrid methods for a more general case than the Laplace equation. Using a line Jacobi or GS smoother in the x-direction, they proved multigrid convergence for differential operators of the type $-(f(x,y)u_x)_x - (g(x,y)u_y)_y$, where $0 < g(x,y) \leq g_{\max}$ and $0 < f_{\min} < f(x,y) < f_{\max}$, i.e., one of the coefficients can be arbitrarily small. However, both coefficients can be arbitrarily small in (1.1). So a modified smoother handling changing types of anisotropies has to be used. The papers [2], [19], and [1] propose algebraic multilevel iteration (AMLI) preconditioners. In the case of piecewise constant functions $f(x,y)$ and $g(x,y)$, the optimality and robustness of the methods can be shown, i.e., these methods can handle changing types of anisotropies. In our paper, the coefficients are not piecewise constant.

In [6], the special case of the singular weight function $\omega(\xi) = \xi$ in (1.1) is considered. Using the techniques of Braess [10], Schieweck [22], and Pflaum [20], a mesh-size independent multigrid convergence rate $\rho < 1$ has been shown. Moreover, numerical experiments (see [9]) for discretizations of differential operators as (1.1) indicate a mesh-size independent convergence rate $\rho < 1$ for multigrid algorithms with semicoarsening and line smoother. In [5], a BPX-like preconditioner which we call the multiple tridiagonal scaling BPX (MTS-BPX) preconditioner $\hat{C}_{\xi,k}$ for $K_{\xi,k}$ is proposed (i.e., $\omega(\xi) = \xi$). Numerical experiments indicate a small increasing condition number of $\hat{C}_{\xi,k}^{-1} K_{\xi,k}$.

The aim of this paper is to extend the MTS-BPX preconditioner of [5] and the multigrid algorithm of [6] to the more general problem of $K_{\omega,k}\underline{w} = \underline{r}$, where $\omega(\xi)$ satisfies Assumption 1.1.

This paper is organized as follows. In section 2, the multigrid algorithm is considered. We state the main assumptions required for the algebraic convergence theory,

the constant in the strengthened Cauchy inequality, and the smoothing property. Then, the definition of the smoother in [6] for $K_{\xi,k}$ is generalized to a smoother for $K_{\omega,k}$. Moreover, a proof of the smoothing property is given. In section 3, the MTS-BPX preconditioner $\hat{C}_{\omega,k}$ for $K_{\omega,k}$ is defined. Finally, the upper eigenvalue estimate of $\hat{C}_{\omega,k}^{-1}K_{\omega,k}$ is proved and some numerical experiments are given.

Throughout this paper, $\omega(\xi)$ describes the weight function in (1.1). Moreover, the lowest and largest eigenvalues of the matrix A are denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. The integer k is the level number for the refinement of the finite element mesh and $n = 2^k$.

2. Multigrid for degenerate problems. In the typical multigrid proofs (cf. [15]), one splits the multigrid operator into a product of two operators, $\mathcal{A}$ and $\mathcal{B}$. One proves a smoothing property for the operator $\mathcal{A}$, whereas an approximation property has to be shown for $\mathcal{B}$. Helpful tools for this aim are the approximation theorems for finite elements such as the Aubin–Nitsche trick. In order to prove such a result, the boundedness and the uniform ellipticity of the bilinear form are required in the Sobolev space $H^1(\Omega)$. However, the uniform ellipticity of the bilinear form (1.1) cannot be guaranteed (cf. relation (1.2)).

Another technique in order to prove a mesh-size independent convergence rate has been introduced by Braess [10]. In this method, the approximation space $\mathbb{V}_k$ is split into a direct sum of the space $\mathbb{V}_{k-1}$ and a complementary space $\mathbb{W}_k$. One obtains a multiplicative solver for the problem on $\mathbb{V}_k$ by solving the problems on $\mathbb{V}_{k-1}$ and $\mathbb{W}_k$. Schieweck [22] and Pflaum [20] have extended this technique. This method does not require regularity assumptions to the bilinear form. Moreover, for triangulations of simple geometry as for (1.4), the required assumptions are quite simple to handle.

Remark 2.1. Note that the bilinear form $a(\cdot,\cdot)$ is positive definite on the space $\mathbb{V}_k$ (cf. Assumption 1.1).

2.1. Multigrid algorithm and convergence theory. In this subsection, the multigrid algorithm in order to solve (1.5) is introduced. The space $\mathbb{V}_k$ is represented as the direct sum

$$\mathbb{V}_k = \mathbb{V}_{k-1} \oplus \mathbb{W}_k, \quad \text{where} \quad \mathbb{W}_k = \operatorname{span}\{\phi_{ij}^k\}_{(i,j)\in N_k}$$

(see, e.g., [18], [10], [22], [23], and [24]). The index subset $N_k \subset \mathbb{N}^2$ contains the indices of the new nodes on level k and is given by

$$N_k := \{(i,j) \in \mathbb{N}^2, 1 \le i,j \le n-1, i = 2m-1 \text{ or } j = 2m-1, m \in \mathbb{N}\}. \tag{2.1}$$

Let $u_0 \in \mathbb{V}_k$ be the initial guess, and let $\|\cdot\|_a^2 := a(\cdot,\cdot)$ be the energy norm. One step of the multigrid algorithm $u_1 = MULT(k, u_0, g)$ is defined recursively as follows.

ALGORITHM 2.1 ($MULT$). Set $l = k$.

- If $l > 1$, then do
  1. Projection onto $\mathbb{W}_l$: Determine $\tilde{w}_1 \in \mathbb{W}_l$ such that $\| \tilde{w}_1 - w_1 \|_a \le \rho_1 \| w_1 \|_a$, where $w_1 \in \mathbb{W}_l$ is the unique solution of

     $$a(w_1, v) = \langle g, v\rangle - a(u_0, v) = \langle r_1, v\rangle \quad \forall v \in \mathbb{W}_l. \tag{2.2}$$

     Set $u_0^1 = u_0 + \tilde{w}_1$.
  2. Projection onto $\mathbb{V}_{l-1}$ (coarse grid correction): Determine the approximation $\tilde{w}_2 \in \mathbb{V}_{l-1}$ to $w_2 \in \mathbb{V}_{l-1}$ using μ_l steps of the algorithm $MULT(l-1, 0, r_2)$, where w_2 is the unique solution of

     $$a(w_2, v) = \langle g, v\rangle - a(u_0^1, v) = \langle r_2, v\rangle \quad \forall v \in \mathbb{V}_{l-1}.$$

Set $u_0^2 = u_0^1 + \tilde{w}_2$.

3. Projection onto $\mathbb{W}_l$: Determine $\tilde{w}_3 \in \mathbb{W}_l$ such that $\| \tilde{w}_3 - w_3 \|_a \leq \rho_1 \| w_3 \|_a$, where $w_3 \in \mathbb{W}_l$ is the unique solution of

$$a(w_3, v) = \langle g, v \rangle - a(u_0^2, v) = \langle r_3, v \rangle \quad \forall v \in \mathbb{W}_l. \tag{2.3}$$

Set $u_0^3 = u_0^2 + \tilde{w}_3$.

- else
  - Solve $a(w, v) = \langle g, v \rangle$ for all $v \in \mathbb{V}_l$ exactly and set $u_1 = w$.
- end-if.

The algorithm can be interpreted as an approximate alternate projection with respect to $a(\cdot, \cdot)$ onto the subspaces $\mathbb{V}_{k-1}$ and $\mathbb{W}_k$. The exact projections are denoted by w_j, the approximate projections by $\tilde{w}_j$, $j = 1, 2, 3$. For the projection onto $\mathbb{V}_{k-1}$, we apply μ iterations of the algorithm 2.1. For the exact projection onto $\mathbb{W}_k$, we have to solve a system with the matrix

$$K_{\mathbb{W}_k} = \left[a(\phi_{lm}^k, \phi_{ij}^k) \right]_{(i,j),(l,m) \in N_k} \tag{2.4}$$

(compare (2.1)–(2.3) and (1.5)), i.e., the stiffness matrix $K_{\omega,k}$ restricted to the space $\mathbb{W}_k$. In the case of the potential equation, i.e., $\omega(\xi) = 1$, the condition number of $K_{\mathbb{W}_k}$ is bounded independently of the mesh-size h. Hence, in order to solve (2.2), ν iterations $\tilde{w}_{1,0} = 0$, $\tilde{w}_{1,j} = \mathcal{S}\tilde{w}_{1,j-1} + (I - \mathcal{S})w_1$, $j = 1, \ldots, \nu$ of a simple iterative method, i.e., the Jacobi smoother, can be done, where $\mathcal{S}$ denotes the error propagation operator of this method. In each iteration step, the relative error in the energy norm is reduced up to a factor of ρ, where ρ is bounded by $\rho_0 < 1$ from above for all $k \in \mathbb{N}$. So, for all $k \in \mathbb{N}$, it suffices to use $\nu = 1 + \left[\frac{\ln \rho_1}{\ln \rho_0} \right]$ iterations with a Jacobi smoother in order to satisfy $\| \tilde{w}_1 - w_1 \|_a \leq \rho_1 \| w_1 \|_a$ in the case of $\omega(x) = 1$. However, in the case of a general weight function ω, the constant ρ cannot be bounded by $\rho_0 < 1$ independent of k if the Jacobi smoother is used. Then, the main aim of this section is to define a more appropriate preconditioned Richardson iteration, in order to solve (2.2) and (2.3), such that the error propagation operator $\mathcal{S}$ satisfies the estimate

$$\| \mathcal{S}^\nu w \|_a \leq \rho^\nu \| w \|_a \quad \forall w \in \mathbb{W}_k, \nu \in \mathbb{N}. \tag{2.5}$$

The constant ρ in (2.5) is bounded by $\rho_0 < 1$ from above independent of k. Then, the νth iterate of the process $\tilde{w}_{1,0} = 0$, $\tilde{w}_{1,j} = \mathcal{S}\tilde{w}_{1,j-1} + (I - \mathcal{S})w_1$, $j = 1, \ldots, \nu$ satisfies

$$\begin{aligned} \| \tilde{w}_{1,\nu} - w_1 \|_a &= \| \mathcal{S}(\tilde{w}_{1,\nu-1} - w_1) \|_a \\ &\leq \rho \| \tilde{w}_{1,\nu-1} - w_1 \|_a \leq \cdots \leq \rho^\nu \| \tilde{w}_{1,0} - w_1 \|_a = \rho^\nu \| w_1 \|_a . \end{aligned}$$

This means that (2.2), or (2.3), can be solved with a relative accuracy of ρ_1 by a number of ν smoothing steps, where ν is bounded independently of the mesh-size h, i.e., it suffices to verify relation (2.5).

Note that the exact solution w_1 does not have to be determined in order to compute $\tilde{w}_{\nu,1}$ if the smoother is a preconditioned Richardson iteration (cf. (2.20)).

In order to prove the convergence of the multigrid algorithm 2.1 for (1.5), the following convergence theorem is known (see [20] and [22]).

THEOREM 2.1. *Let us assume that the following assumptions are fulfilled.*

- *Let $a(\cdot, \cdot)$ be a symmetric and positive definite bilinear form on $\mathbb{V}_k$.*
- *Let $\mathcal{S}$ be the error propagation operator of a simple iterative method in order to solve* (2.2) *and* (2.3) *which satisfies the estimate* (2.5) *with $0 \leq \rho \leq \rho_0 < 1$ independent of k.*

- *There is a constant* $0 \le \gamma < 1$ *independent of* k *such that*

$$(a(v,w))^2 \le \gamma^2 a(v,v)a(w,w) \quad \forall w \in \mathbb{W}_k, \forall v \in \mathbb{V}_{k-1} \tag{2.6}$$

  holds.
- *Let* $u_{j+1,k} = MULT(k, u_{j,k}, g)$, *let* u^* *be the exact solution of* (1.5), *and let*

$$\sigma_k = \sup_{u_{j,k}-u^* \in \mathbb{V}_\neg} \frac{\| u_{j+1,k} - u^* \|_a}{\| u_{j,k} - u^* \|_a}$$

  be the convergence rate of $MULT$ *in the energy norm with* ν *smoothing operations.*

Then, the recursion formula

$$\sigma_k \le \sigma_{k-1}^{\mu_{k-1}} + (1 - \sigma_{k-1}^{\mu_{k-1}})(\rho^\nu + (1-\rho^\nu)\gamma)^2 \tag{2.7}$$

is valid.

Proof. This theorem has been proved by Schieweck, in Theorem 2.2 of [22], with $\rho^\nu = \rho_1 = \rho_3$, and Pflaum, in Theorem 4 of [20]. □

The following lemma gives conditions on ρ, ν, and γ such that the estimate $\sigma_k < \sigma < 1$ is valid.

LEMMA 2.2. *Let us assume that the assumptions of Theorem* 2.1 *hold. Let* $\kappa^2 = (\rho^\nu + (1-\rho^\nu)\gamma)^2 < \frac{\mu-1}{\mu}$ *with* $\mu = \mu_l$. *If* $\gamma^2 < \frac{\mu-1}{\mu}$, *one has* $\sigma_k < \sigma < 1$ *for*

$$\nu > \ln \frac{\sqrt{\frac{\mu-1}{\mu}} - \gamma}{1-\gamma} / \ln \rho.$$

Proof. The proof is standard (see, e.g., [22]). □

By Remark 2.1, the first assumption of Theorem 2.1 is satisfied for the bilinear form $a(\cdot,\cdot)$ (1.1).

In Theorem 2.2 of [6] we have proved $(a(v,w))^2 \le \frac{95}{176} a(v,v)a(w,w)$ for all $v \in \mathbb{V}_k$ and for all $w \in \mathbb{W}_{k+1}$ for the bilinear form $a(\cdot,\cdot)$ (1.1) with the weight function $\omega(\xi) = \xi$. The techniques described there can be extended to other weight functions. Table 2.1 gives estimates of the strengthened Cauchy inequalities for several weight functions of the type $\omega(\xi) = \xi^\alpha$, $\alpha \ge 0$. For $\alpha = 10$, the constant γ^2 of the strengthened Cauchy inequality is very close to 1. In the cases $\alpha = \frac{1}{2}, 1, 2$, the estimate $\frac{1}{2} < \gamma^2 < \frac{2}{3}$ is valid. Thus, we can prove a mesh-size independent multigrid convergence rate $\sigma_k < 1$ for $\mu = 3$, if $\tilde{w}_1$ and $\tilde{w}_3$ are close to w_1 and w_3 in Algorithm 2.1 (cf. (2.2) and (2.3)), i.e., $\nu > \ln \frac{\sqrt{2/3}-\gamma}{1-\gamma} / \ln \rho$.

TABLE 2.1
Estimates of the constant in strengthened Cauchy inequality for several weight functions.

weight function	$\omega = 1$	$\omega(\xi) = \sqrt{\xi}$	$\omega(\xi) = \xi$	$\omega(\xi) = \xi^2$	$\omega(\xi) = \xi^{10}$
γ^2 in (2.6)	$\frac{1}{2}$	$\frac{81}{158}$	$\frac{95}{176}$	$\frac{2195375}{3508896}$	≈ 0.9929

In subsection 2.3, we define an iterative method with an error propagation operator $\mathcal{S}$ satisfying relation (2.5) with $\rho < \rho_0 < 1$ for more general weight functions as in [6]. It will be shown that the constant ρ in (2.5) is independent of the choice of the weight function.

Therefore, the stiffness matrices restricted to the elements $\tau_{ij}^{1,k}$ and $\tau_{ij}^{2,k}$ are required. This is done in subsection 2.2.

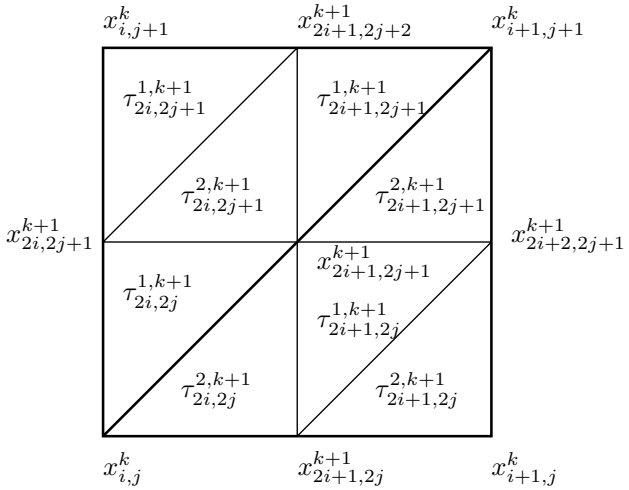


FIG. 2.1. *Local numbering of the nodes and subtriangles of* $\mathcal{E}^k_{ij}$.

2.2. Calculation of the macroelement stiffness matrices. In this subsection, we determine the stiffness matrix on the macroelements $\mathcal{E}^k_{ij}$ with respect to the basis functions of $\mathbb{W}_{k+1}\,|_{\mathcal{E}^k_{ij}}$. We start with the introduction of the basis functions on $\mathcal{E}^k_{ij}$. Note that the triangle $\tau^{2,k}_{ij}$ is the union of the triangles $\tau^{2,k+1}_{2i,2j}$, $\tau^{1,k+1}_{2i+1,2j}$, $\tau^{2,k+1}_{2i+1,2j}$, and $\tau^{2,k+1}_{2i+1,2j+1}$, and the triangle $\tau^{1,k}_{ij}$ is the union of the triangles $\tau^{1,k+1}_{2i,2j}$, $\tau^{1,k+1}_{2i,2j+1}$, $\tau^{2,k+1}_{2i,2j+1}$, and $\tau^{1,k+1}_{2i+1,2j+1}$. The nodes x^k_{ij}, $x^k_{i,j+1}$, $x^k_{i+1,j}$, and $x^k_{i+1,j+1}$ are the coarse grid nodes, the nodes $x^{k+1}_{2i+1,2j}$, $x^{k+1}_{2i,2j+1}$, $x^{k+1}_{2i+2,2j+1}$, $x^{k+1}_{2i+1,2j+2}$, and $x^{k+1}_{2i+1,2j+1}$ are new in level $k+1$ (cf. Figure 2.1). Using this notation, we have

$$\mathbb{W}_{k+1}\,|_{\mathcal{E}^k_{ij}} = \operatorname{span}\{\phi^{k+1}_{lm}\}_{(l,m)\in N^{\mathbb{W}_{k+1}}_{ij}}. \tag{2.8}$$

For reasons of simplicity, we write only ϕ^{k+1}_{lm} instead of $\phi^{k+1}_{lm}\,|_{\mathcal{E}^k_{ij}}$ for the restriction of ϕ^{k+1}_{lm} on $\mathcal{E}^k_{ij}$. The index set in (2.8) is given by

$$N^{\mathbb{W}_{k+1}}_{ij} = N_{k+1} \cap \{(l,m) \in \mathbb{N}_0^2, 2i \le l \le 2i+2, 2j \le m \le 2j+2\},$$

where N_{k+1} was defined in (2.1). Since $\mathbb{V}_k \subset H^1_0(\Omega)$, some modifications are necessary for boundary macroelements $\mathcal{E}^k_{ij}$, i.e., with $i=0$, $j=0$, $i=n-1$, or $j=n-1$.

On the elements $\tau^{s,k}_{ij}$, $s=1,2$, we introduce the matrices

$$J_{s,ij} := \left[a^{\tau^{s,k}_{ij}}(\phi^{k+1}_{lm}, \phi^{k+1}_{rq})\right]_{(r,q),(l,m)\in N^{s,\mathbb{W}_{k+1}}_{ij}}$$

with $N^{s,\mathbb{W}_{k+1}}_{ij} := T^s_{ij} \cap N^{\mathbb{W}_{k+1}}_{ij}$, where $T^1_{ij} := \{(l,m) \in \mathbb{N}_0^2, l-m \le i-j\}$ and $T^2_{ij} := \{(l,m) \in \mathbb{N}_0^2, l-m \ge i-j\}$. Namely, we obtain

$$N^{2,\mathbb{W}_{k+1}}_{ij} = \{(2i+1,2j),(2i+2,2j+1),(2i+1,2j+1)\} \quad \text{and}$$
$$N^{1,\mathbb{W}_{k+1}}_{ij} = \{(2i,2j+1),(2i+1,2j+2),(2i+1,2j+1)\}.$$

The entries of the matrices $J_{q,ij}$ can be determined by a straightforward calculation. We compute those for the case of a general weight function $\omega(\xi)$. The following

positive parameters depending on the integer j are introduced:

$$d_j = \frac{1}{4}\int_{\tau_{2i,2j}^{1,k+1}\cup\tau_{2i,2j+1}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y), \quad e_j = \frac{1}{4}\int_{\tau_{2i,2j}^{2,k+1}\cup\tau_{2i+1,2j}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y),$$

$$f_j = \frac{1}{4}\int_{\tau_{2i,2j+1}^{1,k+1}\cup\tau_{2i+1,2j+1}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y). \tag{2.9}$$

Note that d_j, e_j, and f_j are independent of the integer i. The values d_i, e_i, and f_i are defined by a permutation of x with y, and of each triangle $\tau_{ij}^{2,k}$ with $\tau_{ji}^{1,k}$ in (2.9). One obtains the following proposition.

PROPOSITION 2.3. *Let $0 \le i < n-1$ and $0 < j \le n-1$. Then, one has*

$$J_{2,ij} = 4\begin{bmatrix} d_i+e_j & 0 & -d_i \\ 0 & f_i+d_j & -d_j \\ -d_i & -d_j & d_i+d_j \end{bmatrix}. \tag{2.10}$$

Due to the boundary condition $u\,|_{\partial\Omega} = 0$, the second row and column of $J_{2,ij}$ has to be canceled for $i = n-1$, whereas the first row and column in $J_{2,ij}$ has to be canceled for $j = 0$ in (2.10).

By exchanging the indices i and j in (2.10), one derives the matrices $J_{1,ij} = J_{2,ji}$.

2.3. Construction of the smoother. In order to apply multigrid to the linear system (1.5), we need an efficient smoother. This smoother will be constructed by the local behavior of the differential operator. An idea of Axelsson and Padiy [2] (see also [19]) for anisotropic problems is extended to bilinear forms as in problem (1.1). This smoother operates only on the space $\mathbb{W}_{k+1}$.

Consider the triangle $\tau_{ij}^{s,k}$. For our discussion, only the submatrices $J_{s,ij}$, where $0 \le i,j \le n-1$, and $s = 1,2$, are required which correspond to the nodal basis functions on $\mathbb{W}_{k+1}$. The two cases $i < j$ and $i \ge j$ are discussed. We start with $i < j$ and $s = 2$. By Proposition 2.3,

$$J_{2,ij} = 4\begin{bmatrix} d_i+e_j & 0 & -d_i \\ 0 & f_i+d_j & -d_j \\ -d_i & -d_j & d_i+d_j \end{bmatrix}.$$

The index k is omitted. For $i < j$, the matrix

$$M_{2,ij} = 4\begin{bmatrix} d_i+e_j & 0 & 0 \\ 0 & f_i+d_j & -d_j \\ 0 & -d_j & d_i+d_j \end{bmatrix} \tag{2.11}$$

is introduced. In the matrix $M_{2,ij}$, we set all off-diagonal entries of $J_{2,ij}$ to 0 which have relatively small absolute values in comparison to the corresponding main diagonal entries. Since ω is monotonically increasing, the relation $d_i < d_j$ is valid for $i < j$. Thus, we set the $-d_i$ entries of $J_{2,ij}$ in $M_{2,ij}$ to 0. We now prove the following lemma.

LEMMA 2.4. *For $0 \le i < j < n$, the eigenvalue estimates*

$$\lambda_{\min}\left({M_{2,ij}}^{-1}J_{2,ij}\right) \ge 1 - \frac{1}{2}\sqrt{2} \quad \textit{and} \quad \lambda_{\max}\left({M_{2,ij}}^{-1}J_{2,ij}\right) \le 1 + \frac{1}{2}\sqrt{2}$$

hold.

Proof. Let $\beta = d_i f_i + d_i d_j + f_i d_j$. Then, we have

$$M_{2,ij}^{-1} J_{2,ij} = \begin{bmatrix} 1 & 0 & \frac{-d_i}{d_i+e_j} \\ \frac{-d_i d_j}{\beta} & 1 & 0 \\ \frac{-d_i f_i - d_i d_j}{\beta} & 0 & 1 \end{bmatrix}.$$

This matrix has the characteristic polynomial

$$\det(\lambda I - M_{2,ij}^{-1} J_{2,ij}) = (\lambda - 1)\left((1-\lambda)^2 - \frac{d_i}{d_i+e_j}\frac{d_i f_i + d_i d_j}{d_i f_i + d_i d_j + f_i d_j}\right).$$

The roots λ_i, $i = 1, 2, 3$, of this polynomial are $\lambda_1 = 1$ and $\lambda_{2,3} = 1 \pm \sqrt{\rho}$, where

$$\rho = \frac{d_i}{d_i+e_j}\frac{d_i f_i + d_i d_j}{d_i f_i + d_i d_j + f_i d_j}. \tag{2.12}$$

Note that for all i and j, the values d_j, e_j, and f_j are mean values of the positive function $(\omega(y))^2$ over the union of two triangles having a volume of $\frac{1}{8n^2}$. By $i \le j-1$ and the monotonicity of the weight function, one has $\omega(x) \le \omega(y)$ for all $x, y \in \tau_{ij}^{2,k}$. Thus, by integration over subtriangles of $\tau_{ij}^{2,k}$ with volume $\frac{1}{8n^2}$ (cf. Figure 2.1),

$$d_i = \frac{1}{4}\int_{\tau_{2i,2j}^{2,k+1} \cup \tau_{2i+1,2j}^{1,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) \le \frac{1}{4}\int_{\tau_{2i,2j}^{2,k+1} \cup \tau_{2i+1,2j}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) = e_j,$$

$$\text{or} \quad \frac{d_i}{d_i+e_j} \le \frac{1}{2}.$$

Since $d_i, d_j, f_i, f_j > 0$ (the weight function is positive), we have

$$\frac{d_i fi + d_i d_j}{d_i f_i + d_i d_j + f_i d_j} < 1. \tag{2.13}$$

Inserting the above estimates into (2.12), one has

$$1 - \sqrt{\frac{1}{2}} \le \lambda_2 \le \lambda_1 \le \lambda_3 \le 1 + \sqrt{\frac{1}{2}}.$$

Hence, the assertion follows immediately. ∎

Remark 2.2. The estimate (2.13) is not sharp. With $f_i \le d_j$ and $d_i < d_j$, one can show the stronger estimate $\frac{d_i f_i + d_i d_j}{d_i f_i + d_i d_j + f_i d_j} < \frac{2}{3}$ (cf. [8]).

Now, consider the case $i \ge j$. Introducing the matrix

$$M_{2,ij} = 4\begin{bmatrix} d_i + e_j & 0 & -d_i \\ 0 & f_i + d_j & 0 \\ -d_i & 0 & d_i + d_j \end{bmatrix}, \tag{2.14}$$

we will show that $\kappa\left({M_{2,ij}}^{-1} J_{2,ij}\right) \le c$ independent of the parameters j, i, and n.

LEMMA 2.5. *For $0 \le j \le i < n$, one has*

$$\lambda_{\min}\left({M_{2,ij}}^{-1} J_{2,ij}\right) \ge 1 - \frac{1}{2}\sqrt{2} \quad \textit{and} \quad \lambda_{\max}\left({M_{2,ij}}^{-1} J_{2,ij}\right) \le 1 + \frac{1}{2}\sqrt{2}.$$

Proof. We start with the case $i < n-1$ and $j > 0$. The proof is similar to the proof of Lemma 2.4. A short calculation yields

$$\det(\lambda I - M_{2,ij}^{-1} J_{2,ij}) = (\lambda - 1)\left((\lambda - 1)^2 - \frac{d_j}{d_j + f_i}\frac{d_i d_j + e_j d_j}{d_i e_j + d_i d_j + e_j d_j}\right).$$

By $i \geq j$ and the monotonicity of the weight function ω, we have

$$\begin{aligned}\int_{\tau_{2i+1,2j}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) &= \int_{\tau_{2i+1,2j+1}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y)\\ &\geq \int_{\tau_{2j+1,2i}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y)\\ &= \int_{\tau_{2i,2j+1}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) \geq \int_{\tau_{2i+1,2j}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y). \end{aligned} \tag{2.15}$$

For the same reason,

$$\int_{\tau_{2i,2j}^{2,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) \leq \int_{\tau_{2i+1,2j}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y). \tag{2.16}$$

Using (2.15) and (2.16), we have

$$f_i = \int_{\tau_{2i+1,2j}^{2,k+1} \cup \tau_{2i+1,2j+1}^{2,k+1}} (\omega(x))^2 \,\mathrm{d}(x,y) \geq \int_{\tau_{2i,2j}^{2,k+1} \cup \tau_{2i+1,2j}^{1,k+1}} (\omega(y))^2 \,\mathrm{d}(x,y) = d_j$$
$$\text{and} \quad \frac{d_j}{d_j + f_i} \leq \frac{1}{2}.$$

Together with $e_j d_j + d_j d_i < e_j d_i + e_j d_j + d_j d_i$, the assertion follows as in the proof of Lemma 2.4.

Consider now $i = n-1$. Then, the second row and column of $M_{2,ij}$ and $J_{2,ij}$ has to be canceled. Thus, $M_{2,n-1,j} = J_{2,n-1,j}$ and

$$\lambda_1(M_{2,n-1,j}^{-1} J_{2,n-1,j}) = \lambda_2(M_{2,n-1,j}^{-1} J_{2,n-1,j}) = 1.$$

The last case is $j = 0$. We have to omit the first row and column in $M_{2,i,0}$ and $J_{2,i,0}$. By a short calculation the estimates $\frac{1}{2} \leq \lambda_2 < \lambda_1 \leq \frac{3}{2}$ are obtained for the roots of the characteristic polynomial of the matrix $M_{2,i,0}^{-1} J_{2,i,0}$ (cf. [8]). ∎

In (2.11) and (2.14), we have defined a local preconditioner $M_{2,ij}$ for the element stiffness matrix $J_{2,ij}$ corresponding to the triangle $\tau_{ij}^{2,k}$. On the triangles $\tau_{ij}^{1,k}$, we define matrices $M_{1,ij}$ in the same way as $M_{2,ij}$ for $\tau_{ij}^{2,k}$:

$$M_{1,ij} = \begin{cases} 4\begin{bmatrix} e_i + d_j & 0 & -d_j \\ 0 & d_i + f_j & 0 \\ -d_j & 0 & d_i + d_j \end{bmatrix} & \text{for} \quad i \leq j, \\ 4\begin{bmatrix} e_i + d_j & 0 & 0 \\ 0 & d_i + f_j & -d_i \\ 0 & -d_i & d_i + d_j \end{bmatrix} & \text{for} \quad i > j. \end{cases}$$

Remark 2.3. By the symmetry of the differential operator with respect to the variables x and y, we obtain the same results for the triangles $\tau_{ij}^{1,k}$ as in Lemmas 2.4 and 2.5.

Following [6], a global preconditioner $M_{\mathbb{W}_{k+1}}$ for $K_{\mathbb{W}_{k+1}}$ (2.4) is defined using the local matrices $M_{s,ij}$, where $0 \le i,j \le n-1$, $s = 1,2$. The matrix $K_{\mathbb{W}_{k+1}}$ is the result of assembling the local stiffness matrices $J_{s,ij}$, $s = 1,2$, and $i,j = 0,\dots,n-1$, i.e.,

$$K_{\mathbb{W}_{k+1}} = \sum_{s=1}^{2} \sum_{i,j=0}^{n-1} L_{s,ij}^T J_{s,ij} L_{s,ij}. \tag{2.17}$$

The matrices $L_{s,ij} \in \mathbb{R}^{3\times 3\cdot 4^{k-1}-2^k}$ are the usual finite element connectivity matrices.

DEFINITION 2.6. *We define the matrix* $M_{\mathbb{W}_{k+1}}$ *by*

$$M_{\mathbb{W}_{k+1}} = \sum_{s=1}^{2} \sum_{i,j=0}^{n-1} L_{s,ij}^T M_{s,ij} L_{s,ij}. \tag{2.18}$$

Because of the properties of the local preconditioners $M_{s,ij}$, the matrix $M_{\mathbb{W}_{k+1}}$ is a good preconditioner for $K_{\mathbb{W}_{k+1}}$. This result is stated as the main theorem of this subsection.

THEOREM 2.7. *Let* $\omega(\xi)$ *satisfy Assumption* 1.1*, let* $M_{\mathbb{W}_{k+1}}$ *and* $K_{\mathbb{W}_{k+1}}$ *be defined in* (2.18) *and* (2.17)*, respectively. Then, one obtains*

$$\lambda_{\min}\left((M_{\mathbb{W}_{k+1}})^{-1} K_{\mathbb{W}_{k+1}}\right) \ge 1 - \frac{1}{2}\sqrt{2}, \quad \lambda_{\max}\left((M_{\mathbb{W}_{k+1}})^{-1} K_{\mathbb{W}_{k+1}}\right) \le 1 + \frac{1}{2}\sqrt{2}.$$

In order to prove the assertion, we use the following result (cf. Lemma 2.5 of [6]; see also [2], [25]).

LEMMA 2.8. *Let* $\{A_i \in \mathbb{R}^{m_i,m_i}\}_{i=1}^s$ *be a finite set of symmetric positive definite matrices. Let* $A = \sum_{i=1}^s L_i^T A_i L_i$*, where* $L_i \in \mathbb{R}^{m_i,m}$ *and* $A \in \mathbb{R}^{m,m}$*. Furthermore, let* C_i *be a preconditioner for the matrix* A_i*, i.e., for all* $\underline{w} \in \mathbb{R}^{m_i}$ *the relations*

$$\underline{\lambda}_i (C_i \underline{w}, \underline{w}) \le (A_i \underline{w}, \underline{w}) \le \overline{\lambda}^i (C_i \underline{w}, \underline{w}) \tag{2.19}$$

with $0 < \overline{\lambda^i}$ *and* $0 \le \underline{\lambda_i}$ *hold. Let* $C = \sum_{i=1}^s L_i^T C_i L_i$*. Then,* $\forall \underline{v} \in \mathbb{R}^m$

$$\underline{\lambda}(C\underline{v}, \underline{v}) \le (A\underline{v}, \underline{v}) \le \overline{\lambda}(C\underline{v}, \underline{v})$$

is valid with

$$\underline{\lambda} = \min_i \underline{\lambda}_i, \quad \overline{\lambda} = \max_i \overline{\lambda}^i.$$

Proof of Theorem 2.7. By Lemmas 2.4 and 2.5 and Remark 2.3, the assumptions (2.19) are satisfied. Using Lemma 2.8, the assertions follow. ∎

Applying Theorem 2.7, a preconditioned Richardson iteration can be built as a preconditioned simple iteration method. The error propagation operator $\mathcal{S}_{\omega,k+1}$ of this method is defined by

$$S_{\omega,k+1} = I - \zeta (M_{\mathbb{W}_{k+1}})^{-1} K_{\mathbb{W}_{k+1}}, \tag{2.20}$$

where $S_{\omega,k+1}$ denotes the matrix representation of $\mathcal{S}_{\omega,k+1}$ by the usual fem-isomorphism. This smoother $\mathcal{S} = \mathcal{S}_{\omega,k+1}$ can be used for Algorithm 2.1.

COROLLARY 2.9. *Let* $\| w \|_a^2 = a(w,w)$ *be the energy norm of the bilinear form* a*. Then, for all* $w \in \mathbb{W}_{k+1}$*,*

$$\| \mathcal{S}_{\omega,k+1}^{\nu} w \|_a \le \rho^{\nu} \| w \|_a$$

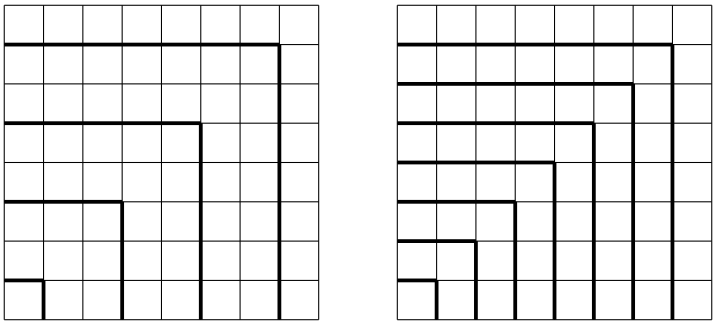


FIG. 2.2. *Lines of the smoothers $S_{\omega,k}$ (left) and $\tilde{S}_{\omega,k}$ (right).*

holds, where $\zeta = 1$ and $\rho = \frac{1}{2}\sqrt{2}$.

Proof. The proof is similar to the proof of Corollary 2.3 in [6]. ∎

The smoother $S_{\omega,k+1}$ can be interpreted as a line smoother (see [6] and the left-hand picture of Figure 2.2). Then, using Cholesky or Crout decomposition, the operation $S_{\omega,k}\underline{w} = \underline{r}$ can be done in $\mathcal{O}(m_k)$ arithmetical operations, where m_k is the number of unknowns on level k.

Additionally, we build a smoother $\tilde{S}_{\omega,k} = I - \zeta L_{\omega,k}^{-1} K_{\omega,k}$ which uses the ideas of (2.20). This smoother operates on the space $\mathbb{V}_k$. Let

(2.21)
$$L_{\omega,k} = \operatorname{diag}(K_{\omega,k}) + \tilde{R}, \quad \text{where} \quad \tilde{R} = \left[\tilde{b}(\phi_{ij}^k, \phi_{lm}^k) + \tilde{b}(\phi_{lm}^k, \phi_{ij}^k)\right]_{(i,j),(l,m)=(1,1)}^{(n-1,n-1)}$$

with the bilinear form $\tilde{b} : \mathbb{V}_k \times \mathbb{V}_k \to \mathbb{R}$,

$$\tilde{b}(\phi_{ij}^k, \phi_{lm}^k) = \begin{cases} a(\phi_{ij}^k, \phi_{lm}^k) & \text{if} \quad \begin{array}{lll} i = l = r, & j = 2, \dots, i, & m = j-1, \\ \text{or} \quad j = m = r, & i = 2, \dots, j, & l = i-1, \end{array} \\ 0 & \text{otherwise} \end{cases}$$

for $r = 1, \dots, n-1$. As well as $S_{\omega,k}$ (2.20), $\tilde{S}_{\omega,k}$ is a line smoother (cf. [6] and the right-hand picture of Figure 2.2). Analogous to $S_{\omega,k}$, the operation $\tilde{S}_{\omega,k}\underline{w} = \underline{r}$ can be done in $\mathcal{O}(n^2)$ flops using Cholesky or Crout decomposition.

3. BPX preconditioner. Recall the finite element discretization of problem (1.1).

Find $u \in \mathbb{V}_k$ such that

$$\int_\Omega \left(\omega^2(y) u_x v_x + \omega^2(x) u_y v_y\right) \, \mathrm{d}(x,y) = \int_\Omega f v \, \mathrm{d}(x,y) \tag{3.1}$$

holds for all $v \in \mathbb{V}_k$ with a weight function $\omega(\xi)$ satisfying Assumption 1.1.

For the efficient solution of systems of linear equations arising from discretizations of uniformly elliptic problems by finite elements, Bramble, Pasciak, and Xu [11] have developed a preconditioner which has been called the BPX preconditioner. For this preconditioner, the spectral equivalence to the original stiffness matrix can be shown. Later, this preconditioner has been improved by the multiple diagonal scaling version (see [27]). As mentioned in [4], a BPX preconditioner with multiple diagonal scaling

TABLE 3.1
Lower (bottom) and upper (top) eigenvalue bounds of the MTS-BPX preconditioned system matrix.

Level	$\overline{c}$				
	$\omega(\xi)=1$	$\omega(\xi)=\sqrt{\xi}$	$\omega(\xi)=\xi$	$\omega(\xi)=\xi^2$	$\omega(\xi)=\xi^{10}$
2	1.86	1.80	1.77	1.82	2.00
3	2.73	2.65	2.59	2.51	2.93
4	3.44	3.41	3.39	3.34	3.75
5	4.00	4.01	4.03	4.06	4.59
6	4.45	4.47	4.52	4.70	5.50
7	4.81	4.85	4.91	5.34	6.44
8	5.11	5.14	5.23	6.03	7.40
9	5.35	5.39	5.59	6.70	8.37
10	5.55	5.59	6.11	7.42	9.35
Level	$\underline{c}$				
	$\omega(\xi)=1$	$\omega(\xi)=\sqrt{\xi}$	$\omega(\xi)=\xi$	$\omega(\xi)=\xi^2$	$\omega(\xi)=\xi^{10}$
2	0.607	0.687	0.747	0.822	0.977
3	0.522	0.607	0.647	0.690	0.844
4	0.495	0.554	0.583	0.619	0.716
5	0.489	0.527	0.543	0.569	0.664
6	0.488	0.513	0.524	0.538	0.611
7	0.488	0.504	0.512	0.522	0.569
8	0.488	0.498	0.504	0.511	0.541
9	0.488	0.495	0.498	0.503	0.524
10	0.488	0.493	0.495	0.498	0.513

does not show good numerical results in order to solve $K_{\omega,k}\underline{u} = \underline{g}_k$, the system of linear algebraic equations resulting from the finite element discretization of (3.1). One reason is that this preconditioner cannot handle the anisotropies resulting from the degenerate elliptic operator. However, with a modification, the so-called MTS-BPX, this behavior of the BPX preconditioner can be improved (see [5]).

In subsection 2.3, the smoother $\tilde{S}_{\omega,k} = I - \zeta L_{\omega,k}^{-1} K_{\omega,k}$ has been considered as smoother for $K_{\omega,k}$. In this smoother, the matrix $L_{\omega,k}$ is a preconditioner for $K_{\omega,k}$ which can handle anisotropies. The idea now is to apply the matrix $L_{\omega,k}$ as "scaling" on each level instead of a diagonal scaling. We expect a stabilization of the BPX preconditioner. The following MTS-BPX preconditioner is now defined. Let Q_l^k, $l = 0, \ldots, k$ be the basis interpolation matrix from the basis $\{\phi_{ij}^l\}_{i,j}^{n_l-1} \in \mathbb{V}_l$ to the basis $\{\phi_{ij}^k\}_{i,j=1}^{n_k-1} \in \mathbb{V}_k$, where $n_j = 2^j$. Let Q_k^l be the transposed matrix. Furthermore, let $L_{\omega,k}$ be the matrix in (2.21). Then, we define the preconditioner

$$\hat{C}_{\omega,k}^{-1} = \sum_{l=1}^{k} Q_l^k L_{\omega,l}^{-1} Q_k^l. \tag{3.2}$$

This preconditioner is called the MTS-BPX preconditioner for $K_{\omega,k}$. In the case of $\omega(\xi) = \xi$, this definition corresponds to the definition of the MTS-BPX preconditioner given in [5].

Concerning the quality of $\hat{C}_{\omega,k}$ as preconditioner for $K_{\omega,k}$, the following result has been proved in [8].

LEMMA 3.1. *The eigenvalue estimate* $\lambda_{\max}\left(\hat{C}_{\omega,k}^{-1} K_{\omega,k}\right) \leq 2\,k$ *for the MTS-BPX preconditioner* (3.2) *is valid.*

For the MTS-BPX preconditioner $\hat{C}_{\omega,k}$ (3.2), Table 3.1 gives the lower and upper

constants for the spectral equivalence relations

$$\underline{c}\left(\hat{C}_{\omega,k}\underline{v},\underline{v}\right) \leq (K_{\omega,k}\underline{v},\underline{v}) \leq \overline{c}\left(\hat{C}_{\omega,k}\underline{v},\underline{v}\right) \quad \forall \underline{v},$$

computed by a vector iteration and inverse vector iteration for the corresponding matrices and the weight functions $\omega(\xi) = \xi^\alpha$, $(\alpha = 0, \frac{1}{2}, 1, 2, 10)$. One can see that the constant $\overline{c}$ is proportional to the level number for all considered weight functions which indicates that the estimate of Lemma 3.1 is sharp. The lower constant $\underline{c}$ seems to be bounded from below by a constant of about 0.488 uniformly with respect to α. However, we cannot prove the boundedness of $\underline{c}$ from below.

4. Concluding remarks. In this paper, we have proposed two methods in order to solve the system $K_{\omega,k}\underline{u}_k = \underline{g}_k$ (1.5). The first method is the multigrid algorithm $MULT$ with an appropriate line smoother $\mathcal{S}$. We have given the proof of the smoothing property (2.5) for a general class of weight functions ω. Together with the estimate for the constant γ^2 of the strengthened Cauchy inequality, the constant lies between $\frac{1}{2}$ and $\frac{2}{3}$ for $\omega(x) = x^\alpha$, $\alpha = 0.5, 1, 2$, and we can prove a mesh-size independent convergence rate for $\mu \geq 3$. Hence, the proposed method solves (1.5) in $\mathcal{O}(n^2)$ operations if $\gamma^2 < \frac{2}{3}$. If $\gamma^2 > \frac{2}{3}$, one can obtain an optimal method with two modified approaches. On the one hand, a multigrid with varying μ_l can be used, i.e., $\mu_l = 3$ if l is even and $\mu_l = 4$ if l is odd. On the other hand, a PCG method with AMLI preconditioner $C_{am,k,\omega}$ is used. Then, using the ingredients of the multigrid algorithm, i.e., $M_{\mathbb{W}_{k+1}}$ (2.17) as preconditioner for $K_{\mathbb{W}_{k+1}}$, and a polynomial iteration of degree $\mu = 3$, one can prove that the condition number of $C^{-1}_{am,k,\omega}K_{\omega,k}$ is bounded independently of the mesh-size h (see [7] for the special case $\omega(\xi) = \xi$, if $\gamma^2 < \frac{8}{9}$; cf. [3]). The second method is the PCG method with the MTS-BPX preconditioner. Due to Lemma 3.1, this method is not optimal. However, the numerical experiments indicate that we obtain a robust fast solver for weight functions as, e.g., $\omega(\xi) = \xi^{10}$ as well.

REFERENCES

[1] O. Axelsson and S. D. Margenov, *On multilevel preconditioners which are optimal with respect to both problem and discretization parameter*, Comput. Meth. Appl. Math., 3 (2003), pp. 6–22.

[2] O. Axelsson and A. Padiy, *On the additive version of the algebraic multilevel iteration method for anisotropic elliptic problems*, SIAM J. Sci. Comput., 20 (1999), pp. 1807–1830.

[3] O. Axelsson and P. S. Vassilevski, *Algebraic multilevel preconditioning methods* II, SIAM J. Numer. Anal., 27 (1990), pp. 1569–1590.

[4] S. Beuchler, *Lösungsmethoden bei der p-version der fem*, Diplomarbeit, Technische Universität Chemnitz, Chemnitz, Germany, 1999.

[5] S. Beuchler, *The MTS-BPX-Preconditioner for the p-Version of the FEM*, Technical report SFB393 01-16, Technische Universität Chemnitz, Chemnitz, Germany, 2001.

[6] S. Beuchler, *Multigrid solver for the inner problem in domain decomposition methods for p-FEM*, SIAM J. Numer. Anal., 40 (2002), pp. 928–944.

[7] S. Beuchler, *AMLI preconditioner for the p-version of the FEM*, Numer. Linear Algebra Appl., 10 (2003), pp. 721–732.

[8] S. Beuchler, *Fast Solvers for Degenerated Problems*, Technical report SFB393 03-04, Technische Universität Chemnitz, Chemnitz, Germany, 2003.

[9] S. Börm and R. Hiptmair, *Analysis of tensor product multigrid*, Numer. Algorithms, 26 (2001), pp. 219–234.

[10] D. Braess, *The contraction number of a multigrid method for solving the Poisson equation*, Numer. Math, 37 (1981), pp. 387–404.

[11] J. Bramble, J. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1991), pp. 1–22.

[12] J. Bramble and X. Zhang, *Uniform convergence of the multigrid V-cycle for an anisotropic problem*, Math. Comp., 70 (2001), pp. 453–470.
[13] A. George, *Nested dissection of a regular finite-element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.
[14] A. George and J. W.-H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
[15] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Heidelberg, 1985.
[16] W. Hackbusch and U. Trottenberg, *Multigrid Methods, Proceedings of the Conference held at Köln-Porz, November* 23-27, 1981, Lect. Notes Math. 960, Springer-Verlag, Berlin, 1982.
[17] A. Kufner and A. M. Sändig, *Some Applications of Weighted Sobolev Spaces*, Teubner, Leipzig, 1987.
[18] J. F. Maitre and F. Musy, *The contraction number of a class of two-level methods, and exact evaluation for some finite element subspaces and model problems*, in Multigrid Methods, Proceedings of the Conference held at Köln-Porz, November 23-27, 1981, Lect. Notes Math. 960, W. Hackbusch and U. Trottenberg, eds., pp. 535–544, Springer-Verlag, Berlin, 1982.
[19] S. D. Margenov and P. S. Vassilevski, *Algebraic multilevel preconditioning of anistropic elliptic problems*, SIAM J. Sci. Comput., 15 (1994), pp. 1026–1037.
[20] Ch. Pflaum, *Fast and Robust Multilevel Algorithms*, Habilitationsschrift, Universität Würzburg, Würzburg, Germany, 1998.
[21] O. Pironneau and F. Hecht, *Mesh adaption for the Black and Scholes equations*, East-West J. Numer. Math., 8 (2000), pp. 25–36.
[22] N. Schieweck, *A multi-grid convergence proof by a strengthened Cauchy-inequality for symmetric elliptic boundary value problems*, in Second Multigrid Seminar, Garzau 1985, Report R-Math 08-86, G. Telschow, ed., pp. 49–62, Karl-Weierstraß-Institut für Mathematik, Berlin, 1986.
[23] C. A. Thole, *Beiträge zur Fourieranalyse von Mehrgittermethoden: V-cycle, ILU-Glättung, anisotrope Operatoren*, Diplomarbeit, Universität Bonn, Bonn, Germany, 1983.
[24] R. Verfürth, *The contraction number of a multigrid method with mesh ratio* 2 *for solving Poisson's equation*, Numer. Math., 60 (1984), pp. 113–128.
[25] A. J. Wathen, *An analysis of some element-by-element techniques*, Comput. Methods Appl. Mech. Engrg., 74 (1989), pp. 271–287.
[26] H. Yserentant, *On the multi-level-splitting of the finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.
[27] X. Zhang, *Multilevel Schwarz methods*, Numer. Math., 63 (1992), pp. 521–539.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1357–1393

STABILITY AND CONVERGENCE OF A CLASS OF FINITE ELEMENT SCHEMES FOR HYPERBOLIC SYSTEMS OF CONSERVATION LAWS*

CHRISTOS ARVANITIS†, CHARALAMBOS MAKRIDAKIS‡, AND ATHANASIOS E. TZAVARAS§

Abstract. We propose a class of finite element schemes for systems of hyperbolic conservation laws that are based on finite element discretizations of appropriate relaxation models. We consider both semidiscrete and fully discrete finite element schemes and show that the schemes are stable and, when the compensated compactness theory is applicable, do converge to a weak solution of the hyperbolic system. The schemes use piecewise polynomials of arbitrary degree and their consistency error is of high order. We also prove that the rate of convergence of the relaxation system to a smooth solution of the conservation laws is of order $O(\varepsilon)$.

Key words. stability and convergence, finite element schemes, hyperbolic conservation laws, adaptive schemes

AMS subject classifications. 65M60, 65M12, 65M15, 35L65

DOI. 10.1137/S0036142902420436

1. Introduction.

The problem of numerical approximation of nonlinear hyperbolic systems of conservation laws,

$$\begin{aligned} &\partial_t u + \sum_{j=1}^{d} \partial_{x_j} F_j(u) = 0, \quad x \in \mathbb{R}^d, \;\; u = u(x,t) \in \mathbb{R}^n, \quad t > 0\,, \\ &u(\cdot,0) = u_0(\cdot), \end{aligned} \tag{1.1}$$

is a challenging area testing the performance of various numerical methods. Such methods need to resolve accurately the shock regions and at the same time approximate with high accuracy the smooth parts of the solution.

It is a widely held belief that to achieve this goal one has to impose extraneous stabilization mechanisms, such as shock capturing terms or limiters (depending on the parameters of the problem, on the order of the method, on the particular form of the system, etc.). This approach seems to hold for those finite element or high-order finite volume methods previously developed [21, 10, 19, 11]. We refer to [11] for a comprehensive review of the current state of the art on the high-order finite difference, finite volume, and finite element methods for hyperbolic conservation laws; see also [17, 26].

*Received by the editors December 23, 2002; accepted for publication (in revised form) March 8, 2004; published electronically December 16, 2004. This work was partially supported by the European Union RTN-network HYKE.

http://www.siam.org/journals/sinum/42-4/42043.html

†Department of Applied Mathematics, University of Crete, 71409 Heraklion, Crete, Greece, and Institute of Applied and Computational Mathematics, FORTH, 71110 Heraklion, Crete, Greece (arvas@tem.uoc.gr).

‡Department of Applied Mathematics, University of Crete, 71409 Heraklion, Crete, Greece, and Institute of Applied and Computational Mathematics, FORTH, 71110 Heraklion, Crete, Greece (makr@math.uoc.gr, http://www.tem.uoc.gr/~makr).

§Department of Mathematics, University of Wisconsin, Madison, WI 53706, and Institute of Applied and Computational Mathematics, FORTH, 71110 Heraklion, Crete, Greece (tzavaras@math.wisc.edu, http://www.math.wisc.edu/~tzavaras).

Our motivation is to consider schemes designed to be used in conjunction with appropriate mesh refinement. It is conceivable that successful adaptive schemes may not need to be stabilized by using extra stabilization operators (such as limiters or shock capturing terms) accounting in turn for stabilization by the natural diffusion or relaxation mechanisms of the problem plus an appropriate mesh selection. A successful application of this idea requires one to have at hand a stable, robust, and flexible method. Indeed, toward this goal finite elements are a natural choice, since the development of supportive structures (finite element spaces of any order, flexibility in mesh construction, etc.) in adaptive finite element literature and software implementation is at a remarkable level.

In this article we propose a class of finite element methods based on relaxation models and address stability and convergence issues. For these relaxation finite element schemes the stabilization mechanisms are the regularization by wave operators (coming from the relaxation model) and appropriate mesh refinement in the shock areas. Our adaptive finite element schemes are of the general type introduced in [4] and further developed in [2, 3]. There, alternative methods and mesh refinement strategies are extensively tested computationally. Preliminary results indicate that the adaptive relaxation finite element schemes are a robust and reliable alternative for shock computations.

1.1. Relaxation finite element approximations. Relaxation models that approximate (1.1) are the basis of our schemes. In particular, the model suggested in [20],

$$\begin{aligned} &\partial_t\, u \;+\sum_{j=1}^{d}\partial_{x_j} v_j \;=\; 0,\\ &\partial_t\, v_i \;+A_i\,\partial_{x_i}\, u \;=\; -\frac{1}{\varepsilon}\big(v_i - F_i(u)\big), \quad i=1,\dots,d, \end{aligned} \tag{1.2}$$

corresponds to the regularization of (1.1) by a wave operator of order ε. Here A_i are symmetric, positive definite matrices with constant coefficients that are selected to satisfy certain stability conditions, *the subcharacteristic conditions*; see [20, 43] and the next sections. This relaxation model induces a regularization mechanism with *finite speed of propagation* that results in a partial differential equation with linear principal part. In return, the number of unknowns is increased. Nevertheless, in schemes based on the discretization of (1.2) the extra cost is compensated for by the simplicity and the natural implicit-explicit discretization that this model admits. The relaxation finite element schemes are based on the direct finite element approximation of (1.2).

Let $\mathcal{T}_h = \{K\}$ be a decomposition of $\mathbb{R}^d$ into elements with the usual properties [7]. We use the notation $h_K = \operatorname{diam}(K)$, $h = \sup_{K\in\mathcal{T}_h} h_K < 1$, and $\underline{h} = \min_{K\in\mathcal{T}_h} h_K$. The standard conforming finite element space S_k is defined by

$$S_k = \{\phi \in C^0\,(\mathbb{R}^d)^n \;:\; \phi\,|_K \in \mathbb{P}_k,\; K\in\mathcal{T}_h,\;\; \phi\,|_{\Omega^C} \equiv\; 0\}. \tag{1.3}$$

Here we assume that the initial values have compact support and thus, for all $t\in[0,T]$, our solution will vanish outside some compact set $\Omega\subset\mathbb{R}^n$. Clearly, $S_k \subset H_0^1(\Omega)$; see [7] for the approximation properties of S_k into Sobolev spaces. Further, we introduce a finite element space consisting of piecewise discontinuous polynomials:

$$V_{k-1} = \{\psi \in L^2(\mathbb{R}^d)^n \;:\; \psi\,|_K \in \mathbb{P}_{k-1},\; K\in\mathcal{T}_h,\;\; \psi\,|_{\Omega^C} \equiv\; 0\}. \tag{1.4}$$

By construction $\partial_{x_i}\phi \in V_{k-1}$ for all $\phi\in S_k$.

The schemes under consideration are obtained by a direct discretization (without adding additional diffusion terms) of (1.2). The approximation of u is sought in the space S_k and the approximations of the relaxation variables v_i in V_{k-1}; that is, find $(u_h, v_{h,1}, \dots, v_{h,d}) : (0,T] \to S_k \times (V_{k-1})^d$ such that

(1.5)
$$\begin{aligned}
&(\partial_t\, u_h, \phi) \;-\sum_{j=1}^{d} (v_{h,j}, \partial_{x_j}\, \phi) \;=\; 0 \quad \forall\, \phi \in S_k,\\
&(\partial_t\, v_{h,i}, \psi) \;+\; (A_i\, \partial_{x_i}\, u_h, \psi) \;=\; -\frac{1}{\varepsilon}\big(v_{h,i} - F_i(u_h), \psi\big) \quad \forall\, \psi \in V_{k-1}, i = 1, \dots, d,
\end{aligned}$$

with initial conditions $u_h(0) = \Pi_S u_0$ and $v_{h,i}(0) = \Pi_V F_i(u_0)$, where Π_S and Π_V are nodal interpolants on S_k and V_{k-1}, respectively. We note that (1.5) is a semidiscrete scheme since we have discretized only the spatial variable, in the sense that for any fixed $t \in [0,T]$, $u_h(\cdot, t) \in S_k$. In section 2 we show that if u_h solves (1.5), then it satisfies

(1.6)
$$\begin{aligned}
&(\partial_t\, u_h, \phi) + \sum_{i=1}^{d} \big(\partial_{x_i} F_i(u_h), \phi\big)\\
&\qquad +\; \varepsilon\, \left((\partial_{tt}\, u_h, \phi) + \sum_{i=1}^{d} (A_i\, \partial_{x_i}\, u_h, \partial_{x_i}\, \phi) \right) \;= 0 \quad \forall \phi \in S_k.
\end{aligned}$$

In the stability analysis we work with (1.6) but note that (1.5) is better suited to explicit-implicit one-step discretizations in time. Time discretizations based on (1.6) are also possible; see section 3.

The method is comparable, in terms of computational performance, with the fully conforming discretization of the relaxation model considered in [4]: find $(u_h, v_{h,1}, \dots, v_{h,d}) : (0,T] \to (S_k)^{d+1}$ such that

(1.7)
$$\begin{aligned}
&(\partial_t\, u_h, \phi) \;-\sum_{j=1}^{d} (v_{h,j}, \partial_{x_j}\, \phi) \;=\; 0 \quad \forall\, \phi \in S_k,\\
&(\partial_t\, v_{h,i}, \psi) \;+\; (A_i\, \partial_{x_i}\, u_h, \psi) \;=\; -\frac{1}{\varepsilon}\big(v_{h,i} - F_i(u_h), \psi\big) \quad \forall\, \psi \in S_k, i = 1, \dots, d.
\end{aligned}$$

The corresponding one field equation to (1.7) takes the form

(1.8)
$$\begin{aligned}
&(\partial_t\, u_h, \phi) + \sum_{i=1}^{d} \big(\partial_{x_i} P\, F_i(u_h), \phi\big),\\
&\qquad +\; \varepsilon\, \left((\partial_{tt}\, u_h, \phi) + \sum_{i=1}^{d} (A_i\, P\, \partial_{x_i}\, u_h, P\, \partial_{x_i}\, \phi) \right) \;= 0 \quad \forall \phi \in S_k,
\end{aligned}$$

where P is the L^2-projection operator onto S_k. Note that, when discretizing (1.7) in time with an explicit scheme, the computation of u_h will require the inversion of $d+1$ systems with the same mass matrix. The same procedure in (1.5) will require only the inversion of one mass matrix.

Based on the semidiscrete schemes one can devise various one-step implicit-explicit Runge–Kutta time discretizations [40, 4, 2, 3]. In the following sections we analyze the stability properties of semidiscrete as well as fully discrete schemes.

1.2. Stabilization by mesh refinement. Schemes (1.5) and (1.7) are indeed simple, but the relaxation mechanism alone does not provide the necessary stabilization required in the shock regions. Indeed, this is confirmed by coarse mesh numerical experiments; see section 6 and [4]. This also becomes evident by further examination of properties of the schemes. Consider the one-space dimensional ($d = 1$) system

$$\begin{aligned} &\partial_t u + \partial_x F(u) = 0, \quad x \in \mathbb{R},\ t > 0,\ \ u = u(x,t) \in \mathbb{R}^n, \\ &u(\cdot, 0) = u_0(\cdot) \end{aligned} \tag{1.9}$$

with u_0 of compact support and the associated finite element relaxation scheme. Following the argument in [4], it is seen that the effective equation of both schemes (1.5) and (1.7) in the case $n = 1,\ d = 1,\ q = 1$ is

$$\partial_t u + F(u)_x + \varepsilon\big[\partial_{tt} u - A\partial_{xx} u\big] + \beta\, h_{\mathrm{loc}}^2 F(u)_{xxx} = 0 \tag{1.10}$$

for some positive constant β. As expected, the finite element discretization induces a dispersion term which is linear in the flux variable. Applying the Chapman–Enskog expansion to (1.10) we obtain

$$\partial_t u + F(u)_x - \varepsilon\partial_x\big((c^2 - F'(u)^2)\,\partial_x u\big) + \beta\, h_{\mathrm{loc}}^2 F(u)_{xxx} = 0.$$

It is evident that to exclude approximations with oscillatory character near shocks or to avoid computing nonentropic solutions, the diffusion term should be dominant; see the relevant numerical example in section 6 and the literature on diffusion-dispersion approximations of conservation laws [28, 29]. This will enforce a condition of the form

$$h_{\mathrm{loc}} < o(\varepsilon), \tag{1.11}$$

where h_{loc} is the local mesh size close to the shock. On the other hand, the theoretical analysis in sections 2–4 provides convergence results under the slightly weaker condition

$$h_{\mathrm{loc}} < \gamma\,\varepsilon \tag{1.12}$$

for some constant γ. That is, the convergence results include even certain cases pertaining to nonclassical shocks. However, in practice typically mesh adaptivity selects the entropic solution, since it applies mesh refinement in a neighborhood of the shock. The extensive numerical experiments in [4, 2] and section 6 show that appropriate mesh refinement indeed stabilizes in a robust way the finite element relaxation schemes. Since the focus of this paper is the theoretical justification of the above schemes, we will not insist on the important problem of identifying appropriate mesh refinement strategies and refer to [4, 2, 3].

1.3. Stability and related properties. In what follows, we investigate the theoretical properties of the relaxation finite element schemes (1.5). It is shown that for a wide class of one-dimensional but also of multidimensional systems (1.1), the schemes are stable in the sense that they satisfy certain strong dissipation estimates; see Propositions 2.1, 2.3, 2.6, 3.1, 3.3, and 3.5. Similar estimates are satisfied by the relaxation model (1.2) [43, 18]. The strong dissipation estimates for relaxation approximations introduced in [43] are a basic tool in our analysis. In addition, nonstandard stability estimates for appropriate finite element projections are used in an essential way. The stability results are of interest since they justify the dissipative character of our schemes.

The stability estimates will also be used in conjunction with the compensated compactness framework to derive compactness conditions. Recall that a pair of functions $\eta = \eta(u)$, $q = q(u)$ are called an entropy-entropy flux pair (or entropy pair for short) if (η, q) solve the linear hyperbolic system

$$\nabla q = \nabla \eta \cdot \nabla F.$$

The existence and properties of entropy pairs have been extensively investigated (e.g., [15, 38]), and entropy pairs are used to describe the compactness properties of approximate solutions for certain one-dimensional systems of two conservation laws [42, 15, 38, 37].

In fact, we show that for the finite element relaxation scheme (1.5) with $d = 1$, the approximations u_h satisfy

$$\partial_t \eta(u_h) + \partial_x q(u_h) \subset \text{ compact set of } H^{-1}_{\text{loc}}(\mathcal{O}).$$

This condition suffices to apply the compensated compactness program for certain one-dimensional equations and systems (see section 4) and to obtain convergence for semidiscrete or fully discrete finite element schemes. Similar results appear to hold for the fully conforming methods (1.7), (1.8), but their verification requires additional technical estimations. This is largely because the presence of the projection P in the one field equation (1.8) will result in extra error terms in the stability analysis. This case will not be pursued here.

The estimates derived in the following sections are rather complicated. To focus on the ideas and to present the material in a readable way, we have chosen to work step by step to distinguish the cases:

- semidiscrete schemes with symmetric flux F',
- semidiscrete schemes and the system admits a convex entropy function,
- fully discrete schemes with symmetric flux F',
- fully discrete schemes and the system admits a convex entropy function, and
- semidiscrete and fully discrete schemes for multidimensional systems that admit a convex entropy function.

In summary, the results provide theoretical support to the use of finite element relaxation schemes by establishing stability for a wide class of systems and convergence in various cases.

1.4. Error estimates for smooth solutions. Since the schemes are based on the discretization of model (1.2), in section 5 we address the problem of error estimates for relaxation approximations. We consider a system endowed with a convex entropy. Let u be a smooth solution of (1.1) defined on a maximal interval of existence, and let U_ε be the smooth solution of the relaxation approximation (1.2). We show that

$$\|U_\varepsilon(t) - u(t)\|_{L^2} \le C(t,u)\,\varepsilon, \tag{1.13}$$

where the constant $C(t,u)$ depends on a strong norm of u and blows up at the critical time. The proof is based on a novel application of an idea of Dafermos [14, Thm. 5.2.1] to an error estimation. The difficulty posed by the relaxation approximation is handled by introducing a modified functional, corresponding to the relative entropy

$$H_R(u, U_\varepsilon) = \eta\big(U_\varepsilon + \varepsilon\partial_t(U_\varepsilon - u)\big) - \eta(u) - \eta'(u)\big(U_\varepsilon - u + \varepsilon\partial_t(U_\varepsilon - u)\big) \tag{1.14}$$

in the place of

$$H(u,w) = \eta(w) - \eta(u) - \eta'(u)(w-u) \tag{1.15}$$

used in [14]; see section 5 for details.

The finite element relaxation schemes are related to the central difference schemes of [33, 27]. One of their main common properties is that both schemes are Riemann solvers free and thus they combine high accuracy with simplicity. Finite element methods for hyperbolic conservation laws were considered in [21, 39, 22, 23, 19, 12, 10]. The theoretical properties of the streamline diffusion method were analyzed extensively (convergence, error estimates) in the scalar case [21, 39, 9]. The case of systems admitting entropy pairs is considered in [23] and it is shown that, for a streamline diffusion shock capturing method defined using the entropy variables, the bounded a.e. converging limits of approximations are weak entropy solutions of the system.

Finite element methods with discontinuous elements were proposed in [19] and [12]. In [12] stabilization is enforced by applying projection operators based on limiters. The above methods use piecewise polynomials of arbitrary degree and are formally of high order. Adaptive finite element methods based on a posteriori estimates have been considered in [22] for the ϵ-viscous approximation of one-dimensional systems of conservation laws. There exists a large literature on finite difference relaxation schemes; see, e.g., [20, 1, 25, 18] and [24] for relaxation schemes on unstructured grids.

The article is organized as follows. In section 2 we consider semidiscrete schemes and show stability and compactness of the dissipation measure for (i) case $d = 1$, F' is symmetric; (ii) case $d = 1$ and the system admits a convex entropy; and (iii) the multidimensional case. Section 3 is devoted to the analysis of implicit-explicit fully discrete schemes. The proofs are presented in a compact way, avoiding repetition of arguments already used in the semidiscrete case. In section 4 we discuss issues related to the application of compensated compactness to certain specific systems in order to conclude convergence of the schemes to a weak solution of (1.1). Section 5 is devoted to the error estimation between a smooth solution of (1.1) and the relaxation model (1.2). We conclude in section 6 with a discussion of implementation issues and present indicative examples reflecting the numerical performance of the method in two test cases.

2. Semidiscrete schemes: Stability estimates. We start by showing that the scheme (1.5) admits a field equation that is in fact a standard finite element discretization of the conservation law perturbed by a wave operator.

LEMMA 2.1. *If u_h solves* (1.5)*, then it satisfies* (1.6).

Proof. Select $\psi = \partial_{x_i}\phi, \quad \phi \in S_k$ in (1.5). Since $\psi \in V_{k-1}$ we have on summing with respect to $i,\ i = 1, \dots, d$,

$$\sum_{i=0}^{d}(\partial_t\, v_{h,i}, \partial_{x_i}\, \phi)\ +\ \sum_{i=1}^{d}(A_i\, \partial_{x_i}\, u_h, \partial_{x_i}\, \phi) = \ -\frac{1}{\varepsilon}\sum_{i=0}^{d}(v_{h,i} - F_i(u_h), \partial_{x_i}\, \phi).$$

Differentiating the first equation of (1.5) with respect to t we get

$$(\partial_{tt}\, u_h, \phi) - \ \sum_{j=1}^{d}(\partial_t v_{h,j},\ \partial_{x_j}\phi)\ =\ 0.$$

Hence,

$$\varepsilon(\partial_{tt}\, u_h, \phi)\ +\ \varepsilon\ \sum_{i=1}^{d}(A_i\, \partial_{x_i}\, u_h, \partial_{x_i}\, \phi) + \ \sum_{i=1}^{d}(v_{h,i}, \partial_{x_i}\, \phi)\ -\ \sum_{i=1}^{d}\big(F_i(u_h), \partial_{x_i}\, \phi\big)\ =\ 0.$$

Then by (1.5) we get the desired relation. □

In what follows, we establish stability properties for the finite element scheme (1.6). The stability estimates are proved consecutively for (i) case $d = 1$, F', symmetric; (ii) case $d = 1$, and the system admits a convex entropy; and (iii) the multidimensional case.

The one-dimensional semidiscrete finite element scheme takes the form

$$(\partial_t\, u_h, \phi) - \big(F(u_h), \partial_x \phi\big) + \varepsilon \Big((\partial_{tt}\, u_h, \phi) + (A\partial_x u_h, \partial_x\, \phi)\Big) = 0 \quad \forall \phi \in S_k. \tag{2.1}$$

For (2.1), we also prove compactness of the dissipation measure so as to apply the compensated compactness program and deduce convergence of the scheme in section 4. In the proof we use Murat's lemma [32].

LEMMA 2.2 (see Murat [32]). *Let $\mathcal{O}$ be an open subset of $\mathbb{R}^m$ and $\{\phi_j\}$ a bounded sequence of $W^{-1,p}(\mathcal{O})$ for some $p > 2$. In addition let $\phi_j = \chi_j + \psi_j$, where $\{\chi_j\}$ belongs in a compact set of $H^{-1}(\mathcal{O})$ and $\{\psi_j\}$ belongs in a bounded set of the space of measures $M(\mathcal{O})$. Then $\{\phi_j\}$ belongs in a compact set of $H^{-1}(\mathcal{O})$.*

2.1. The case $d = 1$ and F' is symmetric. Let $\phi = u_h$ in (2.1) and use $\big(F(u_h), \partial_x u_h\big) = 0$ to get

$$\partial_t \left[\int_\Omega \left(\frac{1}{2}|u_h|^2 + \varepsilon\, u_h\, \partial_t u_h\right) dx\right] + \varepsilon \int_\Omega [A\partial_x\, u_h \cdot \partial_x\, u_h - (\partial_t u_h)^2] dx = 0. \tag{2.2}$$

To estimate $\varepsilon \int_\Omega (\partial_t u_h)^2 dx$ let $\phi = \partial_t u_h$ in (2.1). Then,

$$\begin{aligned} &\|\partial_t\, u_h\|_{L^2}^2 + (F'(u_h)\, \partial_x\, u_h, \partial_t\, u_h) \\ &\qquad + \varepsilon \frac{1}{2} \partial_t \|\partial_t\, u_h\|_{L^2}^2 + \varepsilon \frac{1}{2} \partial_t (A\partial_x\, u_h, \partial_x\, u_h) = 0. \end{aligned} \tag{2.3}$$

Adding (2.2) with 2ε times (2.3) yields

$$\begin{aligned} &\frac{1}{2}\, \partial_t\, \|u_h + \varepsilon \partial_t\, u_h\|_{L^2}^2 + \varepsilon\, (A\partial_x\, u_h, \partial_x\, u_h) + 2\varepsilon (F'(u_h)\, \partial_x\, u_h, \partial_t\, u_h) \\ &\qquad + \varepsilon\, \|\partial_t\, u_h\|_{L^2}^2 + \frac{1}{2}\, \varepsilon^2\, \partial_t \Big\{\|\partial_t\, u_h\|_{L^2}^2 + 2(A\partial_x\, u_h, \partial_x\, u_h)\Big\} = 0. \end{aligned}$$

Since F' is symmetric, we have

$$\|\partial_x F(u_h)\|_{L^2}^2 = \big({F'}^2(u_h) \partial_x u_h, \partial_x u_h\big),$$

and we obtain

$$\begin{aligned} &\frac{1}{2}\, \partial_t \Big\{\|u_h + \varepsilon \partial_t\, u_h\|_{L^2}^2 + \varepsilon^2 \|\partial_t u_h\|_{L^2}^2 + 2\varepsilon^2 (A\partial_x\, u_h, \partial_x\, u_h)\Big\} \\ &\qquad + \varepsilon \|\partial_t\, u_h + \partial_x F(u_h)\|_{L^2}^2 + \varepsilon \big([A - {F'}^2(u_h)] \partial_x u_h, \partial_x u_h\big) = 0. \end{aligned}$$

We conclude with the following.

PROPOSITION 2.1. *Assume that $F'(u), A$ are symmetric and satisfy for some $\nu > 0$*

$$A - F'(u)^2 \geq \nu\, I, \qquad u \in \mathbb{R}^n. \tag{2.4}$$

Then the finite element approximation (2.1) *satisfies*

$$\begin{aligned}\int_\Omega \Big(|u_h + \varepsilon\, \partial_t\, u_h|^2 \;+\; \varepsilon^2 |\partial_t\, u_h|^2 \;+\; 2\varepsilon^2 A\partial_x\, u_h \;\cdot\; \partial_x\, u_h\Big)\\ + 2\int_0^t \int_\Omega \Big(\varepsilon|\partial_t u_h + F'(u_h)\partial_x\, u_h|^2 \;+\; \varepsilon\,\nu|\partial_x\, u_h|^2\Big)\\ \leq \int_\Omega |u_h^0 + \varepsilon\,\partial_t\, u_h(0)|^2 + \;\varepsilon^2|\partial_t\, u_h(0)|^2 + \;2\varepsilon^2 A\partial_x\, u_h^0 \cdot \partial_x\, u_h^0 \;=: C(u_h^0).\end{aligned}$$

In what follows we prove the next proposition.

PROPOSITION 2.2. *Let (η, q) be an entropy pair satisfying*

$$\|\eta\|_{L^\infty},\ \|q\|_{L^\infty},\ \|\eta'\|_{L^\infty},\ \|\eta''\|_{L^\infty} \;\leq\; C.$$

Then, for $h \leq \gamma\,\varepsilon$, there holds

$$\eta(u_h)_t \;+\; q(u_h)_x \text{ lies in a compact set of } H^{-1}_{loc}(\mathbb{R}\times\mathbb{R}^+).$$

Proof. Let (η, q) be an entropy pair and $\phi \in C_c^\infty(\mathbb{R}\times[0,\infty))$ a test function, and $\operatorname{supp}\phi \subset \tilde\Omega \times [0,\tilde T] =: Q$. We denote by $\Pi : L^2(\Omega) \to S_k$ a projection operator onto the finite element space of u_h to be determined later. Using the definition of the scheme we have

$$\begin{aligned}\Big(\eta(u_h)_t \;+\; q(u_h)_x, \phi\Big) &= \Big(\eta'(u_h)\big[u_{h,t} \;+\; F'(u_h)u_{h,x}\big], \phi\Big)\\ &= \Big(\big[u_{h,t} + F'(u_h)u_{h,x}\big], \Pi\big(\eta'(u_h)\,\phi\big)\Big)\\ &\quad + \Big(\big[u_{h,t} + F'(u_h)u_{h,x}\big], \eta'(u_h)\phi \;-\; \Pi(\eta'(u_h)\phi)\Big)\\ &= -\,\varepsilon\Big(A\partial_x\, u_h, \big[\Pi\big(\eta'(u_h)\phi\big)\big]_x\Big) - \varepsilon\Big(u_{h,tt}, \Pi\big(\eta'(u_h)\phi\big)\Big)\\ &\quad + \Big(\big[u_{h,t} + F'(u_h)u_{h,x}\big],\ \eta'(u_h)\phi - \Pi\big(\eta'(u_h)\phi\big)\Big).\end{aligned} \tag{2.5}$$

We select now $\Pi : L^2(\Omega) \to S_k$ to be the L^2-projection onto S_k. Π satisfies

$$(\Pi\omega, \phi) \;=\; (\omega, \phi) \qquad \forall \phi \in S_k,\ \omega \in L^2(\Omega), \tag{2.6}$$

$$\|\Pi\omega - \omega\|_{L^2(\Omega)} \;=\; \inf_{\chi\in S_k} \|\omega - \chi\|_{L^2(\Omega)} \;\leq\; C\,h\,\|\omega_x\|_{L^2(\Omega)}, \quad \omega \in H^1(\Omega), \tag{2.7}$$

as well as the stability estimate [13]

$$\|(\Pi\omega)_x\|_{L^2(\Omega)} \;\leq\; C\|\omega_x\|_{L^2(\Omega)}, \quad \omega \in H^1(\Omega). \tag{2.8}$$

We are ready to bound the terms in the right-hand side of (2.5). Indeed, (2.8) implies

$$\begin{aligned}&\varepsilon\Big|\Big(A\,u_{h,x}, \big[\Pi\big(\eta'(u_h)\phi\big)\big]_x\Big)\Big| \leq \varepsilon\, C\|u_{h,x}\|_{L^2(\Omega)}\,\|\big(\eta'(u_h)\phi\big)_x\|_{L^2(\Omega)}\\ &\leq C\Big(\varepsilon\int_\Omega |u_{h,x}|^2\Big)\|\eta''\|_{L^\infty(\Omega)}\|\phi\|_{C^0(\Omega)} + \varepsilon^{1/2}C\Big(\varepsilon\int_\Omega |u_{h,x}|^2\Big)^{1/2}\|\eta'\|_{L^\infty(\Omega)}\|\phi_x\|_{L^2(\Omega)}.\end{aligned} \tag{2.9}$$

Next, since $u_{h,tt} \in S_k$ and by (2.6),

$$
\begin{aligned}
&- \varepsilon \int_0^t \int_\Omega u_{h,tt}\, \Pi\Big(\eta'(u_h)\phi\Big) dxds \;=\; -\varepsilon \int_0^t \int_\Omega u_{h,tt}\, \eta'(u_h)\phi dxds \\
&= \varepsilon \int_0^t \int_\Omega u_{h,t}\big(\eta'(u_h)\phi\big)_t dxds + \varepsilon \int_\Omega u_{h,t}\, \eta'(u_h)\phi\Big|_{s=0} dx - \varepsilon \int_\Omega u_{h,t}\, \eta'(u_h)\phi\Big|_{s=t} dx.
\end{aligned}
$$

By Proposition 2.1 we have

(2.10)
$$
\varepsilon\Big|\int_\Omega u_{h,t}\, \eta'(u_h)\phi(t)dx\Big| \le \varepsilon\Big(\int_\Omega u_{h,t}^2\Big)^{1/2} \|\eta'\|_{L^\infty(\Omega)}\, \|\phi\|_{C^0(\Omega)}\, m(\Omega)^{1/2} \le \; C_\Omega \|\phi\|_{C^0(\Omega)},
$$

and as before

(2.11)
$$
\begin{aligned}
\varepsilon\Big|\int_0^t \int_\Omega u_{h,t}\big(\eta'(u_h)\phi\big)_t dxdt\Big| \;\le\; & C\Big(\varepsilon \int_0^t \int_\Omega |u_{h,t}|^2\Big)\|\eta''\|_{L^\infty(Q)}\, \|\phi\|_{C^0(Q)} \\
& + \; \varepsilon^{1/2}\Big(\varepsilon \int_0^t \int_\Omega |u_{h,t}|^2\Big)^{1/2} \|\eta'\|_{L^\infty(Q)}\, \|\phi_t\|_{L^2(Q)}.
\end{aligned}
$$

To estimate the last term in (2.5), note that $\eta'(u_h)\phi \in H^1(\Omega)$ and thus

$$
\begin{aligned}
\|\eta'(u_h)\phi - \Pi\big(\eta'(u_h)\phi\big)\|_{L^2(\Omega)} \;&\le Ch\|\big(\eta'(u_h)\phi\big)_x\|_{L^2(\Omega)} + Ch\|\eta'(u_h)\phi_x\|_{L^2(\Omega)} \\
&\le Ch\|\eta''\|_{L^\infty(\Omega)}\, \|u_{h,x}\|_{L^2(\Omega)}\, \|\phi\|_{C^0(\Omega)} + Ch\|\eta'\|_{L^\infty(\Omega)}\, \|\phi_x\|_{L^2(\Omega)}.
\end{aligned}
$$

By (2.4) we have $\|F'(u_h)^2\|_{L^\infty(\Omega)} \le C$; therefore

$$
\begin{aligned}
&\Big|\Big(\big[u_{h,t} + F'(u_h)u_{h,x}\big],\, \eta'(u_h)\phi - \Pi\big(\eta'(u_h)\phi\big)\Big)\Big| \\
&\qquad \le\; C\Big(h \int_\Omega (|u_{h,t}|^2 dx \;+\; |\partial_x u_h|^2) dx\Big)\|\phi\|_{C^0(\Omega)} \\
&\qquad\quad + h\,\Big(\int_\Omega (|u_{h,t}|^2 \;+\; |\partial_x u_h|^2)\, dx\Big)^{1/2} \|\phi_x\|_{L^2(\Omega)}.
\end{aligned} \tag{2.12}
$$

Combining (2.9)–(2.12) and using Murat's Lemma 2.2 (in our case, $\chi_h \to 0$ in H^{-1} and is thus precompact in H^{-1}), we complete the proof. ∎

2.2. The case $d = 1$, and the system admits a convex entropy. The case that F' is not necessarily symmetric but the system is equipped with a convex entropy η is examined next. In this case the system is symmetrizable. The finite element approximations (1.5) enjoy the same a priori bounds with the continuous solution of the relaxation model considered in [43]. Indeed, the following proposition holds.

PROPOSITION 2.3. *Let* (1.9) *be equipped with a strictly convex entropy* $\eta(u)$ *satisfying for some* $\alpha > 0$

$$
\frac{1}{\alpha} I \le \eta''(u) \le \alpha\, I, \quad u \in \mathbb{R}^n. \tag{2.13}
$$

Assume for some $M > 0$ *we have* $|F'(u)| \le M$ *for* $u \in \mathbb{R}^n$ *and that the positive definite, symmetric matrix* A *is selected to satisfy, for* $\overline{\alpha} = 2\alpha \max\{\beta, 1\}$, β *as in* (2.22) *and some* $\nu > 0$,

$$
\frac{1}{2}\big((\eta''(u)A)^T + \eta''(u)A\big) - \overline{\alpha} F'(u)^T F'(u) \ge \nu\, I \qquad \text{for } u \in \mathbb{R}^n. \tag{2.14}
$$

Then there is $\gamma = \gamma(\alpha, \beta, M, \nu) > 0$ *such that, for*

$$h \leq \gamma\, \varepsilon \tag{2.15}$$

and for some positive constants c_1, c_2, *and* c_3, *the finite element approximation* (2.1) *satisfies the stability estimate*

$$\begin{aligned}
&\int_\Omega \Big(\eta(u_h + \varepsilon\, \partial_t\, u_h) + \varepsilon^2 c_1 \big[|\partial_t\, u_h|^2 + A\partial_x\, u_h\, \cdot\, \partial_x\, u_h\big]\Big) dx \\
&\quad + \varepsilon c_2 \int_0^t \int_\Omega \Big(|\partial_t u_h + F'(u_h)\partial_x\, u_h|^2 + |\partial_x\, u_h|^2 + |\partial_t u_h|^2\Big) dx dt \\
&\quad \leq \int_\Omega \Big(\eta(u_h^0 + \varepsilon\, \partial_t\, u_h(0)) + \varepsilon^2 c_3 \big[|\partial_t\, u_h(0)|^2 + A\partial_x\, u_h^0 \cdot \partial_x\, u_h^0\big]\Big) dx.
\end{aligned} \tag{2.16}$$

Remark 2.1. We are interested here in data and associated finite element approximations u_h that are of compact support. It is thus natural to normalize η so that $\eta(0) = 0$ and $\eta'(0) = 0$. This can always be achieved, because if (η, q) is an entropy pair, then

$$\eta(u) - \eta(0) - \eta'(0)u, \quad q(u) - q(0) - \eta'(0)(F(u) - F(0))$$

is also an entropy pair. In view of (2.13), the normalized η is equivalent to the Euclidean norm, $\eta(u) \sim |u|^2$. Thus the stability framework in Proposition 2.3 is that of L^2.

Using the stability estimate, it is easy to see that strong convergence of the finite element approximations gives a weak solution that satisfies the integral version of the entropy inequality.

PROPOSITION 2.4. *Under the hypotheses of Proposition* 2.3, *if*

$$u_h \to u \quad \text{in } L^2_{x,t} \text{ and a.e.,} \tag{2.17}$$

then u *is a weak solution of* (1.9) *that satisfies*

$$\int_\Omega \eta(u(x,t))dx \leq \int_\Omega \eta(u^0(x))dx \qquad \text{for a.e. } t. \tag{2.18}$$

Proof. We assume with no loss of generality that $F(0) = 0$ and note that $|F(u)| \leq M|u|$. Let $u^0 \in H_0^1$ and be of compact support, let $v^0 = F(u^0) \in H_0^1$ and be of compact support, and define the approximations $u_h^0 \in S_k$ and $v_h^0 \in V_{k-1}$ defined by $v_h^0 = \Pi_{V_{k-1}} F(u_h^0)$ with $\Pi_{V_{k-1}}$ the L^2-projection. Let $u_h = u_h(x,t)$, $v_h = v_h(x,t)$ be the solution of the semidiscrete scheme. Note that $\partial_t u_h(0) = \Pi_{S_k} \partial_x F(u_h^0)$, where Π_{S_k} is the L^2-projection onto S_k.

For $\phi(x) \in S_k$ and $\theta(t) \in C_c^\infty([0,\infty))$ we have

$$\begin{aligned}
&-\int_0^t \int_\Omega [u_h \phi \partial_t \theta + F(u_h)\partial_x \phi\theta - \epsilon A \partial_x u_h \cdot \partial_x \phi\theta + \epsilon \partial_t u_h \cdot \phi \partial_t \theta] dx dt \\
&\quad - \int_\Omega (u_h^0 \phi\theta(0) + \epsilon \partial_t u_h(0) \phi\theta(0)) dx = 0.
\end{aligned} \tag{2.19}$$

Note that

$$\begin{aligned}
&u_h \to u, \;\; F(u_h) \to F(u) \quad \text{in } L^2_{x,t} \text{ and a.e.,} \\
&u_h^0 \to u^0, \;\; \varepsilon \partial_t u_h(0) \to 0 \quad \text{in } L^2_x \text{ and (along a subsequence) a.e.,} \\
&\varepsilon^{\frac{1}{2}} \|\partial_x u_h\|_{L^2_{x,t}} + \varepsilon^{\frac{1}{2}} \|\partial_t u_h\|_{L^2_{x,t}} \leq O(1).
\end{aligned}$$

Using that tensor products $\phi(x) \otimes \theta(t)$, $\phi \in S_k$, $\theta \in C_c^\infty([0,\infty))$ are dense as $h \to 0$ in $C^2(\bar{\Omega})$ for Ω bounded, we pass to the limit in (2.19) and obtain that u is a weak solution of (1.9). Using Fatou's lemma, we pass to the limit ε, $h \to 0$ in (2.16) to deduce

$$\int_\Omega \eta(u(x,t))\,dx \le \liminf_{h\to 0,\,\varepsilon\to 0} \int_\Omega \eta(u_h + \varepsilon \partial_t u_h)\,dx \le \int_\Omega \eta(u_0(x))\,dx$$

and conclude. ∎

To show the stability estimate we use the elliptic projection operator onto S_k and its approximation and stability properties. To this end let $P_1 : H_0^1 \to S_k$ be the Ritz (elliptic) projection defined by

$$(A\partial_x\, P_1 v, \partial_x \phi) = (A\partial_x\, v, \partial_x \phi) \qquad \forall \phi \in S_k,\ v \in H_0^1. \tag{2.20}$$

It is a standard result that P_1 satisfies

$$\begin{aligned} \|P_1\omega - \omega\|_{L^2(\Omega)} &\le C\,h\,\|\omega_x\|_{L^2(\Omega)}, \quad \omega \in H_0^1, \\ \|(P_1\omega)_x\|_{L^2(\Omega)} &\le C\|\omega_x\|_{L^2(\Omega)}, \quad \omega \in H_0^1. \end{aligned} \tag{2.21}$$

The second bound is a direct consequence of the definition and the first is obtained by a standard duality argument using once more the second bound (see [7, Thm. 5.4.8]). The following nonstandard stability property of P_1 will be crucial in the proof of Proposition 2.3. It uses in an essential way the stability analysis of the finite element method by mesh-dependent norms due to Babuška and Osborn [5].

LEMMA 2.3. *Let η be a strictly convex entropy and $v_h \in S_k$. Under hypothesis* (2.13)*, there exists a positive constant β such that*

$$(v_h, P_1\,[\,\eta''(w)(v_h)\,]\,) \le \beta\,\|\eta''(w)\|_{L^\infty(\Omega)}\,\|v_h\|^2_{L^2(\Omega)} \quad \forall w \in S_k. \tag{2.22}$$

Proof. It is known that P_1 is not stable with respect to $L^2(\Omega)$ [5]. Its stability with respect to the mesh-dependent L^2-like norm

$$\|v\|_{0,h,\Omega} = \left(\|v\|^2_{L^2(\Omega)} + \sum_j \delta_j |v(x_j)|^2 \right)^{1/2}, \tag{2.23}$$

where x_j are the nodes of the partition and $\delta_j = (x_{j+1} - x_{j-1})/2$ is as shown in [5], and

$$\|P_1\,v\|_{0,h,\Omega} \le \beta_1 \|v\|_{0,h,\Omega}, \tag{2.24}$$

where β_1 is a positive constant independent of h. Thus, (2.24) implies

$$\|P_1\,[\,\eta''(w)(v_h)\,]\,\|_{L^2(\Omega)} \le \beta_1 \|\eta''(w)\|_{L^\infty(\Omega)}\,\|v_h\|_{0,h,\Omega}. \tag{2.25}$$

But in the finite element space local inverse inequalities imply

$$\|v_h\|_{0,h,\Omega} \le \beta_2 \|v_h\|_{L^2(\Omega)} \qquad \forall v_h \in S_k, \tag{2.26}$$

with β_2 independent of h [7, 5]. Therefore, (2.22) follows with $\beta = \beta_1\,\beta_2$. ∎

Proof of Proposition 2.3. The finite element approximation u_h satisfies (2.1). Setting $\phi = P_1\,\eta'(u_h)$ and using (2.20), we obtain after a rearrangement

$$\begin{aligned}(\partial_t\, u_h, \eta'(u_h))+\big(\partial_x F(u_h), \eta'(u_h)\big) + \varepsilon(\partial_{tt} u_h, P_1\eta'(u_h)) + \varepsilon(A\partial_x u_h, \partial_x \eta'(u_h))\\ = (\partial_t\, u_h, \eta'(u_h) - P_1\,\eta'(u_h))+\big(\partial_x F(u_h), \eta'(u_h) - P_1\,\eta'(u_h)\big)\\ =: Z_1 + Z_2.\end{aligned} \tag{2.27}$$

The terms in the right-hand side will be estimated in what follows. First we examine the stability properties of the left-hand side. Since P_1 commutes with time differentiation,

$$\begin{aligned}\varepsilon(\partial_{tt}\, u_h, P_1\,\eta'(u_h)) &= \varepsilon\partial_t\,(\partial_t\, u_h, P_1\,\eta'(u_h)) - \varepsilon(\partial_t\, u_h, P_1\,[\,\eta''(u_h)\partial_t\, u_h\,])\\ &= \varepsilon\partial_t\,(\partial_t\, u_h, \eta'(u_h)) - \varepsilon(\partial_t\, u_h, P_1\,[\,\eta''(u_h)\partial_t\, u_h\,])\\ &\quad - \varepsilon\partial_t\,(\partial_t\, u_h, \eta'(u_h) - P_1\eta'(u_h)).\end{aligned} \tag{2.28}$$

We thus have

$$\begin{aligned}\partial_t \int_\Omega \eta(u_h)+ \int_\Omega \partial_x q(u_h) + \varepsilon\partial_t\,(\partial_t\, u_h, \eta'(u_h))\\ +\varepsilon(A\partial_x\, u_h, \eta''(u_h)\partial_x\, u_h) - \varepsilon(\partial_t\, u_h, P_1\,[\,\eta''(u_h)\partial_t\, u_h\,])\\ = Z_1 + Z_2 + Z_3,\end{aligned} \tag{2.29}$$

where the new term Z_3 is given by

$$Z_3 = \varepsilon\partial_t\,(\partial_t\, u_h, \eta'(u_h) - P_1\eta'(u_h)) = \varepsilon\partial_t Z_1. \tag{2.30}$$

As in [43] the following identity will be important:

$$\begin{aligned}\int_\Omega \eta(u_h + \varepsilon\partial_t\, u_h)dx = \int_\Omega \eta(u_h)dx + \varepsilon(\eta'(u_h), \partial_t\, u_h)\\ + \varepsilon^2\left(\partial_t\, u_h, \left\{\int_0^1\int_0^s \eta''(u_h + \varepsilon\,\tau\partial_t\, u_h)d\tau ds\right\}\partial_t\, u_h\right).\end{aligned} \tag{2.31}$$

It is evident that we need to estimate $\varepsilon(\partial_t u_h, P_1[\,\eta''(u_h)\partial_t u_h\,])$. This is done by Lemma 2.3, which gives

$$\varepsilon|(\partial_t\, u_h, P_1\,[\,\eta''(u_h)\partial_t\, u_h\,])| \le \varepsilon\beta\,\|\eta''(u_h)\|_{L^\infty(\Omega)}\,\|\partial_t\, u_h\|^2_{L^2(\Omega)}. \tag{2.32}$$

We proceed to handle $\varepsilon\int_\Omega(\partial_t u_h)^2dx$. Observe that setting $\phi = \partial_t u_h$ in (2.1) gives

$$\|\partial_t\, u_h\|^2_{L^2(\Omega)} + (F'(u_h)\partial_x\, u_h, \partial_t\, u_h) + \varepsilon\frac{1}{2}\partial_t\|\partial_t\, u_h\|^2_{L^2(\Omega)} + \varepsilon\frac{1}{2}\partial_t(A\partial_x\, u_h, \partial_x\, u_h) = 0. \tag{2.33}$$

Next, define

$$\begin{aligned}\overline{\beta} &= \beta\,\|\eta''(u_h)\|_{L^\infty(\Omega)},\\ \overline{\eta''} &= \left\{\int_0^1\int_0^s \eta''(u_h + \varepsilon\,\tau\partial_t\, u_h)d\tau ds\right\},\\ \overline{\alpha} &= \max\{2\overline{\beta}, 2\alpha\}\end{aligned} \tag{2.34}$$

and note that $\overline{\beta} = \beta\alpha$, $\overline{\alpha} = 2\alpha \max\{1, \beta\}$. After summing (2.29) with $2\varepsilon\,\overline{\alpha}$ times (2.33), we arrive at

(2.35)
$$\begin{aligned}\partial_t \int_\Omega \Big(\eta(u_h + \varepsilon\partial_t u_h) + \varepsilon^2 \partial_t u_h \cdot \{\overline{\alpha}\, I - \overline{\eta''}\}\partial_t u_h + \varepsilon^2 \overline{\alpha} A\partial_x\, u_h \cdot \partial_x u_h\Big) dx - \varepsilon \partial_t Z_1 \\ + \varepsilon\,(\overline{\alpha} - \overline{\beta})\|\partial_t\, u_h\|^2_{L^2(\Omega)} + \varepsilon\overline{\alpha}\|\partial_t\, u_h\|^2_{L^2(\Omega)} + 2\varepsilon\overline{\alpha}(F'(u_h)\partial_x u_h, \partial_t u_h\,) \\ + \varepsilon(A\partial_x\, u_h, \eta''(u_h)\partial_x\, u_h) \le Z_1 + Z_2.\end{aligned}$$

But since

$$\begin{aligned}&\|\partial_t\, u_h\|^2_{L^2(\Omega)} + 2(F'(u_h)\;\partial_x\, u_h, \partial_t\, u_h\,) \\ &\quad = \|\partial_t u_h + F'(u_h)\partial_x u_h\|^2_{L^2(\Omega)} - (F'(u_h)^T F'(u_h)\partial_x u_h, \partial_x\, u_h)\end{aligned} \tag{2.36}$$

and

$$(A\partial_x\, u_h, \eta''(u_h)\partial_x\, u_h) = \frac{1}{2}\,\big((\eta''(u_h)A + (\eta''(u_h)A)^T\big)\partial_x\, u_h, \partial_x\, u_h\big),$$

we conclude by (2.13) and (2.14) that

(2.37)
$$\begin{aligned}\partial_t\Big\{ \int_\Omega \eta(u_h + \varepsilon\partial_t u_h)dx + \varepsilon^2\overline{\alpha}\|\partial_t u_h\|^2_{L^2(\Omega)} + \varepsilon^2\overline{\alpha}(A\partial_x u_h, \partial_x u_h) - \varepsilon Z_1\Big\} \\ + \varepsilon\overline{\beta}\|\partial_t\, u_h\|^2_{L^2(\Omega)} + \varepsilon\,\overline{\alpha}\,\|\partial_t\, u_h + F'(u_h)\;\partial_x\, u_h\|^2_{L^2(\Omega)} + \varepsilon\nu\;\|\partial_x\, u_h\|^2_{L^2(\Omega)} \\ \le Z_1 + Z_2.\end{aligned}$$

At this point $\overline{\alpha}$, $\overline{\beta}$, and ν are fixed. We now turn to the estimation of the Z_i. Observe that, by (2.13) and (2.21),

$$\begin{aligned}Z_1 &= (\partial_t\, u_h, \eta'(u_h) - P_1\,\eta'(u_h)) \\ &\le C\,h\,\|\partial_t\, u_h\|_{L^2(\Omega)}\;\|\partial_x\,\eta'(u_h)\|_{L^2(\Omega)} \\ &\le C\,h\,\|\partial_t\, u_h\|_{L^2(\Omega)}\;\|\eta''(u_h)\|_{L^\infty(\Omega)}\;\|\partial_x\, u_h\|_{L^2(\Omega)} \\ &\le Ch\,\alpha\,\|\partial_t\, u_h\|_{L^2(\Omega)}\;\|\partial_x\, u_h\|_{L^2(\Omega)},\end{aligned} \tag{2.38}$$

while

$$\begin{aligned}Z_2 &= \big(\partial_x F(u_h), \eta'(u_h) - P_1\,\eta'(u_h)\big) \\ &\le C\,h\,\|\partial_x\, u_h\|_{L^2(\Omega)}\;\|F'(u_h)\|_{L^\infty(\Omega)}\;\;\|\eta''(u_h)\|_{L^\infty(\Omega)}\;\|\partial_x\, u_h\|_{L^2(\Omega)} \\ &\le C\,h\,\alpha\,M\,\|\partial_x\, u_h\|^2_{L^2(\Omega)}.\end{aligned} \tag{2.39}$$

Next, we select h so that (i) the quadratic form in the first term of (2.37) is positive definite, and (ii) the terms Z_1 and Z_2 on the right of (2.37) can be absorbed to the left. This can be done provided $h \le \gamma\varepsilon$ for some $\gamma = \gamma(\alpha, \beta, M, \nu)$ positive and small. This gives (2.16) and concludes the proof. ∎

The compactness of the dissipation measure for the scheme is obtained by an argument similar to that in the symmetric case.

PROPOSITION 2.5. *For entropy pairs (η, q) satisfying*

$$\|\eta\|_{L^\infty},\ \|q\|_{L^\infty},\ \|\eta'\|_{L^\infty},\ \|\eta''\|_{L^\infty}\ \le\ C \tag{2.40}$$

and for $h \le \gamma\,\varepsilon$,

$$\eta(u_h)_t + q(u_h)_x \quad \textit{lies in a compact set of } H^{-1}_{loc}(\mathbb{R}\times\mathbb{R}^+). \tag{2.41}$$

Remark 2.2. Proposition 2.5 and the analogous statement for the symmetric case (Proposition 2.2) state that for entropy pairs satisfying (2.40) the entropy dissipation measure is controlled. They are used in section 4 to prove compactness of relaxation finite element approximations for the system (4.1). We note that entropy pairs (η, q) that satisfy (2.40) are constructed in [38] for the system (4.1) under hypotheses (4.3)–(4.4).

2.3. The multidimensional case. Next we consider multidimensional systems (1.1) for which the system is endowed with a uniformly convex entropy η. Let $(q_1,\dots,q_d)$ be the associated entropy flux, and

$$\begin{aligned} &q_i'(u) = \eta'(u)\,F_i'(u), \quad i=1,\dots,d,\\ &\eta''(u)\,F_i'(u) = F_i'(u)^T\,\eta''(u), \quad i=1,\dots,d. \end{aligned} \tag{2.42}$$

Still, in this case the finite element approximations (1.5) satisfy similar a priori bounds with the one-dimensional case, provided that each A_i is chosen to satisfy certain subcharacteristic conditions.

PROPOSITION 2.6. *Assume that* (1.1) *is equipped with a strictly convex entropy* $\eta(u)$ *that satisfies for some* $\alpha > 0$

$$\frac{1}{\alpha}\,I \le \eta''(v) \le \alpha\,I, \quad v\in\mathbb{R}^n; \tag{2.43}$$

let $\overline{\alpha} = 2\alpha\max\{1,\beta\}$ *with* β *as in* (2.46), *and assume that the symmetric, positive definite matrices* A_i *satisfy, for some* $\nu > 0$,

$$\sum_{j=1}^d \frac{1}{2}\big(A_j\eta''(v) + (A_j\eta''(v))^T\big)\xi_j\cdot\xi_j - \overline{\alpha}\left|\sum_{j=1}^d F_j'(v)\xi_j\right|^2 \ge \nu\sum_{j=1}^d|\xi_j|^2 \tag{2.44}$$
$$\forall \xi_1,\dots,\xi_d\in\mathbb{R}^n,\ v\in\mathbb{R}^n.$$

If $h \le \gamma\varepsilon$ *for some* $\gamma > 0$, *then the finite element approximations* (1.5) *satisfy, for some* $c_1,\ c_2 > 0$, *the stability estimate*

$$\begin{aligned} &\int_\Omega\left(\eta(u_h+\varepsilon\partial_t u_h) + \varepsilon^2 c_1\left[|\partial_t u_h|^2 + \sum_{i=1}^d A_i\partial_{x_i}u_h\cdot\partial_{x_i}u_h\right]\right)\\ &\qquad + \varepsilon c_2\int_0^t\int_\Omega\left(\left|\partial_t u_h + \sum_{i=1}^d F_i'(u_h)\partial_x u_h\right|^2 + \sum_{i=1}^d|\partial_{x_i}u_h|^2 + |\partial_t u_h|^2\right)\\ &\qquad \le C(u_h^0, \partial_t u_h(0)\,). \end{aligned}$$

The proof is entirely similar to the one-dimensional case presented before and therefore it will be omitted. Still, an essential tool in the analysis will be the elliptic projection $P_1 : H^1 \to S_k$ defined by

$$\sum_{i=1}^d(A_i\partial_{x_i}\,P_1 v, \partial_{x_i}\phi) = \sum_{i=1}^d(A_i\partial_{x_i}\,v, \partial_{x_i}\phi) \qquad \forall\phi\in S_k. \tag{2.45}$$

The multidimensional analogue of Lemma 2.3 still holds:

$$(v_h, P_1\,[\,\eta''(u_h)(v_h)\,]\,) \;\le\; \beta\,\|\eta''(w)\|_{L^\infty(\Omega)}\,\|v_h\|^2_{L^2(\Omega)}. \tag{2.46}$$

Its proof is based on the stability analysis of the finite element method by mesh-dependent norms [6]; see [16] for related results on stability of the elliptic projection in $L^2(\Omega)$. The quasi-uniformity assumption on the mesh in [6] needed to verify (2.24) can be relaxed along the lines of arguments presented in [16].

3. Fully discrete schemes. There are many alternative ways to perform the time discretization of (1.5) at the discrete time nodes $0, \kappa, 2\kappa, \dots$. In this section we consider a simple implicit-explicit time discretization. Seek $(u_h^n, v_{h,1}^n, \dots, v_{h,d}^n) \in S_k \times V_{k-1}^d$, $n = 0, 1, \dots$,

$$\begin{aligned}
&\left(\frac{u_h^{n+1} - u_h^n}{\kappa},\ \phi\right) - \sum_{i=1}^{d}(v_{h,i}^n,\ \partial_{x_i}\phi) = 0 \quad \forall \phi \in S_k,\\
&\left(\frac{v_{h,i}^{n+1} - v_{h,i}^n}{\kappa},\ \psi\right) + (A_i \partial_{x_i} u_h^{n+1},\ \psi) = -\frac{1}{\varepsilon}\big(v_{h,i}^{n+1} - F_i(u_h^{n+1}),\ \psi\big),\\
&\qquad\qquad \forall \psi \in V_{k-1}, \quad i = 1, \dots, d,
\end{aligned} \tag{3.1}$$

where $u_h^0 = u_0,\ v_{h,i}^0 = F_i(u_0)$, and $i = 1, \dots, d$.

When $d = 1$, the scheme takes the form

$$\begin{aligned}
&\left(\frac{u_h^{n+1} - u_h^n}{\kappa},\ \phi\right) - (v_h^n,\ \partial_x \phi) = 0 \quad \forall \phi \in S_k,\\
&\left(\frac{v_h^{n+1} - v_h^n}{\kappa},\ \psi\right) + (A \partial_x u_h^{n+1},\ \psi) = -\frac{1}{\varepsilon}\big(v_h^{n+1} - F(u_h^{n+1}),\ \psi\big) \quad \forall \psi \in V_{k-1}.
\end{aligned} \tag{3.2}$$

3.1. Properties of the scheme. For any sequence $\{\,Y^n\} \subset L^2(\Omega)$, define the operators $\overline{\partial}_t,\ \overline{\partial}_{tt}$:

$$\overline{\partial}_t Y^n := \frac{1}{\kappa}(Y^{n+1} - Y^n), \qquad \overline{\partial}_{tt} Y^n := \overline{\partial}_t \overline{\partial}_t Y^n.$$

Then the centered difference quotient that corresponds to the second time derivative at t^n is

$$\overline{\partial}_{tt} Y^{n-1} = \frac{1}{\kappa^2}(Y^{n+1} - 2Y^n + Y^{n-1}).$$

The following properties will prove useful (L^2 stands for $L^2(\Omega)$):

$$\begin{aligned}
(\overline{\partial}_t Y^n,\ Y^{n+1}) &= \frac{1}{2\kappa}\big[\,\|\,Y^{n+1}\|^2_{L^2} - \|\,Y^n\|^2_{L^2} + \|\,Y^{n+1} - Y^n\|^2_{L^2}\big]\\
&= \frac{1}{2}\big[\,\overline{\partial}_t\|\,Y^n\|^2_{L^2} + \kappa\|\,\overline{\partial}_t Y^n\|^2_{L^2}\big],
\end{aligned} \tag{3.3}$$

$$(\overline{\partial}_t Y^n,\ Y^n) = \frac{1}{2}\big[\,\overline{\partial}_t\|\,Y^n\|^2_{L^2} - \kappa\|\,\overline{\partial}_t Y^n\|^2_{L^2}\big], \tag{3.4}$$

$$\begin{aligned}
(\overline{\partial}_{tt} Y^n,\ \overline{\partial}_t Y^{n+1}) &= (\overline{\partial}_t W^n,\ W^{n+1}), \qquad W^n := \overline{\partial}_t Y^n,\ n = 0, 1, 2, \dots,\\
&= \frac{1}{2}\big[\,\overline{\partial}_t\|\,\overline{\partial}_t Y^n\|^2_{L^2} + \kappa\|\,\overline{\partial}_{tt} Y^n\|^2_{L^2}\big].
\end{aligned} \tag{3.5}$$

In addition one can verify that

$$
\begin{aligned}
(\overline{\partial}_{tt}Y^{n-1},\ Y^{n+1}) = \kappa(\overline{\partial}_{tt}Y^{n-1},\ \overline{\partial}_t Y^n) \\
+\overline{\partial}_t(\overline{\partial}_t Y^{n-1},\ Y^n)\ - \|\ \overline{\partial}_t Y^n\|^2_{L^2}.
\end{aligned} \tag{3.6}
$$

Now we have the following lemma.

LEMMA 3.1. *If u_h^n solves* (3.1)*, then it satisfies*

$$
(\overline{\partial}_t u_h^n,\ \phi) - \sum_{i=1}^d (F_i(u_h^n),\ \partial_{x_i}\phi) + \varepsilon\left((\overline{\partial}_{tt}u_h^{n-1},\ \phi) + \sum_{i=1}^d (A_i\partial_{x_i}u_h^n,\ \partial_{x_i}\phi)\right) = 0. \tag{3.7}
$$

Proof. For $\phi\in S_k$, we see that the solution of (3.1) satisfies

$$
\begin{aligned}
\sum_{i=1}^d (\overline{\partial}_t v_{h,i}^{n-1},\ \partial_{x_i}\phi) &= \sum_{i=1}^d \left(\frac{v_{h,i}^n - v_{h,i}^{n-1}}{\kappa},\ \partial_{x_i}\phi\right) \\
&\overset{(3.1)}{=} \left(\frac{\overline{\partial}_t u_h^n - \overline{\partial}_t u_h^{n-1}}{\kappa},\ \phi\right) = (\overline{\partial}_{tt}u_h^{n-1},\ \phi).
\end{aligned}
$$

Next, summing $i = 1,\dots,d$, (3.1), and using that $\partial_{x_i}\phi\in V_{k-1}$, we get

$$
\begin{aligned}
0 =& \sum_{i=1}^d (v_{h,i}^n,\ \partial_{x_i}\phi) - \sum_{i=1}^d (F_i(u_h^n),\ \partial_{x_i}\phi) + \varepsilon\sum_{i=1}^d (\overline{\partial}_t v_{h,i}^{n-1} + A_i\partial_{x_i}u_h^n,\ \partial_{x_i}\phi) \\
&\overset{(3.1)}{=} (\overline{\partial}_t u_h^n,\ \phi) - \sum_{i=1}^d (F_i(u_h^n),\ \partial_{x_i}\phi) + \varepsilon\sum_{i=1}^d (\overline{\partial}_t v_{h,i}^{n-1} + A_i\partial_{x_i}u_h^n,\ \partial_{x_i}\phi)
\end{aligned} \tag{3.8}
$$

and the result follows. ∎

In the case $d = 1$, we have

$$
(\overline{\partial}_t v_h^n,\ \partial_x\phi) = (\overline{\partial}_{tt}u_h^n,\ \phi), \tag{3.9}
$$

$$
(\overline{\partial}_t u_h^n,\ \phi) - (F(u_h^n),\ \partial_x\phi) + \varepsilon\big((\overline{\partial}_{tt}u_h^{n-1},\ \phi) + (A\partial_x u_h^n,\ \partial_x\phi)\big) = 0. \tag{3.10}
$$

3.2. The case $d = 1$ and F' symmetric. Let $\phi = 2u_h^{n+1} + 4\varepsilon\ \overline{\partial}_t u_h^n,$ in (3.10). Then

$$
\begin{aligned}
0 &= 2(\overline{\partial}_t u_h^n,\ u_h^{n+1})\ + 2(\partial_x F(u_h^n),\ u_h^{n+1}) \\
&+ 2\varepsilon(\overline{\partial}_{tt}u_h^{n-1},\ u_h^{n+1})\ + 2\varepsilon(A\ \partial_x u_h^n,\ \partial_x u_h^{n+1}) \\
&+ 4\varepsilon(\overline{\partial}_t u_h^n,\ \overline{\partial}_t u_h^n)\ + 4\varepsilon(\partial_x F(u_h^n),\ \overline{\partial}_t u_h^n) \\
&+ 4\varepsilon^2(\overline{\partial}_{tt}u_h^{n-1},\ \overline{\partial}_t u_h^n)\ + 4\varepsilon^2(A\ \partial_x u_h^n,\ \partial_x\overline{\partial}_t u_h^n).
\end{aligned} \tag{3.11}
$$

Using the properties of the discrete time operators listed above, the terms of (3.11) are handled as follows. First note

$$
2(\overline{\partial}_t u_h^n,\ u_h^{n+1}) = \overline{\partial}_t\|\ u_h^n\|^2_{L^2} + \kappa\|\ \overline{\partial}_t u_h^n\|^2_{L^2}.
$$

Also,

$$
2(\partial_x F(u_h^n),\ u_h^{n+1})\ = 2\ \kappa(F'(u_h^n)\ \partial_x u_h^n,\ \overline{\partial}_t u_h^n).
$$

The next term is estimated as

$$2\varepsilon(\overline{\partial}_{tt}u_h^{n-1},\ u_h^{n+1}) \overset{(3.6)}{=} 2\varepsilon\overline{\partial}_t(\overline{\partial}_t u_h^{n-1},\ u_h^n)\ - 2\varepsilon\|\overline{\partial}_t u_h^n\|_{L^2}^2\ + 2\varepsilon\kappa(\overline{\partial}_{tt}u_h^{n-1},\ \overline{\partial}_t u_h^n)$$
$$\geq 2\varepsilon\overline{\partial}_t(\overline{\partial}_t u_h^{n-1},\ u_h^n) - 2\varepsilon\|\overline{\partial}_t u_h^n\|_{L^2}^2 - 2\varepsilon^2\kappa\|\overline{\partial}_{tt}u_h^{n-1}\|_{L^2}^2 - \frac{\kappa}{2}\|\overline{\partial}_t u_h^n\|_{L^2}^2.$$

In addition,

$$2\varepsilon(A\ \partial_x u_h^n,\ \partial_x u_h^{n+1})\ = 2\varepsilon(A\ \partial_x u_h^n,\ \partial_x u_h^n)\ + 2\varepsilon\ \kappa(A\ \partial_x u_h^n,\ \partial_x\overline{\partial}_t u_h^n).$$

For the terms with coefficient 4ε we first note

$$4\varepsilon^2(\overline{\partial}_{tt}u_h^{n-1},\ \overline{\partial}_t u_h^n) \overset{(3.5)}{=} 2\varepsilon^2\ \overline{\partial}_t\|\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2 + 2\varepsilon^2\ \kappa\|\ \overline{\partial}_{tt}u_h^{n-1}\|_{L^2}^2$$

and

$$4\varepsilon^2(A\ \partial_x u_h^n,\ \partial_x\overline{\partial}_t u_h^n)\ = 4\varepsilon^2(W^n,\ \overline{\partial}_t W^n), \qquad W^n := A^{1/2}\partial_x u_h^n,\ n = 0, 1, 2, \dots,$$
$$\overset{(3.4)}{=} 2\varepsilon^2\ \overline{\partial}_t\|\ W^n\|_{L^2}^2 - 2\varepsilon^2\ \kappa\|\ \overline{\partial}_t W^n\|_{L^2}^2$$
$$= 2\varepsilon^2\ \overline{\partial}_t(A\ \partial_x u_h^n,\ \partial_x u_h^n)\ - 2\varepsilon^2\ \kappa(A\ \partial_x\overline{\partial}_t u_h^n,\ \partial_x\overline{\partial}_t u_h^n).$$

Summarizing, the terms with discrete time derivative that will appear in (3.11) are

$$\overline{\partial}_t\Big[\|\ u_h^n\|_{L^2}^2\ + 2\varepsilon(\overline{\partial}_t u_h^{n-1},\ u_h^n)\ + 2\varepsilon^2\|\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ + 2\varepsilon^2(A\ \partial_x u_h^n,\ \partial_x u_h^n)\Big]$$
$$= \overline{\partial}_t\Big[\|\ u_h^n\ + \varepsilon\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ + \varepsilon^2\|\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ + 2\varepsilon^2(A\ \partial_x u_h^n,\ \partial_x u_h^n)\Big].$$

In addition, the following calculation is useful:

$$2\varepsilon\|\overline{\partial}_t u_h^n\|_{L^2}^2 + 4\varepsilon(\partial_x F(u_h^n),\ \overline{\partial}_t u_h^n)$$
$$= \varepsilon\|\overline{\partial}_t u_h^n\|_{L^2}^2 + 2\varepsilon\|\ \frac{1}{\sqrt{2}}\ \overline{\partial}_t u_h^n + \sqrt{2}\ F'(u_h^n)\ \partial_x u_h^n\|_{L^2}^2$$
$$- 4\varepsilon((F'(u_h^n))^2\ \partial_x u_h^n,\ \partial_x u_h^n).$$

We conclude, therefore, that

$$\begin{aligned}
&\overline{\partial}_t\Big[\ \|\ u_h^n\ + \varepsilon\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ + \varepsilon^2\|\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ + 2\varepsilon^2(A\ \partial_x u_h^n,\ \partial_x u_h^n)\Big] \\
&\quad + \varepsilon\|\overline{\partial}_t u_h^n\|_{L^2}^2 + \frac{\kappa}{2}\|\overline{\partial}_t u_h^n\|_{L^2}^2\ + 2\varepsilon((A - 2(F'(u_h^n))^2)\ \partial_x u_h^n,\ \partial_x u_h^n) \\
&\quad \leq |2\kappa(\partial_x F(u_h^n),\ \overline{\partial}_t u_h^n)| + |2\varepsilon\ \kappa(A\ \partial_x u_h^n,\ \partial_x\overline{\partial}_t u_h^n)| \\
&\quad + 2\varepsilon^2\ \kappa(A\ \partial_x\overline{\partial}_t u_h^n,\ \partial_x\overline{\partial}_t u_h^n).
\end{aligned} \tag{3.12}$$

Next,

$$|2\kappa(\partial_x F(u_h^n),\ \overline{\partial}_t u_h^n)| \leq 4\kappa((F'(u_h^n))^2\ \partial_x u_h^n,\ \partial_x u_h^n) + \frac{\kappa}{4}\|\overline{\partial}_t u_h^n\|_{L^2}^2.$$

We will use the inverse inequality in S_k [7],

$$\|\partial_x\varphi\|_{L^2} \leq C_I\underline{h}^{-1}\|\varphi\|_{L^2} \qquad \forall\varphi \in S_k, \tag{3.13}$$

to obtain

$$|2\varepsilon\ \kappa(A\ \partial_x u_h^n,\ \partial_x \overline{\partial}_t u_h^n)| \le \varepsilon\ C_I\ \|A\|\ \frac{\kappa}{\underline{h}}\ \|\partial_x u_h^n\|_{L^2}^2 + \varepsilon\ C_I\ \|A\|\ \frac{\kappa}{\underline{h}}\ \|\overline{\partial}_t u_h^n\|_{L^2}^2,$$

$$2\varepsilon^2\ \kappa(A\ \partial_x \overline{\partial}_t u_h^n,\ \partial_x \overline{\partial}_t u_h^n) \le \varepsilon\ \frac{\varepsilon}{\underline{h}}\ \left(C_I^2\ \|A\|\ \frac{\kappa}{\underline{h}}\right)\ \|\overline{\partial}_t u_h^n\|_{L^2}^2.$$

Multiplying (3.12) by κ, and summing we finally conclude with the following proposition.

PROPOSITION 3.1. *We assume that $F'(u)$ is symmetric and that for given $\tilde{\beta}$ there holds*

$$\kappa \le \tilde{\beta}\varepsilon. \tag{3.14}$$

Assume further that we can choose A symmetric so that for some ν,

$$A - (\,2 + 4\tilde{\beta}\,)\,F'(u)^2 \ge \nu\ I \quad \textit{for}\ \ u \in \mathbb{R}^n. \tag{3.15}$$

Let $\gamma_{CFL} = C_I^2\,\|A\|\frac{\kappa}{\underline{h}}$ and assume that γ_{CFL} is sufficiently small and that

$$\varepsilon \le\ \frac{1}{2\,\gamma_{CFL}}\ \underline{h}.$$

Then the approximations of the fully discrete schemes satisfy the stability estimate

$$\begin{aligned}\|\ u_h^n\ &+\varepsilon\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ +\varepsilon^2\|\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ + 2\varepsilon^2(A\ \partial_x u_h^n,\ \partial_x u_h^n)\\ &+\sum_{j=1}^{n-1}\varepsilon\ \kappa\|\,\overline{\partial}_t u_h^j\|_{L^2}^2 + \sum_{j=1}^{n-1}\kappa^2\|\,\overline{\partial}_t u_h^j\|_{L^2}^2\ + \sum_{j=1}^{n-1}\varepsilon\ \kappa\ \|\partial_x u_h^j\|_{L^2}^2\ \le\ C(u_h^0).\end{aligned}$$

In what follows we study the compactness properties of the dissipation measure associated to the scheme. To this end we use the notation

$$\begin{aligned}&u_h \quad \text{denotes the piecewise linear in time}\\ &\quad\ \text{function such that}\quad u_h(t^n) = u_h^n,\\ &\overline{u}_h \quad \text{denotes the piecewise constant in time}\\ &\quad\ \text{function such that}\quad \overline{u}_h(t^n) = u_h^n,\ I_n\ =\ (t^n, t^{n+1}].\end{aligned} \tag{3.16}$$

PROPOSITION 3.2. *Under the assumptions of Proposition* 3.1*, for entropy pairs (η, q) such that*

$$\|\eta\|_{L^\infty},\ \|q\|_{L^\infty},\ \|\eta'\|_{L^\infty},\ \|\eta''\|_{L^\infty}\ \le\ C$$

and for $h\ \le\ C\,\varepsilon$ there holds

$$\eta(u_h)_t\ +\ q(u_h)_x \quad \textit{lies in a compact set of}\ H_{loc}^{-1}(\mathbb{R}\times\mathbb{R}^+), \tag{3.17}$$

where u_h is defined by (3.16).

Proof. Let (η, q) be an entropy pair and $\phi \in C_c^\infty(\mathbb{R}\times[0,\infty))$ a test function, and $\operatorname{supp}\phi \subset \tilde{\Omega}\times[0,\tilde{T}] =: Q$. Without loss of generality assume that $\tilde{T} = t^{m+1}$. Let

$\Pi : H^1 \to S_k$ be the L^2-projection onto the finite element space defined in (2.6). Then using the definition of the scheme we obtain

(3.18)
$$\begin{aligned}\int_0^{\tilde{T}} \Big(\partial_t \eta(u_h) \,+\, \partial_x q(u_h), \phi\Big) dt \;&=\; \int_0^{\tilde{T}} \Big(\eta'(u_h)\big[\partial_t u_h \,+\, \partial_x F'(u_h) u_h\big], \phi\Big) dt \\ &= -\varepsilon \sum_{j=0}^{j=m} \int_{I_j} \Big\{ \Big(A \partial_x\, u_h^j, \partial_x \big[\Pi\big(\eta'(u_h)\phi\big)\big]\Big) + \Big(\overline{\partial}_{tt} u_h^{j-1}, \Pi\big(\eta'(u_h)\phi\big)\Big)\Big\} dt \\ &+ \sum_{j=0}^{j=m} \int_{I_j} \Big\{ \Big(\big[\,\overline{\partial}_t u_h^j \,+\, \partial_x F'(\overline{u}_h)\big],\; \eta'(u_h)\phi - \Pi\big(\eta'(u_h)\phi\big)\Big) \\ &+ \Big(\eta'(u_h)\partial_x\big[\,F(u_h) - F(\overline{u}_h)\big], \phi\Big)\Big\} dt.\end{aligned}$$

Note here that for notational simplicity when we use $\overline{\partial}_t u_h^j, u_h^j, \overline{\partial}_{tt} u_h^{j-1}$ we mean the piecewise constant (with respect to t) functions that have these values in I_j. To proceed with the estimates, note that using (2.8) one obtains

(3.19)
$$\begin{aligned}&\varepsilon \sum_{j=0}^{j=m} \int_{I_j} \Big|\, \Big(A \partial_x\, u_h^j, \partial_x \big[\Pi\big(\eta'(u_h)\phi\big)\big]\Big) \,\Big| \, dt \\ &\qquad \le \varepsilon\, C \left(\kappa \sum_{j=0}^{m} \|\partial_x\, u_h^j\|^2_{L^2(\Omega)} \right)^{1/2} \|\partial_x\, \big(\eta'(u_h)\phi\big)\|_{L^2(Q)} \\ &\qquad \le C \left(\varepsilon\, \kappa \sum_{j=0}^{m} \|\partial_x\, u_h^j\|^2(\Omega) \right) \cdot \|\eta''\|_{L^\infty}\, \|\phi\|_{C^0(Q)} \\ &\qquad\quad + \varepsilon^{1/2}\, C \left(\varepsilon\, \kappa \sum_{j=0}^{m} \|\partial_x\, u_h^j\|^2(\Omega) \right)^{1/2} \|\eta\|_{L^\infty}\, \|\partial_x \phi\|_{L^2(Q)}.\end{aligned}$$

In addition, using the notation

$$\overline{v}_j = \kappa^{-1} \int_{I_j} v\, dt,$$

we have by (2.6),

(3.20)
$$\begin{aligned}-\varepsilon \sum_{j=0}^{m} \int_{I_j} \int_\Omega \overline{\partial}_{tt} u_h^{j-1}\, \Pi\Big(\eta'(u_h)\phi\Big) \;&=\; -\varepsilon \sum_{j=0}^{m} \int_{I_j} \overline{\partial}_{tt} u_h^{j-1} \int_\Omega \eta'(u_h)\phi \\ &= -\varepsilon \int_\Omega \sum_{j=0}^{m} \big(\overline{\partial}_t u_h^j - \overline{\partial}_t u_h^{j-1}\big)\, \overline{(\eta'(u_h)\phi)}_j \\ &= \varepsilon \int_\Omega \sum_{j=0}^{m-1} \overline{\partial}_t u_h^j \,\big(\, \overline{(\eta'(u_h)\phi)}_{j+1} - \overline{(\eta'(u_h)\phi)}_j \,\big) - \varepsilon \int_\Omega \overline{\partial}_t u_h^m \int_\Omega \overline{(\eta'(u_h)\phi)}_m.\end{aligned}$$

The stability in Proposition 3.1 implies, since $|\overline{v}_j| \le \| v \|_\infty$,

$$\begin{aligned} \varepsilon\Big|\int_\Omega \overline{\partial}_t u_h^m \, \overline{(\eta'(u_h)\phi)}_m \Big| &\le \varepsilon \, \|\overline{\partial}_t u_h^m\|_{L^2(\Omega)} \, \|\eta'\|_{L^\infty} \, \|\phi\|_{C^0(\Omega)} \, m(\Omega)^{1/2} \\ &\le C_\Omega \|\phi\|_{C^0(Q)}. \end{aligned} \tag{3.21}$$

Observing that $|\overline{v}_{j+1} - \overline{v}_j| = \frac{1}{\kappa} \left|\int_{I_j} \int_t^{t+\kappa} v_t \, ds \, dt\right| \le \int_{t_j}^{t_{j+2}} |\, v_t \,| \, dt$, we conclude

$$\begin{aligned} &\varepsilon\left|\int_\Omega \sum_{j=0}^{m-1} \overline{\partial}_t u_h^j \, (\overline{(\eta'(u_h)\phi)}_{j+1} - \overline{(\eta'(u_h)\phi)}_j)\right| \\ &\qquad \le C\left(\varepsilon\,\kappa \sum_{j=0}^{m} \|\overline{\partial}_t \, u_h^j\|^2_{L^2(\Omega)}\right)\|\eta''\|_{L^\infty} \, \|\phi\|_{C^0(Q)} \\ &\qquad + \varepsilon^{1/2}\left(\varepsilon\,\kappa \sum_{j=0}^{m} \|\overline{\partial}_t \, u_h^j\|^2_{L^2(\Omega)}\right)^{1/2} \|\eta'\|_{L^\infty} \, \|\partial_t\phi\|_{L^2(Q)}. \end{aligned} \tag{3.22}$$

Next,

$$\begin{aligned} &\|\eta'(u_h)\phi - \Pi\big(\eta'(u_h)\phi\big)\|_{L^2(\Omega)} \\ &\qquad \le Ch\|\eta''\|_{L^\infty} \, \|\partial_x u_h\|_{L^2(\Omega)} \, \|\phi\|_{C^0(\Omega)} + Ch\|\eta'\|_{L^\infty} \, \|\partial_x\phi\|_{L^2(\Omega)} \end{aligned}$$

and $\|F'(u)^2\|_{L^\infty} \le C$ (see (2.4)); therefore,

$$\begin{aligned} &\sum_{j=0}^{m} \int_{I_j} \Big|\Big(\big[\,\overline{\partial}_t u_h^j \;+\; F'(\overline{u}_h)\partial_x\overline{u}_h\big],\; \eta'(u_h)\phi - \Pi\big(\eta'(u_h)\phi\big)\Big)\Big| \\ &\qquad \le C\left(h\,\kappa \sum_{j=0}^{m} \|\overline{\partial}_t \, u_h^j\|^2_{L^2(\Omega)} + \|\partial_x \, u_h^j\|^2_{L^2(\Omega)}\right)\|\phi\|_{C^0(Q)} \\ &\qquad + h\left(\kappa \sum_{j=0}^{m} \|\overline{\partial}_t \, u_h^j\|^2_{L^2(\Omega)} + \|\partial_x \, u_h^j\|^2_{L^2(\Omega)}\right)^{1/2} \|\partial_x\phi\|_{L^2(Q)}. \end{aligned} \tag{3.23}$$

Finally, using the fact that $|u_h - \overline{u}_h| \le C\kappa|\partial_t u_h| = C\kappa|\overline{\partial}_t \, u_h^n|$, we have by using (3.14) and (3.15),

(3.24)

$$\begin{aligned} \sum_{j=0}^{m} \int_{I_j} \Big(\eta'(u_h)\partial_x\big[\,F(u_h) - F(\overline{u}_h)\big], \phi\Big) &= -\sum_{j=0}^{m} \int_{I_j} \Big(\big[\,F(u_h) - F(\overline{u}_h)\big], \partial_x(\eta'(u_h)\,\phi)\Big) \\ &\le C\left(\varepsilon\,\kappa \sum_{j=0}^{m} \|\overline{\partial}_t \, u_h^j\|^2_{L^2(\Omega)} + \|\partial_x \, u_h^j\|^2_{L^2(\Omega)}\right)\|\phi\|_{C^0(Q)} \\ &+ \varepsilon\left(\kappa \sum_{j=0}^{m} \|\overline{\partial}_t \, u_h^j\|^2_{L^2(\Omega)} + \|\partial_x \, u_h^j\|^2_{L^2(\Omega)}\right)^{1/2} \|\partial_x\phi\|_{L^2(Q)}. \end{aligned}$$

Combining (3.19)–(3.24), we obtain the desired result in view of Lemma 2.2. ∎

3.3. The case $d = 1$, and the system admits a convex entropy. The case that F' is not necessarily symmetric but the system is equipped with a convex entropy η will be briefly examined here. The analysis in this case mainly uses a combination of arguments from the corresponding semidiscrete case and the analysis of the fully discrete scheme in the symmetric case. For this reason we will present briefly the basic steps of the proof, explaining only the new estimates. The following proposition holds.

PROPOSITION 3.3. *Assume that* (1.9) *admits a convex entropy* $\eta(u)$ *satisfying* (2.13), *and the symmetric, positive definite matrix* A *satisfies* (2.14) *for some* $\nu > 0$ *where the constant* $\overline{\alpha}$ *depends on* α, β, *and* $\tilde{\beta}$; *see* (2.13), (2.22), *and* (3.14). *Under similar conditions on* κ, ε, h *as in Proposition* 3.1 *(with possibly different constants), and if* $h \le \gamma\,\varepsilon$ *for some* $\gamma > 0$, *the fully discrete finite element approximations satisfy*

$$\begin{aligned} &\| u_h^n + \varepsilon\,\overline{\partial}_t u_h^{n-1}\|_{L^2}^2 + \varepsilon^2 \| \overline{\partial}_t u_h^{n-1}\|_{L^2}^2 + 2\varepsilon^2 (A\,\partial_x u_h^n,\ \partial_x u_h^n) \\ &\quad + \sum_{j=1}^{n-1} \varepsilon\,\kappa \| \overline{\partial}_t u_h^j\|_{L^2}^2 + \sum_{j=1}^{n-1} \kappa^2 \| \overline{\partial}_t u_h^j\|_{L^2}^2 + \sum_{j=1}^{n-1} \varepsilon\,\kappa\,\|\partial_x u_h^j\|_{L^2}^2 \ \le\ C(u_h^0). \end{aligned} \tag{3.25}$$

Proof. The fully discrete finite element approximation u_h^n satisfies

$$(\overline{\partial}_t u_h^n,\ \phi) - (F(u_h^n),\ \partial_x \phi) + \varepsilon\big((\overline{\partial}_{tt} u_h^{n-1},\ \phi) + (A \partial_x u_h^n,\ \partial_x \phi)\big) = 0. \tag{3.26}$$

Let $\phi = P_1\,\eta'(u_h^{n+1})$ in (3.26), where $P_1 : H^1 \to S_k$ is the elliptic projection defined in (2.20). Then

$$\begin{aligned} &(\overline{\partial}_t u_h^n, \eta'(u_h^{n+1})) + \big(\partial_x F(u_h^n), \eta'(u_h^{n+1})\big) \\ &\quad + \varepsilon(\overline{\partial}_{tt} u_h^{n-1}, P_1, \eta'(u_h^{n+1})) + \varepsilon(A \partial_x\, u_h^n, \partial_x \eta'(u_h^{n+1})) \\ &\quad = (\overline{\partial}_t u_h^n, \eta'(u_h^{n+1}) - P_1\,\eta'(u_h^{n+1})) \\ &\quad + \big(\partial_x F(u_h^n), \eta'(u_h^{n+1}) - P_1\,\eta'(u_h^{n+1})\big) =: Z_1 + Z_2. \end{aligned} \tag{3.27}$$

The terms in the right-hand side will be estimated as in the semidiscrete case. We start by examining the stability that is inherited in the left-hand side. In a way similar to (3.6) one can show

$$\begin{aligned} (\overline{\partial}_{tt} Y^{n-1},\ W^{n+1}) &= \kappa(\overline{\partial}_{tt} Y^{n-1},\ \overline{\partial}_t W^n) \\ &\quad + \overline{\partial}_t(\overline{\partial}_t Y^{n-1},\ W^n) - (\overline{\partial}_t Y^n,\ \overline{\partial}_t W^n). \end{aligned} \tag{3.28}$$

Therefore,

$$\begin{aligned} &\varepsilon(\overline{\partial}_{tt} u_h^{n-1}, P_1\,\eta'(u_h^{n+1})) \\ &\quad = \varepsilon\kappa(\overline{\partial}_{tt} u_h^{n-1},\ \overline{\partial}_t P_1\,\eta'(u_h^n)) + \varepsilon\overline{\partial}_t(\overline{\partial}_t u_h^{n-1},\ P_1\,\eta'(u_h^n)) - \varepsilon(\overline{\partial}_t u_h^n,\ \overline{\partial}_t P_1\,\eta'(u_h^n)) \\ &\quad = \varepsilon\overline{\partial}_t(\overline{\partial}_t u_h^{n-1},\ \eta'(u_h^n)) + \varepsilon\kappa(\overline{\partial}_{tt} u_h^{n-1},\ \overline{\partial}_t P_1\,\eta'(u_h^n)) \\ &\qquad - \varepsilon(\overline{\partial}_t u_h^n,\ \overline{\partial}_t P_1\,\eta'(u_h^n)) + \varepsilon\overline{\partial}_t(\overline{\partial}_t u_h^{n-1},\ P_1\,\eta'(u_h^n) - \eta'(u_h^n)). \end{aligned} \tag{3.29}$$

Taylor's formula implies

$$\begin{aligned} \int_\Omega \eta(u_h^n)dx &= \int_\Omega \eta(u_h^{n+1})dx - \kappa(\eta'(u_h^{n+1}), \overline{\partial}_t u_h^n) \\ &\quad + \kappa^2 \left(\overline{\partial}_t u_h^n, \left\{ \int_0^1 \int_0^s \eta''(u_h^{n+1} - \kappa\,\tau \overline{\partial}_t u_h^n) d\tau ds \right\} \overline{\partial}_t u_h^n \right), \end{aligned} \tag{3.30}$$

i.e.,

$$
\begin{aligned}
(\overline{\partial}_t u_h^n, \eta'(u_h^{n+1})) &= \overline{\partial}_t \int_\Omega \eta(u_h^n) dx \\
&\quad + \kappa \left(\overline{\partial}_t u_h^n, \left\{ \int_0^1 \int_0^s \eta''(u_h^{n+1} - \kappa\, \tau \overline{\partial}_t u_h^n) d\tau ds \right\} \overline{\partial}_t u_h^n \right).
\end{aligned} \tag{3.31}
$$

Further, since (η, q) is an entropy pair,

$$
\begin{aligned}
(F'(u_h^n)\, \partial_x u_h^n,\ \eta'(u_h^{n+1})) &= (F'(u_h^n)\, \partial_x u_h^n,\ \eta'(u_h^n)) \\
&\quad + (F'(u_h^n)\, \partial_x u_h^n,\ \eta'(u_h^{n+1}) - \eta'(u_h^n)) \\
&= \kappa (F'(u_h^n)\, \partial_x u_h^n,\ \overline{\partial}_t \eta'(u_h^n)).
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\overline{\partial}_t \int_\Omega \eta(u_h^n) dx + \varepsilon \overline{\partial}_t (\overline{\partial}_t u_h^{n-1},\ \eta'(u_h^n)) \\
&\qquad + \varepsilon (A \partial_x\, u_h^n, \eta''(u_h) \partial_x\, u_h^n) - \varepsilon (\overline{\partial}_t\, u_{h^n}, P_1\, \overline{\partial}_t \eta'(u_h^n)) \\
&\qquad + \kappa \left(\overline{\partial}_t u_h^n, \left\{ \int_0^1 \int_0^s \eta''(u_h^{n+1} - \kappa\, \tau \overline{\partial}_t u_h^n) d\tau ds \right\} \overline{\partial}_t u_h^n \right) \\
&\qquad = Z_1 + Z_2 + Z_3,
\end{aligned} \tag{3.32}
$$

where the new term Z_3 is given by

$$
\begin{aligned}
Z_3 = &- \varepsilon \kappa (\overline{\partial}_{tt} u_h^{n-1},\ \overline{\partial}_t P_1\, \eta'(u_h^n)) \\
&- \varepsilon \overline{\partial}_t (\overline{\partial}_t u_h^{n-1},\ P_1\, \eta'(u_h^n) - \eta'(u_h^n)) - \kappa (F'(u_h^n)\, \partial_x u_h^n,\ \overline{\partial}_t \eta'(u_h^n)).
\end{aligned} \tag{3.33}
$$

Using once more Taylor's formula we obtain,

$$
\begin{aligned}
\int_\Omega \eta(u_h^n + \varepsilon \overline{\partial}_t u_h^{n-1}) dx &= \int_\Omega \eta(u_h^n) dx + \varepsilon \overline{\partial}_t (\overline{\partial}_t u_h^{n-1},\ \eta'(u_h^n)) \\
&\quad + \varepsilon^2 \left(\overline{\partial}_t u_h^{n-1}, \left\{ \int_0^1 \int_0^s \eta''(u_h^n + \varepsilon\, \tau \overline{\partial}_t u_h^{n-1}) d\tau ds \right\} \overline{\partial}_t u_h^{n-1} \right).
\end{aligned} \tag{3.34}
$$

By a slight modification of the proof of Lemma 2.3 we have

$$
\varepsilon |(\overline{\partial}_t\, u_h^n, P_1\, \overline{\partial}_t \eta'(u_h^n))| \le \beta\, \|\eta''\|_{L^\infty}\, \|\overline{\partial}_t\, u_h^n\|^2_{L^2(\Omega)}. \tag{3.35}
$$

Essentially what remains now is an estimate of $\|\overline{\partial}_t\, u_h^n\|_{L^2(\Omega)}$. As in the symmetric case we use the test function $\phi = \overline{\partial}_t\, u_h^n$ and we conclude the proof by combining arguments from the semidiscrete case (see (2.33)–(2.37)), and the fully discrete case with symmetric F' (cf. the terms with coefficient $4\varepsilon^2$), and by estimating of course the terms Z_i. It is to be noted, finally, the essential role of the estimate

$$
\begin{aligned}
&\kappa \left(\overline{\partial}_t u_h^n, \left\{ \int_0^1 \int_0^s \eta''(u_h^{n+1} - \kappa\, \tau \overline{\partial}_t u_h^n) d\tau ds \right\} \overline{\partial}_t u_h^n \right) \\
&\qquad \ge \mu \kappa \|\overline{\partial}_t u_h^n\|^2_{L^2}, \qquad \mu > 0,
\end{aligned} \tag{3.36}
$$

in the stability analysis. ∎

Remark 3.1 (mesh conditions). Proposition 3.3 holds under the assumptions for the mesh stated in Proposition 3.1, assuming in addition that $h \leq \gamma\varepsilon$. Combining these conditions we conclude that we need to have a CFL condition with small constant γ_{CFL} and in addition $h \leq \frac{\gamma}{2\gamma_{CFL}}\underline{h}$. This last relation is a quasi-uniformity condition on the mesh, the constant of which depends on how strong the CFL condition is. It seems that it is a weakness of our proof to assume $h \leq \gamma\varepsilon$ rather than $h_{\text{loc}} \leq \gamma\varepsilon$, where h_{loc} is the local mesh size close to the shock; see section 1.2. If this were the case this would not be a restriction since h_{loc} is naturally of the order of $\underline{h}$. Nevertheless, the above conditions provide enough room for computations compatible with the principle to have finer mesh in the shock areas and coarser mesh in the smooth parts of the solution. See also the related discussion in section 6.

We conclude with the following proposition.

PROPOSITION 3.4. *For entropy pairs* (η, q) *such that*

$$\|\eta\|_{L^\infty},\ \|q\|_{L^\infty},\ \|\eta'\|_{L^\infty},\ \|\eta''\|_{L^\infty}\ \leq\ C$$

and under the hypotheses of Proposition 3.3*, we have*

$$\eta(u_h)_t\ +\ q(u_h)_x\ \subset\ \textit{lies in a compact set of } H^{-1}_{loc}(\mathbb{R}\times\mathbb{R}^+),$$

where u_h *and* u_h^n *are related by* (3.16).

3.4. Estimates in the multidimensional case. Let (1.1) be endowed with a uniformly convex entropy η; the fluxes q_i are given by (2.42) [14, sec. IV.4.3]. The finite element approximations defined by (3.1) satisfy similar a priori bounds with the one-dimensional case. The matrices A_i should now satisfy the analogue of (2.44). We state the stability estimate; its proof is a modification of the proof of Proposition 3.3 and is omitted.

PROPOSITION 3.5. *Assume that* (1.1) *is equipped with a convex entropy* $\eta(u)$ *satisfying* (2.6)*. If the symmetric, positive definite matrices* A_i *satisfy* (2.44)*, then, under similar conditions on* κ, ε, h *as in Proposition* 3.1 *(with possibly different constants), and for* $h \leq \gamma\varepsilon$ *for some* $\gamma > 0$*, the fully discrete finite element approximations* (3.1) *satisfy*

$$\begin{aligned}\|\ u_h^n\ +\varepsilon\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ +\varepsilon^2\|\ \overline{\partial}_t u_h^{n-1}\|_{L^2}^2\ +2\varepsilon^2\sum_{i=1}^{d}(A_i\ \partial_{x_i}u_h^n,\ \partial_{x_i}u_h^n)\\ +\sum_{j=1}^{n-1}\varepsilon\ \kappa\|\overline{\partial}_t u_h^j\|_{L^2}^2+\sum_{j=1}^{n-1}\kappa^2\|\overline{\partial}_t u_h^j\|_{L^2}^2+\sum_{j=1}^{n-1}\varepsilon\ \kappa\ \sum_{i=1}^{d}\|\partial_{x_i}u_h^j\|_{L^2}^2\ \leq\ C(u_h^0).\end{aligned}$$

4. Convergence of finite element schemes for one-dimensional systems. The compactness of the dissipation measure (2.41) or (3.17) is central in establishing compactness of approximate solutions for systems of conservation laws via the program of compensated compactness. Such results are available (in a one-dimensional context) for the scalar conservation law, the equations of elastodynamics, the equations of isentropic gas dynamics, and the class of rich systems (see [42, 15] and the references in [14, Chap. XV]). One difficulty in applying the compensated compactness framework is that, while several of the existing compactness theorems are valid in the presence of uniform L^∞-estimates, the available estimates in applications are often just in the energy norm. In particular, this is the case for the approximations arising via semidiscrete (2.1) or fully discrete (3.2) finite element schemes. Note that

under the additional hypothesis of uniform L^∞ bounds for the approximations, one would conclude directly convergence toward a weak solution for all the aforementioned systems.

Our results can be applied to systems where the compensated compactness program has been carried out in the energy-norm framework. Such results are available for the scalar conservation law in the L^p framework (e.g., [35], [34, Thm. 2.3]) and for the equations of one-dimensional elasticity,

$$\begin{aligned} u_{1,t} - u_{2,x} &= 0, \\ u_{2,t} - \sigma(u_1)_x &= 0, \end{aligned} \tag{4.1}$$

in the energy norm [31, 38, 37]. In both cases one can deduce compactness of semidiscrete or fully discrete finite element schemes and conclude with a convergence result.

We consider here as a paradigm the system (4.1). For $\sigma'(u_1) > 0$, it is strictly hyperbolic with wave speeds $\lambda_{1,2} = \pm\sqrt{\sigma'(u_1)}$. It admits an infinite number of entropy pairs, of which the special pair

$$\eta = \frac{1}{2}u_2^2 + \int_0^{u_1} \sigma(\tau)d\tau, \quad q = -u_2\sigma(u_1) \tag{4.2}$$

is associated with the mechanical energy and the work of contact forces, and η is strictly convex. We assume that σ satisfies the subcharacteristic condition

$$0 < s \le \sigma'(u) \le S, \quad u \in \mathbb{R}, \tag{4.3}$$

with s, S positive constants. One easily checks that the matrix A can be selected so that all conditions in Propositions 2.3 and 3.3 hold.

We need a second hypothesis on σ that allows us to apply the results of [38, 37]. We assume either that (4.1) is genuinely nonlinear with

$$\sigma''(u) \neq 0 \quad \text{and} \quad \sigma'', \sigma''' \in L^2 \cap L^\infty(\mathbb{R}) \tag{4.4}$$

or that σ has precisely one inflection point at u_0 with

$$\begin{aligned} &(u - u_0)\sigma''(u) \neq 0 \quad \text{for } u \neq u_0 \\ &\quad \text{and } \ \sigma'', \sigma''' \in L^2 \cap L^\infty(\mathbb{R}). \end{aligned} \tag{4.5}$$

We then have the following theorem.

THEOREM 4.1. *Let $\sigma \in C^3$ satisfy hypotheses* (4.3), (4.4) *(or* (4.3), (4.5)*). Let $(u_1^{\varepsilon,h}, u_2^{\varepsilon,h})$ be a family of solutions of* (2.1)*, and let A be a symmetric, positive definite matrix satisfying* (2.14)*. Then, for $h \le \gamma\varepsilon$ (with γ as in Proposition* 2.3*) and along a subsequence,*

$$u_{1,h} \to u_1\,, \quad u_{2,h} \to u_2\,, \quad \text{a.e. } (x,t) \text{ and in } L^p_{loc}(\mathbb{R}\times(0,T)) \text{ for } p < 2,$$

and (u_1, u_2) is a weak solution of (4.1)*.*

Proof. The proof uses the theory of compensated compactness and proceeds by controlling the dissipation measure

$$\partial_t\eta(u_1^{\varepsilon,h}, u_2^{\varepsilon,h}) + \partial_x q(u_1^{\varepsilon,h}, u_2^{\varepsilon,h}) \quad \text{lies in a compact of } H^{-1}_{loc}, \tag{4.6}$$

for entropy pairs $(\eta(u,v)\,,\, q(u,v))$ for the equations of elasticity. In the presence of uniform L^∞-bounds, the theorem of DiPerna [15] would guarantee compactness of

approximate solutions and imply that, along a subsequence, $u_1^{\varepsilon,h} \to u_1$ and $u_2^{\varepsilon,h} \to u_2$ a.e. (x,t).

In the current case, uniform L^∞-estimates are not available and the natural stability framework is in the energy norm (see Proposition 2.3). Nevertheless, under hypothesis (4.3) and by Proposition 2.5, the dissipation measure is controlled for a class of entropy-flux pairs $(\eta(u,v)\,,\, q(u,v))$ satisfying the growth restrictions

$$\eta,\ q,\ \eta_u,\ \eta_v,\ \eta_{uu},\ \eta_{uv},\ \eta_{vv} \in L^\infty(\mathbb{R}^2). \tag{4.7}$$

This class of entropy pairs contains sufficient test pairs in order to achieve the reduction of the generalized Young measure to a point mass and to show strong convergence in L^p_{loc} for $p < 2$. The hypotheses (4.3)–(4.4) allow us to apply the result of Shearer [38], where the reduction is performed for the genuine nonlinear case, while the hypotheses (4.3)–(4.5) allow us to apply the corresponding reduction in Serre and Shearer [37] applicable to the case of elasticity with one inflection point. □

In a similar manner we can prove convergence of fully discrete finite element approximations (3.2) for the equations (4.1).

THEOREM 4.2. *Let σ be as in Theorem* 4.1 *and let A satisfy the hypotheses of Proposition* 3.3. *Let $(u_{1,h}, u_{2,h})$ be the fully discrete finite element approximations defined in* (3.16). *If the parameters κ, h, and ε are restricted by* (3.14) *and $h \le \gamma\varepsilon$ for some $\gamma > 0$, then along a subsequence*

$$u_{1,h} \to u_1, \quad u_{2,h} \to u_2, \quad \text{a.e. } (x,t) \text{ and in } L^p_{loc}(\mathbb{R}\times(0,T)) \text{ for } p<2,$$

and (u,v) is a weak solution of (4.1).

5. Error estimates for smooth solutions. In this section we consider the system of conservation laws

$$\partial_t u + \partial_x F(u) = 0 \tag{5.1}$$

and assume that (5.1) is endowed with a convex entropy $\eta(u)$. We let u be a classical solution of (5.1) defined on a maximal interval of existence and let U_ε be the smooth solution of the relaxation approximation

$$\partial_t U_\varepsilon + \partial_x F(U_\varepsilon) = \varepsilon A \partial_{xx} U_\varepsilon - \varepsilon \partial_{tt} U_\varepsilon\,. \tag{5.2}$$

We show

$$\|U_\varepsilon(t) - u(t)\|_{L^2} \le C(t,u)\,\varepsilon\,, \tag{5.3}$$

where the constant $C(t,u)$ depends on a strong norm of u and blows up at the critical time.

5.1. Motivation. It was established in Theorem 5.2.1 of [14] that the classical solution of (1.1) is unique among the class of admissible weak solutions in the case where the system admits a convex entropy. The result follows by showing a stability estimate in L^2:

$$\|u(t) - w(t)\|_{L^2} \le C(t,u)\,\|u(0) - w(0)\|_{L^2}. \tag{5.4}$$

Here u is the classical and w an admissible weak solution of (1.1). The main idea of the proof is to control the spatial integral of the quadratic in the $u - w$ function

$$H(u,w) = \eta(w) - \eta(u) - \eta'(u)(w-u). \tag{5.5}$$

This is made possible by the observation that certain quantities arising in the proof vanish when u is a classical solution and thus satisfies the entropy inequality as equality. Our idea is to use a similar approach to show the error estimate (5.3). A difficulty arises (except for handling the error terms in an appropriate way) that it is no longer possible to work with the same function H as in (5.5). On the other hand, the estimates in [43] and in section 2 suggest that when the system admits a convex entropy, we are able to control the quantity

$$\int \eta(U_\varepsilon + \varepsilon \partial_t U_\varepsilon) dx.$$

Motivated by these considerations, we introduce the functions

$$H_R(u, U_\varepsilon) = \eta\big(U_\varepsilon + \varepsilon \partial_t (U_\varepsilon - u)\big) - \eta(u) - \eta'(u)\big(U_\varepsilon - u + \varepsilon \partial_t (U_\varepsilon - u)\big), \tag{5.6}$$

$$Q\,(u, U_\varepsilon) = q(U_\varepsilon) \;-\; q(u) - \eta'(u)\;(F(U_\varepsilon) - F(u)). \tag{5.7}$$

The function H_R is the relaxational correction of (5.5) and is of quadratic order in the quantity $\big(U_\varepsilon - u + \varepsilon \partial_t (U_\varepsilon - u)\big)$. Control of $\|u(t) - U_\varepsilon(t)\|^2_{L^2}$ is achieved through the additional control of $\varepsilon^2 \|\partial_t\,(U_\varepsilon - u)\|^2_{L^2}$ that is obtained from a separate estimate natural for approximations by wave equation (5.2).

5.2. The decay functional. The first objective is to establish that H_R is a Lyapunov functional. We begin with the derivation of the main decay identity.

Let η be the convex entropy with q the corresponding flux. The classical solution u satisfies

$$\partial_t \eta(u) + \partial_x q(u) = 0.$$

The approximate solution of (5.2) will henceforth be denoted by $U \equiv U_\varepsilon$. It satisfies the identities

$$\begin{aligned}
&\partial_t (U - u) + \partial_x (F(U) - F(u)) = \varepsilon A U_{xx} - \varepsilon U_{tt},\\
&\partial_t \eta'(u)(U - u) + \partial_x \eta'(u)(F(U) - F(u))\\
&\quad = \eta''(u) u_x \cdot [F(U) - F(u) - F'(u)(U - u)] + \varepsilon \eta'(u) \cdot A U_{xx} - \varepsilon \eta'(u) \cdot U_{tt},
\end{aligned}$$

where we use (5.1) and the fact that η is an entropy if and only if $(\eta'' F')^T = \eta'' F'$; see (2.42). Combining the above, we deduce

$$\begin{aligned}
&\partial_t [\eta(U) - \eta(u) - \eta'(u)(U - u)] + \partial_x [q(U) - q(u) - \eta'(u)(F(U) - F(u))]\\
&\qquad = -\eta''(u) u_x \cdot [F(U) - F(u) - F'(u)(U - u)]\\
&\qquad\quad + \varepsilon\big(\eta'(U) - \eta'(u)\big) \cdot A U_{xx} - \varepsilon\big(\eta'(U) - \eta'(u)\big) \cdot U_{tt}.
\end{aligned} \tag{5.8}$$

We now use (5.8) in conjunction with the identities

$$\begin{aligned}
(\eta'(U) - \eta'(u)) \cdot U_{tt} &= \partial_t [(\eta'(U) - \eta'(u)) \cdot (U_t - u_t)] - \eta''(U)(U_t - u_t) \cdot (U_t - u_t)\\
&\quad - (\eta''(U) - \eta''(u)) u_t \cdot (U_t - u_t) + (\eta'(U) - \eta'(u)) \cdot u_{tt},\\
(\eta'(U) - \eta'(u)) \cdot A U_{xx} &= \partial_x [(\eta'(U) - \eta'(u)) \cdot A (U - u)_x]\\
&\quad - \eta''(U)(U - u)_x \cdot A (U - u)_x\\
&\quad - (\eta''(U) - \eta''(u)) u_x \cdot A (U - u)_x + (\eta'(U) - \eta'(u)) \cdot A u_{xx}
\end{aligned}$$

and

$$\eta\big(U+\varepsilon\partial_t(U-u)\big)=\eta(U)+\eta'(U)\varepsilon\partial_t(U-u)+\varepsilon^2\partial_t(U-u)\cdot\overline{\eta''}\partial_t(U-u)$$
$$\text{with}\quad \overline{\eta''}=\int_0^1\int_0^s \eta''(U+\varepsilon\,\tau\partial_t\,(U-u))d\tau ds$$

to conclude

(5.9)
$$\begin{aligned}&\partial_t\{\eta(U+\varepsilon\partial_t(U-u))-\eta(u)-\eta'(u)[U-u+\varepsilon\partial_t(U-u)]\\&\qquad\qquad-\varepsilon^2\partial_t(U-u)\cdot\overline{\eta''}\partial_t(U-u)\}\\&\quad+\partial_x\{q(U)-q(u)-\eta'(u)(F(U)-F(u))\}\\&\quad+\varepsilon\{\eta''(U)(U-u)_x\cdot A(U-u)_x-\eta''(U)(U-u)_t\cdot(U-u)_t\}\\&=\partial_x\{\varepsilon(\eta'(U)-\eta'(u))\cdot A(U-u)_x\}-\eta''(u)u_x\cdot[F(U)-F(u)-F'(u)(U-u)]\\&\quad+a_{1t}+a_{2t}+b_{1x}+b_{2x}.\end{aligned}$$

The error terms a_{1t}, a_{2t}, b_{1x}, and b_{2x} are defined by

$$\begin{aligned}a_{1t}&=\varepsilon\big(\eta''(U)-\eta''(u)\big)u_t\cdot(U_t-u_t),\\a_{2t}&=-\varepsilon\big(\eta'(U)-\eta'(u)\big)\cdot u_{tt},\\b_{1x}&=-\varepsilon\big(\eta''(U)-\eta''(u)\big)u_x\cdot A(U-u)_x,\\b_{2x}&=\varepsilon\big(\eta'(U)-\eta'(u)\big)\cdot Au_{xx}\end{aligned}\tag{5.10}$$

and will be estimated in what follows.

Identity (5.9) is supplemented by a correction accounting for the fact that the third term is indefinite. The correcting identity is obtained by multiplying the equation

$$(U-u)_t+F'(U)(U-u)_x=\varepsilon A(U-u)_{xx}-\varepsilon(U-u)_{tt}+\varepsilon(Au_{xx}-u_{tt})-(F'(U)-F'(u))u_x$$

by $(U-u)_t$ and integrating by parts to deduce

$$\begin{aligned}&\partial_t\Big\{\frac{1}{2}\varepsilon|U_t-u_t|^2+\frac{1}{2}\varepsilon(U-u)_x\cdot A(U-u)_x\Big\}+|(U-u)_t|^2\\&\qquad+F'(U)(U-u)_x\cdot(U-u)_t=\partial_x\Big\{\varepsilon A(U-u)_x\cdot(U-u)_t\Big\}+c_{1t}+c_{2t},\end{aligned}\tag{5.11}$$

where c_{1t}, c_{2t} are given by

$$\begin{aligned}c_{1t}&=\varepsilon(Au_{xx}-u_{tt})\cdot(U-u)_t,\\c_{2t}&=-(F'(U)-F'(u))u_x\cdot(U-u)_t.\end{aligned}\tag{5.12}$$

Next, we multiply (5.11) by $2\alpha\varepsilon$, add the resulting identity to (5.9), and use (5.6)

and (5.7) to arrive at

$$
\begin{aligned}
&\partial_t \mathcal{G}(u,U) + \partial_x Q(u,U) + \alpha\varepsilon\Big|(U-u)_t + F'(U)(U-u)_x\Big|^2 \\
&\quad + \varepsilon\Big\{\eta''(U)(U-u)_x \cdot A(U-u)_x - \alpha F'(U)(U-u)_x \cdot F'(U)(U-u)_x\Big\} \\
&\quad + \varepsilon\Big\{\big(\alpha I - \eta''(U)\big)(U-u)_t \cdot (U-u)_t\Big\} \\
&= \partial_x\Big\{\varepsilon(\eta'(U)-\eta'(u)\big)\cdot A(U-u)_x + 2\alpha\varepsilon^2 A(U-u)_x \cdot (U-u)_t\Big\} \\
&\quad - \eta''(u)u_x \cdot \Big[F(U) - F(u) - F'(u)(U-u)\Big] \\
&\quad + a_{1t} + a_{2t} + b_{1x} + b_{2x} + 2\alpha\varepsilon(c_{1t}+c_{2t}),
\end{aligned}
\tag{5.13}
$$

where

$$
\begin{aligned}
\mathcal{G}(u,U) &= H_R(u,U) \\
&\quad + \varepsilon^2\big[\alpha I - \overline{\eta''}\,\big](U-u)_t \cdot (U-u)_t + \varepsilon^2\alpha A(U-u)_x \cdot (U-u)_x.
\end{aligned}
\tag{5.14}
$$

5.3. The error estimate. Equation (5.13) is the basic decay identity. We see below that, under certain conditions on the entropy η, the quantity $\mathcal{G}(u,U)$ becomes a Lyapunov functional and leads to an error estimate.

PROPOSITION 5.1. *Assume that* (5.1) *is equipped with a strictly convex entropy* η *that satisfies, for some* $\alpha > 0$,

$$
\frac{1}{\alpha}\, I \le \eta''(u) \le \alpha\, I, \quad u \in \mathbb{R}^n,
\tag{5.15}
$$

and the positive definite, symmetric matrix A *can be selected so that for some* $\nu > 0$ *we have*

$$
\frac{1}{2}\big((\eta''(u)A)^T + \eta''(u)A\big) - \alpha F'^T(u)F'(u) \ge \nu I, \quad u \in \mathbb{R}^n.
\tag{5.16}
$$

Let u *be a smooth solution of* (5.1), *let* U_ε *be a smooth solution of* (5.2), *and suppose that both* u, U_ε *decay sufficiently fast at infinity.*

(i) *Then* $\mathcal{G}(u,U)$ *is positive definite and*

$$
\begin{aligned}
&\frac{d}{dt}\int_{\mathbb{R}} \mathcal{G}(u,U_\varepsilon)dx + \frac{1}{c}\varepsilon\int_{\mathbb{R}} |(U_\varepsilon-u)_x|^2 + |(U_\varepsilon-u)_t|^2 dx \\
&\le \int_{\mathbb{R}} \Big\{|\eta''(u)u_x\big(F(U_\varepsilon) - F(u) - F'(u)(U_\varepsilon-u)\big)| \\
&\qquad\qquad + |a_{1t} + a_{2t} + b_{1x} + b_{2x} + 2\alpha\varepsilon(c_{1t}+c_{2t})|\Big\}dx
\end{aligned}
\tag{5.17}
$$

for some constant c *independent of* ε.

(ii) *If in addition for some* $M > 0$

$$
|F''(u)| \le M, \quad |\eta'''(u)| \le M, \quad u \in \mathbb{R}^n,
\tag{5.18}
$$

then

$$
\begin{aligned}
&\|(U_\varepsilon - u)(t)\|_{L^2} + \varepsilon\|(\partial_x U_\varepsilon - \partial_x u)(t)\|_{L^2} + \varepsilon\|(\partial_t U_\varepsilon - \partial_t u)(t)\|_{L^2} \\
&\quad \le C(t,u)(\|(U_\varepsilon - u)(0)\|_{L^2} + \varepsilon\|(\partial_x U_\varepsilon - \partial_x u)(0)\|_{L^2} + \varepsilon\|(\partial_t U_\varepsilon - \partial_t u)(0)\|_{L^2} + \varepsilon),
\end{aligned}
\tag{5.19}
$$

where $C(t,u)$ is a constant depending on t and norms of the smooth solution u.

Proof. Integrating (5.13) over $\mathbb{R}$ and using the hypotheses (5.15) and (5.16), we obtain (5.17). By (5.15),

$$\alpha I - \overline{\eta''} = \alpha I - \int_0^1 \int_0^s \eta''(U + \varepsilon\,\tau\partial_t\,(U-u))d\tau ds \geq \frac{1}{2}\alpha I.$$

Moreover, the function $H_R(u,U)$ defined in (5.6) is strictly convex and thus $\mathcal{G}(u,U)$ in (5.14) is positive definite.

Under (5.15), (5.18) and for

$$\varphi(t) = \int_{\mathbb{R}} |U-u|^2 + \varepsilon^2|U_t - u_t|^2 + \varepsilon^2|U_x - u_x|^2 dx,$$

we have

$$\frac{1}{C}\varphi(t) \leq \int_{\mathbb{R}} \mathcal{G}(u,U)dx \leq C\varphi(t)\,.$$

The error terms in (5.10) are estimated by

$$\|a_{1t}\|_{L^1} \leq \varepsilon C\|u_t\|_{L^\infty}\|U-u\|_{L^2}\|U_t-u_t\|_{L^2}, \qquad \|a_{2t}\|_{L^1} \leq \varepsilon C\|u_{tt}\|_{L^2}\|U-u\|_{L^2},$$
$$\|b_{1x}\|_{L^1} \leq \varepsilon C\|u_x\|_{L^\infty}\|U-u\|_{L^2}\|U_x-u_x\|_{L^2}, \qquad \|b_{2x}\|_{L^1} \leq \varepsilon C\|u_{xx}\|_{L^2}\|U-u\|_{L^2},$$

while the ones in (5.12) are estimated by

$$\|\varepsilon c_{1t}\|_{L^1} \leq \varepsilon^2 C\big(\|u_{tt}\|_{L^2} + \|u_{xx}\|_{L^2}\big)\|U_t-u_t\|_{L^2},$$
$$\|\varepsilon c_{2t}\|_{L^1} \leq \varepsilon C\|u_x\|_{L^\infty}\|U-u\|_{L^2}\|U_t-u_t\|_{L^2},$$

where C is a generic constant depending on α, M, and norms of u.

From (5.17) we obtain

(5.20)
$$\begin{aligned}
&\frac{d}{dt}\int_{\mathbb{R}} \mathcal{G}(u,U_\varepsilon)dx + \frac{1}{C}\varepsilon\left(\|U_t-u_t\|_{L^2}^2 + \|U_x-u_x\|_{L^2}^2\right)\\
&\leq C(\|U-u\|_{L^2}^2 + \varepsilon\|U-u\|_{L^2}(1 + \|U_t-u_t\|_{L^2} + \|U_x-u_x\|_{L^2}) + \varepsilon^2\|U_t-u_t\|_{L^2})\\
&\leq C(\|U-u\|_{L^2}^2 + \varepsilon^2\|U_t-u_t\|_{L^2}^2 + \varepsilon^2\|U_x-u_x\|_{L^2}^2 + \varepsilon^2).
\end{aligned}$$

This in turn gives

$$\varphi(t) \leq \varphi(0) + \varepsilon^2 Ct + C\int_0^t \varphi(s)ds$$

and we conclude from Gronwall's inequality that

$$\varphi(t) \leq C(t,u)\big(\varphi(0) + \varepsilon^2\big). \tag{5.21}$$

Then (5.19) follows. ∎

Remark 5.1. As an example where Proposition 5.1 applies, consider the equations of elastodynamics (4.1). This system admits the entropy pair (4.2). One checks that if

$$0 < s \leq \sigma'(u) \leq S, \quad |\sigma''(u)| \leq M,$$

for some constants s, S, and $M > 0$, then (5.15), (5.16), and (5.18) are fulfilled and we obtain the relevant stability estimate.

Remark 5.2. Proposition 5.1 can be extended for multidimensional hyperbolic systems. In this case, condition (5.16) should be replaced by the analogue of (2.44).

6. Implementation issues. We include here a short discussion on the implementation of the schemes and we present indicative numerical examples that relate to our results.

Adaptivity and mesh reconstruction. The basic principles of our mesh reconstruction policy are

(a) locate the regions of space where increased accuracy is demanded, through a positive functional g;

(b) find a partition of space with predefined constant cardinality and density that follows the estimator function g; and

(c) reconstruct the solution on the finite element space which corresponds to that partition and advance to the next time step by applying the finite element scheme.

These steps are studied, introducing appropriate estimator functions for finite element methods of systems of hyperbolic conservation laws. Among others, estimator functions g are proposed which are based on a posteriori estimates or on the curvature of the approximate solution [4, 2, 3]. This approach yields a dynamic mesh construction which is combined with finite element schemes in what follows, but the mesh selection according to the basic properties of the solution is independent of the particular method used.

Mesh conditions. The mesh conditions needed in the stability analysis in section 3 are somewhat restrictive regarding the flexibility in the selection of the mesh, especially for small values of ε. The main reason is that the time step κ should be chosen very small if ε is very small. (The restrictions on the spatial mesh discussed in Remark 3.1 are not present in the numerical experiments.) In fact, the computational examples show that certain mesh conditions that relate the mesh size and ε are indeed needed and thus for fixed number of spatial mesh points and fixed κ we cannot take ε close to zero; see the following examples and [2, 3].

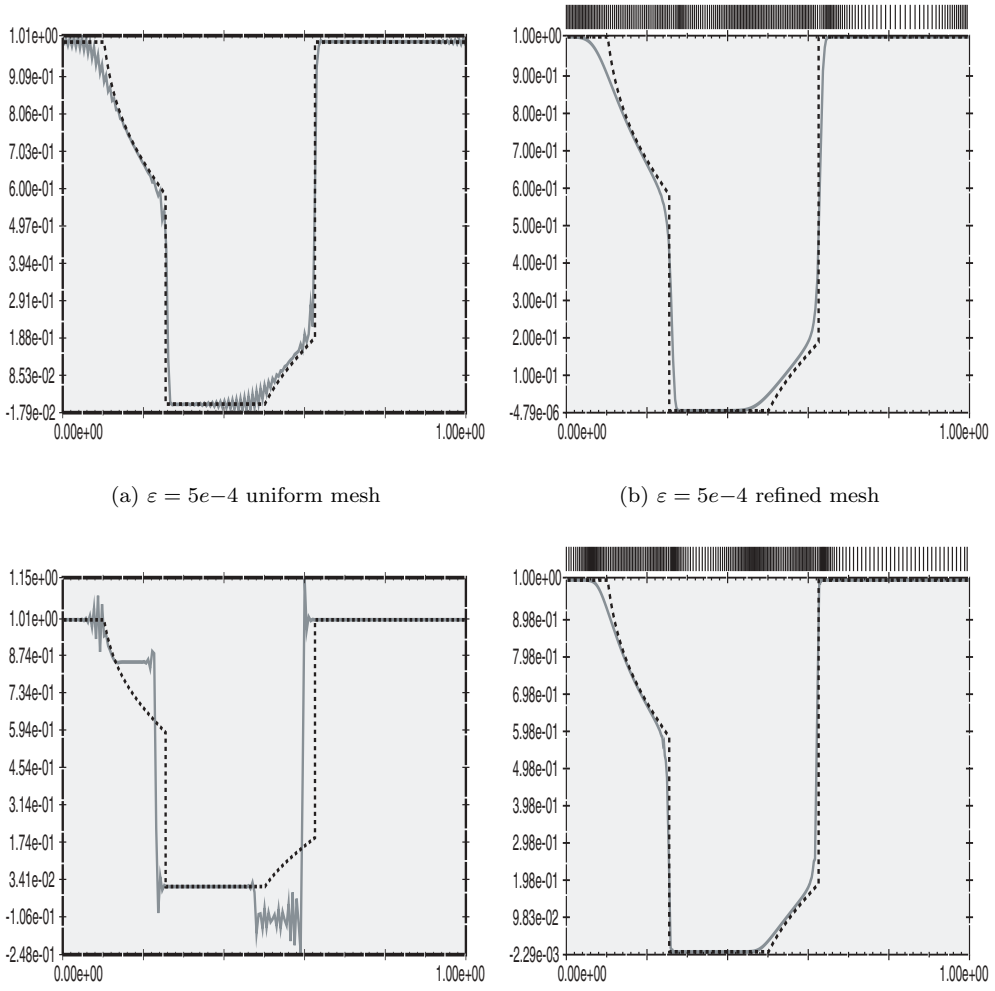
An alternative that completely bypasses this problem is provided by a modification of the finite element relaxation schemes developed in [2, 3]. The alternative is a class of finite element schemes based on the finite element discretization of a modified model with *switched relaxation.* These are schemes in which the application of a Runge–Kutta scheme uses the relaxation finite element model (1.5) for the calculation of the intermediate stages and of u_h^{n+1} and then $v_{h,i}^{n+1}$ is calculated as $v_{h,i}^{n+1} = \Pi F_i(u_h^{n+1})$. This enforces the projection to the equilibrium manifold $v = F(u)$ in each time step. The resulting schemes (switched relaxation finite element schemes) show remarkable stability even for extremely small values of ε. This is illustrated in the examples presented below.

CFL conditions. A common problem in explicit schemes with mesh refinement is to require strong CFL conditions, reflecting the relation of the time step κ to the minimum spatial mesh size $\underline{h}$. This problem appears in the computational examples of [4, 2, 3] but it is not very essential. A computationally more attractive idea would be to use time steps variable with x, or space-time elements, but this will remain for a future work.

Two-phase flow scalar problem. As a scalar example we chose the Buckley–Leverett equation [30] as a model of a two-phase flow in a porous medium. Here the flux F is not convex and is given by

$$F(u) = \frac{u^2}{u^2 + 0.5(1-u)^2}. \tag{6.1}$$

We compute the (periodic) Riemann problem in $[0,1]$ with $u_0 = 1$ on $[0, 0.1] \cup [0.5, 1]$



(a) $\varepsilon = 5e{-}4$ uniform mesh

(b) $\varepsilon = 5e{-}4$ refined mesh

(c) $\varepsilon = 5e{-}6$ uniform mesh

(d) $\varepsilon = 5e{-}6$ refined mesh

FIG. 1. *Buckley–Leverett two-phase flow problem:* 200 *nodes on* $[0,1]$. *The effect of the relationship of* h *and* ε *and of the stabilization by mesh refinement. Dotted line: exact solution; gray line: approximation. The distribution of the nodes in the refined mesh is displayed at top in* (b) *and* (d).

and $u_0 = 0$ on $(0.1, 0.5)$. In Figure 1 we display the results of application of our schemes in this problem for 200 nodes in $[0,1]$ with and without mesh refinement. For $\varepsilon = 5e-4$ the uniform mesh approximation has oscillations, while the corresponding approximation with mesh refinement provides an acceptable solution free of oscillations. Next for $\varepsilon = 5e-6$ the uniform mesh finite element solution seems to approximate a nonclassical weak solution. Thus the restrictions in our stability results on the relationship of h and ε are necessary. In this case the corresponding finite element approximation with mesh refinement not only eliminates the oscillations but resumes into the approximation of the entropy solution.

It is interesting to note that the method with uniform mesh, although oscillatory, seems to converge (weakly) as $h \to 0$. Moreover, this is also true in the example above where a nonclassical shock for (6.1) is captured. This is an indication that relaxation finite element schemes may conceivably be used to compute nonclassical shocks; compare to [29]. This interesting issue will be examined in a forthcoming work.

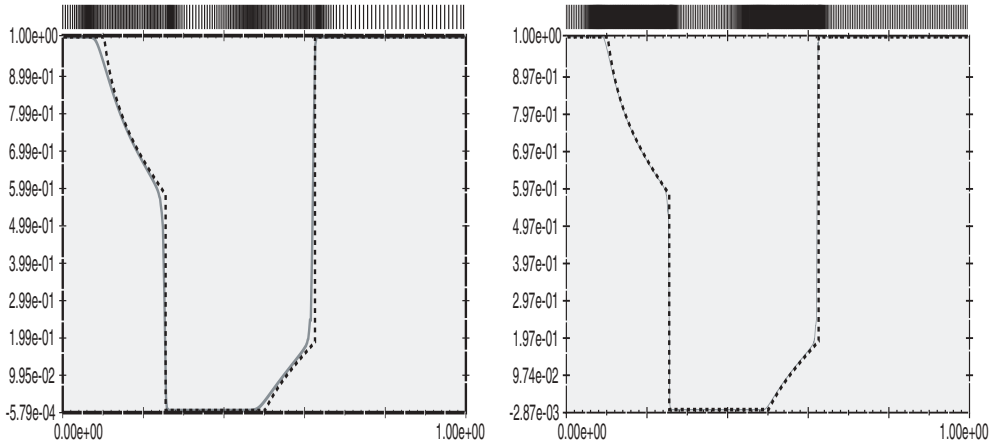
For 200 points we cannot take ε smaller unless we use the modified method based on the switched relaxation parameter. In Figure 2 we display the switched relaxation finite element schemes mentioned above. (Here the parameter $\varepsilon = \varepsilon(t)$ is a function of time that vanishes only on discrete time steps and elsewhere has a constant value ε.) Now we can have acceptable approximations for extremely small values of ε. This is a further indication of the strong regularization inherited by the adaptive mesh refinement.

System of elastodynamics. The one-dimensional system of elastodynamics is a particular case where all the results of this paper apply. We consider

$$\begin{aligned} u_{1,t} - u_{2,x} &= 0, \\ u_{2,t} - \sigma(u_1)_x &= 0 \end{aligned}$$

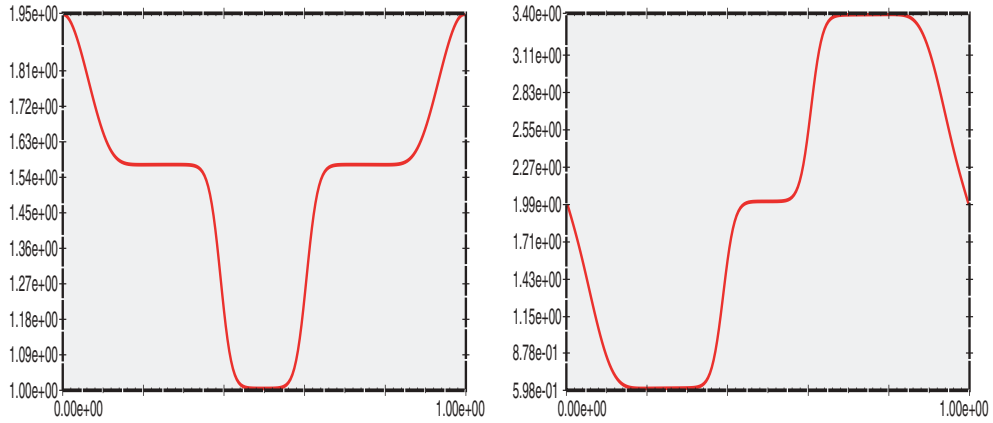
with $\sigma(v) = v + v^3$. We compute the relaxation finite element approximations with Riemann data $u_1(0) = 2$ on $[0, 1/4] \cup [3/4, 1]$ and $u_1(0) = 1$ on $[1/4, 3/4]$ and $u_2(0) = 2$ on $[0, 1]$ extended periodically. Figure 3 displays the approximations for 200 nodes in $[0, 1]$ with mesh refinement for $\varepsilon = 5e-5$. As before we use the modified method with switched relaxation parameter to compute the approximations still with 200 nodes but taking much smaller ε; Figure 4 displays the corresponding results. Figure 5 shows the improvement of the approximations if we use 400 points. In Figure 6 we see the dramatic difference of the approximations with uniform mesh and adaptive mesh refinement still with 400 nodes in $[0, 1]$. For further numerical results and detailed discussion on the adaptive mesh refinement strategies and on implementation issues for the schemes, see [4, 2, 3].

Acknowledgment. A. E. Tzavaras acknowledges the support of the National Science Foundation.



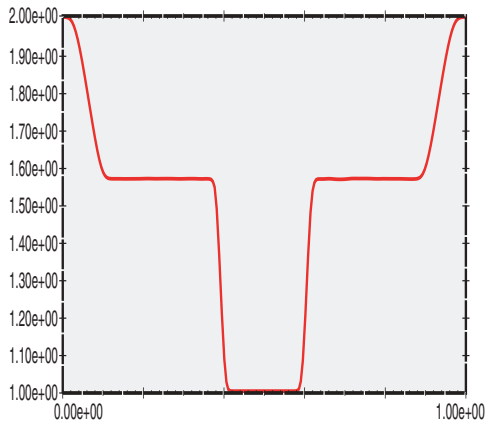
(a) $\varepsilon = 5e{-}7$, $t = 0.112$, 200 nodes (b) $\varepsilon = 5e{-}9$, $t = 0.112$, 400 nodes

FIG. 2. *Buckley–Leverett two-phase flow problem: switched relaxation finite elements with stabilization by mesh refinement. Dotted line: exact solution; gray line: approximation. The distribution of the nodes in the refined mesh is displayed at top.*

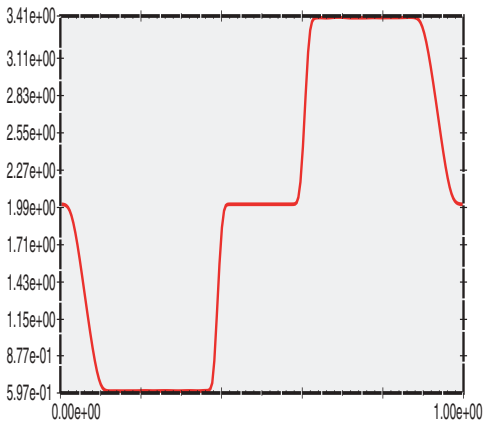


(a) u_1 : $\varepsilon = 5e{-}5$, $t = 0.058$, (b) u_2 : $\varepsilon = 5e{-}5$, $t = 0.058$

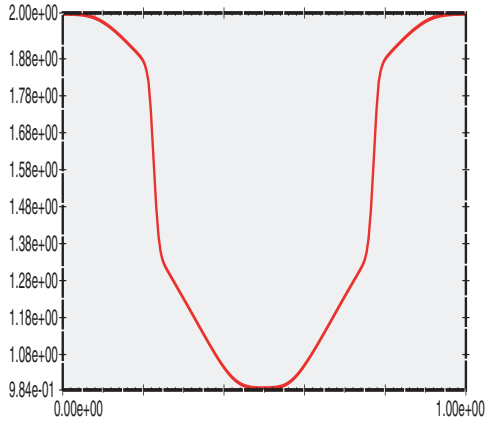
FIG. 3. *System of elastodynamics:* $q = 1$, 200 *nodes in* $[0, 1]$ *with adaptive mesh refinement.*



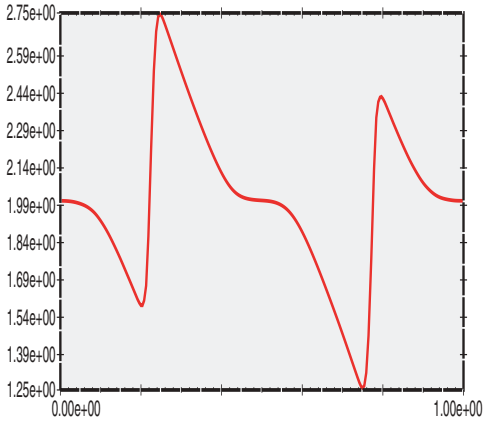
(a) u_1 : $\varepsilon = 5e{-}9$, $t = 0.058$ with refinement



(b) u_2 : $\varepsilon = 5e{-}9$, $t = 0.058$ with refinement

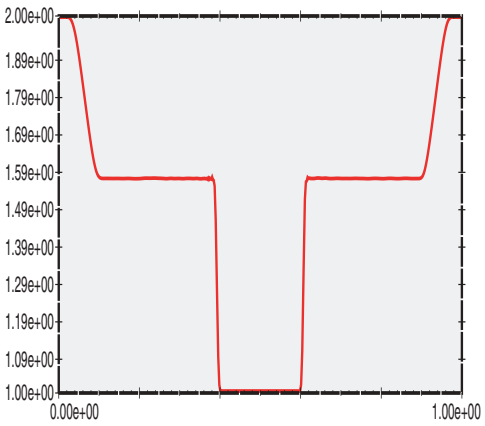


(c) u_1 : $\varepsilon = 5e{-}9$, $t = 0.35$ with refinement

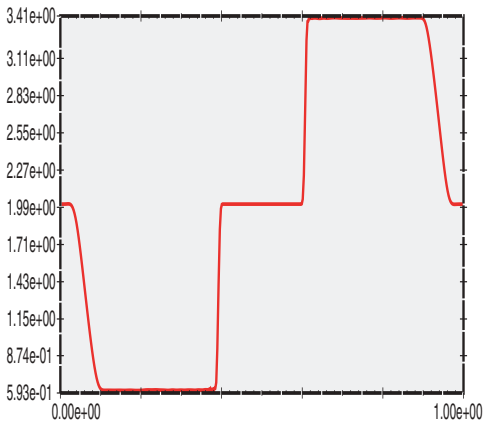


(d) u_2 : $\varepsilon = 5e{-}9$, $t = 0.35$ with refinement

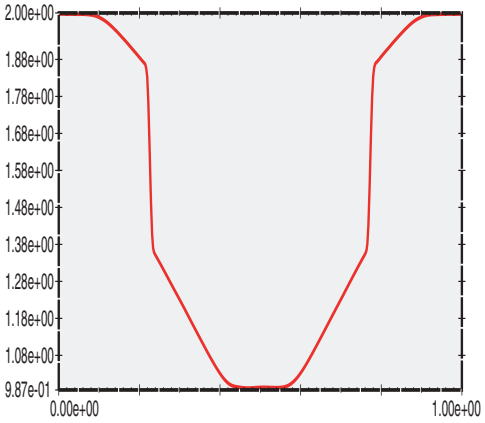
FIG. 4. *System of elastodynamics:* $q = 1$, 200 *nodes in* $[0, 1]$ *with adaptive mesh refinement.*



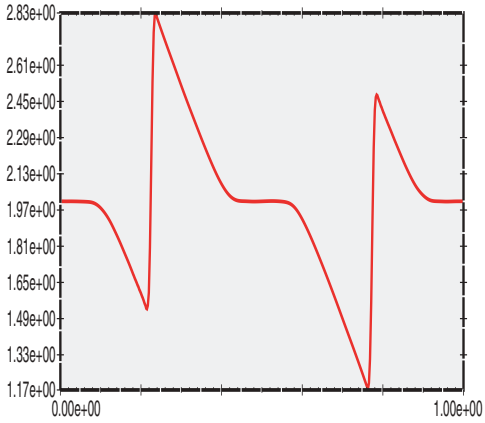
(a) u_1 : $\varepsilon = 5e{-}8$, $t = 0.058$



(b) u_2 : $\varepsilon = 5e{-}8$, $t = 0.058$

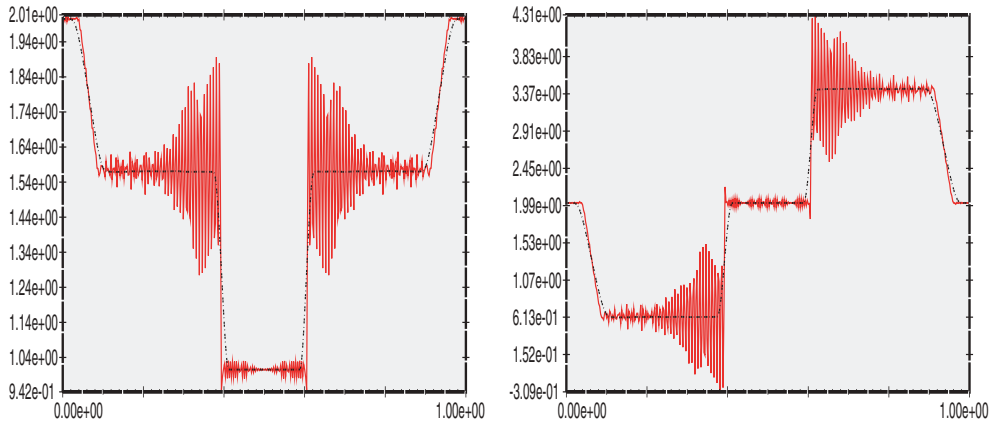


(c) u_1 : $\varepsilon = 5e{-}8$, $t = 0.35$



(d) u_2 : $\varepsilon = 5e{-}8$, $t = 0.35$

FIG. 5. *System of elastodynamics:* $q = 1$, 400 *nodes in* $[0, 1]$ *with adaptive mesh refinement.*



(a) u_1 : $\varepsilon = 5e{-}9$, $t = 0.058$ (b) u_2 : $\varepsilon = 5e{-}9$, $t = 0.058$

FIG. 6. *System of elastodynamics:* $q = 1$, 400 *nodes in* $[0, 1]$ *with uniform mesh (solid lines) and adaptive mesh refinement (dotted lines).*

REFERENCES

[1] D. AREGBA-DRIOLLET AND R. NATALINI, *Discrete kinetic schemes for multidimensional systems of conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 1973–2004.

[2] CH. ARVANITIS, *Finite Elements for Hyperbolic Systems of Conservation Laws: New Methods and Computational Techniques*, Ph.D. thesis, University of Crete, Crete, Greece, 2002.

[3] CH. ARVANITIS, *Mesh Refinement Strategies and Finite Element Schemes for Hyperbolic Systems of Conservation Laws*, preprint, University of Crete, Crete, Greece, 2004.

[4] CH. ARVANITIS, TH. KATSAOUNIS, AND CH. MAKRIDAKIS, *Adaptive finite element relaxation schemes for hyperbolic conservation laws*, Math. Model. Numer. Anal., 35 (2001), pp. 17–33.

[5] I. BABUŠKA AND J. E. OSBORN, *Analysis of finite element methods for second order boundary value problems using mesh dependent norms*, Numer. Math., 34 (1980), pp. 41–62.

[6] I. BABUŠKA, J. E. OSBORN, AND J. PITKÄRANTA, *Analysis of mixed methods using mesh dependent norms*, Math. Comp., 35 (1980), pp. 1039–1062.

[7] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[8] G.-Q. CHEN, C. D. LEVERMORE, AND T.-P. LIU, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 789–830.

[9] B. COCKBURN AND P.-A. GREMAUD, *Error estimates for finite element methods for scalar conservation laws*, SIAM J. Numer. Anal., 33 (1996), pp. 522–554.

[10] B. COCKBURN, S. HOU, AND C.-W. SHU, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws.* IV. *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[11] B. COCKBURN, C. JOHNSON, C.-W. SHU, AND E. TADMOR, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, A. Quarteroni, ed., Lecture Notes in Math. 1697, Springer-Verlag, New York, 1998.

[12] B. COCKBURN AND C.-W. SHU, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws.* II. *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[13] M. CROUZEIX AND V. THOMÉE, *On the stability in* L_p *and* W_p^1 *of the* L_2 *projection into finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.

[14] C.M. DAFERMOS, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.

[15] R. DIPERNA, *Convergence of approximate solutions to conservation laws*, Arch. Ration. Mech. Anal., 60 (1983), pp. 75–100.

[16] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* II: *Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.
[17] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Ser. 118, Springer-Verlag, Berlin, 1996.
[18] L. Gosse and A. Tzavaras, *Convergence of relaxation schemes to the equations of elastodynamics*, Math. Comp., 70 (2001), pp. 555–577.
[19] J. Jaffré, C. Johnson, and A. Szepessy, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci., 5 (1995), pp. 367–386.
[20] S. Jin and Z. Xin, *The relaxing schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–277.
[21] C. Johnson and A. Szepessy, *On the convergence of a finite element method for a nonlinear hyperbolic conservation law*, Math. Comp., 49 (1987), pp. 427–444.
[22] C. Johnson and A. Szepessy, *Adaptive finite element methods for conservation laws. Part* I: *The general approach*, Comm. Pure Appl. Math., 48 (1995), pp. 199–234.
[23] C. Johnson, A. Szepessy, and P. Hansbo, *On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws*, Math. Comp., 54 (1990), pp. 107–129.
[24] T. Katsaounis and Ch. Makridakis, *Finite volume relaxation schemes for multidimensional conservation laws*, Math. Comp., 70 (2001), pp. 533–553.
[25] M.A. Katsoulakis, G.T. Kossioris, and Ch. Makridakis, *Convergence and error estimates of relaxation schemes for multidimensional conservation laws*, Comm. Partial Differential Equations, 24 (1999), pp. 395–424.
[26] D. Kröner, *Numerical Schemes for Conservation Laws*, John Wiley, New York, 1997.
[27] A. Kurganov and E. Tadmor, *New high resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.
[28] Ph.G. LeFloch, *Hyperbolic systems of conservation laws. The theory of classical and nonclassical shock waves*, Lectures in Mathematics ETH Zurich, Birkhäuser Verlag, Basel, Switzerland, 2002.
[29] Ph.G. LeFloch and C. Rohde, *High-order schemes, entropy inequalities, and nonclassical shocks*, SIAM J. Numer. Anal., 37 (2000), pp. 2023–2060.
[30] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Lectures Math. ETH Zurich, Birkhäuser, Zurich, 1992.
[31] P. Lin, *Young measures and an application of compensated compactness to one-dimensional nonlinear elastodynamics*, Trans. Amer. Math. Soc., 329 (1992), pp. 377–413.
[32] F. Murat, *L'injection du cone positif de H^{-1} dans $W^{-1,q}$ est compacte pour tout $q < 2$*, J. Math. Pures Appl., 60 (1981), pp. 309–322.
[33] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.
[34] B. Perthame and A.E. Tzavaras, *Kinetic formulation for systems of two conservation laws and elastodynamics*, Arch. Ration. Mech. Anal., 155 (2000), pp. 1–48.
[35] M.E. Schonbek, *Convergence of solutions to nonlinear dispersive equations*, Comm. Partial Differential Equations, 7 (1982), pp. 959–1000.
[36] D. Serre, *Relaxation semi linéaire et cinétique des systèmes de lois de conservation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 169–192.
[37] D. Serre and J. Shearer, *Convergence with Physical Viscosity for Nonlinear Elasticity*, manuscript, 1993.
[38] J.W. Shearer, *Global existence and compactness in L^p for the quasi-linear wave equation*, Comm. Partial Differential Equations, 19 (1994), pp. 1829–1877.
[39] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 1073–1084.
[40] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[41] A. Szepessy, *Convergence of a shock-capturing streamline diffusion finite element method for a scalar conservation law in two space dimensions*, Math. Comp., 53 (1989), pp. 527–545.
[42] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, R. J. Knops, ed., Res. Notes in Math. 39, Pitman, Boston, 1979, pp. 136–212.
[43] A. Tzavaras, *Materials with internal variables and relaxation to conservation laws*, Arch. Ration. Mech. Anal., 146 (1999), pp. 129–155.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1394–1414

A LOCAL A POSTERIORI ERROR ESTIMATOR BASED ON EQUILIBRATED FLUXES*

R. LUCE† AND B. I. WOHLMUTH‡

Abstract. We present and analyze a new a posteriori error estimator for lowest order conforming finite elements. It is based on Raviart–Thomas finite elements and can be obtained locally by a postprocessing technique involving for each vertex a local subproblem associated with a dual mesh. Under certain regularity assumptions on the right-hand side, we obtain an error estimator where the constant in the upper bound for the true error tends to one. Replacing the conforming finite element solution by a postprocessed one, the error estimator is asymptotically exact. The local equivalence between our estimator and the standard residual-based error estimator is established. Numerical results illustrate the performance of the error estimator.

Key words. a posteriori error estimator, adaptive refinement, Raviart–Thomas finite elements

AMS subject classifications. 65N15, 65N30, 65N50

DOI. 10.1137/S0036142903433790

1. Introduction.

Many efficient numerical algorithms for the solution of partial differential equations are based on adaptive techniques. Here, we consider conforming finite elements for the numerical solution of scalar second order elliptic partial differential equations. A posteriori error estimators are very often used to control the adaptive refinement process of the triangulations. Moreover, upper bounds for the discretization error in terms of the error estimator can be obtained. The quality of the error estimator depends highly on the constants in the upper and lower bounds. Unfortunately, the reliability and efficiency constants very often can be quite large. Here, we propose a new approach such that the discretization error is bounded by our a posteriori error estimator. We refer to [3, 8, 33] for a good overview and some recent techniques and ideas.

There are many different possibilities for constructing locally defined error estimators. One possibility is to use residual-type error estimators which measure locally the jump of the discrete flux; see, e.g., [4, 5, 6, 7, 11, 31, 32]. A different approach is to solve local subproblems by using higher order finite elements, e.g., [9, 12, 18, 20, 25]. The starting point for the construction of these hierarchical error estimators is a saturation assumption. Very simple and extremely cheap a posteriori error estimators can be based on averaging techniques; see, e.g., [14, 29, 34, 35]. Error estimators for more general norms are very often based on duality techniques [10, 26]. Convergent adaptive algorithms without explicit knowledge of constants are proposed and analyzed in [19, 23].

Here, we propose a new approach which can be interpreted as a combination of equilibration and averaging techniques. The error estimator can be obtained locally and is based on a local subproblem for the flux. On each patch, we compute a

*Received by the editors August 26, 2003; accepted for publication (in revised form) February 18, 2004; published electronically December 16, 2004.

http://www.siam.org/journals/sinum/42-4/43379.html

†Laboratoire de Mathematiques Appliquées, Université de Pau et des Pays de l'Adour, France (Robert.Luce@univ-pau.fr). The work of this author was supported in part by the Groupement De Recherche MoMas (CNRS).

‡Institute of Applied Analysis and Numerical Simulations, University of Stuttgart, Germany (wohlmuth@ians.uni-stuttgart.de). The work of this author was supported in part by the Deutsche Forschungsgemeinschaft, SFB 404, C12 and the Lab. de Math. Appliquées de l'Université de Pau.

Raviart–Thomas finite element approximation. The error estimator is then defined in terms of the L^2-norm of the difference between the discrete flux approximation and the Raviart–Thomas finite element solution.

We consider the following elliptic second order boundary value problem with homogeneous Dirichlet boundary conditions on $\partial\Omega$:

$$-\text{div}\,(a\nabla p) = f \quad \text{on} \quad \Omega, \tag{1.1}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded polygonal domain. The symmetric tensor a is assumed to be uniformly positive definite and $f \in L^2(\Omega)$. Associated with (1.1) is the symmetric bilinear form $a(p,q) := \int_\Omega \nabla q \cdot a\nabla p\, dx$.

To discretize (1.1), we use conforming finite elements of lowest order on a shape regular family of simplicial triangulations $\mathcal{T}_h$. We use standard nodal hat functions ϕ_x associated with the vertices of $\mathcal{T}_h$ to define V_h. Then, the hat functions associated with the interior vertices of $\mathcal{T}_h$ form a basis of the finite element space $V_h \subset H_0^1(\Omega)$. The weak solution of (1.1) is denoted by $p \in H_0^1(\Omega)$ and the discrete weak solution by $p_h \in V_h$. We assume that the tensor a restricted to $T \in \mathcal{T}_h$ is constant, and $a_D \in \mathbb{R}$ stands for the lower bound of the eigenvalues of a restricted to the subdomain $D \subset \Omega$. The space of polynomials of order less than or equal to $k \geq 0$ on D is denoted by $P_k(D)$. In the following the generic constants $0 < c, C < \infty$ do not depend on the mesh size but only on the shape regularity of the triangulations and possibly on the variation of the tensor a restricted to suitable local patches. We use the notation $y \equiv z$ as the abbreviation for the equivalence $cy \leq z \leq Cz$.

The rest of this paper is organized as follows. In section 2, we define a Raviart–Thomas finite element by solving local systems, and we introduce our a posteriori error estimator. Global upper and local lower bounds are established for the discretization error in the energy norm in section 3, where we also show the local equivalence with the jump term of the standard residual-based error estimator. In section 4, we introduce a postprocessed flux approximation. It can be shown that our error estimator is asymptotically exact for this discrete flux. Finally, in section 5, we provide some numerical results illustrating the performance of our approach.

2. An a posteriori error estimator. In this section, we introduce our a posteriori error estimator. It is based on a postprocessing of the finite element solution p_h. We consider a dual mesh $\mathcal{K}_h$ which is defined by the centers of gravity of the triangles and the midpoints of the edges; see Figure 2.1.

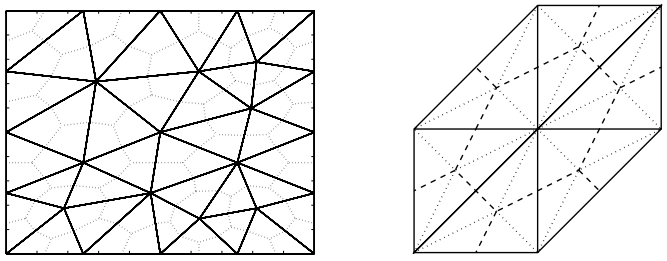


FIG. 2.1. *Primal $\mathcal{T}_h$ and dual $\mathcal{K}_h$ mesh (left) and local construction of the fine mesh $\mathcal{S}_h$ (right).*

The number of elements in the dual mesh $\mathcal{K}_h$ is equal to the number of vertices in the primal mesh $\mathcal{T}_h$. We denote the set of vertices of the primal mesh by $\mathcal{X}_h$. The left picture in Figure 2.1 shows the primal (solid lines) and the dual mesh (dashed

lines). We note that each $K \in \mathcal{K}_h$ is associated with one center vertex $x_K \in \mathcal{X}_h$. The corners of an element K of the dual mesh are the midpoints of the edges of the primal mesh which have the vertex x_K as an endpoint and the centers of gravity of all triangles $T \in \mathcal{T}_h$ with x_K as vertex. This dual mesh is well known and widely used in the context of finite volume methods, e.g., [24]. The intersection of the primal $\mathcal{T}_h$ and the dual $\mathcal{K}_h$ mesh yields a third mesh. To obtain a simplicial triangulation $\mathcal{S}_h$, we have to introduce additional edges. The elements $t \in \mathcal{S}_h$ are subtriangles of the elements $T \in \mathcal{T}_h$. Each element $T \in \mathcal{T}_h$ is decomposed into six subtriangles, as follows. The center of gravity of T is connected with the three vertices and with the three midpoints of the edges of T, yielding six subelements; see the right picture of Figure 2.1. By construction the area of each subelement t of T is equal, and thus $|t| = |T|/6$. Moreover, the number of elements of $\mathcal{S}_h$ is $6n_T$, where n_T is the number of elements in $\mathcal{T}_h$. The shape regularity of $\mathcal{S}_h$ follows from the shape regularity of $\mathcal{T}_h$. We note that the primal $\mathcal{T}_h$ and the fine $\mathcal{S}_h$ mesh are simplicial triangulations but that the dual mesh $\mathcal{K}_h$ is not. Figure 2.2 illustrates the construction of an element $K \in \mathcal{K}_h$.

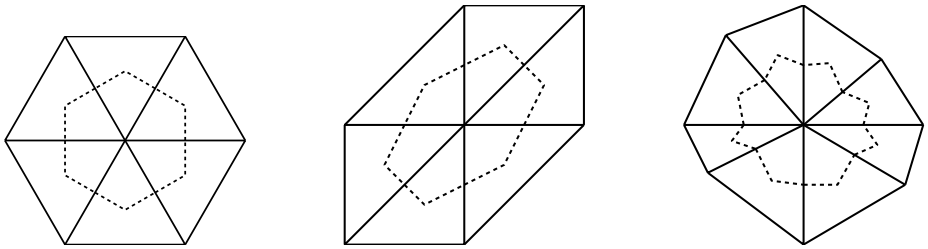


FIG. 2.2. *Different examples for an element* $K \in \mathcal{K}_h$.

We now can introduce a new finite element space U_h associated with the fine mesh $\mathcal{S}_h$. This new space is a subspace of $H(\text{div};\Omega) := \{v \in (L^2(\Omega))^2,\ \text{div}\ v \in L^2(\Omega)\}$ and is based on mixed finite elements. We denote by $RT_0(t)$ the local space of Raviart–Thomas finite elements of lowest order on t and by $RT_h \subset H(\text{div};\Omega)$ the global space of Raviart–Thomas finite elements of lowest order associated with the fine triangulation $\mathcal{S}_h$; see, e.g., [13, 27]. Locally the elements of RT_h can be written as $v_{|_t} = (a_t, b_t)^T + c_t(x, y)^T$ with some constants $a_t, b_t, c_t \in \mathbb{R}$, $t \in \mathcal{S}_h$. The finite element space U_h is defined by

$$U_h := \{v_h \in RT_h;\ \text{div}\ v_h \in Q_h\} \subset RT_h,$$

where $Q_h := \{v \in L^2(\Omega);\ v_{|_K} \in P_0(K), K \in \mathcal{K}_h\} \subset W_h := \{v \in L^2(\Omega);\ v_{|_t} \in P_0(t), t \in \mathcal{S}_h\}$. We denote by $\mathcal{E}_h$ the set of edges of the fine triangulation $\mathcal{S}_h$ which belong to the boundary ∂K of at least one element $K \in \mathcal{K}_h$, and we denote by $\mathcal{K}_h^0 \subset \mathcal{K}_h$ the subset of elements $K \in \mathcal{K}_h$ such that $\partial K \cap \partial\Omega = \emptyset$. Associated with the edge e is the unit vector n_e, which is orthogonal on e. The orientation is arbitrary but should be fixed. We define two local spaces $RT_K := \{v_h \in RT_h;\ \text{supp}\ v_h \subset \bar{K}\}$ and $W_K := \{w_h \in W_h;\ \text{supp}\ w_h \subset \bar{K}, \int_K w_h\, dx = 0\}$, $K \in \mathcal{K}_h$. Then it is well known that the differential operator $Lv := \text{div}\ v$ is a surjective mapping from RT_K onto W_K; see, e.g., [13]. Moreover, the dimension of its kernel is equal to one if $K \in \mathcal{K}_h^0$ and to zero if $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$. For each $K \in \mathcal{K}_h^0$, we set $w_K := \beta_K \text{curl}\ \psi_K$, where ψ_K is the nodal P1 conforming basis function associated with the vertex x_K and the fine triangulation $\mathcal{S}_h$, and the scaling parameter β_K is given by $\beta_K := (\text{curl}\ \psi_K, \text{curl}\psi_K)_0^{-0.5}$. It is easy

to verify that the support of w_K is K, div $w_K = 0$ and $(w_K, w_K)_0 = 1$. Given w_K, we define $w_e \in U_h$, $e \in \mathcal{E}_h$, by

$$w_e n_{\hat{e}|_{\hat{e}}} = \frac{1}{h_e}\delta_{e\hat{e}},\ \hat{e} \in \mathcal{E}_h, \quad \text{and} \quad (w_e, w_K)_0 = 0,\ K \in \mathcal{K}_h^0, \tag{2.1}$$

where h_e denotes the length of the edge e. We note that each $w_e \in U_h$ is uniquely defined by (2.1). Let $v_h^1, v_h^2 \in U_h$ satisfy (2.1); then $\Delta v_h := v_h^1 - v_h^2$ can be written as a linear combination of elements in RT_K, $K \in \mathcal{K}_h$, i.e.,

$$\Delta v_h = \sum_{K \in \mathcal{K}_h} v_K, \quad v_K \in RT_K.$$

Observing that $\mathrm{div} v_K \in P_0(K)$ and $\int_K \mathrm{div} v_K\, dx = 0$, we find that $\mathrm{div} v_K = 0$. The orthogonality on w_K yields $v_K = 0$. The existence follows now from the uniqueness and the dimension of the space. For each inner edge, we find that supp $w_e = \overline{K}_1 \cup \overline{K}_2$, $K_1, K_2 \in \mathcal{K}_h$ such that $e = \partial K_1 \cap \partial K_2$.

It is easy to see that w_e, $e \in \mathcal{E}_h$ and w_K, $K \in \mathcal{K}_h^0$ form a set of linear independent functions. Let w_h be an element of U_h; then $v_h := w_h - \sum_{e \in \mathcal{E}_h} \beta_e w_e \in U_h$ with $\beta_e := h_e w_h n_{e|_e}$. Moreover, by definition $v_h n_{e|_e} = 0$, $e \in \mathcal{E}_h$ and thus div $v_h = 0$. Observing that $v_{h|_K} \in U_h$, we find that $v_h \in$ span $\{w_K, K \in \mathcal{K}_h^0\}$. Thus w_e, $e \in \mathcal{E}_h$ and w_K, $K \in \mathcal{K}_h^0$ form a basis of U_h and

$$U_h = \sum_{e \in \mathcal{E}_h} \text{span } \{w_e\} \oplus \sum_{K \in \mathcal{K}_h^0} \text{span } \{w_K\}.$$

To introduce our error estimator, we define $u_h \in U_h$ in terms of p_h and f. Each element $u_h \in U_h$ can be written as a linear combination of the given basis functions, i.e., $u_h = \sum_{e \in \mathcal{E}_h} \alpha_e w_e + \sum_{K \in \mathcal{K}_h^0} \alpha_K w_K$. We set α_e and α_K,

$$\alpha_e := \int_e a \nabla p_h n_e\, d\sigma + \hat{\alpha}_{K_e}, \quad \alpha_K := \int_K a \nabla p_h\ w_K\ dx, \tag{2.2}$$

where $\hat{\alpha}_{K_e} := 0$ if $e \not\subset \partial\Omega$. Otherwise, we set

$$\hat{\alpha}_{K_e} := \frac{1}{2}\left(\int_\Omega -f \phi_{x_{K_e}}\ dx - \int_{\partial K_e} a \nabla p_h n_{K_e}\ d\sigma \right), \tag{2.3}$$

where K_e is the unique element of the dual mesh such that $e \subset \partial K_e \cap \partial\Omega$. By using the orthogonality between w_e and w_K and definition (2.2), we find

$$\int_K u_h w_K\ dx = \int_K a \nabla p_h w_K\ dx, \quad K \in \mathcal{K}_h^0. \tag{2.4}$$

The analysis of the properties of u_h is based on two operators P_Q and I_Q. For each $K \in \mathcal{K}_h$, we introduce two sets of edges $\mathcal{E}_K$, $\mathcal{E}_K^x$ and a set of triangles $\mathcal{S}_K := \{t \in \mathcal{S}_h;\ t \subset K\}$. The set of edges $\mathcal{E}_K$ contains all edges of the triangulation $\mathcal{S}_h$ which are in the interior of K, and the set of $\mathcal{E}_K^x$ is the union of all edges of the primal triangulation $\mathcal{T}_h$ having x_K as a vertex; see Figure 2.3. The edges in $\mathcal{E}_K^x$ are marked by dashed lines and the elements in $\mathcal{E}_K$ by solid lines.

We remark that the number of elements in $\mathcal{S}_K$ and $\mathcal{E}_K$ is the same for $K \in \mathcal{K}_h^0$. By definition, the divergence of u_h is constant on each element $K \in \mathcal{K}_h$. Moreover, it

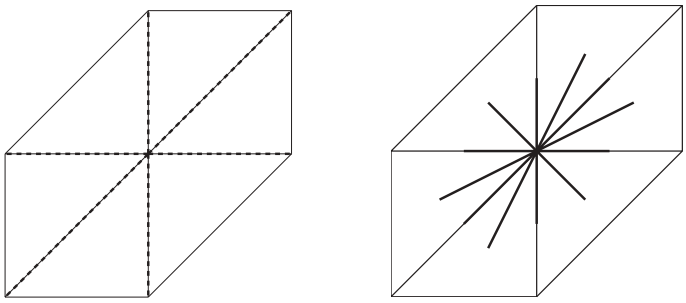


FIG. 2.3. *The two sets of edges $\mathcal{E}_K^x$ (left) and $\mathcal{E}_K$ (right).*

depends only on the right-hand side and can be obtained as a weighted mean value of f. To see this, we introduce two quasi-interpolants. The first, $P_Q : L^2(\Omega) \longrightarrow Q_h$, is defined in terms of a weighted mean value of v on K,

$$P_Q v_{|_K} := \frac{1}{|K|} \int_\Omega v \phi_{x_K} \, dx, \quad v \in L^2(\Omega),$$

where ϕ_{x_K} is the nodal P1 conforming basis function associated with the vertex x_K and the primal triangulation $\mathcal{T}_h$. If v is constant on each element $T \in \mathcal{T}_h$, then $P_Q v = \Pi_Q v$, where Π_Q denotes the L^2-projection on Q_h. The second, $I_Q : C(\Omega) \longrightarrow Q_h$, is given in terms of a nodal value,

$$I_Q v_{|_K} := v(x_K), \quad v \in C(\Omega).$$

It is easy to verify that

$$\int_\Omega w v_h \, dx = \int_\Omega P_Q w \; I_Q v_h \, dx, \quad v_h \in V_h, w \in L^2(\Omega). \tag{2.5}$$

Furthermore, we have the orthogonality relation

$$\int_\Omega q_h (I_Q - \mathrm{Id}) v_h \, dx = 0, \quad q_h \in W_h, v_h \in V_h, \tag{2.6}$$

where $W_h := \{v \in L^2(\Omega);\ v_{|_T} \in P_0(T), T \in \mathcal{T}_h\}$. This results from the fact that $|T \cap K| = 1/3|T|$ for each T such that $T \cap K$ has a nonzero measure. Additionally, the operators P_Q, I_Q satisfy a local approximation property in the L^2-norm,

$$\begin{aligned} \|v - I_Q v\|_{0;K} &\le C h_K |v|_{1;K}, \quad v \in V_h, \\ \|v - P_Q v\|_{0;K} &\le C h_K |v|_{1;\omega_K}, \quad v \in H^1(\Omega), \end{aligned}$$

where ω_K is the union of all elements $T \in \mathcal{T}_h$ such that K and T have a nonempty intersection. The first inequality results from a discrete norm equivalence. We next observe that $P_Q v = v$ on K if v is constant on ω_K. The second inequality is then a consequence of the L^2-stability of P_Q. In terms of P_Q, we find a relation between the divergence of u_h and f.

LEMMA 2.1. $\operatorname{div} u_h = -P_Q f$.

Proof. In a first step, we consider a $K \in \mathcal{K}_h^0$. The definition of the basis functions of U_h and (2.2) yield

$$\begin{aligned}\int_K \operatorname{div} u_h \, dx &= \int_{\partial K} u_h n_K \, d\sigma = \int_{\partial K} a\nabla p_h n_K \, d\sigma \\ &= \sum_{t\in\mathcal{S}_K} \int_{\partial t} a\nabla p_h n_t \, d\sigma - \sum_{e\in\mathcal{E}_K} \int_e [a\nabla p_h n_e] \, d\sigma,\end{aligned}$$

where the jump on e is defined by $[a\nabla p_h n_e] := (a\nabla p_{h|_{t_1}} - a\nabla p_{h|t_2})n_e$ and $\partial t_1 \cap \partial t_2 = e$ such that n_e is the outer unit normal on ∂t_1 and n_t is the outer unit normal on ∂t and n_K on ∂K. The first term on the right side is zero. Using the observation that $[a\nabla p_h n_e] = 0$ for each e which is an interior edge of a primal element $T \in \mathcal{T}_h$ and that the jump of the discrete flux is constant on each edge, we get

$$\begin{aligned}\int_K \operatorname{div} u_h \, dx &= -\frac{1}{2} \sum_{E\in\mathcal{E}_K^x} \int_E [a\nabla p_h n_E] \, d\sigma = - \sum_{E\in\mathcal{E}_K^x} \int_E [a\nabla p_h n_E]\phi_{x_K} \, d\sigma \\ &= - \sum_{T\in\mathcal{T}_K} \int_{\partial T} a\nabla p_h n_T \phi_{x_K} \, d\sigma = - \sum_{T\in\mathcal{T}_K} \int_T a\nabla p_h \nabla \phi_{x_K} \, dx \\ &= - \sum_{T\in\mathcal{T}_K} \int_T f\phi_{x_K} \, dx = - \int_\Omega f\phi_{x_K} \, dx = - \int_K P_Q f \, dx,\end{aligned}$$

where $\mathcal{T}_K$ contains all elements of $\mathcal{T}_h$ such that $K \cap T \neq \emptyset$. Using the definition (2.3), the assertion follows for $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$ from

$$\int_K \operatorname{div} u_h \, dx = \int_{\partial K} a\nabla p_h n_K \, d\sigma + 2\hat{\alpha}_K. \qquad \square$$

We define our a posteriori error estimator in terms of u_h. For each element K in $\mathcal{K}_h$, the local contribution is given by

$$\eta_K^2 := \|a^{-\frac{1}{2}}(u_h - a\nabla p_h)\|_{0;K}^2, \tag{2.7}$$

and the global error estimator is defined as $\eta^2 := \sum_{K\in\mathcal{K}_h} \eta_K^2$. This error estimator is related to averaging techniques, to the equilibrated residual method, and to a better flux approximation; see, e.g., [2, 3, 22, 28, 30]. It can be evaluated easily and is cheap. The coefficients of u_h with respect to the basis functions w_e and w_K are defined by scalar equations. To obtain the basis functions w_e, we have to solve low dimensional local systems on K.

However, it is more efficient to compute u_h directly as an element in RT_h. Working with a scaled nodal basis for Raviart–Thomas finite elements of lowest order, the coefficients of u_h can be obtained locally by solving for each K a low dimensional system. We denote by q_e the scaled nodal basis function of RT_h associated with the edge e. The scaling is done such that q_e satisfies, for all edges $\hat{e}$ of the fine triangulation $\mathcal{S}_h$,

$$(q_e n_{\hat{e}})_{|\hat{e}} = \frac{\delta_{e\hat{e}}}{h_e}.$$

The values of $u_h n_K$ on ∂K in combination with $\operatorname{div} u_h \in P_0(K)$ define u_h on $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$ uniquely. Moreover, for $K \in \mathcal{K}_h^0$, these properties define u_h up to a term

of the form $c_K \mathrm{curl}\, \psi_{x_K}$. Using the definition of the scaled nodal basis functions q_e and of w_e and $U_h \subset RT_h$, we find that u_h restricted to K can be written as $u_h = \sum_{e\subset\partial K} \alpha_e q_e + \sum_{e\in\mathcal{E}_K} \beta_e q_e$, where α_e is defined by (2.2). The coefficients β_e can now be obtained from the fact that div u_h is constant on K and from (2.4). We enumerate the small triangles and the edges in K clockwise. Let m_K be the number of triangles in $\mathcal{S}_K$. The edges on $\partial K \setminus \partial\Omega$ are denoted by $\hat{e}_i$ and the edges in the interior of K or on $\partial K \cap \partial\Omega$ by e_i. For $K \in \mathcal{K}_h$, t_i has the edges $\hat{e}_i$, e_i, and e_{i+1}, $1 \le i \le m_K$, where $e_{m_K+1} := e_1$ for $K \in \mathcal{K}_h^0$; see Figure 2.4. The orientation of the normals is chosen such that $-n_{e_i}$ is the outer normal on t_i and $n_{\hat{e}_i}$ is the outer normal on K.

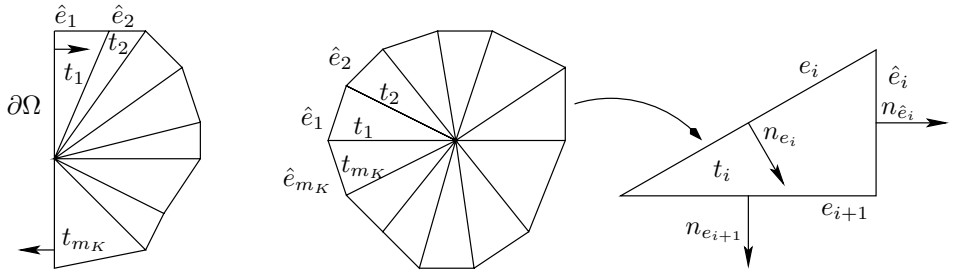


FIG. 2.4. *An element $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$ (left) and an element $K \in \mathcal{K}_h^0$ (right).*

In the case of $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$, we have to find β_{e_i}, $2 \le i \le m_K$, and for $K \in \mathcal{K}_h^0$, we have to find β_{e_i}, $1 \le i \le m_K$. In a first step, we define coefficients γ_i, $2 \le i \le m_K$, such that

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \\ 0 & \ddots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \gamma_2 \\ \gamma_3 \\ \vdots \\ \vdots \\ \vdots \\ \gamma_{m_K} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ \vdots \\ b_{m_K-1} \end{pmatrix},$$

where the right side b depends only on p_h and f. We set

$$b_i := -P_Q f(x_K)|t_i| - \alpha_{\hat{e}_i} + d_K \delta_{i1} \alpha_{e_1},$$

where $d_K = 0$ if $K \in \mathcal{K}_h^0$ and $d_K = 1$ if $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$. For this choice, we find div u_h = div $\tilde{u}_h$, where $\tilde{u}_h := \sum_{i=1}^{m_K} \alpha_{\hat{e}_i} q_{\hat{e}_i} + \sum_{i=2}^{m_K} \gamma_i q_{e_i} + d_K(\alpha_{e_1} q_{e_1} + \alpha_{e_{m_K+1}} q_{e_{m_K+1}})$, and $\tilde{u}_h n_K = u_h n_K$ restricted on ∂K. We recall that Lemma 2.1 yields $\sum_{i=1}^{m_K} \alpha_{\hat{e}_i} = -P_Q f(x_K)|K|$, $K \in \mathcal{K}_h^0$, and $\sum_{i=1}^{m_K} \alpha_{\hat{e}_i} - \alpha_{e_1} + \alpha_{e_{m_K+1}} = -P_Q f(x_K)|K|$, $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$. Due to the uniqueness of u_h, we get that $u_h = \tilde{u}_h$ on $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$ and $u_h = \tilde{u}_h + \gamma_K w_K$ on $K \in \mathcal{K}_h^0$, where γ_K is given by

$$\gamma_K := \int_K (a\nabla p_h - \tilde{u}_h) w_K \, dx.$$

Remark 2.2. We note that in the more general case of Neumann boundary conditions with data g_N, we have to add the term $h_e/a_{t_e} \|a\nabla p_h - g_N\|_{0;e}^2$, where $e = t_e \cap \partial\Omega$ to our estimator. This term is well known for the standard residual-based error estimator.

3. Upper and lower bounds for the discretization error. In this section, we establish upper and lower bounds for the discretization error in terms of the error estimator. The following theorem provides an upper bound for the error in the energy norm in terms of the error estimator. To start, we introduce a higher order term

$$\xi^2 := \sum_{K\in\mathcal{K}_h} \frac{h_K^2}{a_K}\|f - P_Q f\|_{0;K}^2 + \sum_{T\in\mathcal{T}_h} \frac{h_T^2}{a_T}\|f - \Pi_W f\|_{0;T}^2,$$

where Π_W denotes the L^2-projection on W_h.

THEOREM 3.1. *The error in the energy norm is bounded by*

$$|\!|\!| e_h |\!|\!| := a(p - p_h, p - p_h)^{\frac{1}{2}} \le \eta + C\xi.$$

Proof. Integration by parts and Lemma 2.1 yield for the error in the energy norm

$$\begin{aligned}
|\!|\!| e_h |\!|\!|^2 &= \int_\Omega (a\nabla p - u_h + u_h - a\nabla p_h)(\nabla p - \nabla p_h)\ dx \\
&= \int_\Omega \operatorname{div}\ (a\nabla p - u_h)(p_h - p)\ dx + \int_\Omega (u_h - a\nabla p_h)(\nabla p - \nabla p_h)\ dx \\
&\le \|a^{-\frac{1}{2}}(u_h - a\nabla p_h)\|_0 |\!|\!| e_h |\!|\!| + \int_\Omega (f - P_Q f)(p - p_h)\ dx.
\end{aligned}$$

We now have to consider the second term on the right side in more detail. Using the projection operator P_h of Scott and Zhang [21] and the identity (2.5), we find

$$\begin{aligned}
\int_\Omega (f - P_Q f)(p - p_h)\ dx &= \int_\Omega (f - P_Q f)(p - p_h - P_h(p - p_h))\ dx \\
&\quad + \int_\Omega (f - P_Q f) P_h(p - p_h)\ dx \\
&\le C \sum_{K\in\mathcal{K}_h} h_K \|f - P_Q f\|_{0;K} \|\nabla p - \nabla p_h\|_{0;\tilde{\omega}_K} \\
&\quad + \int_\Omega P_Q f (I_Q - \mathrm{Id}) P_h(p - p_h)\ dx,
\end{aligned}$$

where $\tilde{\omega}_K$ is a suitable local neighborhood of K. The number of elements $T \subset \tilde{\omega}_K$ is bounded independently of the mesh size. In terms of the orthogonality (2.6), the second term can be bounded by

$$\begin{aligned}
\int_\Omega P_Q f (I_Q - \mathrm{Id}) P_h(p - p_h)\ dx &= \int_\Omega (P_Q f - \Pi_W f)(I_Q - \mathrm{Id}) P_h(p - p_h)\ dx \\
&\le \sum_{K\in\mathcal{K}_h} h_K \|\Pi_W f - P_Q f\|_{0;K} |P_h(p - p_h)|_{1;K}.
\end{aligned}$$

The definition of the energy norm and the H^1-stability of P_h yield

$$|\!|\!| e_h |\!|\!| \le \eta + C\left(\sum_{K\in\mathcal{K}_h} \frac{h_K^2}{a_K}\|f - P_Q f\|_{0;K}^2 + \sum_{T\in\mathcal{T}_h} \frac{h_T^2}{a_T}\|f - \Pi_W f\|_{0;T}^2\right)^{\frac{1}{2}}.$$

We remark that ξ is a higher order term and can be neglected asymptotically. ∎

For a reliable and efficient error estimator, it is not sufficient to guarantee a global upper bound for the discretization error. Additionally, we have to consider the local contributions of the error estimator. We define for each edge $e \in \mathcal{E}_K$ the coefficient a_e by $a_e := 0.5(a_{t_1} + a_{t_2})$ if $e = \partial t_1 \cap \partial t_2$, $t_1, t_2 \in \mathcal{S}_h$ and $a_e := a_t$ if $e = \partial t \cap \partial\Omega$, $t \in \mathcal{S}_h$.

LEMMA 3.2. *The local contribution η_K, $K \in \mathcal{K}_h^0$, of the error estimator is equivalent to the jump of the discrete flux $a\nabla p_h$, i.e.,*

$$c \sum_{e \in \mathcal{E}_K} \frac{h_e}{a_e} \|[a\nabla p_h]\|_{0;e}^2 \le \eta_K^2 \le C \sum_{e \in \mathcal{E}_K} \frac{h_e}{a_e} \|[a\nabla p_h]\|_{0;e}^2, \quad K \in \mathcal{K}_h^0.$$

Proof. To obtain the lower bound, we use a discrete norm equivalence for lowest order Raviart–Thomas finite elements. The L^2-norm of $v \in RT_0(t)$ on $t \in \mathcal{S}_h$ is equivalent to a weighted L^2-norm on ∂t of the normal component

$$c\|v\|_{0,t}^2 \le \sum_{e \subset \partial t} h_e \|v n_e\|_{0;e}^2 \le C\|v\|_{0,t}^2, \quad v \in RT_0(t). \tag{3.1}$$

Using this norm equivalence and observing that locally $a\nabla p_{h|_t}$ is in $RT_0(t)$, we find for $K \in \mathcal{K}_h$

$$\begin{aligned}
\eta_K^2 &= \sum_{t \in \mathcal{S}_K} \|a^{-\frac{1}{2}}(u_h - a\nabla p_h)\|_{0;t}^2 \equiv \sum_{t \in \mathcal{S}_K} \frac{h_t}{a_t} \|(u_h - a\nabla p_h) n_t\|_{0;\partial t}^2 \\
&\equiv \sum_{e \in \mathcal{E}_K} \frac{h_e}{a_e} \left(\|[(u_h - a\nabla p_h) n_e]\|_{0;e}^2 + \|\{u_h - a\nabla p_h\} n_e\|_{0;e}^2 \right) \\
&\quad + \sum_{e \subset \partial K} \frac{h_e}{a_e} \|(u_h - a\nabla p_h) n_K\|_{0;e}^2,
\end{aligned}$$

where $[\cdot]$ denotes the jump and $\{\cdot\}$ denotes the average over the edge e. The orientation is arbitrary but should be fixed. By definition, $u_h n_K = a\nabla p_h n_K$ on $\partial K \setminus \partial\Omega$. Thus, in the second term on the right-hand side, we can replace the sum by $\sum_{e \subset \partial K \cap \partial\Omega}$. Observing that $u_h \in H(\mathrm{div}\,;\Omega)$ results in $[u_h n_e] = 0$, we find

$$\eta_K^2 \equiv \sum_{e \in \mathcal{E}_K} \frac{h_e}{a_e} \left(\|[a\nabla p_h n_e]\|_{0;e}^2 + \|\{u_h - a\nabla p_h\} n_e\|_{0;e}^2 \right) + \frac{\hat{\alpha}_K^2}{a_K}. \tag{3.2}$$

This norm equivalence guarantees the lower bound for η_K^2. To establish the upper bound, it is sufficient to bound $\hat{\alpha}_K$ and $\|\{u_h - a\nabla p_h\} n_e\|_{0;e}$. Let K be in $\mathcal{K}_h^0$; then $\partial K \cap \partial\Omega = \emptyset$ and $\hat{\alpha}_K = 0$. In a first step, we introduce $w_h \in RT_h$ locally on K by

$$w_h n_K = a\nabla p_h n_K \quad \text{on} \quad \partial K, \quad w_h n_e = \{a\nabla p_h n_e\} \quad \text{on} \quad e \in \mathcal{E}_K.$$

Then, in general, div w_h is not constant on K and thus $w_h \notin U_h$. We can now decompose u_h restricted to K in terms of w_h by $u_h = w_h + v_h$ and find

$$\sigma_{\text{average}}^2 := \sum_{e \in \mathcal{E}_K} \frac{h_e}{a_e} \|\{u_h - a\nabla p_h\} n_e\|_{0;e}^2 = \sum_{e \in \mathcal{E}_K} \frac{h_e}{a_e} \|v_h n_e\|_{0;e}^2 \le C\|a^{-\frac{1}{2}} v_h\|_{0;K}^2.$$

The number of elements t in K is bounded independently of the mesh size. The bound depends only on the shape regularity of the triangulation. A generalized Poincaré–

Friedrichs-type inequality for Raviart–Thomas elements yields

$$\|a^{-\frac{1}{2}} v_h\|_{0;K}^2 \le C \sum_{t\in\mathcal{S}_K} \frac{|t|}{a_t} \|\text{div } v_h\|_{0;t}^2 + \frac{C}{a_K} \left(\int_K v_h w_K \, dx \right)^2$$
$$= C \sum_{t\in\mathcal{S}_K} \frac{|t|}{a_t} \|\text{div } (u_h - w_h)\|_{0;t}^2 + \frac{C}{a_K} \left(\int_K v_h w_K \, dx \right)^2.$$

To obtain the first inequality, we used $v_h n_K = 0$ on ∂K. In a next step, we consider the two terms on the right-hand side separately. We start with the second term and find

$$\left(\int_K v_h w_K \, dx \right)^2 = \left(\int_K (a\nabla p_h - w_h) w_K \, dx \right)^2 \le \|a\nabla p_h - w_h\|_{0;K}^2.$$

Using the definition of w_h, we obtain that $(a\nabla p_h - w_h)_{|t} \in RT_0(t)$ and that the normal component depends only on the jump of $a\nabla p_h n$. Thus, the L^2-norm of $a\nabla p_h - w_h$ can be bounded by the jump, and we get

$$\frac{1}{a_K} \left(\int_K v_h w_K \, dx \right)^2 \le C \sum_{e\in\mathcal{E}_K} \frac{h_e}{a_e} \|[a\nabla p_h n_e]\|_{0;e}^2.$$

We note that each element $t \in \mathcal{S}_K$ has at least one edge such that $[a\nabla p_h n_e] = 0$. This results from the fact that p_h is linear on the triangles of the primal mesh. Let $e_1 := \partial t \cap \partial K$, e_3 stands for the edge of t which is a subset of an edge in $\mathcal{E}_K^p$, and e_2 is the edge of ∂t in the interior of a $T \in \mathcal{T}_h$; see Figure 3.1.

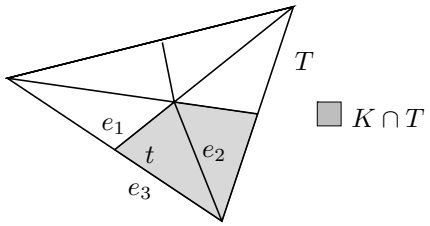


FIG. 3.1. *Numeration of the three edges of ∂t.*

We then get for the divergence of w_h on t,

$$|t| \, \|\text{div } w_h\|_{0;t}^2 = \left(\int_t \text{div } w_h \, dx \right)^2 = \left(\sum_{i=1}^3 \int_{e_i} w_h n_t \, d\sigma \right)^2$$
$$= \left(\int_{e_1} a\nabla p_h n_t \, d\sigma + \int_{e_2} a\nabla p_h n_t \, d\sigma + \int_{e_3} \{a\nabla p_h n_t\} \, d\sigma \right)^2$$
$$= \left(\int_{\partial t} a\nabla p_h n_t \, d\sigma - \frac{1}{2} \int_{e_3} [a\nabla p_h n_t] \, d\sigma \right)^2 = \frac{h_{e_3}}{4} \|[a\nabla p_h n_{e_3}]\|_{0;e_3}^2.$$

In terms of the jump and the divergence of u_h, we can now bound $\sigma^2_{\text{average}}$ by

$$\sigma^2_{\text{average}} \le C \left(\sum_{t\in\mathcal{S}_K} \frac{|t|}{a_t} \|\text{div } u_h\|_{0;t}^2 + \sum_{e\in\mathcal{E}_K} \frac{h_e}{a_e} \|[a\nabla p_h n_e]\|_{0;e}^2 \right).$$

In a last step, we have to consider the divergence of u_h in more detail. To bound div u_h, we use the proof of Lemma 2.1 and the fact that it is constant on K. We find that

$$\sum_{t\in\mathcal{S}_K}\frac{|t|}{a_t}\|\text{div } u_h\|^2_{0;t} \le \sum_{t\in\mathcal{S}_K}\frac{1}{a_t}\left(\int_K \text{div } u_h\, dx\right)^2$$
$$\le C\sum_{t\in\mathcal{S}_K}\frac{1}{a_t}\left(\sum_{e\in\mathcal{E}_K}\int_e [a\nabla p_h n_e]\, d\sigma\right)^2 \le C\sum_{e\in\mathcal{E}_K}\frac{h_e}{a_e}\|[a\nabla p_h n_e]\|^2_{0;e}. \square$$

We note that the jump of the discrete flux $a\nabla p_h n_e$ is one component of the standard residual-based error estimator; see, e.g., [33]. In particular, this term can be bounded locally by the exact error and a higher order term. Moreover, we find that $\eta_K = 0$, $K \in \mathcal{K}_h^0$ if and only if p_h is linear on K. The second term in the standard residual error estimator is a weighted L^2-norm of $\Pi_W f$. An inverse estimate for Raviart–Thomas finite elements yields

$$\eta_K^2 \ge C\sum_{t\in\mathcal{S}_K}\frac{h_t^2}{a_t}\|\text{div } (u_h - a\nabla p_h)\|^2_{0;t} \ge C\frac{h_K^2}{a_K}\|P_Q f\|^2_{0;K}.$$

By means of Lemma 3.2, we find that $h_K^2/a_K\|P_Q f\|^2_{0;K}$, $K \in \mathcal{K}_h^0$, is bounded by the jump of the discrete flux approximation $a\nabla p_h$. We cannot establish this result for an element $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$. The following lemma provides an equivalence for these elements of the dual mesh.

LEMMA 3.3. *The local contribution η_K, $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$, of the error estimator is equivalent to*

$$c\eta_K^2 \le \frac{h_K^2}{a_K}\|P_Q f\|^2_{0;K} + \sum_{e\in\mathcal{E}_K}\frac{h_e}{a_e}\|[a\nabla p_h n_e]\|^2_{0;e} \le C\eta_K^2, \quad K \in \mathcal{K}_h \setminus \mathcal{K}_h^0.$$

Proof. The starting point for the lower bound is the norm equivalence (3.1). We recall that if $w_h \in U_h$ and div $w_h = 0$ on $K \in \mathcal{K}_h \setminus \mathcal{K}_h^0$ and $w_h n_K = 0$ on ∂K, then $w_h = 0$. Due to the shape regularity, the number of elements t in K is bounded, and thus we obtain

$$\eta_K^2 \le C\sum_{t\in\mathcal{S}_K}\frac{1}{a_t}\|u_h - a\nabla p_h\|^2_{0;t}$$
$$\le C\left(\sum_{t\in\mathcal{S}_K}\frac{h_t^2}{a_t}\|\text{ div } u_h\|^2_{0;t} + \sum_{e\in\mathcal{E}_K}\frac{h_e}{a_e}\|[a\nabla p_h n_e]\|^2_{0;e} + \frac{\hat{\alpha}_K^2}{a_K}\right)$$
$$\le C\left(\frac{h_K^2}{a_K}\|P_Q f\|^2_{0;K} + \sum_{e\in\mathcal{E}_K}\frac{h_e}{a_e}\|[a\nabla p_h n_e]\|^2_{0;e} + \frac{\hat{\alpha}_K^2}{a_K}\right).$$

We now have to consider $\hat{\alpha}_K$ in more detail:

$$\begin{aligned}\hat{\alpha}_K^2 &\le \left(\int_K P_Q f \; dx + \int_{\partial K} a\nabla p_h n_K \; d\sigma\right)^2 \\ &= \left(\int_K P_Q f \; dx + \sum_{t\in\mathcal{S}_K}\int_{\partial t} a\nabla p_h n_t \; d\sigma - \sum_{e\in\mathcal{E}_K}\int_e [a\nabla p_h n_e] \; d\sigma\right)^2 \\ &\le C\left(h_K^2\|P_Q f\|_{0;K}^2 + \sum_{e\in\mathcal{E}_K} h_e\|[a\nabla p_h n_e]\|_{0;e}^2\right).\end{aligned}$$

The upper bound can be easily established by using the norm equivalence (3.2), the definition (2.3), and the observation that $\int_{\partial K} a\nabla p_h n_K \; d\sigma$ is bounded by the jump. As a consequence, we find that the weighted L^2-norm of $P_Q f$ on K can be bounded by a weighted sum of the boundary term $|\hat{\alpha}_K|$ and the jump. ∎

By using Lemmas 3.2 and 3.3, we find that our error estimator can be bounded locally, up to higher order terms, by the discretization error.

THEOREM 3.4. *The local contribution η_K of the error estimator can be bounded locally by the exact discretization error*

$$\eta_K \le C\left(|\!|\!| e_h |\!|\!|_K + \frac{h_K^2}{a_K}\|f - P_Q f\|_{0;K}\right).$$

Proof. We do not provide details. Using Lemmas 3.2 and 3.3, it is sufficient to bound the jump terms and $P_Q f$. This can be done by standard techniques; see, e.g., [3, 8, 33]. To get the local bound on K for the jump, we have to use quadratic edge bubble functions associated with the small edges $e \in \mathcal{E}_K$. ∎

Remark 3.5. Although the starting point for the construction of our error estimator is quite different from an equilibrated error estimator, we can interpret our estimator in that framework. In the case of an equilibrated error estimator, a linear approximation of the flux is locally computed in a postprocessing step on each edge. This can be done by using biorthogonal basis functions and solving, for each vertex, a low dimensional system; see, e.g., [30]. Here, we use an approximation of the flux by Raviart–Thomas elements. Each edge of the primal triangulation is decomposed into two subedges, and for each subedge we compute a constant approximation of the flux. As in the equilibrated situation, we have two degrees of freedom per edge and the arising linear system has the same algebraic structure.

4. A postprocessed flux approximation. In this section, we introduce a postprocessing step. To start, we define an approximation of the flux by

$$u_h^p := \frac{a\nabla p_h + u_h}{2}. \tag{4.1}$$

The following theorem guarantees that our error estimator is asymptotically exact for the flux u_h^p.

THEOREM 4.1. *There exists a constant $0 < C < \infty$ independent of the mesh size such that for all $\epsilon > 0$*

$$\frac{1}{4}(1-\epsilon)\eta^2 - C\left(1+\frac{1}{\epsilon}\right)\xi^2 \le \|a^{\frac{1}{2}}\nabla p - a^{-\frac{1}{2}}u_h^p\|_0^2 \le \frac{1}{4}(1+\epsilon)\eta^2 + C\left(1+\frac{1}{\epsilon}\right)\xi^2.$$

Proof. We start with the definition of the approximation u_h^p and set $u := a\nabla p$,

$$\begin{aligned} 4\|a^{\frac{1}{2}}\nabla p - a^{-\frac{1}{2}}u_h^p\|_0^2 &= \|a^{\frac{1}{2}}(\nabla p - \nabla p_h) + a^{-\frac{1}{2}}(u - u_h)\|_0^2 \\ &= \|a^{\frac{1}{2}}(\nabla p - \nabla p_h)\|_0^2 + \|a^{-\frac{1}{2}}(u - u_h)\|_0^2 + 2(\nabla p - \nabla p_h, u - u_h)_0 \\ &= \|a^{\frac{1}{2}}\nabla p_h - a^{-\frac{1}{2}}u_h\|_0^2 + 4(\nabla p - \nabla p_h, u - u_h)_0 \\ &= \eta^2 + 4\int_\Omega (f - P_Q f)(p - p_h)\, dx. \end{aligned}$$

Using Theorem 3.1 and the results of its proof, the second term on the right side can be bounded by

$$-C(\eta + C\xi)\xi \le 4\int_\Omega (f - P_Q f)(p - p_h)\, dx \le C(\eta + C\xi)\xi.$$

Young's inequality provides upper and lower bounds,

$$\eta^2 - \epsilon\eta^2 - 4C\left(1 + \frac{1}{\epsilon}\right)\xi^2 \le 4\|a^{\frac{1}{2}}\nabla p - a^{-\frac{1}{2}}u_h\|_0^2 \le \eta^2 + \epsilon\eta^2 + 4C\left(1 + \frac{1}{\epsilon}\right)\xi^2. \qquad \square$$

We note that ξ, compared to η, is a higher order term which can be neglected asymptotically. If $f \in H^s(\Omega)$, $0 < s \le 1$, we find that ξ is of order h^{1+s}. Setting $\epsilon = h^s$ and assuming $\eta \ge ch$, we obtain

$$1 - \mathcal{O}(h^s) \le \frac{4\|a^{\frac{1}{2}}\nabla p - a^{-\frac{1}{2}}u_h^p\|_0^2}{\eta^2} \le 1 + \mathcal{O}(h^s).$$

Remark 4.2. The term $a_e/h_e\|p_h - g_D\|_{0;e}^2$, $e \subset \partial\Omega$, enters into the a posteriori estimates, in the case of nonhomogeneous Dirichlet boundary conditions g_D. If the boundary data is smooth enough, this is also a higher order term which can be neglected asymptotically.

Figure 4.1 illustrates the postprocessing in the case that $P_Q f = f$ and homogeneous Dirichlet boundary conditions. In that case, we have $2\|a^{\frac{1}{2}}\nabla p - a^{-\frac{1}{2}}u_h^p\|_0 = \eta$.

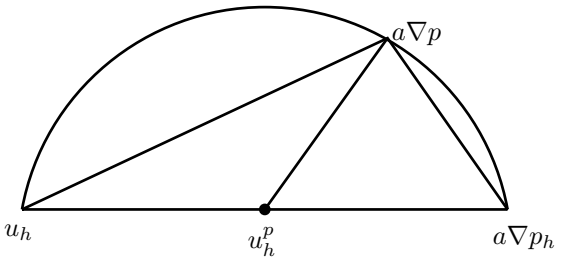


FIG. 4.1. *The discrete flux u_h and the postprocessed flux u_h^p.*

We note that our approach is not restricted to conforming finite elements of lowest order in two dimensions. It can be extended to higher order elements by using higher order Raviart–Thomas finite elements, to the three-dimensional situation and to nonconforming finite elements. However, the construction is then more technical. Nonconforming Crouzeix–Raviart elements are of special interest. Residual- and averaging-based a posteriori error estimators for nonconforming finite elements can

be found in [15, 16, 17]. We do not work out all the details but only point out the basic ideas. In this situation, the construction of a dual mesh is very simple. Each edge of the primal mesh is associated with an element of the dual mesh; see Figure 4.2. The elements of the primal mesh are marked by solid lines and those of the dual and fine mesh by dashed lines, respectively. The shadowed region in the left picture shows an element of the dual mesh.

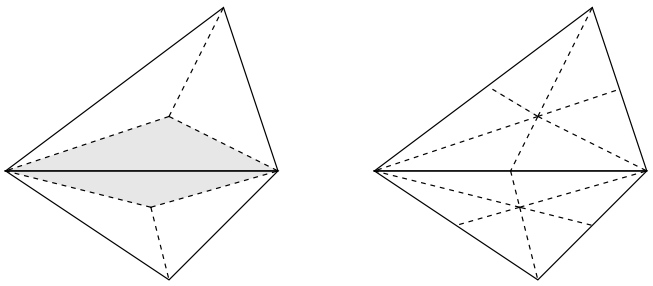


FIG. 4.2. *Primal and dual mesh (left) and primal and fine mesh (right).*

Proceeding as in the conforming setting yields a locally defined Raviart–Thomas finite element of lowest order. However, there is one essential difference. In contrast to conforming finite elements, the Crouzeix–Raviart finite element solution p_h is, in general, not continuous. Thus, applying integration by parts produces an additional term. This term reflects the nonconformity of the finite element solution and can be bounded by the weighted L^2-norm of the jump of p_h. As a consequence, we find that the discretization error cannot be bounded by $\|a^{-\frac{1}{2}}(u_h - a\nabla p_h)\|_0^2$. We must add one term to bound the discretization error from above and below, and we can define the local component of the error estimator on K by

$$\|a^{-\frac{1}{2}}(u_h - a\nabla p_h)\|_{0;K}^2 + \frac{a_{e_K}}{h_{e_K}}\|[p_h]\|_{0;e_K}^2,$$

where e_k is the edge of the primal triangulation which is associated with K. Upper and lower bounds with constants independent of the mesh size can be established. Unfortunately, we cannot guarantee that the upper constant in the bound is one. However, a sharper result can be found in [1]. Here, the same patch as in the conforming situation is used to construct a local error estimator. To estimate the nonconformity, a smoothing step must be carried out; see [1] for details. It can be easily verified that the difference between the nonconforming finite element solution p_h and the smoothed solution is locally equivalent to the jump of p_h on the edges.

5. Numerical results. In this section, we show some numerical results illustrating the efficiency of our error estimator. We provide the global ratio between the exact discretization error and the error estimator for different test examples. Additionally, we consider the local ratio on each element of the dual mesh. The adaptive refinement process is controlled by a standard mean value strategy. We start with an initial coarse triangulation and show the adaptively generated triangulations. We denote the square of the different error components by

$$e_p := \left\|a^{\frac{1}{2}}(\nabla p - \nabla p_h)\right\|_0^2, \quad e_{p;K} := \left\|a^{\frac{1}{2}}(\nabla p - \nabla p_h)\right\|_{0,K}^2,$$
$$e_u := \left\|a^{\frac{1}{2}}\nabla p - a^{-\frac{1}{2}}u_h^p\right\|_0^2, \quad e_{u,K} := \left\|a^{\frac{1}{2}}\nabla p - a^{-\frac{1}{2}}u_h^p\right\|_{0,K}^2.$$

In the case of inhomogeneous Dirichlet boundary conditions, we set

$$e_\Gamma := \sum_{e\in\partial\Omega} \frac{a_e}{h_e} \|p - p_h\|^2_{0;e}\,.$$

To compute the error contributions, we use on each element of the fine triangulation $\mathcal{S}_h$ a Gaussian quadrature formula of higher order or a semianalytical integration scheme if ∇p has a singularity. We note that asymptotically the quadrature error can be neglected. To illustrate the performance of the a posteriori error estimator, we consider the ratio between the exact and the estimated error. We set $\sigma_p := \sqrt{e_p}/\eta$ and $\sigma_u := 2\sqrt{e_u}/\eta$.

In our first example, Table 5.1, we use $\Omega := (0,1)^2$, $a = 1$. The source term and the boundary condition are chosen to match the exact solution:

$$p(x,y) = x(x-1)y(y-1)\exp\left(-100\left(x-\frac{1}{2}\right)^2 - 100\left(y-\frac{117}{1000}\right)^2\right).$$

TABLE 5.1
Example 1: *Error estimator and global errors.*

Nodes	η^2	e_p	σ_p	e_u	σ_u	ξ^2
5	$9.156\ 10^{-2}$	$2.268\ 10^{-1}$	1.574	$1.591\ 10^{-1}$	2.637	$2.316\ 10^{-1}$
13	$9.569\ 10^{-3}$	$2.920\ 10^{-2}$	1.747	$2.408\ 10^{-2}$	3.173	$1.078\ 10^{-1}$
35	$1.126\ 10^{-3}$	$3.417\ 10^{-3}$	1.742	$2.872\ 10^{-3}$	3.194	$3.530\ 10^{-2}$
81	$3.859\ 10^{-4}$	$7.183\ 10^{-4}$	1.364	$6.170\ 10^{-4}$	2.529	$3.320\ 10^{-3}$
158	$2.362\ 10^{-4}$	$2.083\ 10^{-4}$	0.939	$1.181\ 10^{-4}$	1.414	$2.731\ 10^{-4}$
329	$8.167\ 10^{-5}$	$5.717\ 10^{-5}$	0.866	$2.515\ 10^{-5}$	1.110	$2.072\ 10^{-5}$
843	$2.634\ 10^{-5}$	$1.734\ 10^{-5}$	0.811	$7.038\ 10^{-6}$	1.034	$2.203\ 10^{-6}$
2175	$1.061\ 10^{-5}$	$6.031\ 10^{-6}$	0.808	$2.793\ 10^{-6}$	1.033	$5.067\ 10^{-7}$
5091	$4.551\ 10^{-6}$	$2.928\ 10^{-6}$	0.802	$1.165\ 10^{-6}$	1.021	$1.171\ 10^{-7}$
12311	$1.775\ 10^{-6}$	$1.145\ 10^{-6}$	0.803	$4.480\ 10^{-7}$	1.011	$1.847\ 10^{-8}$
29976	$7.583\ 10^{-7}$	$4.901\ 10^{-7}$	0.804	$1.907\ 10^{-7}$	1.003	$4.464\ 10^{-9}$

We observe that σ_u tends asymptotically to one, as predicted by our theory. The ratio between η and $\sqrt{e_p}$ is asymptotically ≈ 0.80. We remark that on the coarse levels, the error e_p is underestimated by the error estimator. This is because the higher order term ξ^2 is dominant at the beginning. The last column of the table shows that ξ^2 can be neglected asymptotically. We recall that ξ^2 measures the quality of the approximation of the right-hand side by the triangulation and has to be controlled during the adaptive refinement process.

Figure 5.1 shows the local distribution of the different error terms and the adaptively generated triangulations on level six. The left and middle pictures have the same scaling and compare the local contribution between η_K^2 and $4e_{u,K}$. To compare η_K^2 and $e_{p,K}$, a rescaling is carried out. It shows that our error estimator captures the local distribution of the error very well.

In the second example, we use an L-shaped domain and set the exact solution of the inhomogeneous Dirichlet problem to be $p = r^{2/3}\sin(2/3\theta)$, where (r,θ) are the polar coordinates and the center is located at the corner of the L-shaped domain. Here, the geometry of the domain and the singularity of the solution at the corner dominate the adaptive refinement process. We note that the higher order term ξ^2 is

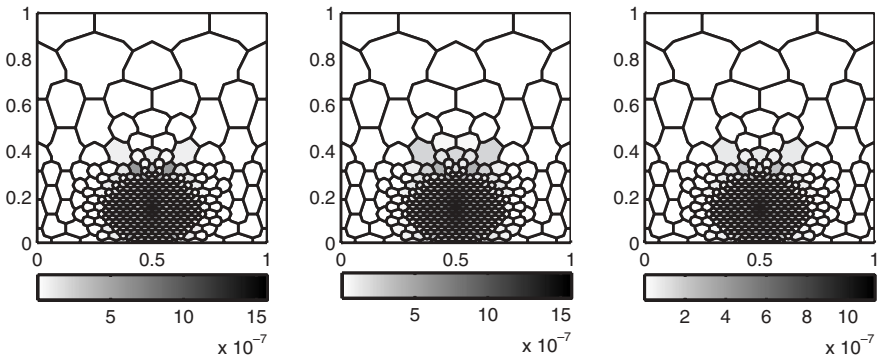


FIG. 5.1. *Example* 1: *Local contributions of the error estimator* η_K^2 *(left), the postprocessed error* $4e_{u,K}$ *(middle), and the error* $e_{p,K}$ *(right).*

TABLE 5.2
Example 2: *Error estimator and global errors.*

Nodes	η^2	e_p	σ_p	e_u	σ_u	e_Γ
11	$3.314\ 10^{-1}$	$1.340\ 10^{-1}$	0.636	$8.622\ 10^{-2}$	1.020	$3.429\ 10^{-3}$
22	$1.487\ 10^{-1}$	$6.445\ 10^{-2}$	0.659	$2.938\ 10^{-2}$	1.028	$2.835\ 10^{-3}$
50	$6.395\ 10^{-2}$	$2.846\ 10^{-2}$	0.667	$1.638\ 10^{-2}$	1.008	$4.459\ 10^{-4}$
82	$3.360\ 10^{-2}$	$1.618\ 10^{-2}$	0.694	$8.615\ 10^{-3}$	1.012	$4.459\ 10^{-4}$
176	$1.473\ 10^{-2}$	$7.258\ 10^{-3}$	0.702	$3.700\ 10^{-3}$	1.002	$5.626\ 10^{-5}$
306	$7.481\ 10^{-3}$	$3.822\ 10^{-3}$	0.715	$1.885\ 10^{-3}$	1.003	$5.626\ 10^{-5}$
566	$3.746\ 10^{-3}$	$1.971\ 10^{-3}$	0.725	$9.401\ 10^{-4}$	1.002	$1.696\ 10^{-5}$
1129	$1.710\ 10^{-3}$	$9.242\ 10^{-4}$	0.736	$4.282\ 10^{-2}$	1.001	$6.478\ 10^{-6}$
2236	$8.342\ 10^{-4}$	$4.596\ 10^{-4}$	0.742	$2.094\ 10^{-4}$	1.002	$2.878\ 10^{-6}$
4669	$3.835\ 10^{-4}$	$2.149\ 10^{-4}$	0.748	$9.600\ 10^{-5}$	1.003	$1.929\ 10^{-6}$
9272	$1.870\ 10^{-4}$	$1.061\ 10^{-4}$	0.753	$4.680\ 10^{-5}$	1.000	$4.004\ 10^{-7}$
18942	$8.830\ 10^{-5}$	$5.066\ 10^{-5}$	0.758	$2.208\ 10^{-5}$	1.000	$1.096\ 10^{-7}$

zero but that we have a nontrivial boundary contribution e_Γ due to the inhomogeneous Dirichlet boundary condition.

Table 5.2 reports the values of the different error terms. Since $f = 0$ and thus $\xi^2 = 0$, we show the influence of the inhomogeneous Dirichlet condition in the last column of Table 5.2. We note that e_Γ does not decrease if no element touching the boundary is refined in the next adaptive step. Compared to η^2, the boundary term e_Γ is of smaller magnitude. The ratio between e_Γ and η^2 tends slowly to zero with the increasing number of refinement steps. During the adaptive refinement process, we have to control this additional term.

From the beginning, σ_u is close to one. The ratio between e_p and η^2 is increasing but it is bounded by one. In this example the asymptotic rate starts late and cannot be observed. We note that the asymptotic ratio σ_u is independent of the problem setting, whereas σ_p depends on the given data. Our theoretical results yield that asymptotically $0 < c \leq \sigma_p \leq 1$.

In Figure 5.2, the local distribution of the errors is given for Example 2. As expected, the highest value is located at the corner singularity. In the case of η_K^2 and $4e_{u,K}$ the scaling is the same, and the local values are quite close.

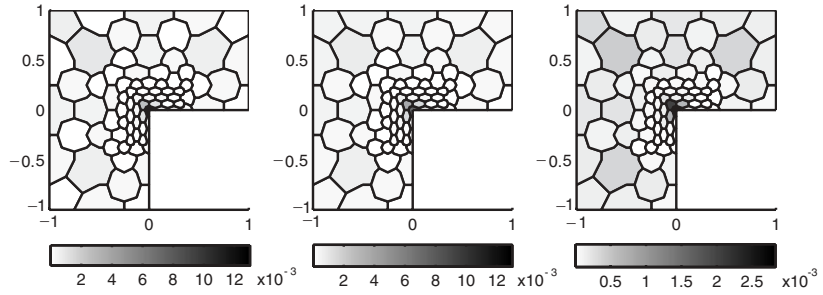


FIG. 5.2. *Example* 2: *Local contributions of the error estimator* η_K^2 *(left), the postprocessed error* $4e_{u,K}$ *(middle), and the error* $e_{p,K}$ *(right).*

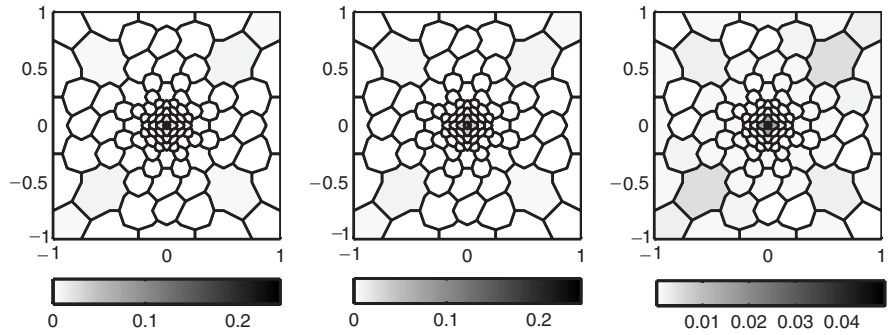


FIG. 5.3. *Example* 3: *Local contributions of the error estimator* η_K^2 *(left), the postprocessed error* $4e_{u,K}$ *(middle), and the error* $e_{p,K}$ *(right).*

In the following example, we use a discontinuous coefficient. The domain Ω is decomposed into four subdomains, $\Omega_{ij} := (-i, -i+1) \times (-j, -j+1)$, $0 \le i, j \le 1$. On each subdomain a constant coefficient a_{ij} is used. We set $a_{10} := a_{01}$, $a_{00} := a_{11}$ and assume that the analytical solution has the form $p = r^\alpha(\beta_{ij}\sin(\alpha\theta)+\gamma_{ij}\cos(\alpha\theta))$, where (r, θ) are the polar coordinates in Ω and β_{ij}, γ_{ij} are constants depending on the subdomains Ω_{ij}. The interface conditions $[p] = 0$ and $[a\nabla p] = 0$ yield $-\mathrm{div}(a\nabla p) = 0$. The coefficients α, β_{ij}, γ_{ij} are uniquely defined in terms of a_{01} and a_{00}. In the case $a_{01} := 1$, $a_{00} := 5$, we find $\alpha = 0.53544094560$, $\beta_{10} = -0.7453559925$, $\gamma_{10} = 2.333333333$, $\beta_{00} = 0.4472135955$, $\gamma_{00} = 1$, $\beta_{11} = -0.9441175905$, $\gamma_{11} = 0.555555556$, $\beta_{01} = -2.401702643$, and $\gamma_{01} = -0.4814814815$.

We report the numerical results for e_p, e_u, and e_Γ in Table 5.3. The ratio σ_u tends asymptotically to one as in the other examples. In contrast to our first example, it is close to one from the first step on. This is due to the fact that $\xi = 0$ and that e_Γ is of smaller magnitude.

Figure 5.3 shows that the local error estimator captures very well the location and the amplitude of the singularity. The isolines of the solution are given in the right picture of Figure 5.4.

In all our examples the local contributions of $4e_{u,K}$ and η_K^2 are quite close; see Figures 5.1–5.3. The error estimator captures the local contributions of $e_{u,K}$ better than the ones of $e_{p,K}$. We note that we compare the local contributions on the dual mesh and not on the primal triangulation.

TABLE 5.3
Example 3: *Error estimator and global errors.*

Nodes	η^2	e_p	σ_p	e_u	σ_u	e_Γ
17	3.797	1.502	0.629	1.001	1.027	$2.500\ 10^{-2}$
49	1.742	$7.251\ 10^{-1}$	0.645	$4.394\ 10^{-1}$	1.004	$2.906\ 10^{-3}$
77	$8.962\ 10^{-1}$	$3.954\ 10^{-1}$	0.664	$2.268\ 10^{-1}$	1.006	$2.906\ 10^{-3}$
151	$4.509\ 10^{-1}$	$2.098\ 10^{-1}$	0.682	$1.150\ 10^{-1}$	1.010	$2.906\ 10^{-3}$
333	$2.205\ 10^{-1}$	$1.022\ 10^{-1}$	0.681	$5.533\ 10^{-2}$	1.002	$3.607\ 10^{-4}$
767	$1.057\ 10^{-1}$	$4.983\ 10^{-2}$	0.686	$2.658\ 10^{-2}$	1.003	$2.674\ 10^{-4}$
1775	$4.974\ 10^{-2}$	$2.346\ 10^{-2}$	0.687	$1.254\ 10^{-2}$	1.004	$2.053\ 10^{-4}$
4151	$2.368\ 10^{-2}$	$1.112\ 10^{-2}$	0.685	$5.930\ 10^{-3}$	1.001	$2.676\ 10^{-5}$
9621	$1.119\ 10^{-2}$	$5.256\ 10^{-3}$	0.686	$2.779\ 10^{-3}$	1.000	$4.874\ 10^{-6}$
22397	$5.261\ 10^{-3}$	$2.461\ 10^{-3}$	0.684	$1.315\ 10^{-3}$	1.000	$7.083\ 10^{-7}$

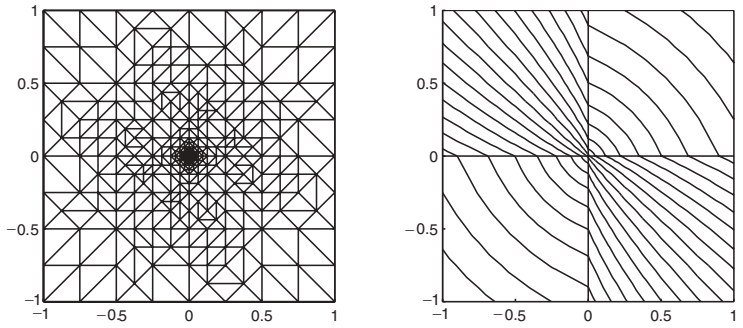


FIG. 5.4. *Example* 3: *Adaptively generated mesh (step* 6*) and isolines of* p_h.

We compare the standard residual-based error estimator with our estimator. Figure 5.5 shows the errors e_p and e_u with respect to the number of unknowns. In the two left pictures, we use $a_{01} := 1$ and $a_{00} := 5$. The two right pictures show the results for $a_{01} := 1$ and $a_{00} := 100$. The efficiency of both error estimators is almost the same for the first setting. Increasing a_{00} from 5 to 100 demonstrates the quality of our error estimator. We note that in the pictures the discretization errors are shown, not the error estimators. In the case $a_{00} = 100$, the solution has a higher singularity at the center and our new error estimator yields better results than the standard residual-based estimator. To obtain the same accuracy, considerably more degrees of freedom are required if the adaptive refinement process is controlled by the residual-based error estimator. We remark that in contrast to the residual-based error estimator, the tensor a enters by the bilinear form $a(\cdot,\cdot)$ in the calculation of η.

In our next example, we consider the domain Ω defined in Figure 5.6. Ω is decomposed into three subdomains Ω_i, $0 \le i \le 2$. On each subdomain a constant coefficient a_i is used. We set $a_2 = a_0$, $a_1 = 1$ and $f = 100$.

Figure 5.7 shows the adaptive meshes on levels three and four. For $a_0 = 10$, there is no significant difference in the mesh size on the three different subdomains. As a_0 increases, we observe stronger adaptive refinement in Ω_1. In the case $a_0 = 1000$, the mesh size is considerably smaller in the subdomain Ω_1 than in Ω_0 and Ω_2.

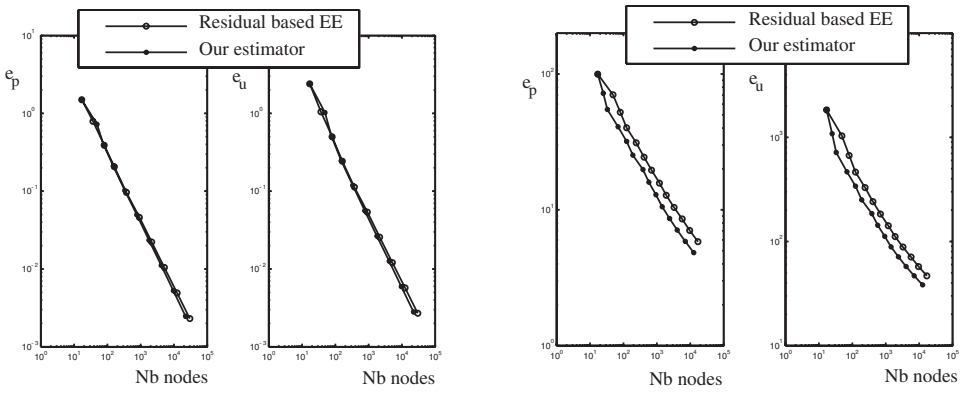


FIG. 5.5. *Discretization errors e_p and e_u for $a_{00} = 5$ (left) and $a_{00} = 100$ (right).*

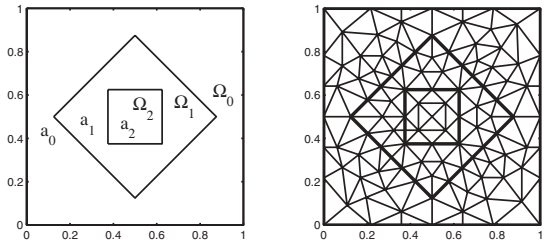


FIG. 5.6. *Example* 4: *Geometry and initial coarse mesh.*

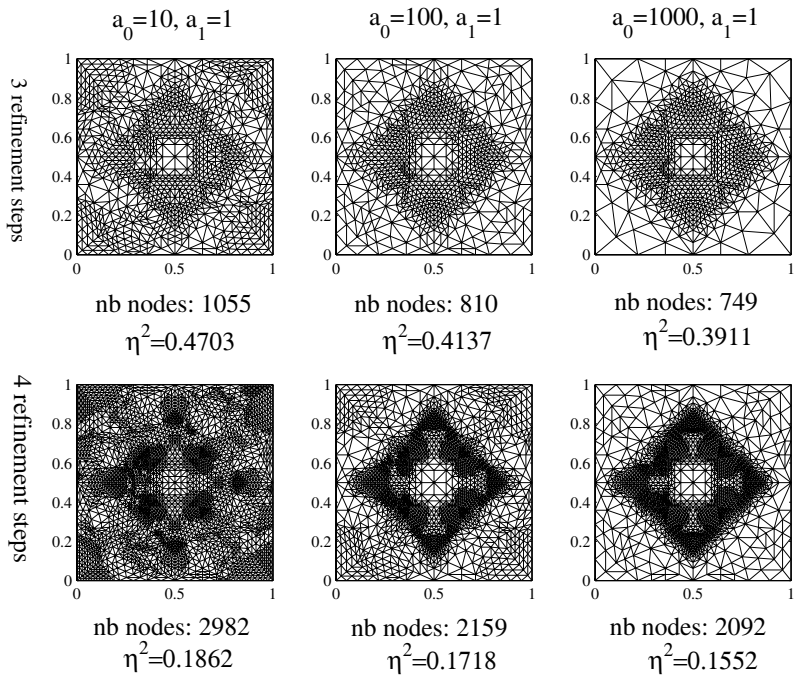


FIG. 5.7. *Example* 4: *Adaptive refinement on Level* 3 *(upper) and Level* 4 *(lower).*

REFERENCES

[1] M. Ainsworth, *Robust A Posteriori Error Estimation for Non-Conforming Finite Element Approximation*, Tech. Report NI03008, Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, 2003.

[2] M. Ainsworth and I. Babuška, *Reliable and robust a posteriori error estimation for singularly perturbed reaction-diffusion problems*, SIAM J. Numer. Anal., 36 (1999), pp. 331–353.

[3] M. Ainsworth and J. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley, Chichester, UK, 2000.

[4] I. Babuška, R. Durán, and R. Rodríguez, *Analysis of the efficiency of an a posteriori error estimator for linear triangular finite elements*, SIAM J. Numer. Anal., 29 (1992), pp. 947–964.

[5] I. Babuška and A. Miller, *A feedback finite element method with a posteriori error estimation.* I: *The finite element method and some basic properties of the a posteriori error estimator*, Comput. Methods Appl. Mech. Engrg., 61 (1987), pp. 1–40.

[6] I. Babuška and W. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[7] I. Babuška and W. Rheinboldt, *A posteriori error estimates for the finite element method*, Int. J. Numer. Methods Engrg., 12 (1978), pp. 1597–1615.

[8] I. Babuška and T. Strouboulis, *The Finite Element Methods and Its Reliability*, Clarendon Press, Oxford, UK, 2001.

[9] R. Bank and A. Weiser, *Some a posteriori error estimates for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[10] R. Becker and R. Rannacher, *A feed-back approach to error control in finite element methods: Basic analysis and examples*, East-West J. Numer. Math., 4 (1996), pp. 237–264.

[11] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Numer. Math., 85 (2000), pp. 579–608.

[12] F. A. Bornemann, B. Erdmann, and R. Kornhuber, *A posteriori error estimates for elliptic problems in two and three spaces dimensions*, SIAM J. Numer. Anal., 33 (1996), pp. 1188–1204.

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

[14] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids.* I: *Low order conforming, nonconforming, and mixed FEM*, Math. Comp., 71 (2002), 945–969.

[15] C. Carstensen, S. Bartels, and S. Jansche, *A posteriori error estimates for nonconforming finite element methods*, Numer. Math., 92 (2002), pp. 233–256.

[16] C. Carstensen and S. Funken, *Constants in Clément-interpolation error and residual based a posteriori error estimates in finite element methods*, East-West J. Numer. Anal., 8 (2000), pp. 153–175.

[17] E. Dari, R. Durán, C. Padra, and V. Vampa, *A posteriori error estimates for nonconforming finite element methods*, Math. Model. Numer. Anal., 30 (1996), pp. 385–400.

[18] P. Deuflhard, P. Leinen, and H. Yserentant, *Concepts of an adaptive hierarchical finite element code*, IMPACT Comput. Sci. Engrg., 1 (1989), pp. 3–35.

[19] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[20] R. Durán and R. Rodríguez, *On the asymptotic exactness of Bank-Weiser's estimator*, Numer. Math., 62 (1992), pp. 297–303.

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

[22] P. Ladevèze and D. Leguillon, *Error estimate procedure in the finite element method and applications*, SIAM J. Numer. Anal., 20 (1983), pp. 485–509.

[23] P. Morin, R. H. Nochetto, and K. G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.

[24] M. Ohlberger, *A posteriori estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations*, *M2AN* Math. Model. Numer. Anal., 35 (2001), pp. 355–387.

[25] M. Petzoldt, *A posteriori error estimators for elliptic equations with discontinuous coefficients*, Adv. Comput. Math., 16 (2002), pp. 47–45.

[26] R. Rannacher and F.-T. Suttmeier, *A feed-back approach to error control in finite element methods: Application to linear elasticity*, Comput. Mech., 19 (1997), pp. 434–446.

[27] P. Raviart and J. Thomas, *A mixed finite element method for 2nd order elliptic problems*, Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.

[28] S. Repin, S. Sauter, and A. Smolianski, *A posteriori error estimation for the Dirichlet problem with account of the error in the approximation of boundary conditions*, Computing, 70 (2003), pp. 205–233.

[29] R. Rodríguez, *Some remarks on the Zienkiewicz-Zhu estimator*, Numer. Methods Partial Differential Equations, 10 (1994), pp. 625–636.

[30] E. Stein and S. Ohnimus, *Equilibrium method for postprocessing and error estimation in the finite element method*, Comput. Assist. Mech. Engrg. Sci., 4 (1997), pp. 645–666.

[31] R. Verfürth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Math. Comp., 62 (1994), pp. 445–475.

[32] R. Verfürth, *A posteriori error estimation and adaptive mesh-refinement techniques*, J. Comp. Appl. Math., 50 (1994), pp. 67–83.

[33] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Ser. Adv. Numer. Math., John Wiley, Chichester, UK, 1996.

[34] J. Zhu and O. Zienkiewicz, *Adaptive techniques in the finite element method*, Commun. Appl. Numer. Methods, 4 (1988), pp. 197–204.

[35] O. Zienkiewicz and J. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, J. Numer. Meth. Engrg., 28 (1987), pp. 28–39.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1415–1434

TENSORIAL RATIONAL SURFACES WITH BASE POINTS VIA MASSIC VECTORS*

OLIVIER GIBARU†

Abstract. Let S be a tensorial rational surface defined by a rational function from $[0,1]^2$ onto $\mathbb{R}^3$ with a base point at $(u,v)=(0,0)$. We demonstrate that the image of this base point is a set of rational curves; a base point is a parameter value for which the rational parametrization takes the value $(\frac{0}{0},\frac{0}{0},\frac{0}{0})$. This result was established by Clebsch [*Ueber die abbildung algebraischer flächen, insbesondere der vierten und fünften ordnung,* Math. Ann., 1 (1869), pp. 253–316]. Base points were first introduced in the context of geometric design by Chionh [*Base Points, Resultants and the Implicit Representation of Rational Surfaces,* Ph.D. thesis, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1990] and Manocha and Canny [*Implicitizing Rational Parametric Surfaces*, Tech. report 90/592, Computer Science Division, University of California, Berkeley, 1990] and were used by Warren [*ACM Trans. Graphics,* 11 (1992), pp. 127–139] to define multisided rational Bézier patches. We give here a constructive approach of this result to exploit it directly in the industrial scope of computer aided design. We show that these rational curves are placed end to end by using the formalism of massic vectors introduced by Fiorot and Jeannin [*Courbes et Surfaces Rationnelles. Applications à la CAO*, R.M.A. 12, Masson, Paris, 1989]. Furthermore, we give the relations between the massic vectors which define these curves and the massic vectors which define S. Finally, we give an algorithm to draw on a computer a surface having base points.

Key words. base point, Newton polygon, SBR-surfaces, BR-curves, blow up, blow down

AMS subject classifications. 65D17, 14B05, 65D10, 65D05, 68U05

DOI. 10.1137/S0036142903420972

1. Introduction.

The aim of this article is to study the image of the base points of rational surfaces put on the form of SBR-surfaces (see [11], [10], [24]). The representation of tensor product rational surfaces in the SBR-form is quite relevant. Indeed, the control of the classical rational Bézier surfaces (see [9]) is obtained only via weighted points where the masses of the points are positive reals. Hence, there is no identity between rational surfaces and rational Bézier surfaces. This is why Fiorot and Jeannin introduced the massic vector space to control any rational surfaces. In addition, they showed that one needs pure vectors and weighted points with negative masses to obtain this identity with rational surfaces. The elements of the massic vector space are called massic vectors; they are either weighted points or pure vectors.

For applications, rational surfaces are usually defined on the square domain $[0,1]^2$. Therefore, we shall state some results on the images of base points which are located on the vertices of $[0,1]^2$. The formalism of massic vectors will help us to obtain these images more precisely. We shall demonstrate that the image of a base point of a rational surface is a set of rational curves whose massic vectors are obtained from the massic vectors of the initial SBR-surface.

We shall see that this type of singularity, which one wishes to avoid in general, is very useful for applications in computer aided geometric design.

Indeed, the users of computer aided design systems are often faced with the problem of filling in a many-sided hole, for example, when trying to define the body of a car or joint surface of a mould in smelting.

*Received by the editors January 7, 2003; accepted for publication (in revised form) December 18, 2003; published electronically December 16, 2004.
http://www.siam.org/journals/sinum/42-4/42097.html
†Laboratoire L2MA, Ecole Nationale Supérieure d'Arts et Métiers de Lille, 59046 Lille Cedex, France (olivier.gibaru@lille.ensam.fr).

Given an n-sided hole left by n contiguous connecting surfaces, either along a common boundary curve or by a point, our aim is to build a surface, called a filling surface, to fill in this hole. The surfaces needed are either Bézier–de Casteljau [1], [7] polynomial surfaces or SBR-surfaces (see [11], [10]). In this case we have to define an n-sided filling SBR-surface to join these n given surfaces along the curves which surround the hole. In [31] and [30], Warren used base points to create n-sided rational Bézier patches. Consequently, to create such a filling surface as a single entity we need to introduce base points at the vertices of the parameter domain $[0,1]^2$ of a rational surface to obtain this n-sided SBR-surface ($n \geq 5$). The additional boundary curves resulting from the images of these base points are obtained by the use of the fundamental Theorem 3.5. As a consequence, this filling n-sided SBR-surface will be defined in one piece on $[0,1]^2$.

In [22], the author introduced another approach via toric patches associated to lattice polygon to deal with this problem of multisided patches (see [20] for related algorithms). Nevertheless, this model of surface takes into account only positive weights.

In section 2, we recall some results concerning massic vectors which allow us to control rational curves or rational surfaces, as was done with points by Bézier–de Casteljau for polynomial curves or surfaces [1], [7]. We give the explicit form of the rational curves and surfaces in terms of massic vectors. The main theorem demonstrated in section 3 is that the image of a base point of a rectangular SBR-surface is a set of rational curves. This result was established by Clebsch in [5] and mentioned in [27]. We state this result via the formalism of massic vectors. Therefore, we consider a more general framework than did Warren. Furthermore, we need to define the Newton polygon of an SBR-surface from the set of indices of the nonzero massic vectors. We show that the massic vectors of the BR-curve images of a base point are those (within a constant) which belong to boundaries of this Newton polygon. With an assumption on the masses of the massic vectors we demonstrate that the curves are placed end to end. A corollary to this theorem is that the image of a base point may be either a curve or reduced to a point. In section 4 we shall examine the problem of drawing rectangular SBR-surfaces with base points at corners on a computer. In [22], the author used the concept of facet variables (see [6]) associated to the lattice polygon of a toric patch (which is in fact the Newton polygon of the patch) to draw it. The method is based on a subdivision process where an n-sided patch is divided into n adjacent tensor product rational Bézier surfaces. Contrary to this strategy (under the conditions of Theorem 3.5), we propose to draw the surface by also using the properties of the Newton polygon of the surface, but we apply successive quadratic changes of variables to obtain n rectangular SBR surfaces without base points. It should be noted that in [24], [25], and [23], the authors used implicit representation to represent a tensor product rational surface. They developed a simple algorithm for implicitizing a rational surface with base points based on perturbation and symbolic manipulation. As we can see, the Newton polygon plays a key role for drawing a rational surface with base points and multisided toric patches. It is also the key for other problems like the efficient computation of sparse mixed resultants (see [2]).

In section 5, we construct from a given set of seven rational BR-curves a seven-sided SBR-surface defined via a rectangular one containing one base point at $(0,0)$. We get three consecutive rational curves arbitrarily taken among the set of given curves as the image of this base point.

2. BR-form and SBR-form of rational curves and surfaces. We know that any rational curve of an affine space in three dimensions (respectively, n), denoted $\mathcal{E}$, is the conic projection of a polynomial curve of an affine space $\mathcal{F}$ in fourth dimension (respectively, $(n+1)$th). The latter can be expressed as a polynomial Bézier–de Casteljau curve. The conic projection of the affine space $\mathcal{F}$ onto $\mathcal{E}$ of a polynomial Bézier–de Casteljau curve whose control points are projected to a finite distance is a rational Bézier–de Casteljau curve controlled by weighted points. A particular problem arises when certain control points of the polynomial Bézier–de Casteljau curve of $\mathcal{F}$ are projected to infinity. In this case, it is necessary to add vectors of $\vec{\mathcal{E}}$ (the vector space associated with the affine space $\mathcal{E}$) to the weighted points to control any rational curve of $\mathcal{E}$. We call massic vectors either weighted points of $\mathcal{E}$ or vectors of $\vec{\mathcal{E}}$. They lead to the BR-form of any rational curve.

To understand these notions, according to [11], we recall some basic definitions and results concerning the representation of rational curves (respectively, rational surfaces) by BR-curves (respectively, SBR-surfaces).

Let us define $\mathcal{E}$ (respectively, $\mathcal{F}$) a real affine space, $\vec{\mathcal{E}}$ (respectively, $\vec{\mathcal{F}}$) its associated linear vector space such that $\mathcal{E}$ (respectively, $\vec{\mathcal{E}}$) is a hyperplane of $\mathcal{F}$ (respectively, $\vec{\mathcal{F}}$), and $\widetilde{\mathcal{E}}$ the projective completion of $\mathcal{E}$. Let Ω be a point of $\mathcal{F}$ not belonging to $\mathcal{E}$.

We define the linear vector space $\widehat{\mathcal{E}} = (\mathcal{E} \times \mathbb{R}^*) \cup \vec{\mathcal{E}}$ called the massic vector space : $\theta \in \widehat{\mathcal{E}}$ is called a massic vector; it is either a weighted point of $\mathcal{E}$ denoted by the couple (P, α), where $P \in \mathcal{E}$ and $\alpha \in \mathbb{R}^*$, or a pure vector, denoted by $\vec{u} \in \vec{\mathcal{E}}$. The one-to-one map $\widehat{\Omega} : \widehat{\mathcal{E}} \to \vec{\mathcal{F}}$ defined by $\widehat{\Omega}(P, \alpha) = \alpha.\overrightarrow{\Omega P}$, $\widehat{\Omega}(\vec{u}) = \vec{u}$ induces an addition operator and an external multiplication in $\widehat{\mathcal{E}}$, respectively denoted $\oplus$ and $*$, such that $\widehat{\mathcal{E}}$ is a linear space and $\widehat{\Omega}$ is an isomorphism.

For any $\theta, \theta' \in \widehat{\mathcal{E}}$ and $\lambda \in \mathbb{R}^* : \theta \oplus \theta' = \widehat{\Omega}^{-1}(\widehat{\Omega}(\theta) + \widehat{\Omega}(\theta'))$ and $\lambda * \theta = \widehat{\Omega}^{-1}(\lambda.\widehat{\Omega}(\theta))$. These operations on massic vectors were first introduced by Grassmann in [21]; see also the survey by Cartan in [3] and the book by Goldman [20].

We shall define the linear form $\chi : \widehat{\mathcal{E}} \to \mathbb{R}$; $\chi(P, \alpha) = \alpha$, $\chi(\vec{u}) = 0$. $\chi(\theta)$ is called the mass of θ. Let $\Pi\Omega : \vec{\mathcal{F}} \backslash \{\vec{0}\} \to \widetilde{\mathcal{E}}$ be the conic projection of apex Ω and $\Pi : \widehat{\mathcal{E}} \backslash \{\vec{0}\} \to \widetilde{\mathcal{E}}$ the natural projection : $\Pi(P, \alpha) = P$, $\Pi(\vec{u}) = (\vec{u})_\infty$, where $(\vec{u})_\infty$ designates the point at infinity of $\mathcal{E}$ along the direction $\vec{u}$. These projections are linked by the relation $\Pi = \Pi\Omega \circ \widehat{\Omega}$. We have $\Pi(\lambda * \theta) = \Pi(\theta)$ for all $\lambda \in \mathbb{R}^*$. We can now define a BR-curve.

DEFINITION 2.1. *Let $\theta_0, \theta_1, \ldots, \theta_n$ be $n+1$ massic vectors not simultaneously null ; a BR-curve of $\widetilde{\mathcal{E}}$, of controlling massic polygon $\theta = \{\theta_i, 0 \leqslant i \leqslant n\}$, denoted by* $\mathrm{BR}[\theta_0, \theta_1, \ldots, \theta_n]$ *or* $\mathrm{BR}[\theta]$*, is parametrized by*

$$\mathrm{BR}[\theta](t) = \Pi(\mathrm{BP}[\theta_0, \theta_1, \ldots, \theta_n; [0,1]](t)),$$

where $BP[\theta_0, \theta_1, \ldots, \theta_n; [0,1]](t) = \sum_{i=0}^{n} B_i^n(t) * \theta_i$ is the Bézier–de Casteljau curve in $\widehat{\mathcal{E}}$, n is called the length of the BR-curve.

For $i = 0, \ldots, n$, the $B_i^n(t) = \binom{n}{i}(1-t)^{n-i} t^i$ are Bernstein's polynomials of degree n relative to $[0,1]$ where $\binom{n}{i} = \frac{n!}{(n-i)!i!}$ denotes the binomial coefficients.

PROPOSITION 2.2. *Let $\theta = (\theta_i)_{0 \leq i \leq n}$ be a set of massic vectors which define a BR-curve. Let us define the sets $I := \{i : \theta_i = (P_i, \beta_i) \in \mathcal{E} \times \mathbb{R}^*\}$ and $\bar{I} := \{i : \theta_i = \vec{U}_i \in \vec{\mathcal{E}}\}$ such that $I \cup \bar{I} = \{0, 1, \ldots, n\}$ and $I \cap \bar{I} = \emptyset$ and the mass of the curve denoted by $\beta(t) = \sum_{i \in I} \beta_i B_i^n(t)$. Hence the explicit form of this BR-curve is*

(a) *if* $\beta(t) \neq 0$, *then*

$$\mathrm{BR}[\theta](t) = \frac{\sum_{i\in I}\beta_i B_i^n(t) P_i}{\beta(t)} + \frac{\sum_{i\in \bar{I}} B_i^n(t)\vec{U}_i}{\beta(t)};$$

(b) *if* $\beta(t) = 0$ *and* $\vec{V}(t) = \sum_{i\in I}\beta_i B_i^n(t) P_i + \sum_{i\in\bar{I}} B_i^n(t)\vec{U}_i \neq \vec{0}$, *then* $\mathrm{BR}[\theta](t)$ *defined the point at infinity*

$$\mathrm{BR}[\theta](t) = \left(\sum_{i\in I}\beta_i B_i^n(t) P_i + \sum_{i\in\bar{I}} B_i^n(t)\vec{U}_i\right)_\infty$$

along the direction given by the vector $\sum_{i\in I}\beta_i B_i^n(t) P_i + \sum_{i\in\bar{I}} B_i^n(t)\vec{U}_i$;

(c) *if* $\vec{V}(t_0) = \vec{0}$ (*hence* $\beta(t_0) = 0$), *then* $\mathrm{BR}[\theta](t_0) = \lim_{t\to t_0}\mathrm{BR}[\theta](t)$.

Proof. For the proof, see Proposition 1.3 in [11]. □

DEFINITION 2.3. *Let* $\theta = (\theta_{ij})_{\substack{0\le i\le n\\ 0\le j\le p}}$ *be a set of massic vectors. A rectangular SBR-surface, denoted by* $\mathrm{SBR}[\theta]$, *is defined by*

$$\mathrm{SBR}[\theta](u,v) = \Pi\left(\sum_{i=0}^{n}\sum_{j=0}^{p} B_i^n(u) B_j^p(v) * \theta_{ij}\right) \text{ for all } (u,v)\in[0,1]^2.$$

This SBR-surface is said to be of length n *and of width* p.

THEOREM 2.4. *There is an identity between the set of rational surfaces (respectively, rational curves) and rectangular SBR-surfaces (respectively, BR-curves).*

Proof. For the proof, see Propositions 2.2.2.2 and 6.3.1.3 in [11]. □

PROPOSITION 2.5. *Let* $\theta = (\theta_{ij})_{\substack{0\le i\le n\\ 0\le j\le p}}$ *be a set of massic vectors which define a SBR-surface. Let us define* $I := \{(i,j) : \theta_{ij} = (P_{ij}, \beta_{ij}) \in \mathcal{E}\times\mathbb{R}^*\}$, $\bar{I} := \{(i,j) : \theta_{ij} = \vec{U}_{ij} \in \vec{\mathcal{E}}\}$, *and*

$$\beta(u,v) := \chi\left(\sum_{i=0}^{n}\sum_{j=0}^{p} B_i^n(u) B_j^p(v) * \theta_{ij}\right) = \sum_{(i,j)\in I} B_i^n(u) B_j^p(v)\beta_{ij}$$

the mass of S. *The explicit form of this rectangular SBR-surface is*

(a) *if* $\beta(u,v) \neq 0$, *then*

$$\mathrm{SBR}[\theta](u,v) = \frac{\sum_{(i,j)\in I}\beta_{ij} B_i^n(u) B_j^p(v) P_{ij}}{\beta(u,v)} + \frac{\sum_{(i,j)\in\bar{I}} B_i^n(u) B_j^p(v)\vec{U}_{ij}}{\beta(u,v)};$$

(b) *if* $\beta(u,v) = 0$ *and* $\vec{V}(u,v) = \sum_{(i,j)\in I}\beta_{ij} B_i^n(u) B_j^p(v) P_{ij} + \sum_{(i,j)\in\bar{I}} B_i^n(u) B_j^p(v)\vec{U}_{ij} \neq \vec{0}$, *then* $\mathrm{SBR}[\theta](u,v)$ *defines the point at infinity*

$$\left(\sum_{(i,j)\in I}\beta_{ij} B_i^n(u) B_j^p(v) P_{ij} + \sum_{(i,j)\in\bar{I}} B_i^n(u) B_j^p(v)\vec{U}_{ij}\right)_\infty;$$

(c) *if* $\vec{V}(u_0,v_0) = \vec{0}$ (*hence* $\beta(u_0,v_0) = 0$), *then* $\mathrm{SBR}[\theta](u_0,v_0)$ *is not defined.* $\mathrm{SBR}[\theta](u,v)$ *takes on the value* $\left(\frac{0}{0},\frac{0}{0},\frac{0}{0}\right)$ *and we say that the SBR-surface has a base point at* $(u,v) = (u_0,v_0)$.

Proof. For the proof, see Proposition 6.3.1.5 in [11]. □

From now on, we assume that the base point is at $(u,v) = (0,0)$. If this is not the case, we apply an affine change of variables to make it so.

LEMMA 2.6. *Let $\hat{f}:[0,1]^2 \to \widehat{\mathcal{E}}$ be a polynomial function defined by*

$$\hat{f}(u,v) = \sum_{i=0}^{n}\sum_{j=0}^{p} B_i^n(u)\, B_j^p(v) * \theta_{ij}, \tag{2.1}$$

where $\theta = (\theta_{ij})_{\substack{0\le i\le n\\ 0\le j\le p}}$ is a set of massic vectors. Then we have $\vec{V}(0,0) = \vec{0} \Leftrightarrow \hat{f}(0,0) = \vec{0} \Leftrightarrow \theta_{00} = \vec{0}$.

Proof. We remark that $\vec{V}(u,v) = \widehat{\Omega}(\hat{f}(u,v))$ and that $\hat{f}(0,0) = \theta_{00}$. □

The main subject of this paper is to handle this singularity as thoroughly as possible for further applications in the industrial scope of computer aided design.

3. Image of a rectangular SBR-surface base point. The aim of this section is to show that there is a deep connection between the lattice polytope (see [6]), here defined by the indices of the nonzero massic vectors of a rectangular SBR-surface, and the image of its base points.

The following theorem highlights the link between the null derivatives of $\hat{f}$ defined by (2.1) at a base point $(u,v) = (0,0)$ (i.e., $\hat{f}(u,v) = \vec{0}$) of the rational surface issued from the Π-projection of $\hat{f}$ and the null massic vectors of $\hat{f}$ in the vicinity of $\theta_{00} = \vec{0}$.

THEOREM 3.1. *Let $\hat{S}$ be a polynomial surface of $\widehat{\mathcal{E}}$ defined by (2.1). Consider $n_0 < n$ and let $q(0) \ge q(1) \ge \cdots \ge q(n_0) \ge 0$ be n_0+1 given integers with $p > q(0)$. Hence, $\partial^{k+l}\hat{f}(0,0)/\partial u^k \partial v^l = \vec{0}$ for $k = 0,\dots,n_0$ and $l = 0,\dots,q(k)$ if and only if $\theta_{kl} = \vec{0}$ for $k = 0,\dots,n_0$ and $l = 0,\dots,q(k)$.*

Moreover, if we put $\hat{f}$ in the monomial form as $\hat{f}(u,v) = \sum_{i=0}^{n}\sum_{j=0}^{p} u^i v^j * \alpha_{ij}$, we obtain that $\alpha_{kl} = \vec{0}$ for $k = 0,\dots,n_0$ and $l = 0,\dots,q(k)$ if and only if $\theta_{kl} = \vec{0}$ for $k = 0,\dots,n_0$ and $l = 0,\dots,q(k)$.

Proof. We shall use the following relation: for any $(k,l) \in \mathbb{R}^2$ we have

$$\frac{\partial^{k+l}\hat{f}(0,0)}{\partial u^k \partial v^l} = \frac{n!p!}{(n-k)!\,(p-l)!} * \Delta^{kl}\theta_{00}, \tag{3.1}$$

where $\Delta^{kl}\theta_{00}$ is the forward difference operator defined by

$$\Delta^{kl}\theta_{00} = \sum_{i=0}^{k}\sum_{j=0}^{l}\binom{k}{i}\binom{l}{j}(-1)^{k+l-i-j} * \theta_{ij}. \tag{3.2}$$

We shall demonstrate the necessary condition by induction on k. Assuming that $\partial^{k+l}\hat{f}(0,0)/\partial u^k \partial v^l = \vec{0}$, it follows from (3.1) that $\Delta^{kl}\theta_{00} = \vec{0}$. For $k = 0$ and $l = 0,\dots,q(0)$, the set of relations $\Delta^{0l}\theta_{00} = \vec{0}$ is a triangular linear system which directly gives $\theta_{00} = \theta_{01} = \cdots = \theta_{0q(0)} = \vec{0}$. Assuming that $\theta_{jl} = \vec{0}$ for $j = 0,\dots,k$ ($k < n_0$) and $l = 0,\dots,q(j)$, then the conditions $\Delta^{k+1,l}\theta_{00} = \vec{0}$ for $l = 0,\dots,q(k+1)$ with $q(k+1) \le q(k) \le \cdots \le q(0)$ imply that $\theta_{k+1,l} = \vec{0}$ for $l = 0,\dots,q(k+1)$, so completing the induction proof. The sufficient condition is obvious via (3.1) and (3.2). The previous result and the relation $\alpha_{ij} = \binom{n}{i}\binom{p}{j} * \Delta^{ij}\theta_{00}$ enable us to conclude the proof. □

Remark. Obviously, in the previous theorem we may change the role of k and l.

Let us consider the following grid of massic vectors:

θ_{0p}	$\cdots$	$\cdots$	$\cdots$	$\cdots$	θ_{np}
$\vdots$					$\vdots$
θ_{03}	θ_{13}				$\vdots$
$\vec{0}$	$\vec{0}$	θ_{22}			$\vdots$
$\vec{0}$	$\vec{0}$	θ_{21}			$\vdots$
$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{30}	$\cdots$	θ_{n0}

,

where at least $\theta_{00} = \vec{0}$. Following the previous theorem for $k = 0, 1, 2$ and $l = 0, \ldots, q(k)$ with $q(0) = q(1) = 2$, $q(2) = 0$, we have $\partial^{k+l}\hat{f}(0,0)/\partial u^k \partial v^l = \vec{0}$. Of course, partial derivatives of $\hat{f}$ can be equal to zero for orders (k, l) different from those above given by the theorem; for instance, we may have $\partial^4 \hat{f}(0,0)/\partial u^2 \partial v^2 = \vec{0}$ if $\theta_{22} = 2 * \theta_{21}$.

As we shall see later, knowing the minimum orders of the nonzero partial derivatives of $\hat{f}$ at a base point is not sufficient to determine the images of this base point. In the previous example the minimum orders are $(k, q(k) + 1)$, $k = 0, 1, 2$. If $\chi(\theta_{03}) > 0$, $\chi(\theta_{30}) > 0$, and $\chi(\theta_{21}) \geq 0$, then we shall show that the image of the base point $(u, v) = (0, 0)$ is the rational cubic defined by the massic vectors θ_{03}, $\theta_{12} = \vec{0}$, θ_{21}, θ_{30} (we have, respectively, $\partial^3 \hat{f}(0,0)/\partial v^3 \neq \vec{0}$, $\partial^3 \hat{f}(0,0)/\partial u \partial v^2 = \vec{0}$, $\partial^3 \hat{f}(0,0)/\partial u^2 \partial v \neq \vec{0}$, $\partial^3 \hat{f}(0,0)/\partial u^3 \neq \vec{0}$); θ_{13} corresponding to $\partial^3 \hat{f}(0,0)/\partial u^1 \partial v^3 \neq \vec{0}$ is not concerned.

Consider a rational surface defined by $\theta = (\theta_{ij})_{\substack{0 \leq i \leq n \\ 0 \leq j \leq p}}$, a set of massic vectors.

Assumption 1. We assume that at least $\theta_{00} = \vec{0}$ and among the massic vectors θ_{i0}, $i \in \{1, \ldots, n\}$ (respectively, among the massic vectors θ_{0j}, $j \in \{1, \ldots, p\}$), there exists at least a nonzero massic vector.

We define the set $\mathcal{A} := \{p_{ij} = (i, j) \in \mathbb{R}^2 : \theta_{ij} \neq \vec{0}\}$ so as to create the Newton polygon of an SBR-surface.

DEFINITION 3.2. *The point $Q \in \mathcal{A}$ is an extreme point if it does not exist two different points q_1 and q_2 belonging to $\mathcal{A}$ such that $Q \in]q_1, q_2[$.*

DEFINITION 3.3. *Let $Q_0, Q_1, \ldots, Q_l$ be the extreme points of $\mathcal{A}$. The convex hull of $\mathcal{A}$ is defined by* $\mathrm{Conv}(\mathcal{A}) := \{\lambda_0 Q_0 + \cdots + \lambda_l Q_l,\ \lambda_i \in \mathbb{R}^+ : \sum_{i=0}^{l} \lambda_i = 1\} \subset \mathbb{R}^2$.

DEFINITION 3.4. *The Newton polygon (see* [29], [8], [28], [6]*) of the rectangular SBR-surface* $\mathrm{SBR}[\theta]$*, denoted by Newton* $(\mathrm{SBR}[\theta])$*, is defined as the convex hull of $\mathcal{A}$. We denote by $\{\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_q\}$ the set of extreme points of* $\mathrm{Conv}(\mathcal{A})$ *$(q \leq l)$. These points are included in the set $\{Q_0, Q_1, \ldots, Q_l\}$ of the extreme points of $\mathcal{A}$.*

We denote by (α_i, β_i) the integer coordinates of the point $\mathcal{P}_i$ for $i = 0, \ldots, q$. Let us assume that $\mathcal{P}_0 = (\alpha_0 \neq 0, \beta_0 = 0)$ is the point such that α_0 is the lowest value with $y = 0$ and $\mathcal{P}_m = (\alpha_m = 0, \beta_m \neq 0)$ is the point such that β_m is the lowest value with $x = 0$. This implies that $\frac{\beta_{i+1} - \beta_i}{\alpha_i - \alpha_{i+1}} > \frac{\beta_i - \beta_{i-1}}{\alpha_{i-1} - \alpha_i}$ for $i = 1, \ldots, m-1$, where $\alpha_i > \alpha_{i+1}$ and $\beta_{i+1} > \beta_i$ for $i = 0, \ldots, m-1$. We need to define the complement of the Newton polygon with respect to the positive orthant in $\mathbb{R}^2$ which contains the point $(0, 0)$, denoted $\overline{\mathrm{Newton}}^{00}(\mathrm{SBR}[\theta])$.

Notice that each edge $[\mathcal{P}_k, \mathcal{P}_{k+1}]$, for $k = 0, \ldots, m-1$, of $\mathrm{Conv}(\mathcal{A})$ has a normal vector $\vec{n}_k$ such that it is inwardly oriented and it is taken as the shortest vector in this direction with integer coordinates. Actually, the coordinates of $\vec{n}_k$ are $(\frac{\beta_k - \beta_{k+1}}{d_k}, \frac{\alpha_{k+1} - \alpha_k}{d_k})$, where d_k is the greatest common divisor of the integers $\beta_{k+1} - \beta_k$

and $\alpha_k - \alpha_{k+1}$. It is observed that $d_k + 1$ is the number of integer points which belong to the segment $[\mathcal{P}_k, \mathcal{P}_{k+1}]$.

Assumption 2. *We assume that the mass of* $\theta_{ij} \geq 0$ *for all* $(i,j) \in \mathcal{A}$ *except for the* $(i,j) \in \{\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_q\}$, *where the mass of* $\theta_{ij} > 0$.

We come to the fundamental theorem.

THEOREM 3.5. *With Assumptions* 1 *and* 2 *we obtain that the image of the base point* $(u,v) = (0,0)$ *of the rational surface*

$$\begin{aligned} &\mathrm{SBR}\,[\theta] : [0,1]^2 / D \to \tilde{\mathcal{E}} \\ &(u,v) \mapsto \mathrm{SBR}\,[\theta]\,(u,v) = \Pi\left(\sum_{i=0}^{n}\sum_{j=0}^{p} B_i^n\,(u)\,B_j^p\,(v) * \theta_{ij}\right) \end{aligned} \tag{3.3}$$

is m *consecutive rational curves (*D *being the points of* $[0,1]^2$ *where* $\mathrm{SBR}\,[\theta]$ *is not defined). They are respectively defined by the massic vectors* θ_{ij} *whose indices* (i,j) *respectively belong to the* m *segments* $[\mathcal{P}_k, \mathcal{P}_{k+1}]$ *for* $k = 0, \ldots, m-1$. *More precisely, these* m *BR-curves are defined by* $\mathrm{BR}\left[\tau^{(k+1)}\right](t) = \Pi(\sum_{i=0}^{d_k} B_i^{d_k}\,(t) * \tau_i^{(k+1)})$, *where the massic vectors* $\tau_i^{(k+1)}$ *of these rational curves are defined via the massic vectors of* $\mathrm{SBR}\,[\theta]$ *as follows:*

$$\tau_i^{(k+1)} = \frac{\dbinom{n}{\alpha_k - i\frac{\alpha_k - \alpha_{k+1}}{d_k}}\dbinom{p}{\beta_k + i\frac{\beta_{k+1} - \beta_k}{d_k}}}{\dbinom{d_k}{i}} * \theta_{\alpha_k - i\frac{\alpha_k - \alpha_{k+1}}{d_k},\, \beta_k + i\frac{\beta_{k+1} - \beta_k}{d_k}}. \tag{3.4}$$

Proof. We use fractional changes of variables to obtain any image curve directly without having to apply successive changes of variables, as is usually done in algebraic geometry [26]. For each value of k in $\{0, \ldots, m-1\}$ we apply the change of variables

$$\varphi_k\,(r,t) = \left(u = r^{(\beta_{k+1} - \beta_k)}\,(1-t)^{\frac{d_k}{\alpha_k - \alpha_{k+1}}}, v = r^{(\alpha_k - \alpha_{k+1})} t^{\frac{d_k}{\beta_{k+1} - \beta_k}}\right)$$

to the function $\mathrm{SBR}\,[\theta]$. Therefore we have

$$\begin{aligned} &\mathrm{SBR}\,[\theta]\,(\varphi_k\,(r,t)) \\ &= \Pi\left(\sum_{(i,j)\in\mathcal{A}} \tbinom{n}{i}\tbinom{p}{j}\left(1 - r^{(\beta_{k+1} - \beta_k)}\,(1-t)^{\frac{d_k}{\alpha_k - \alpha_{k+1}}}\right)^{n-i} \times \left(1 - r^{(\alpha_k - \alpha_{k+1})} t^{\frac{d_k}{\beta_{k+1} - \beta_k}}\right)^{p-j}\right. \\ &\quad \left. \times\, r^{i(\beta_{k+1} - \beta_k) + j(\alpha_k - \alpha_{k+1})}\,(1-t)^{\frac{i d_k}{\alpha_k - \alpha_{k+1}}}\, t^{\frac{j d_k}{\beta_{k+1} - \beta_k}} * \theta_{ij}\right) \text{ for all } (r,t) \in [0,1]^2. \end{aligned}$$

Pairs (i,j) of integers belonging to $[\mathcal{P}_k, \mathcal{P}_{k+1}]$ satisfy $i\,(\beta_{k+1} - \beta_k) + j\,(\alpha_k - \alpha_{k+1}) = \alpha_k\beta_{k+1} - \alpha_{k+1}\beta_k$. Consequently for all $(i,j) \in \mathcal{A}$ we have $i\,(\beta_{k+1} - \beta_k) + j\,(\alpha_k - \alpha_{k+1}) \geq \alpha_k\beta_{k+1} - \alpha_{k+1}\beta_k$. Therefore in the previous sum we factorize with $r^{(\alpha_k\beta_{k+1} - \alpha_{k+1}\beta_k)}$, which is eliminated by a property of the Π-projection (see section 2). Finally, with the assumption on the masses it is possible to set $r = 0$. It gives

$$\mathrm{SBR}\,[\theta]\,(\varphi_k\,(0,t)) = \Pi\left(\sum_{(i,j)\in[\mathcal{P}_k,\mathcal{P}_{k+1}]\cap\mathcal{A}} \binom{n}{i}\binom{p}{j}\,(1-t)^{\frac{i d_k}{\alpha_k - \alpha_{k+1}}}\, t^{\frac{j d_k}{\beta_{k+1} - \beta_k}} * \theta_{ij}\right).$$

Actually, pairs (i,j) of integers belonging to $[\mathcal{P}_k,\mathcal{P}_{k+1}]$ are given by $i=\alpha_k-l\frac{\alpha_k-\alpha_{k+1}}{d_k}$ and $j=\beta_k+l\frac{\beta_{k+1}-\beta_k}{d_k}$ for $l=0,\ldots,d_k$. Again via the Π-projection we simplify with $(1-t)^{\frac{\alpha_{k+1}d_k}{\alpha_k-\alpha_{k+1}}}\,t^{\frac{\beta_k d_k}{\beta_{k+1}-\beta_k}}$ in the previous relation.

We have

$$\begin{aligned}&\mathrm{SBR}\,[\theta]\,(\varphi_k\,(0,t))\\&=\Pi\left(\sum_{l=0}^{d_k}\binom{n}{\alpha_k-l\frac{\alpha_k-\alpha_{k+1}}{d_k}}\binom{p}{\beta_k+l\frac{\beta_{k+1}-\beta_k}{d_k}}(1-t)^{d_k-l}\,t^l\right.\\&\qquad\left.*\,\theta_{\alpha_k-l\frac{\alpha_k-\alpha_{k+1}}{d_k},\beta_k+l\frac{\beta_{k+1}-\beta_k}{d_k}}\right).\end{aligned}$$

Therefore we obtain for $k=0,\ldots,m-1$ the following BR-curves:

$$\mathrm{SBR}\,[\theta]\,(\varphi_k\,(0,t))=\Pi\left(\sum_{i=0}^{d_k}B_i^{d_k}\,(t)*\tau_i^{(k+1)}\right),\ t\in[0,1]\,,$$

where

$$\tau_i^{(k+1)}=\frac{\binom{n}{\alpha_k-i\frac{\alpha_k-\alpha_{k+1}}{d_k}}\binom{p}{\beta_k+i\frac{\beta_{k+1}-\beta_k}{d_k}}}{\binom{d_k}{i}}*\theta_{\alpha_k-i\frac{\alpha_k-\alpha_{k+1}}{d_k},\beta_k+i\frac{\beta_{k+1}-\beta_k}{d_k}}.$$

As we have $\tau_{d_{k-1}}^{(k)}=\tau_0^{(k+1)}$ for $k=1,\ldots,m-1$, it comes that the BR-curves images of the base point are consecutive. Hence, the image of the base point $(u,v)=(0,0)$ is a set made of these m consecutive rational curves whose massic vectors are defined by (3.4). □

Since Clebsch [5], it has been known that the image of a base point can be blown up into a set of rational curves. The main result of this theorem is that via relation (3.4) we explicitly give the link between the massic vectors defining this set of rational curves and those of indices belonging to $\mathrm{Newton}\,(\mathrm{SBR}\,[\theta])\cap\overline{\mathrm{Newton}}^{00}\,(\mathrm{SBR}\,[\theta])$. Moreover, we point out that these BR-curves are placed end to end, which is very important for applications in geometric design.

In Figure 1, the image of the base point $(u,v)=(0,0)$ is represented by two consecutive rational curves respectively defined up to multiplicative constants by the massic vectors: $\theta_{8,0}$, $\theta_{3,2}$; and $\theta_{3,2}$, $\theta_{2,4}$, $\theta_{1,6}$, $\theta_{0,8}$.

Remark. The assumption that $\chi\,(\theta_{ij})\geq 0$ for all $(i,j)\in\mathcal{A}$ except for $(i,j)\in\{\mathcal{P}_0,\mathcal{P}_1,\ldots,\mathcal{P}_q\}$ where $\chi\,(\theta_{ij})>0$ prevents the SBR-surface from having base points elsewhere than at the vertices of each definition domains obtained after applying changes of variables to the surface. We could have admitted negative and positive masses but the image curves of the base point $(u,v)=(0,0)$ of $\mathrm{SBR}\,[\theta]$ are not necessarily given by the massic vectors whose indices belong to $\mathrm{Newton}\,(\mathrm{SBR}\,[\theta])\cap\overline{\mathrm{Newton}}^{00}\,(\mathrm{SBR}\,[\theta])$. Moreover, they are not necessarily end to end and their number can be greater than the number of edges of $\mathrm{Newton}\,(\mathrm{SBR}\,[\theta])\cap\overline{\mathrm{Newton}}^{00}\,(\mathrm{SBR}\,[\theta])$. For instance, in Figure 2 the image of the base point $(0,0)$ is two orthogonal segments. In a future work, we shall tackle this difficult problem.

If we want the image of the base point $(u,v)=(0,0)$ to be a unique rational curve or a point then we have the following.

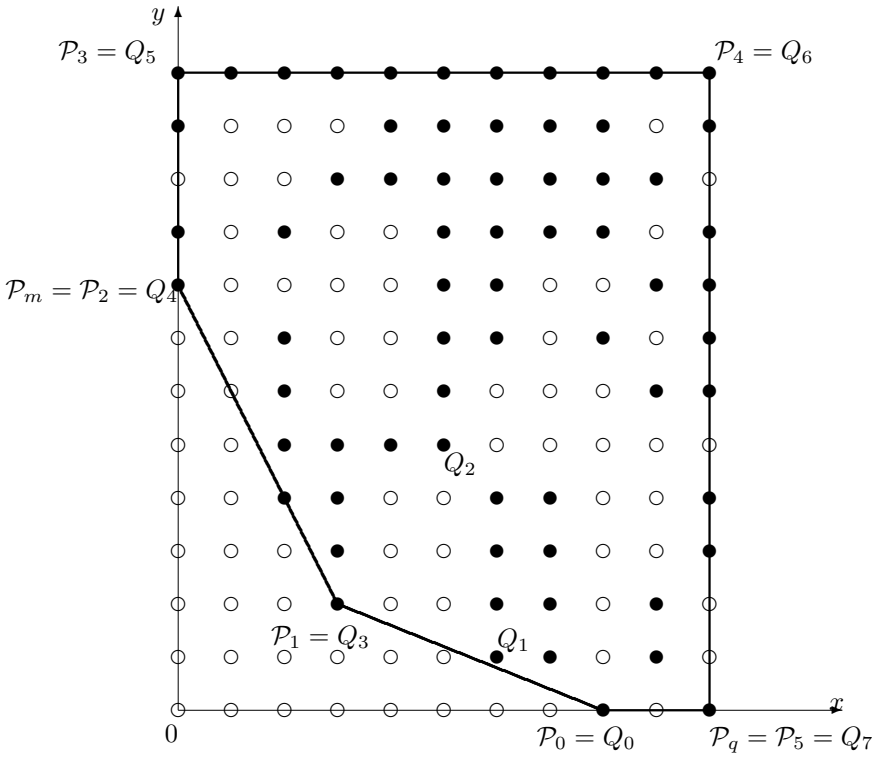


FIG. 1. *An example of a convex hull of a set of integer points associated with the nonzero massic vectors of an SBR-surface.* • *(respectively,* ○*) denotes integer points* (i,j) *such that* $\theta_{ij} \neq \vec{0}$ *(respectively,* $\theta_{ij} = \vec{0}$*);* Q_i *= extreme points of* $\mathcal{A}$*;* $\mathcal{P}_i$ *= extreme points of* $\mathrm{Conv}(\mathcal{A})$.

COROLLARY 3.6. *Let* S *be a SBR-surface defined by* (3.3) *where* $\chi(\theta_{ij}) \geq 0$ *for all* $(i,j) \in \mathcal{A}$ *except for* $(i,j) \in \{\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_q\}$ *where* $\chi(\theta_{ij}) > 0$*. If* $\mathcal{P}_0 = (\alpha_0, 0)$ *and* $\mathcal{P}_1 = (0, \beta_1)$ $(\alpha_0 < n$ *and* $\beta_1 < p)$ *such that* $\theta_{ij} = \vec{0}$ *for all* $(i,j) \in \mathbb{R}^2$ *with* $i\beta_1 + j\alpha_0 < \alpha_0\beta_1$*, then the image of the base point* $(u,v) = (0,0)$ *is the BR-curve defined by*

$$\Pi\left(\sum_{i=0}^{d_0=\gcd(\alpha_0,\beta_1)} B_i^{d_0}(t)\, \frac{\binom{n}{\alpha_0 - i\frac{\alpha_0}{d_0}}\binom{p}{i\frac{\beta_1}{d_0}}}{\binom{d_0}{i}} * \theta_{\alpha_0 - i\frac{\alpha_0}{d_0}, i\frac{\beta_1}{d_0}}\right) \text{ for all } t \in [0,1]. \tag{3.5}$$

Furthermore, if $\Pi(\theta_{\alpha_0,0}) = \Pi(\theta_{0,\beta_1})$ $\left(\text{where } \Pi(\theta_{0,\beta_1}) = P \in \mathbb{R}^3\right)$ *and either* $\Pi(\theta_{\alpha_0 - i\frac{\alpha_0}{d_0}, i\frac{\beta_1}{d_0}}) = P$ *or* $\theta_{\alpha_0 - i\frac{\alpha_0}{d_0}, i\frac{\beta_1}{d_0}} = \vec{0}$ *for* $i = 1, \ldots, d_0 - 1$*, then* $\mathrm{SBR}[\theta]$ *is continuous at* $(0,0)$.

Proof. Relation (3.5) follows from the application of the previous theorem with $\mathcal{P}_0 = (\alpha_0, 0)$ and $\mathcal{P}_1 = (0, \beta_1)$. The assumption on the masses compels us to demonstrate the continuity of $\mathrm{SBR}[\theta]$ at $(0,0)$. Let us define the sets $I = \{(i,j) : \theta_{ij} =$

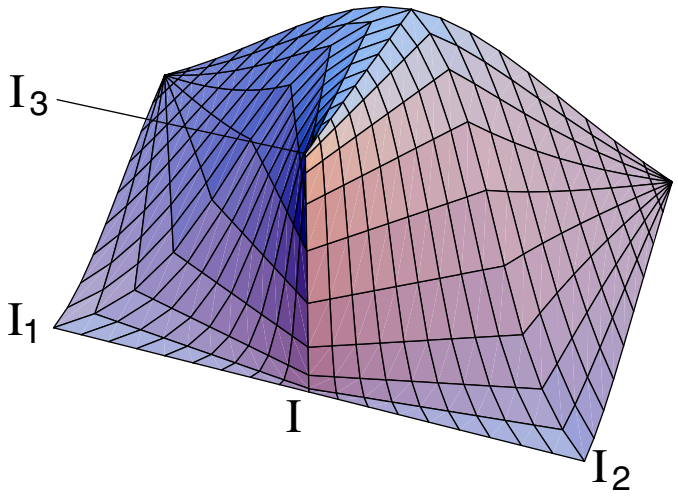


FIG. 2. *The image of the base point is the two orthogonal segments* $[I_1, I_2]$ *and* $[I, I_3]$ *which intersect at the point* I.

$(P_{ij}, \beta_{ij}) \in \mathcal{E} \times \mathbb{R}^*\}$ and $\bar{I} = \{(i,j) : \theta_{ij} = \vec{U}_{ij} \in \vec{\mathcal{E}}\}$; then we obtain via the explicit form of an SBR-surface (see Proposition 2.5(a)) that for all $(u,v) \in [0,1]^2$,

$$\mathrm{SBR}\,[\theta]\,(u,v) - P = \frac{\sum_{(i,j)\in I} \beta_{ij} B_i^n(u) B_j^p(v).(P_{ij}-P) + \sum_{(i,j)\in \bar{I}} B_i^n(u) B_j^p(v).\vec{U}_{ij}}{\sum_{(i,j)\in I} \beta_{ij} B_i^n(u) B_j^p(v)}.$$

When applying the change of variables $u = r^{\beta_1}(1-t)^{\frac{d_0}{\alpha_0}}$, $v = r^{\alpha_0} t^{\frac{d_0}{\beta_1}}$ (with $r \geq 0$ and $t \in [0,1]$) to the right side of the previous expression and symplifying by $r^{\alpha_0\beta_1}$ it follows that

$$\begin{aligned}
&\mathrm{SBR}\,[\theta]\,(u,v) - P \\
&= \frac{\sum_{(i,j)\in I} \binom{n}{i}\binom{p}{j} r^{i\beta_1 + j\alpha_0 - \alpha_0\beta_1} \beta_{ij} . \left(1 - r^{\beta_1}(1-t)^{\frac{d_0}{\alpha_0}}\right)^{n-i} \left(1 - r^{\alpha_0} t^{\frac{d_0}{\beta_1}}\right)^{p-j} (1-t)^{i\frac{d_0}{\alpha_0}} t^{j\frac{d_0}{\beta_1}} (P_{ij}-P)}{\sum_{(i,j)\in I} \binom{n}{i}\binom{p}{j} r^{i\beta_1 + j\alpha_0 - \alpha_0\beta_1} \beta_{ij} . \left(1 - r^{\beta_1}(1-t)^{\frac{d_0}{\alpha_0}}\right)^{n-i} \left(1 - r^{\alpha_0} t^{\frac{d_0}{\beta_1}}\right)^{p-j} (1-t)^{i\frac{d_0}{\alpha_0}} t^{j\frac{d_0}{\beta_1}}} \\
&\quad + \frac{\sum_{\substack{(i,j)\in \bar{I} \\ i+j>m}} \binom{n}{i}\binom{p}{j} r^{i\beta_1 + j\alpha_0 - \alpha_0\beta_1} \left(1 - r^{\beta_1}(1-t)^{\frac{d_0}{\alpha_0}}\right)^{n-i} \left(1 - r^{\alpha_0} t^{\frac{d_0}{\beta_1}}\right)^{p-j} (1-t)^{i\frac{d_0}{\alpha_0}} t^{j\frac{d_0}{\beta_1}} . \vec{U}_{ij}}{\sum_{(i,j)\in I} \binom{n}{i}\binom{p}{j} r^{i\beta_1 + j\alpha_0 - \alpha_0\beta_1} \beta_{ij} . \left(1 - r^{\beta_1}(1-t)^{\frac{d_0}{\alpha_0}}\right)^{n-i} \left(1 - r^{\alpha_0} t^{\frac{d_0}{\beta_1}}\right)^{p-j} (1-t)^{i\frac{d_0}{\alpha_0}} t^{j\frac{d_0}{\beta_1}}}.
\end{aligned} \tag{3.6}$$

The assumption on the masses of the massic vectors implies that the denominator is strictly greater than zero for all $r \geq 0$ and $t \in [0,1]$. Consequently as $P_{i,j} = P$ for all $(i,j) \in I$ such that $i\beta_1 + j\alpha_0 = \alpha_0\beta_1$ we can factorize the numerator of (3.6) by r and we obtain

$$\|\mathrm{SBR}\,[\theta]\,(u,v) - P\| = \mathcal{O}(r).$$

θ_{05}	θ_{15}	θ_{25}	θ_{35}	θ_{45}
θ_{04}	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{44}
θ_{03}	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{43}
$\vec{0}$	θ_{12}	$\vec{0}$	$\vec{0}$	θ_{42}
$\vec{0}$	$\vec{0}$	θ_{21}	$\vec{0}$	θ_{41}
$\vec{0}$	$\vec{0}$	θ_{20}	θ_{30}	θ_{40}

FIG. 3. *An SBR-surface with a corner of null massic vectors.*

Hence,

$$\lim_{\substack{(u,v)\to(0,0)\\(u,v)\in[0,1]^2}} \mathrm{SBR}\left[\theta\right](u,v) = P$$

and the continuity of $\mathrm{SBR}\left[\theta\right]$ at $(0,0)$ holds. □

Figure 3 illustrates the case of an SBR-surface having a base point at $(u,v)=(0,0)$ whose image is a unique rational boundary curve of S. This image is the segment defined by the massic vectors θ_{03} and θ_{20}.

Remark. If

$$\lim_{\substack{(u,v)\to(0,0)\\(u,v)\in[0,1]^2}} \mathrm{SBR}\left[\theta\right](u,v) = P$$

we say that the image of the base point $(u,v)=(0,0)$ blows down into a point.

4. Algorithm for drawing a rectangular SBR-surface with base points. To draw a rectangular SBR-surface S which is the image of $[0,1]^2$, where S has a base point at $(u,v)=(0,0)$ whose image is m consecutive rational curves, we may apply a finite number of quadratic changes of variables of type $u=r(1-t), v=rt$ to the function $\mathrm{SBR}\left[\theta\right]$ (see [29]). Another strategy is to apply the same kind of change of variables as in the proof of Theorem 3.5. In each case, we clearly show the m additional boundary curves of S. Assuming that we apply the first strategy, the drawing of S will be a set of SBR-surfaces without any base point. We shall not apply the second strategy. Indeed, in that case, the regular SBR-surfaces we obtain by this method overlap, which is unsatisfactory for applications. The following proposition will help us to highlight the image curves of a base point.

PROPOSITION 4.1. *Let S be an SBR-surface defined by* (3.3). *Hence, by applying the change of variables $u=r(1-t), v=rt$ to function* $\mathrm{SBR}\left[\theta\right]$ *we obtain*

$$\mathrm{SBR}\left[\theta\right](u(r,t),v(r,t)) = \Pi\left(\sum_{k=0}^{n+pn+p}\sum_{l=0} B_k^{n+p}(r)\, B_l^{n+p}(t) * \omega_{kl}\right),$$

where for all $(k,l)\in\{0,\ldots,n+p\}^2$

$$\omega_{kl} = \sum_{i=0}^{k}\sum_{j=\max(0,i-n)}^{\min(i,p)} \frac{\binom{k}{i}\binom{n}{i-j}\binom{p}{j}\binom{n+p-i}{l-j}}{\binom{n+p}{i}\binom{n+p}{l}} * \Delta^{i-j,j}\theta_{00} \ \textit{if } n\le p,$$

$$\omega_{kl} = \sum_{i=0}^{k}\sum_{j=\max(0,i-p)}^{\min(i,n)} \frac{\binom{k}{i}\binom{n}{j}\binom{p}{i-j}\binom{n+p-i}{l-j}}{\binom{n+p}{i}\binom{n+p}{l}} * \Delta^{j,i-j}\theta_{00} \ \textit{else.}$$

Proof. We can write $\mathrm{SBR}\,[\theta]$ in the following monomial form:

$$\mathrm{SBR}\,[\theta]\,(u,v)=\Pi\left(\sum_{i=0}^{n}\sum_{j=0}^{p}\binom{n}{i}\binom{p}{j}u^{i}v^{j}*\Delta^{ij}\theta_{00}\right).$$

With the assumption that $n\leq p$, letting $u=r\,(1-t)\,,v=rt$, and $k=i+j$, we obtain

$$\mathrm{SBR}\,[\theta]\,(u\,(r,t)\,,v\,(r,t))=\Pi\left(\sum_{k=0}^{n+p}\sum_{j=\max(0,k-n)}^{\min(k,p)}r^{k}B_{j}^{k}\,(t)\,\frac{\binom{n}{k-j}\binom{p}{j}}{\binom{k}{j}}*\Delta^{k-j,j}\theta_{00}\right).$$

To provide the expression of $\mathrm{SBR}\,[\theta]$ in the Bernstein polynomial bases of degree $n+p$ in t, one can apply $n+p-k$ degree of elevations in t (see [9]), which gives

$$\begin{aligned}&\mathrm{SBR}\,[\theta]\,(u\,(r,t)\,,v\,(r,t))\\&\quad=\Pi\left(\sum_{k=0}^{n+p}\sum_{l=0}^{n+p}r^{k}B_{l}^{n+p}\,(t)\sum_{j=\max(0,k-n)}^{\min(k,p)}\frac{\binom{n}{k-j}\binom{p}{j}\binom{n+p-k}{l-j}}{\binom{n+p}{l}}*\Delta^{k-j,j}\theta_{00}\right).\end{aligned}$$

We are now able to write $\mathrm{SBR}\,[\theta]\,(u\,(r,t)\,,v\,(r,t))$ in the Bernstein polynomial bases of degree $n+p$ in r as follows:

$$\mathrm{SBR}\,[\theta]\,(u\,(r,t)\,,v\,(r,t))=\Pi\left(\sum_{k=0}^{n+p}\sum_{l=0}^{n+p}B_{k}^{n+p}\,(r)\,B_{l}^{n+p}\,(t)\,\omega_{kl}\right),$$

where $\Delta^{k0}\omega_{0l}=\sum_{j=\max(0,k-n)}^{\min(k,p)}\frac{\binom{n}{k-j}\binom{p}{j}\binom{n+p-k}{l-j}}{\binom{n+p}{k}\binom{n+p}{l}}*\Delta^{k-j,j}\theta_{00}$. Moreover, as we have $\omega_{kl}=\sum_{i=0}^{k}\binom{k}{i}*\Delta^{i0}\omega_{0l}$, then for all $(k,l)\in\{0,\ldots,n+p\}^{2}$,

$$\omega_{kl}=\sum_{i=0}^{k}\sum_{j=\max(0,i-n)}^{\min(i,p)}\frac{\binom{k}{i}\binom{n}{i-j}\binom{p}{j}\binom{n+p-i}{l-j}}{\binom{n+p}{i}\binom{n+p}{l}}*\Delta^{i-j,j}\theta_{00}.$$

The case $n>p$ is obtained by replacing n by p and conversely replacing $\Delta^{i-j,j}$ by $\Delta^{j,i-j}$. ∎

Remark. We can use this proposition for any base point at the corners of the definition domain. It suffices to apply an affine change of variables to the function (3.3).

The application of a quadratic change of variables to a rational surface with a base point at $(0,0)$ provides a new SBR-surface with certain complete lines of zero massic vectors which come from the diagonal lines of null massic vectors of the initial rectangular SBR-surface. Therefore, the aim of Proposition 4.2 is to eliminate these complete lines of null massic vectors.

PROPOSITION 4.2. *Let S be an* SBR*-surface defined by* (3.3)*, where for any given integers k_0,k_1,l_0,l_1 such that $n-k_0-k_1\geq 0$ and $p-l_0-l_1\geq 0$, we have $\theta_{ij}=\vec{0}$ for all $(i,j)\in\{0,\ldots,k_0-1\}\times\{0,\ldots,l_0-1\}$ and for all $(i,j)\in\{n-k_1+1,\ldots,n\}\times\{p-l_1+1,\ldots,p\}$. Then we have*

$$\mathrm{SBR}\,[\theta]\,(u,v)=\Pi\left(\sum_{i=0}^{n-k_0-k_1}\sum_{j=0}^{p-l_0-l_1}B_{i}^{n-k_0-k_1}\,(u)\,B_{j}^{p-l_0-l_1}\,(v)*\omega_{ij}\right),$$

where $\omega_{ij} = \frac{\binom{n}{i+k_0}\binom{p}{j+l_0}}{\binom{n-k_0-k_1}{i}\binom{p-l_0-l_1}{j}} * \theta_{i+k_0,j+l_0}$ *for all* $(i,j) \in \{0,\ldots,n-k_0-k_1\} \times \{0,\ldots,p-l_0-l_1\}$.

Proof. With these assumptions we have

$$\begin{aligned}&\mathrm{SBR}\,[\theta]\,(u,v)\\ &= \Pi\left(\sum_{j=0}^{p} B_j^p\,(v)\left(\binom{n}{k_0}(1-u)^{n-k_0}\,u^{k_0} * \theta_{k_0,j} \oplus \cdots \oplus (1-u)^{k_1}\,u^{n-k_1} * \theta_{n-k_1,j}\right)\right).\end{aligned}$$

Consequently, we can factorize by $u^{k_0}\,(1-u)^{k_1}$ and write $\mathrm{SBR}\,[\theta]$ in the following form:

$$\mathrm{SBR}\,[\theta]\,(u,v) = \Pi\left(\sum_{i=0}^{n-k_0-k_1}\sum_{j=0}^{p} B_i^{n-k_0-k_1}\,(u)\,B_j^p\,(v)\,\frac{\binom{n}{i+k_0}}{\binom{n-k_0-k_1}{i}} * \theta_{i+k_0,j}\right).$$

Again we factorize by $v^{l_0}\,(1-v)^{l_1}$ and we obtain the result. □

To draw on a computer a rectangular SBR-surface S having a base point at $(0,0)$ which satisfies the assumptions of Theorem 3.5, we need to remove the singularity. This can be done via the use of quadratic changes of variables and a subdivision process. Hence, the initial singular rectangular SBR-surface will be drawn by a net of regular SBR-surfaces defined by massic vectors. To do so, we define the following algorithm:

Step 0: Initialize the set C to $\{S\}$

Step 1: Do while there is a null massic vector either in the top left corner or in the bottom left corner of one rectangular SBR-surface of C

If there is a null massic vector at the two left corners of a surface of C then

- Subdivide the set of the massic vectors of this surface into four sets of massic vectors via the use of the algorithm of de Casteljau for SBR-surfaces [10] evaluated with the following parameters: $\left(\frac{1}{2},\frac{1}{2}\right)$.
- Remove the surface from C
- Add the two right regular surfaces to C
- Apply four quadratic changes of variables (see Propositions 4.1 and 4.2) to the two left surfaces:
  - For the bottom left SBR-surface, one at $(0,0)$ and one at $(1,1)$ (see Figure 5)
  - For the top left SBR-surface, one at $(0,1)$ and one at $(1,0)$
- Add these four new surfaces to C

End if

If there is only a null massic vector at the top left corner of a surface of C, then

- Remove the surface from C
- Apply two quadratic changes of variables to this surface, one at $(0,1)$ and one at $(1,0)$
- Add the two resulting sets of massic vectors to C

End if

If there is only a null massic vector at the bottom left corner of a surface of C, then

- Remove the surface from C
- Apply two quadratic changes of variables to this surface, one at $(0,0)$ and one at $(1,1)$
- Add the two resulting sets of massic vectors to C

End if

Step 2: Draw all the regular SBR-surfaces to the set C.

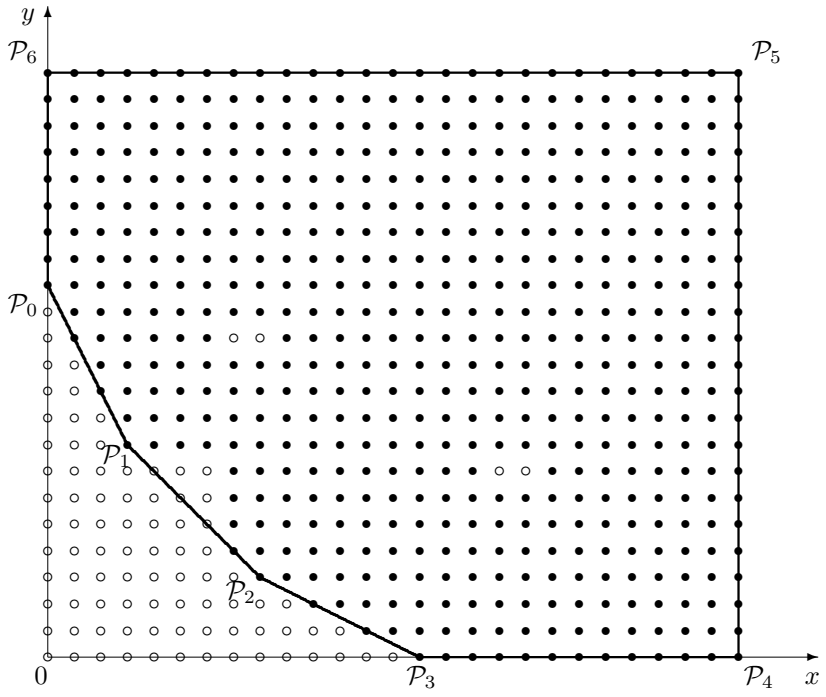


FIG. 4. *Example of a rectangular SBR-surface with a base point at* $(0,0)$. $\circ$ *denotes points associated with null massic vectors;* $\bullet$ *denotes points associated with non-zero massic vectors.*

For instance, in the case of Figure 4, we first apply the change of variables $u = r(1-t), v = rt$ to $\mathrm{SBR}[\theta]$ (see Figure 5). Thus, we divide the domain $[0,1]^2$ into two triangles. The upper triangle defines the domain of the surface SBR-surface called S_1 which does not have any base points at corners. The image of the lower one is, in variables (r,t), the square $[0,1]^2$, where the image of $(u,v) = (0,0)$ is $(0,t)$ for all $t \in [0,1]$. As the SBR-surface image of this domain has two base points (one at $t=0$ and the other one at $t=1$), we subdivide it into four SBR-patches by applying the classical algorithm of de Casteljau [7], [11], [10] at $(r,t) = \left(\frac{1}{2}, \frac{1}{2}\right)$ so that two of them (S_4 and S_5) are free from base points at corners. However, the two other do have a base point. By applying again a quadratic change of variables, $r = r_1(1-t_1)$, $t = r_1 t_1$ and $r = r_2(1-t_2)$, $t = 1 - r_2 t_2$, we obtain two additional SBR-patches S_6 and S_7, which do not present any base points on $[0,1]^2$. This method is illustrated by Figure 5.

5. Example of a seven-sided SBR-surface. We are now able to construct an n-sided SBR-surface which interpolates n given consecutive rational curves via the introduction of base points at the vertices of the definition domain of a rectangular SBR-surface. This can be done via relation (3.4), which gives the relations between the massic vectors whose indices belong to $\mathrm{Newton}(\mathrm{SBR}[\theta]) \cap \overline{\mathrm{Newton}}^{00}(\mathrm{SBR}[\theta])$ of the initial SBR-surface, and the massic vectors of the given BR-curves. Naturally, the massic vectors whose indices are inside points of the Newton polygon can be chosen arbitrarily with positive or null masses. This brings new possibilities to control the form of surfaces. The framework thus considered is more significant than that

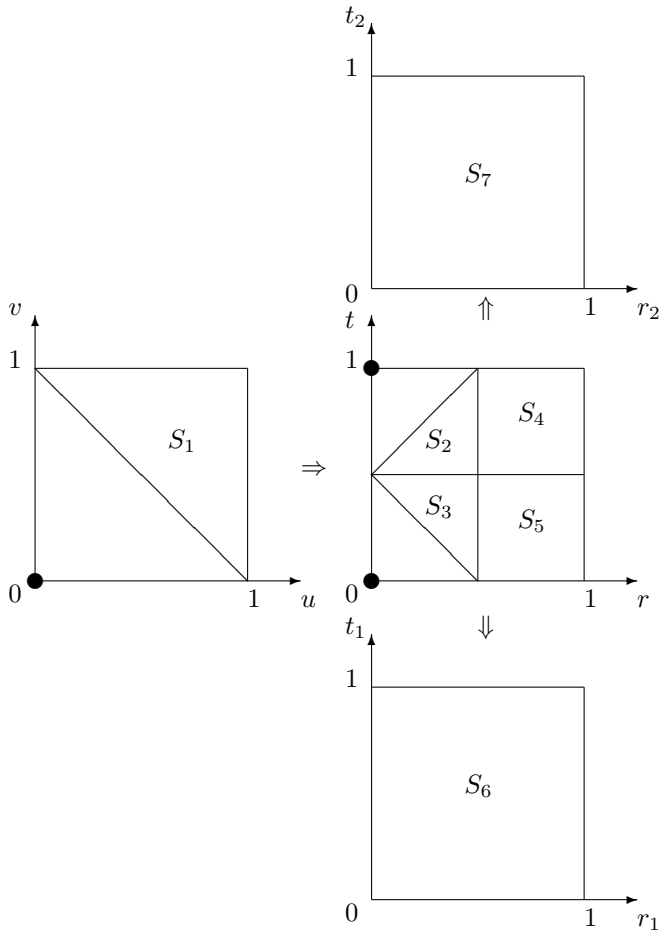


FIG. 5. *The diagram of the consecutive change of variables.*

considered in [30] and [31]. We shall now study an example with seven BR-curves.

Let $(\Omega, \vec{i}, \vec{j}, \vec{k})$ be a Cartesian frame. Let us consider the seven following BR-curves which create a hole (see Figure 6):

$$\Gamma_1(t) = BR[\tau_0^1 = (A_1; 1), \tau_1^1 = 2\vec{k}, \tau_2^1 = (A_2; 1)](t),$$

$$\Gamma_2(t) = BR[\tau_0^2 = (A_2; 1), \tau_1^2 = -2\vec{k}, \tau_2^2 = (A_3; 1)](t),$$

$$\Gamma_3(t) = BR[\tau_0^3 = (A_3; 1), \tau_1^3 = (A_4; 1)](t),$$

$$\Gamma_4(t) = BR\left[\tau_0^4 = (A_4; 1), \tau_1^4 = \left(\begin{pmatrix} 12.5 \\ 0 \\ -4 \end{pmatrix}; 2\right), \tau_2^4 = (A_5; 1)\right](t),$$

$$\Gamma_5(t) = BR\left[\tau_0^5 = (A_5; 1), \tau_1^5 = (A_6; 1)\right](t),$$

$$\Gamma_6(t) = BR\left[\tau_0^5 = (A_6; 1), \tau_1^5 = (A_7; 1)\right](t),$$

$$\Gamma_7(t) = BR[\tau_0^4 = (A_7; 1), \tau_1^4 = -2\vec{k}, \tau_2^4 = (A_1; 1)](t),$$

where

$$A_1 = \begin{pmatrix} -5 \\ 10 \\ 0 \end{pmatrix},\ A_2 = \begin{pmatrix} -3 \\ 5 \\ 0 \end{pmatrix},\ A_3 = \begin{pmatrix} 0 \\ 3 \\ 1 \end{pmatrix},\ A_4 = \begin{pmatrix} 10 \\ 0 \\ 0 \end{pmatrix},$$

$$A_5 = \begin{pmatrix} 15 \\ 0 \\ 0 \end{pmatrix},\ A_6 = \begin{pmatrix} 15 \\ 15 \\ 1 \end{pmatrix},\ A_7 = \begin{pmatrix} -5 \\ 15 \\ 0 \end{pmatrix}.$$

As $\chi\left(\tau_1^1\right)^2 - \chi\left(\tau_0^1\right)\chi\left(\tau_2^1\right) < 0$, $\chi\left(\tau_1^2\right)^2 - \chi\left(\tau_0^2\right)\chi\left(\tau_2^2\right) < 0$, and $\chi\left(\tau_1^7\right)^2 - \chi\left(\tau_0^7\right)\chi\left(\tau_2^7\right) < 0$, we can conclude via Proposition 5.1.6 in [11] that Γ_1, Γ_2, and Γ_7 are arcs of ellipses. As $\chi\left(\tau_1^4\right)^2 - \chi\left(\tau_0^4\right)\chi\left(\tau_2^4\right) > 0$ via the same proposition, we conclude that Γ_4 is a hyperbolic arc.

Moreover, we can show that Γ_1, Γ_2 join C^1 at A_2 and Γ_1, Γ_7 join C^1 at A_1 via Proposition 4.1.6 in [11].

We can now construct the grid of massic vectors of a rectangular SBR-surface, denoted by S, which fills the previous hole. We have to satisfy some relations so that the given BR-curves are the boundary curves of this surface. We arbitrarily choose to define the massic vectors of S as follows.

The massic vectors θ_{i9}, $i = 0, \ldots, 8$ (respectively, θ_{8j}, $j = 0, \ldots, 9$), are obtained from seven (respectively, eight) elevations of the length of the BR-curve Γ_6 (respectively, Γ_5). This allows us to obtain a better parametrization of the surface.

We are now searching for the relations between the massic vectors of the rectangular SBR-surface and the massic vectors of the boundary BR-curves Γ_1, Γ_2, Γ_3 so that they are the image of the base point at $(0,0)$. With that aim in view, we use the relation (3.4) and we obtain

$$\text{for } \Gamma_1\colon \theta_{07} = \frac{\binom{2}{2}}{\binom{8}{0}\binom{9}{7}} * (A_1;1), \quad \theta_{15} = \frac{\binom{2}{1}}{\binom{8}{1}\binom{9}{5}} 2\vec{k}, \quad \theta_{23} = \frac{\binom{2}{0}}{\binom{8}{2}\binom{9}{3}} * (A_2;1)\,;$$

$$\text{for } \Gamma_2\colon \theta_{23} = \frac{\binom{2}{2}}{\binom{8}{2}\binom{9}{3}} * (A_2;1), \quad \theta_{32} = -\frac{\binom{2}{1}}{\binom{8}{3}\binom{9}{2}} 2\vec{k}, \quad \theta_{41} = \frac{\binom{2}{0}}{\binom{8}{4}\binom{9}{1}} * (A_3;1)\,;$$

$$\text{for } \Gamma_3\colon \theta_{41} = \frac{\binom{1}{1}}{\binom{8}{4}\binom{9}{1}} * (A_3;1), \quad \theta_{60} = \frac{\binom{1}{0}}{\binom{8}{6}\binom{9}{0}} * (A_4;1)\,.$$

Moreover, Γ_4 (respectively, Γ_7) has to be the image of $(u,0)$, $u \in [0,1]$ (respectively, of $(0,v)$, $v \in [0,1]$). So we have

$$\text{for } \Gamma_4\colon \theta_{60} = \frac{\binom{2}{2}}{\binom{8}{6}} * (A_4;1), \quad \theta_{70} = \frac{\binom{2}{1}}{\binom{8}{7}} * \left(\begin{pmatrix} 12.5 \\ 0 \\ -4 \end{pmatrix};2\right), \quad \theta_{80} = \frac{\binom{2}{0}}{\binom{8}{8}} * (A_5;1)\,;$$

$$\text{for } \Gamma_7\colon \theta_{07} = \frac{\binom{2}{2}}{\binom{9}{7}} * (A_1;1), \quad \theta_{08} = -\frac{\binom{2}{1}}{\binom{9}{8}} 2\vec{k}, \theta_{09} = \frac{\binom{2}{0}}{\binom{9}{9}} * (A_7;1)\,.$$

As these relations are compatible, we have Figure 8.

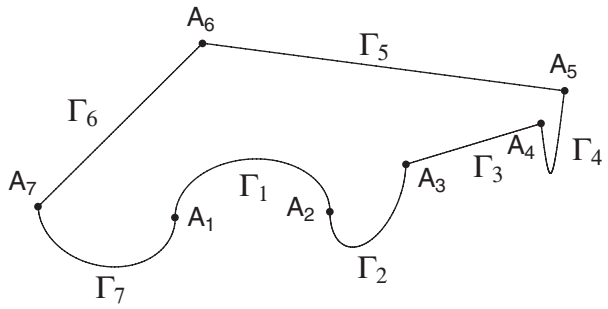


FIG. 6. *The seven given boundary BR-curves of the hole.*

θ_{09}	θ_{19}	θ_{29}	θ_{39}	θ_{49}	θ_{59}	θ_{69}	θ_{79}	θ_{89}
θ_{08}	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{88}
θ_{07}	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{87}
$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{86}
$\vec{0}$	θ_{15}	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{85}
$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{84}
$\vec{0}$	$\vec{0}$	θ_{23}	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{83}
$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{32}	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{82}
$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{41}	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{81}
$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	$\vec{0}$	θ_{60}	θ_{70}	θ_{80}

FIG. 7. *A grid of the massic vectors of S.*

To draw this seven-sided SBR-surface S we use the results of the previous section. It is composed of seven SBR-surfaces which join each other naturally with the G^∞-continuity. Notice that the C^1-continuity of Γ_7 and Γ_1 at A_1 (respectively, the C^1-continuity of Γ_1 and Γ_2 at A_2) is satisfied and the null massic vectors whose indices are in the interior of $\mathrm{Conv}(\mathcal{A})$ (see Figure 7) induce a tension effect on the surface.

This seven-sided SBR-surface could have been defined with a base point, for instance, at $(0,0)$, $(1,0)$, $(1,1)$, with the constraint that the image of each base point is a unique BR-curve.

For instance, in the case of a five-sided filling surface we use the grid of massic vectors defined in Figure 3. The image curve of the base point is given by the massic vectors,

$$\theta_{20} = \left(\begin{pmatrix} 5 \\ 0 \\ 0 \end{pmatrix} ; 1 \right), \theta_{03} = \left(\begin{pmatrix} 0 \\ 5 \\ 0 \end{pmatrix} ; 1 \right)$$

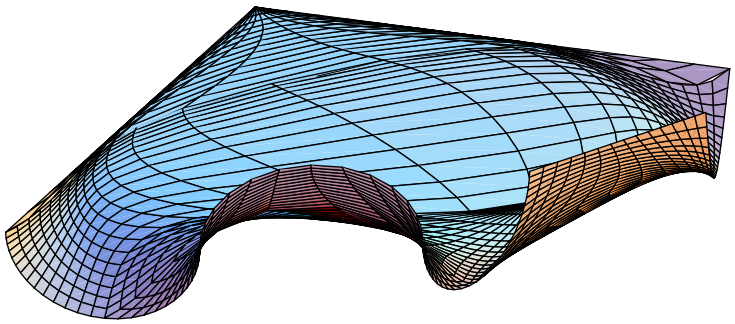


FIG. 8. *A seven sided SBR-surface defined from a rectangular one with a base point.*

(see Corollary 3.6), and the other boundaries are defined via the following four arcs of ellipses:

$$\theta_{20},\theta_{30}=\vec{k},\theta_{40}=\left(\begin{pmatrix}10\\0\\0\end{pmatrix};1\right);$$

$$\theta_{40},\ \theta_{41}=\left(\begin{pmatrix}10\\0\\\frac{2}{3}\end{pmatrix};\frac{3}{5}\right),\ \theta_{42}=\left(\begin{pmatrix}10\\\frac{5}{2}\\\\\frac{3}{2}\end{pmatrix};\frac{2}{5}\right),\ \theta_{43}=\left(\begin{pmatrix}10\\\frac{15}{2}\\\\\frac{3}{2}\end{pmatrix};\frac{3}{5}\right),$$

$$\theta_{44}=\left(\begin{pmatrix}10\\10\\\frac{2}{3}\end{pmatrix};\frac{3}{5}\right),\ \theta_{45}=\left(\begin{pmatrix}10\\10\\0\end{pmatrix};1\right);$$

$$\theta_{03},\theta_{04}=\vec{k},\theta_{05}=\left(\begin{pmatrix}0\\10\\0\end{pmatrix};1\right);$$

$$\theta_{05},\theta_{15}=\left(\begin{pmatrix}0\\10\\\frac{1}{2}\end{pmatrix};\frac{1}{2}\right),\ \theta_{25}=\left(\begin{pmatrix}5\\10\\1\end{pmatrix};\frac{1}{3}\right),\ \theta_{35}=\left(\begin{pmatrix}10\\10\\\frac{1}{2}\end{pmatrix};\frac{1}{2}\right),\ \theta_{45};$$

where the vector

$$\vec{k}=\begin{pmatrix}0\\0\\1\end{pmatrix}$$

and the massic vectors $\theta_{12}=\theta_{21}=\frac{1}{4}\vec{k}$.

If we modify θ_{12} and θ_{21}, the surface preserves its boundaries but the shape of the surface is modified (see Figures 9 and 10).

In a future work, we will apply the previous results to define an n-sided filling SBR-surface which joins G^1 or G^2 continuously with n given polynomial or rational surfaces. More generally (see [12], [13], [15], [14], [17], [16], [19]), we have dealt with

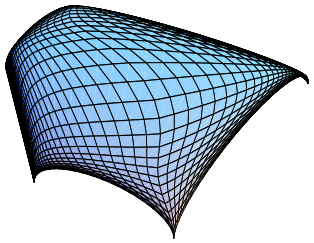


FIG. 9. *A five-sided rational surface where the segment is the image of the base point.*

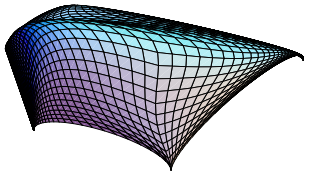


FIG. 10. *A five-sided rational surface with* $\theta_{12} = \vec{0}$ *and* $\theta_{21} = 2\vec{k}$.

this n-sided filling problem with parametric pole-functions of two variables where the image of a base point is several parametric boundary curves.

Acknowledgments. The author is grateful to Professor J. C. Fiorot for his valuable comments and suggestions. He would also like to thank A. Massabo for having brought to his attention the significance of base points for the applications.

REFERENCES

[1] P. BÉZIER, *Définition numérique des courbes et des surfaces* I, Automatisme, 11 (1966), pp. 625–632.
[2] J. F. CANNY AND I. EMIRIS, *An efficient algorithm for the sparse mixed resultant*, in Proceedings of AAECC, G. Cohen, T. Mora, and O. Moreno, eds., Springer-Verlag, Berlin, 1993, pp. 89–104.
[3] E. CARTAN, *Nombres complexes*, in Oeuvres Complètes, Partie II, Vol. I, Gauthier-Villars, Paris, 1953, pp. 107–246.
[4] E. W. CHIONH, *Base Points, Resultants and the Implicit Representation of Rational Surfaces*, Ph.D. thesis, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1990.
[5] A. CLEBSCH, *Ueber die abbildung algebraischer flächen, insbesondere der vierten und fünften ordnung*, Math. Ann., 1 (1869), pp. 253–316.
[6] D. COX, J. LITTLE, AND D. O'SHEA, *Using Algebraic Geometry*, Springer-Verlag, Berlin, 1998.
[7] P. DE CASTELJAU, *Outillage, Méthode de Calcul*, André Citroën Automobiles, Paris, 1959.
[8] J. DIEUDONNÉ, *Calcul Infinitésimal*, Hermann, Paris, 1968.
[9] G. FARIN, *Curves and Surfaces for Computer Aided Geometric Design*, 2nd ed., Academic Press, New York, 1992.
[10] J. C. FIOROT AND P. JEANNIN, *Courbes Splines Rationnelles. Applications à la CAO*, R.M.A. 24, Masson, Paris, 1992.
[11] J. C. FIOROT AND P. JEANNIN, *Courbes et Surfaces Rationnelles. Applications à la CAO*, R.M.A. 12, Masson, Paris, 1989.
[12] J. C. FIOROT AND O. GIBARU, *Blowing up: Application to* G^2*-continuous* 8*-sided filling patch*, Numer. Math., 92 (2002), pp. 257–287.

[13] J. C. Fiorot and O. Gibaru, *A rectangular G^m-continuous filling surface patch and some improvements at corners*, Comput. Aided Geom. Design, 18 (2001), pp. 175–194.
[14] J. C. Fiorot and O. Gibaru, *Modèle de surface de remplissage G^2-continu via un éclatement multiple en géométrie de la CAO*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 1113–1118.
[15] J. C. Fiorot and O. Gibaru, *Blowing up method. Application to G^2-continuous filling surfaces in CAD*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 623–628.
[16] J. C. Fiorot and O. Gibaru, *Surface de remplissage G^2-continu à n côtés*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 307–312.
[17] J. C. Fiorot and O. Gibaru, *The smooth G^2 filling of a rectangular patch hole joining four heterogeneous surfaces*, in Mathematical Methods for Curves and Surfaces II, M. Dæhlem, T. Lyche, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, pp. 175–182.
[18] O. Gibaru and J. C. Fiorot, *Détermination des images d'un point de base d'une surface rationnelle tensorielle définie par des vecteurs massiques*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 283–288.
[19] O. Gibaru, *Surfaces de remplissage G^k-continues à n-côtés sur des géométries non nécessairement compatibles*, Ph.D. thesis, Ecole Nationale Supérieure d'Arts et Métiers, Lille, France, 1997.
[20] R. Goldman, *Pyramid Algorithms*, Morgan Kaufmann, San Francisco, 2003.
[21] H. Grassmann, *Die Lineale Ausdehnungslehere*, 1844, 1863, Engel, Leipzig, 1894–1896.
[22] R. Krasauskas, *Toric surface patches*, Adv. Comput. Math., 17 (2002), pp. 89–113.
[23] D. Manocha and J. F. Canny, *Algorithm for implicitizing rational parametric surfaces*, Comput. Aided Geom. Design, 9 (1992), pp. 25–50.
[24] D. Manocha and J. F. Canny, *The implicit representation of rational parametric surfaces*, J. Symbolic Comput., 13 (1992), pp. 485–510.
[25] D. Manocha and J. F. Canny, *Implicitizing Rational Parametric Surfaces*, Tech. report 90/592, Computer Science Division, University of California, Berkeley, 1990.
[26] J. G. Semple and L. Roth, *Introduction to Algebraic Geometry*, Oxford University Press, London, 1949.
[27] V. Snyder, A. Black, A. Coble, L. Dye, A. Emch, S. Lefschetz, F. Sharpe, and C. Sisam, *Selected Topics in Algebraic Geometry*, 2nd ed., Chelsea, New York, 1970.
[28] J. Verschelde, P. Verlinden, and R. Cools, *Homotopies exploiting Newton polytopes for solving sparse polynomial systems*, SIAM J. Numer. Anal., 31 (1994), pp. 915–930.
[29] R. J. Walker, *Algebraic Curves*, Dover Publications, New York, 1950.
[30] J. Warren, *A bound on the implicit degree of polygonal Bézier surfaces*, in Algebraic Geometry and Its Applications, C. Bajaj, ed., Springer, New York, 1994, pp. 513–525.
[31] J. Warren, *Creating multisided rational Bézier surfaces using base points*, ACM Trans. Graphics, 11 (1992), pp. 127–139.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1435–1451

A POSTERIORI ESTIMATION OF DIMENSION REDUCTION ERRORS FOR ELLIPTIC PROBLEMS ON THIN DOMAINS*

SERGEY REPIN†, STEFAN SAUTER‡, AND ANTON SMOLIANSKI‡

Abstract. A new a posteriori error estimator is presented for the verification of the dimensionally reduced models stemming from the elliptic problems on thin domains. The original problem is considered in a general setting, without any specific assumptions on the domain geometry, coefficients, and the right-hand sides. For the energy norm of the error of the zero-order dimension reduction method, the proposed estimator is shown to always provide a guaranteed upper bound. In the case when the original domain has constant thickness (but, possibly, nonplane upper and lower faces), the estimator demonstrates the optimal convergence rate as the thickness tends to zero. It is also flexible enough to successfully cope with infinitely growing right-hand sides in the equation when the domain thickness tends to zero. The numerical tests indicate the efficiency of the estimator and its ability to accurately represent the local error distribution needed for an adaptive improvement of the reduced model.

Key words. dimension reduction, thin domain, a posteriori error estimate, reliability, efficiency, local error distribution

AMS subject classifications. 35J20, 65N15, 65N30

DOI. 10.1137/030602381

1. Introduction.

The method of dimension reduction is a popular approach frequently used by engineers for the approximate solution of the problems posed in *thin* domains. The term "thin" means that the size of the original physical domain along one coordinate direction is much smaller than along the others; this allows us to make some simplifying assumptions on the behavior of the exact solution and to replace the original high-dimensional problem with a lower-dimensional one. For instance, such a situation arises if a three-dimensional problem is analyzed with the help of a two-dimensional model. It is, however, clear that the solution of the new, "reduced" problem will, in general, differ from the solution to the original high-dimensional problem. Thus, the dimension reduction method unavoidably produces an error that can be referred to as the dimension reduction or the *modeling error*. The essential part of the model verification is, hence, a reliable *a posteriori* control of the dimension reduction error.

Despite the practical importance of the topic, only a few a posteriori estimators for the dimension reduction error have been introduced so far. In [15] and [3] (see also [2]) the residual-type estimators were proposed and proved reliable and efficient under the assumptions that the right-hand side of the given equation is zero and the original domain is a plate with plane parallel faces. In [5] and [12] the implicit estimators based on the solution of local three-dimensional Neumann problems were developed for the hierarchical modeling of complex elastic plates. In [1] the estimator of Babuška and Schwab (see [2], [3]) was extended to take into account the discretization error stemming from the approximate solution of the reduced problem. In this respect, we

*Received by the editors May 5, 2004; accepted for publication (in revised form) May 13, 2004; published electronically December 16, 2004.
http://www.siam.org/journals/sinum/42-4/60238.html
†V. A. Steklov Institute of Mathematics, Fontanka 27, 191 011 St. Petersburg, Russia (repin@pdmi.ras.ru).
‡Institute of Mathematics, Zurich University, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland (stas@amath.unizh.ch, antsmol@amath.unizh.ch).

have to notice that the present work is focused on the estimation of the modeling error; i.e., we assume, exactly as in [2], [3], that the error of discretizing the reduced problem is negligible. The work on the simultaneous a posteriori estimation of both the modeling error and the discretization error will be reported in a forthcoming paper.

In this work we propose a reliable and efficient a posteriori estimator for the dimension reduction error in the energy norm, having no specific assumptions on the right-hand side of the given equation and considering a general geometry of the given domain. In contrast to the above-mentioned papers, which deal with the hierarchical modeling of the problems in thin domains, we consider only the so-called *zero-order method* of dimension reduction that is, however, very popular owing to its simplicity and purely two-dimensional formulation. At the same time, this method forms a basis for the hierarchical modeling of three-dimensional plates (see, e.g., [14], [3], [12]). It is also worth noting that the zero-order method of dimension reduction does not cover the important Kirchhoff plate model in linear elasticity. The presented approach can, however, be extended to this case; the work on this subject is underway.

We advocate the functional-type a posteriori error estimation approach (see [7], [8], [9], [10]) that essentially differs from the approaches taken in the aforementioned articles; however, surprisingly enough, it is possible to show that Babuška and Schwab's estimator for the zero-order reduced problem can be obtained as a particular case of our estimator when the right-hand side of the equation is zero and the original domain is a plate with plane parallel faces. It must also be noticed that the treatment of the case with nonzero right-hand side may require special care, as we are about to see in one of the numerical examples; the presented estimator exhibits sufficient flexibility to remain efficient in this case.

The paper is set out as follows. Section 2 contains the geometric definitions and the problem statement. In section 3 we derive the reduced problem. Section 4 is devoted to the derivation of the a posteriori error estimate, while in section 5 we consider two particular cases and analyze the behavior of the estimator. The numerical examples are considered in section 6, and we draw the conclusions in section 7.

2. Problem setting. We consider three-dimensional Lipschitz domains, which can be given in the form

$$\Omega := \{x \in \mathbb{R}^3 \mid (x_1, x_2) \in \widehat{\Omega}\,,\ d_\ominus(x_1, x_2) < x_3 < d_\oplus(x_1, x_2)\}\,,$$

where $\widehat{\Omega} \subset \mathbb{R}^2$ is the orthogonal projection of Ω on the (x_1, x_2)-plane ($\widehat{\Omega}$ has the Lipschitz boundary $\widehat{\Gamma}$) and $d_\ominus$ and $d_\oplus$ are Lipschitz continuous functions defined on $\overline{\widehat{\Omega}}$. The lower and upper faces of Ω are denoted by

$$\Gamma_\ominus := \{x \in \mathbb{R}^3 \mid (x_1, x_2) \in \widehat{\Omega}\,,\ x_3 = d_\ominus(x_1, x_2)\}$$

and

$$\Gamma_\oplus := \{x \in \mathbb{R}^3 \mid (x_1, x_2) \in \widehat{\Omega}\,,\ x_3 = d_\oplus(x_1, x_2)\}\,;$$

the lateral boundary by

$$\Gamma_0 := \{x \in \mathbb{R}^3 \mid (x_1, x_2) \in \widehat{\Gamma}\,,\ d_\ominus(x_1, x_2) < x_3 < d_\oplus(x_1, x_2)\}$$

(see Figure 1).

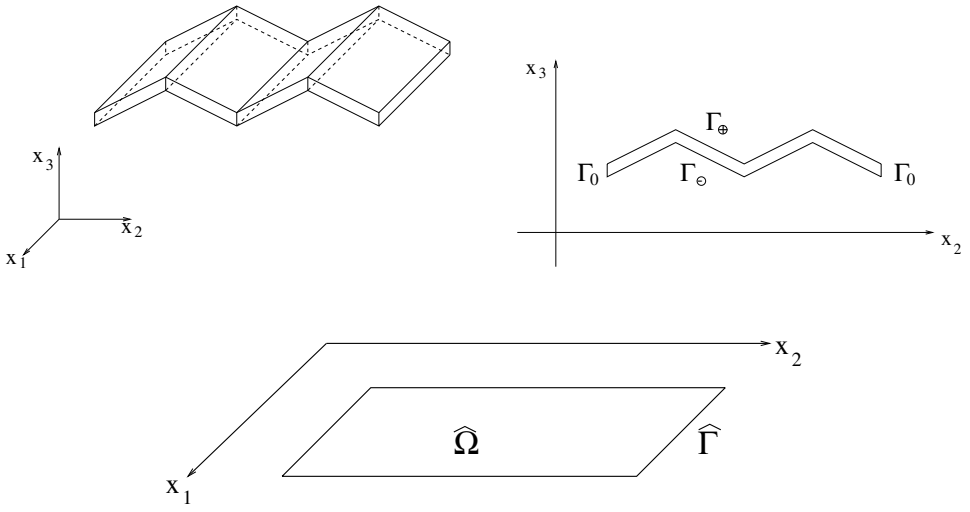


FIG. 1. *Sketch of the domain geometry.*

Remark 2.1. We consider $d_\ominus$ and $d_\oplus$ as explicit functions of (x_1, x_2)-coordinates only for the sake of simplicity. The generalization of the theory to the case of an arbitrary Lipschitz domain Ω presents no difficulty from the conceptional point of view.

The assumption that the given domain Ω is "thin" can now be written as

$$\operatorname{diam}\widehat{\Omega} \gg \max_{(x_1,x_2)\in\overline{\widehat{\Omega}}} d\,(x_1, x_2)\,, \tag{2.1}$$

where $d = d_\oplus - d_\ominus$ is the domain thickness, $d\,(x_1, x_2) \geq d_* > 0 \;\; \forall (x_1, x_2) \in \overline{\widehat{\Omega}}$. Although the assumption is of a purely qualitative nature, it will motivate the derivation of the corresponding two-dimensional reduced model in the next section. We also have to notice that Figure 1 depicts a simplified case; in the geometrical definitions we do not assume the domain thickness $d\,(x_1, x_2)$ to be a constant.

In the domain Ω we consider a model elliptic problem

$$-\operatorname{Div}(\mathbf{A}\nabla u) = f \quad \text{in } \Omega\,, \tag{2.2}$$

$$u = 0 \quad \text{on } \Gamma_0\,, \tag{2.3}$$

$$\mathbf{A}\nabla u \cdot \boldsymbol{\nu}_\ominus = F_\ominus \text{ on } \Gamma_\ominus\,, \tag{2.4}$$

$$\mathbf{A}\nabla u \cdot \boldsymbol{\nu}_\oplus = F_\oplus \text{ on } \Gamma_\oplus\,, \tag{2.5}$$

where $f \in L_2(\Omega)$, $F_\ominus \in L_2(\Gamma_\ominus)$, $F_\oplus \in L_2(\Gamma_\oplus)$, and $\boldsymbol{\nu}_\ominus$ and $\boldsymbol{\nu}_\oplus$ are outward normal vectors at $\Gamma_\ominus$ and $\Gamma_\oplus$, respectively. The matrix $\mathbf{A} = (a_{ij}(x))_{i,j=\overline{1,3}}$ with the components from $L_\infty(\Omega)$ is symmetric and uniformly positive definite; i.e., there exist constants $0 < c < C < \infty$ such that

$$c|\xi|^2 \leq \mathbf{A}(x)\xi \cdot \xi \leq C|\xi|^2 \quad \forall \xi \in \mathbb{R}^3, \text{ a.e. in } \Omega. \tag{2.6}$$

If the space of admissible functions is denoted by

$$V_0 := \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_0\}\,, \tag{2.7}$$

the weak form of problem (2.2)–(2.5) reads as follows.

Problem (P). Find $u \in V_0$ such that

$$\int_\Omega \mathbf{A}\nabla u \cdot \nabla w \, dx = \int_\Omega f\, w \, dx + \int_{\Gamma_\ominus} F_\ominus \, w \, ds + \int_{\Gamma_\oplus} F_\oplus \, w \, ds \qquad \forall w \in V_0 \,. \tag{2.8}$$

From now on we will frequently use the notation $\widehat{x} = (x_1, x_2)$, $\widehat{x} \in \widehat{\Omega}$, and all functions depending only on (x_1, x_2) will be marked by $\widehat{\ }$; in addition, we will distinguish between the three- and two-dimensional divergence operators:

$$\mathrm{Div}\, \boldsymbol{\tau} = \frac{\partial \boldsymbol{\tau}_1}{\partial x_1} + \frac{\partial \boldsymbol{\tau}_2}{\partial x_2} + \frac{\partial \boldsymbol{\tau}_3}{\partial x_3} \,, \quad \mathrm{div}\, \widehat{\boldsymbol{\tau}} = \frac{\partial \widehat{\boldsymbol{\tau}}_1}{\partial x_1} + \frac{\partial \widehat{\boldsymbol{\tau}}_2}{\partial x_2} \,.$$

We also denote $\widehat{F}_\ominus(\widehat{x}) := F_\ominus(\widehat{x}, d_\ominus(\widehat{x}))$, $\widehat{F}_\oplus(\widehat{x}) := F_\oplus(\widehat{x}, d_\oplus(\widehat{x}))$ for any $\widehat{x} \in \widehat{\Omega}$. Finally, we define the energy norm

$$|||v||| := \left(\int_\Omega \mathbf{A}(x) \nabla v \cdot \nabla v \, dx \right)^{1/2} \qquad \forall v \in V_0 \,. \tag{2.9}$$

3. The reduced problem. In view of (2.1), it is reasonable to consider the hypothesis that

(3.1) the exact solution u is almost *constant* with respect to the x_3-coordinate.

This gives rise to the so-called *zero-order reduced model* for the original problem (2.8). The model is very popular due to its simplicity and purely two-dimensional formulation. A discussion on the hierarchy of reduced models of different orders can be found in, e.g., [14], [3].

With (3.1) in mind, one can expect that the exact solution u may be well approximated by the functions from the subspace

(3.2)
$$\widehat{V}_0 := \{ v \in V_0 \mid \exists\, \widehat{v} \in H^1_0(\widehat{\Omega}) \text{ such that } v(x) = \widehat{v}(\widehat{x}) \text{ for a.e. } x = (\widehat{x}, x_3) \in \Omega \} \,.$$

Thus, any function from $\widehat{V}_0$ can be identified with the corresponding function $\widehat{v} \in H^1_0(\widehat{\Omega})$ (and vice versa: for any $\widehat{v} \in H^1_0(\widehat{\Omega})$ one can reconstruct $v \in \widehat{V}_0 \subset V_0$ by the constant extension as in the definition of $\widehat{V}_0$). Then, the energy-norm projection of u onto the subspace $\widehat{V}_0$ yields the following *reduced problem* (the zero-order reduced model).

Problem ($\widehat{P}$). Find $\widehat{u} \in \widehat{V}_0$ such that

$$\int_\Omega \mathbf{A}\nabla \widehat{u} \cdot \nabla \widehat{w} \, dx = \int_\Omega f\, \widehat{w} \, dx + \int_{\Gamma_\ominus} F_\ominus \, \widehat{w} \, ds + \int_{\Gamma_\oplus} F_\oplus \, \widehat{w} \, ds \qquad \forall \widehat{w} \in \widehat{V}_0 \,. \tag{3.3}$$

Now we can define the dimension reduction error (the *modeling error*) as the difference $e := u - \widehat{u}$ between the solution to the original problem (2.8) and the solution to the reduced problem (3.3).

Remark 3.1. It may be noticed that assumption (2.1) (and, consequently, (3.1)) serves only as an intuitive motivation for the introduction of the approximation subspace $\widehat{V}_0$ and the reduced problem (3.3). Since the assumption cannot be quantified,

the real error of "replacing" u with $\widehat{u}$ may be large; a robust a posteriori error estimator should, however, measure this error sufficiently accurately even in the cases when assumption (2.1) is virtually unsatisfied.

Remark 3.2. The asymptotic behavior of the modeling error e was analyzed in [14] (see also [2]) for the case of a plate with plane parallel faces $\Gamma_\ominus$ and $\Gamma_\oplus$ (i.e., when $d_\ominus = -\frac{d_0}{2}$, $d_\oplus = \frac{d_0}{2}$, $d_0 = \text{const} > 0$ is the plate thickness) and $f = 0$. It was proved that

$$|||e||| \le C\, d_0^{1/2} \left(\|\widehat{F}_\ominus\|_{L_2(\widehat{\Omega})} + \|\widehat{F}_\oplus\|_{L_2(\widehat{\Omega})} \right) \quad \text{as } d_0 \to 0\,.$$

Remark 3.3. We have to note that the third component of the vector $\nabla\widehat{u}$ is zero (since $\widehat{u}$ does not depend on x_3) and, thus, the vector will sometimes be considered as a two-component vector when no confusion is possible.

In order to see that the reduced problem (3.3) is, in fact, a two-dimensional problem, we define the operation $(\widetilde{\ })$ of averaging in the x_3-direction as follows:

$$\forall g \in L_1(\Omega)\ : \quad \widetilde{g}(\widehat{x}) := \frac{1}{d\,(\widehat{x})} \int_{d_\ominus(\widehat{x})}^{d_\oplus(\widehat{x})} g(\widehat{x}, x_3)\, dx_3 \ \text{ for a.e. } \widehat{x} \in \widehat{\Omega}\,,$$

and, having noticed that

$$\int_{\Gamma_\ominus} F_\ominus\, \widehat{w}\, ds = \int_{\widehat{\Omega}} \widehat{F}_\ominus\,(\widehat{x})\, \widehat{w}\,(\widehat{x}) \sqrt{1 + |\nabla d_\ominus(\widehat{x})|^2}\, d\widehat{x} \quad \left(\text{analogously for } \int_{\Gamma_\oplus} F_\oplus\, \widehat{w}\, ds \right),$$

we can rewrite problem (3.3) as follows.

Find $\widehat{u} \in \widehat{V}_0$ such that

$$\int_{\widehat{\Omega}} d\,(\widehat{x}) \widetilde{\mathbf{A}}_p(\widehat{x}) \nabla\widehat{u} \cdot \nabla\widehat{w}\, d\widehat{x} = \int_{\widehat{\Omega}} d\,(\widehat{x}) \widehat{f}(\widehat{x}) \widehat{w}\, d\widehat{x} \quad \forall \widehat{w} \in \widehat{V}_0\,. \tag{3.4}$$

Here $\widetilde{\mathbf{A}}_p(\widehat{x}) = (\widetilde{a}_{ij}(\widehat{x}))_{i,j=\overline{1,2}}$ is the averaged "plane" part $\mathbf{A}_p(x)$ $(\mathbf{A}_p(x) = (a_{ij}(x))_{i,j=\overline{1,2}})$ of the matrix $\mathbf{A}$ and

$$\widehat{f}(\widehat{x}) = \widetilde{f}(\widehat{x}) + \frac{\widehat{F}_\ominus(\widehat{x}) \sqrt{1 + |\nabla d_\ominus(\widehat{x})|^2} + \widehat{F}_\oplus(\widehat{x}) \sqrt{1 + |\nabla d_\oplus(\widehat{x})|^2}}{d\,(\widehat{x})}\,.$$

It is clear that problem (3.4) is a two-dimensional elliptic problem with the homogeneous Dirichlet boundary condition

$$-\mathrm{div}\,(d(\widehat{x})\, \widetilde{\mathbf{A}}_p(\widehat{x}) \nabla\widehat{u}) = d(\widehat{x})\, \widehat{f}(\widehat{x}) \quad \text{in } \widehat{\Omega}, \tag{3.5}$$

$$\widehat{u} = 0 \quad \text{on } \widehat{\Gamma}\,. \tag{3.6}$$

4. A posteriori estimation of the modeling error. In order to control the dimension reduction error, we apply the functional-type a posteriori error estimate derived in [10] (see also [7] and [9]) to the original three-dimensional problem (2.8). The estimate reads as follows.

For all $\gamma > 0$, $\delta > 0$, and $y^* \in H_*(\Omega, \mathrm{Div})$ there holds

$$\begin{aligned} |||u - v|||^2 \le{}& (1+\gamma)\, M_1^2\,(v, y^*) + \left(1 + \frac{1}{\gamma}\right)(1+\delta)\, C_\Omega^2\, M_2^2\,(y^*) \\ &+ \left(1 + \frac{1}{\gamma}\right)\left(1 + \frac{1}{\delta}\right) C_\Gamma^2 (1 + C_\Omega^2)\, M_3^2\,(y^*), \end{aligned} \tag{4.1}$$

where v is any function from the energy space V_0, C_Ω is the constant from Friedrichs' inequality,

$$C_\Omega^{-2} = \inf_{w \in V_0 \setminus \{0\}} \frac{|||w|||^2}{\|w\|^2_{L_2(\Omega)}}\,, \tag{4.2}$$

C_Γ is the constant from the trace inequality,

$$C_\Gamma^2 = \sup_{w \in V_0 \setminus \{0\}} \frac{\|w\|^2_{L_2(\Gamma_\oplus)} + \|w\|^2_{L_2(\Gamma_\ominus)}}{|||w|||^2 + \|w\|^2_{L_2(\Omega)}}\,, \tag{4.3}$$

the space $H_*(\Omega, \mathrm{Div})$ is defined as

$$H_*(\Omega, \mathrm{Div}) := \{y^* \in L_2(\Omega, \mathbb{R}^3) \mid \mathrm{Div}\, y^* \in L_2(\Omega)\,,\ y^* \cdot \boldsymbol{\nu}_\ominus \in L_2(\Gamma_\ominus)\,,\ y^* \cdot \boldsymbol{\nu}_\oplus \in L_2(\Gamma_\oplus)\}\,,$$

and the functionals $M_1^2\,(v, y^*)$, $M_2^2\,(y^*)$, $M_3^2\,(y^*)$ are defined by

$$\begin{aligned}
M_1^2\,(v, y^*) &:= \int_\Omega (\nabla v - \mathbf{A}^{-1} y^*) \cdot (\mathbf{A}\nabla v - y^*)\, dx\,,\\
M_2^2\,(y^*) &:= \|\mathrm{Div}\, y^* + f\|^2_{L_2(\Omega)}\,,\\
M_3^2\,(y^*) &:= \|F_\ominus - y^* \cdot \boldsymbol{\nu}_\ominus\|^2_{L_2(\Gamma_\ominus)} + \|F_\oplus - y^* \cdot \boldsymbol{\nu}_\oplus\|^2_{L_2(\Gamma_\oplus)}\,.
\end{aligned}$$

In what follows, we will denote the functionals simply by M_1^2, M_2^2, M_3^2. Since estimate (4.1) holds true for any "approximate solution" v from V_0 and since the solution $\widehat{u}$ of the reduced problem is in $\widehat{V}_0 \subset V_0$, we can simply plug $\widehat{u}$ into estimate (4.1) to obtain an upper bound of the modeling error. We also emphasize that the estimate is valid for any positive numbers γ and δ and for any vector-function y^* from the space $H_*(\Omega, \mathrm{Div})$. While the best possible option would be to take as y^* the exact flux $\mathbf{A}\nabla u$ (then M_2 and M_3 would vanish and M_1 would give us the energy norm of the exact error), we have to restrict ourselves to choosing some computable quantity, i.e., not containing the unknown exact solution u. We approximate the flux by

$$y^* = \widetilde{\mathbf{A}}_p \nabla \widehat{u} + \boldsymbol{\tau}^*\,, \tag{4.4}$$

with $\boldsymbol{\tau}^* = \{0\,,\ 0\,,\ \psi(x)\}^T$. Here ψ is the auxiliary function from $L_2(\Omega)$ satisfying the conditions $\frac{\partial \psi}{\partial x_3} \in L_2(\Omega)\,$, $\psi \in L_2(\Gamma_\ominus)$, and $\psi \in L_2(\Gamma_\oplus)$. The concrete form of the function ψ will be given later. Its meaning becomes clear in the case of the Poisson equation (i.e., if $\mathbf{A}$ is the identity matrix), where ψ should, obviously, approximate the derivative $\frac{\partial u}{\partial x_3}$ of the exact solution in the x_3-direction. Using (3.5), it is easy to verify that y^* from (4.4) belongs to $H_*(\Omega, \mathrm{Div})$.

Remark 4.1. If in (4.1) we take y^* from the set

$$Q^*_{f,F} := \{q^* \in L_2(\Omega, \mathbb{R}^3) \mid \mathrm{Div}\, q^* = -f \text{ in } \Omega\,,\ q^* \cdot \boldsymbol{\nu}_{\ominus,\oplus} = F_{\ominus,\oplus} \text{ on } \Gamma_{\ominus,\oplus}\}\,,$$

we obtain the dual-formulation-based error estimate of [4] (see also [6] and [13]). Since it is not easy to satisfy the constraints of the set $Q^*_{f,F}$, the estimate (4.1) with y^* from $H_*(\Omega, \mathrm{Div})$ seems to be more practical. In particular, for the estimation of the modeling error under consideration we essentially exploit the freedom of choosing y^* in the whole space $H_*(\Omega, \mathrm{Div})$.

Remark 4.2. The estimate (4.1) possesses the property of asymptotic exactness (see [10]) but, if we choose y^* as in (4.4), this property might be lost, since the

only remaining "degree of freedom" is the function ψ and the approximate plane flux $\widetilde{\mathbf{A}}_p\nabla\widehat{u}$ may not sufficiently represent the first two components of the exact flux $\mathbf{A}\nabla u$. On the other hand, if we did not fix the first two components of y^*, the process of estimation would require the minimization of the right-hand side of (4.1) with respect to those components, which is, in principle, equivalent to solving a three-dimensional problem. However, our goal is to avoid any truly three-dimensional calculations in the evaluation of the error estimator (this process should not be more expensive than the solution of the reduced problem). Fortunately, in most of the situations, $\widetilde{\mathbf{A}}_p\nabla\widehat{u}$ is a good approximation to the "plane" part of the exact flux, and the modeling-error estimate with y^* as in (4.4) exhibits both efficiency and flexibility, as the numerical tests of section 6 show.

In order to rewrite estimate (4.1) in a more convenient form, we introduce the notation

$$\mathbf{B} := \mathbf{A}^{-1} \quad (\mathbf{B}(x) = (b_{ij}(x))_{i,j=\overline{1,3}}\,,\ \ \mathbf{B} = \mathbf{B}^T)\,, \tag{4.5}$$

$$\mathbf{B}_p := (b_{ij})_{i,j=\overline{1,2}}\,, \tag{4.6}$$

$$\mathbf{b}_3 := \{b_{31}\,,\ b_{32}\}^T\,. \tag{4.7}$$

The term M_1^2 with $v = \widehat{u}$ reads

$$M_1^2 = \int_\Omega (\nabla\widehat{u} - \mathbf{B}y^*)\cdot(\mathbf{A}\nabla\widehat{u} - y^*)\,dx = \int_\Omega (\mathbf{A}\nabla\widehat{u}\cdot\nabla\widehat{u} - 2y^*\cdot\nabla\widehat{u} + \mathbf{B}y^*\cdot y^*)\,dx\,. \tag{4.8}$$

For the first term in (4.8), one immediately obtains

$$\int_\Omega \mathbf{A}\nabla\widehat{u}\cdot\nabla\widehat{u}\,dx = \int_{\widehat{\Omega}} d\,(\widehat{x})\widetilde{\mathbf{A}}_p(\widehat{x})\nabla\widehat{u}\cdot\nabla\widehat{u}\,d\widehat{x}\,. \tag{4.9}$$

The second term in (4.8) can be further rewritten if one notices that (recall $\frac{\partial\widehat{u}}{\partial x_3} = 0$)

$$y^*\cdot\nabla\widehat{u} = (\widetilde{\mathbf{A}}_p\nabla\widehat{u} + \tau^*)\cdot\nabla\widehat{u} = \widetilde{\mathbf{A}}_p\nabla\widehat{u}\cdot\nabla\widehat{u}\,.$$

Thus,

$$\int_\Omega y^*\cdot\nabla\widehat{u}\,dx = \int_{\widehat{\Omega}} d\,(\widehat{x})\widetilde{\mathbf{A}}_p(\widehat{x})\nabla\widehat{u}\cdot\nabla\widehat{u}\,d\widehat{x}\,. \tag{4.10}$$

For the third term in (4.8) we have

$$\begin{aligned}\mathbf{B}y^*\cdot y^* &= (\mathbf{B}\widetilde{\mathbf{A}}_p\nabla\widehat{u} + \mathbf{B}\tau^*)\cdot(\widetilde{\mathbf{A}}_p\nabla\widehat{u} + \tau^*) = \mathbf{B}\widetilde{\mathbf{A}}_p\nabla\widehat{u}\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u} + \mathbf{B}\widetilde{\mathbf{A}}_p\nabla\widehat{u}\cdot\tau^*\\ &\quad + \mathbf{B}\tau^*\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u} + \mathbf{B}\tau^*\cdot\tau^* = \mathbf{B}_p\widetilde{\mathbf{A}}_p\nabla\widehat{u}\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u} + 2(\mathbf{b}_3\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u})\psi + b_{33}\psi^2\end{aligned}$$

that yields

$$\int_\Omega \mathbf{B}y^*\cdot y^*\,dx = \int_{\widehat{\Omega}} d\,(\widehat{x})\widetilde{\mathbf{B}}_p\widetilde{\mathbf{A}}_p\nabla\widehat{u}\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u}\,d\widehat{x} + \int_\Omega (b_{33}\psi(x)^2 + 2(\mathbf{b}_3\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u})\psi(x))\,dx\,, \tag{4.11}$$

where $\widetilde{\mathbf{B}}_p$ is the averaged "plane" part $\mathbf{B}_p(x)$ of the matrix $\mathbf{B}(x)$.

Substituting (4.9), (4.10), and (4.11) into (4.8), one obtains

(4.12)
$$M_1^2 = \int_{\widehat{\Omega}} d\,(\widehat{x})\,(\widetilde{\mathbf{B}}_p\widetilde{\mathbf{A}}_p - \mathbf{I})\nabla\widehat{u}\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u}\,d\widehat{x} + \int_\Omega (b_{33}\psi(x)^2 + 2(\mathbf{b}_3\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u})\psi(x))\,dx\,,$$

where $\mathbf{I}$ is the identity (2×2)-matrix. It is interesting to note that the first integral in (4.12) represents the error in averaging the coefficient matrix $\mathbf{A}(x)$; this becomes fully transparent in the case of a block-diagonal matrix $\mathbf{A}$, i.e., when $a_{31} = a_{32} = 0$ (then $\mathbf{B}_p = \mathbf{A}_p^{-1}$ and, without the averaging, the integral would be identically zero).

The functional M_2^2 of (4.1) also can be rearranged if one takes y^* as in (4.4). First, note that

$$\operatorname{Div} y^* = \operatorname{div}\widetilde{\mathbf{A}}_p\nabla\widehat{u} + \frac{\partial\psi}{\partial x_3}\,.$$

From (3.5) one can deduce

$$\operatorname{div}\widetilde{\mathbf{A}}_p\nabla\widehat{u} = -\widehat{f} - \frac{\nabla d}{d}\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u}\,.$$

Hence,

(4.13)
$$M_2^2 = \left\| f - \widetilde{f} - \frac{\widehat{F}_\ominus\sqrt{1+|\nabla d_\ominus|^2} + \widehat{F}_\oplus\sqrt{1+|\nabla d_\oplus|^2}}{d} - \frac{\nabla d}{d}\cdot\widetilde{\mathbf{A}}_p\nabla\widehat{u} + \frac{\partial\psi}{\partial x_3}\right\|^2_{L_2(\Omega)}\,.$$

The term M_3^2 with y^* from (4.4) reads

(4.14)
$$M_3^2 = \|F_\ominus - \widetilde{\mathbf{A}}_p\nabla\widehat{u}\cdot\boldsymbol{\nu}_\ominus - \psi\nu_{\ominus 3}\|^2_{L_2(\Gamma_\ominus)} + \|F_\oplus - \widetilde{\mathbf{A}}_p\nabla\widehat{u}\cdot\boldsymbol{\nu}_\oplus - \psi\nu_{\oplus 3}\|^2_{L_2(\Gamma_\oplus)}\,,$$

where $\widetilde{\mathbf{A}}_p\nabla\widehat{u}$ is considered as a vector in $\mathbb{R}^3$ with the third component equal to zero, and

$$\nu_{\ominus 3} = \frac{-1}{\sqrt{1+|\nabla d_\ominus|^2}}\,,\quad \nu_{\oplus 3} = \frac{1}{\sqrt{1+|\nabla d_\oplus|^2}}$$

are the third components of the normal vectors $\boldsymbol{\nu}_\ominus$ and $\boldsymbol{\nu}_\oplus$.

Now we can write the general a posteriori estimate for dimension reduction error as follows.

For all $\gamma > 0$ and $\delta > 0$ there holds

(4.15)
$$|||u-\widehat{u}|||^2 \le (1+\gamma)\,M_1^2 + \left(1+\frac{1}{\gamma}\right)(1+\delta)\,C_\Omega^2\,M_2^2 + \left(1+\frac{1}{\gamma}\right)\left(1+\frac{1}{\delta}\right)C_\Gamma^2(1+C_\Omega^2)\,M_3^2\,,$$

where the constants C_Ω and C_Γ are as above (see (4.2) and (4.3)) and the functionals M_1^2, M_2^2, and M_3^2 are given by (4.12), (4.13), and (4.14).

In estimate (4.15) we still have the freedom of choosing the auxiliary function ψ. The simplest choice is to take such a ψ so that the term M_3 (i.e., the residual on the

Neumann boundary condition) would be identically zero. To do so, we first rewrite the L_2-norms on $\Gamma_{\oplus,\ominus}$ in (4.14) as the integrals over $\widehat{\Omega}$:

$$\begin{aligned}&\| F_{\ominus} - \widetilde{\mathbf{A}}_p \nabla \widehat{u} \cdot \boldsymbol{\nu}_{\ominus} - \psi \nu_{\ominus 3} \|^2_{L_2(\Gamma_{\ominus})} \\ &\quad = \int_{\widehat{\Omega}} (\widehat{F}_{\ominus}(\widehat{x}) - \widetilde{\mathbf{A}}_p \nabla \widehat{u} \cdot \boldsymbol{\nu}_{\ominus} - \psi(\widehat{x}, d_{\ominus}(\widehat{x})) \nu_{\ominus 3})^2 \sqrt{1 + |\nabla d_{\ominus}|^2}\, d\widehat{x}\end{aligned}$$

(analogously for the norm in $L_2(\Gamma_{\oplus})$). Then, we denote

$$\widehat{G}_{\oplus,\ominus} := \widehat{F}_{\oplus,\ominus} - \widetilde{\mathbf{A}}_p \nabla \widehat{u} \cdot \boldsymbol{\nu}_{\oplus,\ominus}$$

and set

$$\psi_1(x) = \widehat{\alpha}(\widehat{x})\, x_3 + \widehat{\beta}(\widehat{x})\,, \tag{4.16}$$

where the functions $\widehat{\alpha}$ and $\widehat{\beta}$ ($\widehat{\alpha}\,,\ \widehat{\beta} \in L_2(\widehat{\Omega})$) are chosen so that

$$\psi_1 \nu_{\oplus 3} = \widehat{G}_{\oplus} \ \text{ at } \ x_3 = d_{\oplus}\,, \quad \psi_1 \nu_{\ominus 3} = \widehat{G}_{\ominus} \ \text{ at } \ x_3 = d_{\ominus}\,. \tag{4.17}$$

As $\nu_{\ominus 3}$, $\nu_{\oplus 3}$ belong to $L_\infty(\widehat{\Omega})$ and cannot be zero in $\widehat{\Omega}$, the functions $\widehat{\alpha}$ and $\widehat{\beta}$ are uniquely defined by conditions (4.17):

$$\widehat{\alpha} = \frac{1}{d}\left(\frac{\widehat{G}_{\oplus}}{\nu_{\oplus 3}} - \frac{\widehat{G}_{\ominus}}{\nu_{\ominus 3}}\right), \tag{4.18}$$

$$\widehat{\beta} = \frac{1}{d}\left(\frac{\widehat{G}_{\ominus}}{\nu_{\ominus 3}}\, d_{\oplus} - \frac{\widehat{G}_{\oplus}}{\nu_{\oplus 3}}\, d_{\ominus}\right). \tag{4.19}$$

It is obvious that the function ψ_1, as well as its derivative in the x_3-direction, belongs to $L_2(\Omega)$, and ψ_1 belongs to $L_2(\Gamma_{\oplus})$ and $L_2(\Gamma_{\ominus})$ (since $\psi_1\big|_{x_3 = d_{\oplus,\ominus}(\widehat{x})} \in L_2(\widehat{\Omega})$). Moreover, with such a function ψ the term M_3 becomes zero.

Remark 4.3. One can also consider a quadratic (with respect to x_3) function

$$\psi_2(x) = \psi_1(x) + \widehat{\eta}(\widehat{x})(x_3 - d_{\oplus}(\widehat{x}))(x_3 - d_{\ominus}(\widehat{x}))$$

with $\widehat{\eta}$ being an arbitrary function from $L_2(\widehat{\Omega})$. The substitution of ψ_2 instead of ψ into (4.14) will evidently imply $M_3 = 0$. In the second numerical example of section 6 we will use ψ_2 because of the freedom in the choice of the function $\widehat{\eta}$. It is clear that one can quite analogously construct the functions $\{\psi_m\}$, $m = 3, 4, \dots$, which would make the M_3-term vanish and could, possibly, allow us to approximate the third component of the exact flux $\mathbf{A}\nabla u$ with a higher accuracy.

Having chosen the function ψ such that $M_3 = 0$, one can obtain from (4.15) the following estimate for the squared energy norm of the modeling error:

$$|||u - \widehat{u}|||^2 \le (1 + \gamma)\, M_1^2 + \left(1 + \frac{1}{\gamma}\right) C_\Omega^2\, M_2^2\,, \tag{4.20}$$

where γ is any positive number, C_Ω is the Friedrichs constant, and M_1^2 and M_2^2 are given by (4.12) and (4.13). Minimizing the right-hand side of (4.20) with respect to the scalar parameter $\gamma > 0$, we immediately arrive at the estimate for the energy norm of the modeling error,

$$|||u - \widehat{u}||| \le M := M_1 + C_\Omega\, M_2, \tag{4.21}$$

with M_1 and M_2 defined by (4.12) and (4.13).

The rest of the paper will be devoted to the analysis of the properties of estimates (4.20), (4.21).

5. Particular cases. The error majorant M in (4.21) has been derived for quite general geometry of Ω and coefficient matrix $\mathbf{A}(x)$; to make the estimate more transparent, we consider two particular cases.

5.1. Plate of constant thickness. We assume that

$$d_\oplus = d_\ominus + d_0 \quad (d_0 = \text{const} > 0) \tag{5.1}$$

and, in addition, that

$$\mathbf{A} = \mathbf{A}(\widehat{x}) \ \text{ (this immediately implies } \mathbf{B} = \mathbf{B}(\widehat{x})\text{)}, \tag{5.2}$$

$$a_{31} = a_{32} = 0 \ \text{ (this yields } \mathbf{B}_p = \mathbf{A}_p^{-1},\ b_{33} = a_{33}^{-1},\ b_{31} = b_{32} = 0\text{)}. \tag{5.3}$$

With these assumptions and the choice $\psi = \psi_1$ (see (4.16)) the terms M_1 and M_2 in estimate (4.21) become simpler:

$$M_1 = \left(\int_\Omega a_{33}^{-1} \psi_1^2 \, dx \right)^{1/2}, \quad M_2 = \|f - \widetilde{f}\|_{L_2(\Omega)}. \tag{5.4}$$

One may notice that the integral in the first term M_1 of the error majorant M can be rewritten as

$$\int_\Omega a_{33}^{-1} \psi_1^2 \, dx = d_0 \cdot \int_{\widehat{\Omega}} a_{33}^{-1} \left(\widehat{\alpha}^2 \frac{d_\oplus^2 + d_\oplus d_\ominus + d_\ominus^2}{3} + \widehat{\alpha}\widehat{\beta}(d_\oplus + d_\ominus) + \widehat{\beta}^2 \right) d\widehat{x},$$

which means that the term M_1 is of order $\mathcal{O}(d_0^{1/2})$ when the plate thickness d_0 tends to zero. If $f \in L_\infty(\Omega)$, the second term M_2 is obviously of the same order $\mathcal{O}(d_0^{1/2})$; i.e., the whole estimator M converges to zero with the rate $\mathcal{O}(d_0^{1/2})$ as $d_0 \to 0$. This is the optimal convergence rate for the modeling error e in the energy norm, as was shown in [14] for the simpler case of a plate with plane parallel faces and $f = 0$ (see Remark 3.2). It is worth noting that, if $f \in C^1(\Omega)$, the second term in M is of higher order $\mathcal{O}(d_0^{3/2})$ as compared to the first term.

5.2. Plate with plane parallel faces. If in addition to (5.2), (5.3) we strengthen assumption (5.1) by replacing it with

$$d_\oplus = \frac{d_0}{2},\ d_\ominus = -\frac{d_0}{2} \quad (d_0 = \text{const} > 0), \tag{5.5}$$

then the function ψ_1 takes the simple form

$$\psi_1(x) = \frac{\widehat{F}_\oplus(\widehat{x}) + \widehat{F}_\ominus(\widehat{x})}{d_0} \, x_3 + \frac{\widehat{F}_\oplus(\widehat{x}) - \widehat{F}_\ominus(\widehat{x})}{2}$$

and the error estimate (4.21) reduces to

$$|||u - \widehat{u}||| \le \sqrt{\frac{d_0}{3}} \left(\int_{\widehat{\Omega}} a_{33}^{-1} (\widehat{F}_\oplus^2 + \widehat{F}_\ominus^2 - \widehat{F}_\oplus \widehat{F}_\ominus) \, d\widehat{x} \right)^{1/2} + C_\Omega \, \|f - \widetilde{f}\|_{L_2(\Omega)}. \tag{5.6}$$

If we set here $f = 0$, $a_{33} = 1$, and $\widehat{F}_\oplus = \widehat{F}_\ominus = \widehat{F}$, we obtain

$$|||u - \widehat{u}||| \le \sqrt{\frac{d_0}{3}} \, \|\widehat{F}\|_{L_2(\widehat{\Omega})}, \tag{5.7}$$

which is exactly the estimator of Babuška and Schwab (see [2]) for the zero-order reduced model. Thus, the latter estimator can be obtained as a particular case of the error majorant (4.21) if one makes the assumptions (5.2), (5.3), (5.5) and sets $f = 0$. This fact is especially interesting, since we advocate the estimation approach (see the details in [10]) that is completely different from the one utilized in [2].

Remark 5.1. The error estimate (4.21) contains the Friedrichs constant C_Ω that must be, in general, evaluated numerically. The constant depends solely on the geometry of the domain Ω and can be computed as $1/\sqrt{\lambda}$, where λ is the minimal eigenvalue of the elliptic operator $-\mathrm{Div}\,(\mathbf{A}\nabla\cdot\,)$ equipped with the homogeneous Dirichlet condition on Γ_0 and homogeneous Neumann conditions on $\Gamma_{\oplus,\ominus}$ (see (4.2)). It is clear that, in the case of a plate with plane parallel faces, C_Ω can be easily estimated from above if one computes the Friedrichs constant in a larger domain obtained by embedding the cross section $\widehat{\Omega}$ of Ω into some rectangle; the faces of this larger domain are then obtained by the extension of plane faces of Ω. Yet a simpler, but rougher, upper estimate for C_Ω in the case of a plate with plane parallel faces is given by $(\mathrm{diam}\,\widehat{\Omega})/c$, where c is the lower bound of the minimal eigenvalue of the matrix $\mathbf{A}(x)$ in Ω (see (2.6)). It is worth noticing that the constant C_Ω multiplies in the majorant the term M_2, which is often of higher order as compared to the first term M_1 (it is so, for example, in the particular cases considered above, when the function f is smooth). Then, the possible error of overestimation of C_Ω is harmless for the majorant accuracy.

6. Numerical examples.

6.1. Numerical test 1. In order to analyze the performance of the proposed error estimator, we consider a two-dimensional test problem in the "sine-shape" domain (see Figure 2 (left)) whose upper and lower faces are given by

$$d_{\oplus,\ominus}(x) = \sin(k\pi x) \pm \frac{d_0}{2}\,, \quad k = 1, 2, \dots\,,$$

where $d_0 > 0$ is the domain thickness. In this example, $\widehat{\Omega} = (0, 1)$ and $\Omega = \{(x, y) \in \mathbb{R}^2 \mid x \in \widehat{\Omega}\,,\ d_\ominus(x) < y < d_\oplus(x)\}$. The considered problem is

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega\,,\\ u &= 0 && \text{at } x = 0 \text{ and } x = 1\,,\\ \nabla u \cdot \boldsymbol{\nu}_{\oplus,\ominus} &= F_{\oplus,\ominus} && \text{at } y = d_{\oplus,\ominus}\,, \end{aligned}$$

and the right-hand sides of the equation and of the boundary condition are computed using the exact solution

$$u(x, y) = \sin(\pi x) \cdot y^m \quad (m = 1, 2, \dots).$$

The reduced problem (3.3) is, in this case, a one-dimensional Dirichlet problem that, of course, can be solved very accurately (in the present work, we address the estimation of the modeling error only, assuming that the discretization error stemming from the solution of the reduced problem is negligible). The Friedrichs constant C_Ω was evaluated by computing the minimal eigenvalue of the Laplace operator with the corresponding homogeneous Dirichlet/Neumann boundary conditions (see Remark 5.1). We found that, for each $k = 1, 2, \dots$, C_Ω is an increasing function of the thickness d_0 as $d_0 \to 0$. There always exists, however, a clear upper bound for C_Ω; in particular, the estimates $C_\Omega \le \sqrt{2}$ for $k = 2$ and $C_\Omega \le 3$ for $k = 4$ hold true.

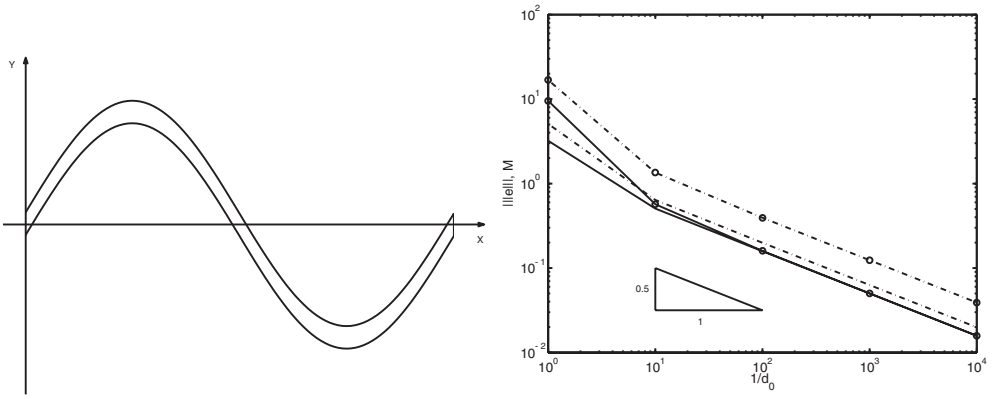


FIG. 2. *Left: The domain geometry. Right: Convergence rate of the exact modeling error and of the error majorant, $k = 2$, $m = 4$ (solid lines) and $m = 5$ (dash-dot lines). The majorant is indicated by "$\circ$."*

TABLE 1

Convergence of the exact modeling error in the energy norm ($|||e|||$) and of the error majorant (M) as $d_0 \to 0$ ($k = 2$); the results are rounded up to 10^{-4}.

d_0^{-1}	$m = 4$			$m = 5$		
	$\|\|\|e\|\|\|$	M	$\frac{M}{\|\|\|e\|\|\|}$	$\|\|\|e\|\|\|$	M	$\frac{M}{\|\|\|e\|\|\|}$
10^0	3.2108	9.5598	2.9774	5.0842	16.8434	3.3129
10^1	0.5058	0.5690	1.1250	0.6399	1.3481	2.1066
10^2	0.1581	0.1598	1.0106	0.1991	0.3937	1.9770
10^3	0.0500	0.0501	1.0010	0.0630	0.1237	1.9650
10^4	0.0158	0.0158	1.0000	0.0199	0.0391	1.9638

Figure 2 (right) shows the convergence rates of the exact modeling error in the energy norm ($|||e|||$) and of the error majorant (M) as the domain thickness d_0 tends to zero (the analysis here corresponds to the case $k = 2$, when the domain Ω has the shape depicted in Figure 2 (left)). It is clear that both the exact error and the majorant vanish with the theoretically predicted, optimal rate $\mathcal{O}(d_0^{1/2})$. However, the behavior of the majorant is different for even and odd values of degree m determining the polynomial growth of the exact solution u in the y-direction. The typical picture corresponding to an even value of the parameter m is well represented by the case $m = 4$ in Figure 2 (right); in this case, the majorant M demonstrates the asymptotic exactness, and, moreover, the *effectivity index* $\frac{M}{|||e|||}$ behaves like $1 + \mathcal{O}(d_0)$ (see Table 1). In the case of an odd value of m (represented by $m = 5$ in Figure 2 (right)), the majorant loses the property of asymptotic exactness, although the effectivity index remains stable and behaves, approximately, like $1.963 + \mathcal{O}(d_0)$ (see Table 1). This problem was addressed in Remark 4.1 and is caused by the fact that the approximate flux computed in the reduced model does not provide sufficient information on the corresponding components of the exact flux. We may note, however, that the effectivity index is still quite acceptable in this case. Finally, it is worth noticing that the presented error estimator provides a reliable upper bound for the exact error at any positive values of the domain thickness d_0, i.e., also in the cases when the domain is not "thin" at all.

The local error distributions provided by the exact error and by the first M_1-term of the majorant M (see (5.4)) are depicted in Figure 3 (here we consider the

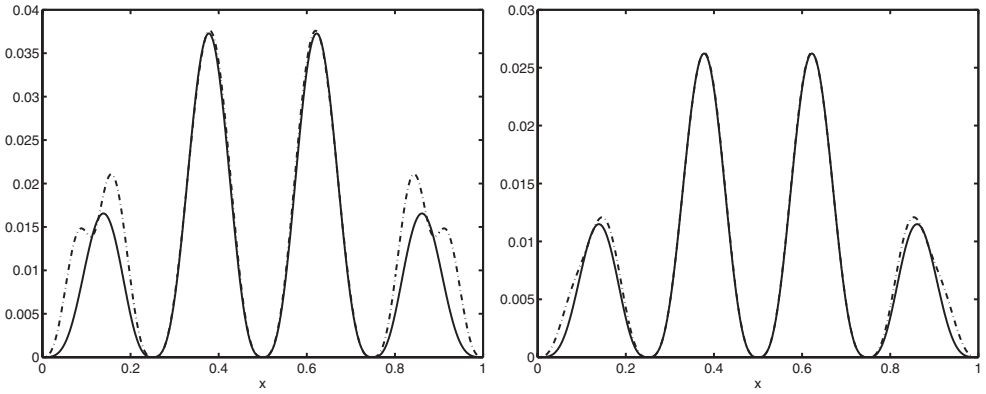


FIG. 3. *Local error distribution provided by the exact modeling error (solid line) and by the M_1-term of the majorant (dash-dot line), $k = 4$, $m = 4$. Left: $d_0 = 0.1$. Right: $d_0 = 0.05$.*

case $k = 4$, when the functions $d_{\oplus,\ominus}$ defining the shape of the domain have 4 extrema). The figure shows that already for rather large values of the domain thickness $d_0 = 0.1$ the majorant delivers sufficiently accurate information on the location of the regions of the biggest modeling error, while for $d_0 = 0.05$ the exact and the estimated error distributions are practically coincident.

6.2. Numerical test 2. The previous test shows that in the standard situations the proposed error estimator performs well. The example in this section demonstrates the performance of the estimator in a relatively difficult case when the right-hand side of the equation grows infinitely as the domain thickness tends to zero.

In this test, we consider a very simple geometry (see Figure 4 (left)), namely

$$d_{\oplus,\ominus} = \pm\frac{d_0}{2}\,,$$

where $d_0 > 0$ is the given thickness of the domain, $\widehat{\Omega} = (0,1)$ and $\Omega = \{(x,y) \in \mathbb{R}^2 \mid x \in \widehat{\Omega}\,,\ -\frac{d_0}{2} < y < \frac{d_0}{2}\}$. The considered problem reads

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega\,,\\ u &= 0 && \text{at } x = 0 \text{ and } x = 1\,,\\ \frac{\partial u}{\partial y} &= \pm F_{\oplus,\ominus} && \text{at } y = \pm\frac{d_0}{2}\,, \end{aligned}$$

and the right-hand sides of the equation and of the boundary condition are computed using the exact solution

$$u(x,y) = \sin(\pi x)\cdot\frac{y^m}{d_0^{m-1}} \quad (m = 1, 2, \dots).$$

The scaling factor d_0^{m-1} makes this test essentially different from the previous one: while the Neumann boundary data $F_{\oplus,\ominus}$ remain of order $\mathcal{O}(1)$ as $d_0 \to 0$, the right-hand side of the equation f exhibits the behavior $f \sim \mathcal{O}(d_0) + \mathcal{O}\big(\frac{1}{d_0}\big)$, i.e., unboundedly grows when d_0 tends to zero. The unbounded growth of f may yield serious problems for an a posteriori error estimator, as we are about to see. We also note that the

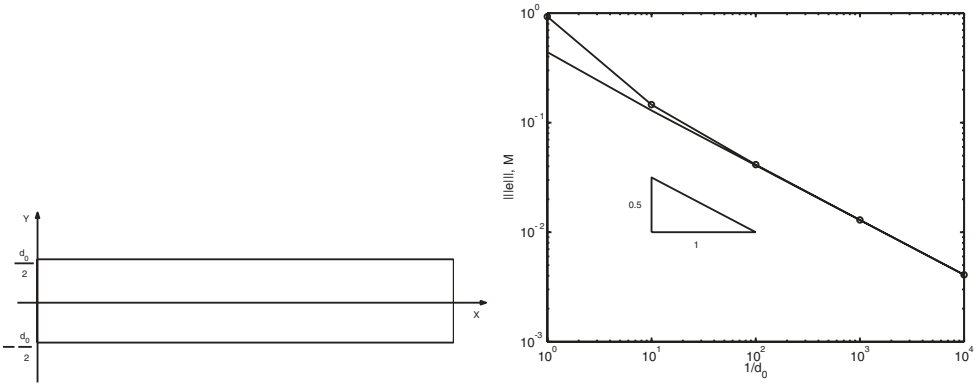


FIG. 4. *Left: The domain geometry. Right: Convergence rate of the exact modeling error and of the error majorant, $m = 2$; the majorant is indicated by* "$\circ$."

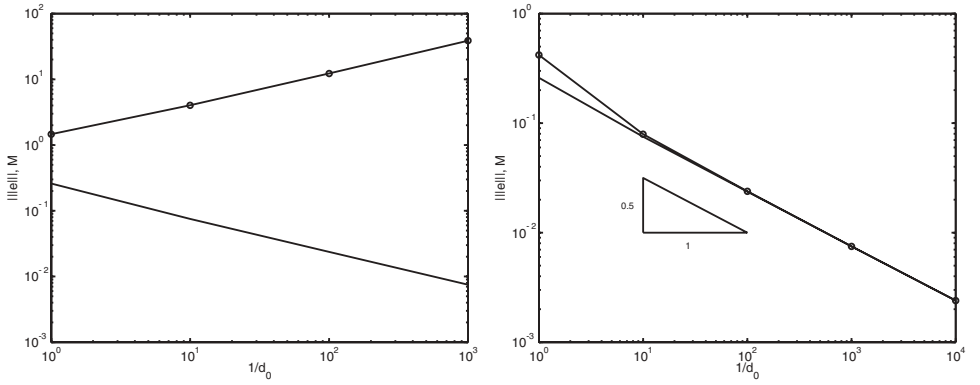


FIG. 5. *The case $m = 3$. Left: Divergence of the majorant $M(\psi_1)$ as $d_0 \to 0$. Right: Convergence of the improved majorant $M(\psi_2)$.*

constant C_Ω can be computed exactly in this example: $C_\Omega = \frac{1}{\pi}$ for all values of the thickness d_0.

First, we take $m = 2$ and observe the convergence of the exact modeling error in the energy norm and of the error majorant as d_0 tends to zero; see Figure 4 (right). As in the preceding example, the error majorant provides a reliable upper bound for the exact error at any values of the thickness d_0, both the exact error and the majorant demonstrate the optimal convergence rate $\mathcal{O}(d_0^{1/2})$ and, moreover, the error majorant shows the asymptotic exactness in this case (the effectivity index $\frac{M}{|||e|||} = 1 + \mathcal{O}(d_0)$; see the column under "$m = 2$, $M(\psi_1)$," in Table 2). However, if we set $m = 3$, the second term of the majorant M (i.e., $\|f - \widetilde{f}\|_{L_2(\Omega)}$, see (5.6)) becomes dominant and the whole estimator grows unboundedly, as can be seen in Figure 5 (left). The estimator becomes, of course, useless as it dramatically overestimates the exact error for small values of d_0. It is rather clear that the problem originates from the poor choice of the auxiliary function ψ that is supposed to approximate $\frac{\partial u}{\partial y}$; for $m = 3$ the derivative is quadratic and cannot be adequately represented by the linear function ψ_1.

The situation may be improved by invoking the quadratic function $\psi = \psi_2$ (see Remark 4.3), $\psi_2(x,y) = \psi_1(x,y) + \widehat{\eta}(x)\,(y^2 - \frac{d_0^2}{4})$, where $\widehat{\eta}$ is an arbitrary function from $L_2(\widehat{\Omega})$. The possibility of choosing a suitable $\widehat{\eta}$ enables us to suppress the unbounded growth of f in the M_2-term of the majorant and makes the majorant flexible enough to efficiently reproduce the behavior of the exact error.

If we plug ψ_2 into the estimate (4.20), we obtain

$$|||u - \widehat{u}|||^2 \le M^2(\widehat{\eta}, \gamma) \qquad \forall \widehat{\eta} \in L_2(\widehat{\Omega})\,,\ \forall \gamma > 0\,, \tag{6.1}$$

where

$$\begin{aligned} M^2(\widehat{\eta},\gamma) :&= (1+\gamma)\|\psi_2\|^2_{L_2(\Omega)} + \left(1 + \frac{1}{\gamma}\right) C_\Omega^2 \left\| f - \widetilde{f} + \frac{\partial \psi_2}{\partial y} \right\|^2_{L_2(\Omega)} \\ &= (1+\gamma) \int_\Omega \left(\psi_1(x,y) + \widehat{\eta}(x)\,\left(y^2 - \frac{d_0^2}{4}\right)\right)^2 dxdy \\ &\quad + \left(1 + \frac{1}{\gamma}\right) C_\Omega^2 \int_\Omega (f(x,y) - \widetilde{f}(x) + \widehat{\eta}(x)\cdot 2y)^2\, dxdy\,. \end{aligned}$$

Since estimate (6.1) is valid for any $\gamma > 0$ and $\widehat{\eta}$ from $L_2(\widehat{\Omega})$, one can minimize the functional $M^2(\widehat{\eta},\gamma)$ with respect to these parameters. In particular, one can set $\gamma = \gamma_* < 1$ (the concrete value of γ_* does not matter, as the numerical experiments show; we used the value $\gamma_* = 0.5$) and find $\widehat{\eta}_{\min}$ as the minimizer of $M^2(\widehat{\eta},\gamma_*)$ over the space S of piecewise-constant functions defined on some finite subdivision of $\widehat{\Omega}$ (obviously, $S \subset L_2(\widehat{\Omega})$). The minimization problem is just an L_2-projection onto the space of functions defined on $\widehat{\Omega}$ and amounts to the solution of a linear system with the diagonal matrix.

The properties of the improved majorant $M(\psi_2) = M_1(\psi_2) + C_\Omega\, M_2(\psi_2)$, where

$$\begin{aligned} M_1(\psi_2) &:= \left\| \psi_1 + \widehat{\eta}_{\min}\,\left(y^2 - \frac{d_0^2}{4}\right) \right\|_{L_2(\Omega)}\,, \\ M_2(\psi_2) &:= \|f - \widetilde{f} + \widehat{\eta}_{\min}\cdot 2y\|_{L_2(\Omega)}, \end{aligned}$$

can be observed in Figure 5 (right). We see that the improved majorant vanishes with the optimal rate $\mathcal{O}(d_0^{1/2})$ as $d_0 \to 0$, remains a reliable upper bound for the exact error at any values of the thickness d_0, and even demonstrates the asymptotic exactness with the effectivity index behaving like $1 + \mathcal{O}(d_0)$ (see Table 2).

We may note that in the case of larger values of m ($m > 3$) the higher degree function ψ_{m-1} might be needed (see Remark 4.3); the function will contain several free parameters which are the functions from $L_2(\widehat{\Omega})$, and, hence, the minimization should be performed with respect to all of them. However, as this always remains a least-squares minimization problem, the total complexity for the moderate values of m will not be greater than the complexity of solving the reduced problem. In general, if the right-hand side f exhibits an unbounded growth for $d_0 \to 0$ and no a priori information on the behavior of the exact solution is available, one has to choose the function ψ in an adaptive way; i.e., starting with ψ_1, increase the polynomial degree of the function until the difference between the two successive majorants $M(\psi_{n-1})$ and $M(\psi_n)$ becomes small enough.

TABLE 2

Convergence of the exact modeling error in the energy norm ($|||e|||$) and of the error majorant (M) as $d_0 \to 0$; the results are rounded up to 10^{-4}.

d_0^{-1}	$m=2,\ M(\psi_1)$			$m=3,\ M(\psi_2)$		
	$\|\|\|e\|\|\|$	M	$\frac{M}{\|\|\|e\|\|\|}$	$\|\|\|e\|\|\|$	M	$\frac{M}{\|\|\|e\|\|\|}$
10^0	0.4405	0.9284	2.1074	0.2594	0.4187	1.6142
10^1	0.1291	0.1461	1.1265	0.0751	0.0793	1.0562
10^2	0.0408	0.0414	1.0127	0.0237	0.0239	1.0056
10^3	0.0129	0.0130	1.0013	0.0075	0.0076	1.0006
10^4	0.0041	0.0041	1.0001	0.0024	0.0024	1.0001

7. Conclusions. For the zero-order dimension reduction method, the new a posteriori error estimator has been derived in a general geometrical setting of the problem and without any specific assumptions on the given data. In particular, the estimator reduces to the Babuška–Schwab estimator when the physical domain Ω is a plate with plane parallel faces and the equation has zero right-hand side. It has been demonstrated, both theoretically and numerically, that also in a more complicated case of a plate having constant thickness but nonplane faces and for a general right-hand side $f \in L_\infty(\Omega)$ the proposed estimator vanishes with the optimal rate $\mathcal{O}(d_0^{1/2})$ as the plate thickness d_0 tends to zero. Since the estimator always provides an upper bound for the exact modeling error, the latter convergence result can be considered as the generalization of the result on the convergence of the dimension reduction error proved in [14] (see also [2]) for the case of a plate with plane parallel faces and zero right-hand side f.

The presented estimator cannot, however, be considered as just a generalization of the explicit residual-type error estimator to the case of more complicated geometry, coefficients, and right-hand side. As numerical test 2 shows, in the problem with the right-hand side f infinitely growing as the plate thickness tends to zero, some additional "degree of freedom" should be introduced into the estimator to suppress the unbounded growth of f. Thus, it seems that any error estimator that cannot be adjusted to the particular problem will fail in such a case. The proposed estimator is sufficiently flexible to allow the modification necessary for capturing the behavior of the exact error. The recovered efficiency of the estimator manifests itself in the asymptotics of the effectivity index $\frac{M}{|||e|||} = 1 + \mathcal{O}(d_0)$ when d_0 tends to zero. We have to note that such an asymptotics may not always be observed if the domain Ω has nonplane faces; however, even in the latter case, the effectivity index of the estimator remains stable (i.e., does not grow with the decreasing domain thickness) and stays at the acceptable level.

The computational cost of evaluating the presented error majorant is typically smaller than or, in the worst case, of the same order as the cost of solving the reduced, lower-dimensional problem. Finally, the numerical results show that the proposed estimator is capable of an accurate indication of the local error distribution and, hence, may be utilized not only for the verification of the dimensionally reduced model but also for its adaptive improvement.

REFERENCES

[1] M. AINSWORTH, *A posteriori error estimation for fully discrete hierarchic models of elliptic boundary value problems on thin domains*, Numer. Math., 80 (1998), pp. 325–362.

[2] I. BABUŠKA, I. LEE, AND C. SCHWAB, *On the a posteriori estimation of the modeling error for the heat conduction in a plate and its use for adaptive hierarchical modeling*, Appl. Numer. Math., 14 (1994), pp. 5–21.

[3] I. Babuška and C. Schwab, *A posteriori error estimation for hierarchic models of elliptic boundary value problems on thin domains*, SIAM J. Numer. Anal., 33 (1996), pp. 221–246.
[4] P. Ladeveze and D. Leguillon, *Error estimate procedure in the finite element method and applications*, SIAM J. Numer. Anal., 20 (1983), pp. 485–509.
[5] J. T. Oden and J. R. Cho, *Adaptive* hpq*-finite element methods of hierarchical models for plate- and shell-like structures*, Comput. Methods Appl. Mech. Engrg., 136 (1996), pp. 317–345.
[6] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math., 5 (1947), pp. 241–269.
[7] S. I. Repin, *A posteriori error estimation for variational problems with uniformly convex functionals*, Math. Comp., 69 (2000), pp. 481–600.
[8] S. I. Repin, *Estimates for errors in two-dimensional models of elasticity theory*, J. Math. Sci. (New York), 106 (2001), pp. 3027–3041.
[9] S. I. Repin, S. A. Sauter, and A. A. Smolianski, *A posteriori error estimation for the Dirichlet problem with account of the error in the approximation of boundary conditions*, Computing, 70 (2003), pp. 205–233.
[10] S. I. Repin, S. A. Sauter, and A. A. Smolianski, *A posteriori error estimation for the Poisson equation with mixed Dirichlet/Neumann boundary conditions*, J. Comput. Appl. Math., 164/165 (2004), pp. 601–612.
[11] S. I. Repin, S. A. Sauter, and A. A. Smolianski, *A posteriori estimation of dimension reduction errors*, in Proceedings of the 5th European Conference on Numerical Mathematics and Advanced Applications, (ENUMATH 2003), Springer-Verlag, New York, 2004, to appear.
[12] E. Stein and S. Ohnimus, *Coupled model- and solution-adaptivity in the finite-element method*, Comput. Methods Appl. Mech. Engrg., 150 (1997), pp. 327–350.
[13] J. L. Synge, *The Hypercircle in Mathematical Physics*, Cambridge University Press, Cambridge, UK, 1957.
[14] M.Vogelius and I. Babuška, *On a dimensional reduction method* I. *The optimal selection of basis functions*, Math. Comp., 37 (1981), pp. 31–46.
[15] M.Vogelius and I. Babuška, *On a dimensional reduction method* III. *A posteriori error estimation and an adaptive approach*, Math. Comp., 37 (1981), pp. 361–384.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1452–1478

ORDER OF CONVERGENCE ESTIMATES FOR AN EULER IMPLICIT, MIXED FINITE ELEMENT DISCRETIZATION OF RICHARDS' EQUATION*

FLORIN RADU†, IULIU SORIN POP‡, AND PETER KNABNER†

Abstract. We analyze a discretization method for a class of degenerate parabolic problems that includes the Richards' equation. This analysis applies to the pressure-based formulation and considers both variably and fully saturated regimes. To overcome the difficulties posed by the lack in regularity, we first apply the Kirchhoff transformation and then integrate the resulting equation in time. We state a conformal and a mixed variational formulation and prove their equivalence. This will be the underlying idea of our technique to get error estimates.

A regularization approach is combined with the Euler implicit scheme to achieve the time discretization. Again, equivalence between the two formulations is demonstrated for the semidiscrete case. The lowest order Raviart–Thomas mixed finite elements are employed for the discretization in space. Error estimates are obtained, showing that the scheme is convergent.

Key words. error estimates, Euler implicit scheme, mixed finite elements, regularization, degenerate parabolic problems, porous media, Richards' equation

AMS subject classifications. 65M12, 65M15, 65M60, 76S05, 35K65, 35K55

DOI. 10.1137/S0036142902405229

1. Introduction.

A commonly accepted mathematical model of water flow in porous media is the Richards' equation, a nonlinear, possibly degenerate, parabolic differential equation. In the pressure formulation, Richards' equation [5] is expressed as

$$\partial_t \Theta(\psi) - \nabla \cdot K(\Theta)\nabla(\psi + z) = 0, \tag{1.1}$$

where ψ is the pressure head, Θ the saturation, K the conductivity, and z the height against the gravitational direction. The equation (1.1) models the flow of a wetting fluid (water) in a porous media in the presence of a nonwetting fluid (air) supposed to be at constant pressure, 0. In the saturated region (where only water is present) we have $\psi \geq 0$, while $\psi < 0$ in the unsaturated domain. Different functional dependencies (retention curves) between ψ, K and Θ are proposed in the literature. These are provided essentially by soil particularities and allow reducing all the unknowns in the above equation to a single one. Here we are interested in both partially saturated and saturated flow, therefore we retain the pressure ψ as primary unknown.

As suggested in [1], applying the Kirchhoff transformation

$$\begin{aligned} \mathcal{K} : \mathbb{R} &\longrightarrow \mathbb{R}, \\ \psi &\longmapsto \int_0^{\psi} K(\Theta(s))\, ds \end{aligned} \tag{1.2}$$

*Received by the editors April 9, 2002; accepted for publication (in revised form) October 29, 2003; published electronically December 16, 2004. This work was supported by the Netherlands Organization for Scientific Research (NWO) through project 809.62.010 of Earth and Life Sciences (ALW).

http://www.siam.org/journals/sinum/42-4/40522.html

†Institute of Applied Mathematics, University Erlangen-Nürnberg, Martensstr. 3, D-91058 Erlangen, Germany (raduf@am.uni-erlangen.de, knabner@am.uni-erlangen.de).

‡Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (I.Pop@tue.nl).

leads to unknowns that are more regular. Since $K(\Theta(s))$ is positive, this transformation can be inverted and equation (1.1) can be rewritten in terms of a new variable, $u := \mathcal{K}(\psi)$. Now defining

$$\begin{aligned} b(u) &:= \Theta \circ \mathcal{K}^{-1}(u), \\ k(b(u)) &:= K \circ \Theta \circ \mathcal{K}^{-1}(u) \end{aligned} \tag{1.3}$$

and letting e_z denote the vertical unit vector, (1.1) becomes

$$\partial_t b(u) - \nabla \cdot (\nabla u + k(b(u))\ e_z) = 0 \quad \text{in} \quad (0,T] \times \Omega. \tag{1.4}$$

By the above transformation, diffusion becomes linear in equation (1.1). However, the problem may still remain degenerate, leading to solutions lacking regularity. Since this equation models important practical problems, several papers are dealing with analysis and numerical methods for it. Euler methods are often employed for the discretization in time. Adaptive time stepping is studied in [25], [14], or [28]. In case of an implicit discretization, iterative methods are considered (see, for example, [16], [8], whose method was already proposed in [12] and used also in [15], and [14]).

For the spatial discretization, mixed finite elements or finite volumes provide a good approximation of the solution [17], [4], [6], [10]. The most comprehensive algorithmic approach has been presented in the thesis [25], where hybrid mixed finite elements and an implicit Euler discretization are used. The set of nonlinear equations is solved by a Newton/multigrid method, while time and space adaptive strategies are constructed on the basis of rigorous error indicators. However, most of the authors are mainly interested in computational aspects and less concerned with rigorous convergence results. With respect to this last aspect we mention [2], where a model nonlinear, degenerate, advection-diffusion equation is considered. Through time integration a mixed variational formulation respecting the known minimal regularity of the solution is obtained. Raviart–Thomas lowest order finite elements are used. A priori error estimates are derived for the time integral of the flux and for the saturation. The estimates are optimal for the semidiscrete (continuous in time), noncomputable scheme. In the degenerate case, an explicit order of convergence for the fully discrete scheme can be deduced only by assuming extra (nonrealistic) regularity for the solution. Using similar techniques, [26] proved also some a priori error estimates for a mixed finite element discretization of Richards' equation. Unfortunately, again an explicit order of convergence for the scheme can be derived only in the nondegenerate case. Another important paper is [21], which deals with a class of multidimensional degenerate parabolic equations, including Richards' equation. A fully discrete scheme based on C^0 piecewise linear finite elements in space and a semi-implicit discretization in time is proposed and analyzed. An explicit order of convergence $(\tau^{1/2} + h)$ is proved. The techniques used here to cope with degenerate parabolic equations, which have been not used so far for discretizations based on mixed finite element method, will permit us to extend the results in [2] to the general, degenerate case. As in [26], error bounds for the time integral of the pressure will be also derived. We note also the recent paper [29], where the techniques from [2] are used for the numerical analysis of an expanded mixed finite element discretization of the Richards' equation. Also employed here are the Raviart–Thomas lowest order finite elements. Convergence rates depending on the Hölder continuity of the capacity term are derived for the entire regime of fully saturated to fully unsaturated flow. Nevertheless, the expanded mixed finite element method is not equivalent with the standard mixed finite element method and their results cannot be simply transferred to our method. Finally, we

mention also [10] (where convergence of an implicit finite volume method is proven by compactness arguments), [13] (for a relaxation scheme that applies to this equation too), and [23] (where error estimates are obtained for the unsaturated regime).

Here we consider an increasing and Lipschitz continuous b. Nevertheless, $b'(u)$ may be 0 for some values of u (not necessary isolated). Our numerical approach employs the lowest order Raviart–Thomas finite elements in space and Euler implicit in time, together with a regularization step. Specifically, with $N > 0$ integer, set $\tau = T/N$ and let $\mathcal{T}_h$ be a decomposition of Ω into closed d-simplices; h stands for the mesh-size. In a formal writing, the numerical scheme under consideration reads as

$$\begin{aligned} b_\epsilon(p_h^n) + \tau \nabla \cdot q_h^n &= b_\epsilon(p_h^{n-1}), \\ q_h^n + \nabla p_h^n + k(b(p_h^n))e_z &= 0 \end{aligned}$$

for $n = \overline{1, N}$; p_h^0 approximates u^0 in the finite dimensional approximation space. The term ∇p_h^n should be understood in a weak sense. Here b_ϵ is a regular approximation of b depending on the small parameter $\epsilon > 0$. By p_h^n we denote a piecewise constant approximation of u and q_h^n is a Raviart–Thomas (RT_0) approximation of the flux $-(\nabla u + k(b(u))e_z)$, based on $\mathcal{T}_h$, both at $t = n\tau$.

As suggested in [2], to overcome the difficulties posed by the lack in regularity, equation (1.4) is first integrated in time. For the resulting problem a mixed variational formulation is stated.

Convergence is shown by obtaining first error estimates for the time discrete scheme, by following the ideas in [21]. Since we work in a slightly more general framework, we include for completeness the proof for the conformal formulation. Next, using the procedure described in [2, 26], error estimates for the fully discrete scheme are obtained. In this setting, the equivalence between the mixed and conformal formulations becomes essential since, in this way, results obtained for one case can be transferred to the other one.

The outline of the paper is as follows. First, we state the main assumptions and notations used throughout the paper, define the problem to be solved, and discuss questions regarding existence and regularity of a solution. In section 2 the equivalence between a conformal and a mixed variational formulations is proved, for the continuous case as well as for the time discrete one. In section 3 we investigate the stability of the numerical scheme, while error estimates are derived in section 4.

1.1. Notations and assumptions. In what follows we let Ω be a domain in $\mathbb{R}^d$ (with $d = 1, 2,$ or 3). Let $J = (0, T]$ be a finite time interval. We are interested in solving (1.4) endowed with initial and boundary conditions,

$$\begin{aligned} \partial_t b(u) - \nabla \cdot (\nabla u + k(b(u))e_z) &= 0 \qquad \text{in} \quad J \times \Omega, \\ u &= u^0 \qquad \text{in} \quad 0 \times \Omega, \\ u &= 0 \qquad \text{on} \quad J \times \Gamma. \end{aligned} \tag{1.5}$$

Throughout this paper we make use of the following assumptions.

(A1) $\Omega \subset \mathbb{R}^d$ is bounded with Lipschitz continuous boundary.

(A2) $b \in C^1$ is nondecreasing and Lipschitz continuous.

(A3) $k(b(z))$ is continuous and bounded in z and satisfies, for all $z_1, z_2 \in \mathbb{R}$, $| k(b(z_2)) - k(b(z_1)) |^2 \leq C_k(b(z_2) - b(z_1))(z_2 - z_1)$.

(A4) $b(u_0)$ is essentially bounded (by 0 and 1) in Ω and $u_0 \in L^2(\Omega)$.

REMARK 1.1. *By* (A3)*, the convection term is bounded. This restriction is not unrealistic since, for Richards' equation, k stands for the conductivity of the medium.*

This assumption makes our analysis easier, but can be avoided. Moreover, the growth condition on $k(b(\cdot))$ (see also [11], [27], [30], *or* [23]*) relaxes the more often assumed Lipschitz continuity of k (see, e.g.,* [21], [2]*). It gives uniqueness for the weak solution, as shown in* [1]*. In addition, source terms can also be considered here, provided that they satisfy a similar growth condition as $k(b(u))$.*

REMARK 1.2. *In the transformed version, Richards' equation fits in our framework. However, since b is Lipschitz, a vanishing permeability in* (1.1) *is not allowed, meaning that our analysis is valid in the variably saturated to fully saturated flow regimes, but not in the completely air saturated one.*

REMARK 1.3. *For the sake of simplicity, we deal with homogeneous Dirichlet boundary conditions. More general situations can be included in a straightforward manner, with similar results. Here nonlinearities depend only on the unknown u, not on x and t. For more general situations, techniques developed in* [2] *can be employed.*

Because of its degenerate character, we do not expect smooth solutions for problem (1.5). For defining a solution in a weak sense we let $(\cdot,\cdot)$ stand for the inner product on $L^2(\Omega)$ or the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, $\|\cdot\|$ for the norm in $L^2(\Omega)$, and $\|\cdot\|_1$ and $\|\cdot\|_{-1}$ for the norms in $H^1(\Omega)$ and $H^{-1}(\Omega)$, respectively. We use analogous notations for the inner product and the corresponding norm on $L^2(J;\mathcal{H})$, with $\mathcal{H}$ being either $L^2(\Omega)$, $H^1(\Omega)$, or $H^{-1}(\Omega)$. In addition, we often write u or $u(t)$ instead of $u(t,x)$ and use C to denote a generic positive constant, not depending on the discretization or regularization parameters.

A weak solution for problem (1.5) is defined as follows.

DEFINITION 1.4. *A function u is called a weak solution for equation* (1.5) *iff $b(u) \in H^1(J;H^{-1}(\Omega))$, $u \in L^2(J;H_0^1(\Omega))$, $u(0) = u_0$ (in H^{-1} sense), and for all $\varphi \in L^2(J;H_0^1(\Omega))$ it holds that*

$$\int_0^T (\partial_t b(u(t)),\varphi(t)) \;+\; (\nabla u(t) + k(b(u(t)))e_z, \nabla\varphi(t))dt = 0. \tag{1.6}$$

Existence, uniqueness, and essential bounds for a weak solution of the above problem are studied in several papers (see, for example, [1], [22], [27] and the references therein). In [1] the following regularity result is obtained:

$$b(u) \in L^\infty(J;L^1(\Omega)), \tag{1.7}$$

$$q := -(\nabla u + k(b(u))e_z) \in L^2(J;(L^2(\Omega))^d). \tag{1.8}$$

Here $b(u)$ models the water content, hence it is natural to assume that, after scaling, it lies between 0 and 1 for almost every $(t,x) \in J \times \Omega$. For the same reason, in (A4) similar restrictions are imposed to the initial data. Such essential estimates can be shown, for example, if b and k do not depend explicitly on x, or if $k(b(u))$ is constant for $u = 0$ and $u = 1$. Moreover, $u \in L^2(J;H_0^1(\Omega))$ yields $b(u) \in L^2(J;H_0^1(\Omega))$ due to the Lipschitz continuity of b. Since $b(u) \in H^1(J;H^{-1}(\Omega))$ we have $b(u) \in C([0,T];L^2(\Omega))$ (see [20, Chapter I]), allowing a simplified mixed variational formulation. Following [2] or [29] we integrate (1.5) in time and obtain, for every $t \in J$,

$$b(u(t)) + \nabla \cdot \int_0^t q\ (s)\, ds = b(u^0) \tag{1.9}$$

in L^2 sense. It follows (see [2] or [25]) that the flux $\vec{q}$ defined in (1.8) satisfies

$$\int_0^t q\, d\tau \in H^1(J;(L^2(\Omega))^d) \cap L^2(J;(H^1(\Omega))^d) =: X. \tag{1.10}$$

2. Equivalent formulations. In this section we give the mixed variational formulations and study the equivalence with the conformal ones in both continuous and time discrete cases.

2.1. The continuous case. Integrated in time, problem (1.5) becomes the following.

Problem 1. Find $u \in L^2(J, H_0^1(\Omega))$ such that $b(u) \in L^\infty(J \times \Omega)$, and for all $t \in J$ and $\phi \in H_0^1(\Omega)$ it holds that

$$(b(u(t)) - b(u^0), \phi) + \int_0^t (\nabla u(s) + k(b(u(s)))e_z, \nabla\phi)ds = 0. \tag{2.1}$$

As mentioned in the previous section, this stronger formulation makes sense since $b(u) \in C(J; L^2(\Omega))$.

A mixed formulation for problem (1.5) reads as follows.

Problem 2. Find $(p, \tilde{q}) \in L^2(J \times \Omega)) \times X$ such that $b(p) \in L^\infty(J \times \Omega)$ and for all $t \in J$ the equations

$$(b(p(t)) - b(p^0), w) + (\nabla \cdot \tilde{q}(t), w) = 0, \tag{2.2}$$

$$(\tilde{q}(t), v) - \int_0^t (p(s), \nabla v)ds + \int_0^t (k(b(p(s)))e_z, v)ds = 0 \tag{2.3}$$

hold for all $w \in L^2(\Omega)$ and $v \in H(\mathrm{div}, \Omega)$, with $p^0 = u^0 \in L^2(\Omega)$.

The two problems are equivalent, as shown in Proposition 2.2. In the proof we use the following lemma [7, p. 91].

LEMMA 2.1. *Let $v \in H(\mathrm{div}, \Omega)$ and $\vec{n}$ denote the outer normal to Γ. Then $v \cdot \vec{n}$ is defined in $H^{-1/2}(\Gamma)$ (in the sense of traces) and Green's formula applies for all $p \in H^1(\Omega)$*

$$\int_\Omega \nabla \cdot v \; p \; dx + \int_\Omega v \cdot \; \nabla p \; dx = \int_\Gamma \vec{n} \cdot v \; p \; ds. \tag{2.4}$$

PROPOSITION 2.2. *$u \in L^2(J, H_0^1(\Omega))$ solves Problem 1 iff $(p, \tilde{q}) \in L^2(J \times \Omega)) \times X$ defined as*

$$(p, \tilde{q}) = \left(u, -\int_0^t (\nabla u(s) + k(b(u(s)))e_z)ds\right) \tag{2.5}$$

solves Problem 2. Moreover, in this case we have $p \in L^2(J, H_0^1(\Omega))$.

Proof. We use some ideas from [18].

"$\Rightarrow$" Let $u \in L^2(J, H_0^1(\Omega))$ be a solution of Problem 1 and $(p, \tilde{q})$ defined in (2.5). By (1.10) we have $(p, \tilde{q}) \in L^2(J, H_0^1(\Omega)) \times X$. Fixing now $t > 0$, for any $v \in H(\mathrm{div}, \Omega)$, using Green's formula we get

$$\begin{aligned}(\tilde{q}(t), v) &= -\int_0^t (\nabla u(s) + k(b(u(s)))e_z, v)ds \\ &= \int_0^t (p(s), \nabla v) - (k(b(p(s)))e_z, v)ds,\end{aligned}$$

so (2.3) is proven.

Next, taking any $\phi \in C_0^\infty(\Omega)$ in (2.1) yields

$$(b(u(t)) - b(u^0), \phi) = -\left(\int_0^t (\nabla u(s) + k(b(u(s)))e_z)ds, \nabla\phi\right)$$
$$= (\tilde{q}(t), \nabla\phi) = -(\nabla \cdot \tilde{q}(t), \phi).$$

However, for any $t > 0$, both $b(u(t)) - b(u^0)$ and $\nabla \cdot \tilde{q}(t)$ lie in $L^2(\Omega)$, so the above relations still hold for $\phi \in L^2(\Omega)$, implying (2.2).

"$\Leftarrow$" Let $(p, \tilde{q}) \in L^2(J \times \Omega) \times X$ solving Problem 2 and set $u = p \in L^2(J \times \Omega)$. Taking $v \in (C_0^\infty(\Omega))^d \subset H(\mathrm{div}, \Omega)$ arbitrary, by differentiating (2.3) we get for almost all $t > 0$

$$(\partial_t \tilde{q}(t), v) + (k(b(p(t)))e_z, v) = (p(t), \nabla \cdot v) = -(\nabla p(t), v), \tag{2.6}$$

so $\nabla p = -\partial_t \tilde{q} - k(b(p))e_z$ in a distributional sense. Since both $\partial_t \tilde{q}$ and $k(b(p))e_z$ are in $L^2(J \times \Omega)$, the same holds for ∇p, so $u = p \in L^2(J, H^1(\Omega))$.

Taking now $v \in H(\mathrm{div}, \Omega)$ in (2.3) gives, for every $t \in J$,

$$-\int_0^t (\nabla p, v) \overset{(2.6)}{=} (\tilde{q}(t), v) + \int_0^t (k(b(p))e_z, v) \overset{(2.3)}{=} \int_0^t (p, \nabla \cdot v).$$

In this way, using (2.4) we get

$$\int_0^t \int_\Gamma pv \cdot \vec{n} ds = \int_0^t (\nabla p, v) + \int_0^t (p, \nabla \cdot v) = 0.$$

Here v was chosen arbitrary, so the trace of p on Γ is zero. Thus $p \in L^2(J, H_0^1(\Omega))$ and the same holds for u.

Moreover, taking any $\phi \in H_0^1(\Omega)$ yields, for all $t > 0$,

$$\big(b(u(t)) - b(u^0), \phi\big) \overset{(2.2)}{=} -(\nabla \cdot \tilde{q}(t), \phi) = (\tilde{q}(t), \nabla\phi)$$
$$\overset{(2.3)}{=} -\int_0^t (\nabla u(s) + k(b(u(s)))e_z, \nabla\phi)ds,$$

so u solves (2.1). ∎

2.2. The semidiscrete case. As mentioned in the introduction, for overcoming difficulties due to degeneracy, we first perturb the original equation to obtain a regular parabolic one. Such a technique has been successfully applied in the analysis of degenerate problems and also allows developing effective numerical schemes (see, e.g., [21]).

In problem (1.5) degeneracy appears due to the vanishing of b'. Therefore we approximate this nonlinearity by b_ϵ, with $\epsilon > 0$ a small perturbation parameter. A possible choice reads as

$$b_\epsilon(u) = b(u) + \epsilon u. \tag{2.7}$$

Obviously, b_ϵ is Lipschitz continuous (with the same Lipschitz constant as b, if ϵ is small enough), strictly increasing and its derivative is bounded from below by ϵ. The regularized problem becomes

$$\begin{aligned} \partial_t b_\epsilon(u) - \nabla \cdot (\nabla u + k(b(u))e_z) &= 0 \quad && \text{in} \quad (0;T] \times \Omega, \\ u &= u^0 \quad && \text{in} \quad \Omega, \\ u &= 0 \quad && \text{on} \quad J \times \Gamma. \end{aligned} \tag{2.8}$$

We let $N > 1$ be an integer giving a time step $\tau = T/N$, with $t_n = n\tau$. The regularized semidiscrete conformal problem reads

Problem 3. Let $n = \overline{1, N}$ and u^{n-1} be given. Find $u^n \in H_0^1(\Omega)$ such that, for all $\phi \in H_0^1(\Omega)$,

$$(b_\epsilon(u^n) - b_\epsilon(u^{n-1}), \phi) + \tau(\nabla u^n + k(b(u^n))e_z, \nabla\phi) = 0. \tag{2.9}$$

However, our final aim is a mixed discretization. The time discrete regularized mixed problem becomes the following.

Problem 4. Let $n = \overline{1, N}$ and p^{n-1} given. Find $(p^n, q^n) \in L^2(\Omega) \times H(\mathrm{div}, \Omega)$ such that

$$(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), w) + \tau(\nabla \cdot q^n, w) = 0, \tag{2.10}$$

$$(q^n, v) - (p^n, \nabla v) + (k(b(p^n))e_z, v) = 0, \tag{2.11}$$

for all $w \in L^2(\Omega)$, respectively, $v \in H(\mathrm{div}, \Omega)$, with $p^0 = u^0 \in L^2(\Omega)$.

As in the continuous case, the two problems above are equivalent.

PROPOSITION 2.3. *Let $n = \overline{1, N}$ be fixed and assume $u^{n-1} = p^{n-1}$. Then $u^n \in H_0^1(\Omega)$ solves Problem* 3 *iff $(p^n, q^n) \in L^2(\Omega) \times H(\mathrm{div}, \Omega)$ defined as*

$$(p^n, q^n) = (u^n, -(\nabla u^n + k(b(u^n))e_z)) \tag{2.12}$$

solve Problem 4. *Moreover, we have $p^n \in H_0^1(\Omega)$.*

Proof. "⇒" Let $u^n \in H_0^1(\Omega)$ be a solution of Problem 3 and (p^n, q^n) be defined in (2.12). For all $v \in H(\mathrm{div}, \Omega)$ we have

$$(q^n, v) = -(\nabla u^n + k(b(u^n))e_z, v) = (p^n, \nabla v) - (k(b(p^n))e_z, v),$$

so (p^n, q^n) verify (2.11).

Next, for all $\phi \in C_0^\infty(\Omega)$ (which is dense in $H_0^1(\Omega)$) we get

$$\begin{aligned}(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), \phi) &\overset{(2.9)}{=} -\tau(\nabla u^n + k(b(u^n))e_z, \nabla\phi) = \tau(q^n, \nabla\phi)\\ &= -\tau(\nabla \cdot q^n, \phi).\end{aligned}$$

But $b_\epsilon(p^n) - b_\epsilon(p^{n-1}) \in L^2(\Omega)$, so $\nabla \cdot q^n \in L^2(\Omega)$, implying $q^n \in H(\mathrm{div}, \Omega)$ and that (2.10) holds by density arguments.

"⇐" Let $(p^n, q^n) \in L^2(\Omega) \times H(\mathrm{div}, \Omega)$ be a solution of Problem 4 and $u^n = p^n \in L^2(\Omega)$. For any $v \in (C_0^\infty(\Omega))^d \subset H(\mathrm{div}, \Omega)$ we have

$$\begin{aligned}(q^n, v) &\overset{(2.11)}{=} (p^n, \nabla v) - (k(b(p^n))e_z, v)\\ &= -(\nabla p^n, v) - (k(b(p^n))e_z, v),\end{aligned}$$

implying

$$\nabla p^n + k(b(p^n))e_z = -q^n$$

in distributional sense. Since both q^n and $k(b(p^n))$ are $L^2(\Omega)$ functions it follows that $p^n \in H^1(\Omega)$. As for the continuous case, using Green's formula (2.4), we get actually $u^n = p^n \in H_0^1(\Omega)$.

Finally, (2.9) results by taking any $\phi \in H_0^1(\Omega)$ in (2.10),

$$\begin{aligned}(b_\epsilon(u^n) - b_\epsilon(u^{n-1}), \phi) &= -\tau(\nabla \cdot q^n, \phi) = \tau(q^n, \nabla\phi)\\ &= -\tau((\nabla p^n + k(b(p^n))e_z), \nabla\phi).\end{aligned}$$

□

As resulting from the equivalencies proven above, stability and error estimates for the time discrete mixed formulation can be obtained by analyzing the Euler implicit scheme applied to Problem 3. This is the underlying idea in the forthcoming section.

3. Stability estimates. In this section we investigate the stability of our numerical approach. We make use of the lemmas below.

LEMMA 3.1. *For any vectors* $a_k, b_k \in \mathbb{R}^q$ $(k = \overline{1,N}, q \geq 1)$ *we have*

$$2\sum_{n=1}^{N} a_n \sum_{k=1}^{n} a_k = \left(\sum_{n=1}^{N} a_n\right)^2 + \sum_{n=1}^{N} (a_n)^2, \tag{3.1}$$

$$2\sum_{n=1}^{N} (a_n - a_{n-1}, a_n) = |a_N|^2 - |a_0|^2 + \sum_{n=1}^{N} |a_n - a_{n-1}|^2, \tag{3.2}$$

$$\sum_{n=1}^{N} (a_n - a_{n-1}, b_n) = a_N b_N - a_0 b_0 - \sum_{n=1}^{N} (b_n - b_{n-1}, a_{n-1}). \tag{3.3}$$

LEMMA 3.2. *Under the assumption* (A1)*, for any real sequence* x^j*,* $j = \overline{1,n}$ *we have*

$$\sum_{j=1}^{n} (b_\epsilon(x^j) - b_\epsilon(x^{j-1}))\ x^j \geq -C|x^0|^2 + \frac{\epsilon}{2}|x^n|^2. \tag{3.4}$$

Proof. Since $b'_\epsilon \geq \epsilon$, one has, for any reals x and y,

$$((b_\epsilon(x) - b_\epsilon(y))x \geq \int_y^x s b'_\epsilon ds \quad \text{and} \quad \int_0^x s b'_\epsilon(s) ds \geq \frac{\epsilon}{2} x^2.$$

Furthermore,

$$\begin{aligned}\sum_{j=1}^{n} (b_\epsilon(x^j) - b_\epsilon(x^{j-1}))x^j &\geq \sum_{j=1}^{n} \int_{x^{j-1}}^{x^j} s b'_\epsilon(s) ds \\ &= \int_0^{x^n} s b'_\epsilon(s) ds - \int_0^{x^0} s b'_\epsilon(s) ds \geq -C|x^0|^2 + \frac{\epsilon}{2}|x^n|^2,\end{aligned}$$

where the constant C is half of the Lipschitz constant of b. ∎

3.1. Stability in the time discrete conformal case.

PROPOSITION 3.3. *Assume* (A1)–(A4)*. If* u^n *solves Problem* 3 *(*$n = \overline{1,N}$*), we have*

$$\tau \sum_{n=1}^{N} \|u^n\|_1^2 \leq C. \tag{3.5}$$

Proof. Taking $\phi = u^n$ in (2.9) and summing up for $n = \overline{1,N}$ give

$$\sum_{n=1}^{N} \left(b_\epsilon(u^n) - b_\epsilon(u^{n-1}), u^n\right) + \sum_{n=1}^{N} \tau \|\nabla u^n\|^2 + \sum_{n=1}^{N} \tau \left(k(b(u^n))e_z, \nabla u^n\right) = 0. \tag{3.6}$$

Now we estimate the terms on the left in the above. By (3.4), since $u^0 \in L^2(\Omega)$,

$$\sum_{n=1}^{N} \left(b_\epsilon(u^n) - b_\epsilon(u^{n-1}), u^n\right) \geq -C.$$

The second term needs no further treatment. Finally, since k is bounded, applying the Cauchy–Schwarz inequality, we get

$$\tau \sum_{n=1}^{N} |(k(b(u^n))e_z, \nabla u^n)| \leq \frac{\tau}{2} \sum_{n=1}^{N} \|k(b(u^n))e_z\|^2 + \frac{\tau}{2} \sum_{n=1}^{N} \|\nabla u^n\|^2$$

$$\leq C + \frac{\tau}{2} \sum_{n=1}^{N} \|\nabla u^n\|^2 .$$

Inserting the last inequalities into (3.6) and using the inequality of Poincaré gives (3.5). ∎

3.2. Stability for the time discrete mixed formulation. By the equivalence of Problems 3 and 4, Proposition 3.3 provides stability for the time discrete solutions p^n and q^n.

PROPOSITION 3.4. *Assuming* (A1)–(A4)*, if, for any* $n = \overline{1, N}$*,* (p^n, q^n) *solve Problem* 4*, we have*

$$\tau \sum_{n=1}^{N} \|p^n\|_1^2 + \tau \sum_{n=1}^{N} \|q^n\|^2 \leq C. \tag{3.7}$$

Proof. The estimate for p^n is a direct consequence of (3.5). Next, taking $w = p^n$ in (2.10) and $v = \tau q^n$ in (2.11) yields

$$(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n) + \tau(\nabla \cdot q^n, p^n) = 0,$$
$$(q^n, \tau q^n) - (p^n, \tau \nabla \cdot q^n) + (k(b(p^n))e_z, \tau q^n) = 0.$$

Adding these two equations and summing up for $n = 1$ to N give

$$\sum_{n=1}^{N} (b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n) + \tau \sum_{n=1}^{N} \|q^n\|^2 + \tau \sum_{n=1}^{N} (k(b(p^n))e_z, q^n) = 0,$$

and the rest of the proof follows exactly as in the previous proposition. ∎

Other stability estimates can be obtained defining an initial flux $q^0 \in [L^2(\Omega)]^d$. In doing so we take $\rho \in C_0^\infty(B_d(0,1))$ ($B_d(0,1)$ being the unit ball in $\mathbb{R}^d$) so that $\int_{B_d(0,1)} \rho(x)dx = 1$ and consider the mollifier sequence $\{\rho_\mu(x) = \frac{1}{\mu^d}\rho(\frac{x}{\mu})\}_{1>>\mu>0}$. Defining q^0 as

$$q^0 = -\nabla(\rho_\mu * p^0) - k(b(p^0))e_z, \tag{3.8}$$

with μ to be chosen further and $*$ denoting the convolution operator, for any $v \in H(\mathrm{div}, \Omega)$ we have

$$(q^0, v) - (\rho_\mu * p^0, \nabla \cdot v) + (k(b(p^0))e_z, v) = 0. \tag{3.9}$$

A mollifying of p^0 in the above is necessary for having $q^0 \in [L^2(\Omega)]^d$. However, since $p^0 \in L^2(\Omega)$, $\|p^0 - \rho_\mu * p^0\|$ goes to 0 as $\mu \searrow 0$, so $\|q^0\|$ is uniformly bounded with respect to μ. Now the following estimates can be obtained.

PROPOSITION 3.5. *Assuming* (A1)–(A4)*, if, for all* $n = \overline{1, N}$*,* (p^n, q^n) *solve Problem* 4*, for any* $k > 0$ *we have*

$$\sum_{n=1}^{k} (b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n - p^{n-1}) + \tau\|q^k\|^2 + \tau \sum_{n=1}^{k} \|q^n - q^{n-1}\|^2 \leq C\tau. \tag{3.10}$$

Proof. First we take $w = p^n - p^{n-1} \in L^2(\Omega)$ in (2.10) and subtract equation (2.11) at time step $n-1$ from the one at time step n. Testing with $v = \tau q^n$ in the resulting equality yields

$$(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n - p^{n-1}) + \tau(\nabla \cdot q^n, p^n - p^{n-1}) = 0,$$
$$\tau(q^n - q^{n-1}, q^n) - \tau(p^n - p^{n-1}, \nabla \cdot q^n) + \tau((k(b(p^n)) - k(b(p^{n-1})))e_z, q^n) = 0.$$

For $n = 1$ the second equation above reads as

$$\tau(q^1 - q^0, q^1) - \tau(p^1 - p^0, \nabla \cdot q^1) + \tau((k(b(p^1)) - k(b(p^0)))e_z, q^1) = \tau(p^0 - \rho_\mu * p^0, \nabla \cdot q^1).$$

Adding the above pairs of equalities and summing the result up for $n = \overline{1,k}$ yields

$$\begin{aligned} &\sum_{n=1}^{k}(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n - p^{n-1}) + \tau\sum_{n=1}^{k}(q^n - q^{n-1}, q^n) \\ &\quad + \tau\sum_{n=1}^{k}((k(b(p^n)) - k(b(p^{n-1})))e_z, q^n) = \tau(p^0 - \rho_\mu * p^0, \nabla \cdot q^1). \end{aligned} \tag{3.11}$$

Denoting the terms above by $T_1, \dots, T_4$, we first notice that T_1 is positive by the monotonicity of b_ϵ. Next, by (3.2),

$$\begin{aligned} T_2 &= \tau\sum_{n=1}^{k}(q^n - q^{n-1}, q^n) \\ &= \frac{\tau}{2}\|q^k\|^2 - \frac{\tau}{2}\|q^0\|^2 + \frac{\tau}{2}\sum_{n=1}^{k}\|q^n - q^{n-1}\|^2. \end{aligned}$$

Recalling (A3) and the Cauchy–Schwarz inequality, for T_3 we get

$$\begin{aligned} |T_3| &\le \frac{\delta_1}{2}\sum_{n=1}^{k}\|(k(b(p^n)) - k(b(p^{n-1})))e_z\|^2 + \frac{\tau^2}{2\delta_1}\sum_{n=1}^{k}\|q^n\|^2 \\ &\le \frac{\delta_1 C_k}{2}\sum_{n=1}^{k}(b(p^n) - b(p^{n-1}), p^n - p^{n-1}) + \frac{\tau^2}{2\delta_1}\sum_{n=1}^{k}\|q^n\|^2. \end{aligned}$$

Estimating T_4 follows as before,

$$|T_4| \le \tau\|p^0 - \rho_\mu * p^0\|\|\nabla \cdot q^1\| \le \delta_2\|p^0 - \rho_\mu * p^0\|^2 + \frac{\tau^2}{4\delta_2}\|\nabla \cdot q^1\|^2.$$

To estimate $\|\nabla \cdot q^1\|$ we use (2.10) for $n = 1$, test with $w = \nabla \cdot q^1 \in L^2(\Omega)$ and obtain

$$\tau\|\nabla \cdot q^1\|^2 \le \|b_\epsilon(p^1) - b_\epsilon(p^0)\|\|\nabla \cdot q^1\| \le \frac{C}{2\tau}\left(b_\epsilon(p^1) - b_\epsilon(p^0), p^1 - p^0\right) + \frac{\tau}{2}\|\nabla \cdot q^1\|^2$$

by the Lipschitz continuity of b_ϵ. In this way we get

$$\tau\|\nabla \cdot q^1\|^2 \le \tfrac{C}{\tau}(b_\epsilon(p^1) - b_\epsilon(p^0), p^1 - p^0).$$

Using these estimates in (3.11) and choosing the δ's properly give

$$\begin{aligned} &\sum_{n=1}^{k}(b_\epsilon(p^n) - b_\epsilon(p^{n-1}), p^n - p^{n-1}) + \tau\|q^k\|^2 + \tau\sum_{n=1}^{k}\|q^n - q^{n-1}\|^2 \\ &\quad \le C_1\tau + C_2\|p^0 - \rho_\mu * p^0\|^2 + C_3\tau^2\sum_{n=1}^{k}\|q^n\|^2. \end{aligned}$$

We still have to choose μ in (3.8). Since $\|p^0 - \rho_\mu * p^0\|$ converges to 0, taking μ sufficiently small, the right term in the above becomes

$$C_4\tau + C_3\tau^2 \sum_{n=1}^{k} \|q^n\|^2.$$

Now (3.10) follows by the discrete Gronwall lemma. ∎

REMARK 3.6. *If $p^0 \in H^1(\Omega)$, q^0 can be defined without using a mollifier,*

$$q^0 = -\nabla p^0 - k(b(p^0))e_z. \tag{3.12}$$

Then $T_4 = 0$ in (3.11)*, without changing* (3.10).

A direct consequence of the stability estimates above follows.

PROPOSITION 3.7. *In the setting of Proposition* 3.5 *we have*

$$\sum_{n=1}^{N} \tau\|\nabla \cdot q^n\|^2 \le C. \tag{3.13}$$

Proof. Taking $w = \nabla \cdot q^j$ in equation (2.10) and applying the Cauchy–Schwarz inequality one gets

$$\tau\|\nabla \cdot q^j\|^2 \le \frac{1}{2\tau}\|b_\epsilon(p^j) - b_\epsilon(p^{j-1})\|^2 + \frac{\tau}{2}\|\nabla \cdot q^j\|^2,$$

so

$$\tau\|\nabla \cdot q^j\|^2 \le \frac{1}{\tau}\|b_\epsilon(p^j) - b_\epsilon(p^{j-1})\|^2.$$

Summing up the above for $j = \overline{1,N}$, using the Lipschitz continuity of b_ϵ and (3.10) leads to (3.13). ∎

4. Error estimates. In this section we obtain a priori error estimates for both time discrete scheme, as well as for the fully discrete one.

4.1. Error estimates for the semidiscrete approximation. To obtain error estimates for the time discrete scheme we employ techniques developed in [21] and make use of the Green operator $G : H^{-1}(\Omega) \to H_0^1(\Omega)$ defined as

$$(\nabla(G\psi), \nabla\phi) = (\psi, \phi) \quad \text{for all} \quad \phi \in H_0^1(\Omega). \tag{4.1}$$

Obviously, G is linear and self-adjoint. Moreover, by the Cauchy–Schwarz inequality, using (3.3) yields the following lemma.

LEMMA 4.1. *For all $f, f_k \in H^{-1}(\Omega)$ $(k = \overline{1,N})$ and $g \in H^1(\Omega)$ we have*

$$(f, g) \le \|f\|_{-1}\|\nabla g\|,$$

$$\|\nabla Gf\|^2 = (f, Gf) = \|f\|_{-1}^2,$$

$$2\sum_{k=1}^{N}(f_k - f_{k-1}, Gf_k) = \|f_k\|_{-1,\Omega}^2 - \|f_0\|_{-1,\Omega}^2 + \sum_{k=1}^{N}\|f_k - f_{k-1}\|_{-1,\Omega}^2.$$

Further, we use the notations

$$\begin{aligned} \overline{u}^n &= \frac{1}{\tau}\int_{t_{n-1}}^{t_n} u(t)dt, \\ u_\Delta(t) &= u^n \quad \text{for} \quad t \in (t_{n-1}, t_n], \\ e_b(u) &= b(u) - b_\epsilon(u_\Delta), \end{aligned} \tag{4.2}$$

where $n = \overline{1, N}$ and $\overline{u}^0 = u^0$.

It is worth pointing out here that, by Propositions 2.2 and 2.3, estimates obtained for the conformal discretization can be transferred to the mixed case.

PROPOSITION 4.2. *Assuming* (A1)–(A4)*, if* u *is the weak solution of Problem* 1 *and* u^n *solves, for each* $n = \overline{1, N}$*, Problem* 3*, then*

$$\begin{aligned} \max_{n=\overline{1,N}} \left\| \overline{e_b(u)}^n \right\|_{-1}^2 + \|e_b(u)\|_{L^2(J\times\Omega)}^2 + \int_0^T (b_\epsilon(u(t)) - b_\epsilon(u_\Delta), u(t) - u_\Delta)dt \\ \le C\,(\tau + \epsilon). \end{aligned} \tag{4.3}$$

Proof. Subtracting (2.1) at $t = t_{j-1}$ from the one at $t = t_j$ and then subtracting (2.9) with $n = j$ from the result give

$$\begin{aligned} &(b(u(t_j)) - b(u(t_{j-1})) - b_\epsilon(u^j) + b_\epsilon(u^{j-1}), \phi) \\ &\quad + \tau(\nabla(\overline{u}^j - u^j), \nabla\phi) + \tau((\overline{k(b(u))}^j - k(b(u^j)))e_z, \nabla\phi) = 0. \end{aligned}$$

Taking $\phi = G\overline{e_b(u)}^j \in H_0^1(\Omega)$ into above and summing up for $j = \overline{1, n}$ (with $n \le N$) yield

$$\begin{aligned} &\sum_{j=1}^n (b(u(t_j)) - b(u(t_{j-1})) - b_\epsilon(u^j) + b_\epsilon(u^{j-1}), G\overline{e_b(u)}^j) \\ &\quad + \sum_{j=1}^n \tau(\nabla\overline{u}^j - \nabla u^j, \nabla G\overline{e_b(u)}^j) \\ &\quad + \sum_{j=1}^n \tau((\overline{k(b(u))}^j - k(b(u^j)))e_z, \nabla G\overline{e_b(u)}^j) = 0. \end{aligned} \tag{4.4}$$

We estimate now each of terms in (4.4), denoted by T_1, T_2, and T_3:

$$\begin{aligned} T_1 &= \sum_{j=1}^n (b(u(t_j)) - \overline{b(u)}^j - b(u(t_{j-1})) + \overline{b(u)}^{j-1}, G\overline{e_b(u)}^j) \\ &\quad + \sum_{j=1}^n (\overline{b(u)}^j - b_\epsilon(u^j) - \overline{b(u)}^{j-1} + b_\epsilon(u^{j-1}), G\overline{e_b(u)}^j) \\ &=: T_{11} + T_{12}. \end{aligned}$$

Further, by (3.3) and recalling that $b(u(0)) = \overline{b(u)}^0$ we have

$$\begin{aligned} T_{11} &= \sum_{j=1}^n (b(u(t_j)) - \overline{b(u)}^j - b(u(t_{j-1})) + \overline{b(u)}^{j-1}, G\overline{e_b(u)}^j) \\ &= (b(u(t_n)) - \overline{b(u)}^n, G\overline{e_b(u)}^n) \end{aligned}$$

$$
\begin{aligned}
&-\sum_{j=1}^{n}(b(u(t_{j-1}))-\overline{b(u)}^{j-1},G\overline{e_b(u)}^{j}-G\overline{e_b(u)}^{j-1})\\
&=: T_{111}-T_{112}.
\end{aligned}
$$

For T_{111} we make use of Lemma 4.1 and obtain

$$
\begin{aligned}
|T_{111}| &\le \frac{1}{\tau}\int_{t_{n-1}}^{t_n}|(b(u(t_n))-b(u(t)),G\overline{e_b(u)}^{n})|dt\\
&\le \frac{1}{\tau}\int_{t_{n-1}}^{t_n}\int_{t}^{t_n}|(\partial_s b(u(s)),G\overline{e_b(u)}^{n})|dsdt\\
&\le \frac{1}{\tau}\int_{t_{n-1}}^{t_n}\sqrt{\tau}\,\|\partial_t b(u)\|_{L^2(t_{n-1},t_n;H^{-1})}\,\|\overline{e_b(u)}^{n}\|_{-1}dt \qquad (4.5)\\
&\le \sqrt{\tau}\,\|\partial_s b(u)\|_{L^2(t_{n-1},t_n;H^{-1})}\,\|\overline{e_b(u)}^{n}\|_{-1}\\
&\le \tau\,\|\partial_s b(u)\|^2_{L^2(t_{n-1},t_n;H^{-1})}+\frac{1}{4}\|\overline{e_b(u)}^{n}\|^2_{-1}.
\end{aligned}
$$

Proceeding as before, T_{112} can be estimated as

$$
|T_{112}|\le \tau\,\|\partial_t b(u)\|^2_{L^2(0,t_n;H^{-1})}+\frac{1}{4}\sum_{j=1}^{n}\|\overline{e_b(u)}^{j}-\overline{e_b(u)}^{j-1}\|^2_{-1}. \qquad (4.6)
$$

Using Lemma 4.1 again, since $\overline{e_b(u)}^{0}=0$, T_{12} gives

$$
\begin{aligned}
T_{12} &= \sum_{j=1}^{n}(\overline{b(u)}^{j}-b_\epsilon(u^j)-\overline{b(u)}^{j-1}+b_\epsilon(u^{j-1}),G\overline{e_b(u)}^{j}))\\
&= \frac{1}{2}(\overline{e_b(u)}^{n},G\overline{e_b(u)}^{n}) \qquad (4.7)\\
&\quad+\frac{1}{2}\sum_{j=1}^{n}(\overline{e_b(u)}^{j}-\overline{e_b(u)}^{j-1},G\overline{e_b(u)}^{j}-G\overline{e_b(u)}^{j-1})\\
&= \frac{1}{2}\|\overline{e_b(u)}^{n}\|^2_{-1}+\frac{1}{2}\sum_{j=1}^{n}\|\overline{e_b(u)}^{j}-\overline{e_b(u)}^{j-1}\|^2_{-1}.
\end{aligned}
$$

For T_2 we have

$$
\begin{aligned}
T_2 &= \sum_{j=1}^{n}\tau(\nabla\overline{u}^{j}-\nabla u^j,\nabla G\overline{e_b(u)}^{j})\\
&= \tau\sum_{j=1}^{n}\left(\frac{1}{\tau}\int_{t_{j-1}}^{t_j}(u(t)-u^j)dt,\frac{1}{\tau}\int_{t_{j-1}}^{t_j}(b(u(s))-b_\epsilon(u^j))ds\right)\\
&= \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\big(u(t)-u^j,b(u(t))-b_\epsilon(u^j)\big)\,dt\\
&\quad+\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left(u(t)-u^j,\frac{1}{\tau}\int_{t_{j-1}}^{t_j}(b(u(s))-b(u(t)))ds\right)dt\\
&=: T_{21}+T_{22}.
\end{aligned}
$$

T_{21} can be decomposed as follows:

$$T_{21} = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, b(u(t)) - b_\epsilon(u(t))\big)\, dt$$
$$+ \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, b_\epsilon(u(t)) - b_\epsilon(u^j)\big)\, dt \;=: T_{211} + T_{212}.$$

The definition of b_ϵ in (2.7) gives

$$|T_{211}| = \left| \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, \epsilon\, u(t)\big)\, dt \right|$$
$$\leq \epsilon \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big\|u(t) - u^j\big\|\, \|u(t)\|\, dt \tag{4.8}$$
$$\leq \frac{\epsilon}{4} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big\|u(t) - u^j\big\|^2\, dt + \epsilon\, \|u\|^2_{L^2(0,t_n;L^2(\Omega))}.$$

Since b_ϵ is monotone, T_{212} is positive; moreover, it holds

$$T_{212} \geq \frac{1}{2} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, b_\epsilon(u(t)) - b_\epsilon(u^j)\big)\, dt$$
$$+ \frac{\epsilon}{2} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big\|u(t) - u^j\big\|^2\, dt. \tag{4.9}$$

Proceeding as for T_{111} and recalling the a priori estimates in Proposition 3.3, since $b(u) \in H^1(J; H^{-1})$ and $u \in L^2(J; H^1)$ we obtain

$$T_{22} = \frac{1}{\tau} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \int_{t_{j-1}}^{t_j} \big(u(t) - u^j, b(u(s)) - b(u(t))\big)\, ds dt$$
$$= \frac{1}{\tau} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \int_{t_{j-1}}^{t_j} \left(\int_t^s (u(t) - u^j, \partial_r b(u)) dr \right) ds dt$$
$$\leq \frac{1}{\tau} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \int_{t_{j-1}}^{t_j} \int_t^s \|\nabla(u(t) - u^j)\| \|\partial_r b(u)\|_{-1} dr ds dt \tag{4.10}$$
$$\leq \frac{\tau}{2} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \big\|\nabla(u(t) - u^j)\big\|^2 + \frac{\tau}{2} \|\partial_r b(u)\|^2_{L^2(0,t_n;H^{-1})}$$
$$\leq C\, \tau.$$

For the T_3 we proceed as follows:

$$|T_3| \leq \frac{\tau}{4\delta} \sum_{j=1}^{n} \|\overline{k(b(u))}^j - k(b(u^j))\|^2 + \delta\tau \sum_{j=1}^{n} \|\overline{e_b(u)}^j\|^2_{-1}$$
$$= T_{31} + \delta\tau \sum_{j=1}^{n} \|\overline{e_b(u)}^j\|^2_{-1}.$$

Applying (A3) and taking $\delta = C_k$ gives

$$
\begin{aligned}
|T_{31}| &= \frac{\tau}{4\delta}\frac{1}{\tau^2}\sum_{j=1}^{n}\int_\Omega\left(\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))dt\right)^2 dx\\
&\le \frac{1}{4\delta\tau}\sum_{j=1}^{n}\int_\Omega \tau\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))^2 dtdx\\
&\le \frac{1}{4}\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left(b(u)-b(u^j),u-u^j\right)dt \qquad (4.11)\\
&\le \frac{1}{4}\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left(b_\epsilon(u)-b_\epsilon(u^j),u-u^j\right)dt.
\end{aligned}
$$

Since $b(u)\in H^1(J;H^{-1})$ and $u\in L^2(J;H^1(\Omega))$, inserting (4.5)–(4.11) into (4.4) yields

$$
\begin{aligned}
&\|\overline{e_b(u)}^n\|_{-1}^2+\sum_{j=1}^{n}\|\overline{e_b(u)}^j-\overline{e_b(u)}^{j-1}\|_{-1}^2+\epsilon\int_{t_{j-1}}^{t_j}\left\|u(t)-u^j\right\|^2dt\\
&\quad+\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left(u(t)-u^j,b_\epsilon(u(t))-b_\epsilon(u^j)\right)dt\\
&\quad\le C(\tau+\epsilon)+4C_k\tau\sum_{j=1}^{n}\|\overline{e_b(u)}^j\|_{-1}^2,
\end{aligned}
$$

and (4.3) is a direct consequence of the discrete Gronwall lemma. □

Using the above result an error estimate for the L^2 norm of the time integrated gradient can be obtained. Such an estimate is essential for our analysis because it provides also an error estimate for the time integral of the flux in the mixed formulation.

PROPOSITION 4.3. *Under the assumptions in Proposition* 4.2 *we have*

$$
\left\|\int_0^T(u(t)-u_\Delta(t))dt\right\|_1^2\le C\,(\tau+\epsilon). \qquad (4.12)
$$

Proof. Following the ideas in [21], we first add (2.9) for $n=1$ to N, subtract the result from (2.1) at $t=t_N=T$ and end up with

$$
\begin{aligned}
&\left(b(u(T))-b_\epsilon(u^N),\phi\right)-\left(b(u(t_0))-b_\epsilon(u^0),\phi\right)\\
&\quad+\left(\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}\nabla(u(t)-u^j)dt,\nabla\phi\right)\\
&\quad+\left(\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))e_z dt,\nabla\phi\right)=0
\end{aligned}
$$

for all $\phi\in H_0^1(\Omega)$. Now taking $\phi=\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)$ into the total above gives

$$
\begin{aligned}
&\left(b(u(T))-b_\epsilon(u^N),\tau\sum_{j=1}^{N}\left(\overline{u}^j-u^j\right)\right)\\
&-\left(\epsilon u^0,\tau\sum_{j=1}^{N}\left(\overline{u}^j-u^j\right)\right)+\left\|\tau\sum_{j=1}^{N}\nabla\left(\overline{u}^j-u^j\right)\right\|^2\\
&+\left(\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))e_z,\nabla\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right)=0.
\end{aligned}
\tag{4.13}
$$

Denoting the terms in (4.13) by T_1, T_2, T_3, and T_4, we proceed by estimating each of them separately. T_1 yields

$$
T_1=\left(b(u(T))-\overline{b(u)}^N+\overline{b(u)}^N-b_\epsilon(u^N),\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right)=:T_{11}+T_{12}.
$$

As in (4.5), since $\partial_t b(u)\in L^2(J;H^{-1})$, T_{11} gives

$$
\begin{aligned}
|T_{11}|&\le\frac{1}{\tau}\int_{t_{N-1}}^{t_N}\int_{t}^{t_N}\left|\left(\partial_s b(u),\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right)\right|dsdt\\
&\le\frac{C_{11}}{2\delta_{11}}\tau+\frac{\delta_{11}}{2}\left\|\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|_1^2.
\end{aligned}
\tag{4.14}
$$

Applying the Cauchy–Schwarz inequality, for T_{12} we obtain

$$
|T_{12}|\le\frac{1}{2\delta_{12}}\left\|\overline{e_b(u)}^N\right\|_{-1}^2+\frac{\delta_{12}}{2}\left\|\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|_1^2.
\tag{4.15}
$$

Analogously, T_2 gives

$$
|T_2|\le\frac{1}{2\delta_2}\epsilon\left\|u^0\right\|^2+\frac{\delta_2}{2}\left\|\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|_1^2.
\tag{4.16}
$$

For T_3 we recall the inequality of Poincaré:

$$
T_3=\left\|\sum_{j=1}^{N}\tau\nabla\left(\overline{u}^j-u^j\right)\right\|^2\ge C\left\|\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|_1^2.
\tag{4.17}
$$

Analogously, T_4 can be estimated as

$$
|T_4|\le\frac{1}{2\delta_4}\left\|\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}(k(b(u))-k(b(u^j)))e_z dt\right\|^2+\frac{\delta_4}{2}\left\|\nabla\sum_{j=1}^{N}\tau\left(\overline{u}^j-u^j\right)\right\|^2.
$$

For the first term above—denoted by T_{41}—we get, by (A3),

$$
\begin{aligned}
T_{41}&\le N\sum_{j=1}^{N}\tau\int_{t_{j-1}}^{t_j}\|k(b(u))-k(b(u^j))\|^2dt\\
&\le TC_k\sum_{j=1}^{N}\int_{t_{j-1}}^{t_j}\left(b(u(t))-b(u^j),u(t)-u^j\right).
\end{aligned}
\tag{4.18}
$$

Inserting (4.14)–(4.18) into (4.13), choosing the $\delta's$ properly and recalling the estimates in Proposition 4.2 we obtain (4.12). ∎

Propositions 4.2 and 4.3 can be summarized in the following.

THEOREM 4.4. *If u is the solution of Problem 1 and u^n solves Problem 3 ($n = \overline{1,N}$), we have*

$$\max_{n=\overline{1,N}} \|\overline{e_b(u)}^n\|_{-1}^2 + \|e_b(u)\|_{L^2(J\times\Omega)}^2 + \left\|\int_0^T (u(t)-u_\Delta(t))dt\right\|_1^2 + \int_0^T (b_\epsilon(u(t))-b_\epsilon(u_\Delta(t)), u(t)-u_\Delta(t))dt \le C(\tau+\epsilon). \tag{4.19}$$

REMARK 4.5. *The estimates above do not change if we replace the last term on the left by $\int_0^T (b(u(t))-b(u_\Delta(t)), u(t)-u_\Delta(t))dt$.*

Since Problems 3 and 4 are equivalent we immediately obtain the following theorem.

THEOREM 4.6. *In the setting of Theorem 4.4, if (p^n, q^n) solve Problem 4 ($n = \overline{1,N}$), we get*

$$\begin{aligned}&\sum_{n=1}^N \int_{t_{n-1}}^{t_n} (b_\epsilon(u(t))-b_\epsilon(p^n), u(t)-p^n)dt \\ &+ \left\|\sum_{n=1}^N \int_{t_{n-1}}^{t_n} (u(t)-p^n)dt\right\|_1^2 + \left\|\tilde{q}(T) - \tau\sum_{n=1}^N q^n\right\|^2 \\ &\le C(\tau+\epsilon).\end{aligned} \tag{4.20}$$

REMARK 4.7. *As in Remark 4.5, we can replace the scalar product in (4.20) by $\int_0^T (b(u(t))-b(u_\Delta(t)), u(t)-u_\Delta(t))dt$. This immediately implies an error estimate for the saturation,*

$$\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \|b(u(t))-b(p^n)\|^2 dt \le C(\tau+\epsilon).$$

4.2. Error estimates for the fully discrete mixed discretization. The next step in our analysis is proving error estimates for the fully discrete approximation. We first estimate the error for the flux variable and then proceed with estimates for the p unknowns.

In doing so we denote by W and V the spaces $L^2(\Omega)$ and $H(\mathrm{div},\Omega)$. Let $\mathcal{T}_h$ be a regular decomposition of $\Omega \subset \mathbb{R}^d$ into closed d-simplices; h stands for the mesh-size (see [9]). Here we assume $\overline{\Omega} = \cup_{T\in\mathcal{T}_h} T$, hence Ω is polygonal. Thus we neglect the errors caused by an approximation of a nonpolygonal domain, avoiding an excess of technicalities (a complete analysis in this sense can be found in [21]).

The discrete subspaces $W_h \times V_h \subset W \times V$ are defined as

$$\begin{aligned}W_h &:= \{p \in W |\ p \text{ is constant on each element } T \in \mathcal{T}_h\}, \\ V_h &:= \{\overrightarrow{q} \in V |\ \overrightarrow{q}_{|T} = \overrightarrow{a} + b\overrightarrow{x} \ \text{ for all } T \in \mathcal{T}_h\}.\end{aligned} \tag{4.21}$$

So W_h denotes the space of piecewise constant functions, while V_h is the RT_0 space (see [7]). Further we make use of the usual L^2 projector

$$P_h : L^2(\Omega) \to W_h, \quad ((P_h w - w), w_h) = 0 \qquad \forall w_h \in W_h. \tag{4.22}$$

Taking a $\widetilde{V}$ slightly better than V (for example, $V \cap (L^s(\Omega))^d$ with an $s > 2$), a projector Π_h can be defined as (see [7, p. 131])

$$\Pi_h : \widetilde{V} \to V_h, \quad (\nabla \cdot (\Pi_h v - v), w_h) = 0 \tag{4.23}$$

for all $w_h \in W_h$. With $r \in (0, 1]$, for the operators defined above we have

$$\begin{aligned} \|w - P_h w\| &\le C h^r \|w\|_r, \\ \|v - \Pi_h v\| &\le C h^r \|v\|_r \end{aligned} \tag{4.24}$$

for any $w \in H^r(\Omega)$ and $v \in (H^r(\Omega))^d$.

The following technical lemma is proven in [25].

LEMMA 4.8. *Assuming* (A1)*, taking $f_h \in W_h$, a $v_h \in V_h$ exists so that*

$$\begin{aligned} \nabla \cdot v_h &= f_h, \\ \|v_h\| &\le C \|\nabla \cdot v_h\|, \end{aligned}$$

$C > 0$ being a generic constant not depending on h, f_h, or v_h.

Before proceeding with the fully discrete approximation scheme, we rewrite Problem 4 (continuous in space) as

Problem 5. Let $n = \overline{1, N}$. Find $(p^n, q^n) \in W \times V$ such that

$$(b_\epsilon(p^n), w) - (b_\epsilon(p^0), w) + \tau \left(\sum_{j=1}^{n} \nabla \cdot q^j, w \right) = 0, \tag{4.25}$$

$$(q^n, v) - (p^n, \nabla \cdot v) + (k(b(p^n))e_z, v) = 0 \tag{4.26}$$

for all $w \in W$ and $v \in V$, with $p^0 = u^0$.

The fully discrete mixed finite element approximation reads the following.

Problem 6. Let $n = \overline{1, N}$. Find $(p_h^n, q_h^n) \in W_h \times V_h$ such that

$$(b_\epsilon(p_h^n), w_h) + \tau \left(\sum_{j=1}^{n} \nabla \cdot q_h^j, w_h \right) = (b_\epsilon(p_h^0), w_h), \tag{4.27}$$

$$(q_h^n, v_h) - (p_h^n, \nabla \cdot v_h) + (k(b(p_h^n))e_z, v_h) = 0 \tag{4.28}$$

for all $w_h \in W_h$ and $v_h \in V_h$.

Initially we take $p_h^0 = b_\epsilon^{-1}(P_h b_\epsilon(u^0))$. Since $P_h b_\epsilon(u^0)$ is constant on any $T \in \mathcal{T}_h$, the same holds for $b_\epsilon^{-1}(P_h b_\epsilon(u^0))$, so $p_h^0 \in W_h$. Moreover, with this choice, for all $w_h \in W_h$, we obtain

$$(b_\epsilon(p_h^0), w_h) = (b_\epsilon(u^0), w_h) = (b_\epsilon(p^0), w_h).$$

We start with some stability estimates for the fully discrete case.

PROPOSITION 4.9. *Assuming* (A1)–(A4)*, if (p_h^n, q_h^n) solve Problem* 6 *($n = \overline{1, N}$), we have*

$$\begin{aligned} \|p_h^n\|^2 + \|q_h^n\|^2 &\le C, \\ \sum_{k=1}^{n} (b_\epsilon(p_h^k) - b_\epsilon(p_h^{k-1}), p_h^k - p_h^{k-1}) &\le C\tau. \end{aligned} \tag{4.29}$$

Proof. Applying the arguments used in Propositions 3.4 and 3.7 we immediately obtain the estimates (4.29), excepting the one for $\|p_h^n\|$. To complete the proof we apply Lemma 4.8. Consequently, there exists a $v_h \in V_h$ such that $\nabla \cdot v_h = p_h^n$ and $\|v_h\| \leq C\|p_h^n\|$. Using this as a test function in (4.28) gives

$$\|p_h^n\|^2 = (q_h^n, v_h) + (k(b(p_h^n))e_z, v_h) \leq C(\|q_h^n\| + \|k(b(p_h^n))\|)\|p_h^n\|,$$

and the estimate follows from the estimates for $\|q_h^n\|$ and the boundedness of k. ∎

Applying now techniques developed in [2] we estimate the errors induced by the spatial discretization.

PROPOSITION 4.10. *Let $n = \overline{1,N}$. If $(p^n, q^n) \in W \times V$, $(p_h^n, q_h^n) \in W_h \times V_h$ solve Problem 5, respectively 6, assuming* (A1)–(A4) *yields*

$$\sum_{n=1}^{N} \left\{ (b_\epsilon(p^n) - b_\epsilon(p_h^n), p^n - p_h^n) + \tau \|\Pi_h q^n - q_h^n\|^2 \right\} + \tau \left\| \sum_{n=1}^{N} (\Pi_h q^n - q_h^n) \right\|^2$$

$$\leq C \sum_{n=1}^{N} \{ \|q^n - \Pi_h q^n\|^2 + \|P_h p^n - p^n\|^2 \}. \tag{4.30}$$

Proof. Subtracting (4.27) from (4.25) and (4.28) from (4.26) gives

$$(b_\epsilon(p^n) - b_\epsilon(p_h^n), w_h) + \tau \left(\sum_{j=1}^{n} \nabla \cdot (q^j - q_h^j), w_h \right) = 0,$$

$$(q^n - q_h^n, v_h) - (p^n - p_h^n, \nabla \cdot v_h) + ((k(b(p^n)) - k(b(p_h^n)))e_z, v_h) = 0.$$

Taking $w_h = P_h p^n - p_h^n \in W_h$ and $v_h = \tau \sum_{j=1}^{n} (\Pi_h q^j - q_h^j) \in V_h$ into the above leads to

$$(b_\epsilon(p^n) - b_\epsilon(p_h^n), P_h p^n - p_h^n) + \tau \left(\sum_{j=1}^{n} \nabla \cdot (\Pi_h q^j - q_h^j), P_h p^n - p_h^n \right) = 0,$$

$$\tau \left(q^n - q_h^n, \sum_{j=1}^{n} (\Pi_h q^j - q_h^j) \right) - \tau \left(P_h p^n - p_h^n, \nabla \cdot \sum_{j=1}^{n} (\Pi_h q^j - q_h^j) \right)$$

$$+ \tau \left((k(b(p^n)) - k(b(p_h^n)))e_z, \sum_{j=1}^{n} (\Pi_h q^j - q_h^j) \right) = 0.$$

Adding these equalities and summing the result up from 1 to N yields

$$\sum_{n=1}^{N} (b_\epsilon(p^n) - b_\epsilon(p_h^n), P_h p^n - p_h^n) + \sum_{n=1}^{N} \left(q^n - q_h^n, \sum_{j=1}^{n} \tau (\Pi_h q^j - q_h^j) \right)$$

$$+ \sum_{n=1}^{N} \left((k(b(p^n)) - k(b(p_h^n)))e_z, \sum_{j=1}^{n} \tau (\Pi_h q^j - q_h^j) \right) = 0. \tag{4.31}$$

We estimate now each of the terms above, denoted by T_1, T_2, and T_3:

$$T_1 = \sum_{n=1}^{N} (b_\epsilon(p^n) - b_\epsilon(p_h^n), p^n - p_h^n)$$

$$+\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),P_hp^n-p^n)=:T_{11}+T_{12}. \tag{4.32}$$

By (A2) and the definition (2.7) of b_ϵ, a $C>0$ independent of τ and h exists such that

$$\begin{aligned}T_{11}&\geq\frac{1}{2}\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n)\\&\quad+C\left(\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2+\epsilon\sum_{n=1}^{N}\|p^n-p_h^n\|^2\right).\end{aligned} \tag{4.33}$$

Applying the inequality of Cauchy T_{12} yields

$$|T_{12}|\leq\frac{\mu}{2}\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2+\frac{1}{2\mu}\sum_{n=1}^{N}\|P_hp^n-p^n\|^2. \tag{4.34}$$

Rewriting T_2 as

$$\begin{aligned}T_2&=\sum_{n=1}^{N}\left(q^n-\Pi_hq^n,\sum_{j=1}^{n}\tau(\Pi_hq^j-q_h^j)\right)\\&\quad+\sum_{n=1}^{N}\left(\Pi_hq^n-q_h^n,\sum_{j=1}^{n}\tau(\Pi_hq^j-q_h^j)\right)=:T_{21}+T_{22},\end{aligned} \tag{4.35}$$

we estimate T_{21} and T_{22}. For T_{21} we get

$$|T_{21}|\leq\frac{1}{2}\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2+\frac{\tau^2}{2}\sum_{n=1}^{N}\left\|\sum_{j=1}^{n}(\Pi_hq^j-q_h^j)\right\|^2, \tag{4.36}$$

while for T_{22} we use (3.1) to obtain

$$T_{22}=\frac{\tau}{2}\left\|\sum_{n=1}^{N}(\Pi_hq^n-q_h^n)\right\|^2+\frac{\tau}{2}\sum_{n=1}^{N}\|\Pi_hq^n-q_h^n\|^2. \tag{4.37}$$

Using (A3), T_3 gets

$$\begin{aligned}|T_3|&\leq\frac{\delta}{2}\sum_{n=1}^{N}\|k(b(p^n))-k(b(p_h^n))\|^2+\frac{\tau^2}{2\delta}\sum_{n=1}^{N}\left\|\sum_{j=1}^{n}(\Pi_hq^j-q_h^j)\right\|^2\\&\leq\frac{C_k\delta}{2}\sum_{n=1}^{N}(b(p^n)-b(p_h^n),p^n-p_h^n)+\frac{\tau^2}{2\delta}\sum_{n=1}^{N}\left\|\sum_{j=1}^{n}(\Pi_hq^j-q_h^j)\right\|^2.\end{aligned} \tag{4.38}$$

Inserting (4.32)–(4.38) into (4.31) and choosing μ and δ properly gives

$$\begin{aligned}&\sum_{n=1}^{N}\{(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n)+\tau\|\Pi_hq^n-q_h^n\|^2\}+\tau\left\|\sum_{n=1}^{N}(\Pi_hq^n-q_h^n)\right\|^2\\&\quad\leq C\sum_{n=1}^{N}\left\{\|q^n-\Pi_hq^n\|^2+\|P_hp^n-p^n\|^2+\tau^2\left\|\sum_{j=1}^{n}(\Pi_hq^j-q_h^j)\right\|^2\right\}.\end{aligned}$$

Finally, (4.30) follows applying the discrete Gronwall lemma. ∎

REMARK 4.11. *By the equivalence proven in Proposition* 2.3, $p^n \in H^1(\Omega)$ *for all* n. *Now using* (4.24) *and* (3.7) *we get*

$$\sum_{n=1}^{N} \|P_h p^n - p^n\|^2 \le Ch^2 \sum_{n=1}^{N} \|p^n\|_1^2 \le C\frac{h^2}{\tau},$$

and the estimates (4.30) *can be modified accordingly.*

Similar estimates can be obtained for the p-unknowns.

PROPOSITION 4.12. *Under the assumptions of Proposition* 4.10 *we have*

$$\tau \left\| \sum_{n=1}^{N} (P_h p^n - p_h^n) \right\|^2 \le C \Bigg\{ \sum_{n=1}^{N} (b_\epsilon(p^n) - b_\epsilon(p_h^n), p^n - p_h^n) + \tau \left\| \sum_{n=1}^{N} (\Pi_h q^n - q_h^n) \right\|^2 + \sum_{n=1}^{N} \|q^n - \Pi_h q^n\|^2 \Bigg\}. \tag{4.39}$$

Proof. Subtracting (4.28) from (4.26), recalling the definition of P_h, and summing up for $n = 1$ to N yield

$$\left(\sum_{n=1}^{N} (q^n - q_h^n), v_h \right) - \left(\sum_{n=1}^{N} (P_h p^n - p_h^n), \nabla \cdot v_h \right) + \left(\sum_{n=1}^{N} \tau (k(b(p^n)) - k(b(p_h^n)))e_z, v_h \right) = 0 \tag{4.40}$$

for any $v_h \in V_h$. Using now Lemma 4.8, a $v_h \in V_h$ exists such that

$$\nabla \cdot v_h = \sum_{n=1}^{N} \tau (P_h p^n - p_h^n) \tag{4.41}$$

and $\|v_h\| < C\|\tau \sum_{n=1}^{N} (P_h p^n - p_h^n)\|$. In this case (4.40) gives

$$\tau \left\| \sum_{n=1}^{N} (P_h p^n - p_h^n) \right\|^2 = \left(\sum_{n=1}^{N} (q^n - q_h^n), v_h \right) + \left(\sum_{n=1}^{N} (k(b(p^n)) - k(b(p_h^n)))e_z, v_h \right). \tag{4.42}$$

Denoting by T_1 and T_2 the terms on the right into above, applying the inequality of Cauchy and recalling the estimates on $\|v_h\|$ leads to

$$|T_1| \le \frac{\tau}{2\delta_1} \left\| \sum_{n=1}^{N} (q_h^n - q^n) \right\|^2 + \frac{\delta_1}{2\tau} \|v_h\|^2$$

$$\le \frac{\tau}{2\delta_1} \left\| \sum_{n=1}^{N} (q_h^n - q^n) \right\|^2 + \frac{C\tau\delta_1}{2} \left\| \sum_{n=1}^{N} (p_h^n - P_h p_h^n) \right\|^2. \tag{4.43}$$

Similarly, by (A3) we obtain

$$|T_2| \le \frac{\tau}{2\delta_2}\left\|\sum_{n=1}^{N}(k(b(p_h^n))-k(b(p^n)))\right\|^2 + \frac{\delta_2}{2\tau}\|v_h\|^2$$

$$\le \frac{C}{2\delta_2}\sum_{n=1}^{N}(b(p_h^n)-b(p^n),p_h^n-p^n) + \frac{C\tau\delta_2}{2}\left\|\sum_{n=1}^{N}(p_h^n-P_hp_h^n)\right\|^2. \tag{4.44}$$

Choosing δ_1 and δ_2 properly, (4.42)–(4.44) gives

$$\tau\left\|\sum_{n=1}^{N}(P_hp^n-p_h^n)\right\|^2$$

$$\le C\left\{\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n)+\tau\left\|\sum_{n=1}^{N}(q^n-q_h^n)\right\|^2\right\}.$$

The last term above can be rewritten as

$$\tau\left\|\sum_{n=1}^{N}(q^n-q_h^n)\right\|^2 \le \tau\left\|\sum_{n=1}^{N}(\Pi_hq^n-q_h^n)\right\|^2+\tau\left\|\sum_{n=1}^{N}(q^n-\Pi_hq^n)\right\|^2$$

$$\le \tau\left\|\sum_{n=1}^{N}(\Pi_hq^n-q_h^n)\right\|^2+T\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2,$$

which completes the proof. □

The following is a direct consequence of Propositions 4.10 and 4.12.

THEOREM 4.13. *Assuming* (A1)–(A4), *if* $(p^n,q^n)\in W\times V$, $(p_h^n,q_h^n)\in W_h\times V_h$ *solve, for* $n=\overline{1,N}$, *Problems* 5 *and* 6, *we obtain*

$$\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n)+\tau\sum_{n=1}^{N}\|\Pi_hq^n-q_h^n\|^2$$

$$+\tau\left\|\sum_{n=1}^{N}(q^n-q_h^n)\right\|^2+\tau\left\|\sum_{n=1}^{N}(p^n-p_h^n)\right\|^2 \tag{4.45}$$

$$\le C\left(\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2+\sum_{n=1}^{N}\|P_hp^n-p^n\|^2\right).$$

Combining the estimates in Theorems 4.6 and 4.13 and recalling Remark 4.11 we get, for the fully discrete scheme, the following.

THEOREM 4.14. *Assuming* (A1)–(A4), *we get*

$$\left\|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(u(t)-p_h^n)dt\right\|^2+\left\|\tilde{q}(T)-\tau\sum_{n=1}^{N}q_h^n\right\|^2$$

$$\le C\left(\tau+\epsilon+h^2+\tau\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2\right). \tag{4.46}$$

Proof. Let T_1 and T_2 denote the terms on the left in (4.46). For T_1, by the properties of norms we have

$$T_1 \le 2\left\|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(u(t)-p^n)dt\right\|^2 + 2\tau^2\left\|\sum_{n=1}^{N}(p^n-p_h^n)\right\|^2.$$

Estimates (4.20) in Theorem 4.4 give

$$\left\|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(u(t)-p^n)dt\right\|^2 \le C(\tau+\epsilon),$$

which, together with (4.45), imply

$$T_1 \le C(\tau+\epsilon). \tag{4.47}$$

Analogously, for T_2 we obtain

$$\begin{aligned} T_2 &\le 2\left\|\tilde{q}(T)-\tau\sum_{n=1}^{N}q^n\right\|^2 + 2\tau^2\left\|\sum_{n=1}^{N}(q^n-q_h^n)\right\|^2 \\ &\le C_1(\tau+\epsilon) + C_2\left(\tau+\epsilon+h^2+\tau\sum_{n=1}^{N}\|q^n-\Pi_h q^n\|^2\right) \end{aligned} \tag{4.48}$$

by the arguments above. Now (4.46) follows from (4.47) and (4.48). ∎

COROLLARY 4.15. *Under the assumptions of Theorem* 4.14*, for the scalar product we have*

$$\begin{aligned} &\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p_h^n),u(t)-p_h^n)dt \\ &\quad\le C\left(\tau^{\frac12}+\epsilon^{\frac12}+h^2/\tau^{\frac12}+\tau^{\frac12}\sum_{n=1}^{N}\|q^n-\Pi_h q^n\|^2\right). \end{aligned} \tag{4.49}$$

Proof. We decompose the scalar product as follows:

$$\begin{aligned} &\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p_h^n),u(t)-p_h^n)dt \\ &\quad=\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p^n),u(t)-p_h^n)dt \\ &\qquad+\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(p^n)-b_\epsilon(p_h^n),u(t)-p_h^n)dt =: T_1+T_2. \end{aligned}$$

Applying Cauchy's inequality T_1 yields

$$\begin{aligned} |T_1| &\le \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|b_\epsilon(u(t))-b_\epsilon(p^n)\|\|u(t)-p_h^n\|dt \\ &\le \frac{1}{4\delta_1}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|b_\epsilon(u(t))-b_\epsilon(p^n)\|^2dt + \delta_1\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|u(t)-p_h^n\|^2dt. \end{aligned}$$

Since b_ϵ is Lipschitz, using (4.20), the first sum gives

$$\begin{aligned}
&\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|b_\epsilon(u(t))-b_\epsilon(p^n)\|^2dt\\
&\quad\le C\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p^n),u(t)-p^n)dt\\
&\quad\le C(\tau+\epsilon).
\end{aligned}$$

Having $u\in L^2(J\times\Omega)$, by (4.29) we get

$$\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|u(t)-p_h^n\|^2dt\le C.$$

In this way, choosing $\delta_1=(\tau+\epsilon)^{\frac12}$ yields

$$|T_1|\le C(\tau+\epsilon)^{\frac12}\le C(\tau^{\frac12}+\epsilon^{\frac12}). \tag{4.50}$$

For T_2 we obtain

$$\begin{aligned}
|T_2|&\le\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|\|u(t)-p_h^n\|dt\\
&\le\frac{\tau^{\frac12}}{4}\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2+\tau^{\frac12}\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\|u(t)-p_h^n\|^2dt.
\end{aligned}$$

As before, the second sum above is uniformly bounded, while for the first one we write

$$\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2\le C\sum_{n=1}^{N}(b_\epsilon(p^n)-b_\epsilon(p_h^n),p^n-p_h^n).$$

Using (4.45) gives

$$\sum_{n=1}^{N}\|b_\epsilon(p^n)-b_\epsilon(p_h^n)\|^2\le C\left(h^2/\tau+\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2\right),$$

so T_2 is bounded by

$$|T_2|\le C\left(\tau^{\frac12}+h^2/\tau^{\frac12}+\tau^{\frac12}\sum_{n=1}^{N}\|q^n-\Pi_hq^n\|^2\right). \tag{4.51}$$

The result follows now from (4.50) and (4.51). ∎

Assuming additionally

(A5) $q^n\in H^1(\Omega)^d$ for all $n=\overline{1,N}$,

using (4.24) and the estimates in Theorem 4.14 and Corollary 4.15 we obtain the following theorem.

THEOREM 4.16. *Assuming* (A1)–(A5) *we have*

$$\left\|\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(u(t)-p_h^n)dt\right\|^2+\left\|\tilde{q}(T)-\tau\sum_{n=1}^{N}q_h^n\right\|^2\le C(\tau+\epsilon+h^2),$$

$$\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(b_\epsilon(u(t))-b_\epsilon(p_h^n),u(t)-p_h^n)dt\le C(\tau^{\frac{1}{2}}+\epsilon^{\frac{1}{2}}+h^2/\tau^{\frac{1}{2}}). \tag{4.52}$$

REMARK 4.17. *Obviously* (A5) *is fulfilled in one spatial dimension, since then* $H(\mathrm{div},\Omega)$ *and* $H^1(\Omega)$ *coincide. Assumption* (A5) *also holds in the multidimensional case, provided* $\partial\Omega$ *is smooth enough and* k *is differentiable. Using* (A2) *and* (A3)*, since* $k(b(\cdot))\in C^1(0,1)$ *we have*

$$|\partial_u k(b(u))|\le\lim_{\delta\to 0}\left|\frac{k(b(u+\delta))-k(b(u))}{\delta}\right|\le\sqrt{C_k\frac{b(u+\delta)-b(u)}{\delta}}\le C.$$

Following [19, *Chapter* 4, *Theorems* 5.1 *and* 5.2]*, for any* $n=\overline{1,N}$*,* u^n *solving Problem* 3 *is in* $H^2(\Omega)$ *and the corresponding norm is bounded uniformly in* n *by a constant that, nevertheless, may depend on* τ*. Therefore* $q^n\in H^1(\Omega)$ *for all* $n\ge 1$ *and* $\|q^n\|_1\le C(\tau)$.

To confirm our theoretical results we present a numerical test. We consider a problem allowing for a travel wave solution, as proposed in [12] which refers to the Richards' equation in its form after the Kirchhoff transformation (1.4), without gravitation term and with

$$b(u)=\begin{cases}\frac{\pi^2}{2}-\frac{u^2}{2} & \text{for} \quad u\le 0,\\ \frac{\pi^2}{2} & \text{for} \quad u>0.\end{cases}$$

For this problem an exact solution is known:

$$u_{\mathrm{ex}}(t,x,y)=\begin{cases}\frac{-2(e^s-1)}{e^s+1} & \text{for} \quad s\ge 0,\\ -s & \text{for} \quad s<0,\end{cases}$$

where $s=x-y-t$. The equation has been solved in the unit square Ω, with Dirichlet boundary condition given by $u=u_{\mathrm{ex}}$ on $\partial\Omega$ and initial value u_{ex} at $t=0$. For mixed finite element discretizations the emerging system of equations is difficult to solve due to being the solution of a saddle point problem. A common implementation trick is to enlarge the system by adding Lagrange multipliers on edges (hybridization of the method). Briefly, within one timestep the resulting algorithm reads as follows: first the flux variable is eliminated on each element; then the continuity equation is locally solved for pressure by a variably damped Newton's method. The global system is set for the Lagrange multipliers and solved using the Newton method and a multigrid solver for the linear subproblems in the Newton iterations (for details see [25]). An alternative linearization approach is discussed in [24]. The implementation is based on the package UG (version 3.8, see also [3]), and the computations have been done on a SUN workstation.

To verify the theoretical estimates, we have started performing computations on a uniform triangular mesh with $h=0.25$, and a time step $\tau=0.04$. Then τ and h^2 are successively halved, up to $\tau=0.000625$ and $h=0.03125$. The final time was set

TABLE 4.1
Numerical results.

N	τ	h	Error	$\tau + h^2$	Convergence Order
1	0.04	0.25	6.344201e-06	1.025000e-01	—
2	0.02	0.176	3.620119e-06	5.125000e-02	0.81 0.81
3	0.01	0.125	2.057356e-06	2.562500e-02	0.82 0.81
4	0.005	0.088	9.574634e-07	1.281250e-02	1.10 0.91
5	0.0025	0.0625	5.362175e-07	6.406250e-03	0.84 0.89
6	0.00125	0.044	2.431734e-07	3.203250e-03	1.14 0.94
7	0.000625	0.03125	1.355397e-07	1.601562e-03	0.84 0.92

to be 1.0 for all the computations. Knowing the exact solution, the square of the total error (as written in (4.52)) is given by

$$E_{tot}^2 = \left\| \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (u_{\mathrm{ex}}(t) - p_h^n)\, dt \right\|^2 + \left\| \tilde{q}_{\mathrm{ex}}(T) - \tau \sum_{n=1}^{N} q_h^n \right\|^2,$$

where $\tilde{q}_{\mathrm{ex}}(T) = \int_0^T \nabla\, u_{\mathrm{ex}}(t)\, dt$ is the exact flux. The order of convergence (for the squared error) is estimated by dividing the errors above, computed for two sets of parameters (refined according to the procedure mentioned above). Dividing the natural logarithm of the result by the natural logarithm of the refinement ratio yields an approximation of the convergence order. Results are displayed in Table 4.1. As predicted by Theorem 4.14, the order of E_{tot}^2 is between 0.8 and 1.1. Thus we can conclude that the numerical results are in concordance with our theoretical analysis, in particular confirming the convergence of the scheme.

Acknowledgments. We would like to thank Prof. C. J. van Duijn and Dr. E. F. Kaasschieter for useful discussions and suggestions.

REFERENCES

[1] H. W. ALT AND S. LUCKHAUS, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[2] T. ARBOGAST, M. F. WHEELER, AND N. Y. ZHANG, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[3] P. BASTIAN, K. BIRKEN, K. JOHANSSEN, S. LANG, N. NEUSS, H. RENTZ-REICHERT, AND C. WIENERS, *UG–a flexible toolbox for solving partial differential equations*, Comput. Visualiz. Sci., 1 (1997), pp. 27–40.

[4] R. G. BACA, J. N. CHUNG, AND D. J. MULLA, *Mixed transform finite element method for solving the non-linear equation for flow in variably saturated porous media*, Internat. J. Numer. Methods Fluids, 24 (1997), pp. 441–455.

[5] J. BEAR AND Y. BACHMAT, *Introduction to Modelling of Transport Phenomena in Porous Media*, Kluwer Academic, Dordrecht, The Netherlands, 1991.

[6] L. BERGANASCHI AND M. PUTTI, *Mixed finite elements and Newton-type linearizations for the solution of Richards' equation*, Internat. J. Numer. Methods Engrg., 45 (1999), pp. 1025–1046.

[7] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[8] M. A. CELIA, E. T. BOULOUTAS, AND R. L. ZARBA, *A general mass-conservative numerical solution for the unsaturated flow equation*, Water Resour. Res., 26 (1990), pp. 1483–1496.

[9] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[10] R. EYMARD, M. GUTNIC, AND D. HILLHORST, *The finite volume method for Richards equation*, Comput. Geosci., 3 (1999), pp. 259–294.

[11] K. Fadimba and R. Sharpley, *A priori estimates and regularization for a class of porous medium equations*, Nonlinear World, 2 (1995), pp. 13–41.
[12] U. Hornung and W. Messing, *Poröse Medien—Methoden und Simulation*, Verlag Beiträge zur Hydrologie, Kirchzarten, 1984.
[13] W. Jäger and J. Kačur, *Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes*, RAIRO Model. Math. Numer. Anal., 29 (1995), pp. 605–627.
[14] D. Kavetski, P. Binning, and S. W. Sloan, *Adaptive time stepping and error control in a mass conservative numerical solution of the mixed form of Richards equation*, Adv. Water Res. 24 (2001), pp. 595–605.
[15] P. Knabner, *Finite element simulation of saturated-unsaturated flow through porous media*, in Large Scale Scientific Computing, Progress in Scientific Computing 7, P. Deuflhard and B. Engquist, eds., Birkhäuser, Boston, 1987, pp. 83–93.
[16] P. Knabner and E. Schneid, *Adaptive hybrid mixed finite element discretization of instationary variably saturated flow in porous media*, in High Performance Scientific and Engineering Computing, M. Breuer, F. Durst, and C. Zenger, eds., Springer-Verlag, Berlin, 2002, pp. 37–44.
[17] P. Knabner and E. Schneid, *Numerical solution of unsteady saturated/unsaturated flow through porous media*, in Numerical Modelling in Continuum Mechanics, Part II, M. Feistauer, R. Rannacher, and K. Kozel, eds., Matfyzpress, Prague, 1997, pp. 337–343.
[18] P. Knabner and G. Summ, *Efficient realization of the mixed finite element discretization for nonlinear problems*, Math. Comp., submitted.
[19] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, London, 1968.
[20] J. L. Lions and E. Magenes, *Non Homogenous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, Berlin, 1972.
[21] R. H. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal., 25 (1988), pp. 784–814.
[22] F. Otto, *L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations*, J. Differential Equations, 131 (1996), pp. 20–38.
[23] I. S. Pop, *Error estimates for a time discretization method for the Richards' equation*, Comput. Geosci., 6 (2002), pp. 141–160.
[24] I. S. Pop, F. Radu, and P. Knabner, *Mixed finite elements for the Richards' equation: Linearization procedure*, J. Comput. Appl. Math., 168 (2004), pp. 365–373.
[25] E. Schneid, *Hybrid-Gemischte Finite-Elemente-Diskretisierung der Richards-Gleichung* (in German), Ph.D. thesis, University of Erlangen–Nürnberg, 2000; also available at http://www.am.uni-erlangen.de/am1/publications/dipl_phd_thesis/dipl_phd_thesis.html.
[26] E. Schneid, P. Knabner, and F. Radu, *A priori error estimates for a mixed finite element discretization of the Richards' equation*, Numer. Math., 98 (2004), pp. 353–370.
[27] M. Watanabe, *An approach by difference to the porous medium equation with convection*, Hiroshima Math. J., 25 (1995), pp. 623–645.
[28] G. A. Williams and C. T. Miller, *An evaluation of temporally adaptive transformation approaches for solving Richards' equation*, Adv. Water Res. 22 (1999), pp. 831–840.
[29] C. Woodward and C. Dawson, *Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media*, SIAM J. Numer. Anal. 37 (2000), pp. 701–724.
[30] I. Yotov, *A mixed finite element discretization on non–matching multiblock grids for a degenerate parabolic equation arizing in porous media flow*, East-West J. Numer. Math., 5 (1997), pp. 211–230.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1479–1501

ADAPTIVE WAVELET GALERKIN METHODS FOR LINEAR INVERSE PROBLEMS*

ALBERT COHEN†, MARC HOFFMANN‡, AND MARKUS REIß§

Abstract. We introduce and analyze numerical methods for the treatment of inverse problems, based on an adaptive wavelet Galerkin discretization. These methods combine the theoretical advantages of the wavelet-vaguelette decomposition (WVD) in terms of optimally adapting to the unknown smoothness of the solution, together with the numerical simplicity of Galerkin methods. In a first step, we simply combine a thresholding algorithm on the data with a Galerkin inversion on a fixed linear space. In a second step, a more elaborate method performs the inversion by an adaptive procedure in which a smaller space adapted to the solution is iteratively constructed; this leads to a significant reduction of the computational cost.

Key words. statistical inverse problems, Galerkin methods, wavelets and nonlinear methods, Besov spaces, minimax estimation

AMS subject classifications. 65J20, 62G07

DOI. 10.1137/S0036142902411793

1. Introduction.

1.1. Statistical model. We want to recover a function f in $L^2(\Omega_X)$, where Ω_X is a certain bounded domain in $\mathbb{R}^d$, but we are able to observe data about Kf only, where $K : L^2(\Omega_X) \to L^2(\Omega_Y)$ is a compact linear operator and Ω_X and Ω_Y are two bounded domains in $\mathbb{R}^d$ and $\mathbb{R}^q$, respectively. In the following, when mentioning L^2 or more general function spaces, we shall omit the domain Ω_X or Ω_Y when this information is obvious from the context.

We are interested in the statistical formulation of linear inverse problems: we assume that the data are noisy, so that we observe

$$g_\varepsilon = Kf + \varepsilon \dot{W}, \tag{1.1}$$

where $\dot{W}$ is a white noise and ε a noise level. In rigorous probabilistic terms, we observe a Gaussian measure on L^2 with drift Kf and intensity ε^2 (see, e.g., [20]). Observable quantities take the form

$$\langle g_\varepsilon, v\rangle = \langle Kf, v\rangle + \varepsilon\, \eta(v),$$

where $v \in L^2$ is a test function and $\eta(v)$ is a Gaussian centered random variable with variance $\|v\|_{L^2}^2$. For $v_1, v_2 \in L^2$ the covariance between $\eta(v_1)$ and $\eta(v_2)$ is given by the scalar product $\langle v_1, v_2\rangle$. In particular, if v_1 and v_2 are orthogonal, the random variables $\eta(v_1)$ and $\eta(v_2)$ are stochastically independent. The cognitive value of the white noise model (1.1) is discussed in detail in [4], [26] and the references therein.

*Received by the editors July 22, 2002; accepted for publication (in revised form) October 22, 2003; published electronically December 16, 2004. This research was partially supported by the European network Dynstoch and the DFG-Sonderforschungsbereich 383.

http://www.siam.org/journals/sinum/42-4/41179.html

†Laboratoire J.L. Lions, Université Pierre et Marie Curie, 175 rue du Chevaleret, 75013, Paris, France (cohen@ann.jussieu.fr).

‡Laboratoire d'Analyse et de Mathématiques Appliquées, Université Marne-la-Vallée, 5 Blvd. Descartes, 77454, Marne-la-Vallée, Cedex 2, France (hoffmann@math.univ-mlv.fr).

§Institut für Mathematik, Humboldt-Universität zu Berlin, unter den Linden 6, D-10099, Berlin, Germany (reiss@mathematik.hu-berlin.de).

If $f_\varepsilon \in L^2$ is an estimator of f, that is a measurable quantity w.r.t. the data g_ε, we measure its accuracy by the mean-square error $E(\|f_\varepsilon - f\|_{L^2}^2)$ as $\varepsilon \to 0$, with $E(\cdot)$ denoting the expectation operator.

1.2. SVD and Galerkin projection. Among the most popular regularization methods, let us first mention the *singular value decomposition* (SVD); see, e.g., [21], [22], and [24]. Although very attractive theoretically, the SVD suffers from two limitations. First, the singular basis functions may be difficult to determine and manipulate numerically. Second, while these bases are fully adapted to describe the action of K, they might not be appropriate for the accurate description of the solution with a small number of parameters (see, e.g., [16]). Concerning numerical simplicity, *projection methods* are more appealing. Given finite dimensional subspaces $X_h \subset L^2(\Omega_X)$ and $Y_h \subset L^2(\Omega_Y)$ with $\dim(X_h) = \dim(Y_h)$, one defines the approximation f_ε as the solution in X_h of the problem

$$\langle K f_\varepsilon, g_h \rangle = \langle g, g_h \rangle \quad \text{for all} \quad g_h \in Y_h, \tag{1.2}$$

which amounts to solving a linear system (see [25] for a general approach to projection methods). In the case where $\Omega_X = \Omega_Y$ and K is a self-adjoint positive definite operator, we choose $Y_h = X_h$ and the linear system is particularly simple to solve since the corresponding discretized operator K_h is symmetric positive definite: this is the so-called *Galerkin method*. In the case of general $\Omega_X \neq \Omega_Y$ and injective K, one may choose $Y_h := K(X_h)$ and we are led back to the Galerkin method applied to the least squares equation $K^*Kf = K^*g$ with data K^*g_ε, where K^* denotes the adjoint of K. The numerical simplicity of projection methods comes from the fact that X_h and Y_h are typically finite element spaces equipped with standard local bases. As in the SVD method, the discretization parameter h has to be properly chosen. The choice of finite element spaces for X_h and Y_h is also beneficial with respect to the second limitation of SVD, since the approximation properties of finite elements can be exploited when the solution has some smoothness.

However, the Galerkin projection method suffers from two drawbacks which are encountered in all *linear* estimation methods, including, in particular, the SVD. First, the choice of h with respect to the noise level ε depends on the regularity of the solution which is almost always unknown in advance. Second, the use of a finite element space X_h with a fixed uniform mesh size h does not provide any spatial adaptation.

1.3. Wavelet-vaguelette decomposition. In recent years, *nonlinear* methods have been developed, with the objective of automatically adapting to unknown smoothness and locally singular behavior of the solution. In the case of simple denoising, i.e., when K is the identity, wavelet thresholding is probably one of the most attractive nonlinear methods, since it is both numerically straightforward and asymptotically optimal for a large variety of Sobolev or Besov classes as models for the unknown smoothness of the solution; see, e.g., [18]. This success strongly exploits the fact that wavelets provide unconditional bases for such smoothness spaces. In order to adapt this approach to the framework of ill-posed inverse problems, Donoho introduced in [16] a wavelet-like decomposition which is specifically adapted to describe the action of K, the so-called *wavelet-vaguelette decomposition* (WVD), and proposed applying a thresholding algorithm on this decomposition. In [1] Donoho's method was compared with the similar *vaguelette-wavelet decomposition* (VWD) algorithm. Both methods rely on an orthogonal wavelet basis (ψ_λ) and associated Riesz bases of

"vaguelettes" defined as

$$v_\lambda = \beta_\lambda K^{-1}\psi_\lambda \text{ and } u_\lambda = \beta_\lambda (K^*)^{-1}\psi_\lambda, \tag{1.3}$$

where the scaling coefficients β_λ typically depend on the order of ill-posedness of K. We thus have the WVD and VWD decompositions

$$f = \sum_\lambda \beta_\lambda^{-1}\langle Kf, u_\lambda\rangle\psi_\lambda = \sum_\lambda \beta_\lambda^{-1}\langle Kf, \psi_\lambda\rangle v_\lambda. \tag{1.4}$$

The WVD and VWD estimation methods amount to estimating the coefficients in these expansions from the observed data and applying a thresholding procedure. On a theoretical level, similarly to wavelet thresholding in the case of simple denoising, both WVD and VWD allow recovery at the same rate as the projection method, under weaker smoothness conditions, a fact that reflects their ability for spatial adaptivity.

On a more applied level, numerical implementations in [1] and [15] have illustrated the efficiency of both WVD and VWD methods, in the case of operators that behave like integration $Kf(x) = \int_0^x f(t)dt$. For more general operators, however, the assumption that $K^{-1}\psi_\lambda$ or $(K^*)^{-1}\psi_\lambda$ are known for all indices λ may simply result in putting forward the inversion problem: if an integral operator has a kernel with a complicated structure (see [5]), or if this kernel is itself derived from observations (see [27]), this inversion has to be done numerically with additional computational cost and error. In other words, the vaguelettes u_λ and v_λ might be difficult to handle numerically (similar to the SVD functions), and in particular they are not ensured to have compact support.

1.4. Our approach: Adaptive wavelet Galerkin. In this context, a natural goal is to build a method which combines the numerical simplicity of linear Galerkin projection methods with the optimality of adaptive wavelet thresholding methods. This is the goal of the paper.

Adaptive Galerkin methods are well established in the context of solving operator equations *without noise*; typically, the finite element space is locally refined based on *a-posteriori error analysis* of the current numerical solution. Such adaptive algorithms were recently extended to the context of wavelet discretizations, exploiting both the characterization of function spaces and the sparse representation of the operator by wavelet bases; see, e.g., [8]. Our goal is to introduce and analyze similar adaptive wavelet Galerkin algorithms in the context of statistical inverse problems. Such adaptive algorithms involve only the wavelet system (ψ_λ) and are therefore often easier to handle numerically than WVD and VWD. On the other hand, their optimality will essentially rely on the assumption that K has certain mapping properties with respect to the relevant function spaces, a fact which is also implicitly used in the WVD and VWD approaches. Last but not least, one can exploit the fact that the Galerkin discretization of K in the wavelet basis might be sparse even for nonlocal integral operators, in order to improve the computational efficiency of our estimator.

Concerning the organization of the paper, we progressively develop our methodology. In section 2, we introduce general assumptions on model (1.1), in terms of mapping properties of the operator K between smoothness spaces. After a brief recall of the analysis of the linear Galerkin method using a wavelet multiresolution space V_j in section 3.1, a first nonlinear method is proposed in section 3.2, which initially operates in a way similar to the VWD, by thresholding the wavelet coefficients $g_\lambda^\varepsilon := \langle g_\varepsilon, \psi_\lambda\rangle$ with λ restricted up to a maximal scale level $j = j(\varepsilon)$, and then

applies a linear Galerkin inversion of these denoised data on the multiresolution space V_j. As ε decreases, the scale level $j(\varepsilon)$ grows and the Galerkin approximation thus becomes computationally heavy, while the solution could still be well represented by a small adaptive set of wavelets within V_j. Therefore, we propose in section 4 an adaptive algorithm which iteratively produces such a set together with the corresponding Galerkin estimator. This algorithm intertwines the process of thresholding with an iterative resolution of the Galerkin system, and it exploits, in addition, the sparse representation of K in the wavelet basis. As we completed the revision of this paper, we became aware of a related approach recently proposed in [13], based on least squares minimization with a nonquadratic penalization term in a deterministic setting, which results in a similar combination of gradient iteration with a thresholding procedure, yet operating in infinite dimension. Both methods in sections 3.2 and 4 are proved to achieve the same minimax rate as WVD and VWD under the same general assumptions on the operator K. We eventually compare the different estimators in section 5 numerically on the example of a singular integral equation of the first kind. For simplicity, we present our methods and results in the case where K is elliptic and self-adjoint. The extension to more general operators, via a least squares approach, is the object of Appendix A. We also discuss in Appendix B several properties of multiresolution spaces which are used throughout the paper.

2. Assumptions on the operator K. The ill-posed nature of the problem comes from the assumption that K is compact and therefore its inverse is not L^2-continuous. This is expressed by a smoothing action: K typically maps L^2 into H^t for some $t > 0$. More generally we say that K has the smoothing property of order t with respect to some smoothness space H^s (resp., W_p^s, $B_{p,q}^s$) if this space is mapped onto H^{s+t} (resp., W_p^{s+t}, $B_{p,q}^{s+t}$).

The estimator f_ε will be searched within a finite dimensional subspace V of $L^2(\Omega_X)$ based on the projection method. In the case where $\Omega_X = \Omega_Y = \Omega$ and K is self-adjoint positive definite, we shall use the Galerkin method, that is,

$$\text{find } f_\varepsilon \in V \text{ such that } \langle Kf_\varepsilon, v\rangle = \langle g_\varepsilon, v\rangle \text{ for all } v \in V. \tag{2.1}$$

The smoothing property of order t will be expressed by the ellipticity property

$$\langle Kf, f\rangle \sim \|f\|^2_{H^{-t/2}}, \tag{2.2}$$

where $H^{-t/2}$ stands for the dual space of the Sobolev space $H^{t/2}$ appended with boundary conditions that might vary depending on the considered problem (homogeneous Dirichlet, periodic, and so on). The symbol $a \sim b$ means that there exists $c > 0$ independent of f such that $c^{-1}b \le a \le cb$.

In the case where $\Omega_X \neq \Omega_Y$ or when K is not self-adjoint positive definite, we shall consider the least squares method, that is,

$$\text{find } f_\varepsilon \in V \text{ which minimizes } \|Kv - g_\varepsilon\|^2_{L^2} \text{ among all } v \in V. \tag{2.3}$$

As already remarked, this amounts to applying the Galerkin method on the equation $K^*Kf = K^*g$ with data K^*g_ε and trial space $K(V)$. The smoothing property of order t will then be expressed by the ellipticity property

$$\|Kf\|^2_{L^2} = \langle K^*Kf, f\rangle \sim \|f\|^2_{H^{-t}}. \tag{2.4}$$

We shall only deal with this general situation in Appendix A, and we therefore assume for the next sections that K is self-adjoint positive definite and satisfies (2.2).

3. Nonlinear estimation by linear Galerkin. Wavelet bases have been documented in numerous textbooks and survey papers (see [12] for a general treatment). With a little effort, they can be adapted to fairly general domains $\Omega \subset \mathbb{R}^d$ (see [7] for a survey of these adaptations as well as a discussion of the characterizations of function spaces on Ω by wavelet coefficients).

The wavelet decomposition of a function $f \in L^2$ takes the form

$$f = \sum_{\lambda \in \Gamma_{j_0}} \alpha_\lambda \varphi_\lambda + \sum_{j \geq j_0} \sum_{\lambda \in \nabla_j} f_\lambda \psi_\lambda, \tag{3.1}$$

where $(\varphi_\lambda)_{\lambda \in \Gamma_j}$ is the scaling function basis spanning the approximation at level j with appropriate boundary modification, and $(\psi_\lambda)_{\lambda \in \nabla_j}$ is the wavelet basis spanning the details at level j. The index λ concatenates the usual scale and space indices j and k. The coefficients of f can be evaluated according to

$$\alpha_\lambda = \langle f, \tilde{\varphi}_\lambda \rangle \text{ and } f_\lambda = \langle f, \tilde{\psi}_\lambda \rangle,$$

where $\tilde{\varphi}_\lambda$ and $\tilde{\psi}_\lambda$ are the corresponding dual scaling functions and wavelets. In what follows, we shall (merely for notational convenience) always take $j_0 := 0$. To simplify notation even more, we incorporate the first layer of scaling functions $(\varphi_\lambda)_{\lambda \in \Gamma_0}$ into the wavelet layer $(\psi_\lambda)_{\lambda \in \nabla_0}$ and define $\nabla = \cup_{j \geq 0} \nabla_j$, so that, if we write $|\lambda| = j$ if $\lambda \in \nabla_j$, we simply have

$$f = \sum_{\lambda \in \nabla} f_\lambda \psi_\lambda = \sum_{j=0}^{\infty} \sum_{|\lambda| = j} f_\lambda \psi_\lambda.$$

3.1. Preliminaries: Linear estimation by linear Galerkin. We first recall some classical results on the linear Galerkin projection method. For some scale $j > 0$ to be chosen further, let V_j be the linear space spanned by $(\varphi_\lambda)_{\lambda \in \Gamma_j}$. We define our first estimator $f_\varepsilon = \sum_{\gamma \in \Gamma_j} f_{\varepsilon,\gamma} \varphi_\gamma \in V_j$ as the unique solution of the finite dimensional linear problem

$$\text{find } f_\varepsilon \in V_j \text{ such that } \langle K f_\varepsilon, v \rangle = \langle g_\varepsilon, v \rangle \text{ for all } v \in V_j. \tag{3.2}$$

Defining the data vector $G_\varepsilon := (\langle g_\varepsilon, \varphi_\gamma \rangle)_{\gamma \in \Gamma_j}$ and the Galerkin stiffness matrix $K_j := (K\varphi_\gamma, \varphi_\mu\rangle)_{\gamma,\mu \in \Gamma_j}$, the coordinate vector $F_\varepsilon := (f_{\varepsilon,\gamma})_{\gamma \in \Gamma_j}$ of f_ε is therefore the solution of the linear system

$$K_j F_\varepsilon = G_\varepsilon. \tag{3.3}$$

The analysis of the method is summarized in the following classical result; see, for instance, [21], [22], [25], [15], and the references therein.

PROPOSITION 3.1. *Assuming that f belongs to the Sobolev ball $B := \{f \in H^s \,;\, \|f\|_{H^s} \leq M\}$ and choosing $j = j(\varepsilon)$ with $2^{-j(\varepsilon)} \sim \varepsilon^{2/(2s+2t+d)}$, we have*

$$\sup_{f \in B} E(\|f - f_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^{4s/(2s+2t+d)},$$

and this rate is minimax over the class B.

The symbol $\lesssim$ means that the left-hand side is bounded by a constant multiple of the right-hand side where the constant possibly depends on s and M, but not on

ε. In order to be self contained, we give a proof of Proposition 3.1. The techniques we use here will prove helpful in the sequel.

Proof. The analysis of this method can be done by decomposing f_ε according to

$$f_\varepsilon = f_j + h_\varepsilon, \tag{3.4}$$

where the terms f_j and h_ε are, respectively, solutions of (3.2) with Kf and $\varepsilon\dot{W}$ in place of g_ε as the right-hand side. This gives the classical decomposition of the estimation error into a bias and variance term

$$E(\|f - f_\varepsilon\|_{L^2}^2) = \|f - f_j\|_{L^2}^2 + E(\|h_\varepsilon\|_{L^2}^2). \tag{3.5}$$

Both terms are estimated by inverse and direct estimates with respect to Sobolev spaces, which are recalled in Appendix B. The variance term can be estimated as follows: we first use the ellipticity property (2.2), which gives

$$\|h_\varepsilon\|_{H^{-t/2}}^2 \lesssim \langle Kh_\varepsilon, h_\varepsilon\rangle = \varepsilon\langle \dot{W}, h_\varepsilon\rangle \le \varepsilon\|P_j\dot{W}\|_{L^2}\|h_\varepsilon\|_{L^2}. \tag{3.6}$$

Using the inverse inequality $\|g\|_{L^2} \lesssim 2^{tj/2}\|g\|_{H^{-t/2}}$ for all $g \in V_j$ and dividing by $\|h_\varepsilon\|_{L^2}$, we obtain

$$\|h_\varepsilon\|_{L^2} \lesssim \varepsilon 2^{tj}\|P_j\dot{W}\|_{L^2}, \tag{3.7}$$

and therefore

$$E(\|h_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^2 2^{2tj}\dim(V_j) \lesssim \varepsilon^2 2^{(2t+d)j}. \tag{3.8}$$

For the bias term, we take an arbitrary $g_j \in V_j$ and write

$$\begin{aligned}\|f - f_j\|_{L^2} &\le \|f - g_j\|_{L^2} + \|f_j - g_j\|_{L^2}\\ &\lesssim \|f - g_j\|_{L^2} + 2^{tj/2}\|f_j - g_j\|_{H^{-t/2}}\\ &\lesssim \|f - g_j\|_{L^2} + 2^{tj/2}\|f - g_j\|_{H^{-t/2}},\end{aligned}$$

where we have again used the inverse inequality and the fact that the Galerkin projection satisfies $\|f - f_j\|_{H^{-t/2}} \lesssim \|f - g_j\|_{H^{-t/2}}$ for any $g_j \in V_j$. It follows that

$$\|f - f_j\|_{L^2} \lesssim \inf_{g_j\in V_j}[\|f - g_j\|_{L^2} + 2^{tj/2}\|f - g_j\|_{H^{-t/2}}]. \tag{3.9}$$

Assuming that f belongs to B we obtain the direct estimate

$$\inf_{g_j\in V_j}[\|f - g_j\|_{L^2} + 2^{tj/2}\|f - g_j\|_{H^{-t/2}}] \lesssim 2^{-sj}, \tag{3.10}$$

and therefore

$$\|f - f_j\|_{L^2}^2 \lesssim 2^{-2sj}. \tag{3.11}$$

Balancing the bias and variance terms gives the optimal choice of resolution

$$2^{-j(\varepsilon)} \sim \varepsilon^{2/(2s+2t+d)}, \tag{3.12}$$

and the rate of convergence

$$E(\|f - f_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^{4s/(2s+2t+d)}, \tag{3.13}$$

which ends the proof of Proposition 3.1. □

3.2. Nonlinear estimation by linear Galerkin. Our first nonlinear estimator f_ε simply consists of applying a thresholding algorithm on the observed data before performing the linear Galerkin inversion which was described in the previous section: for some $j \geq 0$ to be chosen later, we define $f_\varepsilon = \sum_{|\lambda|<j} f_{\varepsilon,\lambda}\psi_\lambda \in V_j$ such that

$$\langle Kf_\varepsilon, \psi_\lambda\rangle = T_\varepsilon(\langle g_\varepsilon, \psi_\lambda\rangle) \tag{3.14}$$

for all $|\lambda| < j$. Here T_ε is the hard thresholding operator

$$T_\varepsilon(x) = x\chi(|x| \geq t(\varepsilon)) \tag{3.15}$$

(where $\chi(P)$ is 1 if P is true and 0 otherwise), and the threshold $t(\varepsilon)$ has the usual size

$$t(\varepsilon) := 8\varepsilon\sqrt{|\log\varepsilon|}. \tag{3.16}$$

Defining the data vector $G_\varepsilon := (\langle g_\varepsilon, \psi_\lambda\rangle)_{|\lambda|<j}$, and the Galerkin stiffness matrix $K_j := (\langle K\psi_\lambda, \psi_\mu\rangle)_{|\lambda|,|\mu|<j}$, the coordinate vector $F_\varepsilon := (f_{\varepsilon,\lambda})_{|\lambda|<j}$ of f_ε in the wavelet basis $(\psi_\lambda)_{|\lambda|<j}$ is the solution of the linear system

$$K_j F_\varepsilon = T_\varepsilon(G_\varepsilon), \tag{3.17}$$

where $T_\varepsilon(G_\varepsilon) := (T_\varepsilon(\langle g_\varepsilon, \psi_\lambda\rangle))_{|\lambda|<j}$. Note that such an estimator can be viewed as a variant of the vaguelette-wavelet estimator truncated at level j. Such an estimator would indeed be given (in the case where (ψ_λ) is an orthonormal basis) by

$$f_\varepsilon := \sum_{|\lambda|<j} T_\varepsilon(\langle g_\varepsilon, \psi_\lambda\rangle)K^{-1}\psi_\lambda. \tag{3.18}$$

The solution f_ε of (3.14) has a similar form with the vaguelettes $K^{-1}\psi_\lambda$ replaced by their Galerkin approximations $u_\lambda^j \in V_j$ such that

$$\langle Ku_\lambda^j, v\rangle = \langle \psi_\lambda, v\rangle \text{ for all } v \in V_j. \tag{3.19}$$

We therefore expect that this estimator behaves in the same optimal way as the VWD estimator provided that j is large enough. The following theorem shows that this is indeed true if $2^{-j} \leq \varepsilon^{1/t}$ where t is the degree of ill-posedness of the operator. It should be noted that the lower bound on j does not depend on the unknown smoothness of f, in contrast to the classical thresholding for signal denoising.

THEOREM 3.2. *Assume that f belongs to $B := \{f\ ;\ \|f\|_{B^s_{p,p}} \leq M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t + d)$. Assume in addition that K is an isomorphism between L^2 and H^t and that it has the smoothing property of order t with respect to the space $B^s_{p,p}$. Then the estimator from equation* (3.14) *satisfies the minimax rate*

$$\sup_{f\in B} E(\|f - f_\varepsilon\|_{L^2}^2) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}, \tag{3.20}$$

provided that j is such that $2^{-j} \leq \varepsilon^{1/t}$.

Proof. We write again $f_\varepsilon = f_j + h_\varepsilon$, where $f_j \in V_j$ is the solution of the linear problem with data g_ε

$$\text{find } f_j \in V_j \text{ such that } \langle Kf_j, v\rangle = \langle g, v\rangle \text{ for all } v \in V_j, \tag{3.21}$$

where $g = Kf$. Correspondingly, the term h_ε represents the solution of the linear problem with the thresholding error as data, in other words $h_\varepsilon \in V_j$ such that

$$\langle Kh_\varepsilon, \psi_\lambda\rangle = T_\varepsilon(\langle g_\varepsilon, \psi_\lambda\rangle) - \langle g, \psi_\lambda\rangle \tag{3.22}$$

for all $|\lambda| < j$. Similarly to the analysis described in the previous section, we need to estimate $\|f - f_j\|_{L^2}^2$ and $E(\|h_\varepsilon\|_{L^2}^2)$. For the deterministic term, we remark that the space $B_{p,p}^s$ is continuously imbedded in H^α whenever

$$\alpha \le s + d/2 - d/p = 2ts/(2t+d). \tag{3.23}$$

By the same arguments as in the previous section we obtain

$$\|f - f_j\|_{L^2}^2 \lesssim 2^{-4sj\frac{t}{d+2t}}. \tag{3.24}$$

This gives the optimal order $\varepsilon^{4s/(2s+2t+d)}$ if j is large enough so that

$$2^{-j} \le \varepsilon^{\frac{d+2t}{t(2s+2t+d)}}. \tag{3.25}$$

We have $\varepsilon^{1/t} \le \varepsilon^{\frac{d+2t}{t(2s+2t+d)}}$ for all $s \ge 0$. Therefore the choice $2^{-j} \le \varepsilon^{1/t}$ yields

$$\|f - f_j\|_{L^2}^2 \lesssim \varepsilon^{4s/(2s+2t+d)}. \tag{3.26}$$

We next turn to the stochastic term $E(\|h_\varepsilon\|_{L^2}^2)$. If H_ε is the coordinate vector of h_ε in the basis $(\psi_\lambda)_{|\lambda|<j}$, we want to estimate $E(\|H_\varepsilon\|_{\ell^2}^2)$. We write

$$H_\varepsilon = K_j^{-1}(T_\varepsilon(G_\varepsilon) - G_j) \tag{3.27}$$

and remark that $T_\varepsilon(G_\varepsilon) - G_j$ is exactly the error when estimating G_j by the thresholding procedure on the data G_ε. We shall take into account the action of K_j^{-1} by measuring this error in the wavelet version of the H^t norm

$$\|U\|_{h^t}^2 := \sum_{|\lambda|<j} 2^{2t|\lambda|}|u_\lambda|^2. \tag{3.28}$$

Indeed, we shall see that the stability property

$$\|K_j^{-1}U\|_{\ell^2} \lesssim \|U\|_{h^t} \tag{3.29}$$

holds under the assumption that K^{-1} maps H^t onto L^2. Our result will therefore follow from

$$E(\|T_\varepsilon(G_\varepsilon) - G_j\|_{h^t}^2) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}. \tag{3.30}$$

Such a rate is a particular case of classical results on wavelet thresholding, using the fact that g belongs to a Besov ball $\tilde{B} = \{g \in B_{p,p}^{s+t}\ ;\ \|g\|_{B_{p,p}^{s+t}} \le \tilde{M}\}$. For this model, (3.30) follows, e.g., from Theorem 4 in [10]. We are thus left with proving the stability property (3.29). To do so, we remark that if

$$K_j^{-1}U = V = (v_\lambda)_{|\lambda|<j}, \tag{3.31}$$

then the function $v = \sum_{|\lambda|<j} v_\lambda \psi_\lambda$, is the Galerkin approximation of $K^{-1}u$, where u is the function defined by

$$U = (\langle u, \psi_\lambda \rangle)_{|\lambda|<j} \text{ and } \langle u, \psi_\lambda \rangle = 0 \text{ if } |\lambda| \geq j. \tag{3.32}$$

It follows that

$$\|K^{-1}u - v\|_{H^{-t/2}} \lesssim 2^{-jt/2}\|K^{-1}u\|_{L^2} \lesssim 2^{-jt/2}\|u\|_{H^t}. \tag{3.33}$$

For the projection $P_j K^{-1}u$, we also have the error estimate

$$\|K^{-1}u - P_j K^{-1}u\|_{H^{-t/2}} \lesssim 2^{-jt/2}\|K^{-1}u\|_{L^2} \lesssim 2^{-jt/2}\|u\|_{H^t}. \tag{3.34}$$

It follows that

$$\|v - P_j K^{-1}u\|_{H^{-t/2}} \lesssim 2^{-jt/2}\|u\|_{H^t}. \tag{3.35}$$

Using the inverse estimate, we obtain

$$\|v - P_j K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t}, \tag{3.36}$$

so that

$$\|v\|_{L^2} \lesssim \|u\|_{H^t} + \|P_j K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t} + \|K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t}. \tag{3.37}$$

Using the wavelet characterization of L^2 and H^t, this yields (3.29). □

Remark. The assumption that K^{-1} maps H^t into L^2 which we are using in the above result is also implicit in the vaguelette-wavelet method when assuming that the vaguelettes

$$v_\lambda = \beta_\lambda K^{-1}\psi_\lambda = 2^{-t|\lambda|}K^{-1}\psi_\lambda \tag{3.38}$$

constitute a Riesz basis of L^2.

4. Nonlinear estimation by adaptive Galerkin. The main defect of the method described in the previous section remains its computational cost: the dimension of V_j is of order $N_j = 2^{dj} \sim \varepsilon^{-d/t}$ and might therefore be quite large. Moreover, in the case of an integral operator the stiffness matrix K_j might be densely populated. In this section we shall try to circumvent this problem by replacing the full Galerkin inversion by an adaptive algorithm which operates only in subspaces of V_j generated by appropriate wavelets and which exploits, in addition, the possibility of compressing the matrix K_j when discretized in the wavelet basis. Our estimator f_ε will therefore belong to an adaptive subspace of V_j

$$V_{\Lambda_\varepsilon} = \text{Span}\{\psi_\lambda \,;\, \lambda \in \Lambda_\varepsilon\}, \tag{4.1}$$

where Λ_ε is a data-driven subset of $\{|\lambda| < j\}$. A first intuitive guess for Λ_ε is the set obtained by the thresholding procedure applied on g_ε in the previous section, namely

$$\Lambda_\varepsilon := \{|\lambda| < j \,;\, |\langle g_\varepsilon, \psi_\lambda \rangle| \geq t(\varepsilon)\}. \tag{4.2}$$

It would thus be tempting to define $f_\varepsilon \in V_{\Lambda_\varepsilon}$ by applying the Galerkin inversion in this adaptive subspace:

$$\langle f_\varepsilon, \psi_\lambda \rangle = \langle g_\varepsilon, \psi_\lambda \rangle \text{ for all } \lambda \in \Lambda_\varepsilon. \tag{4.3}$$

However, it is by no means ensured that such an estimator f_ε will achieve the optimal convergence rate in the case of nonlocal operators K. Indeed, there are many instances of operator equations $Kf = g$ where the adapted wavelet set for the solution f differs significantly from the adapted set for the data g.

In order to build a better adapted set of wavelets, we shall introduce a level dependent thresholding operator S_ε to be applied in the solution domain (in contrast to T_ε which operates in the observation domain) according to

$$S_\varepsilon(u_\lambda) = u_\lambda \chi(|u_\lambda| \geq 2^{t|\lambda|} t(\varepsilon)). \tag{4.4}$$

The role of the weight $2^{t|\lambda|}$ is to take into account the amplification of the noise by the inversion process. The L^2-approximation error obtained by such level dependent thresholding procedures is well understood; see, in particular, Theorem 7.1 in [9], which implies that for $f = \sum_{\lambda \in \nabla} f_\lambda \psi_\lambda \in B^s_{p,p}$, with $s > 0$ and $1/p = 1/2 + s/(2t+d)$ and $S_\varepsilon(f) = \sum_{\lambda \in \nabla} S_\varepsilon(f_\lambda)\psi_\lambda$ we have

$$\begin{aligned} \|f - S_\varepsilon(f)\|^2_{L^2} &\sim \sum_{|f_\lambda| < 2^{t|\lambda|} t(\varepsilon)} |f_\lambda|^2 \\ &\lesssim \|f\|^2_{B^s_{p,p}} t(\varepsilon)^{2-p} = \|f\|^2_{B^s_{p,p}} t(\varepsilon)^{4s/(2s+2t+d)}. \end{aligned} \tag{4.5}$$

Our first result shows that S_ε is well adapted to build an adaptive solution of the inverse problem in the following sense: if we apply S_ε to the coordinates of the estimator f_ε defined in the previous section by (3.14), then the resulting estimator

$$S_\varepsilon(f_\varepsilon) := \sum_{|\lambda| < j} S_\varepsilon(f_{\varepsilon,\lambda})\psi_\lambda \tag{4.6}$$

still satisfies the optimal convergence rate.

THEOREM 4.1. *Let us assume that f belongs to $B := \{f \ ; \ \|f\|_{B^s_{p,p}} \leq M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t+d)$. Then, we have the estimate*

$$\sup_{f \in B} E(\|f_\varepsilon - S_\varepsilon(f_\varepsilon)\|^2_{L^2}) \ \lesssim \ [\varepsilon \sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)}. \tag{4.7}$$

It follows that the adaptive estimator $S_\varepsilon(f_\varepsilon) = \sum_{|\lambda|<j} S_\varepsilon(f_{\varepsilon,\lambda})\psi_\lambda$ is also rate-optimal.

Proof. We want to estimate the expectation of

$$\|f_\varepsilon - S_\varepsilon(f_\varepsilon)\|^2_{L^2} \ \lesssim \sum_{|\lambda|<j, |f_{\varepsilon,\lambda}| < 2^{t|\lambda|} t(\varepsilon)} |f_{\varepsilon,\lambda}|^2. \tag{4.8}$$

Using the fact that if $|a| \leq \eta$ we have for all real b

$$|a| \leq |a - b\chi(|b| \geq 2\eta)|, \tag{4.9}$$

we derive

$$\begin{aligned} \|f_\varepsilon - S_\varepsilon(f_\varepsilon)\|^2_{L^2} &\lesssim \textstyle\sum_{\lambda \in \nabla} |f_{\varepsilon,\lambda} - f_\lambda \chi(|f_\lambda| \geq 2^{t|\lambda|+1} t(\varepsilon))|^2 \\ &\lesssim \|f - f_\varepsilon\|^2_{L^2} + \textstyle\sum_{|f_\lambda| < 2^{t|\lambda|+1} t(\varepsilon)} |f_\lambda|^2 \\ &\lesssim \|f - f_\varepsilon\|^2_{L^2} + [\varepsilon \sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)}. \end{aligned}$$

Taking the expectation, we obtain (4.7) and

$$E(\|f - S_\varepsilon(f_\varepsilon)\|_{L^2}^2) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)} \tag{4.10}$$

follows by the triangle inequality. □

Of course, computing $S_\varepsilon(f_\varepsilon)$ is more costly than computing f_ε, and we cannot be satisfied with this new estimator. However, it shows us that the level-dependent thresholding operator S_ε maintains optimality. Based on this observation we now build an adaptive procedure which aims at reducing the computational cost. Let us note that many numerical methods are available in order to solve the system

$$K_j F_\varepsilon = T_\varepsilon(G_\varepsilon) \tag{4.11}$$

with the optimal cost $\mathcal{O}(N_j)$, where $N_j = \dim(V_j) \sim 2^{dj}$. In particular, one can rely on multigrid methods [3] in the case of local elliptic operators and fast multipole or wavelet [2] methods in the case of integral operators. However, our goal here is to reduce further the computational cost to the order of the dimension of the compressed solution, i.e., the number of nonzero coefficients in $S_\varepsilon(f_\varepsilon)$. Therefore, we shall rather be inspired by the approach introduced in [8] for adaptively solving operator equations *without noise*: consider a simple method for solving

$$K_j F_\varepsilon = T_\varepsilon(G_\varepsilon), \tag{4.12}$$

namely the fixed step gradient iteration $F_\varepsilon^0 = 0$ and

$$F_\varepsilon^n = F_\varepsilon^{n-1} + \tau(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1}) \tag{4.13}$$

with a sufficiently small enough relaxation parameter $\tau > 0$. The convergence rate of F_ε^n to F_ε might deteriorate for large j due to the bad condition number of K_j. Wavelet discretization is well adapted to circumvent this problem, when using the preconditioned iteration

$$F_\varepsilon^n = F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1}), \tag{4.14}$$

where $D_j = \mathrm{Diag}(2^{-t|\lambda|})$. From the ellipticity of K and the wavelet characterization of $H^{-t/2}$, it follows that the condition number $\kappa(D_j^{-1}K_j)$ remains bounded independently of j, so that a proper choice of τ will ensure a fixed error reduction rate

$$\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2} \le \rho\|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}, \tag{4.15}$$

with $\rho \in]0,1[$ independent of j. The idea is now to perturb this iteration by the thresholding operator S_ε, i.e., define

$$F_\varepsilon^n = S_\varepsilon[F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1})]. \tag{4.16}$$

At each step n, the vector $F_\varepsilon^n = (f_{\varepsilon,\lambda}^n)$ is supported on an adaptive index set Λ_ε^n. The corresponding estimator for f is given as

$$f_\varepsilon^n = \sum_{\lambda\in\Lambda_\varepsilon^n} f_{\varepsilon,\lambda}^n \psi_\lambda. \tag{4.17}$$

When comparing (4.16) with (4.14), we observe a first obvious gain in computational time: the cost of the matrix-vector multiplication $K_j F_\varepsilon^{n-1}$ in (4.16) is of

order $(\dim(V_j))^2 \sim 2^{2dj}$, while the cost of the matrix-vector multiplication $K_j F_\varepsilon^{n-1}$ in (4.16) is of order $\dim(V_j) \times \#(\Lambda_\varepsilon^n) \sim 2^{dj}\#(\Lambda_\varepsilon^n)$. Additional computational time can be gained using the fact that for many relevant instances of operators K, the matrix K_j can be compressed by discarding most of its entries. Such instances include in particular pseudodifferential operators and singular integral operators with Calderon–Zygmund-type kernel; see, e.g., Chapter 4 in [7] and [8]. For such operators, the entries $K_j(\lambda,\mu)$ can be estimated a priori, allowing us to predict in advance those coefficients in $F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1})$ which will be thresholded by S_ε and to avoid their exact computation. With such an approach, the cost of each iteration (4.16) can therefore be pushed down to the order $\#(\Lambda_\varepsilon^n)^2$, and even to $\#(\Lambda_\varepsilon^n)$ using a fast matrix vector multiplication; see Chapter 4 in [7] and [8].

We shall now prove that after a sufficient number of iterations independent of the unknown smoothness, the estimator f_ε^n attains the optimal rate of convergence.

THEOREM 4.2. *Let us assume that f belongs to $B := \{f\ ;\ \|f\|_{B^s_{p,p}} \le M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t+d)$. For $n \ge \log(\varepsilon)/\log(\rho)$, we have*

$$\sup_{f\in B} E(\|f_\varepsilon - f_\varepsilon^n\|_{L^2}^2) \ \lesssim\ [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}. \tag{4.18}$$

It follows that the adaptive estimator f_ε^n is also rate-optimal.

Proof. The result will follow from the reduction estimate

$$E(\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2) \le \tilde\rho^2 E(\|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}^2) + C[\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)} \tag{4.19}$$

for any $\tilde\rho > \rho$, where C depends of the closeness of $\tilde\rho$ to ρ. Indeed, assuming this estimate to hold, from

$$E(\|F_\varepsilon - F_\varepsilon^0\|_{\ell^2}^2) = E(\|F_\varepsilon\|_{\ell^2}^2) \ \lesssim\ \|F\|_{\ell^2}^2 \le C \tag{4.20}$$

we obtain after n steps

$$E(\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2) \ \lesssim\ \max\{\tilde\rho^{2n}, [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}\}. \tag{4.21}$$

Since $4s/(2s+2t+d) < 2$, we have

$$\tilde\rho^{2\log(\varepsilon)/\log(\rho)} = \varepsilon^{2\log(\tilde\rho)/\log(\rho)} \ \lesssim\ [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)} \tag{4.22}$$

if $\tilde\rho$ is chosen close enough to ρ, and (4.18) follows. In order to prove (4.19), we introduce the intermediate vector

$$F_\varepsilon^{n-1/2} = F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1}), \tag{4.23}$$

for which we have

$$\|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2} \le \rho\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}. \tag{4.24}$$

We can then write

$$\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2 = \sum_{|\lambda|<j} |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \ge 2^{t|\lambda|}t(\varepsilon))|^2. \tag{4.25}$$

Denoting by $K > 1$ a constant to be fixed later, we split the above sum into three parts Σ_1, Σ_2, and Σ_3, respectively corresponding to the index sets

$$\begin{aligned}
I_1 &:= \{|\lambda| < j\ ;\ |f_{\varepsilon,\lambda}^{n-1/2}| < 2^{t|\lambda|}t(\varepsilon) \text{ and } |f_{\varepsilon,\lambda}| < K2^{t|\lambda|}t(\varepsilon)\},\\
I_2 &:= \{|\lambda| < j\ ;\ |f_{\varepsilon,\lambda}^{n-1/2}| \ge 2^{t|\lambda|}t(\varepsilon)\},\\
I_3 &:= \{|\lambda| < j\ ;\ |f_{\varepsilon,\lambda}^{n-1/2}| < 2^{t|\lambda|}t(\varepsilon) \text{ and } |f_{\varepsilon,\lambda}| \ge K2^{t|\lambda|}t(\varepsilon)\}.
\end{aligned}$$

If $\lambda \in I_1$, we have $|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda|}t(\varepsilon))| = |f_{\varepsilon,\lambda}|$. Using again the fact that if $|a| \leq \eta$ we have $|a| \leq |a - b\chi(|b| \geq 2\eta)|$ for all b, we can write

$$\begin{aligned} |f_{\varepsilon,\lambda}| &\leq |f_{\varepsilon,\lambda} - f_\lambda \chi(|f_\lambda| \geq 2K2^{t|\lambda|}t(\varepsilon))| \\ &\leq |f_{\varepsilon,\lambda} - f_\lambda| + |f_\lambda - f_\lambda\chi(|f_\lambda| \geq 2K2^{t|\lambda|}t(\varepsilon))|. \end{aligned}$$

It follows that

$$\begin{aligned} \Sigma_1 &= 2\|f_\varepsilon - f\|_{L^2}^2 + 2\textstyle\sum_{|f_\lambda| \leq 2K2^{t|\lambda|}t(\varepsilon)} |f_\lambda|^2 \\ &\lesssim \|f_\varepsilon - f\|_{L^2}^2 + [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}, \end{aligned}$$

so that

$$E(\Sigma_1) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}. \tag{4.26}$$

If $\lambda \in I_2$, we have

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda|}t(\varepsilon))| = |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}|, \tag{4.27}$$

so that

$$\Sigma_2 = \|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2(\Lambda_1).}^2 \tag{4.28}$$

If $\lambda \in I_3$, we have $|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda|}t(\varepsilon))| = |f_{\varepsilon,\lambda}|$ and

$$|f_{\varepsilon,\lambda}| \leq |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| + |f_{\varepsilon,\lambda}^{n-1/2}| \leq |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| + 2^{t|\lambda|}t(\varepsilon). \tag{4.29}$$

On the other hand, since $|f_{\varepsilon,\lambda}| > K2^{t|\lambda|}t(\varepsilon)$ and $|f_{\varepsilon,\lambda}^{n-1/2}| < 2^{t|\lambda|}t(\varepsilon)$, we also have

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| \geq (K-1)2^{t|\lambda|}t(\varepsilon). \tag{4.30}$$

It follows that

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda|}t(\varepsilon))| < \tfrac{K}{K-1}|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}|, \tag{4.31}$$

so that

$$\Sigma_3 \leq (\tfrac{K}{K-1})^2\|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2(\Lambda_3)}^2. \tag{4.32}$$

Combining (4.28) and (4.32), we obtain

$$\Sigma_2 + \Sigma_3 \leq (\tfrac{K}{K-1})^2\|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2}^2 \leq (\tfrac{K}{K-1})^2\rho^2\|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}^2. \tag{4.33}$$

Combined with (4.26), this yields the claimed estimate (4.19) with $\tilde{\rho} = \frac{K}{K-1}\rho$, which can be made arbitrarily close to ρ by taking K large enough, up to enlarging the constant C which comes from the estimation of Σ_1. □

Remark. As was already explained, the cost of each iteration can be, at most, pushed down to $\mathcal{O}(\#(\Lambda_\varepsilon^n))$, which allows the estimate of the computational cost of the algorithm by

$$\mathcal{C}(\varepsilon) \leq \sum_{n \leq \log(\varepsilon)/\log(\rho)} \#(\Lambda_\varepsilon^n). \tag{4.34}$$

In addition, a rough estimate of $\#(\Lambda_\varepsilon^n)$ can be obtained from the smoothness of f, assuming that the number of coefficients retained at each thresholding step is of the same order as the number of coefficients which would be retained when applying S_ε to the exact f. In order to estimate this number, we note that if f belongs to $B := \{f\ ;\ \|f\|_{B^s_{p,p}} \le M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t+d)$, it also belongs to $\tilde{B} := \{f\ ;\ \|f\|_{B^s_{q,q}} \le M\}$ with $1/q = 1/2 + (s+t)/d$ (since $q \le p$), or equivalently

$$\sum \|f_\lambda \psi_\lambda\|_{H^{-t}}^q \sim \sum |f_\lambda 2^{-t|\lambda|}|^q \le M^q. \tag{4.35}$$

It follows that

$$\#(\Lambda_\varepsilon^n) \ \lesssim\ t(\varepsilon)^{-q} \sim [\varepsilon\sqrt{|\log\varepsilon|}]^{-\frac{2d}{d+2s+2t}}, \tag{4.36}$$

and therefore

$$\mathcal{C}(\varepsilon) \ \lesssim\ [\varepsilon\sqrt{|\log\varepsilon|}]^{-\frac{2d}{d+2s+2t}} \log(\varepsilon)/\log(\rho). \tag{4.37}$$

In contrast, the cost of a nonadaptive inversion of (4.11) when using an optimal solver is of order

$$\mathcal{C}(\varepsilon) \sim \dim(V_j) \sim 2^{dj} \ \lesssim\ \varepsilon^{-d/t}. \tag{4.38}$$

Since $\frac{2d}{d+2s+2t} < \frac{d}{t}$, we see that (4.37) always improves (4.38). Let us insist on the fact that this improvement relies on the compressibility of the stiffness matrix in the sense that the adaptive matrix-vector multiplication involved in the iteration only costs $\mathcal{O}(\#(\Lambda_\varepsilon^n))$ operations. For arbitrary noncompressible matrices, this cost should be multiplied by 2^{jd}. Note also that the cost for nonadaptive inversion may then also be substantially higher than $\dim(V_j)$.

Note also that, in addition to the fact that the algorithm adapts to unknown smoothness, its practical computational cost also decreases as the amount of smoothness increases.

5. A numerical example. We focus on a simple example of a logarithmic potential kernel in dimension one and a single test function. The relatively simple analytical form of this operator gives the ability to approximate reasonably well its singular values. Therefore, the SVD method is feasible and can be compared with the Galerkin approach with reasonable accuracy. Our goals are the following:

1. To illustrate on a specific test case that (1) the oracle-SVD and the oracle linear Galerkin methods are comparable; (2) the nonlinear Galerkin method of section 4, obtained by thresholding the data in the image domain, achieves comparable numerical results as the oracle-SVD and the oracle linear Galerkin estimators.
2. To verify on an example that the empirical L^2 error stabilizes beyond a certain resolution level j, which is related to the noise level ε and the degree of ill-posedness of the operator. In theory, we know that the condition $2^{-j} \ \lesssim\ \varepsilon^{1/t}$ is sufficient to obtain optimality and that there is no gain in increasing j further.
3. To verify on an example that if we threshold further by S_ε the estimator obtained by the nonlinear Galerkin inversion of section 4, we still have a good estimator, as predicted by our Theorem 4.1. This suggests that the iterative adaptive Galerkin method described in Theorem 4.2 shall be effective when dealing with more precise numerical studies.

A logarithmic potential integral operator. We consider a single-layer logarithmic potential operator that relates the density of electric charge on an infinite cylinder of a given radius $r > 0$: $\{(z, re^{i2\pi x}), z \in \mathbb{R}, x \in [0,1]\}$ to the induced potential on the same cylinder, when both functions are independent of the variable z. The associated kernel $k(x,y)$ of the operator that we take is

$$Kf(x) = \int_0^1 k(x,y)f(y)dy, \quad k(x,y) = -\log\left(r|e^{i2\pi x} - e^{i2\pi y}|\right) \tag{5.1}$$

for some $r > 0$. We choose $r = \frac{1}{4}$ and we rewrite (5.1) as

$$k(x,y) = -\log\left[\tfrac{1}{2}|\sin\pi(y-x)|\right], \quad x, y \in [0,1], \tag{5.2}$$

so that $k(x,y) \geq 0$ on the unit square. It is singular on the diagonal $\{x = y\}$ but integrable. The single layer potential is known to be an elliptic operator of order -1, which maps $H^{-1/2}$ into $H^{1/2}$ (see [7]). So the assumptions on K are satisfied with $t = 1$.

For the maximal resolution level $J \leq 15$ we discretize K by computing the matrix K_J with entries

$$(K_J)_{m,n=0,\dots,2^J-1} = (\langle K_J\varphi_{J,m}, \varphi_{J,n}\rangle)_{m,n=0,\dots,2^J-1},$$

where the $\varphi_{J,m} = 2^{J/2}1_{[m2^{-J},(m+1)2^{-J})}$ are the Haar functions. Each

$$\langle K_J\varphi_{J,m}, \varphi_{J,n}\rangle = \int_0^1\int_0^1 k(x,y)\varphi_{J,m}(x)\varphi_{J,n}(y)dxdy$$

is computed by midpoint rule at scale 2^{-18}. It is noteworthy that k is a periodic convolution kernel. In turn the discretization K_{15} of K is a Toeplitz cyclic matrix, of the form $K_J(m,n) = K_J\big((m-n)[\mathrm{mod}\ 2^J]\big)$. As a consequence, the fast Fourier transform diagonalizes the matrix K_J, which makes the computation of its singular values an easy numerical task. We take K_{15} as a proxy for K and let the level of analysis of our method vary for $j = 1, \dots, 15$. We consider the test function f, defined for $x \in [0,1]$ by

$$f(x) = \max\{1 - |30(x - \tfrac{1}{2})|, 0\}.$$

The piky function f is badly approximated by the singular functions of K, but it has a sparse representation in a wavelet basis, so the Galerkin method shall be more effective for the estimation problem.

Methodology. We first pick the maximal resolution level $J := 12$, a noise level $\varepsilon := 2 \cdot 10^{-4}$, and a single typical sample of a white noise process $w_{12} = (w_{k,12})_{k=0,\dots,2^{12}-1}$. This means that the $w_{k,12}$ are outcomes of independent identically distributed standard Gaussian random variables that contaminate the action of K_{12} on f, up to the noise level $\varepsilon = 2 \cdot 10^{-4}$. Figure 1 shows the true signal f (dash-dotted) together with the data process.

Let us recall that given a family of estimators $\hat{f}_j$ depending on a tuning constant $j = 1, \dots, j_{max}$ (here, $\hat{f}_j$ is constructed with the SVD or the linear Galerkin method, and j varies from level 1 to level 12), the oracle estimator $\hat{f}^*$ is defined as $\hat{f}^* = \hat{f}_{j_*}$, where

$$j_* := \mathrm{argmin}_{j=1,\dots,j_{max}}\|\hat{f}_j - f\|_{L^2}.$$

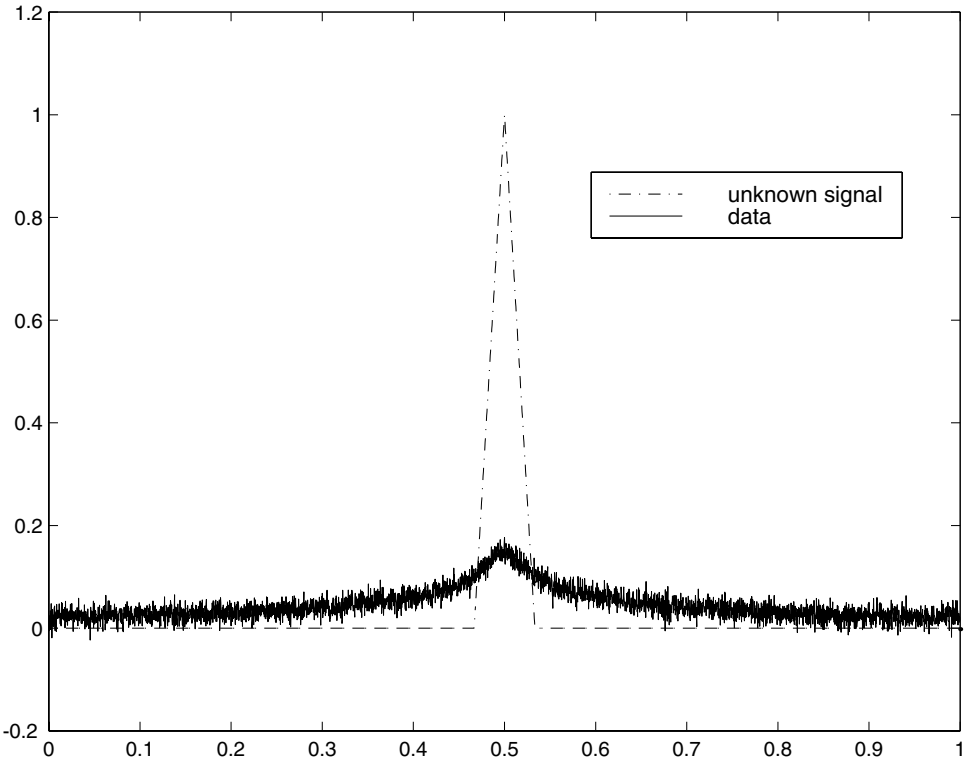


FIG. 1. *True function f and noisy signal.*

TABLE 1

	SVD oracle		Linear Galerkin		Nonlinear Galerkin
	j_*	L^2-error	j_*	L^2-error	L^2-error
f	5	$5.66 \cdot 10^{-4}$	5	$5.31 \cdot 10^{-4}$	$3.75 \cdot 10^{-4}$

Note that, strictly speaking, $\hat{f}^*$ is not an estimator (the ideal level j_* depends on the unknown) and appears as a benchmark for the method at hand. For technical reasons, we have replaced the L^2-risk by its empirical version.

Numerical results. The oracle estimator $\hat{f}^*$, for the SVD is displayed in Figure 2 and the oracle linear Galerkin estimator is displayed in Figure 3. In Figure 4, we show the performance of the nonlinear Galerkin estimator (Method 3), when applying a threshold $T_\varepsilon(\cdot)$ in the observation domain, specified with $t(\varepsilon) := 8 \cdot 10^{-4}$. We take a wavelet filter corresponding to compactly supported Daubechies wavelets of order 14. The numerical results of the three methods are summarized in Table 1.

Compression rate and approximation results. In the same context, we next investigate (see Figure 5) the performance of the nonlinear Galerkin estimator when applying further the level dependent thresholding operator $S_\varepsilon(\cdot)$, recall (4.4), with $t(\varepsilon) := 8 \cdot 10^{-4}$. We also indicate the number of wavelet coefficients put to zero divided by the total number of coefficients. The very high compression rate (see Table 2) that still ensures a small estimation error advocates in favor of the iterative adaptive scheme of section 4.

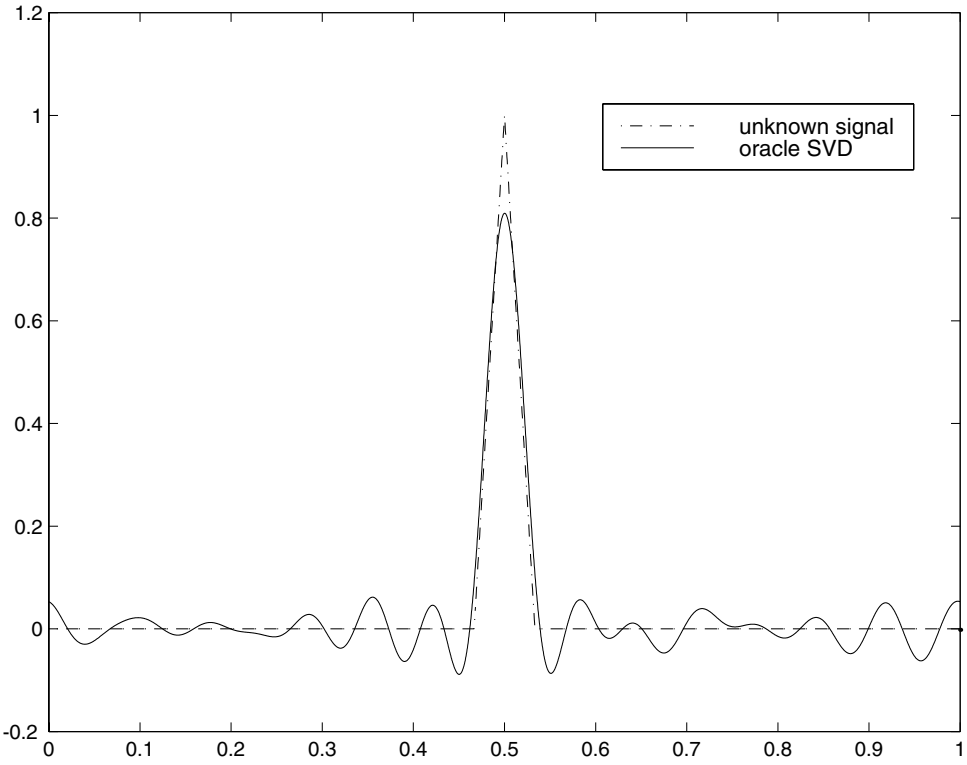


FIG. 2. *Oracle SVD.*

TABLE 2

	L^2-error, T_ε only	L^2-error, T_ε and S_ε	comp. rate
f	$3.75 \cdot 10^{-4}$	$4.12 \cdot 10^{-4}$	0.996

On a visual level, we observe that the nonlinear methods avoid the persistence of high oscillations far away from the singularity, in contrast to the linear methods. However, we still observe an artifact on the right side of the central peak. We hope to remedy this defect by (i) using biorthogonal spline wavelets instead of Daubechies orthogonal wavelets and (ii) apply a translation-invariant processing as introduced in [11] for denoising.

Appendix A. Extension to nonelliptic operators. In this appendix, we shall briefly explain how the methods and results that we have presented throughout can be extended by the mean-square approach to the case where K is not an elliptic operator. Here, the smoothing property of order t is expressed by the ellipticity property of the normal operator

$$\|Kf\|_{L^2}^2 = \langle K^*Kf, f\rangle \sim \|f\|_{H^{-t}}^2. \tag{A.1}$$

We discuss the adaptation of sections 3 and 4 to this more general context.

Linear Galerkin estimation. The method becomes the Galerkin projection

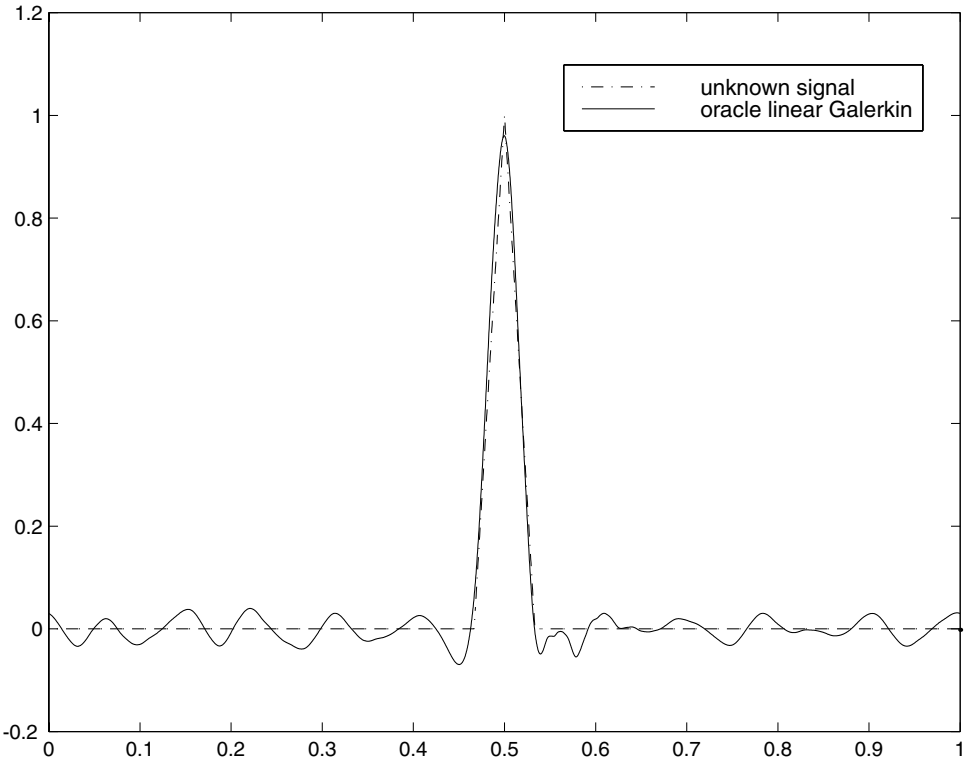


FIG. 3. *Oracle linear Galerkin method.*

method applied to the normal equation $K^*Kf = K^*g$. It therefore reads as follows:

$$\text{find } f_\varepsilon \in V_j \text{ such that } \langle Kf_\varepsilon, Kv\rangle = \langle g_\varepsilon, Kv\rangle \text{ for all } v \in V_j. \tag{A.2}$$

As in the elliptic case, we can use the decomposition $f_\varepsilon = f_j + h_\varepsilon$ in order to estimate the mean-square error according to

$$E(\|f - f_\varepsilon\|_{L^2}^2) \lesssim \|f - f_j\|_{L^2}^2 + E(\|h_\varepsilon\|_{L^2}^2). \tag{A.3}$$

For the variance term, we write

$$\begin{aligned}\|h_\varepsilon\|_{H^{-t}}^2 &\sim \langle K^*Kh_\varepsilon, h_\varepsilon\rangle = \varepsilon\langle \dot{W}, Kh_\varepsilon\rangle \\ &\le \varepsilon\|P_j^K\dot{W}\|_{L^2}\|Kh_\varepsilon\|_{L^2} \lesssim \varepsilon\|P_j^K\dot{W}\|_{L^2}\|h_\varepsilon\|_{H-t},\end{aligned}$$

where P_j^K is the orthogonal projector onto KV_j. Using the inverse inequality which states that $\|h_\varepsilon\|_{L^2} \lesssim 2^{tj}\|h_\varepsilon\|_{H^{-t}}$, we therefore obtain

$$\|h_\varepsilon\|_{L^2} \lesssim \varepsilon 2^{tj}\|P_j^K\dot{W}\|_{L^2}, \tag{A.4}$$

and therefore

$$E(\|h_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^2 2^{2tj}\dim(KV_j) \lesssim \varepsilon^2 2^{(2t+d)j}. \tag{A.5}$$

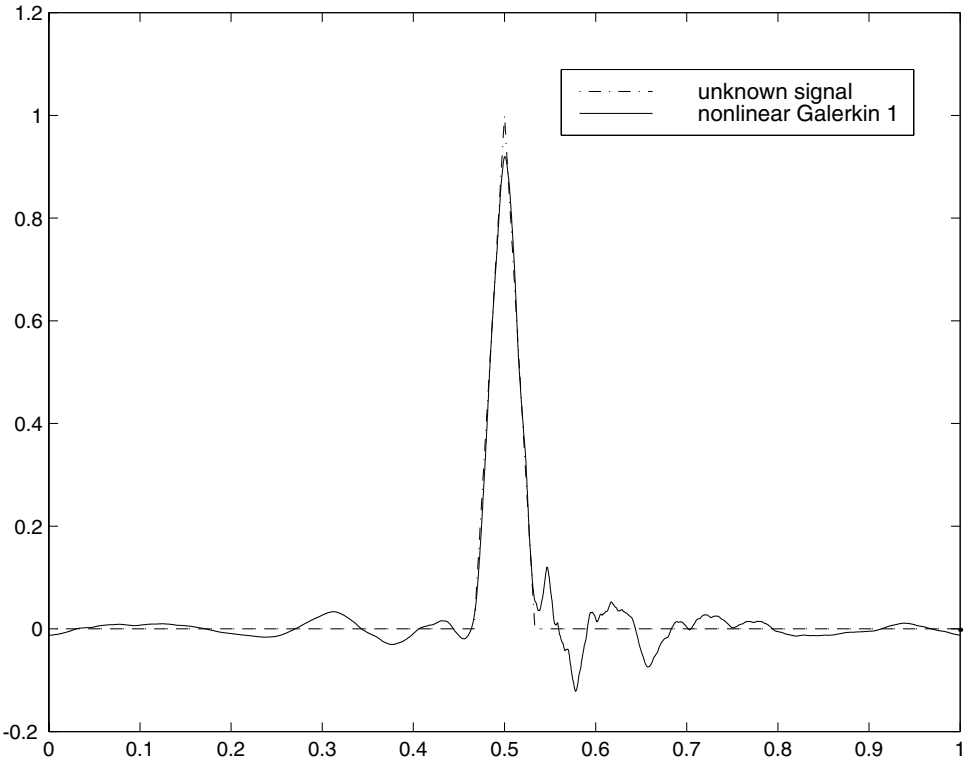


FIG. 4. *Nonlinear Galerkin estimator, thresholding in the image domain.*

For the bias term, we take any $g_j \in V_j$ and write

$$\begin{aligned} \|f - f_j\|_{L^2} &\lesssim \|f - g_j\|_{L^2} + 2^{tj}\|f_j - g_j\|_{H^{-t}} \\ &\lesssim \|f - g_j\|_{L^2} + 2^{tj}\|f - g_j\|_{H^{-t}}, \end{aligned}$$

where we have used the inverse inequality and Galerkin orthogonality. Assuming that f belongs to a Sobolev ball $B := \{f \in H^s \ ; \ \|f\|_{H^s} \leq M\}$, we obtain from approximation theory the direct estimate

$$\|f - f_j\|_{L^2}^2 \lesssim \inf_{g_j \in V_j} [\|f - g_j\|_{L^2} + 2^{tj}\|f - g_j\|_{H^{-t}}] \lesssim 2^{-sj}. \tag{A.6}$$

Proceeding as in section 3, we therefore achieve the same estimate for $E(\|f - f_\varepsilon\|_{L^2}^2)$ under the same assumptions as in the elliptic case.

Nonlinear estimation by linear Galerkin. The method becomes the Galerkin projection applied to the normal equation after thresholding the observed data. It therefore reads as follows: find $f_\varepsilon = \sum_{|\lambda|<j} f_{\varepsilon,\lambda}\psi_\lambda \in V_j$ such that

$$\langle Kf_\varepsilon, K\psi_\lambda\rangle = \tilde{T}_\varepsilon(\langle g_\varepsilon, K\psi_\lambda\rangle) \tag{A.7}$$

for all $|\lambda| < j$. Here the thresholding operator $\tilde{T}_\varepsilon$ differs from T_ε since it is applied to the wavelet coefficients of $K^* g_\varepsilon$. More precisely, we define

$$\tilde{T}_\varepsilon(d_\lambda) = d_\lambda \chi(|d_\lambda| \geq 2^{-t|\lambda|} t(\varepsilon)) = 2^{-t|\lambda|} T_\varepsilon(2^{t|\lambda|} d_\lambda), \tag{A.8}$$

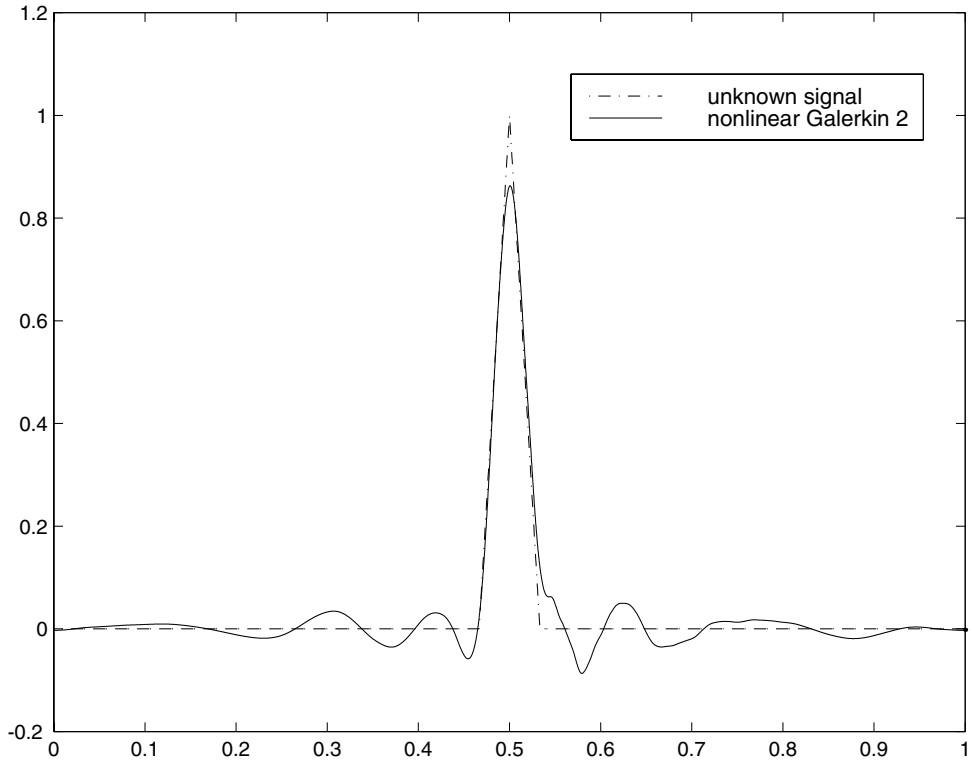


FIG. 5. *Nonlinear Galerkin estimator, thresholding in the image and solution domain.*

again with $t(\varepsilon) = C\varepsilon\sqrt{|\log\varepsilon|}$. With this method, Theorem 3.2 can be extended to the nonelliptic case, provided that we now use the assumption that K^*K is an isomorphism from L^2 to H^{2t} and has the smoothing property of order $2t$ with respect to the space $B^s_{p,p}$. For the proof of this result, we again write $f_\varepsilon = f_j + h_\varepsilon$, where the term h_ε now represents the solution of the linear problem with the thresholding error as data. By the same analysis as in the proof of Theorem 3.2, we obtain

$$\|f - f_j\|^2_{L^2} \lesssim \varepsilon^{4s/(2s+2t+d)} \tag{A.9}$$

if j is large enough so that $2^{-j} \le \varepsilon^{1/t}$. For the stochastic term, if H_ε is the coordinate vector of h_ε in the basis $(\psi_\lambda)_{|\lambda|<j}$, we write

$$H_\varepsilon = L_j^{-1} D_j (T_\varepsilon(\tilde{G}_\varepsilon) - \tilde{G}_j), \tag{A.10}$$

where $L_j := (\langle K\psi_\lambda, K\psi_\mu\rangle)_{|\lambda|,|\mu|<j}$ is now the Galerkin matrix for the least-squares formulation, $\tilde{G}_j := (2^{t|\lambda|}\langle K^*g, \psi_\lambda\rangle)_{|\lambda|<j}$, $\tilde{G}_\varepsilon := (2^{t|\lambda|}\langle K^*g_\varepsilon, \psi_\lambda\rangle)_{|\lambda|<j}$, and again $D_j := \mathrm{Diag}(2^{-t|\lambda|})$. In this case, we invoke the stability property

$$\|L_j^{-1}U\|_{\ell^2} \lesssim \|U\|_{h^{2t}}, \tag{A.11}$$

which is proved by similar argument as in the proof of Theorem 3.2. Since D_j is an i somorphism from h^t to h^{2t}, we are therefore left to prove that

$$E(\|T_\varepsilon(\tilde{G}_\varepsilon) - \tilde{G}_j\|^2_{h^t}) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}. \tag{A.12}$$

The components of $\tilde{G}_\varepsilon$ are related to those of $\tilde{G}_j$ by

$$\tilde{g}_{\varepsilon,\lambda} := 2^{t|\lambda|}\langle K^*Kf, \psi_\lambda\rangle + \varepsilon\langle \dot{W}, 2^{t|\lambda|}K\psi_\lambda\rangle = \tilde{g}_{j,\lambda} + \varepsilon\eta_\lambda, \tag{A.13}$$

where the η_λ are normalized Gaussian variables since

$$\|2^{t|\lambda|}K\psi_\lambda\|_{L^2} \sim 2^{t|\lambda|}\|\psi_\lambda\|_{H^{-t}} \sim 1. \tag{A.14}$$

Therefore, the estimate (A.12) again follows from classical results on wavelet thresholding such as Theorem 4 in [10] using the fact that K^*Kf belongs to a Besov ball $\tilde{B} = \{h \in B^{s+2t}_{p,p}\ ;\ \|h\|_{B^{s+2t}_{p,p}} \le \tilde{M}\}$.

Nonlinear estimation by adaptive Galerkin. The iterative method becomes

$$F^n_\varepsilon = S_\varepsilon[F^{n-1}_\varepsilon + \tau D_j^{-1}(\tilde{T}_\varepsilon(G_\varepsilon) - L_jF^{n-1}_\varepsilon)], \tag{A.15}$$

and the statements of Theorems 4.1 and 4.2 remain valid with the same proof.

Appendix B. Direct and inverse inequalities for multiresolution spaces. Direct and inverse inequalities are a key ingredient in multiresolution approximation theory. In their simplest form, the direct inequality reads as follows:

$$\inf_{g_j\in V_j} \|f - g_j\|_{L^2} \lesssim 2^{-sj}|f|_{H^s}, \tag{B.1}$$

and the inverse estimate states that for all $g_j \in V_j$

$$|g_j|_{H^s} \lesssim 2^{sj}\|g_j\|_{L^2}. \tag{B.2}$$

The proof of such estimates is quite classical and we refer the reader to Chapter 3 in [7]. Basically, the validity of the direct inequality requires that the spaces V_j have enough approximation power, in the sense that polynomials of degree m are contained in V_j for all $m < s$. On the other hand, the validity of the inverse estimate requires that the functions of V_j have enough smoothness in the sense that they are contained in H^s. The direct and inverse estimate which have been used in section 3 are less standard since they involve the Sobolev space of negative order $H^{-t/2}$, and we shall therefore briefly discuss their validity. The inverse estimate states that for all $g_j \in V_j$,

$$\|g_j\|_{L^2} \lesssim 2^{tj/2}\|g_j\|_{H^{-t/2}}. \tag{B.3}$$

We prove it by a duality argument:

$$\begin{aligned}\|g_j\|_{L^2} &= \sup_{f_j\in V_j, \|f_j\|_{L^2}=1} |\langle g_j, f_j\rangle| \\ &\lesssim \sup_{f_j\in V_j, \|f_j\|_{L^2}=1} \|g_j\|_{H^{-t/2}}\|f_j\|_{H^{t/2}} \\ &\lesssim 2^{tj/2}\|g_j\|_{H^{-t/2}},\end{aligned}$$

where we have used the standard inverse estimate (B.2) with $s = t/2$. The direct estimate states that

$$\inf_{g_j\in V_j} [\|f - g_j\|_{L^2} + 2^{tj/2}\|f - g_j\|_{H^{-t/2}}] \lesssim 2^{-sj}\|f\|_{H^s}. \tag{B.4}$$

In order to prove it, we take $g_j = P_jf$ where P_j is the L^2-orthogonal projector onto V_j. Clearly the first part $\|f - P_jf\|_{L^2} \lesssim 2^{-sj}\|f\|_{H^s}$ is simply the standard direct

estimate (B.1). For the second part, we write

$$\begin{aligned}\|f - P_j f\|_{H^{-t/2}} &= \sup_{\|g\|_{H^{t/2}}=1} |\langle f - P_j f, g\rangle| \\ &= \sup_{\|g\|_{H^{t/2}}=1} |\langle f - P_j f, g - P_j g\rangle| \\ &= \|f - P_j f\|_{L^2} \sup_{\|g\|_{H^{t/2}}=1} \|g - P_j g\|_{L^2} \\ &\sim 2^{-sj}\|f\|_{H^s} 2^{-tj/2},\end{aligned}$$

where we have used the fact that $(I - P_j)^2 = (I - P_j)^*(I - P_j) = I - P_j$ and the standard direct estimate (B.1) both for H^s and $H^{t/2}$.

Remark. The type of duality argument that we have used in order to prove both (B.3) and (B.4) can be generalized in such a way that the standard direct and inverse estimate between L^2 and $H^{t/2}$ are invoked for a dual space $\tilde{V}_j$ which might differ from V_j. For the direct estimate, this means that we take for P_j a more general biorthogonal projector, such that P_j^* is a projector onto $\tilde{V}_j$ (see [7] for examples of dual spaces and biorthogonal projectors), so that we are led to apply a standard direct inequality of the type

$$\|g - P_j^* g\|_{L^2} \lesssim 2^{-tj/2}\|g\|_{H^{t/2}} \tag{B.5}$$

which only requires polynomial exactness up to order $t/2$ for $\tilde{V}_j$. For the inverse estimate, we can also use the space $\tilde{V}_j$ in order to evaluate the L^2 norm according to

$$\|g_j\|_{L^2} \lesssim \sup_{\tilde{f}_j \in \tilde{V}_j, \|\tilde{f}_j\|_{L^2}=1} |\langle g_j, \tilde{f}_j\rangle| \lesssim \sup_{\tilde{f}_j \in \tilde{V}_j, \|\tilde{f}_j\|_{L^2}=1} \|g_j\|_{H^{-t/2}} \|\tilde{f}_j\|_{H^{t/2}}, \tag{B.6}$$

so that we are led to apply a standard inverse inequality of the type

$$\|\tilde{f}_j\|_{H^{t/2}} \lesssim 2^{tj/2}\|\tilde{f}_j\|_{L^2}, \tag{B.7}$$

which only requires that the space $\tilde{V}_j$ has $H^{t/2}$ smoothness. This last point is practically important, since it means that we are not enforced to use multiresolution spaces V_j consisting of smooth functions, neither are we forced to use smooth wavelets in the nonlinear methods. In contrast, it is crucial that the spaces V_j have enough polynomial reproduction ($\Pi_m \subset V_j$ for all $m < s$) in order to apply the direct estimate for H^s, and it is crucial for the nonlinear method that the wavelets ψ_λ have enough vanishing moments ($\int x^m \psi_\lambda = 0$ for all $m < s + t$) in order to apply the results on wavelet thresholding such as (3.30).

REFERENCES

[1] F. Abramovich and B.W. Silverman, *Wavelet decomposition approaches to statistical inverse problems*, Biometrika, 85 (1998), pp. 115–129.

[2] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms*, I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[3] J.H. Bramble, *Multigrid Methods*, Longman Scientific and Technical, Harlow, UK, 1993.

[4] L. Brown, T. Cai, M. Low, and C.H. Zhang, *Asymptotic equivalence theory for nonparametric regression with random design*, Ann. Statist., 30 (2002), pp. 688–707.

[5] S. Champier and L. Grammont, *A wavelet-vaguelet method for unfolding sphere size distributions*, Inverse Problems, 18 (2002), pp. 79–94.

[6] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, New York, Oxford, 1978.

[7] A. Cohen, *Wavelet methods in numerical analysis*, in Handbook of Numerical Analysis, Vol. VII, P.G. Ciarlet and J.L. Lions, eds., Elsevier, Amsterdam, 2000, pp. 417–711.

[8] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations: Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.
[9] A. Cohen, R. DeVore, and R. Hochmuth, *Restricted nonlinear approximation*, Constr. Approx., 16 (2000), pp. 85–113.
[10] A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard, *Maximal spaces with given rate of convergence for thresholding algorithms*, Appl. Comput. Harmon. Anal., 11 (2001), pp. 167–191.
[11] R.R. Coifman and D.L. Donoho, *Translation-invariant de-noising*, in Wavelets and Statistics, Lecture Notes in Statist. 103, A. Antoniadis and G. Oppenheim, eds., Springer, New York, 1995, pp. 125–150.
[12] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
[13] I. Daubechies, M. Defrise, and C. De Mol, *An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint*, preprint, Princeton University, Princeton, NJ, 2003.
[14] R. DeVore, *Nonlinear approximation*, Acta Numerica, 7 (1998), pp. 51–150.
[15] V. Dicken and P. Maass, *Wavelet-Galerkin methods for ill-posed problems*, J. Inverse Ill-Posed Probl., 4 (1996), pp. 203–221.
[16] D. Donoho, *Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 101–126.
[17] D.L. Donoho and I.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.
[18] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage: Asymptopia?*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 301–369.
[19] H.W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Press, Dordrecht, The Netherlands, 1996.
[20] T. Hida, *Brownian Motion*, Springer-Verlag, New York, 1980.
[21] I.M. Johnstone and B.W. Silverman, *Speed of estimation in positron emission tomography and related inverse problems*, Ann. Statist., 18 (1990), pp. 251–280.
[22] I.M. Johnstone and B.W. Silverman, *Discretization effects in statistical inverse problems*, J. Complexity, 7 (1991), pp. 1–34.
[23] A. Korostelev and A. Tsybakov, *Minimax Theory of Image Reconstruction*, Lecture Notes in Statist. 82, Springer-Verlag, New York, 1993.
[24] B. Mair and F. Ruymgaart, *Statistical inverse estimation in Hilbert scales*, SIAM J. Appl. Math., 56 (1996), pp. 1424–1444.
[25] F. Natterer, *The Mathematics of Computerized Tomography*, Classics Appl. Math. 32, SIAM, Philadelphia, 2001.
[26] M. Nussbaum and S. Pereverzev, *The Degree of Ill-Posedness in Stochastic and Deterministic Models*, Preprint 509, Weierstraß-Institut, Berlin, 1999.
[27] M. Reiß, *Minimax rates for nonparametric drift estimation in affine stochastic delay differential equations*, Stat. Inference Stoch. Process., 5 (2002), pp. 131–152.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1502–1526

ANALYSIS OF FINITE ELEMENT APPROXIMATION OF EVOLUTION PROBLEMS IN MIXED FORM*

DANIELE BOFFI† AND LUCIA GASTALDI‡

Abstract. This paper deals with the finite element approximation of evolution problems in mixed form. Following [D. Boffi, F. Brezzi, and L. Gastaldi, *Math. Comp.*, 69 (2000), pp. 121–140], we handle separately two types of problems. A model for the first case is the heat equation in mixed form, while the time dependent Stokes problem fits within the second one. For either case, we give sufficient conditions for a good approximation in the natural functional spaces. The results are not obvious in the first situation. In this case, the well-known conditions for the well posedness and convergence of the corresponding steady problem are not sufficient for the good approximation of the time dependent problem. This issue is demonstrated with a numerical (counter-) example and justified analytically.

Key words. mixed finite element, evolution problem, Stokes problem, Laplace problem in mixed form

AMS subject classifications. 65N30, 65M60

DOI. 10.1137/S0036142903431821

1. Introduction.

Mixed finite elements often are used in engineering applications, and their analysis has been considered in several papers, starting from the 1970s (see [10, 3, 13]), mainly for the approximation of steady source problems. For the reader's convenience, we recall here what we mean by a general standard mixed problem. We consider a pair of Hilbert spaces Φ and Ξ and denote by Φ' and Ξ' their dual spaces. Given two data $f \in \Phi'$ and $g \in \Xi'$, a general (steady) mixed problem reads as follows: Find $\psi \in \Phi$ and $\chi \in \Xi$ such that

$$\begin{aligned} a(\psi,\varphi) + b(\varphi,\chi) &= {}_{\Phi'}\!< f, \chi >_{\Phi} \quad && \forall \varphi \in \Phi, \\ b(\psi,\xi) &= {}_{\Xi'}\!< g, \xi >_{\Xi} && \forall \xi \in \Xi, \end{aligned} \tag{1}$$

where $a : \Phi \times \Phi \to \mathbb{R}$ and $b : \Phi \times \Xi \to \mathbb{R}$ are continuous bilinear forms.

When a finite element approximation to problem (1) is considered, it is well known that the necessary and sufficient condition for the well posedness, stability, and convergence of the scheme (for any given data) is that two inf-sup constants are bounded below away from zero independently of the meshsize parameter.

In the 1980s the use of mixed finite elements has been considered also for the approximation of eigenvalue problems (see [18, 4]), and only fairly recently it has been understood (see [8, 7]) that the inf-sup conditions are not the main assumptions in this context. It has been observed that, in most cases, we can distinguish between two families of mixed problems, depending on the role played by the two equations which define the mixed problem itself, and it has been proved that suitable conditions have to be considered in either case for the good convergence of the computed eigenvalues. Namely, we can consider $\binom{f}{0}$-type problems, when in (1) $g = 0$ and $\binom{0}{g}$-type problems when the opposite situation occurs, i.e., $f = 0$. For instance, the Stokes problem

*Received by the editors July 21, 2003; accepted for publication (in revised form) March 17, 2004; published electronically December 16, 2004. This work was supported in part by IMATI-CNR, Italy.
http://www.siam.org/journals/sinum/42-4/43182.html

†Dipartimento di Matematica "F. Casorati", Università di Pavia, 27100 Pavia, Italy (boffi@dimat.unipv.it).

‡Dipartimento di Matematica, Università di Brescia, 25133 Brescia, Italy (gastaldi@ing.unibs.it).

belongs to the first family and the standard mixed formulation for the Laplace problem to the second one.

In this paper, we want to consider the finite element approximation of evolution problems in mixed form. The mathematical literature on this field is mainly related to the approximation of the heat equation by means of Raviart–Thomas (RT) elements (see, e.g., [16]); mixed finite element schemes have been used extensively for the approximation of evolution problems, in particular in fluid dynamic applications (see [14, 15]).

It is not an unexpected result that the theoretical analysis of such approximations strongly relies on the behavior of the corresponding eigenvalue approximations. For this reason, we consider separately the $\binom{0}{g}$- and $\binom{f}{0}$-type formulations. Actually, it is not completely true that, for instance, in the mixed form of the heat equation f has to vanish, being possibly related to some nonhomogeneous boundary conditions. However, in this paper, we shall consider only truly $\binom{0}{g}$ and $\binom{f}{0}$ problems; in the case of the heat equation, for instance, we can reduce the problem into this form via a suitable extension of the boundary trace.

The outline of the paper is as follows. In the next section we recall some known results about the standard Galerkin space semidiscretization of parabolic problems. In section 3 we introduce the problems we are dealing with and present some examples. We are going to use a different notation than the one introduced in (1). In particular, we want to adapt our notation to the mixed Laplace equation for the $\binom{0}{g}$-type system and to the Stokes problem when dealing with the $\binom{f}{0}$ problem. In section 4, we report on some numerical experiments. In particular, we construct test cases and approximating spaces in such a way that the inf-sup conditions hold true (hence the corresponding steady problems are well approximated) but for which we observe that the evolution problem is not well approximated. These results show the need for an accurate analysis of the evolution case, which is not a straightforward extension of the standard steady state analysis. In the next two sections, where the main results of this paper are stated and proved, we give sufficient conditions for the convergence of the approximation of evolution problems of $\binom{0}{g}$- and $\binom{f}{0}$-type, respectively. The last section contains additional remarks related to the counterexample presented in section 4. It follows, in particular, that, if the data are smooth enough, the solution can be accurately approximated even with the scheme used for our counterexample.

2. Galerkin semidiscretization of parabolic problems. In this section we collect some known results on the semidiscrete approximation to parabolic problems with the aim of introducing some basic estimates and of comparing them with those obtained in the case of mixed formulations. The interested reader is referred to, e.g., [23] for a more detailed analysis of the problem under consideration in this section.

We consider two Hilbert spaces V and H, $V \subseteq H$, V dense in H. We identify H with its dual space H'. Let $a : V \times V \to \mathbb{R}$ be a continuous bilinear form, satisfying the following coercivity condition: There exist $\alpha > 0$ and $\mu \geq 0$ such that $a(v,v) + \mu||v||_H^2 \geq \alpha||v||_V^2$ for all $v \in V$. The variational formulation of the parabolic problem (denoting by $(\cdot,\cdot)$ the scalar product in H) reads as follows: Given $T > 0$, $f :]0,T[\to V'$, and $u_0 \in H$, for almost every $t \in]0,T[$ find $u(t) \in V$ such that

$$\frac{d}{dt}(u(t),v) + a(u(t),v) = {}_{V'}\!< f(t), v >_V \quad \forall v \in V; \qquad u(0) = u_0. \tag{2}$$

The following existence and uniqueness theorem for problem (2) is well known (see, e.g., [17]).

THEOREM 1. *Assume that the bilinear form a is continuous and coercive on $V \times V$. Then, given $f \in L^2(]0,T[;V')$ and $u_0 \in H$, there exists a unique solution $u \in L^2(]0,T[;V) \cap C^0([0,T];H)$ to (2), with $\partial u/\partial t \in L^2(]0,T[;V')$. Moreover, the following energy estimate holds true:*

$$\max_{t\in[0,T]} ||u(t)||_H^2 + \alpha \int_0^T ||u||_V^2 \, dt \le ||u_0||_H^2 + C \int_0^T ||f||_{V'}^2 \, dt.$$

A suitable shift reduces our problem to the case $\mu = 0$; for this reason we consider this case in detail.

If the bilinear form a is symmetric and the embedding of V in H is compact, then, for every $f \in L^2(]0,T[;H)$, the solution to (2) can be represented with the following series:

$$u(t) = \sum_{i=1}^{\infty} \left((u_0, w_i) e^{-\lambda_i t} + \int_0^t (f(s), w_i) e^{-\lambda_i (t-s)} \, ds \right) w_i. \tag{3}$$

Here, $\lambda_i \in \mathbb{R}$ and $w_i \in V$, with $w_i \neq 0$, are eigenvalues and eigenvectors of the bilinear form a; that is, for each i they satisfy $a(w_i, v) = \lambda_i (w_i, v)$ for all $v \in V$.

Example 1 (the heat equation). The standard example is given by the following heat equation: Ω is a polygon in $\mathbb{R}^2$ or a Lipschitz polyhedron in $\mathbb{R}^3$, $H = L^2(\Omega)$, $V = H_0^1(\Omega)$, and $a(u,v) = \int_\Omega \operatorname{grad} u \cdot \operatorname{grad} v \, dx$. Clearly, in this case, the coercivity assumption is valid with $\mu = 0$ for the Poincaré inequality.

Approximating the space V by a finite dimensional subspace V_h provides a space semidiscretization of the variational formulation (2). Given $f \in L^2(]0,T[;H)$ and $u_{0,h} \in V_h$, for each $t \in [0,T]$ find $u_h(t) \in V_h$ such that

$$\frac{d}{dt}(u_h(t), v) + a(u_h(t), v) = {}_{V'}\!< f(t), v >_V \quad \forall v \in V_h; \qquad u_h(0) = u_{0,h}. \tag{4}$$

Subtracting (4) from (2), we obtain the error equation

$$\frac{d}{dt}(u(t) - u_h(t), v) + a(u(t) - u_h(t), v) = 0 \quad \forall v \in V_h.$$

Then we can derive the following estimate for all $v_h \in V_h$:

$$\begin{aligned} \max_{t\in[0,T]} ||u(t) - u_h(t)||_H^2 + \alpha \int_0^T ||u(t) - u_h(t)||_V^2 &\le ||u_0 - u_{0,h}||_H^2 \\ + \int_0^T \left(\left(\frac{\partial}{\partial t}(u(t) - u_h(t)), u(t) - v_h \right) + a(u(t) - u_h(t), u(t) - v_h) \right) dt. \end{aligned} \tag{5}$$

Let us try to get some error estimates from (5) in the case of the finite element approximation of the heat equation (see Example 1). If we take $v_h = u^I(t)$, the interpolant of $u(t)$ in V_h, then we get

$$\begin{aligned} \max_{t\in[0,T]} ||u(t) - u_h(t)||_H^2 + \alpha \int_0^T ||u(t) - u_h(t)||_V^2 &\le ||u_0 - u_{0,h}||_H^2 \\ + \int_0^T \left(\left(\left\| \frac{\partial u}{\partial t}(t) \right\|_H + \left\| \frac{\partial u_h}{\partial t}(t) \right\|_H \right) \|u(t) - u^I(t)\|_H + C\|u(t) - u^I(t)\|_V^2 \right) dt. \end{aligned}$$

If V_h is the space of continuous piecewise linear functions and $u_{0,h}$ is the interpolant of u_0, then we have the optimal convergence (with the obvious modifications in the case of nonconvex domains)

$$||u-u_h||_{L^\infty(L^2)}+\alpha||u-u_h||_{L^2(H^1)}\le Ch\left(||u_0||_{H^1}+||f||_{L^2(L^2)}\right). \tag{6}$$

Unfortunately, the generalization of (6) to higher order approximations is not optimal; when kth order elements are used, (5) gives only $O(h^{(k+1)/2})$. On the other hand, we can take in (5) $v_h = Pu(t)$, the L^2-projection of $u(t)$ into V_h. Now the first term in the integral on the right-hand side of (5) reads

$$\begin{aligned}\left(\frac{\partial}{\partial t}u(t)-\frac{\partial}{\partial t}u_h(t),u(t)-Pu(t)\right)&=\left(\frac{\partial}{\partial t}u(t)-P\frac{\partial}{\partial t}u(t),u(t)-Pu(t)\right)\\&=\frac{1}{2}\frac{d}{dt}||u(t)-Pu(t)||_H^2,\end{aligned}$$

since $\partial u_h(t)/\partial t \in V_h$, P is the L^2-projection onto V_h, and P commutes with the time derivative. Moreover, the estimate of the second term in the integral on the right-hand side of (5) involves the term $||u(t)-Pu(t)||_{H^1}$. If V_h is the space of continuous piecewise polynomials of degree k, and if we assume that the mesh is such that we can use an *inverse estimate*, then we can obtain

$$||u-u_h||_{L^\infty(L^2)}+\alpha||u-u_h||_{L^2(H^1)}\le Ch^k\left(||u_0||_{H^k}+||u||_{L^\infty(H^k)}+||u||_{L^2(H^{k+1})}\right). \tag{7}$$

If, instead of using (5), we introduce the *elliptic projection* operator Π defined as

$$\Pi w\in V_h,\quad a(\Pi w,v)=a(w,v)\quad \forall v\in V_h,$$

for each $w\in V$, then the following error estimate holds true:

$$\begin{aligned}&\max_{t\in[0,T]}||u(t)-u_h(t)||_H^2+\alpha\int_0^T||u(t)-u_h(t)||_V^2\,dt\le||u_0-u_{0,h}||_H^2\\&+\max_{t\in[0,T]}||u(t)-\Pi u(t)||_H^2+C\int_0^T\left(\left\|\frac{\partial}{\partial t}\Big(u(t)-\Pi u(t)\Big)\right\|_{V'}^2+||u(t)-\Pi u(t)||_V^2\right)dt.\end{aligned}$$

In the case of the heat equation, the last equation, together with the usual error estimates for elliptic problems, leads to an estimate similar to (7) but without using the inverse inequality (see, e.g., [23, 20]).

3. Setting of the problem.

3.1. $\binom{0}{g}$-type problems. We consider two Hilbert spaces Σ and V and a third Hilbert space H (which will be identified with its dual space H') such that the following standard inclusions hold with dense and continuous embedding: $V\subseteq H\simeq H'\subseteq V'$. When referring to the inner product of H, we shall omit the reference to the space; in the main application we have in mind, H is simply L^2. A model $\binom{0}{g}$-type evolution problem reads as follows: Given $T>0$, $g:\,]0,T[\,\to V'$, and $u_0\in H$, for almost every $t\in\,]0,T[$ find $\sigma(t)\in\Sigma$ and $u(t)\in V$ such that

$$\begin{aligned}&a(\sigma(t),\tau)+b(\tau,u(t))=0 && \forall\tau\in\Sigma,\\&b(\sigma(t),v)-\frac{d}{dt}(u(t),v)=-{}_{V'}\!<g(t),v>_V && \forall v\in V,\\&u(0)=u_0,\end{aligned} \tag{8}$$

where the first two equations are defined in the sense of distributions in $]0,T[$. We make the following assumptions on the involved bilinear forms $a : \Sigma \times \Sigma \to \mathbb{R}$ and $b : \Sigma \times V \to \mathbb{R}$:

$$\begin{array}{l} a(\cdot,\cdot) \text{ and } b(\cdot,\cdot) \text{ are continuous; that is,} \\ \exists M_a > 0 : \forall \sigma, \tau \in \Sigma \; a(\sigma,\tau) \le M_a ||\sigma||_\Sigma ||\tau||_\Sigma, \\ \exists M_b > 0 : \forall \sigma \in \Sigma, \forall v \in V \; b(\sigma, v) \le M_b ||\sigma||_\Sigma ||v||_V. \end{array}$$

We also assume that $a(\cdot,\cdot)$ is symmetric and positive semidefinite and set

$$|\tau|_a := (a(\tau,\tau))^{1/2}.$$

It is immediate to check that $|\cdot|_a$ is a seminorm on Σ and that for all $\sigma, \tau \in \Sigma$ we have

$$a(\sigma,\tau) \le |\sigma|_a |\tau|_a.$$

We suppose that problem (8) is well posed and that the following a priori estimates hold true:

$$\max_{t\in[0,T]} ||u(t)||_H^2 + \int_0^T ||u(t)||_V^2 \, dt \le ||u_0||_H^2 + C \int_0^T ||g(t)||_{V'}^2 \, dt, \tag{9}$$

$$\int_0^T \mu^2(t) ||\sigma(t)||_\Sigma^2 \, dt \le C \left(||u_0||_H^2 + \int_0^T ||g(t)||_{V'}^2 \, dt \right), \tag{10}$$

where $\mu(t)$ is a suitably chosen weight function which might tend to zero as t goes to zero.

Let Σ_h and V_h be finite dimensional subspaces of Σ and V, respectively. The space semidiscretization of problem (8) reads as follows: For almost every $t \in]0,T[$, find $\sigma_h(t) \in \Sigma_h$ and $u_h(t) \in V_h$ such that

$$\begin{array}{ll} a(\sigma_h(t),\tau) + b(\tau, u_h(t)) = 0 & \forall \tau \in \Sigma_h, \\ b(\sigma_h(t), v) - \dfrac{d}{dt}(u_h(t), v) = -_{V'}\!< g(t), v >_V & \forall v \in V_h, \\ u_h(0) = u_{0,h}, & \end{array} \tag{11}$$

where $u_{0,h} \in V_h$ is a suitable approximation of u_0.

Example 2 (the heat equation in mixed form). Let Ω be an open Lipschitz polygon in $\mathbb{R}^2$ or polyhedron in $\mathbb{R}^3$. Setting $\Sigma = H(\mathrm{div};\Omega)$, $V = H = V' = L^2(\Omega)$, the formulation given in (8) is a weak form of the heat equation ($\boldsymbol{\sigma}(t) = \operatorname{grad} u(t)$) with the choices

$$a(\boldsymbol{\sigma}, \boldsymbol{\tau}) = (\boldsymbol{\sigma}, \boldsymbol{\tau}), \qquad b(\boldsymbol{\tau}, v) = (\operatorname{div} \boldsymbol{\tau}, v).$$

It can be directly checked that the a priori estimates (10) hold with $\mu(t) = t$. The estimates can be improved to get $\mu(t) = 1$ if more regularity is assumed on u_0, namely $u_0 \in H_0^1(\Omega)$ or, analogously, $\boldsymbol{\sigma}_0 = \operatorname{grad} u_0 \in L^2(\Omega)$.

For the mixed spatial semidiscretization of the heat equation we construct two sequences of finite element spaces $\Sigma_h \subset H(\mathrm{div};\Omega)$ and $V_h \subset L^2(\Omega)$ and consider

the following discrete problem: For almost every $t \in]0,T[$, find $\boldsymbol{\sigma}_h(t) \in \Sigma_h$ and $u_h(t) \in V_h$ such that

$$
\begin{aligned}
&(\boldsymbol{\sigma}_h(t), \boldsymbol{\tau}) + (\operatorname{div} \boldsymbol{\tau}, u_h(t)) = 0 && \forall \boldsymbol{\tau} \in \Sigma_h,\\
&(\operatorname{div} \boldsymbol{\sigma}_h(t), v) - \frac{d}{dt}(u_h(t), v) = -(g(t), v) && \forall v \in V_h,\\
&u_h(0) = u_{0,h}.
\end{aligned} \tag{12}
$$

Several possible choices for the spaces Σ_h and V_h have been presented in the literature for the corresponding source problem. For instance, in the case of triangular or tetrahedral meshes, we can choose as Σ_h the spaces of the RT elements introduced in [21, 19] or the Brezzi–Douglas–Marini (BDM) and Brezzi–Douglas–Fortin–Marini (BDFM) elements introduced in [12, 11] (see [13] for a unified presentation; see also [16] for the use of RT elements in the context of parabolic problems in mixed form). In all these cases V_h is equal to $\operatorname{div} \Sigma_h$. On quadrilaterals or hexahedrons the situation is more complicated; a two-dimensional theory has been developed recently in [1], showing that the standard families just listed do not achieve optimal approximation properties in $H(\operatorname{div}; \Omega)$ for general quadrilateral meshes. There, a new Arnold–Boffi–Falk (ABF) family has been proposed by adding internal degrees of freedom to the RT elements in order to recover the optimal accuracy. In this case, the inclusion $\operatorname{div} \Sigma_h \subseteq V_h$ is no longer valid (see [1] for the details about the definition of V_h).

All these spaces satisfy the conditions for the well posedness of the steady problem, namely, there exist two positive constants α and β such that

$$
(\boldsymbol{\tau}, \boldsymbol{\tau}) \geq \alpha ||\boldsymbol{\tau}||^2_{H(\operatorname{div};\Omega)} \quad \forall \boldsymbol{\tau} \in \Sigma_h \text{ with } (\operatorname{div} \boldsymbol{\tau}, v) = 0 \quad \forall v \in V_h, \tag{13}
$$

$$
\sup_{\boldsymbol{\tau} \in \Sigma_h} \frac{(\operatorname{div} \boldsymbol{\tau}, v)}{||\boldsymbol{\tau}||_{H(\operatorname{div};\Omega)}} \geq \beta ||v||_{L^2(\Omega)} \quad \forall v \in V_h. \tag{14}
$$

In section 4 we shall test the lowest order RT element for the approximation of the heat equation, together with another less standard element which also satisfies (13) and (14).

3.2. $\binom{f}{0}$-type problems. We consider three Hilbert spaces V, H, and Q such that $V \subseteq H \simeq H' \subseteq V'$ with dense and continuous inclusions. As in the previous section, we shall refer to the scalar product of H with $(\cdot, \cdot)$. A model $\binom{f}{0}$-type evolution problem reads as follows: Given $T > 0$, $f :]0,T[\to V'$, and $u_0 \in H$, for almost every $t \in]0,T[$ find $u(t) \in V$ and $p(t) \in Q$ such that

$$
\begin{aligned}
&\frac{d}{dt}(u(t), v) + a(u(t), v) + b(v, p(t)) = {}_{V'}\!< f, v >_V && \forall v \in V,\\
&b(u(t), q) = 0 && \forall q \in Q,\\
&u(0) = u_0,
\end{aligned} \tag{15}
$$

where the first two equations are defined in the sense of distributions in $]0,T[$. We recall the definitions of the bilinear forms $a : V \times V \to \mathbb{R}$ and $b : V \times Q \to \mathbb{R}$ and make the following standard hypotheses:

$a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are continuous; that is,
$\exists M_a > 0 : \forall u, v \in V \; a(u, v) \leq M_a ||u||_V ||v||_V$,
$\exists M_b > 0 : \forall v \in V, \forall q \in Q \; b(v, q) \leq M_b ||v||_V ||q||_Q$.

Moreover, we assume that the form a is coercive on the kernel of B,

$$K = \{v \in V : b(v,q) = 0 \ \forall q \in Q\},$$

i.e., there exists $\alpha > 0$ such that

$$a(v,v) \geq \alpha ||v||_V^2 \quad \forall v \in K. \tag{16}$$

A weaker ellipticity could be considered when dealing with parabolic problems. However, it can be reduced to (16) with a change of variables. We suppose that problem (15) is well posed and that the following a priori estimates hold true:

$$\max_{t\in[0,T]} ||u(t)||_H^2 + \alpha \int_0^T ||u(t)||_V^2 \, dt \leq ||u_0||_H^2 + C \int_0^T ||f(t)||_{V'}^2 \, dt, \tag{17}$$

$$\int_0^T ||p(t)||_Q^2 \, dt \leq C \left(||u_0||_H^2 + \int_0^T ||f(t)||_{V'}^2 \, dt \right). \tag{18}$$

Let V_h and Q_h be finite dimensional subspaces of V and Q; then the Galerkin space semidiscretization of problem (15) reads as follows: For almost every $t \in]0,T[$, find $u_h(t) \in V_h$ and $p_h(t) \in Q_h$ such that

$$\begin{aligned} &\frac{d}{dt}(u_h(t), v) + a(u_h(t), v) + b(v, p_h(t)) = {}_{V'}\!< f, v >_V \quad && \forall v \in V_h, \\ &b(u_h(t), q) = 0 && \forall q \in Q_h, \\ &u_h(0) = u_{0,h}, \end{aligned} \tag{19}$$

where $u_{0,h} \in V_h$ is an approximation of u_0.

Example 3 (the Stokes equations). Let Ω be an open Lipschitz polyhedron in $\mathbb{R}^n$ (with $n = 2$ or 3); then the Stokes equations fit in a natural way within our setting with the following definitions:

$$\begin{gathered} V = (H_0^1(\Omega))^n, \quad Q = L^2(\Omega)/\mathbb{R}, \quad H = (L^2(\Omega))^n, \\ a(\mathbf{u}, \mathbf{v}) = (\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v})), \qquad b(\mathbf{v}, q) = (\operatorname{div} \mathbf{v}, q), \end{gathered}$$

where $\boldsymbol{\varepsilon}$ is, as usual, the linearized strain tensor. It is known (see, for instance, [22]) that estimates (17) and (18) hold true in this case. Moreover, whenever Ω is convex (with the usual modifications for the nonconvex case), if $f \in L^2(]0,T[\,; H)$ and $u_0 \in K$, then $u \in L^2(]0,T[\,; (H^2(\Omega))^n)$, $\partial u/\partial t \in L^2(]0,T[\,; H)$, and $p \in L^2(]0,T[\,; H^1(\Omega))$.

For the discretization of the steady problem, we refer to [13]. We shall see at the end of section 6 that a good approximation of the steady Stokes equations provides a convergent approximation to the time dependent problem also.

4. Numerical investigations. In this section we report on some numerical tests for various mixed discretization to the heat equation presented in Example 2. The discrete problem we are dealing with is the one presented in (12). We consider two possible choices for the discrete spaces Σ_h and V_h. Given $\Omega =]0,\pi[^2$ and a triangular mesh, the first method consists in choosing Σ_h as the lowest order RT element (see [21]) and as V_h the space of piecewise constant functions. We shall refer to this choice as the RT method. In the second method, which has been analyzed in [8], the space Σ_h consists of continuous piecewise linear (in each component) vector fields and V_h is simply defined as $V_h = \operatorname{div} \Sigma_h$ and turns out to be a subset of the space of piecewise constant functions. We shall refer to this example as the P1 method.

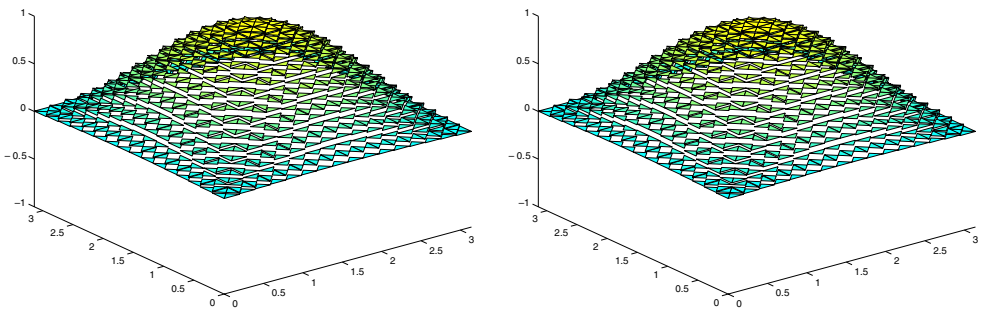


FIG. 1. $u(10)$ *on* 16*-by-*16 *criss-cross mesh computed with the RT (left) and* $P1$ *(right) methods with* $g = 2 \sin x \sin y$.

The RT method is well known to be stable (see [21]), when applied to the steady problem (see (13) and (14)). The P1 method has been introduced in [8] in order to construct a counterexample for the approximation of eigenvalue problems. It has been shown to be stable on a special mesh sequence of the square which is built of uniform subsquares, each of them subdivided into four triangles (criss-cross mesh). In this section, we shall denote by N the number of subdivisions of each side of Ω, so that a criss-cross mesh contains $4N^2$ triangles. On general triangular meshes, the P1 element does not satisfy the inf-sup condition (14) in the sense that the inf-sup constant β tends to zero as h goes to zero.

If not otherwise indicated, we take $T = 10$ and advance in time using an implicit Euler scheme with step $dt = 0.1$.

In all our tests, g does not depend on t and $u_0 \equiv 0$, so that the solution $(\boldsymbol{\sigma}(t), u(t))$ asymptotically tends to the solution $(\boldsymbol{\sigma}_\infty, u_\infty) \in H(\mathrm{div}; \Omega) \times L^2(\Omega)$ of the steady problem

$$\begin{aligned} &(\boldsymbol{\sigma}_\infty, \boldsymbol{\tau}) + (\mathrm{div}\, \tau, u_\infty) = 0 \quad && \forall \boldsymbol{\tau} \in H(\mathrm{div}; \Omega), \\ &(\mathrm{div}\, \boldsymbol{\sigma}_\infty, v) = -(g, v) && \forall v \in L^2(\Omega). \end{aligned}$$

In our first test, we choose $g(x, y) = 2 \sin x \sin y$, so that $u_\infty(x, y) = \sin x \sin y$. In Figure 1 the component $u(t)$ of the solution at time $T = 10$ is plotted for both methods on a criss-cross mesh with $N = 16$. The results look quite similar; the $L^2(\Omega)$ norms of the solution are 1.5725 and 1.5674 (the reference value is $\pi/2 = 1.57079\ldots$) for the RT and P1 methods, respectively, and the corresponding solution values at the center are 0.9957 and 0.9925 (the reference value is 1.0). In the second test, considering again criss-cross meshes, we take the function $g = c(h)$ depending on the meshsize h as a checkerboard function with values ± 1 on the underlying mesh of subsquares. For example, the case $N = 4$ is plotted in Figure 2. Let $u_\infty(h)$ be the asymptotic solution of our problem with datum $g = c(h)$. As h goes to zero, the function $c(h)$ tends weakly to zero in $L^2(\Omega)$, so that, for the compactness of the inverse Laplace operator, we have that $u_\infty(h)$ tends to zero strongly in $L^2(\Omega)$. We expect a good numerical method to provide a solution $u_h(t) \in V_h$ tending to zero as h goes to zero. We explicitly observe that our expectation is related to a sort of uniform convergence, since the solution is computed, for each h, with respect to a different right-hand side. On the other hand, this kind of convergence is what we usually obtain from the error estimates (see (27) and (36)).

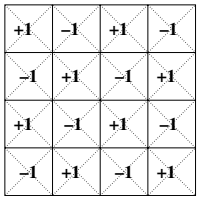


FIG. 2. *The checkerboard $c(h)$ for $N = 4$.*

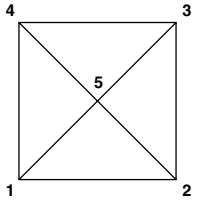


FIG. 3. *The reference criss-cross macro-element.*

TABLE 1
L^2 norms of the solution with $g = c(h)$.

	dt = .1		dt = .01	
N	RT	P1	RT	P1
4	0.080756	0.451091	0.080756	0.451091
8	0.020186	0.430821	0.020186	0.430821
16	0.005047	0.427416	0.005047	0.427416
32	0.001262	0.426687	0.001262	0.426687

TABLE 2
L^2 norms of the solution when g is the 8-by-8 checkerboard.

N	RT	P1
4	0.053830	0.103860
8	0.020186	0.430821
16	0.022569	0.020642
32	0.020689	0.019728

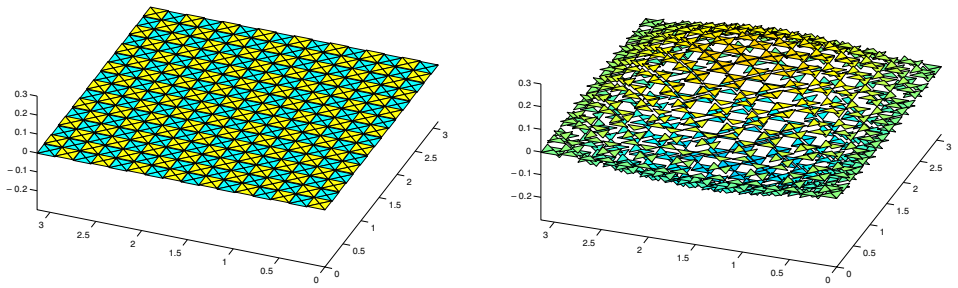


FIG. 4. *$u(10)$ on 16-by-16 criss-cross mesh computed with the RT (left) and P1 (right) methods when g is a checkerboard.*

In Table 1 we report the L^2 norm of $u_h(10)$ for various values of N computed with RT and P1 methods. It is evident that the solution computed with the P1 method does not go to zero. This fact clearly indicates a different behavior for the two methods. In particular, it is evident that the stability of the steady problem (see (13) and (14)) is not sufficient for the good approximation of the evolution problem in mixed form. Here, by *good approximation* we mean a convergence like it comes from estimates (27) or (36). In Table 1 we also show that the bad behavior of the P1 approximation is not related to a poor time discretization. Indeed, a refinement in t does not produce any improvement. At the end of section 5, we shall come back to this example and analyze more deeply the differences between the two methods. In Figure 4 we show the solution $u_h(10)$ computed with the two methods. The resonance induced by the checkerboard in the case of the P1 scheme can be observed. This will

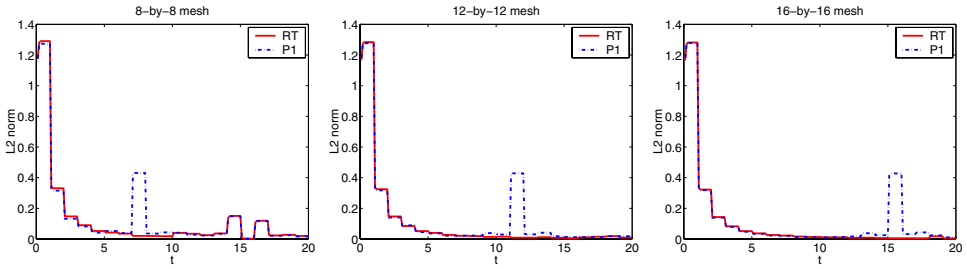


FIG. 5. L^2 *norm of* $u_h(t)$ *for* $t \in [0, 20]$ *with* $g(t)$ *equal to* $N(t)$*-by-*$N(t)$ *checkerboard function and Dirichlet (natural) boundary conditions.*

be related to the presence of spurious eigenvalues (see [8]) at the end of section 5. On the other hand, if we keep g fixed and let h go to zero, then we observe that the solution computed with the P1 method is similar to the one obtained with the RT method (see Table 2), the only exception being when the mesh is the one underlying the structure of g.

We introduce now another example, in which the P1 method is not converging as h goes to zero in the $L^\infty(L^2)$ norm. The load term g is now a function of t; namely, at time t, $g(t)$ is the $N(t)$-by-$N(t)$ checkerboard function with $N(t) = \lceil t \rceil$. The norm of the solution $u(t)$, computed with RT and P1 methods using three different meshes ($N = 8, 12, 16$), is plotted in Figure 5. We notice that, when the P1 method is used, there is a pick with constant height appearing at an increasing value of t as h decreases. This happens when the structure of g *resonates* with the mesh and cannot be avoided even if the mesh is refined.

5. Convergence analysis for the $\binom{0}{g}$-type evolution problems. In this section we shall develop two different error estimates. The first one, stated in Theorem 3, is new and is based on the compatibility assumption (23). This assumption is satisfied for most standard finite elements used for the approximation of the heat equation. The second estimate, given in Theorem 5, is basically equivalent to the one given in Theorem 2.1 of [16]. Our proof is new, and we explicitly observe that, in this case, the regularity hypothesis (34) has to be made.

We start this section by recalling the continuous problem (8) and its space semidiscretization (11). The hypotheses on the involved bilinear forms have been made explicit in section 3.1. Given $T > 0$, $g :]0, T[\to V'$, and $u_0 \in H$, for almost every $t \in]0, T[$ we look for $\sigma(t) \in \Sigma$ and $u(t) \in V$ such that

$$\begin{aligned} &a(\sigma(t), \tau) + b(\tau, u(t)) = 0 && \forall \tau \in \Sigma, \\ &b(\sigma(t), v) - \frac{d}{dt}(u(t), v) = -_{V'}\!< g(t), v >_V && \forall v \in V, \\ &u(0) = u_0. \end{aligned}$$

The semidiscrete counterpart reads as follows: For almost every $t \in]0, T[$, find $\sigma_h(t) \in \Sigma_h$ and $u_h(t) \in V_h$ such that

$$\begin{aligned} &a(\sigma_h(t), \tau) + b(\tau, u_h(t)) = 0 && \forall \tau \in \Sigma_h, \\ &b(\sigma_h(t), v) - \frac{d}{dt}(u_h(t), v) = -_{V'}\!< g(t), v >_V && \forall v \in V_h, \\ &u_h(0) = u_{0,h}. \end{aligned}$$

The error equations are

$$a(\sigma(t)-\sigma_h(t),\tau)+b(\tau,u(t)-u_h(t))=0 \quad \forall \tau\in\Sigma_h,$$
$$b(\sigma(t)-\sigma_h(t),v)-\frac{d}{dt}(u(t)-u_h(t),v)=0 \quad \forall v\in V_h.$$

Given two linear operators $\Pi:\Sigma\to\Sigma_h$ and $P:H\to V_h$, and taking $\tau=\Pi\sigma(t)-\sigma_h(t)$, $v=Pu(t)-u_h(t)$ in the error equations we get, after summation,

(20)
$$\begin{aligned}\left(\frac{\partial}{\partial t}\Big(Pu(t)-u_h(t)\Big),Pu(t)-u_h(t)\right)&+a(\Pi\sigma(t)-\sigma_h(t),\Pi\sigma(t)-\sigma_h(t))\\ &=b(\sigma(t)-\Pi\sigma(t),Pu(t)-u_h(t))-\left(\frac{\partial}{\partial t}\Big(u(t)-Pu(t)\Big),Pu(t)-u_h(t)\right)\\ &\quad-a(\sigma(t)-\Pi\sigma(t),\Pi\sigma(t)-\sigma_h(t))-b(\Pi\sigma(t)-\sigma_h(t),u(t)-Pu(t)).\end{aligned}$$

Let us now take Π as a Fortin operator and P as the H projection onto V_h, namely,

(21)
$$\begin{aligned}&\Pi:\Sigma\to\Sigma_h,\\ &b(\tau-\Pi\tau,v)=0 \quad \forall v\in V_h,\end{aligned}$$

and

$$\begin{aligned}&P:H\to V_h,\\ &(w-Pw,v)=0 \quad \forall v\in V_h.\end{aligned}$$

Then, (20) reduces to

(22)
$$\begin{aligned}\left(\frac{\partial}{\partial t}\Big(Pu(t)-u_h(t)\Big),Pu(t)-u_h(t)\right)&+a(\Pi\sigma(t)-\sigma_h(t),\Pi\sigma(t)-\sigma_h(t))\\ &=a(\sigma(t)-\Pi\sigma(t),\Pi\sigma(t)-\sigma_h(t))-b(\Pi\sigma(t)-\sigma_h(t),u(t)-Pu(t)),\end{aligned}$$

where we used the fact that Π is a Fortin operator (see (21)) and that the projection P commutes with the time derivative.

In several interesting applications, the last term in the right-hand side of (22) vanishes. Let $B:\Sigma\to V'$ denote the canonical operator defined by ${}_{V'}\!<B\sigma,v>_V=b(\sigma,v)$ for all $\sigma\in\Sigma$ and $v\in V$; if

(23)
$$B(\Sigma_h)\subseteq V_h,$$

then we can show that, for any $v\in V$,

$$b(\tau,v-Pv)=0 \quad \forall \tau\in\Sigma_h.$$

Indeed, in this case, $b(\tau,v-Pv)={}_{V'}\!<B\tau,v-Pv>_V=(B\tau,v-Pv)=0$, due to the identification $H\simeq H'$.

Hence, we can easily obtain the following result.

LEMMA 2. *Let (σ,u) and (σ_h,u_h) be the solutions of problems* (8) *and* (11)*, respectively. If the inclusion* (23) *holds, then the following error estimate is true:*

$$\begin{aligned}&\max_{t\in[0,T]}||u(t)-u_h(t)||_H^2+\int_0^T|\sigma(t)-\sigma_h(t)|_a^2\,dt\\ &\qquad\le||Pu_0-u_{0,h}||_H^2+\max_{t\in[0,T]}||u(t)-Pu(t)||_H^2+2\int_0^T|\sigma(t)-\Pi\sigma(t)|_a^2\,dt.\end{aligned}$$

In order to obtain a convergence result, we need some assumptions on the operators P and Π. Let us denote by V^+ a subspace of V such that the second component of the solution u belongs to $L^2(]0,T[;V^+)$ whenever $g \in L^2(]0,T[;H)$ and denote by Σ^+ a space such that the first component of the solution σ belongs to $L^2(]0,T[;\Sigma^+)$ whenever $g \in L^2(]0,T[;H)$. Moreover, let us assume that the following estimate holds:

$$\|\sigma\|_{L^2(\Sigma^+)} + \|u\|_{L^2(V^+)} \le C\|g\|_{L^2(H)}. \tag{24}$$

Then we assume that there exists $\rho_1(h)$, tending to zero as h goes to zero, such that for every $u \in V^+$ it holds that

$$||u - Pu||_H \le \rho_1(h)||u||_{V^+}. \tag{25}$$

In general, estimate (25) can be achieved if P is a suitable approximation operator. As far as the operator Π is concerned, in the spirit of [7], we make the following assumption: There exists $\rho_2(h)$, tending to zero as h goes to zero, such that

$$|\sigma - \Pi\sigma|_a \le \rho_2(h)||\sigma||_{\Sigma^+}. \tag{26}$$

The following theorem summarizes the results obtained so far.

THEOREM 3. *Let (σ,u) and (σ_h,u_h) be the solutions of problems* (8) *and* (11)*, respectively, and suppose that the inclusion* (23) *holds true. If there exist P and Π satisfying* (25) *and* (26)*, then we have the following estimate:*

$$\begin{aligned} ||u(t) - u_h(t)||_{L^\infty(H)} + \left(\int_0^T |\sigma(t) - \sigma_h(t)|_a^2\, dt\right)^{1/2} \\ \le C||Pu_0 - u_{0,h}||_H + C(\rho_1(h) + \rho_2(h))||g||_{L^2(H)}. \end{aligned} \tag{27}$$

A different estimate, which does not require the inclusion (23), but for which some more regularity assumption on the solution is required, can be obtained by using the approach of the *elliptic projection* (see section 2). Let us consider the steady problem associated to (8): Given $g \in V'$, find $(\sigma,u) \in \Sigma \times V$ such that

$$\begin{aligned} a(\sigma,\tau) + b(\tau,u) &= 0 \qquad \forall \tau \in \Sigma, \\ b(\sigma,v) &= -{}_{V'}\!< g, v >_V \quad \forall v \in V, \end{aligned} \tag{28}$$

and the corresponding discrete problem: Find $(\sigma_h,u_h) \in \Sigma_h \times V$ such that

$$\begin{aligned} a(\sigma_h,\tau) + b(\tau,u_h) &= 0 \qquad \forall \tau \in \Sigma_h, \\ b(\sigma_h,v) &= -{}_{V'}\!< g, v >_V \quad \forall v \in V_h. \end{aligned} \tag{29}$$

We make the assumption that (28) and (29) are well posed. Then, given $(\sigma,u) \in \Sigma \times V$ solution of (28), we can define $\Pi\sigma \in \Sigma_h$ and $P\sigma \in V_h$ as the solution of (29) with right-hand side ${}_{V'}\!< g, v >_V = -b(\sigma,v)$, namely,

$$\begin{aligned} a(\Pi\sigma,\tau) + b(\tau,P\sigma) &= 0 \quad \forall \tau \in \Sigma_h, \\ b(\Pi\sigma,v) &= b(\sigma,v) \qquad \forall v \in V_h. \end{aligned} \tag{30}$$

We observe that, thanks to the well posedness of problem (28), the continuous solution of (30) is (σ,u). Since, basically, Π and P depend explicitly only on σ, we omit u from the notation for simplicity. It should be remembered, however, that $P\sigma$ is a

discrete counterpart to u. Moreover, the operator $\Pi : \Sigma \to \Sigma_h$ defined in (30) is a Fortin operator in the sense of (21). Inserting now the *elliptic projections* Π and P defined in (30) into (20), we obtain

$$\left(\frac{\partial}{\partial t}\Big(P\sigma(t)-u_h(t)\Big), P\sigma(t)-u_h(t)\right) + a(\Pi\sigma(t)-\sigma_h(t), \Pi\sigma(t)-\sigma_h(t)) = -\left(\frac{\partial}{\partial t}\Big(u(t)-P\sigma(t)\Big), P\sigma(t)-u_h(t)\right),$$

where we made use of the fact that Π is a Fortin operator and we took advantage of the following error equation:

$$a(\sigma(t)-\Pi\sigma(t),\tau) + b(\tau, u(t)-P\sigma(t)) = 0 \quad \forall \tau \in \Sigma_h.$$

Hence, using standard arguments and Gronwall's lemma, we obtain the following result.

LEMMA 4. *Let (σ, u) and (σ_h, u_h) be the solutions of problems* (8) *and* (11)*, respectively. Let $\Pi : \Sigma \to \Sigma_h$ and $P : V \to V_h$ be defined as in* (30)*. Suppose, moreover, that u_0 is in V and is such that there exists $\sigma_0 \in \Sigma$ with (σ_0, u_0) solution to* (28) *for a suitable $g \in V'$. Then the following error estimate is true:*

$$\max_{t\in[0,T]} ||u(t)-u_h(t)||_H^2 + \int_0^T |\sigma(t)-\sigma_h(t)|_a^2\, dt \le ||P\sigma_0 - u_{0,h}||_H^2 + \max_{t\in[0,T]} ||u(t)-P\sigma(t)||_H^2 + 2\int_0^T |\sigma(t)-\Pi\sigma(t)|_a^2\, dt + C\int_0^T \left\|\frac{\partial}{\partial t}(u(t)-P\sigma(t))\right\|_H^2 dt.$$

In order to get the convergence estimate from the previous lemma, we make the following assumptions. First, we assume that there exists $\omega_1(h)$, going to zero as h goes to zero, such that

$$||u-P\sigma||_H \le \omega_1(h)||u||_{V^+}. \tag{31}$$

Then, we need an estimate for $||\sigma(t)-\Pi\sigma(t)||_a$. In analogy to (26), we suppose that there exists $\omega_2(h)$, going to zero as h goes to zero, such that

$$|\sigma-\Pi\sigma|_a \le \omega_2(h)||\sigma||_{\Sigma^+}. \tag{32}$$

Finally, we suppose the existence of $\omega_3(h)$, going to zero as h goes to zero, such that

$$\left\|\frac{\partial}{\partial t}(u-P\sigma)\right\|_{L^2(H)} \le \omega_3(h)\left\|\frac{\partial u}{\partial t}\right\|_{L^2(V^+)}. \tag{33}$$

Remark 1. We explicitly notice that estimate (33) is not an immediate consequence of (31) unless the operator P commutes with the time derivative. Actually, this property is true provided some regularity assumption is made. Indeed, we need to assume

$$\frac{\partial \sigma}{\partial t} \in L^2(]0,T[;\Sigma) \tag{34}$$

in order to define $P(\partial\sigma(t)/\partial t) \in V_h$, according to (30):

$$\begin{aligned} a\left(\Pi\frac{\partial\sigma(t)}{\partial t},\tau\right) + b\left(\tau, P\frac{\partial\sigma(t)}{\partial t}\right) &= 0 \quad \forall \tau \in \Sigma_h, \\ b\left(\Pi\frac{\partial\sigma(t)}{\partial t}, v\right) &= b\left(\frac{\partial\sigma(t)}{\partial t}, v\right) \quad \forall v \in V_h. \end{aligned}$$

Differentiating (30) with respect to t, we get

$$\begin{aligned} a\left(\frac{\partial\Pi\sigma(t)}{\partial t},\tau\right) + b\left(\tau, \frac{\partial P\sigma(t)}{\partial t}\right) &= 0 \quad \forall \tau \in \Sigma_h, \\ b\left(\frac{\partial\Pi\sigma(t)}{\partial t}, v\right) &= b\left(\frac{\partial\sigma(t)}{\partial t}, v\right) \quad \forall v \in V_h. \end{aligned}$$

Comparing the last two equations, and using the uniqueness hypothesis on problem (29), we finally obtain

$$P\frac{\partial\sigma(t)}{\partial t} = \frac{\partial P\sigma(t)}{\partial t} \quad \text{for a.e. } t \in]0,T[\,. \tag{35}$$

Relation (35), which relies on the regularity assumption (34), can be used to get the estimate

$$\left\| \frac{\partial}{\partial t}(u - P\sigma) \right\|_{L^2(H)} \leq \omega_1(h) \left\| \frac{\partial u}{\partial t} \right\|_{L^2(V^+)}.$$

Assumptions (31), (32), and (33) allow us to state the following theorem.

THEOREM 5. *Let (σ, u) and (σ_h, u_h) be the solutions of problems* (8) *and* (11)*, respectively. Let Π and P be the two components of the elliptic projection* (30)*, and let σ_0 be as in Lemma* 4*. If* (31)*,* (32)*, and* (33) *are satisfied, then we have the following estimate:*

$$\begin{aligned} &||u(t) - u_h(t)||_{L^\infty(H)} + \left(\int_0^T |\sigma(t) - \sigma_h(t)|_a^2 \, dt \right)^{1/2} \\ &\quad \leq C||P\sigma_0 - u_{0,h}||_H + C(\omega_1(h) + \omega_2(h))||g||_{L^2(H)} + C\omega_3(h) \left\| \frac{\partial u}{\partial t} \right\|_{L^2(V^+)}. \end{aligned} \tag{36}$$

Example 4 (convergence analysis for the mixed approximation to the heat equation). We review here the numerical examples presented in section 4. We start with the analysis of the RT, BDM, and BDFM methods for mesh of triangles, tetrahedrons, rectangles, and parallelepipeds. In these cases, Theorem 3 gives the optimal kth order rate of convergence (see [13, 8]).

On general quadrilateral meshes, we can use the ABF family introduced in [1]. In this case we cannot use Theorem 3, since the inclusion (23) is not satisfied. However, we can invoke Theorem 5 and conclude with the optimal estimate, provided some extra regularity on the time derivative of u is assumed.

On the other hand, the P1 scheme does not fit within our results. Indeed, it has been proved in [8] that (32) does not hold. We shall make it clear how things might go wrong (as shown numerically in Figure 4) with the following considerations. In [8] it has been shown that the P1 element, when applied to the mixed eigenvalue problem associated with the Laplace operator, presents spurious eigenvalues. Let Ω be the square of side length π, and define f^h as the eigenfunction associated to the first spurious eigenvalue $\bar{\lambda}_h$. We normalize f^h so that $||f^h||_{L^2} = 1$. This function, in particular, has a clear checkerboard pattern, and numerical evidence shows that $\bar{\lambda}_h$ tends to a number close to six (see [8] and section 7 of this paper). Take $u_0 = 0$, then the continuous solution u^h to the heat equation is (see (3))

$$u^h(t) = \sum_{i=1}^{\infty} w_i \int_0^t (f^h, w_i) e^{\lambda_i (s-t)}\, ds.$$

Let us consider the discrete eigenmodes $\lambda_{i,h} \in \mathbb{R}$ and $w_{i,h} \in V_h$, $\sigma_{i,h} \in \Sigma_h$, $i = 1, \dots, \mathcal{N}(h)$, (where $\mathcal{N}(h)$ is the dimension of V_h) satisfying

$$\begin{aligned} &(\boldsymbol{\sigma}_{i,h}, \boldsymbol{\tau}) + (\operatorname{div} \boldsymbol{\tau}, w_{i,h}) = 0 \quad && \forall \boldsymbol{\tau} \in \Sigma_h, \\ &(\operatorname{div} \boldsymbol{\sigma}_{i,h}, v) = -\lambda_{i,h} (w_{i,h}, v) \quad && \forall v \in V_h. \end{aligned}$$

With this notation, a solution expansion also holds at the discrete level, namely,

$$u_h^h(t) = \sum_{i=1}^{\mathcal{N}(h)} w_{i,h} \int_0^t (f^h, w_{i,h}) e^{\lambda_{i,h}(s-t)}\, ds = \sum_{i=1}^{\mathcal{N}(h)} w_{i,h} (f^h, w_{i,h}) \frac{1 - e^{-\lambda_{i,h} t}}{\lambda_{i,h}}. \tag{37}$$

Since f^h tends to zero weakly in $L^2(\Omega)$, the continuous asymptotic solution u_∞^h tends to zero strongly in $L^2(\Omega)$. On the other hand, from (37) we get

$$u_h^h(t) = f^h \frac{1 - e^{-\bar{\lambda}_h t}}{\bar{\lambda}_h} \quad \text{and} \quad ||u_h^h||_{L^\infty(L^2)} = \frac{1 - e^{-\bar{\lambda}_h T}}{\bar{\lambda}_h}.$$

This last relation implies that an estimate like (27) cannot hold for the P1 method if we can show that $||u^h||_{L^\infty(L^2)}$ tends to zero as h goes to zero. Indeed, we can split u^h as the sum of u_1 and u_2 defined as follows:

$$\begin{aligned} &\frac{\partial u_1}{\partial t} - \Delta u_1 = f^h \quad \text{in } \Omega \times]0, T[\,, && \qquad \frac{\partial u_2}{\partial t} - \Delta u_2 = 0 \quad \text{in } \Omega \times]0, T[\,, \\ &u_1(0) = u_\infty^h \quad \text{in } \Omega, && \qquad u_2(0) = -u_\infty^h \quad \text{in } \Omega. \end{aligned}$$

It is clear that $u_1(t) = u_\infty^h$ for all t, so that

$$||u^h||_{L^\infty(L^2)} \le ||u_1||_{L^\infty(L^2)} + ||u_2||_{L^\infty(L^2)} \le 2||u_\infty^h||_{L^2}.$$

6. Convergence analysis for the $\binom{f}{0}$-type evolution problems. Also in this section we present two different error estimates. In the first one, we shall make the hypothesis that the bilinear form a is coercive on the whole space V (see Theorem 9). This estimate applies, for instance, to the Stokes problem introduced in Example 3 and in this case provides the optimal rate of convergence only when the lowest order elements are used. In the second estimate, presented in Theorem 11, we only assume the ellipticity in the kernel (16), but, in order to get the result, we require an additional approximation property. If applied to the Stokes problem, this estimate turns out to be optimal also for higher order schemes.

We recall the continuous problem (15) and its space semidiscretization (19). The hypotheses on the forms, in particular the ellipticity in the kernel (16), have been made in section 3.2. Given $T > 0$, $f :]0,T[\to V'$, and $u_0 \in H$, for almost every $t \in]0,T[$ find $u(t) \in V$ and $p(t) \in Q$ such that

$$\begin{aligned} &\frac{d}{dt}(u(t),v) + a(u(t),v) + b(v,p(t)) = {}_{V'}\!< f, v >_V \quad && \forall v \in V, \\ &b(u(t),q) = 0 && \forall q \in Q, \\ &u(0) = u_0. \end{aligned}$$

The discrete counterpart reads as follows: For almost every t, find $u_h(t) \in V_h$ and $p_h(t) \in Q_h$ such that

$$\begin{aligned} &\frac{d}{dt}(u_h(t),v) + a(u_h(t),v) + b(v,p_h(t)) = {}_{V'}\!< f, v >_V \quad && \forall v \in V_h, \\ &b(u_h(t),q) = 0 && \forall q \in Q_h, \\ &u_h(0) = u_{0,h}, \end{aligned}$$

where $u_{0,h} \in V_h$ is an approximation of u_0. The error equations are as follows:

$$\begin{aligned} \frac{d}{dt}(u(t) - u_h(t), v) + a(u(t) - u_h(t), v) + b(v, p(t) - p_h(t)) = 0 \quad \forall v \in V_h, \\ b(u(t) - u_h(t), q) = 0 \quad \forall q \in Q_h. \end{aligned}$$

In the next lemma, we shall use the kernel of the discrete operator associated to b, namely, $K_h = \{v_h \in V_h : b(v_h, q_h) = 0 \ \forall q_h \in Q_h\}$.

LEMMA 6. *Let us suppose that the form a is elliptic in V, that is,* (16) *holds for any $v \in V$. Let (u,p) and (u_h,p_h) be the solutions of problems* (15) *and* (19)*, respectively. If the time derivatives of u and u_h are bounded in $L^2(]0,T[;H)$, then the following estimate holds true:*

$$\begin{aligned} \max_{t\in[0,T]} ||u(t) - u_h(t)||_H^2 + \alpha \int_0^T ||u(t) - u_h(t)||_V^2 &\le ||u_0 - u_{0,h}||_H^2 \\ &+ C\left(\int_0^T ||u(t) - \Pi u(t)||_H^2 \, dt\right)^{1/2} + C\int_0^T ||u(t) - \Pi u(t)||_V^2 \, dt \\ &+ \int_0^T b(\Pi u(t) - u_h(t), p(t)) \, dt, \end{aligned} \tag{38}$$

where Π is an operator from V^+ to the discrete kernel K_h.

Proof. From the ellipticity and the error equations we get

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}||u(t) - u_h(t)||_H^2 + \alpha||u(t) - u_h(t)||_V^2 \\ &\quad = \left(\frac{\partial}{\partial t}\Big(u(t) - u_h(t)\Big), u(t) - u_h(t)\right) + a(u(t) - u_h(t), u(t) - u_h(t)) \\ &\quad = \left(\frac{\partial}{\partial t}\Big(u(t) - u_h(t)\Big), u(t) - \Pi u(t)\right) + a(u(t) - u_h(t), u(t) - \Pi u(t)) \\ &\qquad - b(\Pi u(t) - u_h(t), p(t) - p_h(t)). \end{aligned} \tag{39}$$

From our assumption on the time derivative of u and u_h and the fact that $\Pi u(t) \in K_h$, we easily get the result integrating from 0 to T. ∎

Remark 2. From the previous proof it follows that the first constant C appearing on the right-hand side of (38) is related to the bound of the time derivatives of u and u_h in $L^2(]0,T[;H)$. This is a regularity assumption which is in general not too strong if $f \in L^2(]0,T[;H)$. On the other hand, if the time derivatives of u and u_h are only in $L^2(]0,T[;V')$, then a result similar to (38) can be obtained but with the V norm instead of the H one in the second term in the right-hand side.

In order to obtain a rate of convergence from the previous lemma, we consider $V^+ \subseteq V$ and $Q^+ \subseteq Q$ such that the solution to (15) satisfies $u \in L^2(]0,T[;V^+)$ and $p \in L^2(]0,T[;Q^+)$ if $f \in L^2(]0,T[;H)$ with the a priori estimate $\|u\|_{L^2(V^+)} + \|p\|_{L^2(Q^+)} \leq C\|f\|_{L^2(H)}$. Then we introduce the following definitions (see [7]).

DEFINITION 7. *We say that the* weak approximability *property of the space Q^+ holds, if there exists $\rho_1(h)$, tending to zero as h goes to zero, such that*

$$\sup_{v_h \in K_h} \frac{b(v_h,p)}{||v_h||_V} \leq \rho_1(h)||p||_{Q^+} \quad \forall p \in Q^+.$$

Moreover, we also need an approximability property of the kernel; hence, we define the following *strong approximability* of V^+.

DEFINITION 8. *The* strong approximability *of V^+ with respect to Π is satisfied if there exists $\rho_2(h)$ tending to zero as h goes to zero, such that for all $u \in V^+$ it holds that*

$$\|u - \Pi u_h\|_V \leq \rho_2(h)\|u\|_{V^+}.$$

If the above definitions are fulfilled, then the following theorem holds true.

THEOREM 9. *Let us assume that the form a is elliptic in V and that Definitions 7 and 8 hold true. Moreover, let us denote by $\rho_3(h)$ a function going to zero as h tends to zero such that $||u - \Pi u||_H \leq \rho_3(h)||u||_{V^+}$. Let (u,p) and (u_h,p_h) be the solutions of problems* (15) *and* (19)*, respectively. If the time derivatives of u and u_h are bounded in $L^2(]0,T[;H)$, then we have the following error estimate:*

$$\begin{aligned} ||u-u_h||_{L^\infty(H)} + ||u-u_h||_{L^2(V)} &\leq ||u_0 - u_{0,h}||_H \\ &\quad + C\left(\sqrt{\rho_3(h)||f||_{L^2(H)}} + (\rho_1(h)+\rho_2(h))\,||f||_{L^2(H)}\right). \end{aligned} \tag{40}$$

Remark 3. Since $V \subseteq H$ with continuous embedding, we could take $\rho_3(h) = \rho_2(h)$ in the previous theorem. However, we prefer to keep separated the two functions, since in general the approximation in H might be of higher order than in V.

From estimate (40), it is clear that, in order to derive an optimal order of convergence, we need a good balance among the ρ_i, $i = 1,2,3$. We shall discuss this issue in more detail in Example 5.

Going back to the proof of the previous theorem, we notice that, taking in (39) Π as the H projection onto the discrete kernel K_h, we obtain the following different estimate, provided u belongs to $L^\infty(V)$ (see [14] for a similar approach to the analysis of the Navier–Stokes equations):

$$\begin{aligned} &||u-u_h||_{L^\infty(H)} + ||u-u_h||_{L^2(V)} \\ &\qquad \leq ||u_0 - u_{0,h}||_H + C\left(\rho_4(h)||u||_{L^\infty(V)} + \rho_2(h)\|u\|_{L^2(V^+)} + \rho_1(h)||p||_{L^2(Q^+)}\right), \end{aligned} \tag{41}$$

where $\rho_4(h)$, going to zero as h tends to zero, is such that

$$||u - \Pi u||_H \leq \rho_4(h)||u||_V. \tag{42}$$

We notice that this estimate is not fully satisfactory either. Indeed, in some cases, in order to get a good bound for $\rho_4(h)$, one should use an inverse inequality, leading to stronger assumptions on the mesh sequence.

We are now ready to present the second estimate of this section. Let $\Pi : V^+ \times Q^+ \to K_h$ and $P : V^+ \times Q^+ \to Q_h$ denote the elliptic projections; that is, for $u \in V^+$ and $p \in Q^+$,

$$\begin{array}{ll} a(\Pi(u,p),v) + b(v,P(u,p)) = a(u,v) + b(v,p) & \forall v \in V_h, \\ b(\Pi(u,p),q) = 0 & \forall q \in Q_h. \end{array} \tag{43}$$

In order to give sense to (43), we make the assumption that the approximation to the steady equation associated with problem (15) is stable in the sense of [13].

LEMMA 10. *Let (u,p) and (u_h,p_h) be the solutions of problems* (15) *and* (19)*, respectively. Assume that the form a is uniformly elliptic in the discrete kernel; that is,* (16) *is satisfied for any $v \in K_h$ with α independent of h. Then we have*

$$\begin{aligned} &\max_{t\in[0,T]} ||u(t)-u_h(t)||_H^2 + \alpha \int_0^T ||u(t)-u_h(t)||_V^2 \, dt \le ||\Pi(u_0,0)-u_{0,h}||_H^2 \\ &\quad + \max_{t\in[0,T]} ||u(t)-\Pi(u(t),p(t))||_H^2 \\ &\quad + \int_0^T \left(\left\| \frac{\partial}{\partial t}(u(t)-\Pi(u(t),p(t))) \right\|_H^2 + ||u(t)-\Pi(u(t),p(t))||_V^2 \right) dt, \end{aligned} \tag{44}$$

where Π is the elliptic projector mapping $V^+ \times Q^+$ into K_h (see (43)*).*

Proof. Using the ellipticity in the discrete kernel we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}||\Pi(u(t),p(t))-u_h(t)||_H^2 + \alpha||\Pi(u(t),p(t))-u_h(t)||_V^2 \\ \le -\left(\frac{\partial}{\partial t}\Big(u(t)-\Pi(u(t),p(t))\Big), \Pi(u(t),p(t))-u_h(t)\right) \\ - a(u(t)-\Pi(u(t),p(t)), \Pi(u(t),p(t))-u_h(t)) \\ - b(\Pi(u(t),p(t))-u_h(t), p(t)-p_h(t)). \end{aligned}$$

In the last term, p_h can be replaced by $P(u,p) \in Q_h$, since $\Pi(u(t),p(t)) - u_h(t) \in K_h$. Then, using the definition of the elliptic projections (see (43)), we get the desired estimate. ∎

The following theorem gives the convergence result.

THEOREM 11. *Let the hypotheses of Lemma* 10 *be satisfied, and suppose that the* strong approximability *property (see Definition* 8*) is fulfilled. Then, we have the estimate*

$$\begin{aligned} &||u-u_h||_{L^\infty(H)} + \alpha||u-u_h||_{L^2(V)} \\ &\quad \le ||\Pi(u_0,0)-u_{0,h}||_H + \rho_4(h)\left(||u||_{L^\infty(V)} + \left\| \frac{\partial u}{\partial t} \right\|_{L^2(V)} \right) + \rho_2(h)||u||_{L^2(V^+)}, \end{aligned} \tag{45}$$

where $\rho_4(h)$ has been introduced in (42).

Proof. The proof easily follows from the previous lemma, by noticing that the elliptic projection operator Π commutes with the time derivative. ∎

We now present an estimate for the pressure.

THEOREM 12. *Let (u,p) and (u_h,p_h) be the solutions of problems* (15) *and* (19)*, respectively. Let the hypotheses of Lemma* 10 *be satisfied, and suppose that the following discrete inf-sup condition holds with $\beta>0$ independent of h:*

$$\inf_{q_h\in Q_h}\sup_{v_h\in V_h}\frac{b(v_h,q_h)}{||v_h||_V||q_h||_Q}\geq\beta.$$

If there exists $\rho_5(h)$ going to zero as h tends to zero such that

$$||p-P(u,p)||_Q\leq\rho_5(h)||p||_{Q^+}\quad\forall q\in Q^+,$$

then we have the following estimate:

(46)
$$||p-p_h||_{L^2(Q)}\leq C\left(\rho_5(h)||p||_{L^2(Q^+)}+\rho_4(h)\left\|\frac{\partial u}{\partial t}\right\|_{L^2(V)}+||\Pi(u_0,0)-u_{0,h}||_V\right).$$

Proof. For almost any t we have

$$\begin{aligned}\beta||P(u(t),p(t))-p_h(t)||_Q&\leq\sup_{v_h\in V_h}\frac{b(v_h,P(u(t),p(t))-p_h(t))}{||v_h||_V}\\&=\sup_{v_h\in V_h}\frac{b(v_h,P(u(t),p(t))-p(t))-\left(\frac{\partial}{\partial t}\Big(u(t)-u_h(t)\Big),v_h\right)-a(u(t)-u_h(t),v_h)}{||v_h||_V}\\&\leq||p(t)-P(u(t),p(t))||_Q+\left\|\frac{\partial}{\partial t}\Big(u(t)-u_h(t)\Big)\right\|_{V'}+||u(t)-u_h(t)||_V.\end{aligned}$$

In order to conclude the proof we need to estimate the second term. We subtract from the first equation in (43) the first equation in (19) and, taking $v_h\in K_h$, we get

$$\begin{aligned}\left(\frac{\partial}{\partial t}\Big(\Pi(u(t),p(t))-u_h(t)\Big),v_h\right)+a(\Pi(u(t),p(t))-u_h(t),v_h)\\=-\left(\frac{\partial}{\partial t}\Big(u(t)-\Pi(u(t),p(t))\Big),v_h\right).\end{aligned}$$

We take $v_h=\partial(\Pi(u(t),p(t))-u_h(t))/\partial t\in K_h$, and we have

$$\begin{aligned}\left\|\frac{\partial}{\partial t}\Big(\Pi(u(t),p(t))-u_h(t)\Big)\right\|_H^2+\frac{1}{2}\frac{d}{dt}a(\Pi(u(t),p(t))-u_h(t),\Pi(u(t),p(t))-u_h(t))\\\leq\left\|\frac{\partial}{\partial t}\Big(u(t)-\Pi(u(t),p(t))\Big)\right\|_H\left\|\frac{\partial}{\partial t}\Big(\Pi(u(t),p(t))-u_h(t)\Big)\right\|_H.\end{aligned}$$

Integrating over $[0,T]$, we obtain the result in a standard way. ∎

Example 5 (convergence analysis for the approximation to the evolution Stokes problem). Estimate (40) can be used when lowest order elements are used. For instance, if we consider the MINI element (see [2]); then we have $\rho_2(h)=Ch$, $\rho_3(h)=Ch^2$, and $\rho_1(h)$ can be bounded by Ch in a standard way as follows:

$$\sup_{\mathbf{v}_h\in K_h}\frac{b(\mathbf{v}_h,p)}{||\mathbf{v}_h||_{H^1}}=\sup_{\mathbf{v}_h\in K_h}\frac{b(\mathbf{v}_h,p-p^I)}{||\mathbf{v}_h||_{H^1}}\leq Ch||p||_{H^1},$$

where p^I is an approximation of p satisfying $\|p-p^I\|_{L^2} \le Ch\|p\|_{H^1}$. Then, Theorem 9 gives a first order convergence estimate.

Estimate (45) can be used to analyze higher order methods. For instance, when using generalized kth order Hood–Taylor schemes (see [5] and [6]), we have $||\mathbf{u} - \Pi(\mathbf{u},p)||_{L^2} + h||\mathbf{u} - \Pi(\mathbf{u},p)||_{H^1} \le Ch^{k+1}|\mathbf{u}|_{H^{k+1}}$ and Theorem 11 gives a kth order estimate, provided suitable regularity on the solution is assumed, in particular on the time derivative of $\mathbf{u}$. An alternative estimate can be obtained from (41) with no regularity assumptions on $\partial u/\partial t$ but with the need for an inverse inequality. As far as the approximation of the pressure is concerned, the conclusions of Theorem 12 are that $||p - p_h||_{L^2(L^2)}$ is $O(h^k)$, as expected, provided the solution is smooth enough (see (46)).

7. Further considerations on the heat equation in mixed form. In this section we study in more detail the numerical results reported at the end of section 4. We introduce a modified P1 element on criss-cross meshes, which we call P1*, following the notation of [8], and which has a behavior similar to the P1 element with respect to the convergence of eigenmodes. We recall that the criss-cross mesh is constructed by dividing Ω into N-by-N subsquares (macroelements) which are then partitioned into four subtriangles by their diagonals. The elements P1 and P1* present different definitions of both spaces Σ_h and V_h. For the P1* approximation, the space of scalars V_h is made of piecewise constants on the square macroelements and the number of degrees of freedom in Σ_h has been reduced by eliminating the ones corresponding to the centers of the macroelements. The elimination is performed in such a way that the divergences of the elements in Σ_h are constant on each macroelement. We refer the reader to [8] for more details on how to perform the elimination of such degrees of freedom.

To get started, we recall the mixed formulation of the Laplace eigenproblem: Find $\lambda \in \mathbb{R}$ such that there exist $w \in V = L^2_0(\Omega)$ and $\boldsymbol{\sigma} \in \Sigma = H_0(\mathrm{div};\Omega)$ with $w \not\equiv 0$ satisfying

$$\begin{aligned} &(\boldsymbol{\sigma},\boldsymbol{\tau}) + (\mathrm{div}\,\boldsymbol{\tau}, w) = 0 \quad \forall \boldsymbol{\tau} \in \Sigma, \\ &(\mathrm{div}\,\boldsymbol{\sigma}, v) = -\lambda(w,v) \qquad \forall v \in V, \end{aligned} \tag{47}$$

and its numerical approximation with the P1* method: Find $\lambda_h \in \mathbb{R}$ such that there exist $w_h \in V_h$ and $\boldsymbol{\sigma}_h \in \Sigma_h$ with $w_h \not\equiv 0$ satisfying

$$\begin{aligned} &(\boldsymbol{\sigma}_h,\boldsymbol{\tau}) + (\mathrm{div}\,\boldsymbol{\tau}, w_h) = 0 \quad \forall \boldsymbol{\tau} \in \Sigma_h, \\ &(\mathrm{div}\,\boldsymbol{\sigma}_h, v) = -\lambda_h(w_h,v) \quad \forall v \in V_h. \end{aligned} \tag{48}$$

It can be easily shown (in (48), set $w_h = -1/\lambda_h \,\mathrm{div}\,\boldsymbol{\sigma}_h$, which comes from the second equation, into the first equation) that the eigenvalues of (48) correspond to the non-vanishing ones of the following problem: Find $\lambda_h \in \mathbb{R}$ such that there exists $\boldsymbol{\sigma}_h \in \Sigma$ with $\boldsymbol{\sigma}_h \not\equiv 0$ satisfying

$$(\mathrm{div}\,\boldsymbol{\sigma}_h, \mathrm{div}\,\boldsymbol{\tau}) = \lambda_h(\boldsymbol{\sigma}_h,\boldsymbol{\tau}) \quad \forall \boldsymbol{\tau} \in \Sigma_h. \tag{49}$$

In [8] it has been proved and numerically demonstrated that this method does not work. Namely, some *spurious* solutions appear which pollute the numerical spectrum. Here we present the following new result which shows that the only pathology of the method under consideration is the presence of spurious modes; namely, all continuous eigenmodes are correctly approximated.

Let $\Omega =]0,\pi[\times]0,\pi[$; then by separation of variable it is easy to obtain the exact solution to (47):

$$\begin{aligned}\lambda^{mn} &= m^2+n^2, \quad m,n\in\mathbb{N},\ n+m\neq 0,\\ \boldsymbol{\sigma}^{mn}(x,y) &= (m\sin(mx)\cos(ny), n\cos(mx)\sin(nx)),\\ w^{mn} &= -\cos(mx)\cos(ny).\end{aligned}$$

The following proposition and the next corollary give the expressions of the discrete solutions to (49) and (48).

PROPOSITION 13. *Given $N\in\mathbb{N}$, and defined*

$$\begin{aligned}h=\pi/N,\quad \mathrm{cm}=\cos(mh),\quad \mathrm{cn}=\cos(nh),\\ A=(1+1/3\cos(mh)+1/3\cos(nh)+1/3\cos(mh)\cos(nh))(\cos(nh)-\cos(mh)),\\ B_1=\sin(mh)\sin(nh)(4/3+2/3\cos(mh)),\\ B_2=\sin(mh)\sin(nh)(4/3+2/3\cos(nh)),\end{aligned}$$

for $m,n=0,\dots,N-1$, the eigenvalues of scheme (49) *are given by $\lambda_h^{00}=0$,*

(50)
$$\lambda_h^{m0}=\frac{6}{h^2}\frac{1-\cos(mh)}{2+\cos(mh)},\qquad \lambda_h^{0n}=\frac{6}{h^2}\frac{1-\cos(nh)}{2+\cos(nh)},$$
$$\lambda_h^{mn}=\frac{2}{h^2}\frac{4+\mathrm{cm}+\mathrm{cn}-(\mathrm{cm}+\mathrm{cn})^2-\mathrm{cm}\,\mathrm{cn}(\mathrm{cm}+\mathrm{cn})+3\sqrt{A^2+B_1B_2}}{(1+\mathrm{cm}/3+\mathrm{cn}/3+\mathrm{cm}\,\mathrm{cn}/3)(4+\mathrm{cm}+\mathrm{cn})-\sqrt{A^2+B_1B_2}},\qquad mn>0.$$

Setting for $mn>0$,

$$\overline{m}=\sqrt{\frac{A+\sqrt{A^2+B_1B_2}}{B_2}}\sqrt{nm},\qquad \overline{n}=\sqrt{\frac{-A+\sqrt{A^2+B_1B_2}}{B_1}}\sqrt{nm},\tag{51}$$

the eigenfunctions corresponding to (50) *are*

$$\boldsymbol{\sigma}_h^{mn}(x_i,y_j)=\begin{cases}(m\sin(mx_i),0) & \text{if } n=0,\\ (0,n\sin(ny_j)) & \text{if } m=0,\\ (\overline{m}\sin(mx_i)\cos(ny_j),\overline{n}\cos(mx_i)\sin(ny_j)) & \text{if } mn>0.\end{cases}\tag{52}$$

Proof. We provide a sketch of the proof which is mainly a tedious calculation. Considering the reference criss-cross macroelement (based on the square $[0,1]\times[0,1]$), we denote by $\varphi_i(\hat{x},\hat{y})$, $i=1,\dots,5$, the standard continuous piecewise linear functions associated to the nodes represented in Figure 3. It turns out that the eight basis functions for the space Σ_h are

$$\begin{aligned}\psi_1(\hat{x},\hat{y})&=(\varphi_1(\hat{x},\hat{y})+\varphi_5(\hat{x},\hat{y})/4,\varphi_5(\hat{x},\hat{y})/4),\\ \psi_2(\hat{x},\hat{y})&=(\varphi_5(\hat{x},\hat{y})/4,\varphi_1(\hat{x},\hat{y})+\varphi_5(\hat{x},\hat{y})/4),\\ \psi_3(\hat{x},\hat{y})&=(\varphi_2(\hat{x},\hat{y})+\varphi_5(\hat{x},\hat{y})/4,-\varphi_5(\hat{x},\hat{y})/4),\\ \psi_4(\hat{x},\hat{y})&=(-\varphi_5(\hat{x},\hat{y})/4,\varphi_2(\hat{x},\hat{y})+\varphi_5(\hat{x},\hat{y})/4),\\ \psi_5(\hat{x},\hat{y})&=(\varphi_3(\hat{x},\hat{y})+\varphi_5(\hat{x},\hat{y})/4,\varphi_5(\hat{x},\hat{y})/4),\\ \psi_6(\hat{x},\hat{y})&=(\varphi_5(\hat{x},\hat{y})/4,\varphi_3(\hat{x},\hat{y})+\varphi_5(\hat{x},\hat{y})/4),\\ \psi_7(\hat{x},\hat{y})&=(\varphi_4(\hat{x},\hat{y})+\varphi_5(\hat{x},\hat{y})/4,-\varphi_5(\hat{x},\hat{y})/4),\\ \psi_8(\hat{x},\hat{y})&=(-\varphi_5(\hat{x},\hat{y})/4,\varphi_4(\hat{x},\hat{y})+\varphi_5(\hat{x},\hat{y})/4),\end{aligned}$$

where the numbering is taken such that $\psi_{2i-1}(\hat{x},\hat{y})$ and $\psi_{2i}(\hat{x},\hat{y})$ are the two shape functions associated with node i ($i = 1,\dots,4$). Taking $\boldsymbol{\sigma}_h$ as given by (52), it is easy (even if long) to check that the expression for λ_h is the one reported in (50). We acknowledge that, in order to guess the correct expression for (52), we have used Mathematica (see [24]). ∎

COROLLARY 14. *With the same notation as in Proposition* 13*, the solutions* $(\lambda_h, \boldsymbol{\sigma}_h, w_h)$ *of* (48) *are given by* (50), (52)*, and by the following expressions:*

(53)

$$w_h^{m0}|_{K_{ij}} = -\frac{2}{h\lambda_h^{m0}} m \sin\Big(m\frac{h}{2}\Big)\cos\big(mx_{i+\frac{1}{2}}\big),$$

$$w_h^{0n}|_{K_{ij}} = -\frac{2}{h\lambda_h^{0n}} n \sin\Big(n\frac{h}{2}\Big)\cos\big(ny_{j+\frac{1}{2}}\big),$$

$$w_h^{mn}|_{K_{ij}}$$

$$= -\frac{2}{h\lambda_h^{mn}}\Big(\overline{m}\sin\Big(m\frac{h}{2}\Big)\cos\Big(n\frac{h}{2}\Big) + \overline{n}\sin\Big(n\frac{h}{2}\Big)\cos\Big(m\frac{h}{2}\Big)\Big)\cos\big(mx_{i+\frac{1}{2}}\big)\cos\big(ny_{j+\frac{1}{2}}\big),$$

where the last term in (53) *holds for* $mn > 0$ *and* $x_{i+1/2} = x_i + h/2$, $y_{j+1/2} = y_i + h/2$, *and* K_{ij} *is the square of vertices* (x_i, y_j), (x_{i+1}, y_j), (x_{i+1}, y_{j+1}), *and* (x_i, y_{j+1}).

Proof. The proof follows from the expressions for λ_h and $\boldsymbol{\sigma}_h$ given in Proposition 13 and from the formula $w_h = -\operatorname{div}\boldsymbol{\sigma}_h/\lambda_h$. ∎

Let m and n be fixed; then from (50) and (51) we have

$$\lim_{h\to 0}\lambda_h^{mn} = m^2 + n^2 \quad \text{and} \quad \lim_{h\to 0}\overline{m} = m, \quad \lim_{h\to 0}\overline{n} = n,$$

so that it is evident that all continuous eigensolutions are approximated by a suitably chosen discrete one. On the other hand, if we take $m = n = N - 1 = \pi/h - 1$, then

$$\lim_{h\to 0}\lambda_h^{N-1,N-1} = 6,$$

which does not correspond to any continuous eigenvalue of (47). This result is in agreement with the numerical experiments presented in [8], where, in particular, the presence of a sequence of spurious eigenvalues converging to 6 was apparent. For a better understanding of the behavior of the discrete eigenvalues, in Figure 6 we present the graph of λ_h^{mn} as a function of m, n when $N = 100$. The obtained surface is similar to the one obtained in [9] in an analogue situation. We now present the main result of this section, which is closely related to the numerical tests reported in section 4. In that section, we showed two examples in order to demonstrate that the P1 method provides results which are acceptable in one case (*regular* right-hand side) and awful in the other (oscillatory right-hand side). Here we theoretically substantiate those tests, proving that, in general, the P1* methods works if the right-hand side is regular enough. Our theorem is proved under the general hypothesis that any continuous eigensolution to (47) is well approximated by the numerical scheme, no matter whether other spurious solutions are present. For this reason we analyze the P1* method, since to our best knowledge estimates like (50), (52), and (53) are not available for the P1 method. On the other hand, we chose to perform the numerical tests using the P1 method (which seems more natural), even though the P1* method would behave similarly.

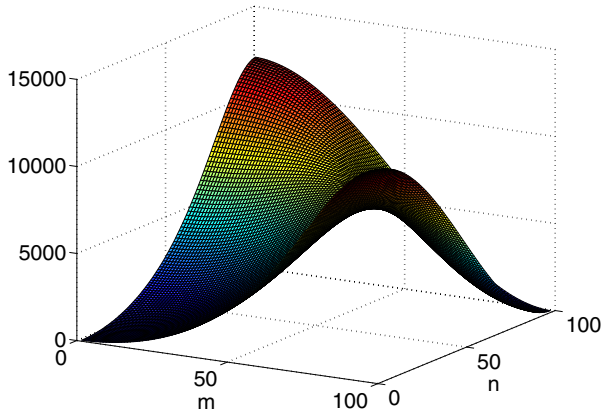


FIG. 6. *Eigenvalues corresponding to formula* (50) *as functions of* m *and* n *when* $N = 100$.

THEOREM 15. *Let us consider the* $P1^*$ *element for the approximation of the heat equation as in scheme* (12)*, and let* $g \in L^\infty(]0,T[;L^2(\Omega))$ *and* $u_0 \in L^2(\Omega)$.

Then, if $(\boldsymbol{\sigma}, u)$ *is the continuous solution and* $(\boldsymbol{\sigma}_h, u_h)$ *the discrete one, we have that* u_h *converges to* u *in* $L^\infty([0,T];L^2(\Omega))$.

Proof. The regularity assumptions on g and u_0 can be written in a more convenient way as follows: For any $\varepsilon > 0$ there exists N such that

$$\sum_{i>N}(g(t),w_i)^2 < \varepsilon, \quad \forall t \in [0,T], \qquad \sum_{i>N}(u_0,w_i)^2 < \varepsilon, \tag{54}$$

where w_i denote the eigenfunctions of problem (47).

The following structure for the continuous and discrete solutions of the mixed heat equation holds:

$$\begin{aligned} u(t) &= \sum_{i=1}^{\infty}\left((u_0,w_i)e^{-\lambda_i t} + \int_0^t (g(s),w_i)e^{-\lambda_i(t-s)}\,ds\right)w_i, \\ u_h(t) &= \sum_{i=1}^{\infty}\left((u_0,w_{i_h})e^{-\lambda_{i,h} t} + \int_0^t (g(s),w_{i,h})e^{-\lambda_{i,h}(t-s)}\,ds\right)w_{i,h}, \end{aligned} \tag{55}$$

where λ_i and w_i (resp., $\lambda_{i,h}$ and $w_{i,h}$ for $i = 1,\ldots,\mathcal{N}(h)$) denote continuous (resp., discrete) eigensolutions of problem (47) (resp., (48)). According to the analysis presented at the beginning of this section (see, in particular, (50) and (53)), they can be ordered in such a way that for any i,

$$\lambda_{i,h} \to \lambda_i \quad \text{and} \quad w_{i,h} \to w_i \quad \text{pointwise in } \Omega \tag{56}$$

as h goes to zero. We explicitly note that in the previous notation we have associated with any eigenvalue λ_i (resp., $\lambda_{i,h}$) a one-dimensional eigenspace spanned by the eigenfunction w_i (resp., $w_{i,h}$); this means in particular that it might be $\lambda_i = \lambda_j$ (and/or, resp., $\lambda_{i,h} = \lambda_{j,h}$) for some $i \neq j$. Moreover, we shall use the orthogonalities $(w_i, w_j) = 0$ (resp., $(w_{i,h}, w_{j,h})$) for $i \neq j$.

The aim of our proof is to show that for any $\varepsilon > 0$ we have $||u(t) - u_h(t)||_0 < \varepsilon$ for any $t \in [0,T]$ when h is small enough. Using (55), we have

$$||u(t) - u_h(t)||_0^2 \leq T_1 + T_2 + T_3 + T_4$$

with

$$T_1 = \sum_{i=N+1}^{\infty} \left((u_0, w_i)^2 e^{-2\lambda_i t} + \int_0^t (g(s), w_i)^2 e^{-2\lambda_i (t-s)}\, ds \right),$$

$$T_2 = \sum_{i=N+1}^{\mathcal{N}(h)} \left((u_0, w_{i,h})^2 e^{-2\lambda_{i,h} t} + \int_0^t (g(s), w_{i,h})^2 e^{-2\lambda_{i,h} (t-s)}\, ds \right),$$

$$T_3 = \left\| \sum_{i=1}^{N} \left((u_0, w_i) w_i e^{-\lambda_i t} - (u_0, w_{i,h}) w_{i,h} e^{-\lambda_{i,h} t} \right) \right\|_0^2,$$

$$T_4 = \int_0^t \left\| \sum_{i=1}^{N} \left((g(s), w_i) w_i e^{-\lambda_i (t-s)} - (g(s), w_{i,h}) w_{i,h} e^{-\lambda_{i,h} (t-s)} \right) \right\|_0^2 ds.$$

Given $\varepsilon > 0$, thanks to the regularity hypotheses (54), we can choose N such that $T_1 < \varepsilon$. The convergence of the eigenvalues and eigenvectors (56) gives that, for h small enough, we also have $T_3 < \varepsilon$ and $T_4 < \varepsilon$. It remains to estimate T_2, which we do now. We shall show that $\sum_{i>N} (u_0, w_{i,h})^2$ can be bounded by 2ε if h is small enough. The term involving $g(s)$ can be handled in the same way and, putting things together, this is what we need in order to get $T_2 < \varepsilon$. The term $\sum_{i>N} (u_0, w_{i,h})^2$ can indeed be estimated in the following way. We have

$$\sum_{i=1}^{N} (u_0, w_{i,h})^2 + \sum_{i=N+1}^{\mathcal{N}(h)} (u_0, w_{i,h})^2 = \sum_{i=1}^{\mathcal{N}(h)} (u_0, w_{i,h})^2$$
$$\leq \sum_{i=1}^{\infty} (u_0, w_i)^2 = \sum_{i=1}^{N} (u_0, w_i)^2 + \sum_{i=N+1}^{\infty} (u_0, w_i)^2 \leq \sum_{i=1}^{N} (u_0, w_i)^2 + \varepsilon.$$

From the convergence of the eigenvectors (56), we have, for h small enough,

$$\sum_{i=1}^{N} (u_0, w_{i,h})^2 - \sum_{i=1}^{N} (u_0, w_i)^2 \leq \varepsilon, \qquad \text{and the bound} \qquad \sum_{i=N+1}^{\mathcal{N}(h)} (u_0, w_{i,h})^2 \leq 2\varepsilon. \qquad \square$$

Remark 4. The proof of the previous theorem strongly relies on (54) and (56). In particular, it shows that any scheme which provides convergent eigenmodes for problem (47) (no matter whether spurious solutions are present) can be successfully applied to the approximation of the heat equation in the case of smooth data (in the sense of (54)).

REFERENCES

[1] D. N. Arnold, D. Boffi, and R. S. Falk, *Quadrilateral H(div) finite elements*, SIAM J. Numer. Anal., to appear.

[2] D. N. Arnold, F. Brezzi, and M. Fortin, *A stable finite element for the Stokes equations*, Calcolo, 21 (1984), pp. 337–344.

[3] I. Babuška, *The finite element method with Lagrangian multipliers*, Numer. Math., 20 (1972/73), pp. 179–192.

[4] I. Babuška and J. Osborn, *Eigenvalue problems*, in Handbook of Numerical Analysis, Vol. II, P. Ciarlet and J. Lions, eds., North–Holland, Amsterdam, 1991, pp. 641–787.

[5] D. Boffi, *Stability of higher order triangular Hood-Taylor methods for the stationary Stokes equations*, Math. Models Methods Appl. Sci., 4 (1994), pp. 223–235.

[6] D. Boffi, *Three-dimensional finite element methods for the Stokes problem*, SIAM J. Numer. Anal., 34 (1997), pp. 664–670.
[7] D. Boffi, F. Brezzi, and L. Gastaldi, *On the convergence of eigenvalues for mixed formulations*, Ann. Scuola. Norm. Sup. Pisa Cl. Sci. (4), 25 (1997), pp. 131–154.
[8] D. Boffi, F. Brezzi, and L. Gastaldi, *On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form*, Math. Comp., 69 (2000), pp. 121–140.
[9] D. Boffi, R. G. Duran, and L. Gastaldi, *A remark on spurious eigenvalues in a square*, Appl. Math. Lett., 12 (1999), pp. 107–114.
[10] F. Brezzi, *On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers*, RAIRO Anal. Numér., 8 (1974), pp. 129–151.
[11] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.
[12] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.
[13] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[14] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* I: *Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
[15] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* II: *Stability of solutions and error estimates uniform in time*, SIAM J. Numer. Anal., 23 (1986), pp. 750–777.
[16] C. Johnson and V. Thomée, *Error estimates for some mixed finite element methods for parabolic type problems*, RAIRO Anal. Numér., 15 (1981), pp. 41–78.
[17] J. L. Lions and E. Magenes, *Problèmes aux limites non-homogènes et applications*, Dunod, Paris, 1968.
[18] B. Mercier, J. Osborn, J. Rappaz, and P.-A. Raviart, *Eigenvalue approximation by mixed and hybrid methods*, Math. Comp., 36 (1981), pp. 427–453.
[19] J.-C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[20] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1994.
[21] P.-A. Raviart and J.-M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, I, Galligani and E. Magenes, eds., Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.
[22] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, rev. ed., North–Holland, Amsterdam, 1979.
[23] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Ser. Comput. Math. 25, Springer-Verlag, Berlin, 1997.
[24] S. Wolfram, *The Mathematica® Book*, 4th ed., Wolfram Media, Inc., Champaign, IL, 1999.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1527–1551

DESIGN OF ABSORBING BOUNDARY CONDITIONS FOR SCHRÖDINGER EQUATIONS IN $\mathbb{R}^{d*}$

JÉRÉMIE SZEFTEL†

Abstract. We construct a family of absorbing boundary conditions for the linear Schrödinger equation on curved boundaries in any dimension which are local both in space and time. We give a convergence result and show that the corresponding initial boundary value problems are well posed. We also prove the nonlinear Schrödinger equation with the absorbing boundary conditions of the linearized problem to be well posed. We finally present numerical results both for the linear and the nonlinear case.

Key words. Schrödinger equation, transparent operator, pseudodifferential calculus, absorbing boundary conditions, well-posedness

AMS subject classifications. 35J10, 58J40, 58J47, 35A07, 65M99

DOI. 10.1137/S0036142902418345

1. Introduction.

The Schrödinger equation on the whole space arises in particular in quantum mechanics and, with some additional nonlinear term, models the propagation of a laser beam. Numerical computations of this equation are therefore often needed. Although the problem is defined in the whole space, it is often sufficient to know the solution only on a bounded domain: the domain of interest. An artificial domain which includes this region of interest is then defined. Inside the domain the equations are discretized in the usual way but there remains the question of the choice of reliable boundary conditions on the artificial boundary. To reduce the computational cost, the numerical domain must be chosen to be slightly larger than the region of interest. Thus the boundary conditions have to be well posed and accurate to be able to approximate the restriction of the solution to the domain of interest.

Several strategies have been developed to find boundary conditions that minimize the reflection of the solution at the artificial boundary. These absorbing boundary conditions have been constructed for hyperbolic problems [12] and parabolic problems [14] with success using pseudodifferential calculus. This is the strategy we develop in this paper for the linear Schrödinger equation.

Engquist and Majda [12] first found a sequence of absorbing boundary conditions for the case of the wave equation in a half space which implies the convergence toward the restriction to this half-space of the solution in the whole space. They also implemented a strategy based on a pseudodifferential factorization of the operator allowing them to derive absorbing boundary conditions for linear hyperbolic equations with variable coefficients.

For linear parabolic equations, Halpern and Rauch [14] also used a pseudodifferential factorization of the operator to construct a family of absorbing boundary conditions on a curved boundary, and they identified the various terms geometrically. In [10], Dubach implemented these boundary conditions in $\mathbb{R}^2$ on a disk.

In the case of the reaction diffusion equations, we used the conditions designed for the linearized heat equation, we proved them to be well posed for the nonlinear

*Received by the editors November 22, 2002; accepted for publication (in revised form) December 5, 2003; published electronically December 16, 2004.
http://www.siam.org/journals/sinum/42-4/41834.html
†LAGA UMR 7539, Institut Galilée, Université Paris 13, 99, avenue J.B. Clément 93430 Villetaneuse, France (szeftel@math.univ-paris.13.fr).

problem and, through numerical experiments, that they were well suited for reaction-diffusion equations (see [27]).

In the case of the linear Schrödinger equation, several authors [6], [8], [25], [4], [5], [22], [3] approximated the transparent condition (the boundary condition satisfied by the exact solution) with a finite difference scheme in one space dimension. The authors in [26], [17], [9], [13], and [1] constructed absorbing boundary conditions for the half-space case, and the author in [9] proved with the Laplace transform in time and Fourier transform in the directions of the boundary that they are well posed. Antoine and Besse [2] wrote absorbing boundary conditions involving the nonlocal operator in time $\sqrt{\partial_t}$ in the case of a curved boundary in two dimensions, and they identified the various terms geometrically. In [7], the authors tested numerically some qualitative properties of absorbing boundary conditions designed for the linear Schrödinger equation on a nonlinear Schrödinger equation.

We extend here all these works for the linear and nonlinear Schrödinger equations. As in [12], we rely on the theory of reflection of singularities we derived in a related paper [28]. The present work consists of two parts. First, we study the linear Schrödinger equation. We recall the expression of the transparent operator obtained in [28] through a result of reflection of singularities for the Schrödinger equation, and we identify geometrically the first two terms in the asymptotic expansion. The transparent condition is not very manageable for numerical simulation and we approximate it with absorbing boundary conditions that are easy to implement. We justify the choice of these absorbing boundary conditions in the case of the half-space with a convergence result and we generalize these conditions for open sets with curved boundary. Then we show the Schrödinger equation to be well posed with these conditions. Finally, we present numerical results in the linear case relying on the optimization of the reflection coefficient.

Second, we study the application to the nonlinear case. We prove the well-posedness of the nonlinear problem with the boundary conditions of the linear one: these boundary conditions have never been studied in the Schrödinger literature. We define an iterative scheme, and we use the estimates of the linear nonhomogeneous problem to show the convergence of the scheme. We finally present numerical results related to the propagation of solitons.

2. The linear Schrödinger equation. We try to approximate the restriction to a bounded region of the solution u_{ex} of the Cauchy problem for the Schrödinger operator:

$$\begin{cases} (i\partial_t + \Delta)u_{ex} = 0 \quad \text{in } \mathbb{R}_t^+ \times \mathbb{R}^d, \\ u_{ex}|_{t=0} = u_0. \end{cases} \tag{2.1}$$

2.1. Absorbing boundary conditions for the Schrödinger equation.

2.1.1. The transparent operator for the Schrödinger equation. We adapt to the Schrödinger equation the work of Engquist and Majda on absorbing boundary conditions for hyperbolic equations [12] and the work of Halpern and Rauch on absorbing boundary conditions for parabolic equations [14]. We first factorize the Schrödinger operator near the boundary of the numerical domain through Nirenberg's method, as in the case of hyperbolic [12] and parabolic [14] equations. We deduce the transparent operator from this factorization using a result of reflection of singularities. This is an important result in itself and was addressed in a separate paper [28].

Let Ω be a bounded convex open subset of $\mathbb{R}^d$ which together with its boundary S is an embedded smooth submanifold with boundary. Let $L = i\partial_t + \Delta$ be the

Schrödinger operator on $\mathbb{R}^d$, u_0 be in $L^2(\mathbb{R}^d)$ with compact support in Ω, and u_{ex} in $C^0(\mathbb{R}_t^+, L^2(\mathbb{R}^d))$ be the unique solution of (2.1).

We consider the Dirichlet problem for L. Let h be in $C_0^\infty(\mathbb{R}\times S)$ so that $h\equiv 0$ for $t\le 0$, and let v be the unique solution of

$$\begin{cases} Lv = 0 \text{ in } \mathbb{R}\times\Omega, \\ v|_{\mathbb{R}\times S} = h, \\ v = 0 \text{ for } t\le 0 \text{ in } \Omega. \end{cases} \tag{2.2}$$

Let ν_Ω be the unit outward pointing normal. We call $N_\Omega : h \to \partial_{\nu_\Omega} v|_S$ the Dirichlet-to-Neumann map for the open set Ω and the operator L.

In $\widetilde{\Omega} = \text{ext}(\Omega)$, u_{ex} is a solution of

$$\begin{cases} Lu_{ex} = 0 \text{ in } \mathbb{R}\times\widetilde{\Omega}, \\ u_{ex} = 0 \text{ for } t\le 0. \end{cases} \tag{2.3}$$

The definition of $N_{\tilde\Omega}$ therefore implies $\partial_{\nu_\Omega} u_{ex} = -N_{\tilde\Omega}(u_{ex}|_{\mathbb{R}\times S})$. Thus u_{ex} is a solution of

$$\begin{cases} Lu_{ex} = 0 \text{ in } \mathbb{R}^+\times\Omega, \\ \partial_{\nu_\Omega} u_{ex} = -N_{\tilde\Omega} u_{ex} \text{ in } \mathbb{R}^+\times S, \\ u_{ex} = u_0 \text{ at } t = 0. \end{cases} \tag{2.4}$$

As in [21], for any even integer r, we define the inhomogeneous Sobolev space,

$$\begin{aligned} H^{r,\frac{r}{2}}(]0,T[\times\Omega) = \{u\in\mathcal{D}'(]0,T[\times\Omega)/\partial_t^l\partial_x^\alpha u \in L^2(]0,T[\times\Omega) \\ \forall(l,\alpha), 2l+|\alpha|\le r\}, \end{aligned} \tag{2.5}$$

and extend it for $r\ge 0$ by interpolation. $u_{ex}|_{\mathbb{R}^+\times\Omega}$ is the unique solution of (2.4).

LEMMA 1. *Let u_0 be in $H^4(\mathbb{R}^d)$ with compact support in Ω and $T>0$. There exists a unique solution w of* (2.4) *in $H^{4,2}(]0,T[\times\Omega)$.*

Proof. The extension of u_0 by 0 outside Ω is in $H^4(\mathbb{R}^d)$ and we still call it u_0. The solution u_{ex} of (2.1) is in $H^{4,2}(]0,T[\times\mathbb{R}^d)$ and taking $w = u_{ex}|_{\mathbb{R}^+\times\Omega}$ imply the existence.

Let w be a solution of (2.4) in $H^{4,2}(]0,T[\times\Omega)$ and let $h = w|_{\mathbb{R}^+\times S}$. The results in [21] imply that h is in $H^{7/2,7/4}(]0,T[\times S)$ and satisfies $h|_{t=0} = \partial_t h|_{t=0} = 0$, and that there exists a unique solution v in $H^{2,1}(]0,T[\times\widetilde{\Omega})$ of (2.2), where $\widetilde{\Omega}$ replaces Ω. Multiplying (2.2) by $\hat v$, taking the imaginary part, and integrating on $\widetilde{\Omega}$ in space and on $]0,T'[$ in time,

$$\|v(T',.)\|^2_{L^2(\Omega)} + \int_0^{T'} [\partial_{\nu_{\tilde\Omega}} v, v] = 0, \tag{2.6}$$

where $0\le T'\le T$, and $[.,.]$ denotes the Hermitian product in $L^2(S)$. Multiplying (2.4) by $\hat w$, taking the imaginary part, and integrating on Ω in space and on $]0,T'[$ in time,

$$\|w(T',.)\|^2_{L^2(\Omega)} - \|u_0\|^2_{L^2(\Omega)} - \int_0^{T'} [N_{\tilde\Omega} w, w] = 0. \tag{2.7}$$

As $w|_{\mathbb{R}^+\times S} = w|_{\mathbb{R}^+\times S} = h$ and $\partial_{\nu_{\tilde\Omega}} v = N_{\tilde\Omega} h$ by definition of $N_{\tilde\Omega}$,

$$\int_0^{T'} [\partial_{\nu_{\tilde\Omega}} v, v] = \int_0^{T'} [N_{\tilde\Omega} w, w],$$

and thus $\int_0^{T'} [N_{\tilde\Omega} w, w] \leq 0$ by (2.6), which implies for all $0 \leq T' \leq T$

$$\|w(T', .)\|^2_{L^2(\Omega)} \leq \|u_0\|^2_{L^2(\Omega)} \tag{2.8}$$

in view of (2.7). (2.8) implies the uniqueness result. □

We first show that $N_{\tilde\Omega}$ is a pseudodifferential operator and compute the first two terms of its symbol. For this we use a change of coordinates which splits the normal coordinate from the tangential ones in L. We recall the expression of the Laplace–Beltrami operator for the Riemannian metric $g = \sum g_{ij} dx_i \otimes dx_j$,

$$\Delta_g v = \sum_{i,j} (\det g)^{-\frac{1}{2}} \partial_{x_i} ((\det g)^{\frac{1}{2}} g^{ij} \partial_{x_j} v),$$

where $g^{ij} = (g^{-1})_{i,j}$. Therefore $L = i\partial_t + \Delta_g$ with g being the canonical scalar product on $\mathbb{R}^d$. In geodesic normal coordinates, g has the form

$$g = (dx_d)^2 + \sum_{\alpha,\beta=1}^{d-1} g_{\alpha\beta} dx_\alpha dx_\beta$$

(i.e., $g_{id} = 0$ for $i = 1, \ldots, d-1$, and $g_{dd} = 1$), and

$$\Delta_g v = \rho^{-1} \partial_{x_d} (\rho \partial_{x_d} v) + \sum_{\alpha,\beta=1}^{d-1} \rho^{-1} \partial_{x_\alpha} (\rho g^{\alpha\beta} \partial_{x_\beta} v),$$

where $\rho = (\det g)^{1/2}$ (with the analyst's sign convention for Δ_g). With the standard notation $D \equiv -i\nabla \equiv (D_1, \ldots, D_d) \equiv (D', D_d)$, the operator L takes the form

$$L = -D_d^2 + i\rho^{-1} \partial_{x_d} \rho D_d + \sum_{\alpha,\beta=1}^{d-1} \rho^{-1} \partial_\alpha (\rho g^{\alpha\beta}) i D_\beta - \sum_{\alpha,\beta=1}^{d-1} g^{\alpha\beta} D_\alpha D_\beta - D_t.$$

Lascar [18] introduces a pseudodifferential algebra well suited to the Schrödinger equation. The symbol class $S^m_{Sch}(\mathbb{R}_t \times \mathbb{R}^d_x)$ is defined by

$$|\partial^\alpha_{t,x} \partial^\beta_{\tau,\xi} p(t, x, \tau, \xi)| \leq C_{\alpha,\beta} (1 + \tau^2 + |\xi|^4)^{\frac{m - 2\beta_\tau - |\beta_\xi|}{4}}. \tag{2.9}$$

Remark. Antoine and Besse [2] use this algebra to compute N_Ω.

We will give an approximation of N_Ω by an operator of $OpS^1_{Sch}(\mathbb{R}_t \times S)$.

Let I be the first fundamental form for the metric g restricted to S. I is given by

$$I(x')(\xi', \xi') = \sum_{\alpha,\beta=1}^{d-1} g_{\alpha\beta} \xi_\alpha \xi_\beta.$$

We define the glancing region $\mathcal{G}$, the hyperbolic region $\mathcal{H}$, and the elliptic region $\mathcal{E}$:

$$\mathcal{G} = \{(t, (x', 0), \tau, \xi') \in T^*(\mathbb{R}_t \times S)/\tau + I(x')(\xi', \xi') = 0\}, \tag{2.10}$$

$$\mathcal{H} = \{(t,(x',0),\tau,\xi') \in T^*(\mathbb{R}_t \times S)/\tau + I(x')(\xi',\xi') < 0\}, \tag{2.11}$$

$$\mathcal{E} = \{(t,(x',0),\tau,\xi') \in T^*(\mathbb{R}_t \times S)/\tau + I(x')(\xi',\xi') > 0\}. \tag{2.12}$$

The following result is proved in [28] using Nirenberg's procedure of factorization (see [23]).

PROPOSITION 1. *Suppose geodesic normal coordinates are introduced as above. Then there are tangential pseudodifferential operators $A(x,D_t,D_{x'})$ and $B(x,D_t,D_{x'})$ in $C^\infty(]-\varepsilon,\varepsilon[_{x_d}, OpS^1_{Sch}(T^*(\mathbb{R}_t\times S)\backslash\mathcal{G}))$ with symbols*

$$A(x,\tau,\xi') \sim \sum_{j\geq 0} A_{1-j}(x,\tau,\xi'),$$
$$B(x,\tau,\xi') \sim \sum_{j\geq 0} B_{1-j}(x,\tau,\xi')$$

with A_k and B_k satisfying $A_k(x,\lambda^2\tau,\lambda\xi') = \lambda^k A_k(x,\tau,\xi')$ and $B_k(x,\lambda^2\tau,\lambda\xi') = \lambda^k B_k(x,\tau,\xi')$ for all $\lambda > 0$ such that

$$\begin{aligned} &L = -(D_{x_d}+A)(D_{x_d}+B) \ \textit{mod}\ C^\infty(]-\varepsilon,\varepsilon[_{x_d}: OpS^{-\infty}_{Sch}(T^*(\mathbb{R}_t\times S)\backslash\mathcal{G})),\\ &B_1 \in i\mathbb{R}^+ \textit{ in } \mathcal{E} \textit{ and } B_1 \in \mathbb{R}^+ \textit{ in } \mathcal{H}. \end{aligned} \tag{2.13}$$

We compute B_1 and B_0:

$$\begin{aligned}(D_{x_d}+A)(D_{x_d}+B) &= D^2_{x_d} + (A+B)D_{x_d} + AB + [D_{x_d},B]\\ &\sim D^2_{x_d} + (A+B)D_{x_d} - i\frac{\partial B}{\partial x_d} + \sum(\partial^\alpha_{\tau,\xi'}A)(D^\alpha_{t,x'}B)/\alpha!.\end{aligned}$$

Comparing with the expression of L, we identify

$$\begin{aligned} B_1 &= \left(-\tau - \sum_{\alpha,\beta} g^{\alpha\beta}\xi_\alpha\xi_\beta\right)^{\frac{1}{2}},\\ B_0 &= -i\frac{\rho^{-1}\partial_{x_d}\rho}{2} + \frac{i\sum \rho^{-1}\partial_\alpha(\rho g^{\alpha\beta})\xi_\beta + D_{x_d}B_1 - \sum \partial_{\xi_\alpha}B_1 D_{x_\alpha}B_1}{2B_1}. \end{aligned} \tag{2.14}$$

The following proposition allows us to compute the Dirichlet-to-Neumann map. It is proved in [28] using a result of reflection of singularities for Schrödinger's equation.

PROPOSITION 2. *Let $N_{\tilde\Omega}$ be the Dirichlet-to-Neumann map for the open set $\widetilde\Omega$ and let B be the operator appearing in* (2.13). *If Ω is convex, then $N_{\tilde\Omega} = -iB + OpS^{-\infty}_{Sch}(T^*(\mathbb{R}_t\times S)\backslash\mathcal{G})$.*

Remark. Antoine and Besse [2] proved the previous factorization in the two-dimensional case, and they formally wrote $N_{\tilde\Omega} = -iB$.

Proposition 2 shows that $N_{\tilde\Omega}$ is a pseudodifferential operator, and we computed the first two terms of its symbol. Consider the Schrödinger operator L_0 on the Riemannian product manifold $S\times\mathbb{R}^+_{x_d}$ with metric equal to $g_S + g_{\mathbb{R}} = g_S + (dx_d)^2$. Then $L_0 = i\partial_t + \Delta_S + \partial^2_d$. For this operator, $N_0^2 = -i\partial_t - \Delta_S$, where N_0 is the Dirichlet-to-Neumann map for L_0 and $S\times\mathbb{R}^+_{x_d}$. (2.14) gives a pseudodifferential asymptotic expansion of N_0:

$$\sigma((-i\partial_t - \Delta_S)^{\frac{1}{2}}) = \sigma(N_0) = -iB_1 + \frac{\sum \rho^{-1}\partial_\alpha(\rho g^{\alpha\beta})\xi_\beta + i\sum \partial_{\xi_\alpha}B_1 D_{x_\alpha}B_1}{2B_1} + S^{-1}_{Sch};$$

thus we obtain

$$\sigma(N_{\tilde{\Omega}}) = \sigma((-i\partial_t - \Delta_S)^{\frac{1}{2}}) - i\frac{D_{x_d}B_1}{2B_1} - \frac{\rho^{-1}\partial_{x_d}\rho}{2} + S_{Sch}^{-1}. \tag{2.15}$$

The geometrical interpretation in [14] remains. Let H be the mean curvature, and let I and II be the two first fundamental forms for the metric g restricted to S. H and II are given by

$$H = \frac{1}{d-1}\rho^{-1}\partial_d\rho, \text{ and } II(\xi',\xi') = -\frac{1}{2}\sum_{\alpha,\beta=1}^{d-1}\partial_{x_d}g^{\alpha\beta}\xi_\alpha\eta_\beta.$$

Then the proposition follows.

PROPOSITION 3. *The symbol of $N_{\tilde{\Omega}}$ is given by*

$$\sigma(N_{\tilde{\Omega}}) = \sigma((-i\partial_t - \Delta_S)^{\frac{1}{2}}) - \frac{d-1}{2}H + \frac{II(\xi',\xi')}{2(\tau + I(\xi',\xi'))} + S_{Sch}^{-1}.$$

Remark. This is a generalization of the result obtained by Antoine and Besse [2] in two dimensions.

2.1.2. A sequence of absorbing boundary conditions for the half-space. Having in mind (2.4), we solve the problem

$$\begin{cases} Lu_{app} = 0 & \text{in } \mathbb{R}^+ \times \Omega, \\ \partial_{\nu_\Omega} u_{app} = F u_{app} & \text{in } \mathbb{R}^+ \times S, \\ u_{app} = u_0 & \text{at } t = 0, \end{cases} \tag{2.16}$$

where the operator F is an approximation of $-N_{\tilde{\Omega}}$ in a sense that we will make more precise later. We may keep only the first terms in the asymptotic expansion of $N_{\tilde{\Omega}}$ so that we let the singular part of the solution leave the domain. (We will keep B_1 or $B_1 + B_0$ in what follows.) But $(-i\partial_t - \Delta_S)^{\frac{1}{2}}$ is nonlocal in space and time and thus is not practical for numerical computations. We need to approximate it with operators local in space and time leading to well-posed initial boundary value problems.

Here we adapt to Schrödinger's equation the strategy used in [12] for the wave equation.

Let d be an integer, $d \geq 2$, and h be a function in $H^{7/2,7/4}(\mathbb{R}_t^+ \times \mathbb{R}_{x'}^{d-1})$. We assume that $h|_{t=0} = 0$ and $\partial_t h|_{t=0} = 0$. For all $T > 0$, there exists a unique solution $\tilde{u}$ in $H^{2,1}(]0,T[\times\mathbb{R}_+^{d-1})$ (see [21]) of (2.2) in the half space $x \geq 0$:

$$\begin{cases} i\partial_t\tilde{u} + \Delta\tilde{u} = 0, \ t \geq 0, \ \ y \in \mathbb{R}^{d-1}, \ \ x \geq 0, \\ \tilde{u} = 0 \text{ for } t \leq 0, \\ \tilde{u}(0,y,t) = h(y,t). \end{cases} \tag{2.17}$$

Remark. Taking $h = u_{ex}|_{x=0}$ and u_0 with support in $x < 0$, we see from (2.3) that $\tilde{u} = u_{ex}|_{x>0}$.

Let $\delta > 0$ be an acceptable error, let T be a strictly positive time, and let a_0 be a strictly positive real number. We look for a boundary operator F on the hyperplane

$x = a$, with $a \geq a_0$ as close to a_0 as possible, such that if u is the solution of

$$\begin{cases} i\partial_t u + \Delta u = 0, \ t \geq 0, \ \ y \in \mathbb{R}^{d-1}, \ \ 0 < x < a, \\ u = 0 \ \text{ for } \ t \leq 0, \\ u(0, y, t) = h(y, t), \\ \partial_x u(a, y, t) = Fu(a, y, t), \end{cases} \tag{2.18}$$

then u satisfies

$$\left(\int_0^T \int_{\mathbb{R}^{d-1}} \int_0^{a_0} |\tilde{u} - u|^2 dx dy dt \right)^{\frac{1}{2}} \leq \delta. \tag{2.19}$$

Moreover, we seek F as a local operator in space and time leading to a well-posed problem for the Schrödinger operator. As in [12], to get (2.19) we will minimize the amplitude of the reflected waves.

Given $\eta > 0$, the Laplace transform in t and the Fourier transform in y are defined for w such that $e^{-\eta t} w$ is in $L^2(]0, +\infty[_t \times \mathbb{R}^{d-1})$ by:

$$\mathcal{L}w(s, \omega) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_0^{+\infty} \int_{\mathbb{R}^{d-1}} \exp(-(\eta + i\tau)t - i\omega y) w(t, y) dt dy,$$

where $s = \eta + i\tau$. For all $\eta > 0$, $e^{-\eta t}\tilde{u}(x, .)$ is in $L^2(\mathbb{R}_t^+ \times \mathbb{R}^{d-1})$ for almost all $x > 0$, and its Fourier–Laplace transform satisfies the following ordinary differential equation:

$$\frac{d^2}{dx^2} \mathcal{L}\tilde{u} - (-is + |\omega|^2) \mathcal{L}\tilde{u} = 0.$$

The Fourier–Laplace transform of $\tilde{u}$ is not more than slowly increasing because it is a tempered distribution. Therefore, $\mathcal{L}\tilde{u} = \exp(-\varrho x)\mathcal{L}h$, where $\varrho = (-is + |\omega|^2)^{1/2}$, and we choose the square root branch well defined for $-\pi < \arg(z) < \pi$. At $x = a$, $\mathcal{L}\tilde{u}$ satisfies

$$\frac{d}{dx} \mathcal{L}\tilde{u} + \varrho \mathcal{L}\tilde{u} = 0.$$

We will thus seek a sequence of local operator F_m such that its symbol approximates $-\varrho$.

To write absorbing boundary conditions for the heat equation on a half-space, Joly used the following sequence of rational functions in [16]:

$$\begin{aligned} f_{m+1}(z) &= 1 + \frac{z}{1 + f_m(z)}, \\ f_0(z) &= 0. \end{aligned} \tag{2.20}$$

In the following lemma, we group together the properties that will be used in what follows.

LEMMA 2. *Let Δ be the set of all real numbers smaller than -1. For all complex numbers z in $\mathbb{C} \setminus \Delta$ we have the equality*

$$\frac{f_m(z) - (1+z)^{\frac{1}{2}}}{f_m(z) + (1+z)^{\frac{1}{2}}} = -\left(\frac{1 - (1+z)^{\frac{1}{2}}}{1 + (1+z)^{\frac{1}{2}}} \right)^m. \tag{2.21}$$

In particular, we have for any set K in $\mathbb{C}\backslash\Delta$

$$\sup_{z\in K}\left|\frac{f_m(z)-(1+z)^{\frac{1}{2}}}{f_m(z)+(1+z)^{\frac{1}{2}}}\right|\leq\gamma(K)^m$$

with $\gamma(K)=\sup_{z\in K}\left|\frac{1-(1+z)^{\frac{1}{2}}}{1+(1+z)^{\frac{1}{2}}}\right|\leq 1$. If K is a compact subset of $\mathbb{C}\backslash\Delta$, then $\gamma(K)<1$. Finally, we have for all integer m

$$f_{2m}(z)=\frac{1}{m}\sum_{k=0}^{m-1}\frac{(z+1)\ \cos^2\left(\frac{(2k+1)\pi}{4m}\right)}{z+1+\tan^2\left(\frac{(2k+1)\pi}{4m}\right)}. \tag{2.22}$$

As $\varrho=(-is+|\omega|^2)^{1/2}=-i(is-|\omega|^2)^{1/2}$ we may approximate it by $\varrho_m=-if_{2m}(is-|\omega|^2-1)$. It is the same as approximating the transparent operator by the operator F_m which symbol is $if_{2m}(is-|\omega|^2-1)$. Using Lemma 2, we can rewrite the symbol of F_m as

$$\sigma(F_m)=i\beta_m+i\sum_{k=0}^{m-1}\frac{a_{km}(is-|\omega|^2)}{is-|\omega|^2+d_{km}}, \tag{2.23}$$

where $\beta_m=0$, $a_{km}=1/\cos^2(\frac{(2k+1)\pi}{4m})$, and $d_{km}=\tan^2(\frac{(2k+1)\pi}{4m})$. In particular, $\beta_m\geq 0$, $a_{km}>0$, and $d_{km}>0$. Following Lindmann [19], we introduce the auxiliary functions φ_{km} defined in the Fourier–Laplace domain by

$$\mathcal{L}\varphi_{km}=\frac{\mathcal{L}v}{is-|\omega|^2+d_{km}}.$$

F_m is then defined by

$$F_mv=i\beta_mv+i\sum_{k=0}^{m-1}a_{km}(i\partial_t+\Delta_y)\varphi_{km},$$

where φ_{km} is defined by

$$\begin{cases}(i\partial_t+\Delta_y+d_{km})\varphi_{km}=v\ \text{ in }\ \mathbb{R}_t^+\times\mathbb{R}_{x'}^{d-1},\\ \varphi_{km}=0\ \text{ at }\ t=0.\end{cases}$$

F_m is therefore local in space and time.

We will show the well-posedness of Schrödinger's equation with boundary conditions F_m in Proposition 6. (We prove it for a bounded domain but it extends easily to the half-space case.) We show here that the solution u_m of (2.18) with $F=F_m$ satisfies (2.19) for sufficiently large m. We have the following proposition.

PROPOSITION 4. *Let δ be an acceptable error, let a_0 be a strictly positive real number, let T be a strictly positive time, let h be in $H^{7/2,7/4}(\mathbb{R}_t^+\times\mathbb{R}_{x'}^{d-1})$ so that $h|_{t=0}=0$ and $\partial_t h|_{t=0}=0$, and let u_m be the solution of* (2.18) *with $F=F_m$. Then u_m is in $H^{2,1}(]0,T[\times\mathbb{R}_+^d)$ and there exists $a\geq a_0$ and an integer m so that u_m satisfies* (2.19).

Proof. We use the results in [21] for the nonhomogeneous Dirichlet condition at $x = 0$ and Proposition 6 for the absorbing boundary conditions at $x = a$ to prove the existence and the regularity of u_m. $e^{-\eta t}\tilde{u}$ and $e^{-\eta t}u_m$ are in $L^2(\mathbb{R}_t^+ \times \mathbb{R}_{x'}^{d-1}\times]0,a[)$ because h is in $H^{7/2,7/4}(\mathbb{R}_t^+ \times \mathbb{R}_{x'}^{d-1})$. Using the Parseval equality,

$$\int_0^T \int_{\mathbb{R}^{d-1}} \int_0^{a_0} |e^{-\eta t}\tilde{u} - e^{-\eta t}u|^2 dx dy dt \le \int_{\mathbb{R}^d} \int_0^{a_0} |\mathcal{L}\tilde{u} - \mathcal{L}u|^2 dx d\omega d\tau. \tag{2.24}$$

The map taking $h \in H^{7/2,7/4}(]0,T[\times\mathbb{R}_{x'}^{d-1})$ to $\tilde{u} \in H^{2,1}(]0,T[\times\mathbb{R}_+^d)$ is continuous. It is therefore sufficient to prove the proposition for h in a dense subset of $H^{7/2,7/4}(]0,T[\times\mathbb{R}_{x'}^{d-1})$. We may assume there exists a real number $0 < \varepsilon < 1$ such that the support of $\mathcal{L}h$ is included in $\{(\tau,\omega)/|\tau + |\omega|^2| \ge \varepsilon \text{ and } \tau + |\omega|^2 \ge -1/\varepsilon\}$.

As $\mathcal{L}\tilde{u}$, $\mathcal{L}u_m$ satisfies a second-order ordinary differential equation with constant coefficients. We compute $\mathcal{L}u_m$ explicitly using boundary conditions at $x = 0$ and $x = a$. We obtain

$$\mathcal{L}u_m - \mathcal{L}\tilde{u} = (\exp(\varrho x) - \exp(-\varrho x))\frac{R_m}{R_m - \exp(2\varrho a)}\mathcal{L}h, \tag{2.25}$$

where R_m is the reflection coefficient defined by

$$R_m(s,i\omega) = \frac{-(-is + |\omega|^2)^{\frac{1}{2}} - if_{2m}(is - |\omega|^2 - 1)}{(-is + |\omega|^2)^{\frac{1}{2}} - if_{2m}(is - |\omega|^2 - 1)}.$$

Lemma 2 implies that $|R_m(s,i\omega)| < 1$ for all $\eta > 0$. We define the following three regions of $\mathbb{R}^d$:

$$\begin{aligned} G_1 &= \{(\tau,\omega)/|\tau + |\omega|^2| \le \varepsilon \text{ or } \tau + |\omega|^2 \le -\tfrac{1}{\varepsilon}\},\\ G_2 &= \{(\tau,\omega)/ -\tfrac{1}{\varepsilon} \le \tau + |\omega|^2 \le -\varepsilon\},\\ G_3 &= \{(\tau,\omega)/\tau + |\omega|^2 \ge \varepsilon\}. \end{aligned}$$

In G_1, $\mathcal{L}(u_n - \tilde{u}) = 0$.

Let γ_ε be equal to $\sup_{z\in K_\varepsilon}\left|\frac{1-(1+z)^{\frac{1}{2}}}{1+(1+z)^{\frac{1}{2}}}\right|$, where K_ε is the compact set of $\mathbb{C} \setminus \Delta$ defined by $\{\theta + i\eta/\varepsilon - 1 \le \theta \le 1/\varepsilon - 1,\ 0 \le \eta \le 1\}$. $0 < \gamma_\varepsilon < 1$ by Lemma 2 and (2.25) implies the following inequality in G_2:

$$|\mathcal{L}u_m - \mathcal{L}\tilde{u}| \le \frac{2\gamma_\varepsilon^{2m}}{1 - \gamma_\varepsilon}|\mathcal{L}h|.$$

In G_3, we have the inequality

$$|\mathcal{L}u_m - \mathcal{L}\tilde{u}| \le \frac{\exp(\sqrt{\varepsilon}(a_0 - 2a))|\mathcal{L}h|}{1 - \exp(-2a_0\sqrt{\varepsilon})}.$$

Summing up, we finally get

$$\begin{aligned} \int_0^{a_0}\int_{\mathbb{R}^+}\int_{\mathbb{R}^{d-1}} |\mathcal{L}u_m - \mathcal{L}\tilde{u}|^2 &= \int_0^{a_0}\int_{G_2} |\mathcal{L}u_m - \mathcal{L}\tilde{u}|^2 + \int_0^{a_0}\int_{G_3} |\mathcal{L}u_m - \mathcal{L}\tilde{u}|^2\\ &\le \frac{4a_0\gamma_\varepsilon^{4m}}{(1-\gamma_\varepsilon)^2}\int_{G_2}|\mathcal{L}h|^2 + \frac{\exp(2\sqrt{\varepsilon}(a_0 - 2a))a_0}{(1 - \exp(-2a_0\sqrt{\varepsilon}))^2}\int_{G_3}|\mathcal{L}h|^2. \end{aligned}$$

We choose m such that $\frac{4a_0\gamma_\varepsilon^{4m}}{(1-\gamma_\varepsilon)^2}\int_{G_2}|\mathcal{L}h|^2 \leq \frac{\delta}{2}$ and a such that

$$\frac{\exp(2\sqrt{\varepsilon}(a_0-2a))a_0}{(1-\exp(-2a_0\sqrt{\varepsilon}))^2}\int_{G_3}|\mathcal{L}h|^2 \leq \frac{\delta}{2}.$$

Using (2.24), we obtain (2.19) by letting η converge toward 0. □

2.1.3. Absorbing boundary conditions in the general case. The study in section 2.1.2 suggests the following approximation of $(-i\partial_t - \Delta_S)^{1/2}$:

$$(-i\partial_t - \Delta_S)^{\frac{1}{2}} \approx i\beta + i\sum_{k=1}^{m} a_k(i\partial_t + \Delta_S)(i\partial_t + \Delta_S + d_k)^{-1},$$

where $\beta \geq 0$, $a_k > 0$, and $d_k > 0$ again. The term $(i\partial_t + \Delta_S + d_k)^{-1}$ can again be handled with auxiliary functions:

$$\begin{cases}(i\partial_t + \Delta_S + d_k)\varphi_k = u \text{ in } \mathbb{R}^+ \times S,\\ \varphi_k = 0 \text{ at } t = 0.\end{cases} \tag{2.26}$$

Remark. In the case of the half-space, the conditions on β, a_k, and d_k imply that the reflection coefficient $R(\eta + i\tau, \omega)$ is less than one when $\eta = 0$ and holomorphic when $\eta > 0$. Thus, the maximum principle implies that $|R| < 1$ when $\eta > 0$ as in the proof of Proposition 4.

If we keep only $B_1 + B_0$ in the pseudodifferential expansion, we approximate $N_{\tilde{\Omega}}$ by $-F$ with F defined by

$$F = i\beta u + i\sum_{k=1}^{m} a_k(i\partial_t + \Delta_S)\varphi_k - \frac{d-1}{2}H_{\tilde{\Omega}}u + \frac{II(D',D')}{2}\varphi_0, \tag{2.27}$$

where

$$\begin{cases}(i\partial_t + \Delta_S)\varphi_0 = u \text{ in } \mathbb{R}^+ \times S,\\ \varphi_0 = 0 \text{ at } t = 0.\end{cases} \tag{2.28}$$

2.2. The linear problem with absorbing boundary conditions.

2.2.1. Properties of the boundary operator F. The auxiliary functions are defined by (2.26) and (2.28). F is the operator defined by

$$F : u \to i\beta u + i\sum_{k=1}^{m} a_k(u - d_k\varphi_k) - \frac{d-1}{2}H_\Omega u + II(D',D')\frac{\varphi_0}{2}.$$

Let η be a strictly positive real number. We introduce the following auxiliary functions:

$$\begin{cases}(i\partial_t + \Delta_S + i\eta + d_k)\varphi_{\eta k} = u \text{ in } \mathbb{R}^+ \times S,\\ \varphi_{\eta k} = 0 \text{ at } t = 0.\end{cases} \tag{2.29}$$

F_η is the operator defined by

$$F_\eta : u \to i\beta u + i\sum_{k=1}^{m} a_k(u - d_k\varphi_{\eta k}) - \frac{d-1}{2}H_\Omega u + II(D',D')\frac{\varphi_{\eta 0}}{2}.$$

Remark. F equals F_η when $\eta = 0$.

Let b be the continuous bilinear form on $H^1(S) \times H^1(S)$ such that

$$\left[-\frac{II(D', D')v}{2}, w\right] = b(v, w) \ \ \forall v \in H^2(S) \quad \forall w \in H^1(S).$$

To obtain energy estimates, we assume that Ω is convex. Therefore, H_Ω is nonnegative on S and b is a nonnegative bilinear form in this case.

Remark. We did not need this additional hypothesis for the heat equation (see [27]) due to the more favorable estimates derived in that case. In any case, the convexity hypothesis has been used to compute the transparent operator (see Proposition 2).

PROPOSITION 5. *Let T be any strictly positive time, let s be any real number, and let u be in $H^1(]0, T[, H^s(S))$. For all $\eta \geq 0$, there exist functions $\varphi_{\eta k}$ in $L^\infty(]0, T[, H^{s+2}(S)) \cap W^{1,\infty}(]0, T[, H^s(S))$ for $k = 0, \ldots, m$ solutions of (2.29). Moreover, F_η satisfies the following inequalities in $\mathcal{D}'(0, T)$:*

$$Im\,([F_\eta u, u]) \geq \frac{1}{2}\frac{d}{dt} b(\varphi_{\eta 0}, \varphi_{\eta 0}) + \eta b(\varphi_{\eta 0}, \varphi_{\eta 0}) - C\|u\|^2_{L^2(S)} \tag{2.30}$$

and

$$Re\,([F_\eta u, \partial_t u]) \leq C(\|u\|^2_{L^2(S)} + \|\partial_t u\|^2_{L^2(S)}). \tag{2.31}$$

Finally, if $\eta > d_k$ for $k = 1, \ldots, m$ we define $c_\eta = \sum_{k=1}^m a_k(1 - d_k/\eta)$. Then $c_\eta > 0$ and F_η satisfies the following inequality:

$$\int_0^T Im\,([F_\eta u, u]) \geq (\beta + c_\eta) \int_0^T \|u\|^2_{L^2(S)} + \frac{1}{2} b(\varphi_{\eta 0}(T), \varphi_{\eta 0}(T)) + \int_0^T \eta b(\varphi_{\eta 0}, \varphi_{\eta 0}). \tag{2.32}$$

Proof. We first prove the existence and regularity of $\varphi_{k\eta}$. We prove it in the case $s = 0$. The case where s is an integer follows by iteration differentiating the equation along any C^∞ vector field on S. The case where s is a nonnegative real number follows by interpolation. (See [27] in the case of the heat equation.)

Let d be a nonnegative real number and let u be in $H^1(]0, T[, L^2(S))$. We will prove the existence and uniqueness of φ_η in $W^{1,\infty}(]0, T[, L^2(S)) \cap L^\infty(]0, T[, H^2(S))$ solution of

$$\begin{cases} \forall \psi \in H^1(S), \\ i\frac{d}{dt}[\varphi_\eta, \psi] - [\nabla_S \varphi_\eta, \nabla_S \psi] + (d + i\eta)[\varphi_\eta, \psi] = [u, \psi] \ \text{ in } \ \mathcal{D}'(0, T), \\ \varphi_\eta = 0 \ \text{ at t = 0.} \end{cases} \tag{2.33}$$

Multiplying the equation by $\overline{\varphi_\eta}$, integrating it in space, and taking the imaginary part, we obtain a first a priori estimate:

$$\frac{1}{2}\frac{d}{dt}\|\varphi_\eta\|^2_{L^2(S)} + \eta\|\varphi_\eta\|^2_{L^2(S)} \leq \frac{1}{2}\|\varphi_\eta\|^2_{L^2(S)} + \frac{1}{2}\|u\|^2_{L^2(S)}. \tag{2.34}$$

Multiplying the equation by $\overline{\partial_t \varphi_\eta}$, integrating it in space, and taking the real part, we obtain a second a priori estimate:

$$\frac{1}{2}\frac{d}{dt}\|\nabla_S \varphi_\eta\|^2_{L^2(S)} \leq C(\|\varphi_\eta\|^2_{L^2(S)} + \|\partial_t \varphi_\eta\|^2_{L^2(S)} + \|u\|^2_{L^2(S)}). \tag{2.35}$$

Differentiating in time the equation satisfied by φ_η, multiplying it by $\overline{\partial_t\varphi_\eta}$, integrating it in space, and taking the imaginary part, we obtain a third a priori estimate:

$$\frac{1}{2}\frac{d}{dt}\|\partial_t\varphi_\eta\|^2_{L^2(S)} + \eta\|\partial_t\varphi_\eta\|^2_{L^2(S)} \leq \frac{1}{2}\|\partial_t\varphi_\eta\|^2_{L^2(S)} + \frac{1}{2}\|\partial_t u\|^2_{L^2(S)}. \tag{2.36}$$

We deduce from these three energy estimates and Galerkin's method the existence and uniqueness of φ_η in $W^{1,\infty}(]0,T[,L^2(S)) \cap L^\infty(]0,T[,H^1(S))$ solution of (2.33).

Since $\Delta_S\varphi_\eta = -i\partial_t\varphi_\eta + u$, $\Delta_S\varphi_\eta$ is in $L^\infty(]0,T[,L^2(S))$. Thus φ_η belongs to $L^\infty(]0,T[,H^2(S))$ by using the regularity of weak solutions of second-order elliptic equations on a manifold, and we obtain the following estimate:

$$\|\varphi_\eta\|_{L^\infty(]0,T[,H^2(S))} \leq C(\|\varphi_\eta\|_{W^{1,\infty}(]0,T[,L^2(S))} + \|u\|_{H^1(]0,T[,L^2(S))}). \tag{2.37}$$

The estimates (2.34), (2.36), and (2.37) and Gronwall's lemma yield

$$\|\varphi_\eta\|_{L^\infty(]0,T[,H^2(S))} + \|\varphi_\eta\|_{W^{1,\infty}(]0,T[,L^2(S))} \leq C\|u\|_{H^1(]0,T[,L^2(S))}. \tag{2.38}$$

It remains to prove (2.30), (2.31), and (2.32). We first recall the following lemma (see [20]).

LEMMA 3. *Let V, H, and V' be three Hilbert spaces such that $V \subset H \subset V'$ and V' is the dual space of V. If φ is in $L^2(]0,T[,V)$ and φ' is in $L^2(]0,T[,V')$, then φ equals almost everywhere a continuous function from $[0,T]$ to H and $\frac{d}{dt}\|\varphi\|^2_H = 2Re < \varphi', \varphi >$.*

Using the previous lemma with $H = H^1(S)$, $V = H^2(S)$ and $V' = H^{-2}(S)$, we obtain $\frac{d}{dt}b(\varphi_{\eta 0},\varphi_{\eta 0}) = 2\mathrm{Re}\, b(\varphi_{\eta 0},\partial_t\varphi_{\eta 0})$. Using the regularity of u and $\varphi_{\eta k}$, we have

$$\begin{aligned}
\mathrm{Im}\,([F_\eta u,u]) &= \beta\|u\|^2_{L^2(S)} + \sum_{k=1}^m a_k \mathrm{Re}\,([u-d_k\varphi_{\eta k},u]) \\
&\quad + \mathrm{Im}\left(\left[II(D',D')\frac{\varphi_{\eta 0}}{2}, i\partial_t\varphi_{\eta 0} + \Delta_S\varphi_{\eta 0} + i\eta\varphi_{\eta 0}\right]\right) \\
&\geq \frac{1}{2}\frac{d}{dt}b(\varphi_{\eta 0},\varphi_{\eta 0}) + \eta b(\varphi_{\eta 0},\varphi_{\eta 0}) + \beta\|u\|^2_{L^2(S)} \\
&\quad - \sum_{k=1}^m a_k\|u-d_k\varphi_{\eta k}\|_{L^2(S)}\|u\|_{L^2(S)},
\end{aligned}$$

which yields (2.30). Then, let e be equal to $\frac{d-1}{2}\|H_\Omega\|_{L^\infty(S)}$:

$$\begin{aligned}
\mathrm{Re}\,([F_\eta u,\partial_t u]) &= -\beta\mathrm{Im}\,([u,\partial_t u]) + \sum_{k=1}^m \mathrm{Re}\,(ia_k[u-d_k\varphi_{\eta k},\partial_t u]) \\
&\quad - \frac{d-1}{2}\mathrm{Re}\,([H_\Omega u,\partial_t u]) + \mathrm{Re}\left(\left[II(D',D')\frac{\varphi_{\eta 0}}{2},\partial_t u\right]\right) \\
&\leq \left(\frac{d-1}{2}\|H_\Omega\|_{L^\infty(S)} + \beta\right)\|u\|_{L^2(S)}\|\partial_t u\|_{L^2(S)} \\
&\quad + \sum_{k=1}^m a_k\|u-d_k\varphi_{\eta k}\|_{L^2(S)}\|\partial_t u\|_{L^2(S)} + C\|\varphi_{\eta 0}\|_{H^2(S)}\|\partial_t u\|_{L^2(S)},
\end{aligned}$$

which yields (2.31) by using (2.38). Finally we have

$$\begin{aligned}\mathrm{Im}\left([F_\eta u,u]\right) &= \beta\|u\|^2_{L^2(S)} + \sum_{k=1}^m a_k \mathrm{Re}\left([u-d_k\varphi_{\eta k},u]\right)\\ &\quad + \mathrm{Im}\left(\left[II(D',D')\frac{\varphi_{\eta 0}}{2}, i\partial_t\varphi_{\eta 0} + \Delta_S\varphi_{\eta 0} + i\eta\varphi_{\eta 0}\right]\right)\\ &\geq \left(\beta + \sum_{k=1}^m a_k\right)\|u\|^2_{L^2(S)} + \frac{1}{2}\frac{d}{dt}b(\varphi_{\eta 0},\varphi_{\eta 0}) + \eta b(\varphi_{\eta 0},\varphi_{\eta 0})\\ &\quad - \sum_{k=1}^m a_k d_k \|\varphi_{\eta k}\|_{L^2(S)}\|u\|_{L^2(S)}.\end{aligned} \tag{2.39}$$

Moreover, we may generalize (2.34) for all $C>0$ to

$$\frac{1}{2}\frac{d}{dt}\|\varphi_\eta\|^2_{L^2(S)} + \eta\|\varphi_\eta\|^2_{L^2(S)} \leq C\|\varphi_\eta\|^2_{L^2(S)} + \frac{1}{4C}\|u\|^2_{L^2(S)}.$$

Taking $C=\eta/2$, we obtain

$$\eta^2\int_0^T \|\varphi_\eta\|^2_{L^2(S)} \leq \int_0^T \|u\|^2_{L^2(S)}, \tag{2.40}$$

and if $\eta > d_k$ for all $k=1,\dots,m$, (2.39) and (2.40) yield (2.32). □

2.2.2. Existence and uniqueness results. We try to approximate the restriction of the solution u_{ex} of (2.1) to the open set Ω. We introduce the following problem:

$$\begin{cases} Lu = 0 & \text{in } \mathbb{R}^+\times\Omega,\\ \partial_{\nu_\Omega}u = Fu & \text{in } \mathbb{R}^+\times S,\\ u = u_0 & \text{at } t=0,\end{cases} \tag{2.41}$$

where F is defined by (2.27).

PROPOSITION 6. *Let g be in $H^1(]0,T[,L^2(\Omega))$ and u_0 be in $H^2(\Omega)$ such that*

$$\partial_{\nu_\Omega}u_0|_S = -\left(i\beta + ia_1 + \cdots + ia_m - \frac{d-1}{2}H_\Omega\right)u_0|_S. \tag{2.42}$$

There exists a unique solution u in the space $W^{1,\infty}(]0,T[,L^2(\Omega))\cap H^1(]0,T[,L^2(S))\cap L^\infty(]0,T[,H^1(\Omega))$ of the following variational problem:

$$\begin{cases}\forall v\in H^1(\Omega):\\ i\frac{d}{dt}(u,v) - (\nabla u,\nabla v) + [Fu,v] = (g,v) \text{ in } \mathcal{D}'(0,T),\end{cases} \tag{2.43}$$

where (u,v) denotes the Hermitian product in $L^2(\Omega)$.

Moreover, suppose g is in $H^2(]0,T[,L^2(\Omega))$ and Δu_0 and $g(0)$ are in $H^2(\Omega)$ and satisfy (2.42). Then u belongs to $W^{2,\infty}(]0,T[,L^2(\Omega))\cap W^{1,\infty}(]0,T[,H^1(\Omega))\cap L^\infty(]0,T[,H^2(\Omega))$ and satisfies the estimate

$$\begin{aligned}\|u\|^2_{L^\infty(]0,T[,H^2(\Omega))} &\leq C(\|u_0\|^2_{H^2(\Omega)} + \|\Delta u_0\|^2_{H^2(\Omega)} + \|g(0)\|^2_{H^2(\Omega)}\\ &\quad + \|\partial_t g(0)\|^2_{L^2(\Omega)} + \|g\|^2_{H^2(]0,T[,L^2(\Omega))}).\end{aligned} \tag{2.44}$$

Remark. If u_0 has compact support in Ω as in section 2.1.1, then u_0 satisfies (2.42).

Remark. The existence result of Proposition 6 with $g = 0$ implies the well-posedness of problem (2.41). The regularity result of Proposition 6 will be used to define and show the convergence of the iterative scheme in section 3.1.2

Proof. Let η be equal to $\max_{1\leq k\leq m}(d_k) + 1$; then $\eta > d_k$ for $k = 1, \dots, m$. Let $g_\eta = \exp(-\eta t)g$ and u_η be equal to $\exp(-\eta t)u$. u is the unique solution in $W^{1,\infty}(]0,T[, L^2(\Omega)) \cap H^1(]0,T[, L^2(S)) \cap L^\infty(]0,T[, H^1(\Omega))$ of (2.43) if and only if u_η is the unique solution in $W^{1,\infty}(]0,T[, L^2(\Omega)) \cap H^1(]0,T[, L^2(S)) \cap L^\infty(]0,T[, H^1(\Omega))$ of the following variational problem:

$$\begin{cases} \forall v \in H^1(\Omega) : \\ i\frac{d}{dt}(u_\eta, v) + i\eta(u_\eta, v) - (\nabla u_\eta, \nabla v) + [F_\eta u_\eta, v] = (g_\eta, v) \text{ in } \mathcal{D}'(0,T). \end{cases} \tag{2.45}$$

We will prove the existence and uniqueness of u_η in the space $W^{1,\infty}(]0,T[, L^2(\Omega)) \cap H^1(]0,T[, L^2(S)) \cap L^\infty(]0,T[, H^1(\Omega))$ solution of (2.45). Multiplying the equation by $\overline{u_\eta}$, integrating it in space, and taking imaginary part yields

$$\frac{d}{dt}\|u_\eta\|^2_{L^2(\Omega)} + 2\eta\|u_\eta\|^2_{L^2(\Omega)} + 2\mathrm{Im}\,([F_\eta u_\eta, u_\eta]) = 2\mathrm{Im}\,(g_\eta, u_\eta). \tag{2.46}$$

Multiplying the equation by $\overline{\partial_t u_\eta}$, integrating it in space, and taking the real part yields

$$\frac{d}{dt}\|\nabla u_\eta\|^2_{L^2(\Omega)} \leq 2\mathrm{Re}\,([F_\eta u_\eta, \partial_t u_\eta]) + \eta\|\partial_t u_\eta\|^2_{L^2(\Omega)} + \eta\|u_\eta\|^2_{L^2(\Omega)} - 2\mathrm{Re}\,(g_\eta, \partial_t u_\eta). \tag{2.47}$$

Differentiating the equation satisfied by u_η in time, multiplying it by $\overline{\partial_t u_\eta}$, integrating it in space, and taking the imaginary part yields

$$\frac{d}{dt}\|\partial_t u_\eta\|^2_{L^2(\Omega)} + 2\eta\|\partial_t u_\eta\|^2_{L^2(\Omega)} + 2\mathrm{Im}\,([F_\eta \partial_t u_\eta, \partial_t u_\eta]) = 2\mathrm{Im}\,(\partial_t g_\eta, \partial_t u_\eta). \tag{2.48}$$

By (2.32), (2.46), and Gronwall's lemma we obtain the following a priori estimate:

$$\|u_\eta\|^2_{L^\infty(]0,T[,L^2(\Omega))} + \|u_\eta\|^2_{L^2(]0,T[,L^2(S))} \leq C(\|u_0\|^2_{L^2(\Omega)} + \|g_\eta\|^2_{L^2(]0,T[,L^2(\Omega))}), \tag{2.49}$$

where $C = \max(1, (2(\beta + c_\eta))^{-1})$. In what follows, we do not make this dependence on η precise because η is fixed. By an inequality similar to (2.32), (2.48), and Gronwall's lemma we obtain the following a priori estimate:

$$\|\partial_t u_\eta\|^2_{L^\infty(]0,T[,L^2(\Omega))} + \|\partial_t u_\eta\|^2_{L^2(]0,T[,L^2(S))} \leq C(\|\partial_t u(0)\|^2_{L^2(\Omega)} + \|\partial_t g_\eta\|^2_{L^2(]0,T[,L^2(\Omega))}). \tag{2.50}$$

We shall find a bound on $\|\partial_t u_\eta(0)\|^2_{L^2(\Omega)}$. Multiplying the equation by $\overline{\partial_t u_\eta}$ and taking the imaginary part, we obtain after an integration by part in space

$$\begin{aligned} &\|\partial_t u_\eta\|^2_{L^2(\Omega)} + \eta\mathrm{Re}\,(u_\eta, \partial_t u_\eta) + \mathrm{Im}\,(\Delta u_\eta, \partial_t u_\eta) \\ &\qquad + \mathrm{Im}\,[F_\eta u_\eta - \partial_{\nu_\Omega} u_\eta, \partial_t u_\eta] = \mathrm{Re}\,(g_\eta, \partial_t u_\eta). \end{aligned} \tag{2.51}$$

By (2.42), $F_\eta u_\eta(0) - \partial_{\nu_\Omega} u_\eta(0) = (i\beta + i\sum_{k=1}^m a_k - (d-1)/2H_\Omega)u_0 - \partial_{\nu_\Omega} u_0 = 0$. Thus, taking $t = 0$ in (2.51),

$$\|\partial_t u_\eta(0)\|_{L^2(\Omega)} \leq \eta\|u_0\|_{L^2(\Omega)} + \|\Delta u_0\|_{L^2(\Omega)} + \|g_\eta(0)\|_{L^2(\Omega)}. \tag{2.52}$$

Remark. We do not use the equality $i\partial_t u_\eta(0) = g(0) - i\eta u_0 - \Delta u_0$ to give a bound on $\|\partial_t u_\eta(0)\|^2_{L^2(\Omega)}$. We rather use the variational formulation so that the estimates remain valid for approximate solutions of Galerkin's method. In fact, we may take initial conditions of these approximate solutions satisfying (2.42) as in [27]. Then, these approximate solutions satisfy (2.52).

Finally, (2.50) and (2.52) yield

$$\begin{aligned}\|\partial_t u_\eta\|^2_{L^\infty(]0,T[,L^2(\Omega))} + \|\partial_t u_\eta\|^2_{L^2(]0,T[,L^2(S))}\\ \le C(\|u_0\|^2_{H^2(\Omega)} + \|g_\eta(0)\|^2_{L^2(\Omega)} + \|\partial_t g_\eta\|^2_{L^2(]0,T[,L^2(\Omega))}).\end{aligned} \tag{2.53}$$

Equations (2.31) and (2.47) yield the following energy estimate:

$$\begin{aligned}\|\nabla u_\eta\|^2_{L^\infty(]0,T[,L^2(\Omega))} \le C(\|u_0\|^2_{H^1(\Omega)} + \|g_\eta\|^2_{L^2(]0,T[,L^2(\Omega))}\\ + \|u_\eta\|^2_{H^1(]0,T[,L^2(\Omega))} + \|u_\eta\|^2_{H^1(]0,T[,L^2(S))}.\end{aligned} \tag{2.54}$$

Thus (2.49) and (2.53) yield

$$\begin{aligned}\|\partial_t u_\eta\|^2_{L^\infty(]0,T[,L^2(\Omega))} + \|\partial_t u_\eta\|^2_{L^2(]0,T[,L^2(S))}\\ \le C(\|u_0\|^2_{H^2(\Omega)} + \|g_\eta(0)\|^2_{L^2(\Omega)} + \|\partial_t g_\eta\|^2_{L^2(]0,T[,L^2(\Omega))}).\end{aligned} \tag{2.55}$$

From the energy estimates (2.49), (2.53), and (2.55) we deduce by using Galerkin's method the existence and uniqueness of u_η in $W^{1,\infty}(]0,T[,L^2(\Omega))\cap H^1(]0,T[,L^2(S))\cap L^\infty(]0,T[,H^1(\Omega))$ solution of (2.45). Therefore, we have proved the existence and uniqueness of u in $W^{1,\infty}(]0,T[,L^2(\Omega))\cap H^1(]0,T[,L^2(S))\cap L^\infty(]0,T[,H^1(\Omega))$ solution of (2.43). Moreover, estimates (2.49), (2.53), and (2.55) remain valid for u (with different constants C).

Moreover, we suppose that g is in $H^2(]0,T[,L^2(\Omega))$ and Δu_0 and $g(0)$ are in $H^2(\Omega)$ and satisfy (2.42). The previous estimates are valid for $\partial_t u$ which is like u in $W^{1,\infty}(]0,T[,L^2(\Omega)) \cap H^1(]0,T[,L^2(S)) \cap L^\infty(]0,T[,H^1(\Omega))$. In particular, u is in $W^{2,\infty}(]0,T[,L^2(\Omega)) \cap W^{1,\infty}(]0,T[,H^1(\Omega))$ and we have a similar estimate to (2.55) for $\partial_t u$:

$$\begin{aligned}\|\nabla\partial_t u\|^2_{L^\infty(]0,T[,L^2(\Omega))} \le C(\|\Delta u_0\|^2_{H^2(\Omega)} + \|g(0)\|^2_{H^2(\Omega)}\\ + \|\partial_t g(0)\|^2_{L^2(\Omega)} + \|\partial_t g\|^2_{H^1(]0,T[,L^2(\Omega))}).\end{aligned} \tag{2.56}$$

It remains to show that u belongs to $L^\infty(]0,T[,H^2(\Omega))$. As g and $\partial_t u$ belong to $L^\infty(]0,T[,L^2(\Omega))$, Δu belongs also to $L^\infty(]0,T[,L^2(\Omega))$. Moreover, u belongs to $W^{1,\infty}(]0,T[,H^1(\Omega))$, and thus Fu belongs to $L^\infty(]0,T[,H^{\frac12}(S))$ according to Proposition 5. This proves that $\partial_{\nu_\Omega} u$ is in $L^\infty(]0,T[,H^{\frac12}(S))$. Standard results for elliptic boundary value problems (see, for example, [21]) imply that u is in $L^\infty(]0,T[,H^2(\Omega))$ and satisfies the following estimate:

$$\begin{aligned}\|u\|^2_{L^\infty(]0,T[,H^2(\Omega))} \le C\big(\|u\|^2_{L^\infty(]0,T[,L^2(\Omega))} + \|\Delta u\|^2_{L^\infty(]0,T[,L^2(\Omega))}\\ + \|\partial_{\nu_\Omega} u\|^2_{L^\infty(]0,T[,H^{\frac12}(S))}\big).\end{aligned} \tag{2.57}$$

Since $\partial_{\nu_\Omega} u = Fu$, Proposition 5 yields

$$\|\partial_{\nu_\Omega} u\|^2_{L^\infty(]0,T[,H^{\frac12}(S))} \le C\|u\|^2_{H^1(]0,T[,H^{\frac12}(S))} \le C\|u\|^2_{H^1(]0,T[,H^1(\Omega))}, \tag{2.58}$$

$$\|g\|_{L^\infty(]0,T[,L^2(\Omega))} \le \|g(0)\|_{L^2(\Omega)} + C\|\partial_t g\|_{L^2(]0,T[,L^2(\Omega))}.$$

Therefore, using $\Delta u = g - i\partial_t u$ we get

(2.59)
$$\|\Delta u\|_{L^\infty(]0,T[,L^2(\Omega))} \leq \|g(0)\|_{L^2(\Omega)} + C\|\partial_t g\|_{L^2(]0,T[,L^2(\Omega))} + \|\partial_t u\|_{L^\infty(]0,T[,L^2(\Omega))}.$$

(2.49), (2.53), (2.55), (2.56), (2.57), (2.58), and (2.59) imply (2.44). □

Remark. We can prove the half-space case in a similar way (see section 2.1.2). Moreover, we could prove a similar result replacing the operator $-\Delta$ by the elliptic operator $-\sum_{i,j=1}^{d} \partial_{x_i} a^{ij}(t,x)\partial_{x_j}$, where a^{ij} are in $C^1(\mathbb{R}\times\bar{\Omega})$ (as in [14] for parabolic operators). Finally, we may prove Schrödinger's equation with the absorbing boundary conditions of Di Menza [9] to be well posed in a similar way. (Di Menza gives a proof in a half-space through a Laplace transform.)

2.3. Numerical results.

2.3.1. The frame. We test our absorbing boundary conditions in the one-dimensional and two-dimensional cases.

In the one-dimensional case, we take $\Omega =]-5,5[$ as computational domain and $]-4.98,4.98[$ as domain of interest. The boundary S reduces then to two points. We compare the solution of (2.41) to the explicit solution of (2.1):

$$u_{ex}(t,x) = \frac{\exp(-i\pi/4)}{\sqrt{4t-i}} \exp\left(\frac{ix^2-6x-36t}{4t-i}\right).$$

This solution is a good test because it has almost compact support in $]-5,5[$ at $t=0$ (it remains under 10^{-10} on the boundary) and crosses the boundary $x=-5$ between $t=0$ and $t=1$. We take the time step $\delta t = 10^{-3}$ and the space step $h = 10^{-2}$. (δt is small compared to h to avoid numerical instabilities pointed out in [1].) We use three auxiliary functions ($m=3$), and the coefficients β, a_k, and d_k are given in section 2.3.2. We use a finite difference scheme with $N = 10/h$, $x_j = jh$ for $0 \leq j \leq N$, and $t_n = n\delta t$. u_j^n approximates $u(t_n,x_j)$, $\varphi_{k,1}^n$ approximates $\varphi_k(t_n,x_1)$, and $\varphi_{k,2}^n$ approximates $\varphi_k(t_n,x_{N-1})$. Inside the computational domain we use Crank–Nicolson's scheme

$$i\frac{u_j^{n+1}-u_j^n}{\delta t} + \frac{u_{j+1}^{n+1/2} - 2u_j^{n+1/2} + u_{j-1}^{n+1/2}}{h^2} = 0,$$

where $u_j^{n+1/2} = (u_j^{n+1}+u_j^n)/2$. The boundary conditions are approximated by

$$\frac{u_0^{n+1/2}-u_2^{n+1/2}}{2h} = i\beta u_1^{n+1/2} + i\sum_{k=1}^{m} a_k(u_1^{n+1/2} - d_k\varphi_{k,1}^{n+1/2}),$$

$$\frac{u_N^{n+1/2}-u_{N-2}^{n+1/2}}{2h} = i\beta u_{N-1}^{n+1/2} + i\sum_{k=1}^{m} a_k(u_{N-1}^{n+1/2} - d_k\varphi_{k,2}^{n+1/2}),$$

$$i\frac{\varphi_{k,1}^{n+1}-\varphi_{k,1}^n}{\delta t} + d_k\varphi_{k,1}^{n+1/2} = u_1^{n+1/2},\ k=1,\dots,m,$$

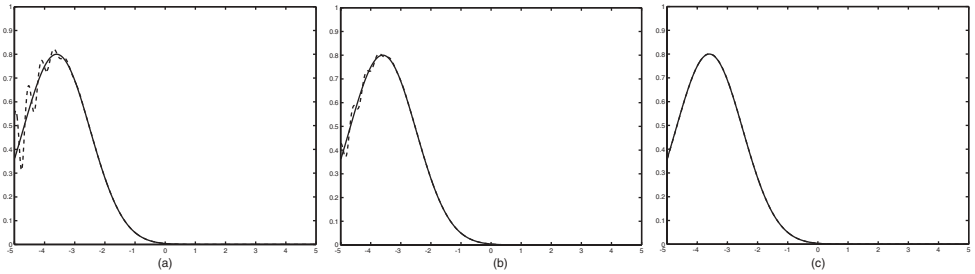


FIG. 1. *The exact solution* (*solid*) *and the approximate ones* (*dashed*) *at* $t = 0.3$, *one-dimensional case.* (*a*) *ABC in* [9], (*b*) *Padé,* (*c*) *present method.*

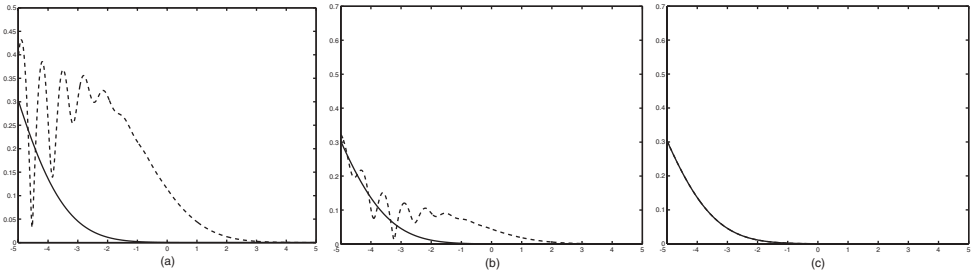


FIG. 2. *The exact solution* (*solid*) *and the approximate ones* (*dashed*) *at* $t = 0.6$, *one-dimensional case.* (*a*) *ABC in* [9], (*b*) *Padé,* (*c*) *present method.*

and

$$i\frac{\varphi_{k,2}^{n+1} - \varphi_{k,2}^{n}}{\delta t} + d_k \varphi_{k,2}^{n+1/2} = u_{N-1}^{n+1/2},\ k = 1, \ldots, m.$$

We give a snapshot of the solution at time $t = 0.3$ and $t = 0.6$ computed, respectively, with the ABC of Di Menza [9] with $m = 3$ (Figures 1(a) and 2(a)), the ABC with the coefficient of (2.23) and $m = 3$, which will be called Padé in what follows (Figures 1(b) and 2(b)), and our method with $m = 3$ (Figures 1(c) and 2(c)) and compared to the exact solution. We give the relative error in L^2 norm on $]-4.98, 4.98[$ as a function of time computed, respectively, with the ABC of Fevens and Jiang [13] with $p = 3$ (Figure 3(a)), the TBC of Arnold and Ehrhardt [5] (Figure 3(b)), and our method with $m = 3$ (Figure 4). We take this time $\delta t = 2.10^{-4}$ and $h = 2.10^{-3}$ (otherwise we compute the errors of the scheme instead of the errors due to our absorbing boundary conditions).

Remark. The TBC of Arnold and Ehrhardt [5] is the only boundary condition we tried that gives results of the same order as our method. However, we must store the boundary values at all previous times and do discrete convolution in time to implement the transparent boundary conditions of [5] (see (3.6)). It is much less costly to approximate the solution with our absorbing boundary conditions.

In the two-dimensional case, we take the disk with radius 1.1 as computational domain and the disk with radius 0.9 as domain of interest. We take the time step $\delta t = 10^{-3}$ and we mesh the two disks with Freefem+ [15] taking 200 points on the boundary of each disk. We use Crank–Nicolson's scheme for the time discretization and for the space discretization P1 finite elements based on the following weak formulation:

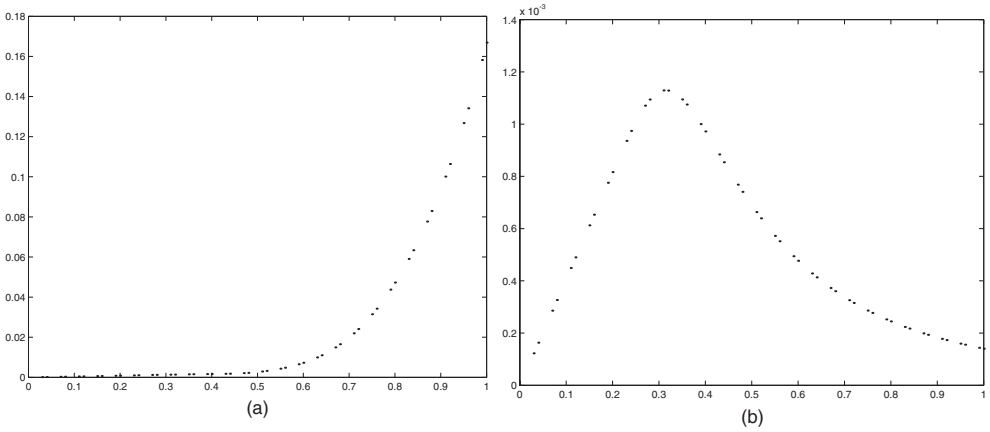


FIG. 3. *Linear case on* $]-4.98, 4.98[$. *Relative error in the* L^2 *norm as a function of time.* (*a*) *ABC in* [13], (*b*) *TBC in* [5].

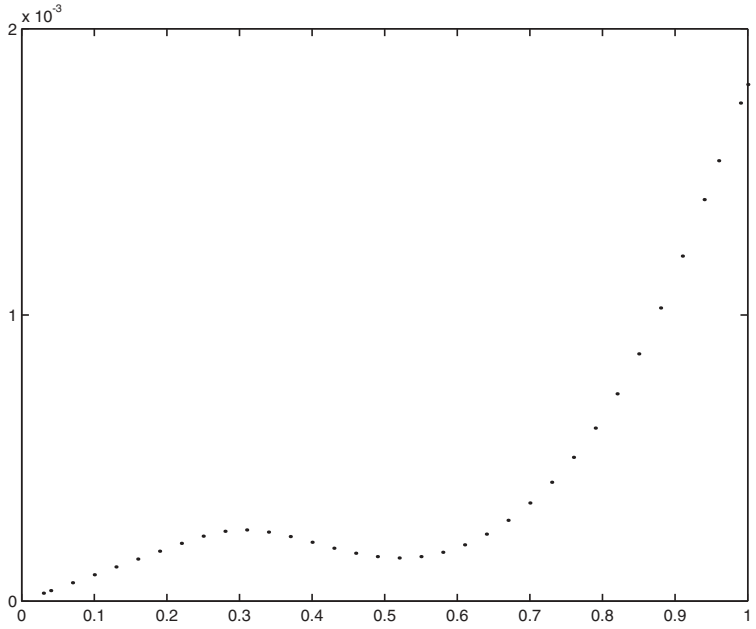


FIG. 4. *Linear case on* $]-4.98, 4.98[$. *Relative error in the* L^2 *norm as a function of time. Present method.*

$$i\left(\frac{u^{n+1}-u^n}{\delta t}, v\right) - (\nabla u^{n+1/2}, \nabla v) + i\beta[u^{n+1/2}, v] + i\sum_{k=1}^{m} a_k[u^{n+1/2} - d_k\varphi_k^{n+1/2}, v]$$
$$- \frac{1}{2}[Hu^{n+1/2}, v] + \left[\frac{II(D', D')}{2}u^{n+1/2}, v\right] = 0$$

and

$$i\left[\frac{\varphi_k^{n+1}-\varphi_k^n}{\delta t}, \theta_k\right] - [\nabla_S\varphi_k^{n+1/2}, \nabla_S\theta_k] + d_k[\varphi_k^{n+1/2}, \theta_k] = [u^{n+1/2}, \theta_k],\ k = 1, \ldots, m,$$

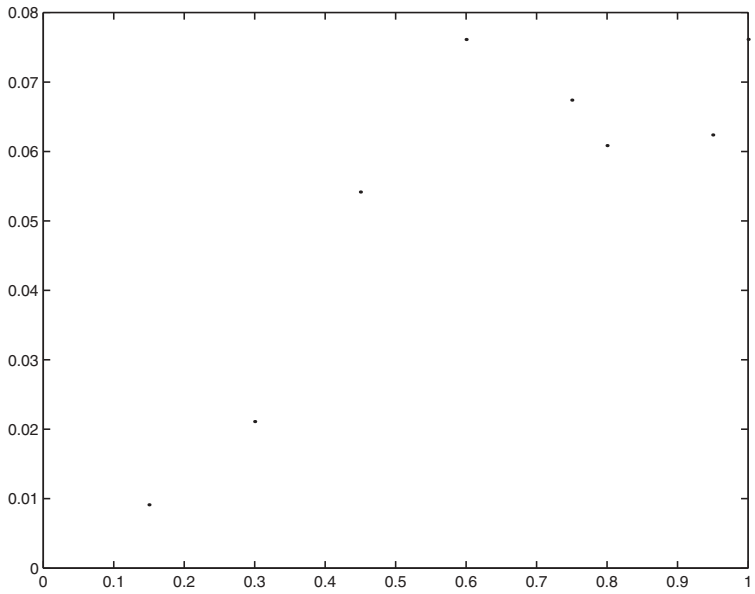


FIG. 5. *Linear case on the disk of radius* 0.9*, relative error in the* L^2 *norm as a function of time.*

TABLE 1
Maximum of the relative error in L^2 *norm for Padé ABC with increasing* m*.*

m	3	5	7	10	12	15
Padé	7.0737	1.9673	0.5647	0.0846	0.0200	0.0085

where $u^{n+1/2} = (u^{n+1} + u^n)/2$, $\varphi_k^{n+1/2} = (\varphi_k^{n+1} + \varphi_k^n)/2$, $k = 1, \ldots, m$, and where v and θ_k, $k = 1, \ldots, m$, are test functions. We compare the solution of (2.41) to the explicit solution of (2.1):

$$u_{ex}(t, x_1, x_2) = \frac{-i}{4t\gamma - i} \exp\left(\frac{i\gamma|x|^2 - \sqrt{\gamma}x_1 - \gamma t}{4t\gamma - i}\right),$$

where $\gamma = 24$ so that the initial condition has almost compact support in Ω (it remains under 10^{-10} on the boundary). This solution crosses the boundary between $t = 0$ and $t = 1$. Figure 5 gives the relative error in L^2 norm on the disk of radius 0.9 as a function of time. We take three auxiliary functions ($m = 3$), and the coefficients β, a_k, and d_k are given in section 2.3.2.

2.3.2. Optimization of the reflection coefficient. *The one-dimensional case.* We proved in section 2.1.2 the convergence of the problem with absorbing boundary conditions (2.18) to the problem (2.17). This happens for a special choice of the coefficients β_m, a_{km}, and d_{km} when letting m converge toward infinity. (See Table 1 for a numerical verification.) However, we are not interested in the case where m is large. In fact, one goal of this work is to reduce the numerical cost as much as possible. Thus we try for a given small m to optimize the absorbing boundary conditions.

Section 2.1.2 emphasized the role of the reflection coefficient. m being fixed, we optimize it for $\beta \geq 0$, $a_k > 0$, and $d_k > 0$, where $k = 1, \ldots, m$. In the one-dimensional case, the reflection coefficient is

$$R(\tau) = \frac{-\sqrt{\tau} - i\beta - i\sum_{k=1}^{m} \frac{a_k(-\tau)}{-\tau+d_k}}{\sqrt{\tau} - i\beta - i\sum_{k=1}^{m} \frac{a_k(-\tau)}{-\tau+d_k}}, \ \tau \in \mathbb{R}.$$

As in section 2.1.2, $|R| = 1$ when $\tau \geq 0$ and $|R| < 1$ when $\tau < 0$ for all $\beta \geq 0$, $a_k > 0$, and $d_k > 0$. We optimize the reflection coefficient in the region $\tau < 0$ (which corresponds to the hyperbolic region G_2 in section 2.1.2) and more precisely in $[-2\pi/\delta t, -2\pi/T]$ to stay in the range of the numerical frequencies. Moreover, we optimize in the L^2 norm with the weight $d\tau/(1+\tau^2)$. Because δt is small we optimize with the simplex method for $r = -1/\tau$ in $[0, T/2\pi]$ the integral

$$\int_0^{T/2\pi} \left| \frac{\sqrt{r} - \beta r - \sum_{k=1}^{m} \frac{a_k r}{1+d_k r}}{\sqrt{r} + \beta r + \sum_{k=1}^{m} \frac{a_k r}{1+d_k r}} \right|^2 \frac{dr}{1+r^2},$$

which is computed with the Weddle–Hardy method. Figure 4 shows the result obtained in the case $m = 3$. The error remains under 0.2% for times between 0 and 1, which is very satisfactory.

The coefficients that allowed us to obtain these results are $\beta = 7.269284.10^{-1}$, $a_1 = 2.142767$, $a_2 = 5.742223$, $a_3 = 4.658032.10^1$, $d_1 = 6.906263$, $d_2 = 6.582243.10^1$, and $d_3 = 1.124376.10^3$.

The two-dimensional case. In the case of the disk of radius $\kappa = 1.1$, the reflection coefficient is

$$R(\tau, \omega) = \frac{-\sqrt{\tau + (\frac{\omega}{\kappa})^2} - i\beta - i\sum_{k=1}^{m} \frac{a_k(-\tau-(\frac{\omega}{\kappa})^2)}{-\tau-(\frac{\omega}{\kappa})^2+d_k}}{\sqrt{\tau + (\frac{\omega}{\kappa})^2} - i\beta - i\sum_{k=1}^{m} \frac{a_k(-\tau-(\frac{\omega}{\kappa})^2)}{-\tau-(\frac{\omega}{\kappa})^2+d_k}}.$$

We optimize the reflection coefficient in the region $\tau + (\frac{\omega}{\kappa})^2 < 0$ (which corresponds to the hyperbolic region G_2 in section 2.1.2). More precisely, we optimize in L^2 norm on $s = -\tau - (\frac{\omega}{\kappa})^2$ in $[b, +\infty[$ with the weight $ds/(1+s^2)$ where $b > 0$. Because the interval is large we optimize for $r = 1/s$ the integral

$$\int_0^{1/b} \left| \frac{\sqrt{r} - \beta r - \sum_{k=1}^{m} \frac{a_k r}{1+d_k r}}{\sqrt{r} + \beta r + \sum_{k=1}^{m} \frac{a_k r}{1+d_k r}} \right|^2 \frac{dr}{1+r^2}.$$

We obtained our best results with $b = 0.5$. Figure 5 shows the results obtained in the case $m = 3$. The error remains under 8% for times between 0 and 1. (The solution computed with Dirichlet conditions on the circle of radius 1.1 gives 2000% error at time $t = 1$.)

Remark. We could use more terms in the asymptotic expansion of the Dirichlet-to-Neumann map $N_{\tilde{\Omega}}$ to improve these results. (We used only two terms in this study.)

The coefficients that allowed us to obtain those results are $\beta = 2.241274.10^{-1}$, $a_1 = 7.062909.10^{-1}$, $a_2 = 2.057994$, $a_3 = 1.705610.10^1$, $d_1 = 6.941414.10^{-1}$, $d_2 = 7.496463$, and $d_3 = 1.449250.10^2$.

3. The nonlinear problem with the absorbing boundary conditions of the linear problem. We wrote absorbing boundary conditions for the linear Schrödinger's equation. The numerical tests were convincing. In the case of the heat equation, we used absorbing boundary conditions of the linear problem for nonlinear ones and the results were also very good (see [27]). Similarly, we will try for nonlinear Schrödinger's equations the absorbing boundary conditions designed for the linear problem.

3.1. Existence results for nonlinear approximate problems.

3.1.1. The exact problem and the approximate problems. We fix an integer d so that $1 \leq d \leq 3$ and let f be a C^2 function from $\mathbb{C}$ to $\mathbb{C}$ (as a function from $\mathbb{R}^2$ to $\mathbb{R}^2$). The operator $-\Delta$ is the generator of a semigroup of contraction on $H^2(\mathbb{R}^d)$ and f is locally Lipschitzian from $H^2(\mathbb{R}^d)$ into itself. Thus (see [24]), for all u_0 in $H^2(\mathbb{R}^d)$, there is a time $T > 0$ and a unique solution u_{ex} in $C([0,T], H^2(\mathbb{R}^d))$ of the problem

$$\begin{cases} Lu_{ex} = f(u_{ex}),\ ]0,T[\times \mathbb{R}^d, \\ u_{ex} = u_0 \ \text{ at } \ t = 0, \end{cases} \tag{3.1}$$

We want to approximate the restriction of u_{ex} to the open set Ω. To generalize problem (2.16), we introduce the following problem:

$$\begin{cases} Lu = f(u) & \text{in } \mathbb{R}^+ \times \Omega, \\ \partial_{\nu_\Omega} u = Fu & \text{in } \mathbb{R}^+ \times S, \\ u = u_0 & \text{at } t = 0, \end{cases} \tag{3.2}$$

where F is an operator equal to the right-hand side of (2.27) in what follows.

We define the solution of (3.2) via the following variational formulation:

$$\begin{cases} u \in L^\infty(]0,T[, L^2(\Omega)) \cap L^2(]0,T[, H^1(\Omega)), \quad f(u) \in L^\infty(]0,T[, L^2(\Omega)), \\ \forall v \in H^1(\Omega), \\ i\frac{d}{dt}(u,v) - (\nabla u, \nabla v) + [Fu, v] = (f(u), v) \ \text{ in } \ \mathcal{D}'(0,T). \end{cases} \tag{3.3}$$

THEOREM 1. *Suppose Ω is convex. For any u_0 in $H^4(\Omega)$ with compact support in Ω, there exists a time $T > 0$ such that problem (3.3) has a unique solution u in $W^{2,\infty}(]0,T[, L^2(\Omega)) \cap W^{1,\infty}(]0,T[, H^1(\Omega)) \cap L^\infty(]0,T[, H^2(\Omega))$.*

Remark. We have already used the compactness of the support of u_0 in Ω to obtain our absorbing boundary conditions (see 2.1.1).

3.1.2. The iterative scheme. First, we define u_{l+1} from u_l, as the solution of the linear problem with the right-hand side $f(u_l)$. Then we prove that for sufficiently small $T > 0$, (u_l) is bounded in the space $W^{2,\infty}(]0,T[, L^2(\Omega)) \cap W^{1,\infty}(]0,T[, H^1(\Omega)) \cap L^\infty(]0,T[, H^2(\Omega))$. Finally, we show that (u_l) is a Cauchy sequence in $C^0(0,T; L^2(\Omega))$ and that the limit u satisfies the nonlinear problem.

To define (u_l), we need the following lemma.

LEMMA 4. *Let v belong to the space $L^\infty(]0,T[, H^2(\Omega)) \cap W^{1,\infty}(]0,T[, H^1(\Omega)) \cap W^{2,\infty}(]0,T[, L^2(\Omega))$, and $g = f(v)$. Then g belongs to $H^2(]0,T[, L^2(\Omega))$ and satisfies the inequality*

$$\|g\|^2_{H^2(]0,T[,L^2(\Omega))} \leq T\theta(\|v\|^2_{L^\infty(]0,T[,H^2(\Omega))}+\|v\|^2_{W^{1,\infty}(]0,T[,H^1(\Omega))}+\|v\|^2_{W^{2,\infty}(]0,T[,L^2(\Omega))}),$$

where θ is a continuous increasing function.

The proof of this lemma relies on the injection of $H^2(\Omega)$ in $L^\infty(\Omega)$, the injection of $H^1(\Omega)$ in $L^4(\Omega)$, and the fact that f is C^2.

Proposition 6, Lemma 4, and Gronwall's lemma yield the following.

COROLLARY 1. *Let v belong to $L^\infty(]0,T[,H^2(\Omega)) \cap W^{1,\infty}(]0,T[,H^1(\Omega)) \cap W^{2,\infty}(]0,T[,L^2(\Omega))$ such that $v(0) = u_0$ and $i\partial_t v(0) = -\Delta u_0 + f(u_0)$, where u_0 is in $H^4(\Omega)$ with compact support in Ω. Let w be the unique solution of Proposition 6 with $g = f(v)$. Then w satisfies following estimate:*

$$\begin{aligned}\|w\|^2_{W^{2,\infty}(]0,T[,L^2(\Omega))} + \|w\|^2_{W^{1,\infty}(]0,T[,H^1(\Omega))} + \|w\|^2_{L^\infty(]0,T[,H^2(\Omega))} \leq e^{CT}(\theta(\|u_0\|_{H^4(\Omega)})\\ + T\theta(\|v\|^2_{W^{2,\infty}(]0,T[,L^2(\Omega))} + \|v\|^2_{W^{1,\infty}(]0,T[,H^1(\Omega))}\\ + \|v\|^2_{L^\infty(]0,T[,H^2(\Omega))})),\end{aligned}$$

where C depends only on Ω, θ and the function f.

Remark. We suppose that u_0 has compact support which is stronger than (2.42). In fact, (2.42) is not preserved under the action of a nonlinear function f.

Now we set $C_1 = \max(\|u_0\|_{H^2(\Omega)}, \theta(\|u_0\|_{H^4(\Omega)}))$ and $T_1 > 0$ so that we have $e^{CT_1}(C_1 + T_1\theta(2C_1)) \leq 2C_1$. Set $C_2 = \sup_{|s|\leq 4C_1} |f'(s)|$ and $T_2 > 0$ so that $CC_2T_2 \leq \frac{1}{2}$.

DEFINITION 1. *Let $T > 0$ be such that $T \leq \max(T_1, T_2)$. We define the sequence (u^l) in $W^{2,\infty}(]0,T[,L^2(\Omega)) \cap W^{1,\infty}(]0,T[,H^1(\Omega)) \cap L^\infty(]0,T[,H^2(\Omega))$ by induction. u^0 is constant in time and is equal to u_0 on $[0,T]$. Suppose u^l is defined in $W^{2,\infty}(]0,T[,L^2(\Omega)) \cap W^{1,\infty}(]0,T[,H^1(\Omega)) \cap L^\infty(]0,T[,H^2(\Omega))$; then $f(u^l)$ is in $H^2(]0,T[,L^2(\Omega))$ by Lemma 4: we may thus define u^{l+1} as the unique solution of problem (2.43) with the right-hand side $f(u^l)$.*

Corollary 1 and the energy estimates of Proposition 6 yield the following.

LEMMA 5. *The sequence (u^l) is bounded in the space $W^{2,\infty}(]0,T[,L^2(\Omega)) \cap W^{1,\infty}(]0,T[,H^1(\Omega)) \cap L^\infty(]0,T[,H^2(\Omega))$ and is a Cauchy sequence in $C([0,T],L^2(\Omega))$.*

The end of the proof of the existence of u is classical (see, for example, [20]). The limit u of the Cauchy sequence (u^l) satisfies the variational formulation (3.3) by using the strong convergence in $C([0,T],L^2(\Omega))$ and the boundedness in $L^\infty(]0,T[,H^2(\Omega))$ for the nonlinear term. The uniqueness follows from the energy estimate (2.49) and the fact that u belongs to $L^\infty(]0,T[,H^2(\Omega))$.

3.2. Numerical results. We test our absorbing boundary conditions in the one-dimensional case. In fact, we will see that the numerical results are poor in this case. It is therefore useless to implement higher-dimensional cases.

3.2.1. The frame. We take $\Omega =]-5,5[$ as computational domain and $]-4.98,4.98[$ as domain of interest. We take the time step $\delta t = 10^{-3}$ and the space step $h = 10^{-2}$. We compare the solution of (3.2) to the soliton solution of (3.1) with $f(u) = -|u|^2u$:

$$u_{ex}(t,x) = \sqrt{2a}\,\mathrm{sech}(\sqrt{a}(x-ct))\exp\left(i\frac{c}{2}(x-ct)+i\theta_0\right)\exp\left(i\left(a+\frac{c^2}{4}\right)t\right),$$

where $a = 27$, $c = 15$, and $\theta_0 = \pi/4$. This solution is a good test because it has almost compact support in $]-5,5[$ at $t = 0$ (it remains under 10^{-10} on the boundary), and it crosses the boundary $x = 5$ between $t = 0$ and $t = 1$.

We use Durán and Sanz-Serna's scheme [11], where the semidiscrete form is

$$\frac{U^{n+1}-U^n}{\delta t} = i\partial_{xx}^2 U^{n+\frac{1}{2}} + if(U^{n+\frac{1}{2}}), \tag{3.4}$$

where $U^{n+1/2} = (U^{n+1}+U^n)/2$ and $n = 0,\dots,T/\delta t - 1$. We solve the nonlinear system with a fixed-point method giving $U^{n+1/2}$:

$$Z = \left(1 - i\frac{\delta t}{2}\partial_{xx}^2\right)^{-1}\left(U^n + i\frac{\delta t}{2}f(Z)\right).$$

We initialize the fixed-point method with U^n. Then $U^{n+1} = 2Z - U^n$.

3.2.2. The discrete transparent operator for the linear case. We approximate (3.2) with the following problem:

$$\begin{cases} Lu = f(u) & \text{in } \mathbb{R}^+ \times \Omega, \\ \partial_{\nu_\Omega} u = Tu & \text{in } \mathbb{R}^+ \times S, \\ u = u_0 & \text{at } t = 0, \end{cases} \tag{3.5}$$

where T is the transparent operator for the linear case (the Dirichlet-to-Neumann map of section 2.1.1). In [5], Arnold and Ehrhardt discretize T (here at $x = -5$) as

$$u_1^{n+1} - s^0 u_0^{n+1} = \sum_{l=1}^{n} s^{n+1-k} u_0^k - u_1^n, \tag{3.6}$$

where $(s^k)_{k\geq 0}$ is computed through a recurrence formula (see Theorem 3.8 of [5]). We implement the solution of (3.5) with the scheme (3.4).

Figure 6 shows that (3.2) approximates (3.5), but not very well. In fact, the error reaches 37% at time 1. In any case, Figure 7 shows that (3.5) does not approximate (3.1) at all. In fact, the soliton leaves Ω, which is not the case of the solution of (3.5). Therefore, the use of any choice of absorbing boundary conditions of the linearized problem does not allow one to approximate the solution of the nonlinear Schrödinger's equation.

Remark. This strategy (to approximate the nonlinear equation with the absorbing boundary conditions of the linearized problem) is efficient in the case of reaction-diffusion equations (see [27]).

Remark. In [7], the authors show that the solution of (3.1) and of (3.5) share some qualitative properties (blow-up, filamentation, etc.). However, Figure 7 shows that the two solutions are quantitatively very different.

4. Conclusion. Our absorbing boundary conditions give an accurate method with which to approximate the solution of Schrödinger's equation on the whole space with a low numerical cost. In the case of curved boundaries, we could still improve our results by computing more terms in the asymptotic expansion of the Dirichlet-to-Neumann map. (we used only two terms in this study.)

Regarding the nonlinear problem, we have proved an existence and uniqueness result for the problem with nonclassical boundary conditions, which is a new result. Our absorbing boundary conditions for the linearized problem allow us to approximate the problem with the transparent boundary conditions for the linearized problem as shown by our numerical computations. However, the nonlinear problem with the transparent boundary conditions of the linearized problem does not approximate the nonlinear problem on the whole space. We should develop a purely nonlinear strategy.

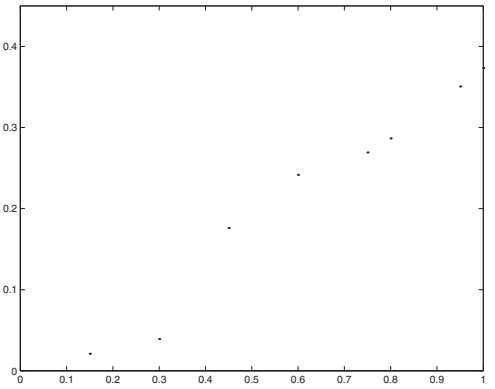


FIG. 6. *nonlinear case: comparison tbc and abc.*

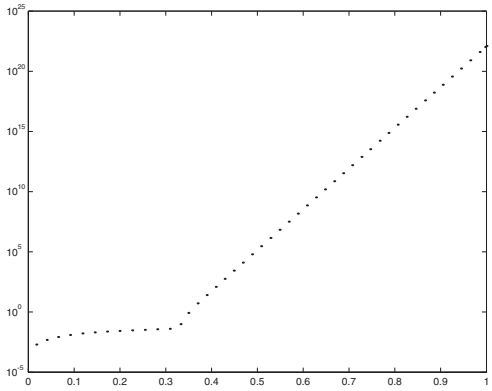


FIG. 7. *nonlinear case: comparison tbc and exact solution.*

Acknowledgments. The author thanks S. Descombes for discussions about the numerical aspects and thanks his Ph.D. advisor L. Halpern.

REFERENCES

[1] I. ALONSO-MALLO AND N. REGUERA, *Weak ill-posedness of spatial discretizations of absorbing boundary conditions for Schrödinger-type equations*, SIAM J. Numer. Anal., 40 (2002), pp. 134–158.

[2] X. ANTOINE AND C. BESSE, *Construction, structure and asymptotic approximations of a microdifferential transparent boundary condition for the linear Schrödinger equation*, J. Math. Pures Appl., (9) 80 (2001), pp. 701–738.

[3] X. ANTOINE AND C. BESSE, *Unconditionally stable discretization schemes of non-reflecting boundary conditions for the one-dimensional Schrödinger equation*, J. Comput. Phys., 188 (2003), pp. 157–175.

[4] A. ARNOLD, *Numerically absorbing boundary conditions for quantum evolution equations*, VSLI Design, 6 (1998), pp. 313–319.

[5] A. ARNOLD AND M. EHRHARDT, *Discrete transparent boundary conditions for the Schrödinger equation*, Rivi. Math. Univ. Parma, 6 (2001), pp. 57–108.

[6] V. A. BASKAKOV AND A. V. POPOV, *Implementation of transparent boundaries for numerical solution of the Schrödinger equation*, Wave Motion, 14 (1991), pp. 123–128.

[7] C. H. BRUNEAU, L. DI MENZA, AND T. LEHNER, *Numerical resolution of some nonlinear Schrödinger-like equations in plasmas*, Numer. Methods Partial Differential Equations, (1999), pp. 672–696.

[8] P. DEUFLHARD AND F. SCHMIDT, *Discrete transparent boundary conditions for the numerical solution of Fresnel's equation*, Comput. Math. Appl., 29 (1995), pp. 53–76.

[9] L. DI MENZA, *Transparent and absorbing boundary conditions for the Schrödinger equation in a bounded domain*, Numer. Funct. Anal. Optim., 18 (1997), pp. 759–775.

[10] E. DUBACH, *Artificial boundary conditions for diffusion equations: Numerical study*, J. Comput. Appl. Math., 70 (1996), pp. 127–144.

[11] A. DURÁN AND J. M. SANZ-SERNA, *The numerical integration of relative equilibrium solutions. The nonlinear Schrödinger equation*, IMA J. Numer. Anal., 20 (2000), pp. 235–261.

[12] B. ENGQUIST AND A. MAJDA, *Radiation boundary conditions for acoustic and elastic wave calculations*, Comm. Pure Appl. Math., 32 (1979), pp. 313–357.

[13] T. FEVENS AND H. JIANG, *Absorbing boundary conditions for the Schrödinger equation*, SIAM J. Sci. Comput., 21 (1999), pp. 255–282.

[14] L. HALPERN AND J. RAUCH, *Absorbing boundary conditions for diffusion equations*, Numer. Math., 71 (1995), pp. 185–224.

[15] F. HECHT AND O. PIRONNEAU, *Multiple Unstructured Meshes and the Design of freefem+*, Tech. report, INRIA, France, 1999.

[16] P. Joly, *Pseudotransparent boundary conditions for the diffusion equation. Part* I, Math. Methods Appl. Sci., 11 (1989), pp. 725–758.
[17] J. P. Kuska, *Absorbing boundary conditions for the Schrödinger equation on finite intervals*, Phys. Rev. B, 46 (1992), pp. 5000–5003.
[18] R. Lascar, *Propagation des singularités des solutions d'équations pseudodifférentielles quasi-homogènes*, Ann. Inst. Fourier (Grenoble), 27 (1977), pp. 79–123.
[19] E. L. Lindmann, *Free-space boundary conditions for the time dependent wave equation*, J. Comput. Phys., 18 (1985), pp. 16–78.
[20] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.
[21] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Dunod, Paris, 1968.
[22] C. Lubich and A. Schädle, *Fast convolution for nonreflecting boundary conditions,* SIAM J. Sci. Comput., 24 (2002), pp. 161–182.
[23] L. Nirenberg, *Lectures on Linear Partial Differential Equations*, CBMS Reg. Conf. Ser. Math. 17, AMS, Providence, RI, 1973.
[24] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
[25] F. Schmidt and D. Yevick, *Discrete transparent boundary conditions for Schrödinger-type equations*, J. Comput. Phys., 134 (1997), pp. 96–107.
[26] T. Shibata, *Absorbing boundary conditions for the finite-difference time-domain calculation of the one-dimensional Schrödinger equation*, Phys. Rev. B, 43 (1991), pp. 6760–6763.
[27] J. Szeftel *Absorbing boundary conditions for reaction diffusion equation*, IMA J. Appl. Math., 68 (2003), pp. 167–184.
[28] J. Szeftel, *Réflexion des singularités pour l'équation de Schrödinger*, Comm. Partial Differential Equations, 29 (2004), pp. 707–761.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1552–1571

HOMOTOPIES FOR INTERSECTING SOLUTION COMPONENTS OF POLYNOMIAL SYSTEMS*

ANDREW J. SOMMESE†, JAN VERSCHELDE‡, AND CHARLES W. WAMPLER§

Abstract. We show how to use numerical continuation to compute the intersection $C = A \cap B$ of two algebraic sets A and B, where A, B, and C are numerically represented by witness sets. En route to this result, we first show how to find the irreducible decomposition of a system of polynomials restricted to an algebraic set. The intersection of components A and B then follows by considering the decomposition of the diagonal system of equations $u - v = 0$ restricted to $\{u, v\} \in A \times B$. An offshoot of this new approach is that one can solve a large system of equations by finding the solution components of its subsystems and then intersecting these. It also allows one to find the intersection of two components of the two polynomial systems, which is not possible with any previous numerical continuation approach.

Key words. components of solutions, embedding, generic points, homotopy continuation, irreducible components, numerical algebraic geometry, polynomial system

AMS subject classifications. Primary, 65H10; Secondary, 13P05, 14Q99, 68W30

DOI. 10.1137/S0036142903430463

1. Introduction.

In a series of papers [12, 13, 14, 15, 16, 17], we proposed numerical continuation algorithms that use witness sets as the basic construct for representing solution components of a system of polynomial equations on $\mathbb{C}^N$. Witness sets are the central concept of a young subject that we call numerical algebraic geometry, which uses numerical continuation [1, 2] and generalizes earlier work in computing isolated solutions of polynomial systems [8, 9]. The main concern of this paper is to provide an algorithm for computing the intersection of two solution components A, B from two possibly identical polynomial systems f, g, whose witness sets have been given. It is important to realize that naively combining f, g into one system $h = \{f, g\}$ is not sufficient, even if we were willing to put aside the potentially prohibitive size of the combined system. For example, suppose A is the line $x_2 = 0$ as a solution component of $f(x) = x_1 x_2$ and B is the line $x_1 - x_2 = 0$ as a solution component of $g(x) = x_1(x_1 - x_2)$. Then, $A \cap B$, which is the isolated point $(0, 0)$, does not appear as an irreducible component of the system $h = \{f, g\}$.

Questions involving intersection of components arise naturally in applications. Just as a single polynomial in one variable has multiple roots, a system of polynomial equations in several variables can have multiple solution components; these compo-

*Received by the editors June 25, 2003; accepted for publication (in revised form) January 28, 2004; published electronically December 16, 2004. This research was supported by Land Baden-Württemberg (Research in Pairs Program at the Mathematical Research Institute in Oberwolfach, Germany).

http://www.siam.org/journals/sinum/42-4/43046.html

†Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556-4618 (sommese@nd.edu, http://www.nd.edu/~sommese). This material is based on work supported by the National Science Foundation under grant 0105653 and the Duncan Chair of the University of Notre Dame.

‡Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 South Morgan (M/C 249), Chicago, IL 60607-7045 (jan@math.uic.edu, jan.verschelde@na-net.ornl.gov, http://www.math.uic.edu/~jan). This material is based on work supported by the National Science Foundation under grants 0105739 and 0134611.

§General Motors Research and Development, Mail Code 480-106-359, 30500 Mound Road, Warren, MI 48090-9055 (Charles.W.Wampler@gm.com).

nents can even appear at different dimensions (points, curves, surfaces, etc.) from the same set of equations. We may wish to find the intersection of just one of those components with another algebraic set. In our new approach, only the degrees of the components being intersected come into play in the determination of the number of paths followed by the homotopies that we use. This is important since the degree of a component of a given system of polynomials is typically much less than the number of paths required to find even all isolated solutions of the given system.

Viewed another way, the intersection operation is required for a Boolean algebra of constructible algebraic sets; a complete Boolean algebra also requires the operations of union and complement. Suppose W is a witness set for a component X. There are several probability-one algorithms for deciding if a point $x \in \mathbb{C}^N$ is a member of X, using numerical continuation and the data in W. (We review witness sets and membership tests in section 2.) The complement operation is just the logical inversion of a membership test, and the union operation is just a union of witness sets, utilizing membership tests to eliminate duplications. However, the operation of intersection is more difficult.

In our previous work, we showed how to find the solution set of a system of polynomial equations as a union of witness sets, and, further, we showed how to decompose these into witness sets for the irreducible components. Said another way, this solves the problem of intersecting a collection of hypersurfaces defined by polynomial equations. But this does not give us an effective means of computing the intersection of two components represented by witness sets.

Our first step in creating an algorithm for the intersection of components is to generalize an earlier algorithm for generating the witness sets for the solution set of a system of polynomial equations on $\mathbb{C}^N$. The generalization instead considers the polynomial equations restricted to an algebraic set. The intersection of components A and B then follows by considering the decomposition of the diagonal system of equations $u - v = 0$ restricted to $\{u, v\} \in A \times B$. Hence, we call the intersection algorithm the diagonal homotopy.

This paper is organized as follows. First, in section 2, we review the definition of a witness set and its role in finding the numerical irreducible decomposition of the solution set of a system of polynomial equations. In section 3 we introduce a slight generalization of the randomization procedure of [17], and in section 4 we give a general construction of homotopies. These sections give the basic definitions and results that will be needed later in the article.

The original algorithm for constructing witness supersets was given in [17]. A much more efficient algorithm for constructing witness supersets was given in [12] by means of an embedding theorem. In section 5, we show how to carry out the generalization of [12] to the case of a system of polynomials on a pure N-dimensional algebraic set $X \subset \mathbb{C}^m$, i.e., an algebraic subset of $\mathbb{C}^m$ all of whose irreducible components are N-dimensional. We call this the abstract embedding theorem because it does not rely on any specific numerical description of X. In this generality we lose some control of multiplicities. However, since our main objective is to find the underlying reduced algebraic solution components, this loss of multiplicity information is of minor importance.

In section 6 we show how to implement the abstract embedding theorem numerically. We need only the information about X that would be produced by the algorithm for the numerical irreducible decomposition of a polynomial system f, for which X is an irreducible component of the solution set of f.

In section 7 we specialize to the situation in which we have two polynomial systems

f and g on $\mathbb{C}^N$ and we wish to describe the irreducible decompositions of $A \cap B$ where A is an irreducible component of $V(f)$ and B is an irreducible component of $V(g)$. Computational experiments are discussed in section 8.

In Appendix A, we give some further discussion of the method of constructing homotopies described in section 4.

In Appendix B, we give the proof of Theorem 5.1.

2. Witness sets. We begin by reviewing the basics of numerical algebraic geometry, wherein the most fundamental concept is a witness set. See [13, 15, 17] for details on irreducible components, the irreducible decomposition, and reduced algebraic sets.

Given a system of polynomials on $\mathbb{C}^N$

$$f(x) := \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}, \tag{2.1}$$

we denote the underlying point set $\{x \in \mathbb{C}^N \mid f(x) = 0\}$ by $V(f)$, i.e., the algebraic set $f^{-1}(0)$ (with all multiplicity information that comes with $f^{-1}(0)$ ignored). A pure i-dimensional algebraic subset X of $V(f)$ is a subset $X \subset V(f)$ equal to the closure of a union of i-dimensional connected components of the smooth points of $V(f)$. We emphasize that X is reduced, i.e., that we are ignoring the multiplicity of X within $f^{-1}(0)$. We represent X numerically by a witness set, defined as follows.

DEFINITION 2.1. *A witness set for a pure i-dimensional algebraic set $X \subset V(f) \in \mathbb{C}^N$ consists of*

1. *the dimension, i, of X;*
2. *the polynomial system $f(x)$;*
3. *a general $(N-i)$-dimensional affine linear subspace $L_{N-i} \subset \mathbb{C}^N$; and*
4. *the set of $\deg X$ distinct points $\mathcal{X} = L_{N-i} \cap X$.*

In other words, a witness set W for X is the ordered set $W = \{i, f, L_{N-i}, \mathcal{X}\}$. We use the notation $V(W)$ to denote the component represented by W; in the current context $V(W) = X$.

This definition is useful because it allows us to numerically represent and manipulate the irreducible decomposition of the solution set of a polynomial system. Let us quickly review that concept before describing our new results. Everything we say is over the complex numbers, e.g., even if the polynomials have real coefficients, we always deal with the sets of solutions on complex Euclidean space.

We start with a system of polynomials f on $\mathbb{C}^N$ as in (2.1). Let $V(f)$ denote the set of solutions of f on $\mathbb{C}^N$, i.e., the set of points $x \in \mathbb{C}^N$ such that $f(x) = 0$. The set $Z := V(f)$ is an affine algebraic set and decomposes into a union of distinct irreducible components. Recall that an algebraic set X is irreducible if and only if the Zariski open dense subset of manifold points on X is connected. We have the decomposition

$$Z = \bigcup_{i=1}^{\dim Z} Z_i = \bigcup_{i=1}^{\dim Z} \left(\bigcup_{j \in \mathcal{I}_i} Z_{i,j} \right), \tag{2.2}$$

where

1. for each i, $Z_i := (\cup_{j \in \mathcal{I}_i} Z_{i,j})$;
2. the sets $\mathcal{I}_i$ are finite and each $Z_{i,j}$ is irreducible of dimension i; and

3. $Z_{i,j}$ is not contained in a union of a collection of the $Z_{a,b}$ unless $Z_{i,j}$ occurs in the collection.

Any collection of irreducible components $Z_{i,j}$ having the same dimension, i, can be numerically represented by a witness set. A numerical irreducible decomposition is a list having one witness set $W_{i,j}$ for the reduction of each irreducible component $Z_{i,j}$.

In a series of papers [12, 13, 14, 15, 16, 17], we showed how to compute a numerical irreducible decomposition of $Z := V(f)$. The approach is to intertwine two numerical algorithms: a witness generating algorithm, which finds a superset of witness points for each pure-dimensional algebraic set Z_i, and a decomposition algorithm, which eliminates spurious points from the superset and breaks it into irreducible components. To be more precise, at each dimension $i = 0, \ldots, \dim Z$, the witness generating algorithm gives a finite set of points $\widehat{W}_i$ satisfying $L_{N-i} \cap Z_i \subset \widehat{W}_i \subset L_{N-i} \cap (\cup_{j \geq i} Z_j)$, where $L_{N-i} \subset \mathbb{C}^N$ is a general $(N-i)$-dimensional affine linear subspace. The second algorithm decomposes the $\widehat{W}_i$. Precisely,

1. $\widehat{W}_i$ decomposes into the disjoint union

$$J_i \cup \left(\bigcup_{j \in \mathcal{I}_i} \mathcal{Z}_{i,j} \right), \tag{2.3}$$

where $J_i \subset \cup_{k>i} Z_k$ and $\mathcal{Z}_{i,j}$ consists of the $\deg Z_{i,j}$ points of $L_{N-i} \cap Z_{i,j}$; and

2. $J_{\dim Z} = \emptyset$.

The points $\mathcal{Z}_{i,j}$, along with the dimension i, the system of equations f, and the linear subspace L_{N-i}, form a witness set for the irreducible component $Z_{i,j}$.

The key theoretical advance of this paper is to observe that the previous algorithms for numerical irreducible decomposition still work with restrictions of a polynomial system to a pure-dimensional algebraic set. Only the first algorithm constructing the witness point supersets $\widehat{W}_i$ needs to be generalized. The decomposition algorithms starting with the witness point supersets $\widehat{W}_i$ are proved in [13, 14, 15] in sufficient generality to cover the present situation.

The above implicitly assumes that the components $Z_{i,j}$ are reduced, i.e., of multiplicity one in $f^{-1}(0)$. The algorithms in [12, 17] in fact produce sets $\mathcal{Z}_{i,j}$ consisting of $\deg Z_{i,j}$ distinct points each repeated $\mu_{i,j}$ times, where $\mu_{i,j}$ is greater than or equal to the multiplicity of $Z_{i,j}$ in $f^{-1}(0)$. Moreover the multiplicity of $Z_{i,j}$ is one if and only if $\mu_{i,j} = 1$. Unfortunately, in the algorithm in this article we can assert only that $\mu_{i,j} > 0$ for any irreducible component $Z_{i,j}$.

As mentioned in the introduction, an important aspect of a witness set $\mathcal{X}$ is that we can use it to test a point for membership in the algebraic set $X = V(\mathcal{X})$ that $\mathcal{X}$ represents. This stems from the fact that we can sample X by continuously perturbing the linear slice and numerically tracking its intersection with X, starting from the witness points in $\mathcal{X}$. Several different membership tests can be employed. At one expensive extreme, by sampling and fitting, we might compute a set of polynomials, whose set of common zeros is exactly X. A much more efficient probability-one test for a point $x \in \mathbb{C}^N$ to be in X is whether the pullback from $\mathbb{C}^{\dim X+1}$ of a $\deg X$ defining polynomial for $\pi(X) \subset \mathbb{C}^{\dim X+1}$ is zero on x, where π is a general linear projection from $\mathbb{C}^N$ to $\mathbb{C}^{\dim X+1}$. Finally, a very different sort, and quite efficient test, for $x \in \mathbb{C}^N$ to be in X is to see whether x is one of the images of the set $\mathcal{X}$ under the homotopy taking L_{N-i} to a general $(N-i)$-dimensional linear subspace of $\mathbb{C}^N$ that contains x. This test depends on the real-one-dimensional path

between the general linear subspaces to remain general, which occurs with probability one.

3. Randomizing systems. Randomization is a key element of our approach. This section introduces some notation for randomized systems and gives a lemma describing their most important properties. Given a system of n equations defined on $\mathbb{C}^N$, as in (2.1), and a positive integer k, we define a randomization operation

$$\mathfrak{R}(f(x);k) := \Lambda f(x), \quad \Lambda \in \mathbb{C}^{k\times n}, \tag{3.1}$$

where Λ is chosen generically from $\mathbb{C}^{k\times n}$. Note that k does not have to equal n. Gaussian elimination does not change the ideal generated by $\Lambda f(x)$. Therefore, when $k \geq n$, $\Lambda f(x)$ is equivalent to the system consisting of $f(x)$ plus $k-n$ identically zero equations. For the same reason, when $k \leq n$, $\Lambda f(x)$ is equivalent to the system

$$\left[\begin{array}{cc} I_k & R \end{array}\right] f(x), \quad R \in \mathbb{C}^{k\times(n-k)}, \tag{3.2}$$

where I_k is the $k\times k$ identity matrix. Consequently, we may without loss of generality assume that $\mathfrak{R}(f(x);k)$ is of the form of (3.2) with a generic choice of $R \in \mathbb{C}^{k\times(n-k)}$. This form allows us to take some advantage of the original equations. For example, if $k = N$ and the original equations had total degrees $d_1 \geq d_2 \geq \cdots \geq d_n$, then the total degree of the original form of $\mathfrak{R}(f(x);N)$ is d_1^N, but the total degree of the modified form is $d_1 d_2 \cdots d_N$.

The following lemma gives the main properties of randomization.

LEMMA 3.1. *Let*

$$f(x) := \left[\begin{array}{c} f_1(x) \\ \vdots \\ f_n(x) \end{array}\right] \tag{3.3}$$

be a system of restrictions of n polynomials on $\mathbb{C}^m$ to a pure N-dimensional affine algebraic set $X \subset \mathbb{C}^m$. Assume that $k \leq \min\{n, N\}$. Assume that f does not vanish on any component of X. These conclusions follow:

1. *The dimension of any component of $V(\mathfrak{R}(f(x);k))$ is $\geq N-k$.*
2. *The irreducible components of $V(\mathfrak{R}(f(x);k))$ and $V(f)$ of dimension greater than $N-k$ are the same, and the irreducible $(N-k)$-dimensional components of $V(f)$ are components of $V(\mathfrak{R}(f(x);k))$.*

Proof. This variant of Bertini's theorem follows by the same type of reasoning as the analogous result in [12, 17] for systems of N polynomials on $\mathbb{C}^N$. For the convenience of the reader, we give a proof.

The first conclusion is simply [11, Corollary 3.14].

Since $V(f) \subset V(\mathfrak{R}(f(x);k))$, the second conclusion will follow if we show that all irreducible components $V(\mathfrak{R}(f(x);k)) \cap (X \setminus V(f))$ have dimension $N-k$. Thus it suffices to show that if $V(f)$ is empty, it follows that all irreducible components $V(\mathfrak{R}(f(x);k))$ have dimension $N-k$. This is immediate from Theorem B.1. ∎

4. Construction of homotopies. Our algorithm for intersecting algebraic varieties is based on constructing homotopies to solve a system of polynomial equations restricted to an algebraic set. This is a generalization of existing homotopies, which have until now always worked on complex Euclidean space, $\mathbb{C}^m$. Accordingly, in this section we give a very general construction for homotopies on varieties.

Let $X \subset \mathbb{C}^m$ be an irreducible N-dimensional affine algebraic set and let Y be an irreducible r-dimensional smooth algebraic set with $r \geq 1$. Let

$$f(x,y) = \begin{bmatrix} f_1(x,y) \\ \vdots \\ f_N(x,y) \end{bmatrix} = 0 \tag{4.1}$$

be a system of N algebraic functions on $X \times Y$. In practice, Y is a parameter space defining a family of systems of interest, and for any one member of the family, we wish to find its solution points in X.

More precisely, suppose we have some parameter value $y^* \in Y$ for which we want to find a finite set $\mathcal{F}^*$ of solutions of the system $f(x,y^*) = 0$, such that all the isolated solutions of $f(x,y^*) = 0$ are contained in $\mathcal{F}^*$. A procedure to do this proceeds in a number of steps in the same manner as if Y is $\mathbb{C}^N$:

1. Choose a point $y' \in Y$ for which we can find the isolated solutions $\mathcal{F}'$ of $f(x,y') = 0$, and the number of isolated solutions is the maximum number D for any system $f(x,y) = 0$ as a system in the x variables. We assume here that $y' \neq y^*$, since otherwise we are done.
2. Construct a smooth connected algebraic curve $B \subset Y$ which contains y^* and y'. (Typically Y is a Euclidean space and we choose B equal to the complex line joining the points y^* and y'.)
3. Construct a differentiable mapping $c : [0,1] \times \Gamma \to B$, where Γ is an interval or the unit circle, $c(0,\Gamma) = y^*$, $c(1,\Gamma) = y'$, and where there is a positive integer K such that given any point $y'' \in c([0,1] \times \Gamma)$ not equal to y^* or y', it follows that $c^{-1}(y'')$ has at most K inverse images.
4. Choose a random $\gamma \in \Gamma$ and starting with the isolated solutions $\mathcal{F}'$ of $f(x,y') = 0$ use homotopy continuation of the system $f(x,c(t,\gamma)) = 0$ to continue from the solutions $\mathcal{F}'$ at $t = 1$ to solutions $\mathcal{F}^*$ at $t = 0$.

Let us show that if we can make the choices specified by this procedure, we will find a finite set $\mathcal{F}^*$ of solutions of the system $f(x,y^*) = 0$, such that all the isolated solutions of $f(x,y^*) = 0$ are contained in $\mathcal{F}^*$. In Appendix A we show how to relax item 2 so that the procedure can be carried in all situations where Y is irreducible.

It may happen that the solution sets of $f(x,y) = 0$ for some or all the $y \in Y$ also contain positive dimensional solution components. Nevertheless, the number D in item 1 exists and is finite by general results, e.g., [10]. Now choose B as in item 2. Lemma A.1 guarantees that for all but a finite number of points $\hat{y} \in B$, $f(x,\hat{y}) = 0$ has D isolated solutions and that the closure of the union of the isolated solutions of $f(x,\hat{y}) = 0$ as $\hat{y}$ runs over B is an algebraic curve $\mathcal{B}$ which surjects generically D-to-one onto B. Since the set of points in B over which this mapping is not a covering is an algebraic set and hence finite, the procedure is seen to work.

Remark 4.1. Algebraic functions on an affine algebraic set $X \subset \mathbb{C}^m$ are the restrictions of polynomials from $\mathbb{C}^m$. If by $\deg f_i$ we denote a degree of a polynomial on $\mathbb{C}^m$ restricting to X, it follows that the number D above is at most $\deg X \times \Pi_{i=1}^N \deg f_i$. From this it further follows that if we can find a y' such that $f(x,y') = 0$ has $\deg X \times \Pi_{i=1}^N \deg f_i$ nonsingular isolated solutions, we can use y'.

Remark 4.2. Lemma A.1, which justifies the above procedure, is strong enough to yield the algorithms we need to construct witness points. However, the lemma is too weak to relate the multiplicity of the points as they appear in these algorithms to the multiplicity of the components that they represent. See Appendix A for more details.

5. An abstract embedding theorem. The object of this section is to present Theorem 5.1, a generalization of the main theorem of [12]. We are aiming for the same results as in that article except that $\mathbb{C}^N$ is replaced by a pure N-dimensional affine algebraic set X. We call the generalization in this section abstract because we do not specify an explicit description of X; a numerical version is the topic of the next section. Since the proof of Theorem 5.1 follows the same line of reasoning of [12], we only state and discuss the parts of that article that need changes. Before we can state the theorem, we need some notation.

5.1. Definitions. Let $X \subset \mathbb{C}^m$ be a reduced pure N-dimensional affine algebraic set, i.e., an affine algebraic subset of $\mathbb{C}^m$, all of whose irreducible components are of multiplicity one and dimension N. We assume that we have a system of restrictions of polynomials on $\mathbb{C}^m$ to X,

$$f(x) := \begin{bmatrix} f_1(x) \\ \vdots \\ f_N(x) \end{bmatrix}. \tag{5.1}$$

We assume that f does not vanish identically on any irreducible component of X. We will occasionally abuse notation and use the same notation f_i to denote the polynomial on $\mathbb{C}^m$ and its restriction to X. In line with this abuse, we let

$$x := \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \tag{5.2}$$

denote both the coordinates on $\mathbb{C}^m$ and the restrictions of the coordinates to X.

Let $\mathcal{Y}$ denote the matrix space $\mathbb{C}^{N\times(1+m+N)}$, with submatrices denoted as

$$\begin{bmatrix} \mathcal{A}_0 & \mathcal{A}_1 & \mathcal{A}_2 \end{bmatrix}, \tag{5.3}$$

where $\mathcal{A}_0 \in \mathbb{C}^{N\times 1}$, $\mathcal{A}_1 \in \mathbb{C}^{N\times m}$, and $\mathcal{A}_2 \in \mathbb{C}^{N\times N}$. We have the stratification of $\mathcal{Y}$,

$$\mathcal{Y}_0 \subset \mathcal{Y}_1 \subset \cdots \subset \mathcal{Y}_N, \tag{5.4}$$

where $\mathcal{Y}_i$ is the subspace of $\mathcal{Y}$ obtained by setting the last $N-i$ rows of $\mathcal{Y}$ equal to 0, and we define $\pi_i : \mathcal{Y} \to \mathcal{Y}_i$ as the corresponding projection. Note in particular that $\mathcal{Y}_N$ is $\mathcal{Y}$, whereas $\mathcal{Y}_0$ is an $N \times (1+m+N)$ matrix of zeros. Defining e_i as the $N \times N$ matrix of all zeros except a 1 in the ith diagonal element and letting $P_i = \sum_{j=1}^{i} e_j$, we can explicitly write $\pi_i(Y) = P_i Y$. We define P_0 to be the $N \times N$ matrix with all entries zero. This notation will be useful in defining a homotopy.

We let

$$z := \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix} \tag{5.5}$$

denote coordinates on $\mathbb{C}^N$.

5.2. Embedding and cascade. For $Y = (\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2) \in \mathcal{Y}$, we define the system

$$\mathcal{E}(f)(x, z, Y) := \begin{bmatrix} f(x) + \mathcal{A}_2^T z \\ z - \mathcal{A}_0 - \mathcal{A}_1 x \end{bmatrix}, \tag{5.6}$$

which admits the embeddings

$$\mathcal{E}_i(f)(x, z, Y) = \mathcal{E}(x, z, \pi_i(Y)) = \mathcal{E}(x, z, P_i Y). \tag{5.7}$$

We often refer to $\mathcal{E}_i(f)(x, z, Y)$ by $\mathcal{E}_i$ or $\mathcal{E}_i(f)$. We regard $\mathcal{E}_i$ as a family of systems of equations on $X \times \mathbb{C}^N$ parameterized by $\mathcal{Y}_i$. Note that

1. the $\mathcal{E}_i$ are the restrictions of systems $\mathcal{E}_N$ to $\mathcal{Y}_i$; and
2. the $\mathcal{E}_i$ can be identified with systems on $X \times \mathbb{C}^i$ with coordinates $z_1, \ldots, z_i$ on $\mathbb{C}^i$. (This is because $z_j = 0$ for $j > i$.) Thus, $\mathcal{E}_0$ is naturally identified with the system f.

For i from 1 to N and $\gamma_i \in \{\gamma \in \mathbb{C} \mid |\gamma| = 1\}$, we define a cascade of homotopies that connect the embedded systems:

$$\mathcal{H}_i(x, z, t, Y, \gamma_i) := \begin{bmatrix} f(x) + \mathcal{A}_2^T z \\ P_{i-1}(z - \mathcal{A}_0 - \mathcal{A}_1 x) \\ \qquad + e_i((1-t)z + \gamma_i t(z - \mathcal{A}_0 - \mathcal{A}_1 x)) \\ \qquad + (I_N - P_i)z \end{bmatrix}. \tag{5.8}$$

The nonzero parts of the three terms in the lower block of this expression occupy separate rows, with only the ith row depending on t. At $t = 1$, $\mathcal{H}_i(x, z, 1, Y, \gamma_i)$ is equivalent to $\mathcal{E}_i(x, z, Y)$ (they differ only in that the ith row of the lower block has been scaled by γ_i), and at $t = 0$, $\mathcal{H}_i(x, z, 0, Y, \gamma_i) = \mathcal{E}_{i-1}(x, z, Y)$. Homotopy $\mathcal{H}_i$ allows us to compute solutions to the embedded systems by continuation, as described in the next paragraph.

For i from 1 to N, $\mathcal{F}_i$ denotes the solutions to $\mathcal{E}_i = 0$ with $z \neq 0$. In the case of $i = 0$, we make the convention that $\mathcal{F}_0$ is the empty set. Of course, like $\mathcal{E}_i$, $\mathcal{F}_i$ depends on $Y \in \mathcal{Y}$. We do not emphasize the dependence since the thrust of the main result is that a generic choice of Y, which is done once and for all using a random number generator in implementations, has a number of nice properties:

1. The solutions $\mathcal{F}_i$ of $\mathcal{E}_i = 0$ are nonsingular and isolated and equal to the set of solutions of $\mathcal{E}_i = 0$ with $z_i \neq 0$.
2. The solutions of $\mathcal{E}_i = 0$ equal $\mathcal{F}_i$ for $i > \dim V(f)$.
3. For all $u \in \mathcal{F}_i$ and but a finite number of γ_i, there is a unique continuous map $s_u(t) : [0, 1] \to X \times \mathbb{C}^i$ such that
   (a) $s_u(1) = u$;
   (b) $\mathcal{H}_i(s_u(t), t, Y, \gamma_i) = 0$; and
   (c) the Jacobian of $\mathcal{H}_i(x, z, t, Y, \gamma_i)$ with respect to (x, z) is invertible at $(s_u(t), t)$ for $t \in (0, 1]$.
4. The limits of the functions $s_u(t)$ as $t \to 0$, which exist by the last properties, consist of the set $\mathcal{F}_{i-1}$ plus a finite set $\widehat{W}_{i-1}$.

The collection of sets $\widehat{W}_i$ for $i = 1, \ldots, N$ contains the witness points for the irreducible decomposition of $f^{-1}(0)$. This is stated formally in the following theorem, a generalization of the main theorem of [12].

THEOREM 5.1. *Let f be the restriction of a system of N polynomials on $\mathbb{C}^m$ to a pure N-dimensional affine algebraic set $X \subset \mathbb{C}^m$. Assume that f is not identically*

zero on any irreducible component of X and that $(\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2)$ is chosen generically. If j is the largest integer with $\widehat{W}_j$ nonempty, then the dimension of $f^{-1}(0)$ is j. Moreover, given any irreducible component W of $f^{-1}(0)$ of dimension $i \leq j$, then the finite set $\widehat{W}_i$ contains $\deg(W_{\rm red})$ *generic points of $W_{\rm red}$, each counted ν_W times, where ν_W is a positive integer, and $W_{\rm red}$ is the reduction of W. The remaining points $J_i \subset \widehat{W}_i$ lie on components of $f^{-1}(0)$ of dimension $> i$.*

Theorem 5.1 is a consequence of Lemma A.1 and Lemmas B.2 and B.4.

6. Numerical embedding. In this section we show how to numerically implement the algorithm of section 5. We assume that we have f and $X \subset \mathbb{C}^m$ as in Theorem 5.1. We assume that we have a system of polynomials on $\mathbb{C}^m$

$$g(x) := \begin{bmatrix} g_1(x) \\ \vdots \\ g_n(x) \end{bmatrix} \tag{6.1}$$

such that X is a union of dimension N irreducible components of $V(g)$. Once and for all choose a randomized system of $m - N$ polynomials $G(x) := \mathfrak{R}(g(x); m - N)$. By Lemma 3.1 we know that X is a union of dimension N irreducible components of $V(G)$.

We further assume that we begin with a witness set for X; that is, we know its dimension N and have found the $\deg X$ smooth isolated witness points $W = \mathcal{L}_{m-N} \cap X$ for a general linear subspace $\mathcal{L}_{m-N}$ of dimension $m - N$. (This will be on hand after computing the numerical irreducible decomposition of $g(x) = 0$.)

To convert the abstract systems of the previous section to systems we can compute with, we append G. Thus regarding the f_i as polynomials on $\mathbb{C}^m$, we replace $\mathcal{E}_i(f)$ by

$$\begin{bmatrix} G(x) \\ \mathcal{E}_i(f)(x, z) \end{bmatrix}, \tag{6.2}$$

which by abuse we still call $\mathcal{E}_i(f)$. We let $\widetilde{\mathcal{E}}_i(f)$ denote the original $\mathcal{E}_i(f)$ without the $G(x)$; we use this only in (6.6) and (6.7).

To start the algorithm we need to solve $\mathcal{E}_N(f) = 0$. Assume the total degree of f_i as a polynomial on $\mathbb{C}^m$ is d_i for each i.

Choosing $d_1 + \cdots + d_N$ general linear forms

$$L_{1,1}(x), \ldots, L_{1,d_1}(x), \ldots, L_{N,1}(x), \ldots, L_{N,d_N}(x) \tag{6.3}$$

on $\mathbb{C}^m$, we want them to have the good property that for any choice of integers i_j in $1, \ldots, d_j$ for each j in $1, \ldots, N$, the solution set $\mathcal{S}_{i_1,\ldots,i_N}$ of the system of restrictions to X of the linear equations

$$\begin{aligned} L_{1,i_1}(x) &= 0 \\ &\vdots \\ L_{N,i_N}(x) &= 0 \end{aligned} \tag{6.4}$$

consists of $\deg X$ nonsingular isolated solutions, and moreover $\mathcal{S}_{i_1,\ldots,i_N} \cap \mathcal{S}_{k_1,\ldots,k_N} = \emptyset$ unless $(i_1, \ldots, i_N) = (k_1, \ldots, k_N)$. Let $\pi : \mathbb{C}^m \to \mathbb{C}^N$ denote a general linear projection. As discussed in [14], π_X is proper and $(\deg X)$-to-one. Let B be the proper algebraic subset such that π_X is an unramified cover when restricted to $X \backslash \pi^{-1}(B)$. By

composing with π we have reduced to the straightforward observation that choosing $d_1+\cdots+d_N$ general linear functions L_i on $\mathbb{C}^N$ for i from 1 to $d_1+\cdots+d_N$, it follows that

1. the unique zero of the linear functions $L_{i_1},\ldots,L_{i_N}$ for distinct $i_1,\ldots,i_N$ between 1 and $d_1+\cdots+d_N$ vanishes at a general point of $\mathbb{C}^N\setminus B$; and
2. given any $N+1$ of the $d_1+\cdots+d_N$ linear functions, there are no solutions on $\mathbb{C}^N$.

The system

$$\widehat{L}(x)=\begin{bmatrix} L_{1,1}(x)\cdots L_{1,d_1}(x)\\ \vdots\\ L_{N,1}(x)\cdots L_{N,d_N}(x)\end{bmatrix}=0 \tag{6.5}$$

has $d_1\cdots d_N\cdot\deg X$ nonsingular isolated solutions w_α contained in X_{reg}, the Zariski open set of smooth points of X. By homotopy continuation tracking from $\mathcal{L}_{m-N}$ to each of the $d_1\cdots d_N$ linear systems that occur in the system $\widehat{L}(x)$, we can compute all the solutions w_α of $\widehat{L}(x)=0$.

Fix the homotopy

$$H(x,z,t):=\begin{bmatrix} G(x)\\ (1-t)\widetilde{\mathcal{E}}_N(f)(x,z)+t\gamma\begin{bmatrix}\widehat{L}(x)\\ z\end{bmatrix}\end{bmatrix}=0, \tag{6.6}$$

where γ is any of all but a finite number of norm one complex numbers. The solutions of $\mathcal{E}_N(f)=0$ are the nonsingular limits as $t\to 0$ on X of paths starting at $t=1$ with the w_α and $z_i=0$ for all i.

Remark 6.1. In practice we often have some estimate, say, $\mathcal{N}-1$, of the largest dimension of any component of the solution set of f on X. This will happen, for example, in section 7. In such a situation we need only start with $\mathcal{E}_{\mathcal{N}}$. In this case we can replace the homotopy (6.6) with

$$H(x,z,t):=\begin{bmatrix} G(x)\\ (1-t)\widetilde{\mathcal{E}}_{\mathcal{N}}(f)(x,z)+t\gamma\begin{bmatrix}\widehat{L}(x)\\ z\end{bmatrix}\end{bmatrix}=0. \tag{6.7}$$

Note that the smooth nonsingular solutions of $\mathcal{E}_i(f)$ on X are generic. Thus they miss $E:=\overline{(G^{-1}(0)\setminus X)}\cap X$ except for a proper algebraic set of parameter values. Thus for a Zariski open dense set of the homotopy parameters the homotopies with G compute the abstract homotopies. Although E may contain the limits of a homotopy, the value of the limit is not influenced by G.

It is important to realize that serious numerical difficulties can arise, even when we are dealing with a smooth reduced component $\mathcal{C}$ of the system f on X. These occur if $\mathcal{C}$ is contained in a component of $V(g)$ other than those in X. If this happens, path tracking to decompose the witness point superset containing generic points of $\mathcal{C}$ will be singular and require the path tracker used in [15].

7. Diagonal homotopies. Assume that A is an irreducible component of the solution set of polynomial system $f_A(u)=0$ in $u\in\mathbb{C}^k$ of dimension $a>0$ and that B is an irreducible component of the solution set of polynomial system $f_B(v)=0$ in $v\in\mathbb{C}^k$ of dimension $b>0$. An important special case of this is when f_A and f_B are the same system and A and B are distinct irreducible components. After renaming if

necessary, we assume $a \geq b$. Moreover, we assume that B is not contained in A, since we would check this at the start of the algorithm and terminate if B was contained in A. Thus all components of $A \cap B$ are of dimension at most $b-1$.

We wish to compute the irreducible decomposition of $A \cap B$. Note that the product $X := A \times B \subset \mathbb{C}^{2k}$ is irreducible of dimension $a+b$. The theory of the preceding sections applies with $m = 2k$ and $N = a+b$. The intersection of A and B can be identified, e.g., [4, Example 13.15], with $X \cap \Delta$, where Δ is the diagonal of $\mathbb{C}^{2k}$ defined by the system on X,

$$\delta(u,v) := \begin{bmatrix} u_1 - v_1 \\ \vdots \\ u_k - v_k \end{bmatrix} = 0. \tag{7.1}$$

Remark 7.1. Notice that $\delta(u,v)$ plays the role of f in (5.1).

If $a+b \geq k$, set $D(u,v)$ equal to $\delta(u,v)$ with $a+b-k$ identically zero equations adjoined. If $k < a+b$, fix a randomization $D(u,v) := \mathfrak{R}(\delta(u,v); a+b)$ once and for all. Note that the smallest dimensional nonempty component of $A \cap B$ is of dimension at least $\max\{0, a+b-k\}$. Thus by Lemma 3.1, we can find the irreducible decomposition of $A \cap B$ by finding the irreducible decomposition of $D(u,v) = 0$ on X.

Fix randomizations $F_A(u) := \mathfrak{R}(f_A(u); k-a)$ and $F_B(v) := \mathfrak{R}(f_B(v); k-b)$ once and for all. We assume that we have already processed f_A and f_B through our numerical irreducible decomposition. Thus our data for A consist of a generic system $L_A(u) = 0$ of $a = \dim A$ linear equations and the $\deg A$ solutions $\{\alpha_1, \ldots, \alpha_{\deg A}\} \in \mathbb{C}^k$ of the system

$$\begin{bmatrix} F_A(u) \\ L_A(u) \end{bmatrix} = 0, \tag{7.2}$$

and the data for B consist of a generic system $L_B(v) = 0$ of $b = \dim B$ linear equations and the $\deg B$ solutions $\{\beta_1, \ldots, \beta_{\deg B}\} \in \mathbb{C}^m$ of the system

$$\begin{bmatrix} F_B(v) \\ L_B(v) \end{bmatrix} = 0. \tag{7.3}$$

Remark 7.2. We are not assuming that A and B occur with multiplicity one. If the multiplicity is greater than one, we must use a singular path tracker [15].

Note that $A \times B$ is an irreducible component of the solution set of the system

$$\mathcal{F}(u,v) := \begin{bmatrix} F_A(u) \\ F_B(v) \end{bmatrix} = 0. \tag{7.4}$$

In the following paragraphs, we write $z_{h:k}$ to mean the column vector of variables $z_h, \ldots, z_k$.

Since we know that all components of $A \cap B$ are of dimension at most $b-1$, the first system of the cascade of homotopies is

$$\mathcal{E}_b(u,v,z_{1:b}) = \begin{bmatrix} \mathcal{F}(u,v) \\ \mathfrak{R}(D(u,v), z_1, \ldots, z_b; a+b) \\ z_{1:b} - \mathfrak{R}(1,u,v;b) \end{bmatrix} = 0. \tag{7.5}$$

This system consists of $k-a+k-b+a+b+b = 2k+b$ equations in $2k+b$ variables.

To start the cascade, we must find the solutions of (7.5). Recall $a \geq b$. Specializing the system from the end of section 6, we have the homotopy

$$\left[\begin{array}{c} \mathcal{F}(u,v) \\ (1-t) \left[\begin{array}{c} \mathfrak{R}(D(u,v), z_1, \ldots, z_b; a+b) \\ z_{1:b} - \mathfrak{R}(1,u,v;b) \end{array} \right] + t\gamma \left[\begin{array}{c} L_A(u) \\ L_B(v) \\ z_{1:b} \end{array} \right] \end{array} \right] = 0. \tag{7.6}$$

At $t = 1$, solution paths start at the $\deg A \times \deg B$ nonsingular solutions

$$\{(\alpha_1, \beta_1), \ldots, (\alpha_{\deg A}, \beta_{\deg B})\} \subset \mathbb{C}^{2k} \tag{7.7}$$

obtained by combining the witness points for A and B. At $t = 0$, the solution paths terminate at the desired start solutions for (7.5).

Since $A \cap B \neq \emptyset$ implies that

$$\dim A \cap B \geq a + b - k, \tag{7.8}$$

we see that when $a + b \geq k$, we do not have to continue the cascade beyond level $a + b - k$. We can codify this into the numerics by noting that the system $\mathcal{E}_b$ is, with probability one, the same as the system

$$\widehat{\mathcal{E}}_b(u, v, z_{b-\bar{a}+1}, \ldots, z_b) = \left[\begin{array}{l} \mathcal{F}(u,v) \\ \mathfrak{R}(\delta(u,v), z_{b-\bar{a}+1}, \ldots, z_b; k) \\ \mathfrak{R}(1,u,v;b-\bar{a}) \\ z_{(b-\bar{a}+1):b} - \mathfrak{R}(1,u,v;\bar{a}) \end{array} \right] = 0, \tag{7.9}$$

where $\bar{a} = k - a$. This system has $(k-a)+(k-b)+k+(k-a)+(a+b-k) = 3k-a$ equations in $3k-a$ variables. Notice that $a + b \geq k$ implies $3k - a \leq 2k + b$. To appreciate this, consider the case when a and b are both $k-1$ and f_A and f_B are each a single equation. In this case the first system of the cascade is

$$\widehat{\mathcal{E}}_1(u, v, z_1) = \left[\begin{array}{r} f_A(u) \\ f_B(v) \\ u - v + R_{k\times 1} z_1 \\ \mathfrak{R}(1,u,v;k-2) \\ z_1 - \mathfrak{R}(1,u,v;1) \end{array} \right] = 0, \tag{7.10}$$

where $R_{k\times 1}$ is a generic complex k-vector.

In the important case when $a + b \geq k$, we want to compute the start solutions for (7.9). Then, letting $\bar{a} = k - a$, the homotopy (7.6) reduces with probability one to

$$\left[\begin{array}{c} \mathcal{F}(u,v) \\ (1-t) \left[\begin{array}{c} \mathfrak{R}(\delta(u,v), z_{b-\bar{a}+1}, \ldots, z_b; k) \\ \mathfrak{R}(1,u,v;b-\bar{a}) \\ z_{(b-\bar{a}+1):b} - \mathfrak{R}(1,u,v;\bar{a}) \end{array} \right] + t\gamma \left[\begin{array}{c} L_A(u) \\ L_B(v) \\ z_{(b-\bar{a}+1):b} \end{array} \right] \end{array} \right] = 0. \tag{7.11}$$

8. Computational experiments. The diagonal homotopies are implemented in the software package PHCpack [18], recently upgraded to deal with positive dimensional solution components. The software is available at http://www.math.uic.edu/~jan.

To compute witness points on all positive dimensional components of the intersection, we distinguish three stages:

1. Given witness points on the two components, construct the top dimensional system in the cascade and the start system to start the cascade.
2. Use polynomial continuation to compute the solutions at the start of the cascade.
3. Follow all paths defined by the cascade, in b stages, until all slack variables in $z_{1:b}$ are eliminated or until no more paths are left to trace. When $a+b \geq k$, we need work only with $z_{(b-\bar{a}+1):b}$.

The complexity of this procedure thus depends on

1. the number of variables (and equations) in the top dimensional system in the cascade,
2. the number of paths it takes to compute the solutions at the start of the cascade, and
3. the number of paths defined by the cascade.

Although we mention timings of runs done on a 2.4 Ghz Linux machine, the numbers describing the complexity are less transient.

8.1. An illustrative example. Consider the following example:

$$f(x,y,z,w) = \begin{bmatrix} xz \\ xw \\ yz \\ yw \end{bmatrix} = 0. \tag{8.1}$$

There are two solution components of dimension two, characterized by the equations $\{x=0, y=0\}$ and $\{z=0, w=0\}$. Pretending we do not know the two components intersect in the origin, we will set up a cascade of homotopies to compute the intersection of the two components.

Since we start out with four variables ($k=4$) and work with two-dimensional components ($a=b=2$), the total number of variables at the start of the cascade is $2k+b=10$. The components are characterized by one witness point each, so there is only one path to trace. Tracing one path to start the cascade takes only 80 milliseconds of CPU time and gives a point with $z_2 \neq 0$, $z_1 \neq 0$. In the first stage of the cascade, we take z_2 to zero, but z_1 remains nonzero, showing that there is not a one-dimensional component. The second stage of the cascade takes z_1 to zero and yields the origin as the point of intersection of the two components, as expected. The two stages of the cascade together take just 30 milliseconds.

8.2. Intersection of a cylinder with a sphere. In Figure 8.1 we see a sphere intersected by a cylinder. The curve C defined by this intersection is

$$C := \{\ (x,y,z) \mid x^2+y^2-1=0\ \cap\ (x+0.5)^2+y^2+z^2-1=0\ \}. \tag{8.2}$$

The total user CPU time of all path tracking is about a tenth of a second. First we track two paths to find a witness set for the cylinder, which takes 20 milliseconds. Then it also takes 20 milliseconds to compute a witness set for the sphere. We have $a=b=2$ and $k=3$; thus $a+b>k$ and the diagonal homotopy requires seven variables, as $7=3k-a$. Tracking the 2×2 paths defined by the diagonal homotopy takes 70 milliseconds of CPU time. At the end of the paths we find four points in the witness set for the curve C.

We may now move the slicing plane of the witness set to find the intersection of C with any desired plane. For example, to find the points on C of the form (x,x,z), we move the slice in a continuous fashion to $x-y=0$. Tracking the four solutions in

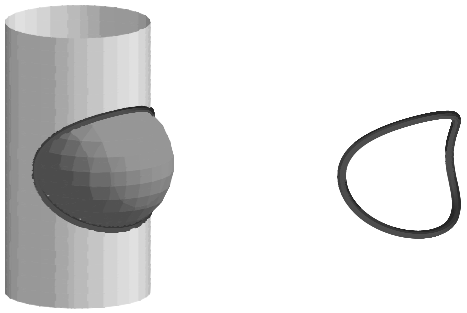


FIG. 8.1. *Intersection of a sphere with a cylinder. At the right we see the curve of degree four defined by the intersection.*

the witness set to this special plane takes only 10 milliseconds of CPU time and gives two real and two complex-conjugate solutions.

8.3. Adding an extra leg to a moving platform. In this section we give an application of the important case where one of the components is a hypersurface. We consider a special case of a Stewart–Gough platform proposed by Griffis and Duffy [6]. When further specialized to have equilateral upper and lower triangles connected by six legs in cyclic fashion from a vertex of one triangle to a midpoint of an edge of the other triangle, and vice versa, the platform permits motion. This property was first identified and analyzed by Husty and Karger [7] and subsequently reexamined in [15].

When the legs of the mechanism described above have general lengths, a formulation of the kinematic equations using Study coordinates has one curve of degree 28 and 12 lines [15]. The lines are mechanically irrelevant, so we ignore them. Suppose we form a tetrahedron by adding a fourth point in general position to the base triangle and similarly for the upper triangle and then add a seventh leg of known length connecting these two points. The condition for assembling the mechanism is equivalent to intersecting the motion curve of degree 28 for the first six legs with a quadratic hypersurface that equates the length of the seventh leg to the distance between its points of connection. This hypersurface is of the same form as the main equations in the system defining the curve. With the addition of the seventh leg, the platform will no longer move but instead will have a finite number of fixed postures.

The number of variables and equations in the original system is eight ($k = 8$). We intersect a one-dimensional component with a hypersurface; for $k = 8$, this hypersurface is of dimension seven. Since $a \geq b$, we have $a = 7$ and $b = 1$. Thus the cascade starts with 17 variables, as $2k + b = 2 \times 8 + 1 = 17$. The hypersurface is represented by two witness points and the curve we intersect has 28 witness points. To start the cascade, we trace $2 \times 28 = 56$ paths in dimension 17, using 20.3 seconds of CPU time. The cascade just has to remove one hyperplane to arrive at the 40 intersection points (16 of the 56 paths diverge), which requires 14.4 seconds of CPU time. Interestingly, a general Stewart–Gough platform also has 40 solution points.

Finally, we point out that the CPU time spent on the diagonal homotopy is considerably less than for solving the system directly. For the direct approach the input is a system in nine equations and eight variables. Before giving it to the blackbox solver of PHCpack, we add to every equation one monomial, which is a new slack variable multiplied with a random constant. The mixed volume of this new nine-dimensional system is 164. The computation of the mixed volume and tracking of all 164 paths

takes 108.5 seconds (1.8 minutes) of CPU time. At the end we find the same 40 intersection points; the other 124 paths diverged to infinity. Notice that in the diagonal homotopy, only 16 paths diverged.

9. Conclusions. In this paper, we extend the cascade of [12] to compute witness points on all components of the intersection of two irreducible varieties. This is done by computing the irreducible decomposition of the diagonal of the product of the two irreducible varieties, and so we call the new procedure a diagonal homotopy. The procedure is justified as a special case of a method, also described herein, for the irreducible decomposition of the solution set of any polynomial system restricted to an irreducible algebraic set.

The diagonal homotopy given here always has at least twice the number of variables as the ambient space of the varieties being intersected. In a sequel to this paper, we will describe a modification to the diagonal homotopy that avoids the explicit doubling of the system, which leads to more efficient computation.

Appendix A. Homotopy on an algebraic set. In section 4, we give a procedure for constructing homotopies to solve a system of parameterized polynomials restricted to an algebraic set. In that procedure, Y is the parameter space and B is a smooth curve in Y, and we compute solution paths along a one-real-dimensional curve in B. Both Y and B are irreducible algebraic sets.

While choosing a smooth B is a difficulty when Y is irreducible and singular, it is always easy to find an irreducible curve B that contains y' and y^* with $y' \notin B_{\text{Sing}}$. If y' is not in the singular set Y_{Sing} of Y, then B is not contained in Y_{Sing}. This is more than enough for the procedure to find a finite set $\mathcal{F}^*$ of solutions of the system $f(x, y^*) = 0$, such that all the isolated solutions of $f(x, y^*) = 0$ are contained in $\mathcal{F}^*$.

In fact, the procedure given in section 4 works with item 2 of that procedure relaxed to the condition of constructing an irreducible curve B containing y' and y^* such that

1. B is not contained in the singular set Y_{Sing} of Y, and
2. $y' \notin B_{\text{Sing}}$.

That the procedure finds a finite set $\mathcal{F}^*$ of solutions of the system $f(x, y^*) = 0$, such that all the isolated solutions of $f(x, y^*) = 0$ are contained in $\mathcal{F}^*$, may be shown by reducing to the nonsingular case:

1. Desingularize Y, i.e., let $\pi : \overline{Y} \to Y$ denote a surjective birational morphism which gives an isomorphism from $\overline{Y} \setminus \pi^{-1}(Y_{\text{Sing}})$ to $Y \setminus Y_{\text{Sing}}$.
2. Note that since $B \not\subset Y_{\text{Sing}}$, there is an algebraic curve $B' \subset \overline{Y}$ that maps generically one to one and onto B. By using embedded resolution of $\overline{Y}$, it can be further assumed that B' is smooth.
3. By composition with π we get algebraic functions f_i' on $X \times \overline{Y}$.
4. Note that given $y^* \in B \subset Y$, there is a point $y'^* \subset B'$ that maps onto y^*.
5. Note that the result shown holds for $X, \overline{Y}, f', B'$ and that (with the obvious identifications) the system $f(x, y^*) = 0$ on $X \times \{y^*\}$ is identical to the system $f'(x, y'^*)$ on $X \times \{y'^*\}$.

The lemma which justifies the homotopy is as follows.

LEMMA A.1. *Let $X \subset \mathbb{C}^m$ be an irreducible N-dimensional affine algebraic set and let Y be an irreducible smooth algebraic set. Let*

$$f(x, y) = \begin{bmatrix} f_1(x, y) \\ \vdots \\ f_N(x, y) \end{bmatrix} = 0 \tag{A.1}$$

be a system of N algebraic functions on $X \times Y$. Let x^* be an isolated solution of $f(x, y^*) = 0$ for a fixed value $y^* \in Y$, i.e., assume that there is an open set $\mathcal{O} \subset X$ containing x^* with x^* the only solution of $f(x, y^*) = 0$ on $\mathcal{O}$. Then there exists a neighborhood V of $y^* \in Y$ such that for any $y \in V$ there exists at least one isolated solution of $x \in \mathcal{O}$ of $f(x, y) = 0$.

Proof. This result is a special case of a basic general result from complex algebraic geometry, e.g., [11, (3.10)]. Any irreducible component of $f(x, y) = 0$ is of dimension $\geq \dim Y$. Choose such a component C through (x^*, y^*). Consider $\overline{C} \subset \overline{X} \times Y$, where we close up X within $\mathbb{P}^m$. Since the induced projection $\pi : \overline{C} \to Y$ is proper, and there is a Euclidean neighborhood $\mathcal{O}$ of x^* as in the lemma with $\pi_{\mathcal{O}}^{-1}(y^*) = x^*$, we conclude that there is a neighborhood $V \subset Y$ of y^* such that $\pi_{C \cap (\mathcal{O} \times V)} : C \cap (\mathcal{O} \times V) \to V$ is proper. By the proper mapping theorem the image is an algebraic subset, and by the upper semicontinuity of fiber dimension it must be surjective. This proves the lemma. □

We conclude this appendix with a few remarks on multiplicity. Lemma A.1 is strong enough to yield the algorithms we need to construct witness points but unfortunately too weak for us to relate the multiplicity of x^* as a solution of $f(x, y^*) = 0$ to the multiplicity of the projection map from C to Y at (x^*, y^*). If X were a local complete intersection, then it would follow that C was Cohen–Macaulay in a neighborhood of (x^*, y^*), and we could use the stronger result [12, Lemma 6], conclude the two multiplicities are the same, and thus have the same multiplicity statements as in [12, Theorem 3].

When we apply Lemma A.1, we know a bit more information, i.e., that for a general point y' near y^*, the solutions of $f(x, y') = 0$ near (x^*, y') are nonsingular. It is worth noting in this case, e.g., using [11, Appendix to Chapter 6] that when we choose a sufficiently generic smooth curve in Y through y^*, e.g., a generic line through y^* when Y is Euclidean space, the number of paths coming into (x^*, y^*) is the multiplicity of the local ring of X at x^* with respect to the ideal generated by the functions $f_i(x, y^*)$. Unfortunately, this multiplicity is in general bounded only by the multiplicity of (x^*, y^*) as a solution of $f(x, y^*) = 0$.

Appendix B. Proof of the main theorem. For algebraic sets, there is a very strong version of Sard's theorem, e.g., [11, Theorem 3.7]. This result has a large number of consequences, going under the name Bertini's theorem, asserting that the zero set of a suitably general function inherits properties of the set the function is defined on. For the convenience of the reader we collect in one place a Bertini theorem of sufficient generality to cover the needs of this article. Given a complex vector space V, we let V^r denote the Cartesian product of V with itself r times, i.e., the space of r-tuples of elements of V. V^r has a natural vector space structure given by addition of r-tuples and multiplication of an r-tuple by a complex number being defined as the r-tuple obtained by componentwise multiplication of elements of the r-tuple by the complex number. V^r with this vector space structure is denoted by $V^{\oplus r}$. In the following theorem, we use notation close to that of Fulton [5, Lemma B.9.1].

THEOREM B.1 (Bertini's theorem). *Let X denote an irreducible algebraic subset of $\mathbb{C}^k$. Let $Z_1, \ldots, Z_q$ denote a finite number of irreducible algebraic subsets of X (with one of the X_i possibly equal to X). Let $\mathbb{V}$ denote a finite dimensional vector space of polynomial functions on $\mathbb{C}^k$. Assume that for each point of X at least one element of $\mathbb{V}$ does not evaluate to zero. Then for any integer $r > 0$, there is a Zariski open dense set $U \subset \mathbb{V}^r$ such that for $f := (f_1, \ldots, f_r) \in U$ it follows for each Z_i*

that

1. *if* $V(f) \cap Z_i$ *is nonempty, then* $\dim V(f) \cap Z_i = \dim Z_i - r$, *and*
2. *letting* $\mathrm{Sing}(Z_i)$ *denote the singular set of* Z_i, $V(f) \cap (Z_i \setminus \mathrm{Sing}(Z_i))$ *is smooth.*

Proof. We first apply [5, Lemma B.9.1]. For the vector bundle E in [5, Lemma B.9.1] take $X \times \mathbb{C}^r$; take $p = 1$ with $C_1 = X \times \{0\}$; for Γ take $\mathbb{V}^{\oplus r}$, i.e., take $\mathbb{V}^r$. The conclusion from [5, Lemma B.9.1] is the existence of a Zariski open dense set Γ^o of Γ such that for $f := (f_1, \ldots, f_s) \in \Gamma^o$, it follows that if $V(f) \cap Z_i$ is nonempty, then

$$\dim(V(f) \cap Z_i) \leq \dim Z_i - r.$$

The opposite inequality is a property of zero sets of functions, e.g., [11, Corollary 3.14].

Since the intersection of a finite number of Zariski open and dense sets is Zariski open and dense, it suffices to show that there is a Zariski open dense set $U_i \subset \mathbb{V}^{\oplus r}$ such that for $f \in U_i$, $V(f) \cap (Z_i \setminus \mathrm{Sing}(Z_i))$ is smooth. For this we use [3, Theorem 1.7.1.1]. Restricting to $(Z_i \setminus \mathrm{Sing}(Z_i))$, we conclude from [3, Theorem 1.7.1.1] that there is a Zariski open dense set $\mathcal{O}_1 \subset \mathbb{V}$ such that for $f_1 \in \mathcal{O}_1$ we have that $V(f_1) \cap (Z_i \setminus \mathrm{Sing}(Z_i))$ is smooth and if nonempty of dimension $\dim Z_i - 1$. Applying [3, Theorem 1.7.1.1] to the restriction of $\mathbb{V}$ to $V(f_1) \cap (Z_i \setminus \mathrm{Sing}(Z_i))$, we conclude that there is a Zariski open dense set $\mathcal{O}_2 \subset \mathbb{V}$ such that for $f_2 \in \mathcal{O}_2$ we have that $V(f_1, f_2) \cap (Z_i \setminus \mathrm{Sing}(Z_i))$ is smooth and if nonempty of dimension $\dim Z_i - 2$. Proceeding this way for j going to r, $U_i := \mathcal{O}_1 \times \cdots \times \mathcal{O}_r \subset \mathbb{V}^{\oplus r}$ is the desired Zariski open dense set. ∎

LEMMA B.2. *Let* f *and* X *be as in Theorem* 5.1. *Assume further that* Z *is an algebraic subset of* X *of dimension* $< N$. *Assume that* f *does not vanish on any component of* X *or of* Z. *There is a Zariski open and dense set* $U \subset \mathcal{Y} = \mathbb{C}^{N \times (1+m+N)}$ *such that*

1. *the solutions* $\mathcal{F}_i$ *of the system* $\mathcal{E}_i(f)(x, z, Y)$ *for* $Y \in U$ *with* $z \neq 0$ *are isolated nonsingular solutions and lie in the set* $(X \setminus Z) \times \mathbb{C}^i$;
2. $U \cap \mathcal{Y}_i$ *is Zariski open and dense for each* $i < N$; *and*
3. *the solutions of* $\mathcal{E}_i(f)(x, z, Y)$ *for* $Y \in U$ *with* $z \neq 0$ *are the same as those with* $z_i \neq 0$.

Proof. Since the following result follows almost verbatim from the reasoning in the first half of the proof of [12, Lemma 2], we give only a brief sketch of the proof. As discussed in section 5, we regard $\mathcal{E}_i$ as a system on $X \times \mathbb{C}^i$.

Consider the vector space V_1 of functions on $X \times \mathbb{C}^i$ generated by

$$f_1, \ldots, f_N, z_1, \ldots, z_i.$$

The common zeros of the functions in V_1 are the points

$$V(V_1) := \left\{(x, 0) \in X \times \mathbb{C}^i | f(x) = 0\right\}.$$

From this we conclude, using Theorem B.1, that for a choice of a system $\mathcal{S}$ in a nonempty Zariski open set of the vector space $V_1^{\oplus N}$, it follows that the common zeros $Z_{\mathcal{S}}$ of $\mathcal{S}$ on $X \times \mathbb{C}^i \setminus V(V_1)$ are pure i-dimensional with singular set of dimension $\leq i-1$. Similarly, $Z_{\mathcal{S}}$ meets $Z \times \mathbb{C}^i \setminus V(V_1)$ in a set of dimension at most $\dim Z + i - N \leq i - 1$.

Now let V_2 be the vector space of functions on $X \times \mathbb{C}^i$ generated by

$$1, x_1, \ldots, x_m, z_1, \ldots, z_i.$$

Since $1 \in V_2$, there are no common zeros of the functions in V_2. Using Theorem B.1 again, we conclude that for a generic choice of a system $\mathcal{S}'$ in a nonempty Zariski

open set of the vector space $V_2^{\oplus i}$, it follows that the common zeros of $\mathcal{S}'$ on $Z_{\mathcal{S}}$ with $z \neq 0$ are a finite set of isolated smooth points not contained in $Z \times \mathbb{C}^i$. The above system

$$\begin{bmatrix} \mathcal{S} \\ \mathcal{S}' \end{bmatrix} = 0 \tag{B.1}$$

of $N + i$ equations is of the form

$$\begin{aligned} B \begin{bmatrix} f_1(x) \\ \vdots \\ f_N(x) \end{bmatrix} + C \begin{bmatrix} z_1 \\ \vdots \\ z_i \end{bmatrix} = 0, \\ D + E \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} + F \begin{bmatrix} z_1 \\ \vdots \\ z_i \end{bmatrix} = 0, \end{aligned} \tag{B.2}$$

where B is an $N \times N$ complex matrix, C is an $N \times i$ complex matrix, D is an $i \times 1$ complex matrix, E is an $i \times m$ complex matrix, and F is an $i \times i$ complex matrix. The above Bertini type results show that the set of

$$(B, C, D, E, F) \in \mathbb{C}^{N \times (N+i) + i \times (1+m+i)}$$

giving rise to systems of the form (B.2) with only isolated nonsingular solutions on $X \times \mathbb{C}^i \setminus X \times \{0\}$ is dense in $\mathbb{C}^{N \times (N+i) + i \times (1+m+i)}$ with respect to the usual Euclidean topology. The set of such (B, C, D, E, F) such that the maximal number of isolated solutions of the associated system (B.2) on $X \times \mathbb{C}^i \setminus X \times \{0\}$ occurs is a dense constructible set and thus by Chevalley's theorem, e.g., [11, Proposition 2.31], contains a dense Zariski open set $\mathcal{O}$. Moreover, we know that the systems of the form (B.2) with only isolated solutions on $X \times \mathbb{C}^i \setminus X \times \{0\}$ form a constructible set $\mathcal{C}$ of (B, C, D, E, F). By the density of the systems (B.1) in the usual Euclidean topology, we conclude that $\mathcal{C}$ is a dense Zariski constructible set and thus contains a dense Zariski open set $\mathcal{O}'$. The systems arising with parameters from the set $U_i' = \mathcal{O} \cap \mathcal{O}'$ have the properties required for the first assertion of the lemma. The matrices (B, C, D, E, F) giving rise to systems with the desired properties are invariant under the action

$$G_1 \times G_2 \times (B, C, D, E, F) \to (G_1^{-1}B, G_1^{-1}C, G_2^{-1}D, G_2^{-1}E, G_2^{-1}F),$$

where G_1 is an invertible $N \times N$ complex matrix and G_2 is an invertible $i \times i$ complex matrix. Thus we can assume that U_i' is invariant under this action. Since the matrices (B, C, D, E, F) with B and F invertible form a Zariski open dense set invariant under the same action, we can assume that the Zariski open set U_i' is chosen so that all (B, C, D, E, F) in the set have B and F invertible. Since $(I_N, B^{-1}C, F^{-1}D, F^{-1}E, I_i)$ is in U_i', we see that the set

$$U_i := U_i' \cap \{(B, C, D, E, F) | B \text{ and } F \text{ invertible}\}$$

is the desired set for conclusion 1 of the lemma.

We have natural projections $\pi_i : \mathcal{Y} \to \mathcal{Y}_i$ obtained by setting the last $N - i$ rows of an element of $\mathcal{Y}$ to 0. The set $U := \cap_{i=1}^N \pi^{-1}(U_i)$ is Zariski open and dense. Noting that since the maps π_i are surjective, the images are Zariski dense constructible sets, we have, on redefining $U_i := U \cap \mathcal{Y}_i$, the first two assertions of the lemma.

For the last assertion we can assume without loss of generality that $i \geq 2$. The desired assertion will follow if we show that the set of $Y \in U$ for which there are solutions of $\mathcal{E}_i(f)(x,z,Y)$ with $z_i = 0$ but $z \neq 0$ is not Zariski dense. Assume it is Zariski dense. Then, for a general $Y \in U$ and a general $(a_i, a_{1,1}, \ldots, a_{i,N}) \in \mathbb{C}^{N+1}$, the system

$$\begin{bmatrix} \mathcal{E}_{i-1}(f)(x,z,Y) \\ a_i + a_{i,1}x_1 + \cdots + a_{i,m}x_m \end{bmatrix} = 0$$

has a solution with $(z_1, \ldots, z_{i-1}) \neq 0$. This is absurd, since we have already shown that for a general $Y \in U$, there are only a finite number of solutions of $\mathcal{E}_{i-1}(f)(x,z,Y)$ with $(z_1, \ldots, z_{i-1}) \neq 0$. ∎

Remark B.3. The condition in Lemma B.2 that the $\mathcal{F}_i$ lie in $(X - Z) \times \mathbb{C}^i$ is important because we typically will not have defining polynomials for X but know only that X is an irreducible component of $V(g)$ for a system of polynomials g. Taking Z equal to the union of the intersections of X with other components of $V(g)$ guarantees with probability one that g will be a set of defining equations for X on a Zariski open set large enough so that all the homotopy continuations that are given in this article will be well defined.

We need some information about the isolated solutions of $\mathcal{E}_i(f)(x,z,Y)$ with $z = 0$. This is the generalization of the last assertion of [12, Lemma 2].

LEMMA B.4. *There is a Zariski open and dense set $U \subset \mathcal{Y} = \mathbb{C}^{N\times(1+m+N)}$ such that the solutions of the system $\mathcal{E}_i(f)(x,z,Y)$ for $Y \in U$ with $z = 0$ consist of*

1. *positive dimensional components all contained in components of $V(f)$ of dimension greater than i, plus*
2. *for each dimension i irreducible component W of $f^{-1}(0)$, isolated solutions consisting of $\deg(W_{\mathrm{red}})$ generic points of W_{red}, the reduction of W, each occurring the same number of times.*

Proof. When $z = 0$, the system $\mathcal{E}_i(f)(x,z,Y)$ reduces to

$$\begin{bmatrix} f(x) \\ \mathcal{A}_0 + \mathcal{A}_1 \cdot x \end{bmatrix}. \tag{B.3}$$

The assertion is contained in the discussion in [17]. ∎

The remaining result from [12] that needs modification is the local extension lemma [12, Lemma 6]. We use Lemma A.1 in its place.

Acknowledgment. We would like to thank the referees for their helpful suggestions.

REFERENCES

[1] E.L. ALLGOWER AND K. GEORG, *Introduction to Numerical Continuation Methods*, Classics Appl. Math. 45, SIAM, Philadelphia, 2003.

[2] E.L. ALLGOWER AND K. GEORG, *Numerical path following*, in Techniques of Scientific Computing (Part 2), Handb. Numer. Anal. 5, P.G. Ciarlet and J.L. Lions, eds., North–Holland, Amsterdam, 1997, pp. 3–207.

[3] M. BELTRAMETTI AND A.J. SOMMESE, *The Adjunction Theory of Complex Projective Varieties*, Expo. Math. 16, Walter De Gruyter, Berlin, 1995.

[4] D. EISENBUD, *Commutative Algebra with a View Toward Algebraic Geometry*, Grad. Texts in Math. 150, Springer-Verlag, New York, 1995.

[5] W. FULTON, *Intersection Theory*, 2nd ed., Ergebnisse der Mathematik und ihrer Grenzgebiete 3, Springer-Verlag, Berlin, 1998.

[6] M. Griffis and J. Duffy, *Method and Apparatus for Controlling Geometrically Simple Parallel Mechanisms with Distinctive Connections*, U.S. Patent 5,179,525, 1993.

[7] M.L. Husty and A. Karger, *Self-motions of Griffis-Duffy type parallel manipulators*, in Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco, 2000, pp. 7–12.

[8] T.Y. Li, *Numerical solution of multivariate polynomial systems by homotopy continuation methods*, Acta Numer., 6 (1997), pp. 399–436.

[9] A. Morgan, *Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[10] A. Morgan and A.J. Sommese, *Coefficient-parameter polynomial continuation*, Appl. Math. Comput., 29 (1989), pp. 123–160; *Errata*, Appl. Math. Comput., 51 (1992), p. 207.

[11] D. Mumford, *Algebraic Geometry* I: *Complex Projective Varieties*, Grundlehren Math. Wiss. 221, Springer-Verlag, Berlin, 1976.

[12] A.J. Sommese and J. Verschelde, *Numerical homotopies to compute generic points on positive dimensional algebraic sets*, J. Complexity, 16 (2000), pp. 572–602.

[13] A.J. Sommese, J. Verschelde, and C.W. Wampler, *Numerical decomposition of the solution sets of polynomial systems into irreducible components*, SIAM J. Numer. Anal., 38 (2001), pp. 2022–2046.

[14] A.J. Sommese, J. Verschelde, and C.W. Wampler, *Using monodromy to decompose solution sets of polynomial systems into irreducible components*, in Application of Algebraic Geometry to Coding Theory, Physics and Computation, C. Ciliberto, F. Hirzebruch, R. Miranda, and M. Teicher, eds., Kluwer Academic Publishers, Norwell, MA, pp. 297–315.

[15] A.J. Sommese, J. Verschelde, and C.W. Wampler, *Symmetric functions applied to decomposing solution sets of polynomial systems*, SIAM J. Numer. Anal., 40 (2002), pp. 2026–2046.

[16] A.J. Sommese, J. Verschelde, and C.W. Wampler, *A method for tracking singular paths with application to the numerical irreducible decomposition*, in Algebraic Geometry, a Volume in Memory of Paolo Francia, M.C. Beltrametti, F. Catanese, C. Ciliberto, A. Lanteri, and C. Pedrini, eds., Walter de Gruyter, Berlin, 2002, pp. 329–345.

[17] A.J. Sommese and C.W. Wampler, *Numerical algebraic geometry*, in The Mathematics of Numerical Analysis, Lectures in Appl. Math. 32, J. Renegar, M. Shub, and S. Smale, eds., Springer-Verlag, New York, 1996, pp. 749–763.

[18] J. Verschelde, *Algorithm* 795: *PHCpack: A general-purpose solver for polynomial systems by homotopy continuation*, ACM Trans. Math. Software, 25 (1999), pp. 251–276.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1572–1591

FINITE ELEMENT METHODS FOR A MODIFIED REISSNER–MINDLIN FREE PLATE MODEL*

L. BEIRÃO DA VEIGA†

Abstract. The solution of the Reissner–Mindlin plate problem with free boundary conditions presents a strong layer effect near the free edges. As a consequence, the solution is not even uniformly bounded even in $H^{3/2}$, which implies that at most an $O(h^{1/2})$ uniform convergence rate can be reached by finite element methods in the H^1 norm. Following instead the modified free boundary model presented by Beirão da Veiga and Brezzi, which gives more regular solutions, better error estimates can be obtained in principle. In this paper we present and analyze the extension of different families of well-known optimal plate methods to this new model. All the modified methods presented are proved to be optimal and free of locking.

Key words. plates, boundary layers, finite element methods

AMS subject classifications. 74K20, 74S05

DOI. 10.1137/S003614290343181X

1. Introduction.

The history of finite element methods for Reissner–Mindlin plates is strictly tied to the well-known shear locking phenomena. This numerical effect, which arises from the natural constraint enforced in the problem when the thickness of the plate tends to zero, can severely deteriorate the convergence of the method at small thicknesses.

Numerous finite element schemes that are completely free of locking are available in the literature. Most of these methods hold a rate of convergence to the continuous solution which is independent of the plate thickness and optimal with respect to the polynomial degree of the discrete spaces and the regularity required for the solution.

On the other hand, this last point introduces another issue which is fundamental in the numerical analysis of plates—the presence of boundary layers. As a consequence of this phenomena, even the solution of Reissner–Mindlin plate problems with C^∞ loads can hold (as the thickness tends to zero) a Sobolev regularity which is not much higher than H^1, the minimum required for the variational problem formulation.

In particular, as proved in [4], even with smooth loads the maximum (uniform in t) H^s regularity which can be expected for the rotations is

$$s < \frac{7}{2} \qquad \text{soft clamped plates,}$$

$$s < \frac{5}{2} \qquad \text{hard clamped, hard supported plates,}$$

$$s < \frac{3}{2} \qquad \text{soft supported, free plates.}$$

This low solution regularity, especially because it is coupled with the locking phenomena, clearly hinders the approximation capabilities of numerical methods on the whole

*Received by the editors July 22, 2003; accepted for publication (in revised form) January 30, 2004; published electronically December 16, 2004.
http://www.siam.org/journals/sinum/42-4/43181.html
†Dipartimento di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy (beirao@dimat.unipv.it).

plate domain (unless some ad hoc mesh refinement is introduced). For example, for the hard clamped plate there are only a few elements (for instance, the nonconforming element of [5] and the MITC4 of Bathe and Dvorkin as proved in [17]) which give an uniform $O(h)$ rate of convergence adopting a basic polynomial degree equal to one and using a solution regularity which satisfies the bounds above. The effect of the boundary layer is particularly severe in the case of soft supported and free plates, where at most an $O(h^{1/2})$ uniform convergence rate can be reached in the H^1 norm. In the case of the supported plates, as there is no clear physical reason whether to use the soft or hard version, this can be avoided using the latter one. This is not true for free plates, where no other immediate choice seems to be viable.

In [9] the authors proposed a new way to model the free plate boundary condition which, holding a deeper consistency with the limit Kirchhoff model, gives a Sobolev order of regularity of up to 5/2. The idea is to minimize the usual Reissner–Mindlin energy functional under the additional condition that the tangential component of the rotations equals the tangential derivative of the transversal displacements on the free edge. Roughly speaking, this additional constraint generates free boundary conditions which are more consistent than the original ones with those of the limit (thickness $= 0$) Kirchhoff model; as a consequence, the respective boundary layer is weaker and the solution gains more regularity.

On the other hand, this new free boundary condition cannot in general be directly enforced in the discrete space (as is done for the clamped and supported conditions) without spoiling its approximation properties. Therefore it may not be immediate to derive a finite element method which shows optimal error bounds. In [9] an optimal finite element method, which was an adaptation of the Duran–Libermann element (see [18]), was presented. The aim of this contribution is to show that several families of finite elements present in the literature can be adapted to efficiently treat this model.

The outline of the paper is as follows. In section 2 we present the problem and the new free boundary model of [9]. In section 3 we treat the discretization of the problem following the Helmholtz decomposition finite element philosophy (see, for example, [11, 12]), which for instance covers well-known plate elements as the MITC. The additional boundary constraint above is here imposed in a relaxed way that involves use of the projection operator already present in this method. We also introduce a new first order nonconforming element and its adaptation to the new model. In section 4 we present the extension of the linked interpolation plate methods (see, for example, [6, 7, 21]). The projection operator which is hidden in the linked formulation is not suitable to directly enforce the free boundary constraint in a relaxed way, as done in section 3. On the other hand, we show that the particular approximation properties of this finite element allow one to enforce this condition by penalization.

For all the extensions presented we prove that the same uniform and optimal approximation properties of the original methods continue to hold. Considering the solution regularity of the modified model, this means that $O(h)$ (respectively, $O(h^{3/2})$) uniform error estimates are provided for all the first order (respectively, higher order) methods analyzed. This is to be compared with an $O(h^{1/2})$ convergence rate, which is the best that can be reached using the original free plate model.

2. The Reissner–Mindlin plate model. We start by introducing the equations in strong form for the Reissner–Mindlin plate model. Let Ω be an open bounded domain in $\mathbb{R}^2$ representing the plate and let g be an assigned sufficiently regular load. Then $(\boldsymbol{\theta}, w, \boldsymbol{\gamma})$, respectively, the rotation, transversal displacements, and scaled shear

stresses, must satisfy the scaled equations

$$-\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta})-\boldsymbol{\gamma}=\mathbf{0}\quad\text{in }\Omega, \tag{2.1}$$

$$-\operatorname{div}\boldsymbol{\gamma}=g\quad\text{in }\Omega, \tag{2.2}$$

$$\boldsymbol{\gamma}=\lambda t^{-2}(\boldsymbol{\nabla}w-\boldsymbol{\theta})\quad\text{in }\Omega. \tag{2.3}$$

In (2.1)–(2.3), $\mathbf{C}$ is the tensor of bending moduli, ε is the usual symmetric gradient operator, $\lambda(=5/6)$ is the shear correction factor, and t is the thickness.

The boundary conditions to be coupled with the above system clearly depend on the physical constraints imposed at the edges of the plate.

We first introduce the operators (defined on any rotation $\boldsymbol{\varphi}$ sufficiently regular)

$$\mathbf{M}[\boldsymbol{\varphi}]:=\mathbf{C}\,\varepsilon(\boldsymbol{\varphi}),\quad \mathbf{M}_{\mathbf{n}}[\boldsymbol{\varphi}]:=\mathbf{M}[\boldsymbol{\varphi}]\cdot\mathbf{n}, \tag{2.4}$$

and

$$M_{nn}[\boldsymbol{\varphi}]:=\mathbf{M}_{\mathbf{n}}[\boldsymbol{\varphi}]\cdot\mathbf{n},\quad M_{ns}[\boldsymbol{\varphi}]:=\mathbf{M}_{\mathbf{n}}[\boldsymbol{\varphi}]\cdot\mathbf{s}, \tag{2.5}$$

where $\mathbf{n}$ and $\mathbf{s}$ are, respectively, the outward unit normal and counterclockwise unit tangent vector to $\partial\Omega$. Here and in what follows, whenever the above operators are applied on the solution $\boldsymbol{\theta}$, this will not be written explicitly; for example, by M_{nn} we intend $M_{nn}[\boldsymbol{\theta}]$.

We also need the following notation: for every vector valued function $\boldsymbol{\eta}$ and for every scalar function v, we will write

$$\eta_n:=\boldsymbol{\eta}\cdot\mathbf{n},\quad \eta_s:=\boldsymbol{\eta}\cdot\mathbf{s},\quad v_{/n}:=\frac{\partial v}{\partial\mathbf{n}},\quad v_{/s}:=\frac{\partial v}{\partial\mathbf{s}}. \tag{2.6}$$

We now assume that the boundary $\partial\Omega$ is the union of three nonoverlapping parts $\partial\Omega=\Sigma_c\cup\Sigma_s\cup\Sigma_f$, corresponding to clamped, (hard) simply supported, and free boundary conditions. More precisely, we require (formally)

$$\boldsymbol{\theta}=\mathbf{0},\qquad w=0\qquad\text{on }\Sigma_c, \tag{2.7}$$

$$\theta_s=0,\quad w=0,\quad M_{nn}=0\qquad\text{on }\Sigma_s, \tag{2.8}$$

and

$$M_{nn}=0,\quad M_{ns}=0,\quad \gamma_n\equiv-(\operatorname{div}\mathbf{M})_n=0\qquad\text{on }\Sigma_f. \tag{2.9}$$

We make also the minimal requirement that every part of $\partial\Omega$ is the union of a finite number of connected components and that every rigid movement $\mathbf{r}$ satisfying $\mathbf{r}=\mathbf{0}$ on Σ_c and $r_s=0$ on Σ_s is necessarily $\mathbf{0}$. This, together with the usual ellipticity assumptions on $\mathbf{C}$, will grant the well-known Korn inequality: there exists a constant $\alpha>0$ such that for every $\boldsymbol{\eta}\in(H^1(\Omega))^2$ satisfying $\boldsymbol{\eta}=\mathbf{0}$ on Σ_c and $\eta_s=0$ on Σ_s we have

$$\alpha\|\boldsymbol{\eta}\|^2_{(H^1(\Omega))^2}\le\int_\Omega\mathbf{C}\,\varepsilon(\boldsymbol{\eta}):\varepsilon(\boldsymbol{\eta})\mathrm{d}x. \tag{2.10}$$

It is well known that, as the thickness $t \to 0$, the solution $(\boldsymbol{\theta}^t, w^t, \boldsymbol{\gamma}^t)$ of the Reissner–Mindlin plate model tends to a finite limit which is the solution of the Kirchhoff model with the corresponding boundary conditions; on the other side, this convergence is well known to take place only in Sobolev spaces of low order. In particular, in [4] it is proved that if $\Sigma_f = \emptyset$, then $\|\boldsymbol{\theta}(t)\|_s$ remains bounded as $t \to 0$ at least for all $s < 5/2$ while, if some free boundary conditions are present ($\Sigma_f \neq \emptyset$), the solution stays uniformly bounded only up to $s < 3/2$.

This additional irregularity of the $\Sigma_f \neq \emptyset$ case is clearly of great hindrance when t-uniform error estimates for finite element methods are sought. It is easily seen that, unless some particular technique is applied to treat the corresponding boundary layer, an error estimate of order no better than $h^{1/2}$ can be obtained in the natural norms of the problem.

This is the reason why in [9] a modified set of conditions to model the plate free boundaries was proposed; the idea is to change (2.9) with a set of boundary conditions holding deeper consistency with the limit Kirchhoff model:

$$\theta_s = w_{/s}, \quad M_{nn} = 0, \quad \text{and} \quad M_{ns/s} - \gamma_n = 0 \quad \text{on } \Sigma_f. \tag{2.11}$$

Introducing the spaces

$$\boldsymbol{\Theta} = \{\boldsymbol{\varphi} \in [H^1(\Omega)]^2 \,:\, \boldsymbol{\varphi} = 0 \text{ on } \Sigma_c,\ \boldsymbol{\varphi}_s = 0 \text{ on } \Sigma_s\}, \tag{2.12}$$

$$W = \{w \in [H^1(\Omega)] \,:\, w = 0 \text{ on } \Sigma_c \cup \Sigma_s\}, \tag{2.13}$$

$$\mathcal{V} = \{(\boldsymbol{\theta}, w) \in \boldsymbol{\Theta} \times W \text{ such that } \theta_s = w_{/s} \text{ on } \Sigma_f\}, \tag{2.14}$$

classical arguments give the following result (see [9]).

PROPOSITION 2.1. *For every $t > 0$, any smooth solution of* (2.1)–(2.3) *with the boundary conditions* (2.7), (2.8), *and* (2.11) *coincides with the unique minimizing argument on $\mathcal{V}$ of the functional*

$$J^t(\boldsymbol{\eta}, v) = \frac{1}{2} a(\boldsymbol{\eta}, \boldsymbol{\eta}) + \frac{\lambda t^{-2}}{2} ||\boldsymbol{\nabla} v - \boldsymbol{\eta}||^2_{0,\Omega} - (g, v), \tag{2.15}$$

where

$$a(\boldsymbol{\theta}, \boldsymbol{\eta}) := \int_\Omega \mathbf{C}\, \varepsilon(\boldsymbol{\theta}) : \varepsilon(\boldsymbol{\eta}) \mathrm{d}x \equiv \int_\Omega \mathbf{M} : \varepsilon(\boldsymbol{\eta}) \mathrm{d}x. \tag{2.16}$$

Conversely, the unique minimizing argument of (2.15) *satisfies* (2.1)–(2.3) *in the distributional sense, and if it is smooth enough it also satisfies the boundary conditions* (2.7), (2.8), *and* (2.11).

Concerning the t-uniform regularity of the solution for this new problem, we have the following.

PROPOSITION 2.2. *Assume, for simplicity, that the material is homogeneous (and hence the system has constant coefficients), that the load g is in $C^\infty(\bar{\Omega})$, and that Ω is a polygon (or that the boundary $\partial\Omega$ is piecewise C^∞). Let D be any open subset of Ω such that $\overline{D}$ that does not contain any vertex or points where the boundary conditions change from one type to another. Then there exists a constant c, independent of t, such that*

$$\|\boldsymbol{\theta}\|_{r+2,D} + t\|\boldsymbol{\gamma}\|_{r+1,D} + \|\boldsymbol{\gamma}\|_{r,D} + \|\mathrm{div}\boldsymbol{\gamma}\|_{r,D} \le c\|g\|_r \quad \forall\ -1 \le r < \frac{1}{2}, \tag{2.17}$$

$$\|w\|_{\rho+2,D} \le \|g\|_\rho \qquad \forall \rho \ge -1. \tag{2.18}$$

Proof. The proof can be easily obtained by adapting the analysis of [4] to the present situation near Σ_f. We see that the most irregular term in the expansion of the solution (as computed by [4]) drops, leaving for $\boldsymbol{\theta}$ a limit regularity of order $s < 5/2$. □

Consequently, following this new model, $\|\boldsymbol{\theta}(t)\|_{s,D}$ stays uniformly bounded for any $s < 5/2$. As addressed in [9], taking the steps from Proposition 2.2 it is not unrealistic to assume

$$\|\boldsymbol{\theta}\|_{r+2,\Omega} + \|w\|_{r+2,\Omega} + t\|\boldsymbol{\gamma}\|_{r+1,\Omega} + \|\boldsymbol{\gamma}\|_{r,\Omega} + \|\mathrm{div}\boldsymbol{\gamma}\|_{r,\Omega} \le c \quad \forall r < 1/2, \tag{2.19}$$

as we will do in what follows. In light of (2.18), it will also be realistic to require (as we will need for the linked interpolation elements) the additional uniform regularity

$$\|w\|_{s,\Omega} \le c \quad \forall s < 7/2. \tag{2.20}$$

This improved uniform regularity of the solution allows in principle a t-independent error estimate of up to order $h^{3/2}$ in the natural norms of the problem. The difficulty here is that the additional constraint $\theta_s = w_{/s}$ on Σ_f appearing in $\mathcal{V}$ cannot in general be directly enforced in the discrete problem (as is done for the clamped and hard simply supported conditions) without spoiling its approximation properties. Therefore it may not be immediate to derive a finite element method which shows the aforementioned optimal error bounds. In [9] an optimal finite element method, which was an adaptation of the Duran–Libermann element (see [18]), was analyzed. The aim here is to show that several families of finite elements present in the literature can be adapted to efficiently treat this model; in particular we will consider all those methods pertaining to the Helmholtz decomposition (see, for instance, [11, 12]) and the linked interpolation (see, for example, [6, 7, 21]) philosophies.

3. Model discretization following the P1–P5 philosophy. In this section we discretize and derive (uniform and optimal) error estimates for our free plate model, taking the steps from the plate finite element methods proposed in [11, 12].

Let $\mathcal{T}_h$ be a regular mesh on Ω (see, for example, [10, 14]) and h the maximum diameter of its elements. Then, we start with three conforming piecewise polynomial spaces on this grid,

$$\boldsymbol{\Theta}_h \subset \boldsymbol{\Theta}, \quad W_h \subset W, \quad \boldsymbol{\Gamma}_h \subset H(\mathrm{rot};\Omega), \tag{3.1}$$

approximating, respectively, the space of rotations, deflections, and shear stresses.

To relax the shear stress constraint, we introduce as usual a linear reduction operator Π_h,

$$\Pi_h \,:\, H(\mathrm{rot};\Omega) \longrightarrow \boldsymbol{\Gamma}_h, \tag{3.2}$$

which is the identity when restricted to $\boldsymbol{\Gamma}_h$.

At this stage we require that our operator Π_h satisfies the approximation property

$$\|\boldsymbol{\varphi} - \Pi_h\boldsymbol{\varphi}\|_0 \le c\,h\|\boldsymbol{\varphi}\|_1, \tag{3.3}$$

which in particular implies continuity as an application from $H^1(\Omega)$ to $L^2(\Omega)$.

Given the spaces $\boldsymbol{\Theta}_h, W_h$ and the operator Π_h above, we set

$$\mathcal{V}_h := \{(\boldsymbol{\eta}_h, v_h) \in \boldsymbol{\Theta}_h \times W_h \text{ such that } (\Pi_h\boldsymbol{\eta}_h)_s = (v_h)_{/s} \text{ on } \Sigma_f\}, \tag{3.4}$$

where we remark that the boundary conditions on Σ_c, Σ_s are already included in (3.1).

We can now define the discrete solution $(\boldsymbol{\theta}_h, w_h)$ as the unique minimizer of the relaxed functional

$$J_h^t(\boldsymbol{\eta}_h, v_h) = \frac{1}{2}a(\boldsymbol{\eta}_h, \boldsymbol{\eta}_h) + \frac{\lambda t^{-2}}{2}||\boldsymbol{\nabla} v_h - \Pi_h \boldsymbol{\eta}_h||^2_{0,\Omega} - (g, v_h) \tag{3.5}$$

over the discrete space $\mathcal{V}_h$. It is then easily seen that, setting

$$\boldsymbol{\gamma}_h := \lambda t^{-2}(\boldsymbol{\nabla} w_h - \Pi_h \boldsymbol{\theta}_h), \tag{3.6}$$

the triple $(\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h)$ coincides with the unique solution of the variational problem

$$\begin{cases} \text{Find } ((\boldsymbol{\theta}_h, w_h), \boldsymbol{\gamma}_h) \in \mathcal{V}_h \times \boldsymbol{\Gamma}_h \text{ such that} \\ a(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) - (\boldsymbol{\gamma}_h, \Pi_h \boldsymbol{\eta}_h - \boldsymbol{\nabla} v_h) = (g, v_h), \qquad (\boldsymbol{\eta}_h, v_h) \in \mathcal{V}_h, \\ \lambda^{-1} t^2 (\boldsymbol{\gamma}_h, \boldsymbol{\delta}_h) - (\boldsymbol{\nabla} w_h, \boldsymbol{\delta}_h) + (\Pi_h \boldsymbol{\theta}_h, \boldsymbol{\delta}_h) = 0, \qquad \boldsymbol{\delta}_h \in \boldsymbol{\Gamma}_h. \end{cases} \tag{3.7}$$

Now let $Q_h \subset L^2_0(\Omega)$ be an auxiliary finite element space such that the five properties introduced in [12] hold for the choice of spaces $(\boldsymbol{\Theta}_h, W_h, \boldsymbol{\Gamma}_h, Q_h)$:

P1. $\boldsymbol{\nabla} W_h \subset \boldsymbol{\Gamma}_h$;

P2. $\operatorname{rot} \boldsymbol{\Gamma}_h \subset Q_h$;

P3. $\operatorname{rot} \Pi_h \boldsymbol{\varphi} = P_h \operatorname{rot} \boldsymbol{\varphi}$, $\boldsymbol{\varphi} \in [H^1_0]^2$, where $P_h : L^2_0 \to Q_h$ denotes the L^2 projection;

P4. $\{\mathbf{s} \in \boldsymbol{\Gamma}_h \ : \ \operatorname{rot} \mathbf{s} = 0\} = \boldsymbol{\nabla} W_h$; and

P5. $(\boldsymbol{\Theta}_h, Q_h)$ is a stable pair of spaces for the Stokes problem.

We also assume the following:

P6. If $\gamma \in H(\operatorname{rot}; \Omega)$ is such that $\gamma_s = 0$ on Σ_f, then the same is true for $\Pi_h \gamma$.

Then we have the following crucial lemma.

LEMMA 3.1. *Consider the couple $(\boldsymbol{\theta}, w) \in \mathcal{V}$. Let the properties* P1 *to* P6 *be satisfied. Then there exists a couple of interpolants $(\boldsymbol{\theta}_I, w_I) \in \mathcal{V}_h$ such that*

$$\Pi_h(\boldsymbol{\nabla} w_I - \boldsymbol{\theta}_I) = \Pi_h(\boldsymbol{\nabla} w - \boldsymbol{\theta}) \tag{3.8}$$

with

$$\| \boldsymbol{\theta} - \boldsymbol{\theta}_I \|_1 \leq c \inf_{\boldsymbol{\varphi} \in \boldsymbol{\Theta}_h} \| \boldsymbol{\theta} - \boldsymbol{\varphi} \|_1, \tag{3.9}$$

$$\| w - w_I \|_1 \leq c \inf_{\boldsymbol{\varphi} \in \boldsymbol{\Theta}_h} \| \boldsymbol{\theta} - \boldsymbol{\varphi} \|_1 + \| \boldsymbol{\nabla} w - \Pi_h(\boldsymbol{\nabla} w) \|_0. \tag{3.10}$$

Proof. A similar proof can be found in [18]. Due to property P5, it follows easily (see [11]) that there exists $\theta_I \in \boldsymbol{\Theta}_h$ such that

$$\int_\Omega \operatorname{rot}(\boldsymbol{\theta} - \boldsymbol{\theta}_I) q = 0 \qquad \forall q \in Q_h \tag{3.11}$$

and (3.9) holds.

From (3.11) and P3 we have

$$0 = P_h \operatorname{rot}(\boldsymbol{\theta} - \boldsymbol{\theta}_I) = \operatorname{rot} \Pi_h(\boldsymbol{\theta} - \boldsymbol{\theta}_I), \tag{3.12}$$

which, recalling P4, grants the existence of $w_1 \in W_h$ such that $\Pi_h(\boldsymbol{\theta} - \boldsymbol{\theta}_I) = \boldsymbol{\nabla} w_1$.

With similar steps, starting from the obvious statement $\mathrm{rot}(\boldsymbol{\nabla} w) = 0$ we obtain $\mathrm{rot}\,\Pi_h(\boldsymbol{\nabla} w) = 0$ and finally the existence of a $w_2 \in W_h$ such that $\Pi_h(\boldsymbol{\nabla} w) = \boldsymbol{\nabla} w_2$. Setting $w_I = w_2 - w_1$ and remembering that, due to P1, $\Pi_h v_h = v_h$ for all $v_h \in W_h$, we immediately have (3.8).

Using the Poincaré inequality and the triangle inequality and recalling the definition of w_1, w_2, it follows that

$$\begin{aligned} &\| w - w_I \|_1 \le c \| \boldsymbol{\nabla} w - \boldsymbol{\nabla} w_I \|_0 \le c(\| \boldsymbol{\nabla} w - \boldsymbol{\nabla} w_2 \|_0 + \| \boldsymbol{\nabla} w_1 \|_0) \\ &= c(\| \boldsymbol{\nabla} w - \Pi_h \boldsymbol{\nabla} w \|_0 + \| \Pi_h(\boldsymbol{\theta} - \boldsymbol{\theta}_I) \|_0). \end{aligned} \tag{3.13}$$

Recalling the continuity observed for Π_h, statements (3.13) and (3.9) immediately give estimate (3.10).

What remains to check is that the couple $(\boldsymbol{\theta}_I, w_I) \in \mathcal{V}_h$, in other words, that it satisfies the boundary condition on Σ_f. But this follows easily if we remember that $(\boldsymbol{\theta}, w) \in \mathcal{V}$, using first P6 on $(\boldsymbol{\nabla} w - \boldsymbol{\theta})$ and then (3.8). □

Under the above hypothesis we have uniform error estimates, as shown in the following proposition.

PROPOSITION 3.2. *Let $((\boldsymbol{\theta}, w), \boldsymbol{\gamma})$ be the solution of the continuous problem (see section 2) while $((\boldsymbol{\theta}_h, w_h), \boldsymbol{\gamma}_h)$ is the solution of the discrete problem* (3.7). *Let the approximating spaces satisfy properties* P1 *to* P6. *Then we have the following error estimates:*

$$\begin{aligned} &\|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_1 + t\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_0 + \|\boldsymbol{\nabla}(w - w_h)\|_0 \\ &\le c \left(\inf_{\boldsymbol{\varphi} \in \boldsymbol{\Theta}_h} \| \boldsymbol{\theta} - \boldsymbol{\varphi} \|_1 + t\|\boldsymbol{\gamma} - \Pi_h \boldsymbol{\gamma}\|_0 + \| \boldsymbol{\nabla} w - \Pi_h \boldsymbol{\nabla} w \|_0 + h\, A_1 + h^{1/2}\, A_2 \right), \end{aligned} \tag{3.14}$$

where

$$A_1 = \sup_{\boldsymbol{\varphi} \in \boldsymbol{\Theta}_h} \frac{(\boldsymbol{\gamma}, (I - \Pi_h)\boldsymbol{\varphi})}{\|(I - \Pi_h)\boldsymbol{\varphi}\|_0}, \tag{3.15}$$

$$A_2 = \sup_{\boldsymbol{\varphi} \in \boldsymbol{\Theta}_h} \frac{\int_{\Sigma_f} M_{ns}(I - \Pi_h)\boldsymbol{\varphi}_s \, ds}{\|(I - \Pi_h)\boldsymbol{\varphi}_s\|_{0,\Sigma_f}}. \tag{3.16}$$

Proof. Using statement (3.8) it follows immediately that

$$\boldsymbol{\gamma}_I := \lambda t^{-2}(\Pi_h \boldsymbol{\nabla} w_I - \Pi_h \boldsymbol{\theta}_I) = \lambda t^{-2}(\Pi_h \boldsymbol{\nabla} w - \Pi_h \boldsymbol{\theta}) \equiv \Pi_h \boldsymbol{\gamma}, \tag{3.17}$$

which will play a fundamental role in our proof.

Before deriving the error equations we first notice that the space $\mathcal{V}_h$, as defined in (3.4), is not a subspace of $\mathcal{V}$, defined in (2.14). As a consequence, for $(\boldsymbol{\eta}, v) \in \mathcal{V}_h$ we have, integrating by parts and using (2.1) and (2.2), using boundary conditions (2.7), (2.8), then using (2.11), and finally (3.4),

(3.18)

$$\begin{aligned} a(\boldsymbol{\theta}, \boldsymbol{\eta}) + (\boldsymbol{\gamma}, \boldsymbol{\nabla} v - \boldsymbol{\eta}) - (g, v) &= \int_{\Sigma_f} M_{ns}\eta_s + \boldsymbol{\gamma}_n v \, \mathrm{d}s \\ &= \int_{\Sigma_f} M_{ns}(\eta_s - v_{/s}) \, \mathrm{d}s = \int_{\Sigma_f} M_{ns}(\eta - \Pi_h \eta)_s \, \mathrm{d}s. \end{aligned}$$

Comparing (3.18) with the first equation of (3.7) we obtain, for $(\boldsymbol{\eta}, v) \in \mathcal{V}_h$, the error equation

$$a(\boldsymbol{\theta} - \boldsymbol{\theta}_h, \boldsymbol{\eta}) + (\boldsymbol{\gamma}, \boldsymbol{\nabla} v - \boldsymbol{\eta}) - (\boldsymbol{\gamma}_h, \boldsymbol{\nabla} v - \Pi_h \boldsymbol{\eta}) = \int_{\Sigma_f} M_{ns}(\eta - \Pi_h \eta)_s \, \mathrm{d}s, \tag{3.19}$$

which can also be rewritten as

$$a(\boldsymbol{\theta} - \boldsymbol{\theta}_h, \boldsymbol{\eta}) + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\nabla} v - \Pi_h \boldsymbol{\eta}) = (\boldsymbol{\gamma}, (I - \Pi_h)\boldsymbol{\eta}) + \int_{\Sigma_f} M_{ns}((I - \Pi_h)\eta)_s \, \mathrm{d}s. \tag{3.20}$$

Using the Korn inequality and adding and subtracting $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ we have

$$\begin{aligned} \alpha \|\boldsymbol{\theta}_I - \boldsymbol{\theta}_h\|_1^2 + \lambda^{-1} t^2 \|\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h\|_0^2 &\le a(\boldsymbol{\theta}_I - \boldsymbol{\theta}_h, \boldsymbol{\theta}_I - \boldsymbol{\theta}_h) + \lambda^{-1} t^2 (\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h, \boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h) \\ &= a(\boldsymbol{\theta}_I - \boldsymbol{\theta}, \boldsymbol{\theta}_I - \boldsymbol{\theta}_h) + a(\boldsymbol{\theta} - \boldsymbol{\theta}_h, \boldsymbol{\theta}_I - \boldsymbol{\theta}_h) \\ &\quad + \lambda^{-1} t^2 (\boldsymbol{\gamma}_I - \boldsymbol{\gamma}, \boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h) + \lambda^{-1} t^2 (\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h). \end{aligned} \tag{3.21}$$

On the other hand, as $\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h = \lambda t^{-2}(\boldsymbol{\nabla}(w_I - w_h) - \Pi_h(\boldsymbol{\theta}_I - \boldsymbol{\theta}_h))$, it follows from the error equation (3.20) (tested with $\eta = \boldsymbol{\theta}_I - \boldsymbol{\theta}_h$ and $v = w_I - w_h$) that

$$\begin{aligned} &a(\boldsymbol{\theta} - \boldsymbol{\theta}_h, \boldsymbol{\theta}_I - \boldsymbol{\theta}_h) + \lambda^{-1} t^2 (\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h) \\ &\quad = (\boldsymbol{\gamma}, (I - \Pi_h)(\boldsymbol{\theta}_I - \boldsymbol{\theta}_h)) + \int_{\Sigma_f} M_{ns}(I - \Pi_h)(\theta_I - \theta_h)_s \, \mathrm{d}s. \end{aligned} \tag{3.22}$$

Combining (3.21) and (3.22) we then have

$$\alpha \|\boldsymbol{\theta}_I - \boldsymbol{\theta}_h\|_1^2 + \lambda^{-1} t^2 \|\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h\|_0^2 = T_1 + T_2 + T_3 + T_4, \tag{3.23}$$

where

$$T_1 = a(\boldsymbol{\theta}_I - \boldsymbol{\theta}, \boldsymbol{\theta}_I - \boldsymbol{\theta}_h), \tag{3.24}$$

$$T_2 = \lambda^{-1} t^2 (\boldsymbol{\gamma}_I - \boldsymbol{\gamma}, \boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h), \tag{3.25}$$

$$T_3 = (\boldsymbol{\gamma}, (I - \Pi_h)(\boldsymbol{\theta}_I - \boldsymbol{\theta}_h)), \tag{3.26}$$

$$T_4 = \int_{\Sigma_f} M_{ns}(I - \Pi_h)(\theta_I - \theta_h)_s \, \mathrm{d}s, \tag{3.27}$$

which we shall bound separately.

By continuity (and remembering (3.17)),

$$T_1 \le c \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_I\|_1 \, \|\boldsymbol{\theta}_I - \boldsymbol{\theta}_h\|_1, \tag{3.28}$$

$$T_2 \le c \, t \, \|\boldsymbol{\gamma} - \Pi_h \boldsymbol{\gamma}\|_0 \, t \, \|\boldsymbol{\gamma}_I - \boldsymbol{\gamma}_h\|_0. \tag{3.29}$$

For the third term, dividing and multiplying by $\|(I - \Pi_h)(\boldsymbol{\theta}_I - \boldsymbol{\theta}_h)\|_0$, using property (3.3), and finally bounding with the supremum, we obtain

$$T_3 \le c \, h \frac{(\boldsymbol{\gamma}, (I - \Pi_h)(\boldsymbol{\theta}_I - \boldsymbol{\theta}_h))}{\|(I - \Pi_h)(\boldsymbol{\theta}_I - \boldsymbol{\theta}_h)\|_0} \|\boldsymbol{\theta}_I - \boldsymbol{\theta}_h\|_1 \le c \, h \sup_{\boldsymbol{\varphi} \in \boldsymbol{\Theta}_h} \frac{(\boldsymbol{\gamma}, (I - \Pi_h)\boldsymbol{\varphi})}{\|(I - \Pi_h)\boldsymbol{\varphi}\|_0} \|\boldsymbol{\theta}_I - \boldsymbol{\theta}_h\|_1. \tag{3.30}$$

To bound the last term, we first recall the well-known Agmon inequality [1]: if e is an edge of a triangle T (with the usual minimum angle condition), $\varphi \in H^1(T)$, and h_T is the diameter of T, then we have

$$\|\varphi\|_{0,e} \le c\,(h_T^{-1/2}\|\varphi\|_{0,T} + h_T^{1/2}\|\varphi\|_{1,T}). \tag{3.31}$$

With the same arguments used in (3.30) we obtain, for any $e \in \Sigma_f$,

$$\int_e M_{ns}(I-\Pi_h)(\theta_I-\theta_h)_s\,\mathrm{d}s \le c \sup_{\boldsymbol{\varphi}\in\boldsymbol{\Theta}_h} \frac{\int_e M_{ns}(I-\Pi_h)\boldsymbol{\varphi}_s\,\mathrm{d}s}{\|(I-\Pi_h)\boldsymbol{\varphi}_s\|_{0,e}}\,\|(I-\Pi_h)(\theta_I-\theta_h)_s\|_{0,e}. \tag{3.32}$$

Using (3.31) and (3.3), from (3.32) we have

$$T_4 \le c\,h^{1/2} \sup_{\boldsymbol{\varphi}\in\boldsymbol{\Theta}_h} \frac{\int_{\Sigma_f} M_{ns}(I-\Pi_h)\boldsymbol{\varphi}_s\,\mathrm{d}s}{\|(I-\Pi_h)\boldsymbol{\varphi}_s\|_{0,\Sigma_f}}\,\|\boldsymbol{\theta}_I-\boldsymbol{\theta}_h\|_1. \tag{3.33}$$

Finally, inserting (3.28)–(3.30) and (3.33) in (3.23), and from the usual arithmetic-geometric inequality, we have

$$\alpha\|\boldsymbol{\theta}_I-\boldsymbol{\theta}_h\|_1^2 + \lambda^{-1}t^2\|\boldsymbol{\gamma}_I-\boldsymbol{\gamma}_h\|_0^2 \le c\,(\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_I\|_1^2 + t^2\|\boldsymbol{\gamma}-\Pi_h\boldsymbol{\gamma}\|_0^2 + h^2\,A_1^2 + h\,A_2^2\,), \tag{3.34}$$

where we recall that A_1 and A_2 are defined in (3.15) and (3.16).

From (3.34) we can then easily obtain an estimate for $\boldsymbol{\nabla}(w_I - w_h)$:

$$\|\boldsymbol{\nabla}(w_I-w_h)\|_0 \le \lambda^{-1}t^2\,\|\boldsymbol{\gamma}_I-\boldsymbol{\gamma}_h\|_0 + \|\boldsymbol{\theta}_I-\boldsymbol{\theta}_h\|_0. \tag{3.35}$$

The proposition is finally proved using (3.34), (3.35), (3.9), (3.10) and the triangle inequality and absorbing λ in the constants. □

3.1. Examples of finite elements. The above analysis works essentially with all the plate finite elements following the P1–P5 philosophy (see [11, 12]).

To fix the ideas, we will apply Proposition 3.2 to one of the two main elements of [11], the MITC7. Then we will show briefly how the same arguments can be applied to the general families of [11, 12]; finally, an adaptation to a new low order nonconforming element will be presented.

3.1.1. The MITC7 element. We quickly present the second order triangular mixed interpolation of tensorial components plate element, MITC7. We assume that we are given a triangulation (into triangles T) of Ω (see, for instance, [10, 14]). We introduce, on every T, the space

$$S_7(T) = P_2(T) \oplus B(T), \tag{3.36}$$

where here and in what follows P_k represents the polynomials of order k, and $B = \lambda_1\lambda_2\lambda_3$ is the cubic bubble on T.

Let also

$$RT_k(T) = [P_k]^2 \oplus P_k\,(y, -x)^T \tag{3.37}$$

be the rotated Raviart–Thomas space of order k ($k \in \mathbf{N}$; see [20]).

The discretization spaces for our problem are

$$\boldsymbol{\Theta}_h = \{\boldsymbol{\varphi} \,:\, \boldsymbol{\varphi} \in \boldsymbol{\Theta},\ \boldsymbol{\varphi}|_T \in S_7(T)\ \forall\, T\}, \tag{3.38}$$

$$W_h = \{v \,:\, v \in W,\ v|_T \in P_2(T)\ \forall\, T\}, \tag{3.39}$$

$$\boldsymbol{\Gamma}_h = \{\boldsymbol{\delta} \,:\, \boldsymbol{\delta} \in H(\mathrm{rot};\Omega),\ \boldsymbol{\delta}|_T \in RT_1(T)\ \forall\, T\}, \tag{3.40}$$

where we note that requiring $\boldsymbol{\delta} \in H(\mathrm{rot};\Omega)$ in (3.40) is equivalent to the continuity of the tangential components at the interelement boundaries.

The approximating space $\mathcal{V}_h$ therefore will be as described in (3.4); we observe that, given the degrees of freedom of the spaces above, the additional condition on the free boundary is indeed quite easy to implement.

We now introduce the reduction operator Π_h. Given a smooth $\boldsymbol{\delta}$ defined on Ω, $\Pi_h\boldsymbol{\delta}$ is the unique element of $\boldsymbol{\Gamma}_h$ such that, on every triangle T,

$$\int_e (\boldsymbol{\delta} - \Pi_h\boldsymbol{\delta})_s p_1(s)\,\mathrm{d}s = 0 \quad \forall e \text{ edge of } T,\ \forall p_1(s) \in P_1(e), \tag{3.41}$$

$$\int_T (\boldsymbol{\delta} - \Pi_h\boldsymbol{\delta})\,\mathrm{d}x\,\mathrm{d}y = \mathbf{0}. \tag{3.42}$$

From (3.41) and (3.42) it can be seen that Π_h is indeed uniquely defined (see, for example, [20]) and with image in $\boldsymbol{\Gamma}_h$. Also, the approximation property (3.3) follows from classical results (see again [20]).

For these elements, it is proved in [11] that, choosing Q_h as the piecewise linear a priori discontinuous functions, properties P1–P5 hold; on the other side, from the definition of Π_h, it can be easily checked that P6 also holds.

Applying the general Proposition 3.2 we then have the following.

PROPOSITION 3.3. *Let $((\boldsymbol{\theta}_h, w_h), \boldsymbol{\gamma}_h)$ be the solution of the discrete problem* (3.7) *using the MITC7 finite element spaces already introduced, while $((\boldsymbol{\theta}, w), \boldsymbol{\gamma})$ is the solution of the continuous problem (see section* 2*). Then, for $1 \le s \le 2$,*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_1 + t\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_0 + \|\boldsymbol{\nabla}(w - w_h)\|_0 \le c\,h^s(\,\|\,\boldsymbol{\theta}\,\|_{s+1} + t\|\boldsymbol{\gamma}\|_s + \|\boldsymbol{\gamma}\|_{s-1}\,), \tag{3.43}$$

which, recalling (2.19)*, gives an error estimate of order $O(h^s)$ for $s < 3/2$ on the full domain.*

Proof. We start applying Proposition 3.2. The bounds for the first three addenda in the left member of (3.14) follow from classical polynomial interpolation theory (see, for instance, [10, 14]). We have

$$\inf_{\boldsymbol{\varphi}\in\boldsymbol{\Theta}_h} \|\,\boldsymbol{\theta} - \boldsymbol{\varphi}\,\|_1 \le ch^2\,\|\,\boldsymbol{\theta}\,\|_3, \tag{3.44}$$

$$t\|\boldsymbol{\gamma} - \Pi_h\boldsymbol{\gamma}\|_0 \le ch^2 t\|\boldsymbol{\gamma}\|_2, \tag{3.45}$$

$$\|\boldsymbol{\nabla} w - \Pi_h\boldsymbol{\nabla} w\|_0 \le ch^2\|w\|_3. \tag{3.46}$$

To treat A_1 (see (3.15)), we use the orthogonality of $(I - \Pi_h)\boldsymbol{\varphi}$ to the piecewise constants (given in (3.42)); taking $\bar{\boldsymbol{\gamma}}$ as the piecewise constant function representing the mean value of $\boldsymbol{\gamma}$ on every triangle, we have

$$A_1 = \sup_{\boldsymbol{\varphi}\in\boldsymbol{\Theta}_h} \frac{(\boldsymbol{\gamma} - \bar{\boldsymbol{\gamma}}, (I - \Pi_h)\boldsymbol{\varphi})}{\|(I - \Pi_h)\boldsymbol{\varphi}\|_0}. \tag{3.47}$$

Now, using the Cauchy–Schwarz inequality on (3.47) and classical polynomial interpolation results, we infer

$$A_1 \leq c\,h\,\|\boldsymbol{\gamma}\|_1. \tag{3.48}$$

Using instead (3.41), with similar reasoning we obtain for A_2

$$A_2 \leq c\,h^{3/2}\,\|M_{ns}\|_{3/2,\Sigma_f}, \tag{3.49}$$

which, recalling the definition of M_{ns} and classical trace operator results for Sobolev spaces, give

$$A_2 \leq c\,h^{3/2}\,\|\boldsymbol{\theta}\|_3. \tag{3.50}$$

The proof in the $s = 2$ case follows immediately combining the above statements, while the $s = 1$ case is done with similar arguments. Finally, the extension to a general $1 \leq s \leq 2$ is done by the usual space interpolation techniques. ∎

3.1.2. Some finite element families. In [11], a recipe for building finite elements that satisfy P1–P5 is given. To apply such methods to this new free plate model, we must check an additional set of conditions.

1. *The validity of* P6. To apply Proposition 3.2, it is essential to check P6. In [11, 12], the analyzed problem is a clamped plate; therefore, the shear stresses being null at the boundary, Π_h is built as an application from $H_0(\mathrm{rot};\Omega)$ to $\boldsymbol{\Gamma}_h \subset H_0(\mathrm{rot};\Omega)$. Consequently, it is easily seen that the natural extension of Π_h to $H(\mathrm{rot};\Omega)$ (i.e., treating boundary edges as internal edges) satisfies P6.

2. *Practical applicability of the free boundary condition.* We must check that the free boundary condition in (3.4) can be expressed as a direct relation between the degrees of freedom of $\boldsymbol{\Theta}_h$ and W_h. But this usually follows naturally from the good properties of the discretization spaces and the operator Π_h, which, roughly speaking, are already built to interact well. For example this holds true for all elements introduced in [11, 12].

3. *Estimates for A_1 and A_2 in Proposition* 3.2. We must check that A_1 and A_2 can be bounded with an order in h which is sufficiently high; we want to avoid deterioration of the global estimate (3.14) by the additional A_1, A_2 terms. This usually can be done using orthogonality properties of Π_h as already shown for the MITC7. These orthogonality properties are, for example, already part of the definition of Π_h in almost all the elements of [11, 12].

Using Proposition 3.2, applying classical polynomial interpolation results, and checking the above points (see the MITC7 example), the following proposition follows easily.

PROPOSITION 3.4. *For the triangular families* I *and* III *of* [12] *of general order k, and for all the quadrilateral elements of* [11] *(both MITC*9 *($k = 2$) and higher order elements ($k = 3$)), the following error estimate holds:*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_1 + t\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_0 + \|\boldsymbol{\nabla}(w - w_h)\|_0 \leq c\,h^s(\,\|\,\boldsymbol{\theta}\,\|_{s+1} + t\|\boldsymbol{\gamma}\|_s + \|\boldsymbol{\gamma}\|_{s-1}\,), \tag{3.51}$$

where $1 \leq s \leq k$.

Remark 1. The triangular elements of [11] are not mentioned here because they are already included in those of [12]. Note also that the quadrilateral elements of [11] can be easily extended to $k \geq 4$.

Remark 2. The finite element family II of [12] was not included because point 3 above does not follow immediately (by orthogonality) as in the other cases. Clearly, this does not imply that the estimate could not be obtained by other means.

Remark 3. Recalling (2.19), the above proposition gives an error estimate of up to order 3/2 on the full domain for our free plate problem. Estimates of up to order k can, however, be expected in parts of the domain where the solution is more regular, for example, in interior subdomains.

Remark 4. The elements above are all at least of order 2. A first order element following this philosophy was presented in [9] and another will be presented in what follows.

3.1.3. A first order nonconforming element. We present here a first order method which follows this philosophy. Given a regular triangulation $\mathcal{T}_h$ of Ω of maximum diameter h, let T_+ and T_- be any two triangles with an edge e in common, and let $\mathbf{n}^+, \mathbf{n}^-$ be their outward normal unit vectors. Furthemore, given a piecewise continuous function φ on Ω (either scalar or vectorial), call φ^+ (respectively, φ^-) the trace of $\varphi\,|_{T_+}$ (respectively, $\varphi\,|_{T_-}$) on e. Then, the jump of φ across e is given by

$$[\varphi] = \varphi^+\mathbf{n}^+ + \varphi^-\mathbf{n}^- \qquad \text{for } \varphi \text{ scalar}, \tag{3.52}$$

$$[\boldsymbol{\varphi}] = (\boldsymbol{\varphi}^+ \otimes \mathbf{n}^+)_{sym} + (\boldsymbol{\varphi}^- \otimes \mathbf{n}^-)_{sym} \qquad \text{for } \varphi \text{ vectorial}, \tag{3.53}$$

where the symbol $_{sym}$ denotes the symmetric part of the tensor. If instead e is a boundary edge, then the jump function is simply $[\varphi] = \varphi\mathbf{n}$ or $[\boldsymbol{\varphi}] = (\boldsymbol{\varphi} \otimes \mathbf{n})_{sym}$, where $\mathbf{n}$ is again the outward normal.

Let Σ_i be the set of all internal edges; we introduce the space $\boldsymbol{\Theta}_h$ as

$$\boldsymbol{\Theta}_h = \left\{ \boldsymbol{\varphi} \,:\, \boldsymbol{\varphi}|_T \in P_1(T)\ \forall\, T,\ \int_e [\boldsymbol{\varphi}] = 0\ \forall e \in \Sigma_c \cup \Sigma_i,\ \int_e [\boldsymbol{\varphi}_s] = 0\ \forall e \in \Sigma_s \right\}. \tag{3.54}$$

This space follows a classical nonconforming approximation of $[H^1(\Omega)]^2$; the basic degrees of freedom are simply the values on the midpoints of the triangle edges, adjusted according to the boundary conditions.

We also define (see also (3.37))

$$W_h = \{v \,:\, v \in W,\ v|_T \in P_1(T)\ \forall\, T\}, \tag{3.55}$$

$$\boldsymbol{\Gamma}_h = \{\boldsymbol{\delta} \,:\, \boldsymbol{\delta} \in H(\mathrm{rot};\Omega),\ \boldsymbol{\delta}|_T \in RT_0(T)\ \forall\, T\}, \tag{3.56}$$

and Π_h as the classical interpolant for RT_0 (see, for example, [20]) uniquely defined on each triangle T by

$$\int_e (\boldsymbol{\delta} - \Pi_h\boldsymbol{\delta})_s \,\mathrm{d}s = 0 \quad \forall e \text{ edge of } T \tag{3.57}$$

for all sufficiently regular $\boldsymbol{\delta}$. Finally, the space $\mathcal{V}_h$ is defined as in (3.4).

We have the following proposition.

PROPOSITION 3.5. *Let $((\boldsymbol{\theta}, w), \boldsymbol{\gamma})$ be the solution of the continuous problem (see section 2). Let $((\boldsymbol{\theta}_h, w_h), \boldsymbol{\gamma}_h)$ be the solution of the discrete problem*

$$\begin{cases} \text{Find } ((\boldsymbol{\theta}_h, w_h), \boldsymbol{\gamma}_h) \in \mathcal{V}_h \times \boldsymbol{\Gamma}_h \text{ such that} \\ a(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) + j(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) - (\boldsymbol{\gamma}_h, \Pi_h\boldsymbol{\eta}_h - \boldsymbol{\nabla} v_h) = (g, v_h), \qquad (\boldsymbol{\eta}_h, v_h) \in \mathcal{V}_h, \\ \lambda^{-1}t^2(\boldsymbol{\gamma}_h, \boldsymbol{\delta}_h) - (\boldsymbol{\nabla} w_h, \boldsymbol{\delta}_h) + (\Pi_h\boldsymbol{\theta}_h, \boldsymbol{\delta}_h) = 0, \qquad \boldsymbol{\delta}_h \in \boldsymbol{\Gamma}_h, \end{cases} \tag{3.58}$$

where the jump penalty

$$j(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) = \sum_{e \in \Sigma_i \cup \Sigma_c} \frac{1}{|\, e \,|} \int_e [\boldsymbol{\theta}_h] : [\boldsymbol{\eta}_h] \, ds + \sum_{e \in \Sigma_s} \frac{1}{|\, e \,|} \int_e [(\theta_h)_s] \cdot [(\eta_h)_s] \, ds. \tag{3.59}$$

Then, we have the error estimate

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_1 + t\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_0 + \|\boldsymbol{\nabla}(w - w_h)\|_0 \le c\, h(\,\|\, \boldsymbol{\varphi} \,\|_2 + t\|\boldsymbol{\gamma}\|_1 + \|\boldsymbol{\gamma}\|_0 \,). \tag{3.60}$$

Proof. The proof will be presented rather briefly. This method is nonconforming because the space $\boldsymbol{\Theta}_h$ is not in $H^1(\Omega)$. We start defining the norm on $\boldsymbol{\Theta}_h$,

$$\|\boldsymbol{\eta}\|_\theta^2 := \|\boldsymbol{\eta}\|_1^2 + j(\boldsymbol{\eta}, \boldsymbol{\eta}), \tag{3.61}$$

which is also well defined for all θ in $[H^1(\Omega)]^2$. Then, it can be proved that the discrete Korn inequality

$$\alpha\|\boldsymbol{\eta}\|_\theta^2 \le a(\boldsymbol{\eta}, \boldsymbol{\eta}) + j(\boldsymbol{\eta}, \boldsymbol{\eta}) \qquad \forall \boldsymbol{\eta} \in \boldsymbol{\Theta}_h \tag{3.62}$$

holds (see, for example, [13, 3]) and that $a(\cdot, \cdot)$ is continuous with respect to the norm (3.61). For these finite elements, if we define w_I as the usual Lagrange interpolant of w and $\boldsymbol{\theta}_I$ as the unique element of $\boldsymbol{\Theta}_h$ satisfying

$$\int_e (\boldsymbol{\theta} - \boldsymbol{\theta}_I) \, \mathrm{d}s = \mathbf{0} \quad \forall e \text{ edge of } \mathcal{T}_h, \tag{3.63}$$

where the integral is to be intended component by component, we easily infer Lemma 3.1.

Consequently, and without referring to P1–P6, we can follow the same steps as in the proof of Proposition 3.2, starting from the ellipticity of $a(\cdot, \cdot)$ in the $\|\cdot\|_\theta$ norm. The main difference is that now, due to the nonconformity of $\boldsymbol{\Theta}_h$, when we test (2.1) and (2.3) on the discrete spaces, we get an additional addendum in the right-hand side of the error equation (3.20). Without showing the calculations, such a term is

$$\sum_{e \in \Sigma_i \cup \Sigma_c} \int_e \mathbf{M}_n \boldsymbol{\eta} \, \mathrm{d}s + \sum_{e \in \Sigma_s} \int_e M_{ns} \eta_s \, \mathrm{d}s, \tag{3.64}$$

where $\mathbf{n}$ is the outward normal to each triangle (see (2.4)). For simplicity, we will now assume $\Sigma_s = \emptyset$; the general case can be treated with the same arguments as below. When tested with $\boldsymbol{\eta} = \boldsymbol{\theta}_I - \boldsymbol{\theta}_h$ and $v = w_I - w_h$ (see (3.22)), the term (3.64) gives the additional addendum in (3.23):

$$T_5 = \sum_{e \in \Sigma_i \cup \Sigma_i} \int_e \mathbf{M}_n (\boldsymbol{\theta}_I - \boldsymbol{\theta}_h) \, \mathrm{d}s. \tag{3.65}$$

It can be checked that

$$T_5 = \sum_{e \in \Sigma_i \cup \Sigma_i} \int_e \mathbf{M} : [\boldsymbol{\theta}_I - \boldsymbol{\theta}_h] \, \mathrm{d}s, \tag{3.66}$$

where we also used that, due to the regularity of $\boldsymbol{\theta}$, $\mathbf{M}$ is continuous across the internal edges (see, for instance, [2, 13] for a similar computation); otherwise we should simply substitute $\mathbf{M}$ above with the average between the traces of $M\,|_{T_+}$ and $M\,|_{T_-}$ on e.

From the definition of $\boldsymbol{\Theta}_h$, the jumps $[\boldsymbol{\theta}_I - \boldsymbol{\theta}_h]$ are orthogonal to the constants. Therefore, for any family $\{\bar{\mathbf{M}}_e\}_{e\in\Sigma_i\cup\Sigma_c}$ of constant tensors

(3.67)
$$T_5 = \sum_{e\in\Sigma_i\cup\Sigma_c}\int_e (\mathbf{M}-\bar{\mathbf{M}}_e) : [\boldsymbol{\theta}_I - \boldsymbol{\theta}_h]\,\mathrm{d}s \le \sum_{e\in\Sigma_i\cup\Sigma_c} \|(\mathbf{M}-\bar{\mathbf{M}}_e)\|_{0,e}\|[\boldsymbol{\theta}_I - \boldsymbol{\theta}_h]\|_{0,e}.$$

Without showing the details, from the Cauchy–Schwarz inequality in l^2, the Agmon inequality (3.31), classical polynomial interpolation results, and the definition of M, we obtain

$$\begin{aligned} T_5 &\le c\left(h\sum_{T\in\mathcal{T}_h}\|\boldsymbol{\theta}\|^2_{2,T}\right)^{1/2}\left(h\sum_{e\in\Sigma_i\cup\Sigma_c}\frac{1}{|\,e\,|}\|[\boldsymbol{\theta}_I-\boldsymbol{\theta}_h]\|^2_{0,e}\right)^{1/2} \\ &\le ch\|\boldsymbol{\theta}\|_2\|\boldsymbol{\theta}_I-\boldsymbol{\theta}_h\|_\theta. \end{aligned} \tag{3.68}$$

All the other terms T_1–T_4 have the same identical form as in (3.24)–(3.27) and can be bounded by an $O(h)$ following the same arguments as in the proof of Propositions 3.2 and 3.3. Therefore, bounding the error for the deflections as already done in (3.35), the proposition is proved. □

4. Discretization following the linked interpolation philosophy. Another class of finite element methods for plates that give good results are the ones based on the kinematic linked interpolation philosophy. We will show how these methods can be adapted to our free plate model, but by following a path different from the previous section. The aim here mainly is to introduce this extension and to show another viable way to enforce our new free boundary condition in a finite element setting; therefore we will be rather brief on some more standard parts of the demonstrations.

We start with a quick introduction to the linked interpolation method, referring, for example, to [6, 7, 21, 22] and references therein for a deeper presentation. Let $\mathcal{T}_h$ be a (regular) grid on the domain Ω. Call h_K the diameter of each element K and h the maximum of all the diameters. With the usual notation, now let $\boldsymbol{\Theta}_h \subset \boldsymbol{\Theta}$, $W_h \subset W$, $\boldsymbol{\Gamma}_h \subset [L^2(\Omega)]^2$ be the finite element spaces adopted; then, the linked interpolation method is based on the introduction of a linear and uniformly bounded operator $L : \boldsymbol{\Theta}_h \to W$.

The final form of the discrete problem takes the form

(4.1)
$$\begin{cases} \text{Find } ((\boldsymbol{\theta}_h, w_h), \boldsymbol{\gamma}_h) \in \boldsymbol{\Theta}_h \times W_h \times \boldsymbol{\Gamma}_h \text{ such that} \\ a(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) - (\boldsymbol{\gamma}_h, \boldsymbol{\eta}_h - \boldsymbol{\nabla}(v_h + L\boldsymbol{\eta}_h)) = (g, v_h + L\boldsymbol{\eta}_h), & (\boldsymbol{\eta}_h, v_h) \in \boldsymbol{\Theta}_h \times W_h, \\ \lambda^{-1}t^2(\boldsymbol{\gamma}_h, \boldsymbol{\delta}_h) + (\boldsymbol{\theta}_h - \boldsymbol{\nabla}(w_h + L\boldsymbol{\theta}_h), \boldsymbol{\delta}_h) = 0, & \boldsymbol{\delta}_h \in \boldsymbol{\Gamma}_h. \end{cases}$$

Then, for all $(\boldsymbol{\theta}, w, \boldsymbol{\gamma}) \in [H^1(\Omega)]^2 \times H^1(\Omega) \times [L^2(\Omega)]^2$, we define the norms

$$\|\boldsymbol{\gamma}\|_h^2 := \sum_{K\in\mathcal{T}_h} h_K^2\|\boldsymbol{\gamma}\|_{0,K}, \tag{4.2}$$

$$|\!|\!|\boldsymbol{\theta}, w, \boldsymbol{\gamma}|\!|\!|^2 := \|\boldsymbol{\theta}\|_1^2 + \|w\|_1^2 + \|\boldsymbol{\gamma}\|_h^2 + t^2\|\boldsymbol{\gamma}\|_0^2. \tag{4.3}$$

Assume also that the following two properties hold for the discretization spaces:

$$\boldsymbol{\nabla} W_h \subset \boldsymbol{\Gamma}_h, \tag{4.4}$$

$$\sup_{(\boldsymbol{\eta}_h, v_h)\in\boldsymbol{\Theta}_h\times W_h} \frac{(\boldsymbol{\delta}_h, \boldsymbol{\eta}_h - \boldsymbol{\nabla}(\boldsymbol{\eta}_h + Lv_h))}{\|\boldsymbol{\eta}_h\|_1 + \|v_h\|_1} \ge \beta\|\boldsymbol{\delta}_h\|_h \quad \forall \boldsymbol{\delta}_h \in \boldsymbol{\Gamma}_h, \tag{4.5}$$

where β is a positive constant independent of h.

Under this hypothesis, optimal error estimates (with respect to the order of the polynomial spaces used) in the $|||\cdot|||$ norm are proved in [6]; the results are shown in the clamped plate case but can be immediately extended to hard simply supported and free plates. The extension to our new free plate model is instead to be handled with some care; the projection operator which is hidden in the linked formulation is not suitable to directly enforce $\theta_s = w_{/s}$ on Σ_f in a relaxed way as done in (3.4). On the other hand, the particular approximation properties of this finite element allow us to enforce the free boundary condition by penalization.

4.1. A new discrete formulation. We start introducing the symmetric bilinear form (see (2.4)–(2.5)),

$$\begin{aligned} C(\boldsymbol{\theta}, w; \boldsymbol{\eta}, v) = &\sum_{e\in\Sigma_f} \int_e M_{ns}[\boldsymbol{\theta}](\boldsymbol{\eta} - \boldsymbol{\nabla} v)_s \,\mathrm{d}s + \int_e (\boldsymbol{\theta} - \boldsymbol{\nabla} w)_s M_{ns}[\boldsymbol{\eta}] \,\mathrm{d}s \\ &+ \frac{1}{|\ e\ |} \int_e (\boldsymbol{\theta} - \boldsymbol{\nabla} w)_s (\boldsymbol{\eta} - \boldsymbol{\nabla} v)_s \,\mathrm{d}s, \end{aligned} \tag{4.6}$$

which in particular is well defined on the discrete space $\boldsymbol{\Theta}_h \times W_h$.

We now remark that L can be extended to all $\boldsymbol{\Theta}$ by composition with the L^2 projection on $\boldsymbol{\Theta}_h$. Then, starting from (4.3), we can define the norm

$$|||\boldsymbol{\eta}, v, \boldsymbol{\delta}|||_*^2 := |||\boldsymbol{\eta}, v, \boldsymbol{\delta}|||^2 + \sum_{e\in\Sigma_f} \frac{1}{|\ e\ |} \int_e (\boldsymbol{\eta} - \boldsymbol{\nabla}(v + L\boldsymbol{\eta}))_s^2 \,\mathrm{d}s, \tag{4.7}$$

defined for all $(\boldsymbol{\eta}, v, \boldsymbol{\delta})$ in the discrete space and, for instance, for all $(\boldsymbol{\eta}, v, \boldsymbol{\delta}) \in [H^1(\Omega)]^2 \times H^2(\Omega) \times [L^2(\Omega)]^2$.

The condition $\theta_s = w_{/s}$ on Σ_f will be enforced through the addition of the bilinear form (4.6) to the original discrete plate problem. In simple words, the first term is introduced to reach consistency with the continuous problem, the second to obtain a symmetric form, and the third to add stability.

Our discrete problem is therefore

(4.8)
$$\begin{cases} \text{Find } ((\boldsymbol{\theta}_h, w_h), \boldsymbol{\gamma}_h) \in \boldsymbol{\Theta}_h \times W_h \times \boldsymbol{\Gamma}_h \text{ such that} \\ a(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) + C(\boldsymbol{\theta}_h, w_h + L\boldsymbol{\theta}_h; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) - (\boldsymbol{\gamma}_h, \boldsymbol{\eta}_h - \boldsymbol{\nabla}(v_h + L\boldsymbol{\eta}_h)) \\ \qquad\qquad = (g, v_h + L\boldsymbol{\eta}_h) \qquad \forall (\boldsymbol{\eta}_h, v_h) \in \boldsymbol{\Theta}_h \times W_h, \\ \lambda^{-1} t^2 (\boldsymbol{\gamma}_h, \boldsymbol{\delta}_h) + (\boldsymbol{\theta}_h - \boldsymbol{\nabla}(w_h + L\boldsymbol{\theta}_h), \boldsymbol{\delta}_h) = 0 \qquad \forall \boldsymbol{\delta}_h \in \boldsymbol{\Gamma}_h. \end{cases}$$

Letting

$$\begin{aligned} &A(\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h; v_h, \boldsymbol{\eta}_h, \boldsymbol{\delta}_h) \\ &\quad := a(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) - (\boldsymbol{\gamma}_h, \boldsymbol{\eta}_h - \boldsymbol{\nabla} v_h) + (\boldsymbol{\theta}_h - \boldsymbol{\nabla} w_h, \boldsymbol{\delta}_h) + \lambda^{-1} t^2 (\boldsymbol{\gamma}_h, \boldsymbol{\delta}_h), \end{aligned} \tag{4.9}$$

problem (4.8) is equivalent to

(4.10)
$$\begin{cases} \text{Find } (\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h) \in \boldsymbol{\Theta}_h \times W_h \times \boldsymbol{\Gamma}_h \text{ such that} \\ A(\boldsymbol{\theta}_h, w_h + L\boldsymbol{\theta}_h, \boldsymbol{\gamma}_h; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h, \boldsymbol{\delta}_h) + C(\boldsymbol{\theta}_h, w_h + L\boldsymbol{\theta}_h; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) \\ \qquad\qquad = (g, v_h + L\boldsymbol{\eta}_h) \quad \forall (\boldsymbol{\eta}_h, v_h, \boldsymbol{\delta}_h) \in \boldsymbol{\Theta}_h \times W_h \times \boldsymbol{\Gamma}_h. \end{cases}$$

We have the following lemma.

LEMMA 4.1. *Let* $\boldsymbol{\theta}, w, \boldsymbol{\gamma}$ *be the solution of the continuous problem of section* 2 ((2.1)–(2.3) *plus boundary conditions* (2.7), (2.8), *and* (2.11)). *Then, it holds that*

$$A(\boldsymbol{\theta}, w, \boldsymbol{\gamma}; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h, \boldsymbol{\delta}_h) + C(\boldsymbol{\theta}, w; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) = (g, v_h + L\boldsymbol{\eta}_h) \tag{4.11}$$

for all $(\boldsymbol{\eta}_h, v_h, \boldsymbol{\delta}_h) \in \boldsymbol{\Theta}_h \times W_h \times \boldsymbol{\Gamma}_h$.

Proof. The proof follows by standard variational means, testing (2.1)–(2.3), respectively, on $\boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h, \boldsymbol{\delta}_h$, then integrating by parts and recalling the boundary conditions (2.7), (2.8), and (2.11). Because the finite element space used does not satisfy a priori the condition on Σ_f, a nonconforming term appears; but this is taken care of by the additional bilinear form C introduced. We note that the regularity of w and $\boldsymbol{\theta}$ (see (2.19)) is sufficient to ensure the well posedness of the form C in the variational equation (4.11). ∎

The following proposition, which states the uniform invertibility of the discrete problem, is fundamental for the following estimates. Its proof relies strongly on the inf-sup condition (4.5) and adopts the same technique of [6, 19, 15]; therefore we will not include it here.

PROPOSITION 4.2. *There exists a positive constant* k *independent of* h *such that for any* $(\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h) \in \boldsymbol{\Theta}_h \times W_h \times \boldsymbol{\Gamma}_h$ *there exists* $(\boldsymbol{\eta}_h, v_h, \boldsymbol{\delta}_h) \in \boldsymbol{\Theta}_h \times W_h \times \boldsymbol{\Gamma}_h$, *giving*

$$\begin{aligned} &A(\boldsymbol{\theta}_h, w_h + L\boldsymbol{\theta}_h, \boldsymbol{\gamma}_h; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h, \boldsymbol{\delta}_h) + C(\boldsymbol{\theta}_h, w_h + L\boldsymbol{\theta}_h; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) \\ &\geq |||\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h|||_*, \end{aligned} \tag{4.12}$$

$$|||\boldsymbol{\eta}_h, v_h, \boldsymbol{\delta}_h|||_* \leq k. \tag{4.13}$$

For any $\boldsymbol{\eta}$ sufficiently regular, we indicate (here and in what follows) by $\boldsymbol{\eta}_{II}$ the interpolant of $\boldsymbol{\eta}$ in $\boldsymbol{\Theta}_h$. Analogously, v_I will indicate the interpolant of v in W_h and $\boldsymbol{\delta}_*$ the interpolant of $\boldsymbol{\delta}$ in $\boldsymbol{\Gamma}_h$.

We will assume, in addition to (4.4) and (4.5), that the property

$$\boldsymbol{\gamma}_s = 0 \text{ on } e \Longrightarrow (L\boldsymbol{\gamma}_{II})_s = 0 \text{ on } e \tag{4.14}$$

holds for all boundary edges e. We remark that, being already implicit in the linking philosophy, this condition is not restrictive; for example, it holds for all the linked elements presented in [6, 7, 8], using the natural interpolant in $\boldsymbol{\Theta}_h$.

Now, following the mainstream of [6], we introduce the operator

$$\Pi v := v_I + L(\boldsymbol{\nabla} v)_{II} \qquad \forall v \text{ sufficiently regular.} \tag{4.15}$$

We can finally state the main result.

PROPOSITION 4.3. *Let* $(\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h)$ *be the solution of problem* (4.8) *and let* $(\boldsymbol{\theta}, w, \boldsymbol{\gamma})$ *be the solution of the continuous problem of section* 2 *((2.1)–(2.3) plus boundary conditions* (2.7), (2.8), *and* (2.11)). *Then, we have the following estimate:*

$$\begin{aligned} &|||\boldsymbol{\theta} - \boldsymbol{\theta}_h, w - w_h, \boldsymbol{\gamma} - \boldsymbol{\gamma}_h|||_* \\ &\quad \leq c[h\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_1 + h^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_0 + \|w - \Pi w\|_2 \\ &\qquad + h^{-1}\|w - \Pi w\|_1 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_*\|_{H^{-1}(\mathrm{div})} + t\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_*\|_0 + t\|\boldsymbol{\nabla} L\gamma_{II}\|_0]. \end{aligned} \tag{4.16}$$

Proof. We start applying Proposition 4.2 to $(\boldsymbol{\theta}_h - \boldsymbol{\theta}_{II}, w_h - w_I, \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_*)$ and find

$(\boldsymbol{\eta}_h, v_h, \boldsymbol{\delta}_h)$ in the discrete space such that

$$
\begin{aligned}
&A(\boldsymbol{\theta}_h - \boldsymbol{\theta}_{II}, w_h - w_I + L(\boldsymbol{\theta}_h - \boldsymbol{\theta}_{II}), \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_*; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h, \boldsymbol{\delta}_h) \\
&\qquad + C(\boldsymbol{\theta}_h - \boldsymbol{\theta}_{II}, w_h - w_I + L(\boldsymbol{\theta}_h - \boldsymbol{\theta}_{II}); \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) \\
&\geq |\!|\!| \boldsymbol{\theta}_h - \boldsymbol{\theta}_{II}, w_h - w_I, \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_* |\!|\!|_*, \qquad (4.17) \\
&|\!|\!| \boldsymbol{\eta}_h, v_h, \boldsymbol{\delta}_h |\!|\!|_* \leq k. \qquad (4.18)
\end{aligned}
$$

Adding and subtracting as usual

$$
A(\boldsymbol{\theta}, w, \boldsymbol{\gamma}; v_h + L\boldsymbol{\eta}_h, \boldsymbol{\eta}_h, \boldsymbol{\delta}_h) + C(\boldsymbol{\theta}, w; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) \qquad (4.19)
$$

in (4.17), and then using the error equation given by the difference of (4.10) and (4.11), we obtain in the end

$$
\begin{aligned}
&|\!|\!| \boldsymbol{\theta}_h - \boldsymbol{\theta}_{II}, w_h - w_I, \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_* |\!|\!|_* \\
&\quad \leq A(\boldsymbol{\theta} - \boldsymbol{\theta}_{II}, w - w_I - L\boldsymbol{\theta}_{II}, \boldsymbol{\gamma} - \boldsymbol{\gamma}_*; v_h + L\boldsymbol{\eta}_h, \boldsymbol{\eta}_h, \boldsymbol{\delta}_h) \\
&\qquad + C(\boldsymbol{\theta} - \boldsymbol{\theta}_{II}, w - w_I - L\boldsymbol{\theta}_{II}; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h). \qquad (4.20)
\end{aligned}
$$

Recalling definition (4.9), and using the Cauchy–Schwarz inequality, the continuity of $a(\cdot,\cdot)$, and the bound (4.18), we can derive by standard means,

$$
\begin{aligned}
&A(\boldsymbol{\theta} - \boldsymbol{\theta}_{II}, w - w_I - L\boldsymbol{\theta}_{II}, \boldsymbol{\gamma} - \boldsymbol{\gamma}_*; v_h + L\boldsymbol{\eta}_h, \boldsymbol{\eta}_h, \boldsymbol{\delta}_h) \leq c[\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_1 \\
&+ \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_*\|_{H^{-1}(\mathrm{div})} + t\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_*\|_0 + (\boldsymbol{\delta}_h, \boldsymbol{\theta} - \boldsymbol{\theta}_{II} - \boldsymbol{\nabla}(w - w_I - L\boldsymbol{\theta}_{II}))]. \qquad (4.21)
\end{aligned}
$$

To bound the last term, we argue as in [6], and we refer there for the details. Starting from (2.3), it is immediate to see that

$$
\boldsymbol{\nabla} L\boldsymbol{\theta}_{II} = \boldsymbol{\nabla} L(\boldsymbol{\nabla} w)_{II} + \lambda^{-1} t^2 \boldsymbol{\nabla} L\gamma_{II}. \qquad (4.22)
$$

We now substitute (4.22) into the last member of (4.21) and use the Cauchy–Schwarz inequality in L^2. Observing that, due to (4.18), $\|\boldsymbol{\delta}_h\|_h + t\|\delta_h\|_0$ is bounded, and recalling (4.15), we obtain

$$
(\boldsymbol{\delta}_h, \boldsymbol{\theta} - \boldsymbol{\theta}_{II} - \boldsymbol{\nabla}(w - w_I - L\boldsymbol{\theta}_{II})) \leq c(h^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_0 + h^{-1}\|w - \Pi w\|_1 + t\|\boldsymbol{\nabla} L\gamma_{II}\|_0). \qquad (4.23)
$$

To estimate the right-hand side of (4.20), we still need to bound the free boundary part (defined in (4.6)),

$$
\begin{aligned}
&C(\boldsymbol{\theta} - \boldsymbol{\theta}_{II}, w - w_I - L\boldsymbol{\theta}_{II}; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) \\
&\quad = \sum_{e \in \Sigma_f} \int_e M_{ns}[\boldsymbol{\theta} - \boldsymbol{\theta}_{II}](\boldsymbol{\eta}_h - \boldsymbol{\nabla}(v_h + L\boldsymbol{\eta}_h))_s \, \mathrm{d}s \\
&\qquad + \int_e (\boldsymbol{\theta} - \boldsymbol{\theta}_{II} - \boldsymbol{\nabla}(w - w_I - L\boldsymbol{\theta}_{II}))_s M_{ns}[\boldsymbol{\eta}_h] \, \mathrm{d}s \\
&\qquad + \frac{1}{|\, e \,|} \int_e (\boldsymbol{\theta} - \boldsymbol{\theta}_{II} - \boldsymbol{\nabla}(w - w_I - L\boldsymbol{\theta}_{II}))_s (\boldsymbol{\eta}_h - \boldsymbol{\nabla}(v_h + L\boldsymbol{\eta}_h))_s \, \mathrm{d}s. \qquad (4.24)
\end{aligned}
$$

We start observing that, recalling the definition of M_{ns} in (2.5) and the Agmon inequality (3.31), we have

$$
\begin{aligned}
\sum_{e \in \Sigma_f} \|M_{ns}[\boldsymbol{\theta} - \boldsymbol{\theta}_{II}]\|^2_{0,e} &\leq c \sum_{K \in \mathcal{T}_h} h^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|^2_{1,T} + h\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|^2_{2,T} \\
&\leq h^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|^2_{1,\Omega} + h\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|^2_{2,\Omega}, \qquad (4.25)
\end{aligned}
$$

while, from the definition (4.7) and bound (4.18), it follows that

$$\sum_{e\in\Sigma_f} \|(\boldsymbol{\eta}_h - \boldsymbol{\nabla}(v_h + L\boldsymbol{\eta}_h))_s\|_{0,e}^2$$
$$\leq h \sum_{e\in\Sigma_f} \frac{1}{|\ e\ |}\|(\boldsymbol{\eta}_h - \boldsymbol{\nabla}(v_h + L\boldsymbol{\eta}_h))_s\|_{0,e}^2 \leq h\|\!|\boldsymbol{\eta}_h, v_h, \boldsymbol{\delta}_h|\!\|_*^2 \leq c\,h. \tag{4.26}$$

Applying the Cauchy–Schwarz inequality (first in $L^2(e)$, then in l^2), we can bound the first term in the right-hand side of (4.24) with the product of the square roots of (4.25) and (4.26). Summing over all $e \in \Sigma_f$, this easily gives

$$\sum_{e\in\Sigma_f} \int_e M_{ns}[\boldsymbol{\theta} - \boldsymbol{\theta}_{II}](\boldsymbol{\eta}_h - \boldsymbol{\nabla}(v_h + L\boldsymbol{\eta}_h))_s\,\mathrm{d}s \leq c[\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_1 + h\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_2]. \tag{4.27}$$

From (4.22) and the definition of Π in (4.15) it follows that

$$\|(\boldsymbol{\theta} - \boldsymbol{\theta}_{II} - \boldsymbol{\nabla}(w - w_I - L\boldsymbol{\theta}_{II}))_s\|_{0,e} \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_{0,e} + |\ w - \Pi w\ |_{1,e} + \lambda^{-1}t^2\|(L\boldsymbol{\gamma}_{II})_s\|_{1,e}, \tag{4.28}$$

where $e \in \Sigma_f$.

Due to the boundary condition (2.11) and property (4.14), the last term in (4.28) is null. Therefore, using the Agmon inequality on each $e \in \Sigma_f$ and then summing over all $K \in \mathcal{T}_h$, we have

$$\sum_{e\in\Sigma_f} \|(\boldsymbol{\theta} - \boldsymbol{\theta}_{II} - \boldsymbol{\nabla}(w - w_I - L\boldsymbol{\theta}_{II}))_s\|_{0,e}^2$$
$$\leq c[h^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_{0,\Omega}^2 + h\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_{1,\Omega}^2 + h^{-1}\|w - \Pi w\|_{1,\Omega}^2 + h\|w - \Pi w\|_{2,\Omega}^2]. \tag{4.29}$$

Using again the Agmon inequality, classical inverse inequalities for piecewise polynomials and then bound (4.18), we easily get

$$\sum_{e\in\Sigma_f} \|M_{ns}[\boldsymbol{\eta}_h]\|_{0,e}^2 \leq ch^{-1} \sum_{K\in\mathcal{T}_h} \|\boldsymbol{\eta}_h\|_{1,K}^2$$
$$\leq ch^{-1}\|\boldsymbol{\eta}_h\|_{1,\Omega}^2 \leq ch^{-1}\|\!|\boldsymbol{\eta}_h, v_h, \boldsymbol{\delta}_h|\!\|_*^2 \leq ch^{-1}. \tag{4.30}$$

Using (4.29) and (4.30) and the Cauchy–Schwarz inequality (as done in (4.27)), we can then bound the second term in the right-hand side of (4.24):

$$\sum_{e\in\Sigma_f} \int_e (\boldsymbol{\theta} - \boldsymbol{\theta}_{II} - \boldsymbol{\nabla}(w - w_I - L\boldsymbol{\theta}_{II}))_s M_{ns}[\boldsymbol{\eta}_h]\,\mathrm{d}s$$
$$\leq c[h^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_0 + \|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_1 + h^{-1}\|w - \Pi w\|_1 + \|w - \Pi w\|_2]. \tag{4.31}$$

The last member of (4.24) can be bounded in a similar way using (4.26) and (4.29); combining (4.27) and (4.31) with this last bound, we finally obtain from (4.24)

$$C(\boldsymbol{\theta} - \boldsymbol{\theta}_{II}, w - w_I - L\boldsymbol{\theta}_{II}; \boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h)$$
$$\leq c[\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_1 + h\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_2 + h^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{II}\|_0 + h^{-1}\|w - \Pi w\|_1 + \|w - \Pi w\|_2]. \tag{4.32}$$

The proof follows by the triangle inequality and combining estimates (4.21), (4.23), and (4.32). □

Proposition 4.3 grants uniform and optimal error estimates for most of the linked elements in the literature applied to our new free boundary model.

Assume that properties (4.4), (4.5), and (4.14) are true, and that, for $k \leq 1$,

- the interpolant $\boldsymbol{\theta}_{II}$ is P_k invariant,
- the interpolant w_I is P_k invariant, and
- the interpolant $\boldsymbol{\gamma}_*$ is P_{k-1} invariant.

Assume also that, for the linear and uniformly bounded operator $L : \boldsymbol{\Theta}_h \to W$, it holds that

- the application Π (see (4.15)) is P_{k+1} invariant and
- $\mid L\boldsymbol{\delta} \mid_1 = O(h^k)$ for all sufficiently regular $\boldsymbol{\delta}$.

Then, it is easy to obtain from Proposition 4.3 and standard polynomial interpolation that, for all $1 \leq s \leq k$,

$$\begin{aligned} &\|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_1 + \|w - w_h\|_1 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_h + t^2\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_0 \\ &\qquad \leq ch^s(\,\|\,\boldsymbol{\varphi}\,\|_{s+1} + \|\,w\,\|_{s+2} + t\|\boldsymbol{\gamma}\|_s + \|\boldsymbol{\gamma}\|_{s-1} + \|\mathrm{div}\boldsymbol{\gamma}\|_{s-1}\,), \end{aligned} \tag{4.33}$$

which, remembering (2.19) and from Remark 6, gives an order of convergence of up to order $\min(3/2, k)$ on the full domain.

Remark 5. The last two conditions above may seem less natural, but they are fundamental requirements for any linked interpolation method and follow in each particular case from the form of the operator L. For example, all the above requirements are met by the triangular elements of [7] with $k = 1$, by the quadrilateral elements of [8] with $k = 1$, and by the higher order triangular elements of [6] with $k = 2$.

Remark 6. We note that for these elements, the higher order (but realistic) regularity (2.20) is required for w.

Remark 7. A similar method was applied and numerically tested in [16] for the PSRI plate elements.

REFERENCES

[1] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, NJ, 1965.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Discontinuous Galerkin methods for elliptic problems*, in Discontinuous Galerkin Methods, Lecture Notes in Comput. Sci. Engrg. 11, Springer-Verlag, Berlin, 2000, pp. 89–101.

[3] D. N. Arnold, F. Brezzi, and L. D. Marini, *New elements for the Reissner-Mindlin plates*, J. Sci. Comput., to appear.

[4] D. N. Arnold and R. S. Falk, *Asymptotic analysis of the boundary layer for the Reissner–Mindlin plate model*, SIAM J. Math. Anal., 27 (1996), pp. 486–514.

[5] D. N. Arnold and R. S. Falk, *A uniformly accurate finite element method for the Reissner–Mindlin plate*, SIAM J. Numer. Anal., 26 (1989), pp. 1276–1290.

[6] F. Auricchio and C. Lovadina, *Analysis of kinematic linked interpolation methods for Reissner-Mindlin plate problems*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2465–2482.

[7] F. Auricchio and R. L. Taylor, *Linked interpolation for Reissner-Mindlin plate elements* II: *A simple triangle*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 3057–3066.

[8] F. Auricchio and R. L. Taylor, *A shear deformable plate element with an exact thin limit*, Comput. Methods Appl. Mech. Engrg., 118 (1994), pp. 393–412.

[9] L. Beirão da Veiga and F. Brezzi, *Reissner-Mindlin plates with free boundary conditions*, Rend. Acad. Lincei, to appear.

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

[11] K. J. Bathe, F. Brezzi, and M. Fortin, *Mixed-interpolated elements for the Reissner-Mindlin plates*, Internat. J. Numer. Methods Engrg., 28 (1989), pp. 1787–1801.

[12] F. Brezzi, M. Fortin, and R. Stenberg, *Error analysis of mixed-interpolated elements for Reissner-Mindlin plates*, Math. Models Methods Appl. Sci., 1 (1991), pp. 125–151.
[13] F. Brezzi and L. D. Marini, *A nonconforming element for the Reissner-Mindlin plate*, Comput. & Structures, 81 (2003), pp. 515–522.
[14] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[15] D. Chapelle and R. Stenberg, *An optimal low-order locking-free finite element method for Reissner-Mindlin plates*, Math. Models Methods Appl. Sci., 8 (1998), pp. 407–430.
[16] C. Chinosi, *PSRI elements for the Reissner-Mindlin free plate*, to appear.
[17] R. G. Duran, E. Hernández, L. Hervella-Nieto, E. Liberman, and R. Rodríguez, *Error estimates for low-order isoparametric quadrilateral finite elements for plates*, SIAM J. Numer. Anal., 41 (2003), pp. 1751–1772.
[18] R. G. Duran and E. Liberman, *On mixed finite element methods for the Reissner-Mindlin plate model*, Math. Comp., 58 (1992), pp. 561–573.
[19] C. Lovadina, *Analysis of a mixed finite element method for Reissner-Mindlin plate problem*, Comput. Methods Appl. Mech. Engrg., 163 (1998), pp. 71–85.
[20] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, Springer-Verlag, New York, 1975, pp. 219–315.
[21] O. C. Zienkiewicz, Z. Xu, L. F. Zeng, A. Samuelsson, and N.-E. Wideberg, *Linked interpolation for Reissner-Mindlin plate elements* II: *A simple quadrilateral*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 3043–3056.
[22] Z. Xu, *A thick-thin triangular plate element*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 963–973.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1592–1609

A TIME-DOMAIN FINITE ELEMENT METHOD FOR MAXWELL'S EQUATIONS*

TRI VAN† AND AIHUA WOOD‡

Abstract. Presented here is a time-domain finite element method for approximating Maxwell's equations. The problem is to approximate the electromagnetic fields scattered by a bounded, inhomogeneous cavity embedded in an infinite ground plane. The time-dependent scattering problem is first discretized in time by Newmark's time-stepping scheme. The resulting semidiscrete problem is proved to be well posed. A nonlocal boundary condition on the cavity aperture is constructed to reduce the computational domain to the cavity itself. Stability analysis and error estimates of the fully discrete problem are provided.

Key words. Maxwell's equations, nonlocal boundary condition, time-domain finite element methods, Newmark scheme, stability, cavity

AMS subject classifications. 65M60, 74J20

DOI. 10.1137/S0036142901387427

1. Introduction.

Time-harmonic (frequency-domain) Maxwell's equations for scattering problems are well studied and documented ([1, 19, 20, 21], to name a few) compared to their time-domain counterparts. This is due, to a large extent, to their obvious advantage of the absence of the time dependence and the limitations of computer power. The recent rapid advances in computing technology have prompted a growing popularity of numerical schemes for simulating electromagnetic transients (time-domain) for their potential to generate wide-band data and model nonlinear materials. Reports of new and faster numerical techniques for electromagnetic analysis have flourished in the engineering literature (see, for example, [18, 23, 12, 25]). However, very little analysis is known in the open literature. To the authors' knowledge, the first mathematical study of time-domain Maxwell's equations for scattering by a bounded perfectly electric conducting (PEC) body was reported in [22], in which a spatially discretized problem is analyzed. More recently, in [4, 5], fully discrete finite element methods for solving Maxwell's equations of bounded PEC scattering bodies are considered. In both cases, the problem is defined in a bounded domain. This paper serves as our first attempt to understand the various stability and convergence issues associated with the finite element method for modelling transient electromagnetic scattering from non-PEC bodies. The problem is defined in an infinite space.

As observed in [7], most numerical schemes for scattering problems are faced with the problem of truncating the infinite domain to a bounded computational domain

*Received by the editors February 22, 2002; accepted for publication (in revised form) September 15, 2003; published electronically December 16, 2004. This research was supported in part by Air Force Office of Scientific Research grant AFOSR-PO-990025 and by a grant from the Joint Air Force Research Lab/Dayton Area Graduate Studies Institute (AFRL/DAGSI) Research Program. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

http://www.siam.org/journals/sinum/42-4/38742.html

†Mission Research Corporation, 3975 Research Blvd., Dayton, OH 45430 (tvan@mrcday.com).

‡Air Force Institute of Technology, 2950 P St., AFIT/ENC, Wright-Patterson AFB, OH 45433-7765 (aihua.wood@afit.af.mil).

without introducing excessive error. This is usually achieved by introducing an artificial boundary with appropriately defined boundary conditions that reduce the reflection of waves incident on this artificial boundary. Among the more popular are the so-called absorbing boundary conditions (ABC), perfectly matched layer (PML), and exact nonreflecting boundary conditions (NRBC) (for a survey of nonreflecting boundary conditions see [14]). An NRBC is nonlocal in both time and space. The nonlocality of NRBC in time requires the storing of the entire history of the solution on the artificial boundary and can be very expensive for large problems. In [15], the numerical methods for solving three-dimensional acoustic and elastic wave problems are derived based on the Kirchhoff-type boundary condition which is nonlocal in both time and space. For electromagnetic scattering problems, exact NRBCs are developed in [16] by using vector spherical harmonic series expansions for the spherical artificial boundary. In this formulation the computational domain can be unnecessarily large when the scatterer geometry is highly irregular, such as a thin and long body. In this paper, we present a finite element/Fourier transform method for analyzing the transient electromagnetic scattering from three-dimensional cavities embedded in the infinite PEC ground plane. Contrary to the conventional approach, we first discretize the problem in time by the Newmark scheme. At each time step, a spatially nonlocal boundary condition is constructed right at the cavity aperture Γ so that the computational domain is reduced to a minimum: the cavity itself. The finite element matrix is precomputed outside of the time loop and is reused at each time step, significantly reducing storage and computational time. Since the memory requirement does not increase with time, one is able to numerically solve larger problems.

The next section describes the model problem. Section 3 demonstrates the time-discretization of the cavity problem by using the Newmark time-stepping scheme. This procedure yields the so-called semidiscrete problem, which is analytically solved in the exterior of the cavity by using Fourier transform. The exterior solution is used to construct a nonlocal boundary condition, which is then used to reduce the semidiscrete problem to a finite domain (the cavity itself). Section 4 establishes the existence and uniqueness of a weak solution of the semidiscrete problem. The semidiscrete problem is fully discretized in section 5 by the curl-conforming finite elements. Motivated by the work of Baker [2] and Cowsar, Dupont, and Wheeler [11] for second-order hyperbolic equations, stability of the time-marching scheme is analyzed. Convergence properties of the finite element method are also addressed. The paper is concluded in section 6.

2. Mathematical formulation. Let Ω denote a bounded Lipschitz domain (cavity) in $\mathbb{R}^3$ such that $\Omega \subset \{\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 \leq 0\}$. The relative electric permittivity of Ω is characterized by the bounded positive function $\varepsilon_r = \varepsilon/\varepsilon_0$. It is assumed that the medium throughout is nonmagnetic, thus, $\mu_r = 1$. The cavity aperture is defined by the surface $\Gamma = \{\boldsymbol{x} \in \partial\Omega : x_3 = 0\}$. We are interested in the scattering properties of the cavity-backed aperture in an infinite ground plane. The infinite ground plane excluding the cavity is assumed to be PEC, hence, the tangential components of the total field and the scattered field vanish there. An incident electromagnetic field $(\boldsymbol{E}^i, \boldsymbol{H}^i)$ is assumed to be the classical solution of the time-domain Maxwell's equations in the free space

$$\begin{cases} \operatorname{curl} \boldsymbol{E}^i(\boldsymbol{x}, t) &= -\dfrac{\partial}{\partial t}\boldsymbol{H}^i(\boldsymbol{x}, t) \ \text{ in } \mathbb{R}^3 \times (0, T), \\ \operatorname{curl} \boldsymbol{H}^i(\boldsymbol{x}, t) &= \dfrac{\partial}{\partial t}\boldsymbol{E}^i(\boldsymbol{x}, t) \ \text{ in } \mathbb{R}^3 \times (0, T), \end{cases}$$

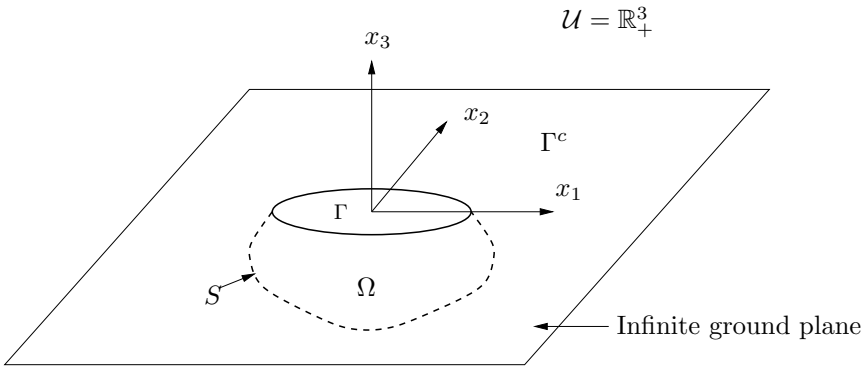


FIG. 2.1. *Cavity Ω in the infinite ground plane.*

where $0 < T < \infty$. The incident field interacts with the cavity to produce the total field $(\boldsymbol{E}, \boldsymbol{H})$ which satisfies the initial-value problem

$$\text{(2.1)} \qquad \begin{cases} \operatorname{curl} \boldsymbol{E}(\boldsymbol{x},t) &= -\dfrac{\partial}{\partial t}\boldsymbol{H}(\boldsymbol{x},t) \ \text{ in } \mathbb{R}^3 \times (0,T), \\ \operatorname{curl} \boldsymbol{H}(\boldsymbol{x},t) &= \varepsilon_r(\boldsymbol{x})\dfrac{\partial}{\partial t}\boldsymbol{E}(\boldsymbol{x},t) \ \text{ in } \mathbb{R}^3 \times (0,T), \\ \boldsymbol{E}(\boldsymbol{x},0) = \boldsymbol{E}_0 \quad &\text{and} \quad \boldsymbol{H}(\boldsymbol{x},0) = \boldsymbol{H}_0, \end{cases}$$

where $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ are initial conditions with compact support. Let $(\boldsymbol{E}^r, \boldsymbol{H}^r)$ denote the field reflected by the ground plane in the absence of the scatterer. The scattered field $(\boldsymbol{E}^s, \boldsymbol{H}^s)$ is defined by the differences

$$\boldsymbol{E}^s = \boldsymbol{E} - (\boldsymbol{E}^i + \boldsymbol{E}^r), \quad \boldsymbol{H}^s = \boldsymbol{H} - (\boldsymbol{H}^i + \boldsymbol{H}^r).$$

Let us denote S as the cavity walls, $\Gamma^c = \{x \in \mathbb{R}^3 : x_3 = 0\} \setminus \Gamma$, and $\mathcal{U}$ as the upper half space $\{\boldsymbol{x} \in \mathbb{R}^3 : x_3 > 0\}$. The domain of interest is $\mathcal{G} := \Omega \cup \mathcal{U}$. The surface $\partial\mathcal{G}$ is perfectly conducting, where $\partial\mathcal{G} = S \cup \Gamma^c$ (see Figure 2.1). Inside the cavity Ω we solve for the total field $\boldsymbol{E}$, while in the upper half space $\mathcal{U}$ we only need to solve for the scattered field $\boldsymbol{E}^s$ since both $\boldsymbol{E}^i$ and $\boldsymbol{E}^r$ are known in $\mathcal{U}$. We eliminate the magnetic field $\boldsymbol{H}$ in (2.1) to obtain a transmission-type problem,

$$\text{(2.2)} \qquad \begin{cases} \operatorname{curl}\operatorname{curl} \boldsymbol{E}(\boldsymbol{x},t) + \varepsilon_r \dfrac{\partial^2}{\partial t^2}\boldsymbol{E}(\boldsymbol{x},t) &= 0 \ \text{ in } \Omega \times (0,T), \\ \operatorname{curl}\operatorname{curl} \boldsymbol{E}^s(\boldsymbol{x},t) + \dfrac{\partial^2}{\partial t^2}\boldsymbol{E}^s(\boldsymbol{x},t) &= 0 \ \text{ in } \mathcal{U} \times (0,T), \end{cases}$$

with the transmission conditions

$$\begin{aligned} \hat{n} \times \boldsymbol{E} &= 0 \ \text{ on } S \times (0,T), \\ \hat{n} \times \boldsymbol{E}^s &= 0 \ \text{ on } \Gamma^c \times (0,T), \\ \hat{n} \times \boldsymbol{E} &= \hat{n} \times (\boldsymbol{E}^s + \boldsymbol{E}^i + \boldsymbol{E}^r) \ \text{ on } \Gamma \times (0,T), \\ \hat{n} \times \operatorname{curl} \boldsymbol{E} &= \hat{n} \times \operatorname{curl}(\boldsymbol{E}^s + \boldsymbol{E}^i + \boldsymbol{E}^r) \ \text{ on } \Gamma \times (0,T) \end{aligned}$$

and the initial conditions

$$\boldsymbol{E}(\boldsymbol{x},0) = \boldsymbol{E}_0(\boldsymbol{x}), \quad \partial_t \boldsymbol{E}(\boldsymbol{x},0) = \boldsymbol{E}_1(\boldsymbol{x}) = \operatorname{curl} \boldsymbol{H}_0(\boldsymbol{x})/\varepsilon_r(\boldsymbol{x}),$$

where $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ are the given electric and magnetic fields at time $t = 0$. In the rest of the paper, we wish to numerically solve the time-dependent problem (2.2).

3. Semidiscrete problem. In this section, we shall first discretize the model problem (2.2) in time using Newmark's time-marching scheme [26, 27]. At each time step, we solve the resulting semidiscrete problem exactly in the exterior of the cavity using Fourier transform method. The solution is then used to construct an exact nonlocal boundary condition over the cavity aperture. This boundary condition is key to reducing the infinite problem to a bounded (interior) computational domain.

3.1. Time-discretization. We begin by describing the Newmark time-stepping scheme. Let $\Delta t > 0$ be the (constant) time step

$$\Delta t = T/\mathcal{N},$$

where $\mathcal{N}$ is an arbitrary positive integer and T the total time. Let $t_n = n\Delta t$ be the nth time step. Newmark's scheme can be defined as follows. Consider the following expansion of a sufficiently regular function $y(t)$:

$$y(t_{n+1}) = y(t_n) + \Delta t y'(t_n) + (\Delta t)^2 \left[\beta y''(t_{n+1}) + \left(\frac{1}{2} - \beta\right) y''(t_n)\right] + \mathcal{O}(\Delta t^3),$$
$$y'(t_{n+1}) = y'(t_n) + \Delta t \left[\gamma y''(t_{n+1}) + (1-\gamma) y''(t_n)\right] + \mathcal{O}(\Delta t^2),$$

where β and γ are parameters. Define the approximations y^{n+1}, $\dot{y}^{n+1}$, and $\ddot{y}^{n+1}$ of $y(t_{n+1})$, $y'(t_{n+1})$, and $y''(t_{n+1})$, respectively, by

$$y^{n+1} = y^n + \Delta t \dot{y}^n + (\Delta t)^2 \left[\beta \ddot{y}^{n+1} + \left(\frac{1}{2} - \beta\right) \ddot{y}^n\right], \quad n = 0, 1, \ldots, \mathcal{N},$$
$$\dot{y}^{n+1} = \dot{y}^n + \Delta t \left[\gamma \ddot{y}^{n+1} + (1-\gamma)\ddot{y}^n\right], \quad n = 0, 1, \ldots, \mathcal{N}.$$

The above recurrence equations are called the Newmark time-stepping scheme. We note that the scheme is *explicit* if $\beta = 0$, i.e., y_{n+1} depends only on the terms at the nth time step, and *implicit* otherwise. The scheme accuracy is $\mathcal{O}(\Delta t^2)$ if $\gamma = 1/2$ and $\mathcal{O}(\Delta t)$ if $\gamma \neq 1/2$, which can be seen from the Taylor expansion

$$y'(t_{n+1}) = y'(t_n) + \frac{\Delta t}{2} y''(t_n) + \mathcal{O}(\Delta t^2).$$

Throughout the paper, we apply the Newmark scheme with $\gamma = 1/2$. By applying the scheme to (2.2) we have

$$\text{(3.1)} \qquad \begin{cases} \operatorname{curl}\operatorname{curl} \boldsymbol{E}^{n+1} + \varepsilon_r \ddot{\boldsymbol{E}}^{n+1} = 0\,, \\ \boldsymbol{E}^{n+1} = \boldsymbol{E}^n + \Delta t \dot{\boldsymbol{E}}^n + \dfrac{(\Delta t)^2}{2}[(1-2\beta)\ddot{\boldsymbol{E}}^n + 2\beta \ddot{\boldsymbol{E}}^{n+1}]\,, \\ \dot{\boldsymbol{E}}^{n+1} = \dot{\boldsymbol{E}}^n + \Delta t \left[\dfrac{\ddot{\boldsymbol{E}}^n}{2} + \dfrac{\ddot{\boldsymbol{E}}^{n+1}}{2}\right], \end{cases}$$

where $\boldsymbol{E}^n = \boldsymbol{E}^n(\boldsymbol{x})$, $\dot{\boldsymbol{E}}^n = \dot{\boldsymbol{E}}^n(\boldsymbol{x})$, and $\ddot{\boldsymbol{E}}^n$, are the *temporal approximations* of $\boldsymbol{E}(\boldsymbol{x}, t_n)$, $\frac{\partial \boldsymbol{E}}{\partial t}(\boldsymbol{x}, t_n)$, and $\frac{\partial^2 \boldsymbol{E}}{\partial t^2}(\boldsymbol{x}, t_n)$, respectively, for $n = 1, 2, \ldots, \mathcal{N}$. If $n = 0$,

we set

$$\boldsymbol{E}^0 = \boldsymbol{E}(0), \quad \dot{\boldsymbol{E}}^0 = \frac{\partial}{\partial t}\boldsymbol{E}(0).$$

From (3.1) we can derive the following *semidiscrete problem* (recursive relations):

$$(3.2)\quad \begin{cases} (\Delta t)^2\beta \operatorname{curl}\operatorname{curl}\boldsymbol{E}^{n+1} + \varepsilon_r \boldsymbol{E}^{n+1} = \varepsilon_r \left[\boldsymbol{E}^n + \Delta t v_n + \dfrac{(\Delta t)^2}{2}(1-2\beta)\ddot{\boldsymbol{E}}^n\right], \\ \ddot{\boldsymbol{E}}^{n+1} = \dfrac{1}{2\beta}\left[\dfrac{2}{(\Delta t)^2}(\boldsymbol{E}^{n+1} - \boldsymbol{E}^n - \Delta t\dot{\boldsymbol{E}}^n) - (1-2\beta)\ddot{\boldsymbol{E}}^n\right], \\ \dot{\boldsymbol{E}}^{n+1} = \dot{\boldsymbol{E}}^n + \Delta t\left[\dfrac{\ddot{\boldsymbol{E}}^n}{2} + \dfrac{\ddot{\boldsymbol{E}}^{n+1}}{2}\right]. \end{cases}$$

We note that the first equation in (3.2) is a partial differential equation in terms of the function $\boldsymbol{E}^{n+1}$, in which the right-hand side is known at the current time step. The discrete version of the boundary condition on $\partial\mathcal{G}$, as well as some radiation condition, must be added to (3.2) to complete the statement of the problem. From an implementation viewpoint, it is beneficial to write these equations in a predictor-corrector form as follows.

Prediction:

$$\tilde{\boldsymbol{E}}^{n+1} = \boldsymbol{E}^n + \Delta t\dot{\boldsymbol{E}}^n + \frac{(\Delta t)^2}{2}(1-2\beta)\ddot{\boldsymbol{E}}^n,$$

$$\tilde{\dot{\boldsymbol{E}}}^{n+1} = \dot{\boldsymbol{E}}^n + \frac{\Delta t}{2}\ddot{\boldsymbol{E}}^n.$$

Solution:

$$(P^{n+1})\begin{cases} (\Delta t)^2\beta \operatorname{curl}\operatorname{curl}\boldsymbol{E}^{n+1} + \varepsilon_r\boldsymbol{E}^{n+1} = \varepsilon_r\tilde{\boldsymbol{E}}^{n+1} & \text{in } \Omega, \\ \hat{n}\times\boldsymbol{E}^{n+1} = 0 & \text{on } S, \\ \hat{n}\times\boldsymbol{E}^{n+1} = \hat{n}\times(\boldsymbol{E}^{s,n+1} + \boldsymbol{E}^{i,n+1} + \boldsymbol{E}^{r,n+1}) & \text{on } \Gamma, \\ \hat{n}\times\operatorname{curl}\boldsymbol{E}^{n+1} = \hat{n}\times\operatorname{curl}(\boldsymbol{E}^{s,n+1} + \boldsymbol{E}^{i,n+1} + \boldsymbol{E}^{r,n+1}) & \text{on } \Gamma. \end{cases}$$

Correction:

$$\ddot{\boldsymbol{E}}^{n+1} = \frac{1}{\beta(\Delta t)^2}(\boldsymbol{E}^{n+1} - \tilde{\boldsymbol{E}}^{n+1}),$$

$$\dot{\boldsymbol{E}}^{n+1} = \tilde{\dot{\boldsymbol{E}}}^{n+1} + \frac{\Delta t}{2}\ddot{\boldsymbol{E}}^{n+1}.$$

Thus, in addition to the updating performed in the *prediction* and the *correction*, we have to solve in each time step the boundary value problem in the *solution*. We also note that the differential operator $(\operatorname{curl}\operatorname{curl} + \varepsilon_r I)$ in (P^{n+1}) is independent of time; hence it need be inverted only once. To solve the sequence of boundary value problems (P^{n+1}), $n = 0, 1, 2, \ldots, \mathcal{N}-1$, we use the standard finite element method with an appropriate nonlocal boundary condition. We now concentrate on the partial

differential equation in (P^{n+1}):

$$
\left\{
\begin{aligned}
\operatorname{curl}\operatorname{curl}\boldsymbol{E}^{n+1}+\alpha^2\varepsilon_r\boldsymbol{E}^{n+1} &= \boldsymbol{h}^{n+1} \quad \text{in } \Omega,\\
\operatorname{curl}\operatorname{curl}\boldsymbol{E}^{s,n+1}+\alpha^2\boldsymbol{E}^{s,n+1} &= \boldsymbol{h}^{s,n+1} \quad \text{in } \mathcal{U},\\
\hat{n}\times\boldsymbol{E}^{n+1} &= 0 \quad \text{on } S,\\
\hat{n}\times\boldsymbol{E}^{s,n+1} &= 0 \quad \text{on } \Gamma^c,\\
\hat{n}\times\operatorname{curl}\boldsymbol{E}^{n+1} &= \hat{n}\times\operatorname{curl}(\boldsymbol{E}^s+\boldsymbol{E}^i+\boldsymbol{E}^r)^{n+1} \quad \text{on } \Gamma,\\
\hat{n}\times\boldsymbol{E}^{n+1} &= \hat{n}\times(\boldsymbol{E}^s+\boldsymbol{E}^i+\boldsymbol{E}^r)^{n+1} \quad \text{on } \Gamma,
\end{aligned}
\right. \tag{3.3}
$$

where

$$\alpha^2=\frac{1}{(\Delta t)^2\beta}, \tag{3.4}$$

$$\boldsymbol{h}^{n+1}=\alpha^2\varepsilon_r(\boldsymbol{x})\tilde{\boldsymbol{E}}^{n+1} \quad \text{in } \Omega, \tag{3.5}$$

$$\boldsymbol{h}^{s,n+1}=\alpha^2\tilde{\boldsymbol{E}}^{s,n+1} \quad \text{in } \mathcal{U}. \tag{3.6}$$

In the next subsection, we wish to solve for the scattered field $\boldsymbol{E}^{s,n+1}$ in (3.3) exactly by using Fourier transformation. We can then construct a nonlocal boundary operator which relates the tangential components of $\operatorname{curl}\boldsymbol{E}$ and $\boldsymbol{E}$ for $\boldsymbol{E}\in H(\operatorname{curl},\Omega)$.

3.2. Exterior problem and nonlocal boundary condition. Let $\tilde{\Gamma}$ be a closed smooth surface such that $\bar{\Gamma}\subset\tilde{\Gamma}$. Let $\operatorname{div}_\Gamma$ and $\operatorname{curl}_\Gamma$ be the surface divergence and the scalar surface rotational, respectively. For the detailed definitions of these operators, the reader is referred to [3, 8]. We define the following Sobolev spaces:

$$
\begin{aligned}
H^{1/2}(\Gamma)&=\{\phi|_\Gamma:\phi\in H^{1/2}(\tilde{\Gamma})\},\\
\tilde{H}^{1/2}(\Gamma)&=\{\phi|_\Gamma:\phi\in H^{1/2}(\tilde{\Gamma}),\ \operatorname{supp}(\phi)\subset\bar{\Gamma}\},
\end{aligned}
$$

and

$$
\begin{aligned}
H^{-1/2}(\Gamma)&=(\tilde{H}^{1/2}(\Gamma))',\\
\tilde{H}^{-1/2}(\Gamma)&=(H^{1/2}(\Gamma))'.
\end{aligned}
$$

We also need

$$
\begin{aligned}
H^{-1/2}(\operatorname{div}_\Gamma,\Gamma)&=\{\Phi\in[H^{-1/2}(\Gamma)]^3:\Phi\cdot\hat{x}_3=0,\ \operatorname{div}_\Gamma\Phi\in H^{-1/2}(\Gamma),\ \text{and } \nu\cdot\Phi|_{\partial\Gamma}=0\},\\
H^{-1/2}(\operatorname{curl}_\Gamma,\Gamma)&:=\{\Phi\in[H^{-1/2}(\Gamma)]^3:\Phi\cdot\hat{x}_3=0,\ \operatorname{curl}_\Gamma\Phi\in H^{-1/2}(\Gamma),\ \text{and } \nu\cdot\Phi|_{\partial\Gamma}=0\},
\end{aligned}
$$

where ν is the norm to the boundary of Γ. Denote

$$\Phi^{n+1}(x_1,x_2)=(-\phi_2^{n+1},\phi_1^{n+1},0)=\hat{x}_3\times\boldsymbol{E}^{s,n+1} \quad \text{on } \{x_3=0\}.$$

Consider the semidiscrete problem (3.3) in the upper half space $\mathcal{U}$:

Given $\boldsymbol{h}^{s,n+1}\in[L^2(\mathcal{U})]^3$ and $\Phi\in H^{-1/2}(\operatorname{div}_\Gamma,\Gamma)$, find $\boldsymbol{E}^{s,n+1}\in H(\operatorname{curl},\mathcal{U})$, $n=0,1,\ldots,\mathcal{N}-1$, such that

$$
\left\{
\begin{aligned}
\operatorname{curl}\operatorname{curl}\boldsymbol{E}^{s,n+1}+\alpha^2\boldsymbol{E}^{s,n+1} &= \boldsymbol{h}^{s,n+1} \quad \text{in } \mathcal{U},\\
\hat{x}_3\times\boldsymbol{E}^{s,n+1} &= \Phi^{n+1} \quad \text{on } \Gamma,\\
\hat{x}_3\times\boldsymbol{E}^{s,n+1} &= 0 \quad \text{on } \Gamma^c.
\end{aligned}
\right. \tag{3.7}
$$

We shall omit the superscript $n+1$ in (3.7) for the rest of this section. We have the following well-posedness result for the exterior problem. The proof of the result is based on the construction of the solution to the Fourier transformed equation of (3.7).

THEOREM 1. *For each* $\boldsymbol{h}^s \in [L^2(\mathcal{U})]^3$ *and* $\Phi \in H^{-1/2}(\mathrm{div}_\Gamma, \Gamma)$, *there exists a unique solution* $\boldsymbol{E}^s \in H(\mathrm{curl}, \mathcal{U})$ *to* (3.7).

Proof. In the homogeneous upper half space $\mathcal{U}$, $\boldsymbol{E}^s$ is divergence-free ($\mathrm{div}\, \boldsymbol{E}^s = 0$). By the identity $\mathrm{curl}\,\mathrm{curl}\, u = -\Delta u + \mathrm{grad}(\mathrm{div}\, u)$, (3.7) becomes

$$\begin{cases} -\Delta \boldsymbol{E}^s + \alpha^2 \boldsymbol{E}^s &= \boldsymbol{h}^s \ \text{ in } \mathcal{U}, \\ \hat{x}_3 \times \boldsymbol{E}^s &= \Phi \ \text{ on } \Gamma, \\ \hat{n} \times \boldsymbol{E}^s &= 0 \ \text{ on } \Gamma^c. \end{cases}$$

Or, for $j = 1, 2$,

$$\begin{cases} -\Delta E_j^s + \alpha^2 E_j^s &= h_j^s \ \text{ in } \mathcal{U}, \\ E_j^s &= \phi_j \ \text{ on } \{x_3 = 0\}, \\ \lim\limits_{r\to\infty} r E_j^s &= 0, \end{cases} \tag{3.8}$$

where

$$E_j^s = \phi_j \text{ iff } \int_{x_3=0} E_j^s \psi = \int_{x_3=0} \phi_j \psi \quad \forall \psi \in \tilde{H}^{1/2}(\Gamma).$$

By the divergence-free property of $\boldsymbol{E}^s$, we only need to solve for the components E_1^s and E_2^s. Taking the two-dimensional Fourier transform of the distributions with respect to (x_1, x_2) of (3.8), we get, for $j = 1, 2$,

$$\begin{cases} \left[\dfrac{\partial^2}{\partial x_3^2} - (\alpha^2 + \xi_1^2 + \xi_2^2)\right] \hat{E}_j^s(\xi, x_3) &= -\hat{h}_j^s(\xi, x_3) \ \text{ in } \{x_3 > 0\}, \\ \hat{E}_j^s(\xi, 0) &= \hat{\phi}_j(\xi) \ \text{ on } \{x_3 = 0\}, \end{cases} \tag{3.9}$$

where $\xi = (\xi_1, \xi_2)$.

REMARK 2. *Since* $\boldsymbol{E}$ *is real, the Fourier transforms of* E_j^s, h_j^s, *and* ϕ_j *are of the form*

$$\begin{aligned} \hat{f}(\xi, x_3) &= \frac{1}{2\pi} \int_{\mathbb{R}^2} f(x_1, x_2, x_3) e^{-i(x_1\xi_1 + x_2\xi_2)} dx_1 dx_2 \\ &= 4\,\mathrm{Re} \left\{ \frac{1}{2\pi} \int_{\mathbb{R}^2_+} f(x_1, x_2, x_3) e^{-i(x_1\xi_1 + x_2\xi_2)} dx_1 dx_2 \right\}. \end{aligned}$$

The solution of the nonhomogeneous differential equation (3.9) is unique and can be expressed as the sum of the homogeneous solution and the particular solution

$$\hat{E}_j^s(\xi, x_3) = w_j(x_3) + \int_0^\infty G(x_3, x_3')(-\hat{h}_j^s(\xi, x_3'))dx_3', \tag{3.10}$$

where w_j is the solution of

$$\begin{cases} w_j'' - \eta^2 w_j &= 0 \quad \text{in } \{x_3 > 0\}, \\ w_j &= \hat{\phi}_j \quad \text{on } \{x_3 = 0\}, \\ \lim\limits_{x_3\to\infty} x_3 w_j &= 0, \end{cases}$$

where

$$\eta = \sqrt{\alpha^2 + \xi_1^2 + \xi_2^2},$$

and $G(x, x')$ is the associated Green function, corresponding to the homogeneous boundary conditions [17, p. 234]. We now derive w_j and $G(x, x')$. It is easy to see that the homogeneous solution w_j is

$$w_j(x_3) = \hat{\phi}_j e^{-\eta x_3}. \tag{3.11}$$

The Green's function is defined by

$$G(x, x') = \begin{cases} \phi(x')\psi(x)W(x')^{-1} & \text{if } x \le x', \\ \psi(x')\phi(x)W(x')^{-1} & \text{if } x \ge x', \end{cases}$$

where

$$\begin{aligned} \phi(x') &= e^{-\eta x'}, \\ \psi(x') &= e^{\eta x'} - e^{-\eta x'} = 2\sinh \eta x', \\ W(x') &= \psi(x')\phi'(x') - \phi(x')\psi'(x') = -2\eta. \end{aligned}$$

Hence,

$$G(x, x') = \begin{cases} -e^{-\eta x'} \sinh(\eta x)/\eta & \text{if } x \le x', \\ -\sinh(\eta x')e^{-\eta x}/\eta & \text{if } x \ge x'. \end{cases} \tag{3.12}$$

Taking the partial derivative of G with respect to $\boldsymbol{x}$ yields

$$\frac{\partial G(x, x')}{\partial x} = \begin{cases} -e^{-\eta x'} \cosh(\eta x) & \text{if } x \le x', \\ \sinh(\eta x')e^{-\eta x} & \text{if } x \ge x'. \end{cases}$$

Hence, $\partial_3 \hat{E}_j^s,\ j = 1, 2$, are

$$\partial_3 \hat{E}_j^s = -\eta \hat{\phi}_j e^{-\eta x_3} - \int_0^\infty \frac{\partial G(x_3, x_3')}{\partial x_3} \hat{h}_j^s(x_3') dx_3'.$$

By taking the inverse Fourier transform of (3.10), we obtain, for $j = 1, 2$,

$$\begin{aligned} E_j^s &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{E}_j^s e^{i(x_1\xi_1 + x_2\xi_2)} d\xi_1 d\xi_2 \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{\phi}_j e^{-\eta x_3} e^{i(\xi_1 x_1 + \xi_2 x_2)} d\xi_1 d\xi_2 \\ &\quad - \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ \int_0^\infty G(x_3, x_3') \hat{h}_j^s dx_3' \right\} e^{i(\xi_1 x_1 + \xi_2 x_2)} d\xi_1 d\xi_2. \end{aligned} \tag{3.13}$$

For the third component, we note that

$$\hat{E}_3^s = \eta^{-2} \left(\frac{\partial^2 \hat{E}_3^s}{\partial x_3^2} + \hat{h}_3 \right).$$

Since $\operatorname{div} \boldsymbol{E}^s = 0$ in $\mathcal{U}$, its Fourier transform also vanishes; thus,

$$\frac{\partial \hat{E}_3^s}{\partial x_3} = -i(\xi_1 \hat{E}_1^s + \xi_2 \hat{E}_2^s).$$

Hence, we get

$$\begin{aligned}\hat{E}_3^s &= \eta^{-2}\left[-i\frac{\partial}{\partial x_3}(\xi_1\hat{E}_1^s+\xi_2\hat{E}_2^s)+\hat{h}_3^s\right],\\ &= \frac{i}{\eta}(\xi_1\hat{\phi}_1+\xi_2\hat{\phi}_2)e^{-\eta x_3}+\frac{i}{\eta^2}\int_0^\infty \frac{\partial G(x_3,x_3')}{\partial x_3}(\xi_1\hat{h}_1^s+\xi_2\hat{h}_2^s)dx_3'+\frac{1}{\eta^2}\hat{h}_3^s.\end{aligned}$$

Taking the inverse Fourier transform gives

$$\begin{aligned}E_3^s &= \frac{i}{2\pi}\int_{-\infty}^\infty\int_{-\infty}^\infty \frac{1}{\eta}(\xi_1\hat{\phi}_1+\xi_2\hat{\phi}_2)e^{-\eta x_3}e^{i(\xi_1x_1+\xi_2x_2)}d\xi_1d\xi_2\\ &+\frac{i}{2\pi}\int_{-\infty}^\infty\int_{-\infty}^\infty \frac{1}{\eta^2}\left\{\int_0^\infty \frac{\partial G(x_3,x_3')}{\partial x_3}(\xi_1\hat{h}_1^s+\xi_2\hat{h}_2^s)dx_3'\right\}e^{i(\xi_1x_1+\xi_2x_2)}d\xi_1d\xi_2\\ &+\frac{1}{2\pi}\int_{-\infty}^\infty\int_{-\infty}^\infty \frac{1}{\eta^2}\hat{h}_3^s e^{i(\xi_1x_1+\xi_2x_2)}d\xi_1d\xi_2. \qquad (3.14)\end{aligned}$$

Therefore, (3.13) and (3.14) yield the solution $\boldsymbol{E}^s$ to (3.7). ∎

From the solution $\boldsymbol{E}^s$, we now construct the nonlocal boundary operator. The tangential components of curl $\boldsymbol{E}^s$ on Γ,

$$\hat{n}\times\operatorname{curl}\boldsymbol{E}^s = (-(\partial_3E_1^s-\partial_1E_3^s),\partial_2E_3^s-\partial_3E_2^s,0),$$

can be expressed in terms of Φ and $\boldsymbol{h}^s$ as follows. Denote $x=(x_1,x_2)$, for $x_3=0$,

$$\begin{aligned}-\partial_3E_1^s+\partial_1E_3^s &= \frac{1}{2\pi}\int_{\mathbb{R}^2}\left\{\eta\hat{\phi}_1-\frac{1}{\eta}(\xi_1\hat{\phi}_1+\xi_2\hat{\phi}_2)\xi_1\right\}e^{i\xi\cdot x}d\xi\\ -\frac{1}{2\pi}\int_{\mathbb{R}^2}e^{i\xi\cdot x}d\xi\int_0^\infty \frac{\partial G(x_3,x_3')}{\partial x_3}&\left[\hat{h}_1^s+\frac{1}{\eta^2}(\xi_1\hat{h}_1^s+\xi_2\hat{h}_2^s)\xi_1\right]_{x_3=0}dx_3' \qquad (3.15)\\ &+\frac{i}{2\pi}\int_{\mathbb{R}^2}\frac{\xi_1}{\eta^2}\hat{h}_3^s|_{x_3=0}\,e^{i\xi\cdot x}d\xi,\end{aligned}$$

and

$$\begin{aligned}-\partial_3E_2^s+\partial_2E_3^s &= \frac{1}{2\pi}\int_{\mathbb{R}^2}\left\{\eta\hat{\phi}_2-\frac{1}{\eta}(\xi_2\hat{\phi}_1+\xi_2\hat{\phi}_2)\xi_2\right\}e^{i\xi\cdot x}d\xi\\ -\frac{1}{2\pi}\int_{\mathbb{R}^2}e^{i\xi\cdot x}d\xi\int_0^\infty \frac{\partial G(x_3,x_3')}{\partial x_3}&\left[\hat{h}_2^s+\frac{1}{\eta^2}(\xi_1\hat{h}_1^s+\xi_2\hat{h}_2^s)\xi_2\right]_{x_3=0}dx_3' \qquad (3.16)\\ &+\frac{i}{2\pi}\int_{\mathbb{R}^2}\frac{\xi_2}{\eta^2}\hat{h}_3^s|_{x_3=0}\,e^{i\xi\cdot x}d\xi.\end{aligned}$$

We note that only the first term on the right-hand side of (3.15) and (3.16) contains the boundary function ϕ. So, we define the operator T by

$$T(\hat{x}_3\times u) = [T^{(1)}(\hat{x}_3\times u),T^{(2)}(\hat{x}_3\times u)] \quad \text{on } \Gamma, \qquad (3.17)$$

where

$$T^{(1)}(\hat{x}_3\times u)(x) = \frac{1}{2\pi}\int_{\mathbb{R}^2}\left\{\eta\hat{u}_1-\frac{1}{\eta}(\xi_1\hat{u}_1+\xi_2\hat{u}_2)\xi_1\right\}e^{i\xi\cdot x}d\xi, \qquad (3.18)$$

$$T^{(2)}(\hat{x}_3\times u)(x) = \frac{1}{2\pi}\int_{\mathbb{R}^2}\left\{\eta\hat{u}_2-\frac{1}{\eta}(\xi_1\hat{u}_1+\xi_2\hat{u}_2)\xi_2\right\}e^{i\xi\cdot x}d\xi. \qquad (3.19)$$

Thus, we have

$$\hat{x}_3 \times \operatorname{curl} \boldsymbol{E}^s = T(\hat{x}_3 \times \boldsymbol{E}^s) + Q\boldsymbol{h}^s, \tag{3.20}$$

where $Q\boldsymbol{h}^s = [Q^{(1)}\boldsymbol{h}^s, Q^{(2)}\boldsymbol{h}^s]$ with

$$Q^{(1)}\boldsymbol{h}^s = -\frac{1}{2\pi}\int_{\mathbb{R}^2} e^{i\xi\cdot x}d\xi \int_0^\infty \frac{\partial G(x_3,x_3')}{\partial x_3}\left[\hat{h}_1^s + \frac{1}{\eta^2}(\xi_1\hat{h}_1^s + \xi_2\hat{h}_2^s)\xi_1\right]dx_3'$$
$$+\frac{i}{2\pi}\int_{\mathbb{R}^2}\frac{1}{\eta^2}\xi_1\hat{h}_3^s e^{i\xi\cdot x}d\xi, \quad \text{for } x_3 = 0,$$

and

$$Q^{(2)}\boldsymbol{h}^s = -\frac{1}{2\pi}\int_{\mathbb{R}^2} e^{i\xi\cdot x}d\xi \int_0^\infty \frac{\partial G(x_3,x_3')}{\partial x_3}\left[\hat{h}_2^s + \frac{1}{\eta^2}(\xi_1\hat{h}_1^s + \xi_2\hat{h}_2^s)\xi_2\right]dx_3'$$
$$+\frac{i}{2\pi}\int_{\mathbb{R}^2}\frac{1}{\eta^2}\xi_2\hat{h}_3^s e^{i\xi\cdot x}d\xi, \quad \text{for } x_3 = 0.$$

LEMMA 3. *The operator $T : H^{-1/2}(\operatorname{div},\Gamma) \to [H^{-1/2}(\operatorname{curl},\Gamma)]^*$, the dual space of $H^{-1/2}(\operatorname{curl},\Gamma)$, is a linear bounded operator. Furthermore,*

$$\langle T(\hat{x}_3 \times u), \hat{x}_3 \times \hat{x}_3 \times u\rangle_\Gamma \le 0.$$

Proof. We shall prove that

$$|\langle T(\hat{x}_3 \times u), \phi\rangle| \le C\|\hat{x}_3 \times u\|_{H^{-1/2}(\operatorname{div},\Gamma)}\|\phi\|_{H^{-1/2}(\operatorname{curl},\Gamma)} \quad \forall \phi \in H^{-1/2}(\operatorname{curl},\Gamma).$$

Recall that

$$\|\hat{x}_3 \times u\|^2_{H^{-1/2}(\operatorname{div},\Gamma)} = \int_{\mathbb{R}^2}\frac{1}{\sqrt{1+|\xi|^2}}\Big[|\hat{u}_1|^2 + |\hat{u}_2|^2 + |-\xi_1\hat{u}_2 + \xi_2\hat{u}_1|^2\Big]d\xi$$

and

$$\|\phi\|^2_{H^{-1/2}(\operatorname{curl},\Gamma)} = \int_{\mathbb{R}^2}\frac{1}{\sqrt{1+|\xi|^2}}\Big[|\hat{\phi}_1|^2 + |\hat{\phi}_2|^2 + |\xi_1\hat{\phi}_2 - \xi_2\hat{\phi}_1|^2\Big]d\xi.$$

By the definition of T, we have

$$\begin{aligned}
&\langle T(\hat{x}_3 \times u), \phi\rangle = \int_\Gamma T(\hat{x}_3 \times u)\cdot\bar{\phi}dx\\
&= \int_{x_3=0} T^{(1)}(\hat{x}_3 \times u)\cdot\bar{\phi}_1 dx + \int_{x_3=0} T^{(2)}(\hat{x}_3 \times u)\cdot\bar{\phi}_2 dx\\
&= \int\left[\eta\hat{u}_1 - \frac{1}{\eta}(\xi_1\hat{u}_1 + \xi_2\hat{u}_2)\xi_1\right]\bar{\tilde{v}}_1 d\xi + \int\left[\eta\hat{u}_2 - \frac{1}{\eta}(\xi_1\hat{u}_1 + \xi_2\hat{u}_2)\xi_2\right]\bar{\tilde{v}}_2 d\xi\\
&= \int\frac{1}{\eta}\Big[\eta^2(\hat{u}_1\bar{\tilde{v}}_1 + \hat{u}_2\bar{\tilde{v}}_2) - (\xi_1\hat{u}_1 + \xi_2\hat{u}_2)(\xi_1\bar{\tilde{v}}_1 + \xi_2\bar{\tilde{v}}_2)\Big]d\xi\\
&= \int\frac{1}{\eta}\Big[\alpha^2(\hat{u}_1\bar{\tilde{v}}_1 + \hat{u}_2\bar{\tilde{v}}_2) + \xi_2^2\hat{u}_1\bar{\tilde{v}}_1 + \xi_1^2\hat{u}_2\bar{\tilde{v}}_2 - \xi_1\xi_2\hat{u}_2\bar{\tilde{v}}_1 - \xi_1\xi_2\hat{u}_1\bar{\tilde{v}}_2\Big]d\xi\\
&= \int\frac{1}{\eta}\Big[\alpha^2(\hat{u}_1\bar{\tilde{v}}_1 + \hat{u}_2\bar{\tilde{v}}_2) + (-\xi_1\hat{u}_2 + \xi_2\hat{u}_1)(\xi_2\bar{\tilde{v}}_1 - \xi_1\bar{\tilde{v}}_2)\Big]d\xi.
\end{aligned}$$

By the Cauchy–Schwarz inequality, we obtain

$$|\langle T(\hat{x}_3 \times u), \phi\rangle| \le C\|\hat{x}_3 \times u\|_{H^{-1/2}(\mathrm{div},\Gamma)}\|\phi\|_{H^{-1/2}(\mathrm{curl},\Gamma)}.$$

To show the operator T is nonpositive, we observe that

$$\begin{aligned}
&-(2\pi)^2\langle T(\hat{x}_3 \times u), \hat{x}_3 \times \hat{x}_3 \times u\rangle_\Gamma \\
&= \int_\Gamma T^{(1)}(\hat{x}_3 \times u)u_1 + \int_\Gamma T^{(2)}(\hat{x}_3 \times u)u_2 \\
&= \int_{\mathbb{R}^2}\left\{\eta\hat{u}_1 - \frac{1}{\eta}(\xi_1\hat{u}_1 + \xi_2\hat{u}_2)\xi_1\right\}\hat{u}_1 d\xi + \int_{\mathbb{R}^2}\left\{\eta\hat{u}_2 - \frac{1}{\eta}(\xi_1\hat{u}_1 + \xi_2\hat{u}_2)\xi_2\right\}\hat{u}_2 d\xi \\
&= \int_{\mathbb{R}^2}\left\{\eta(\hat{u}_1^2 + \hat{u}_2^2) - \frac{1}{\eta}\left(\xi_1\hat{u}_1 + \xi_2\hat{u}_2\right)^2\right\} d\xi \\
&= \int_{\mathbb{R}^2}\frac{1}{\eta}\left\{\eta^2(\hat{u}_1^2 + \hat{u}_2^2) - (\xi_1\hat{u}_1 + \xi_2\hat{u}_2)^2\right\} d\xi.
\end{aligned}$$

Recall that $\eta = (\alpha^2 + \xi_1^2 + \xi_2^2)^{1/2}$; the last integrand is clearly positive. Therefore, we have

$$\langle T(\hat{x}_3 \times u), \hat{x}_3 \times \hat{x}_3 \times u\rangle \le 0$$

as desired. ∎

This lemma will be used in the next section to obtain the existence and uniqueness of the weak solutions to the variational problem defined in the cavity Ω (Lax–Milgram theorem).

4. Interior problem and variational formulation. In this section, we consider the problem defined in the cavity Ω:

$$\left\{\begin{array}{rcl}
\operatorname{curl}\operatorname{curl}\boldsymbol{E}^{n+1} + \alpha^2\varepsilon_r\boldsymbol{E}^{n+1} &=& \boldsymbol{h}^{n+1} \ \text{ in } \Omega, \\
\hat{n} \times \boldsymbol{E}^{n+1} &=& 0 \ \text{ on } S, \\
\hat{x}_3 \times \boldsymbol{E}^{n+1} &=& \hat{x}_3 \times (\boldsymbol{E}^i + \boldsymbol{E}^r + \boldsymbol{E}^s)^{n+1} \ \text{ on } \Gamma,
\end{array}\right. \tag{4.1}$$

where $\boldsymbol{h}^{n+1} = \alpha^2\varepsilon_r\tilde{\boldsymbol{E}}^{n+1}$ as before. We formally multiply the first equation of (4.1) with a suitable test function $\boldsymbol{\phi}$ and integrate by parts over Ω to obtain

$$(\operatorname{curl}\boldsymbol{E}^{n+1}, \operatorname{curl}\boldsymbol{\phi}) + \alpha^2(\varepsilon_r E^{n+1}, \boldsymbol{\phi}) - \left\langle \hat{x}_3 \times \operatorname{curl}\boldsymbol{E}^{n+1}, \hat{x}_3 \times (\hat{x}_3 \times \boldsymbol{\phi})\right\rangle_\Gamma = (\boldsymbol{h}^{n+1}, \boldsymbol{\phi}).$$

We wish to couple the total field $\boldsymbol{E}^{n+1}$ in Ω to the total field in the upper half space through Γ. This can be accomplished by finding a relation between $\hat{x}_3 \times \operatorname{curl}\boldsymbol{E}^{n+1}$ and $\hat{x}_3 \times \boldsymbol{E}^{n+1}$. Since $\hat{x}_3 \times (\boldsymbol{E}^{i,n+1} + \boldsymbol{E}^{r,n+1}) = 0$ on Γ, we have

$$\begin{aligned}
\hat{x}_3 \times \operatorname{curl}\boldsymbol{E}^{s,n+1} &= T(\hat{x}_3 \times (\boldsymbol{E}^{n+1} - \boldsymbol{E}^{i,n+1} - \boldsymbol{E}^{r,n+1})) \\
&\quad + Q(\boldsymbol{h}^{n+1} - \boldsymbol{h}^{i,n+1} - \boldsymbol{h}^{r,n+1}) \\
&= T(\hat{x}_3 \times \boldsymbol{E}^{n+1}) + Q\boldsymbol{h}^{n+1} - Q(\boldsymbol{h}^{i,n+1} + \boldsymbol{h}^{r,n+1}).
\end{aligned}$$

Hence, we obtain the following relation on Γ:

$$\begin{aligned}
\hat{x}_3 \times \operatorname{curl}\boldsymbol{E}^{n+1} &= T(\hat{x}_3 \times \boldsymbol{E}^{n+1}) + \hat{x}_3 \times (\operatorname{curl}\boldsymbol{E}^{i,n+1} + \operatorname{curl}\boldsymbol{E}^{r,n+1}) \\
&\quad + Q\boldsymbol{h}^{n+1} - Q(\boldsymbol{h}^{i,n+1} + \boldsymbol{h}^{r,n+1}) \\
&= T(\hat{x}_3 \times \boldsymbol{E}^{n+1}) + \boldsymbol{M}^{n+1},
\end{aligned} \tag{4.2}$$

where

$$\boldsymbol{M}^{n+1} = \hat{x}_3 \times (\operatorname{curl} \boldsymbol{E}^{i,n+1} + \operatorname{curl} \boldsymbol{E}^{r,n+1}) + Q\boldsymbol{h}^{n+1} - Q(\boldsymbol{h}^{i,n+1} + \boldsymbol{h}^{r,n+1}).$$

The nonlocal boundary condition (4.2) on Γ is exact for each t_{n+1}. Now we can state the interior problem (4.1) as the following.

Given $\boldsymbol{h}^{n+1} \in [L^2(\Omega)]^3$, find $\boldsymbol{E}^{n+1} \in H(\operatorname{curl}, \Omega)$, $n = 0, 1, \ldots, \mathcal{N} - 1$, such that

$$\text{(4.3)} \qquad \begin{cases} \operatorname{curl}\operatorname{curl} \boldsymbol{E}^{n+1} + \alpha^2 \varepsilon_r \boldsymbol{E}^{n+1} &= \boldsymbol{h}^{n+1} \quad \text{in } \Omega, \\ \hat{n} \times \boldsymbol{E}^{n+1} &= 0 \quad \text{on } S, \\ \hat{x}_3 \times \operatorname{curl} \boldsymbol{E}^{n+1} &= T(\hat{x}_3 \times \boldsymbol{E}^{n+1}) + \boldsymbol{M}^{n+1} \quad \text{on } \Gamma. \end{cases}$$

The semidiscrete problem (4.3) is solved numerically by a variational method.

Define the variational space $V \subset H(\operatorname{curl}, \Omega)$ as

$$V = \{\boldsymbol{u} \in H(\operatorname{curl}, \Omega) : \hat{n} \times \boldsymbol{u}|_S = 0\}.$$

Then the corresponding variational form of the interior problem is

$$\text{(4.4)} \qquad a(\boldsymbol{E}^{n+1}, \boldsymbol{\phi}) = b^{n+1}(\boldsymbol{\phi}) \quad \forall \boldsymbol{\phi} \in V, \quad n = 0, 1, \ldots, \mathcal{N} - 1,$$

where

$$a(\boldsymbol{E}^{n+1}, \boldsymbol{\phi}) = (\operatorname{curl} \boldsymbol{E}^{n+1}, \operatorname{curl} \phi) + \alpha^2 (\varepsilon_r \boldsymbol{E}^{n+1}, \boldsymbol{\phi}) - \left\langle T(\hat{x}_3 \times \boldsymbol{E}^{n+1}), \hat{x}_3 \times (\hat{x}_3 \times \boldsymbol{\phi}) \right\rangle_\Gamma \quad \text{(4.5)}$$

and

$$\text{(4.6)} \qquad b^{n+1}(\boldsymbol{\phi}) = (\boldsymbol{h}^{n+1}, \boldsymbol{\phi}) + \left\langle \boldsymbol{M}^{n+1}, \hat{x}_3 \times (\hat{x}_3 \times \boldsymbol{\phi}) \right\rangle_\Gamma.$$

REMARK 4. *The time-stepping scheme can be described as the following.*

1. *Form the matrix A defined by the bilinear form $a(u, v)$ in* (4.5).
   Time-loop: for $n = 0, 1, 2 \ldots$
   2. *Compute the predicted values $\tilde{\boldsymbol{E}}^{n+1}$, $\tilde{\dot{\boldsymbol{E}}}^{n+1}$ in the interior Ω.*
   3. *Compute the predicted values $\tilde{\boldsymbol{E}}^{n+1}$, $\tilde{\dot{\boldsymbol{E}}}^{n+1}$ in the exterior $\mathcal{U}$.*
   4. *Form the right-hand side vector F^{n+1} defined by $b^{n+1}(\boldsymbol{\phi})$ in* (4.6).
   5. *Solve for the unknown expansion coefficients E^{n+1} in $AE^{n+1} = F^{n+1}$ in* (4.4).
   6. *Compute the solution $\boldsymbol{E}^{n+1}$ in the exterior $\mathcal{U}$ by* (3.6).
   7. *Correct $\ddot{\boldsymbol{E}}^{n+1}$ and $\dot{\boldsymbol{E}}^{n+1}$ in Ω.*
   8. *Correct $\ddot{\boldsymbol{E}}^{n+1}$ and $\dot{\boldsymbol{E}}^{n+1}$ in $\mathcal{U}$.*

Note that the matrix A is defined outside of the time loop; thus it can be precomputed and reused for each time step. This significantly reduces the storage and computational time.

We now show that the variational problem (4.4) has a unique solution in $H(\operatorname{curl}, \Omega)$ at each time step t_n.

THEOREM 5. *There exists a unique solution $\boldsymbol{u} \in V$ such that*

$$a(\boldsymbol{u}, \boldsymbol{v}) = b(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V.$$

Proof. Since $\varepsilon_r > 0$ and T is negative, it is easy to see that

$$\begin{aligned} a(\boldsymbol{u},\boldsymbol{u}) &= \|\operatorname{curl}\boldsymbol{u}\|_0^2 + \alpha^2(\varepsilon_r\boldsymbol{u},\boldsymbol{u}) - \langle T(\hat{x}_3\times\boldsymbol{u}),\hat{x}_3\times x_3\times\boldsymbol{u}\rangle_\Gamma \\ &\geq C\|\boldsymbol{u}\|^2_{H(\operatorname{curl},\Omega)}, \quad C>0. \end{aligned}$$

The bilinear form $a(\cdot,\cdot)$ is continuous in V, that is,

$$|a(\boldsymbol{u},\boldsymbol{v})| \leq C\|\boldsymbol{u}\|_{H(\operatorname{curl},\Omega)}\|\boldsymbol{v}\|_{H(\operatorname{curl},\Omega)}, \quad \forall \boldsymbol{u},\boldsymbol{v}\in V.$$

In fact, by the property of T and the trace theorem we have

$$\begin{aligned} |\langle T(\hat{x}_3\times\boldsymbol{u}),\hat{x}_3\times x_3\times\boldsymbol{v}\rangle_\Gamma| &\leq C\|\hat{x}_3\times\boldsymbol{u}\|_{H^{-1/2}(\operatorname{div}_\Gamma,\Gamma)}\|\hat{x}_3\times\hat{x}_3\times\boldsymbol{v}\|_{H^{-1/2}(\operatorname{curl}_\Gamma,\Gamma)} \\ &\leq C\|\boldsymbol{u}\|_{H(\operatorname{curl},\Omega)}\|\boldsymbol{v}\|_{H(\operatorname{curl},\Omega)}. \end{aligned}$$

Therefore, by the Lax–Milgram theorem the variational problem has a unique solution in V. ∎

In the next section, we use finite element methods to numerically solve the variational problem (4.4). We will also analyze the finite element error in $H(\operatorname{curl},\Omega)$-norm and the stability of the Newmark time-stepping scheme.

5. Fully discrete problem and finite element analysis. We assume that the cavity Ω can be decomposed into Lipschitz subregions Ω_i, $i = 1,2,\ldots,M$, so that $\epsilon_r|_{\Omega_i}$ is Lipschitz continuous and uniformly bounded. Let Ω be covered by a tetrahedral mesh $\mathcal{T}_h$ of regular, quasi-uniform finite elements with a maximum diameter $h > 0$. We will use linear edge elements as approximating functions which can be defined as follows. Let $\tilde{K}$ denote the reference tetrahedron

$$\tilde{K} = \{\tilde{\boldsymbol{x}}\in\mathbb{R}^3 : 1-\tilde{x}_1-\tilde{x}_2-\tilde{x}_3\geq 0,\ \tilde{x}_i\geq 0\}.$$

Let K be a tetrahedron in $\mathcal{T}_h$. Then there exists an affine map F_K such that $F_K(\tilde{K}) = K$. We define the set of "linear" functions on the reference tetrahedron

$$\mathcal{R}_1 = \{\tilde{\mathbf{u}} : \tilde{K}\to\mathbb{C}^3 : \tilde{\mathbf{u}} = \mathbf{a}+\mathbf{b}\times\tilde{\mathbf{x}},\ \mathbf{a},\mathbf{b}\in\mathbb{C}^3\},$$

and using this set we define for each $K\in\mathcal{T}_h$

$$\begin{aligned} \mathcal{R}_1(K) = \Big\{ &\mathbf{u}:\mathbf{K}\to\mathbb{C}^{\mathbf{3}} : \mathbf{u}(\mathbf{F_K}(\tilde{\mathbf{x}})) = [\mathbf{DF_K}^{\mathbf{T}}]^{-\mathbf{1}}(\tilde{\mathbf{x}})\tilde{\mathbf{u}}(\tilde{\mathbf{x}}) \\ &\text{for some } \tilde{\mathbf{u}}\in\mathcal{R}_1 \text{ and } \forall\tilde{\mathbf{x}}\in\tilde{K}\Big\}, \end{aligned}$$

where DF_K is the Jacobian of the transformation F_K. Thus, the linear edge element space V_h is defined as

$$V_h := \{\boldsymbol{v}_h\in V : \boldsymbol{v}_h|_K\in\mathcal{R}_1(K) \quad \forall K\in\mathcal{T}_h\}.$$

The degrees of freedom for this space are the moments of the tangential components of the field along the edges in the meshes and are defined by, for each tetrahedron K,

$$M_e(\boldsymbol{u}) = \left\{\int_e(\boldsymbol{u}\cdot\boldsymbol{\tau}_e)q\,ds : q\in P_1(e) \text{ on six edges of } K\right\},$$

where $\boldsymbol{\tau}_e$ is the unit tangential vector on e and $P_1(e)$ is the set of polynomial of degree one defined on e (for more details, see [13, 19]). In order for the integral in M_e to

make sense, it is required that $u \in W^{1,s}(\Omega)^3$ for $s > 2$ [13]. The discrete version of the variational problem (4.4) is the following. Find $\boldsymbol{E}_h \in V_h$ such that

$$a(\boldsymbol{E}_h^{n+1}, \boldsymbol{v}_h) = b(\boldsymbol{v}_h) \quad \forall \boldsymbol{v}_h \in V_h, \quad n = 0, 1, \ldots, \mathcal{N}. \tag{5.1}$$

Again by the Lax–Milgram theorem, we have the following theorem.

THEOREM 6. *The fully discrete problem* (5.1) *has a unique solution* $\boldsymbol{E}_h$ *in* V_h.

Approximation properties of V_h*:* The approximation properties of the finite element subspace V_h are well established (see, e.g., [13]) and are summarized as follows. Let $\boldsymbol{u} \in [W^{1,s}(\Omega)]^3$, $s > 2$ and $r_h : [W^{1,s}(\Omega)]^3 \to V_h$ be the interpolation operator defined by the moments M_e. If $\boldsymbol{u} \in [H^2(\Omega)]^3$,

$$\|\boldsymbol{u} - r_h\boldsymbol{u}\|_{L^2(\Omega)} + h\,\|\boldsymbol{u} - r_h\boldsymbol{u}\|_{H(\mathrm{curl},\Omega)} \le Ch\,\|\boldsymbol{u}\|_{H^2(\Omega)}\,.$$

Error estimate in $H(\mathrm{curl}, \Omega)$ for the finite element method is obtained by Céa's theorem [6].

THEOREM 7. *Let* $\boldsymbol{E}^n \in V$ *and* $\boldsymbol{E}_h^n \in V_h$ *be the semidiscrete solution and the fully discrete solution, respectively, to the variational equation. Then there exists* $C > 0$ *independent on* h *such that*

$$\|\boldsymbol{E} - \boldsymbol{E}_h\|_{H(\mathrm{curl},\Omega)} \le C \inf_{\boldsymbol{v}_h \in V_h} \|\boldsymbol{E} - \boldsymbol{v}_h\|_{H(\mathrm{curl},\Omega)}.$$

By using the approximation property of the linear edge elements we also have the following corollary.

COROLLARY 8. *Let* $\boldsymbol{E} \in V$ *be a solution of* (4.4) *and* $\boldsymbol{E} \in [H^2(\Omega_i)]^3$, $i = 1, 2, \ldots, M$. *Then*

$$\|\boldsymbol{E} - \boldsymbol{E}_h\|_{H(\mathrm{curl},\Omega_i)} \le C_i h\,\|\boldsymbol{E}\|_{H^2(\Omega_i)}\,, \quad i = 1, 2, \ldots, M.$$

We now analyze the stability of the Newmark scheme and the error estimate for the finite element approximation. With $\gamma = 1/2$, the Newmark scheme can also be expressed in terms of

$$\partial_\tau \boldsymbol{E}^n = \frac{\boldsymbol{E}^{n+1} - \boldsymbol{E}^n}{\Delta t}, \quad \partial_\tau^2 \boldsymbol{E}^n = \frac{\boldsymbol{E}^{n+1} - 2\boldsymbol{E}^n + \boldsymbol{E}^{n-1}}{(\Delta t)^2},$$
$$\boldsymbol{E}^{n,\beta} = \beta \boldsymbol{E}^{n+1} + (1 - 2\beta)\boldsymbol{E}^n + \beta \boldsymbol{E}^{n-1},$$

and (4.3) can be rewritten as

$$\begin{aligned}&(\partial_\tau^2 \boldsymbol{E}^n, \boldsymbol{v}) + (\mathrm{curl}\,\boldsymbol{E}^{n,\beta}, \mathrm{curl}\,\boldsymbol{v}) - \left\langle T(\hat{x}_3 \times \boldsymbol{E}^{n,\beta}), \hat{x}_3 \times \hat{x}_3 \times \boldsymbol{v} \right\rangle \\ &= (\boldsymbol{h}^{n,\beta}, \boldsymbol{v}) + \left\langle \boldsymbol{H}^{n,\beta}, \hat{x}_3 \times \hat{x}_3 \times \boldsymbol{v} \right\rangle, \quad n = 0, 1, \ldots, \mathcal{N},\end{aligned}$$

where

$$\boldsymbol{h}^{n,\beta} = \frac{\varepsilon_r}{\Delta t^2 \beta}\left[\boldsymbol{E}^{n-1,\beta} + \Delta t \dot{\boldsymbol{E}}^{n-1,\beta} + \frac{(1 - 2\beta)\Delta t^2}{2}\ddot{\boldsymbol{E}}^{n-1,\beta}\right].$$

To analyze the stability, we consider the homogeneous equation

$$(\partial_\tau^2 \boldsymbol{E}^n, \boldsymbol{v}) + (\mathrm{curl}\,\boldsymbol{E}^{n,\beta}, \mathrm{curl}\,\boldsymbol{v}) - \left\langle T(\hat{x}_3 \times \boldsymbol{E}^{n,\beta}), \hat{x}_3 \times \hat{x}_3 \times \boldsymbol{v} \right\rangle = 0\,. \tag{5.2}$$

We note that

$$\partial_\tau^2 \boldsymbol{E}^n = \frac{1}{\Delta t}(\partial_\tau \boldsymbol{E}^n - \partial_\tau \boldsymbol{E}^{n-1}), \tag{5.3}$$

and we define

$$\delta_\tau \boldsymbol{E}^n = \frac{1}{2}(\partial_\tau \boldsymbol{E}^n + \partial_\tau \boldsymbol{E}^{n-1}), \tag{5.4}$$

or equivalently,

$$\delta_\tau \boldsymbol{E}^n = \frac{1}{2\Delta t}(\boldsymbol{E}^{n+1} - \boldsymbol{E}^{n-1}). \tag{5.5}$$

By setting $\boldsymbol{v} = \delta_\tau \boldsymbol{E}^n$ in the homogeneous equation (5.2), we get

(5.6)

$$\begin{aligned} 0 &= (\partial_\tau^2 \boldsymbol{E}^n, \delta_\tau \boldsymbol{E}^n) + (\operatorname{curl} \boldsymbol{E}^{n,\beta}, \operatorname{curl} \delta_\tau \boldsymbol{E}^n) - \left\langle T(\hat{x}_3 \times \boldsymbol{E}^{n,\beta}), \hat{x}_3 \times \hat{x}_3 \times \delta_\tau \boldsymbol{E}^n \right\rangle \\ &\equiv I + II + III. \end{aligned}$$

We consider the first term I:

$$\begin{aligned} I = (\partial_\tau^2 \boldsymbol{E}^n, \delta_\tau \boldsymbol{E}^n) &= \frac{1}{2\Delta t}((\partial_\tau \boldsymbol{E}^n - \partial_\tau \boldsymbol{E}^{n-1}), (\partial_\tau \boldsymbol{E}^n + \partial_\tau \boldsymbol{E}^{n-1})) \\ &= \frac{1}{2\Delta t}\left(\|\partial_\tau \boldsymbol{E}^n\|_0^2 - \|\partial_\tau \boldsymbol{E}^{n-1}\|_0^2\right). \end{aligned}$$

For the second term II, we write

$$\begin{aligned} \boldsymbol{E}^{n,\beta} &= \beta(\boldsymbol{E}^{n+1} - 2\boldsymbol{E}^n + \boldsymbol{E}^{n-1}) + \boldsymbol{E}^n \\ &= \beta \Delta t^2 \partial_\tau^2 \boldsymbol{E}^n + \boldsymbol{E}^n, \end{aligned} \tag{5.7}$$

and denote

$$\boldsymbol{E}^{n+1/2} = \frac{1}{2}(\boldsymbol{E}^{n+1} + \boldsymbol{E}^n).$$

So, in terms of $\boldsymbol{E}^{n\pm 1/2}$ we have

$$\delta_\tau \boldsymbol{E}^n = \frac{1}{\Delta t}(\boldsymbol{E}^{n+1/2} - \boldsymbol{E}^{n-1/2}), \tag{5.8}$$

$$\boldsymbol{E}^n = \frac{\boldsymbol{E}^{n+1/2} + \boldsymbol{E}^{n-1/2}}{2} - \frac{\Delta t^2}{4}\partial_\tau^2 \boldsymbol{E}^n. \tag{5.9}$$

Substituting (5.9) into (5.7), we get

$$\begin{aligned} \boldsymbol{E}^{n,\beta} &= \beta \Delta t^2 \partial_\tau^2 \boldsymbol{E}^n + \frac{\boldsymbol{E}^{n+1/2} + \boldsymbol{E}^{n-1/2}}{2} - \frac{\Delta t^2}{4}\partial_\tau^2 \boldsymbol{E}^n \\ &= \left(\beta - \frac{1}{4}\right)\Delta t^2 \partial_\tau^2 \boldsymbol{E}^n + \frac{\boldsymbol{E}^{n+1/2} + \boldsymbol{E}^{n-1/2}}{2}. \end{aligned} \tag{5.10}$$

Hence, II becomes

$$\begin{aligned} II &= (\operatorname{curl} \boldsymbol{E}^{n,\beta}, \operatorname{curl} \delta_\tau \boldsymbol{E}^n) \\ &= \left(\beta - \frac{1}{4}\right)\Delta t^2 (\operatorname{curl} \partial_\tau^2 \boldsymbol{E}^n, \operatorname{curl} \delta_\tau \boldsymbol{E}^n) - \frac{1}{2}(\operatorname{curl} \boldsymbol{E}^{n+1/2} + \operatorname{curl} \boldsymbol{E}^{n-1/2}, \operatorname{curl} \delta_\tau \boldsymbol{E}^n) \\ &= \left(\beta - \frac{1}{4}\right)\frac{\Delta t}{2}\left(\|\operatorname{curl} \partial_\tau \boldsymbol{E}^n\|_0^2 - \|\operatorname{curl} \partial_\tau \boldsymbol{E}^{n-1}\|_0^2\right) \\ &\quad + \frac{1}{2\Delta t}\left(\|\operatorname{curl} \boldsymbol{E}^{n+1/2}\|_0^2 - \|\operatorname{curl} \boldsymbol{E}^{n-1/2}\|_0^2\right). \end{aligned}$$

Finally, we consider the boundary term *III*. From (5.3, 5.4, 5.8, 5.10) we have

$$\begin{aligned}
&\left\langle T(\hat{x}_3 \times \boldsymbol{E}^{n,\beta}), \hat{x}_3 \times \hat{x}_3 \times \delta_\tau \boldsymbol{E}^n \right\rangle \\
&= \left(\beta - \frac{1}{4}\right) \frac{\Delta t}{2} \left\langle T(\hat{x}_3 \times (\partial_\tau \boldsymbol{E}^n + \partial_\tau \boldsymbol{E}^{n-1})), \hat{x}_3 \times \hat{x}_3 \times (\partial_\tau \boldsymbol{E}^n - \partial_\tau \boldsymbol{E}^{n-1}) \right\rangle \\
&\quad + \frac{1}{2\Delta t} \left\langle T(\hat{x}_3 \times (\boldsymbol{E}^{n+1/2} + \boldsymbol{E}^{n-1/2})), \hat{x}_3 \times \hat{x}_3 \times (\boldsymbol{E}^{n+1/2} - \boldsymbol{E}^{n-1/2}) \right\rangle.
\end{aligned}$$

Summing (5.6) over n, $n = 0, 1, 2, 3, \ldots, \mathcal{N} - 1$, and by the property of the operator T we obtain

$$\begin{aligned}
&\frac{1}{2\Delta t} \|\partial_\tau \boldsymbol{E}^{\mathcal{N}}\|_0^2 + \left(\beta - \frac{1}{4}\right) \frac{\Delta t}{2} \|\operatorname{curl} \partial_\tau \boldsymbol{E}^{\mathcal{N}}\|_0^2 + \frac{1}{2\Delta t} \|\operatorname{curl} \boldsymbol{E}^{\mathcal{N}+1/2}\|_0^2 \\
&\qquad - \left(\beta - \frac{1}{4}\right) \frac{\Delta t}{2} \left\langle T(\hat{x}_3 \times \partial_\tau \boldsymbol{E}^{\mathcal{N}}), \hat{x}_3 \times \hat{x}_3 \times \partial_\tau \boldsymbol{E}^{\mathcal{N}} \right\rangle \\
&\qquad\quad - \frac{1}{2\Delta t} \left\langle T(\hat{x}_3 \times \boldsymbol{E}^{\mathcal{N}+1/2}), \hat{x}_3 \times \hat{x}_3 \times \boldsymbol{E}^{\mathcal{N}+1/2} \right\rangle \\
&= \frac{1}{2\Delta t} \|\partial_\tau \boldsymbol{E}^0\|_0^2 + \left(\beta - \frac{1}{4}\right) \frac{\Delta t}{2} \|\operatorname{curl} \partial_\tau \boldsymbol{E}^0\|_0^2 + \frac{1}{2\Delta t} \|\operatorname{curl} \boldsymbol{E}^{1/2}\|_0^2 \\
&\qquad - \left(\beta - \frac{1}{4}\right) \frac{\Delta t}{2} \left\langle T(\hat{x}_3 \times \partial_\tau \boldsymbol{E}^0), \hat{x}_3 \times \hat{x}_3 \times \partial_\tau \boldsymbol{E}^0 \right\rangle \\
&\qquad\quad - \frac{1}{2\Delta t} \left\langle T(\hat{x}_3 \times \boldsymbol{E}^{1/2}), \hat{x}_3 \times \hat{x}_3 \times \boldsymbol{E}^{1/2} \right\rangle,
\end{aligned}$$

which implies

$$\begin{aligned}
&\|\partial_\tau \boldsymbol{E}^{\mathcal{N}}\|_0^2 + \left(\beta - \frac{1}{4}\right) \Delta t^2 \|\operatorname{curl} \partial_\tau \boldsymbol{E}^{\mathcal{N}}\|_0^2 + \|\operatorname{curl} \boldsymbol{E}^{\mathcal{N}+1/2}\|_0^2 \\
&\le \|\partial_\tau \boldsymbol{E}^0\|_0^2 + \left(\beta - \frac{1}{4}\right) \Delta t^2 \|\operatorname{curl} \partial_\tau \boldsymbol{E}^0\|_0^2 + \|\operatorname{curl} \boldsymbol{E}^{1/2}\|_0^2 \\
&\quad - \left(\beta - \frac{1}{4}\right) \Delta t^2 \left\langle T(\hat{x}_3 \times \partial_\tau \boldsymbol{E}^0), \hat{x}_3 \times \hat{x}_3 \times \partial_\tau \boldsymbol{E}^0 \right\rangle \\
&\quad - \left\langle T(\hat{x}_3 \times \boldsymbol{E}^{1/2}), \hat{x}_3 \times \hat{x}_3 \times \boldsymbol{E}^{1/2} \right\rangle.
\end{aligned} \tag{5.11}$$

It is clear from the inequality (5.11) that if $1/4 \le \beta \le 1$, there exists a positive constant C_0 such that

$$\|\partial_\tau \boldsymbol{E}^{\mathcal{N}}\|_{H(\operatorname{curl},\Omega)}^2 + \|\operatorname{curl} \boldsymbol{E}^{\mathcal{N}+1/2}\|_0^2 \le C_0 \left(\|\partial_\tau \boldsymbol{E}^0\|_{H(\operatorname{curl},\Omega)}^2 + \|\boldsymbol{E}^{1/2}\|_{H(\operatorname{curl},\Omega)}^2 \right),$$

where C_0 depends only on $\beta, \Delta t$, and Ω. If $0 \le \beta < 1/4$, then we will not have an unconditionally stable scheme. For example, if we choose $\beta = 0$, then the scheme is explicit. Consequently, we need to find a condition imposed on Δt and h (mesh size) to guarantee the stability of the scheme. This can be accomplished by the "inverse assumption" for quasi-uniform triangulations [6],

$$\|\operatorname{curl} \boldsymbol{v}_h\|_0 \le C h^{-1} \|\boldsymbol{v}_h\|_0 \quad \forall \boldsymbol{v}_h \in V_h,\ C > 0.$$

When $\beta = 0$, we have

$$\left(1 - C^2\frac{\Delta t^2}{4h^2}\right)\|\partial_\tau \boldsymbol{E}^{\mathcal{N}}\|_0^2 + \|\operatorname{curl}\boldsymbol{E}^{\mathcal{N}+1/2}\|_0^2 \le C_1(\|\partial_\tau \boldsymbol{E}^0\|_{H(\operatorname{curl},\Omega)}^2 + \|\boldsymbol{E}^{1/2}\|_{H(\operatorname{curl},\Omega)}^2).$$

Hence, if $C\dfrac{\Delta t}{2h} < 1$, then the scheme is stable; that is,

$$\left(1 - C^2\frac{\Delta t^2}{4h^2}\right)\|\partial_\tau \boldsymbol{E}^{\mathcal{N}}\|_0^2 + \|\operatorname{curl}\boldsymbol{E}^{\mathcal{N}+1/2}\|_0^2$$

is bounded by the initial data. Therefore, we have shown that the following is true.

THEOREM 9. *If $1/4 \le \beta \le 1$, the Newmark scheme is unconditionally stable. If $0 \le \beta \le 1$, the Newmark scheme is conditionally stable. In particular, if $\beta = 0$, we impose the stability condition*

$$C\frac{\Delta t}{2h} < 1 \tag{5.12}$$

to have a stable scheme. Here, the constant $C > 0$ is determined by the inverse assumption.

Therefore, we conclude that if the solution $\boldsymbol{E}^n$ is assumed to be in $[H^2(\Omega_i)]^3$ for $i = 1, 2, \ldots, M$ and the stability of the Newmark scheme (with $\gamma = 1/2$) is achieved, then we obtain the optimal error estimate

$$\begin{aligned}\max_{1\le n\le \mathcal{N}}\|\boldsymbol{E}(t^n) - \boldsymbol{E}_h^n\|_{H(\operatorname{curl},\Omega)} &\le \max_{1\le n\le \mathcal{N}}\left\{\|\boldsymbol{E}(t^n) - \boldsymbol{E}^n\|_{H(\operatorname{curl},\Omega)} + \|\boldsymbol{E}^n - \boldsymbol{E}_h^n\|_{H(\operatorname{curl},\Omega)}\right\}\\ &= \mathcal{O}(\Delta t^2) + \mathcal{O}(h).\end{aligned}$$

REMARK 10. *Readers are referred to* [9, 10] *for further details on the regularity property of a solution to Maxwell's equations.*

6. Conclusion. We have presented a finite element/Fourier transform method for analyzing transient electromagnetic scattering from inhomogeneous cavities embedded in the infinite ground plane. Our method is shown to lead a well-posed semidiscrete problem in space. Stability conditions are given for the time-marching scheme. Convergence properties of the finite element scheme for the fully discrete problem are also discussed. Numerical experiments are performed for two-dimensional cavity problems that demonstrate the accuracy and stability of the method and are reported in [24].

We believe this is the first mathematical treatment of time-domain Maxwell's equations for scattering problems defined in an unbounded domain. Computationally, the exact nonlocal boundary condition introduced here has the potential to lead more accurate hybridization schemes (vs. those using ABC, PML, or the like) coupling the integral equation solution to the exterior infinite domain and the finite element treatment of the interior bounded domain.

REFERENCES

[1] H. AMMARI, G. BAO, AND A. WOOD, *A cavity problem for Maxwell's equations*, Methods Appl. Anal., 9 (2002), pp. 249–259.

[2] G. A. BAKER, *Error estimates for finite element methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 13 (1976), pp. 564–576.

[3] M. Cessenat, *Mathematical Methods in Electromagnetism. Linear Theory and Applications*, World Scientific, River Edge, NJ, 1996.
[4] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37 (2000), pp. 1542–1570.
[5] P. Ciarlet and J. Zou, *Fully discrete finite element approaches for time-dependent Maxwell's equations*, Numer. Math., 82 (1999), pp. 193–219.
[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North-Holland, New York, 1978, pp. 36–69.
[7] F. Collino and P. Monk, *The perfectly matched layer in curvilinear coordinates*, SIAM J. Sci. Comput., 19 (1998), pp. 2061–2090.
[8] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.
[9] M. Costabel, *A remark on the regularity of solutions to Maxwell's equations on Lipschitz domains*, Math. Methods Appl. Sci., 12 (1990), pp. 365–368.
[10] M. Costabel and M. Dauge, *Singularities of Electromagnetic Fields in Polyhedral Domains*, Technical report, IRMAR, Université de Rennes 1, France, 1997. Report available at http://www.maths.univ-rennes1.fr/~costabel/.
[11] L. Cowsar, T. F. Dupont, and M. F. Wheeler, *A priori estimates for mixed finite element methods for the wave equation*, Comput. Methods Appl., 82 (1990), pp. 205–222.
[12] S. Dodson, S. Walker, and M. Bluck, *Costs and cost scaling in time-domain integral-equation analysis of electromagnetic scattering*, IEEE Trans. Antennas and Propagation, 40 (1998), pp. 12–21.
[13] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*. Theory and Algorithms, Springer-Verlag, Berlin, 1986.
[14] D. Givoli, *Nonreflecting boundary conditions*, J. Comput. Phys., 94 (1991), pp. 1–29.
[15] D. Givoli and D. Cohen, *Nonreflecting boundary conditions based on Kirchhoff-type formulae*, J. Comput. Phys., 117 (1995), pp. 102–113.
[16] M. Grote and J. B. Keller, *Exact nonreflecting boundary conditions for the time dependent wave equation*, SIAM J. Appl. Math, 55 (1995), pp. 280–297.
[17] F. B. Hildebrand, *Advanced Calculus for Applications*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1976.
[18] D. Jiao, M. Lu, and E. M. and J. Jin, *A Fast Time-Domain Finite Element-Boundary Integral Method for Electromagnetic Analysis*, preprint, 2000.
[19] J. M. Jin, *The Finite Element Method in Electromagnetics*, John Wiley & Sons Inc., New York, 1993.
[20] A. Kirsch and P. Monk, *A finite element/spectral method for approximating the time-harmonic Maxwell system in* $\mathbb{R}^3$, SIAM J. Appl. Math., 55 (1995), pp. 1324–1344.
[21] A. Kirsch and P. Monk, *A finite element method for approximating electromagnetic scattering from a conducting object*, Numer. Math., 92 (2002) pp. 501–534.
[22] P. Monk, *Analysis of a finite element method for Maxwell's equations*, SIAM J. Numer. Anal., 29 (1992), pp. 714–729.
[23] B. Shanker, A. A. Ergin, K. Aygun, and E. Michiessen, *Analysis of transient electromagnetic scattering from closed surfaces using a combined field integral equation*, IEEE Trans. Antennas and Propagation, 48 (2000), pp. 1064–1074.
[24] T. Van and A. Wood, *A time-domain finite element method for Helmholtz equations*, J. Comput. Phys., 183 (2002), pp. 486–507.
[25] S. P. Walker, *Scattering analysis via time-domain integral equations: methods to reduce the scaling of costs with frequency*, IEEE Trans. Antennas and Propagation, 39 (1997), pp. 13–20.
[26] W. L. Wood, *Practical Time-Stepping Schemes*, Clarendon Press, London, 1990.
[27] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method. Solid and Fluid Mechanics. Dynamics and Non-Linearity*, Vol. II, 4th ed., McGraw-Hill, New York, 1991.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1610–1630

COMPUTING BOUNDS FOR LINEAR FUNCTIONALS OF EXACT WEAK SOLUTIONS TO POISSON'S EQUATION*

A. M. SAUER-BUDGE†, J. BONET‡, A. HUERTA§, AND J. PERAIRE†

Abstract. We present a method for Poisson's equation that computes guaranteed upper and lower bounds for the values of piecewise-polynomial linear functional outputs of the exact weak solution of the infinite-dimensional continuum problem with piecewise-polynomial forcing. The method results from exploiting the Lagrangian saddle point property engendered by recasting the output problem as a constrained minimization problem. Localization is achieved by Lagrangian relaxation and the bounds are computed by appeal to a local dual problem. The proposed method computes approximate Lagrange multipliers using traditional finite element approximations to calculate a primal and an adjoint solution along with well-known hybridization techniques to calculate interelement continuity multipliers. The computed bounds hold uniformly for any level of refinement, and in the asymptotic convergence regime of the finite element method, the bound gap decreases at twice the rate of the energy norm measure of the error in the finite element solution. Given a finite element solution and its output adjoint solution, the method can be used to provide a certificate of precision for the output with an asymptotic complexity that is linear in the number of elements in the finite element discretization. The elemental contributions to the bound gap are always positive and hence lend themselves to be used as adaptive indicators, as we demonstrate with a numerical example.

Key words. bounds, PDEs, Poisson, exact solutions, certificates, functional outputs, error estimation

AMS subject classifications. 65N15, 65N30, 49M29

DOI. 10.1137/S0036142903425045

1. Introduction.

Uncertainty about the reliability of numerical approximations frequently undermines the utility of field simulations in the engineering design process: simulations are often not trusted because they lack reliable feedback on accuracy, or are more costly than necessary because they are performed with greater fidelity than necessary in an attempt to bolster trust. In addition to devitalized confidence, numerical uncertainty often causes ambiguity about the source of any discrepancies when using simulation results in concert with experimental measurements. Can the discretization error account for the discrepancies, or is the underlying continuum model inadequate?

While confidence in the precision of a field simulation can be buoyed by performing convergence studies, such studies are computationally very expensive and in practice are often not performed at more than a few conditions, if at all, due to cost and time constraints. For this reason, researchers and practitioners employ adaptive methods to converge the solution in a manner that costs less in time and resources than uniform refinement. Adaptive methods powered by current error estimation technology, however, provide only asymptotic guarantees of precision, at best, and no guarantees of precision, at worse, since the convergence of adaptive methods remains an open question [12].

*Received by the editors March 26, 2003; accepted for publication (in revised form) May 1, 2004; published electronically December 27, 2004.

http://www.siam.org/journals/sinum/42-4/42504.html

†Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 (ambudge@alum.mit.edu, peraire@mit.edu). The research of these authors was supported by the Singapore-MIT Alliance.

‡Department of Civil Engineering, University of Wales Swansea, Singleton Park, Swansea SA2 8PP, Wales, UK (j.bonet@swansea.ac.uk).

§Departamento de Matematica Aplicada III, Universitat Politecnica de Catalunya, Campus Nord UPC, E-08034 Barcelona, Spain (antonio.huerta@upc.es).

Our observations of engineering practice inform us that integrated quantities such as forces and total fluxes are frequently queried quantitative outputs from field simulations and that design and analysis does not always require the full precision available. The primary objective of our method, therefore, is to certify the precision of integrated outputs for low-fidelity simulations as well as high-fidelity simulations. We call our bounds *uniform* to differentiate our goal of obtaining quantitative bounds for all levels of refinement from the lesser goal of obtaining quantitative bounds only asymptotically in the limit of refinement. In this regard, the complete procedure can be viewed as a polynomial time algorithm in the number of mesh elements that provides a certificate of precision for a predicted output. The certificate guarantees a minimum level of precision in the output from a particular *finite*-dimensional approximation with respect to the output from the *infinite*-dimensional model that it is approximating. Furthermore, the procedure provides local information that can be used in conjunction with adaptive meshing to efficiently drive a solution to an arbitrary and guaranteed precision.

Verification and a posteriori error analysis have a long history in the development of the finite element method with many different approaches forwarded and investigated. Ainsworth and Oden give a detailed overview of many of the approaches in [2]. Conceptually, our method descends from a long line of complementary energy methods beginning in the early 1970s when Fraeijs de Veubeke [7] proposed verifying the precision of a simulation by comparing the energy computed from a global primal approximation with the complementary energy computed from a global dual approximation. Global primal-dual methods offer a rich context for approximation, but suffer from the delicate nature of the global dual approximation, relatively high cost, and for verification, from a lack of relevant measure because the upper and lower bounding properties hold only for the total energy.

Much more closely related to our work are the works of Ladevèze and Rougeot [10], Ladeveze and Leguillon [9], Ainsworth and Oden [1], and of Destuynder and Métivet [6], all of which consider local complementary energy problems for developing estimates for the energy norm of the error. In contrast to the work of Ladevèze, we endeavor to compute uniformly guaranteed two-sided bounds on an output, not an estimate of the error in an abstract norm. While the work of Ainsworth and Oden as well as the related work of Cao, Kelly, and Sloan [5] require the exact solution of infinite-dimensional local problems in order to guarantee bounds, our method guarantees bounds uniformly with the solution of a finite-dimensional local problem. Our method differs from that of Destuynder and Métivet in that it is not burdened with the explicit construction of globally conforming approximations to dual admissible vector fields. The work we present here extends earlier work done by Paraschivoiu, Peraire, and Patera [14] and Paraschivoiu and Patera [13] on two-level residual based techniques for computing output bounds.

In this paper, we focus on the overarching structure of the method and do not consider the details of its implementation, nor extensions to non-piecewise-polynomial forcing or curved domains, nor more general equations such as nonsymmetric dissipative operators, which will be presented in a future publication [18]. Section 2 presents the core concepts in the simpler setting of energy bounds, where the method has a clear variational meaning and a direct relationship to hybrid methods. Section 3 recasts the energy bound method as a method for linear functional output bounds, simultaneously carrying out an explicit extension to more relevant error measures and an implicit extension to nonvariational problems. Finally, the last section demon-

strates the method with numerical results for three example problems. The last of the three examples shows how the method can be used to drive an adaptive refinement process.

1.1. Poisson's equation. We consider Poisson's equation posed on polygonal domains, Ω, in d spacial dimensions and, only for the sake of simplicity of presentation, homogeneous Dirichlet boundaries, $\Gamma = \partial\Omega$. The Poisson problem is formulated weakly as: find $u \in \mathcal{U}$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\Omega = \int_\Omega f \, v \, \mathrm{d}\Omega \quad \forall v \in \mathcal{U}, \tag{1.1}$$

where $\mathcal{U}(\Omega) \equiv \{ u \in \mathcal{H}^1(\Omega) \mid u|_\Gamma = 0 \}$ and the domain Ω is assumed when otherwise unspecified, that is, $\mathcal{U} \equiv \mathcal{U}(\Omega)$. As a consequence of all the Dirichlet boundaries being homogeneous, $\mathcal{U}$ serves as both the function set and test space in our presentation. While we present the method for homogeneous Dirichlet data, it can be easily extended to nonhomogeneous data and Neumann boundary conditions.

2. Computing energy bounds. We begin by developing a lower bound on the total energy of the system, $\frac{1}{2}\int_\Omega \nabla u \cdot \nabla u \, \mathrm{d}\Omega - \int_\Omega f \, u \, \mathrm{d}\Omega$, which in the context of heat conduction, combines the heat dissipation energy, $\frac{1}{2}\int_\Omega \nabla u \cdot \nabla u \, \mathrm{d}\Omega$, and the potential energy of the thermal loads, $-\int_\Omega f \, u \, \mathrm{d}\Omega$. There is a well-known physical principle at work in this problem, related to the symmetric positive definite nature of the diffusion operator, which states that the solution, u, is the function that minimizes the total energy with respect to all other candidates in $\mathcal{U}$,

$$u = \arg \inf_{w \in \mathcal{U}} \frac{1}{2} \int_\Omega \nabla w \cdot \nabla w \, \mathrm{d}\Omega - \int_\Omega f \, w \, \mathrm{d}\Omega, \tag{2.1}$$

as can easily be verified by comparing the Euler–Lagrange equation of this minimization statement to Poisson's equation (1.1). This minimization formulation makes it clear that if we look for a discrete approximation of (1.1) in a finite set of conforming functions, $\mathcal{U}_h$, for which $\mathcal{U}_h \subset \mathcal{U}$, then the resulting total energy predicted by the approximation will approach the exact value from above.

While insightful, this upper bound on the total energy has limited usefulness for two primary reasons. First, only rarely will the total energy be relevant to the purpose of solving the original problem. Second, even when it is relevant, the upper bound will most likely not be helpful for managing approximation uncertainty. In an engineering design task, the upper bound usually corresponds to the "best case scenario," as opposed to the "worst case scenario" which would be required to ensure feasibility of the design.

Our strategy for obtaining lower bounds on the energy in a cost efficient manner is to first decompose the global problem into independent local elemental subproblems by relaxing the continuity of the set $\mathcal{U}$ along edges of a triangular partitioning of Ω, using approximate Lagrange multipliers, then accumulate the lower bound from the objective values of approximate local dual subproblems.

2.1. Weak continuity reformulation. We begin by partitioning the domain into a mesh, $\mathcal{T}_h$, of nonoverlapping open subdomains, T, called elements, for which $\bigcup_{T \in \mathcal{T}_h} \overline{T} = \bar{\Omega}$. We denote by ∂T the edges, γ, constituting the boundary of a single element T, and by $\partial\mathcal{T}_h$ the network of all edges in the mesh. We have not yet evoked a

discretization of $\mathcal{U}$, but merely a domain decomposition represented by a mesh. With the broken space

$$\hat{\mathcal{U}} \equiv \left\{ v \in L^2(\Omega) \;\middle|\; v|_T \in \mathcal{H}^1(T)\ \forall T \in \mathcal{T}_h \right\}, \tag{2.2}$$

in which the continuity of $\mathcal{U}$ is broken across the mesh edges, $\partial\mathcal{T}_h$, we can reformulate the energy minimization statement (2.1) by explicitly enforcing continuity

$$\begin{aligned} u = \arg \inf_{\hat{w}\in\hat{\mathcal{U}}} \quad & \frac{1}{2}\int_\Omega \nabla\hat{w}\cdot\nabla\hat{w}\,\mathrm{d}\Omega - \int_\Omega f\,\hat{w}\,\mathrm{d}\Omega \\ \text{s.t.} \quad & \sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T\,\lambda\,\hat{w}\,\mathrm{d}\Gamma = 0 \quad \forall\lambda\in\Lambda, \end{aligned} \tag{2.3}$$

where, for $T, T_{\mathrm{N}} \in \mathcal{T}_h$ and an arbitrary ordering of the elements,

$$\sigma_T(x) = \begin{cases} -1, & x \in \overline{T}\cap\overline{T}_{\mathrm{N}}, T < T_{\mathrm{N}}, \\ +1, & \text{otherwise.} \end{cases} \tag{2.4}$$

Integrals over the broken domain, such as $\int_\Omega \nabla\hat{w}\cdot\nabla\hat{v}\,\mathrm{d}\Omega$, are understood as sums of integrals over the subdomains, such as $\sum_{T\in\mathcal{T}_h}\int_T \nabla\hat{w}|_T\cdot\nabla\hat{v}|_T\,\mathrm{d}\Omega$. As there is no ambiguity, we have suppressed the trace operators from our notation for the boundary integrals to simplify the appearance of the expressions.

To see how the constraint arises, consider a single edge, $\gamma \in \partial\mathcal{T}_h$, with neighboring elements T and T_{N}, for which a strong continuity constraint can be written roughly as $\hat{w}|_{T,\gamma} - \hat{w}|_{T_{\mathrm{N}},\gamma} = 0$ on γ. An integral weak representation is obtained by multiplying by an arbitrary test function, λ_γ, taken from an appropriate space, $\Lambda(\gamma)$, integrating along the edge, and ensuring the resulting integrated quantity is zero for all possible test functions: $\int_\gamma (\hat{w}|_{T,\gamma} - \hat{w}|_{T_{\mathrm{N}},\gamma})\,\lambda_\gamma\,\mathrm{d}\Gamma = 0 \quad \forall\lambda_\gamma\in\Lambda(\gamma)$. The constraint used above is obtained by rewriting the combination of all edge constraints as a combination of elemental contributions, using σ_T to track the sign of the contribution. Since $\hat{w}|_T$ is a member of $\mathcal{H}^1(T)$, the trace of $\hat{w}|_T$ on an edge γ is a member of $\mathcal{H}^{\frac{1}{2}}(\partial T)$. Therefore, λ on γ is a member of the dual of the trace space, $\mathcal{H}^{-\frac{1}{2}}(\gamma)$, and the continuity multiplier space Λ is the corresponding product space taken over all the edges of the mesh.

Notice that we have relaxed the Dirichlet boundary conditions as well as the interior continuity. The homogeneous Dirichlet conditions are weakly enforced implicitly by the continuity constraint. We shall not prove it here, but it is important to know that the minimizer of the constrained minimization problem (2.3) is indeed u, the exact solution of Poisson's equation (1.1) [2, 4].

2.2. Localization by continuity relaxation. Considering the Lagrangian of the constrained minimization (2.3),

$$\mathcal{L}(\hat{w};\lambda) \equiv \frac{1}{2}\int_\Omega \nabla\hat{w}\cdot\nabla\hat{w}\,\mathrm{d}\Omega - \int_\Omega f\,\hat{w}\,\mathrm{d}\Omega - \sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T\,\lambda\,\hat{w}\,\mathrm{d}\Gamma, \tag{2.5}$$

we recall from the saddle point property of Lagrange multipliers and the strong duality of convex minimizations that $\forall\tilde{\lambda}\in\Lambda$ there exists a lower energy bound, ε^-, satisfying

$$\varepsilon^- \le \inf_{\hat{w}\in\hat{\mathcal{U}}}\mathcal{L}(\hat{w};\tilde{\lambda}) \le \sup_{\lambda\in\Lambda}\inf_{\hat{w}\in\hat{\mathcal{U}}}\mathcal{L}(\hat{w};\lambda) = \inf_{\hat{w}\in\hat{\mathcal{U}}}\sup_{\lambda\in\Lambda}\mathcal{L}(\hat{w};\lambda) = \varepsilon,$$

where the value at optimality is the minimum total energy of the continuum system, $\varepsilon = \frac{1}{2}\int_\Omega \nabla u \cdot \nabla u \,\mathrm{d}\Omega - \int_\Omega f\, u \,\mathrm{d}\Omega$. The lower bounding minimization for a given $\tilde{\lambda}$ is separable, an important property allowing us to treat each element independently. In order to obtain a lower bound, $\tilde{\lambda}$ cannot be chosen arbitrarily. We obtain $\tilde{\lambda}$ by approximating the problem using finite elements in a manner that guarantees that the relaxed minimization is bounded from below.

2.2.1. Continuity multiplier approximation. We now introduce the finite element approximation of Poisson's equation (1.1) as means of obtaining an approximate Lagrange multiplier. We first solve the finite-dimensional Poisson problem: find $u_h \in \mathcal{U}_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v \,\mathrm{d}\Omega = \int_\Omega f\, v \,\mathrm{d}\Omega \quad \forall v \in \mathcal{U}_h, \tag{2.6}$$

where $\mathcal{U}_h \equiv \{\, v \in \mathcal{U} \mid v|_T \in \mathbb{P}^p(T)\ \forall T \in \mathcal{T}_h \,\}$ for $\mathbb{P}^p(T)$ is the space of polynomials on element T (in d spacial dimensions) with degree less than or equal to p. Along with $\mathcal{U}_h$, we introduce the broken discrete space $\hat{\mathcal{U}}_h \equiv \{\, v \in \hat{\mathcal{U}} \mid v|_T \in \mathbb{P}^p(T)\ \forall T \in \mathcal{T}_h \,\}$ and the companion discrete Lagrange multiplier space $\Lambda_h \equiv \{\, \lambda \in \Lambda \mid \lambda|_\gamma \in \mathbb{P}^p(\gamma)\ \forall \gamma \in \partial\mathcal{T}_h \,\}$, where $\mathbb{P}^p(\gamma)$ is the space of polynomials on element edge γ (in $d-1$ spacial dimensions) with degree less than or equal to p.

Once we have obtained u_h, we solve the gradient condition of (2.5) to obtain λ_h: find $\lambda_h \in \Lambda_h$ such that

$$\sum_{T\in\mathcal{T}_h} \int_{\partial T} \sigma_T\, \lambda_h\, \hat{v} \,\mathrm{d}\Gamma = \int_\Omega \nabla u_h \cdot \nabla \hat{v} \,\mathrm{d}\Omega - \int_\Omega f\, \hat{v} \,\mathrm{d}\Omega \quad \forall \hat{v} \in \hat{\mathcal{U}}_h. \tag{2.7}$$

We call this the equilibration problem, and we call any compatible Lagrange multiplier "equilibrating," since the problem has a nonunique solution. In the context of hybrid methods [4], this continuity multiplier is often referred to as a hybrid flux. As mentioned previously, this particular choice for the Lagrange multiplier ensures a finite lower bound.

LEMMA 2.1. *If a Lagrange multiplier $\lambda_h \in \Lambda_h$ satisfies the equilibration condition* (2.7), *then* $\inf_{\hat{w}\in\hat{\mathcal{U}}} \mathcal{L}(\hat{w};\lambda_h)$ *is bounded from below.*

Proof. Recall that the null space for the Poisson operator is the one-dimensional space of constants, $\mathbb{P}^0$, and let $\hat{\mathbb{P}}^0 = \prod_{T\in\mathcal{T}_h} \mathbb{P}^0(T)$ denote the null space of the broken operator. Considering $\hat{c} \in \hat{\mathbb{P}}^0 \subset \hat{\mathcal{U}}_h$ in the equilibration problem (2.7) and that any $\hat{w} \in \hat{\mathcal{U}}$ can be represented as $\hat{w}' + \hat{c}$ for $\hat{w}' \in \hat{\mathcal{U}} \setminus \hat{\mathbb{P}}^0$, it is easily shown that $\mathcal{L}(\hat{w}' + \hat{c};\lambda_h) = \mathcal{L}(\hat{w}';\lambda_h)$. For the Poisson equation, equilibration ensures that null space of the operator does not cause the minimization to become unbounded below. The existence of a minimum now follows from the coercivity of the Poisson operator in $\hat{\mathcal{U}} \setminus \hat{\mathbb{P}}^0$. ∎

While not part of the classical finite element problem set, the equilibration problem has been addressed a number of times and in a number of contexts in the finite element community, not the least of which is in the context of error estimation. For our implementation, we use a method due to Ladeveze and Leguillon [9] and Ainsworth and Oden [2] which has an asymptotically linear computational cost in the number of mesh vertices.

2.3. Local dual subproblem. Now that we have successfully decomposed the global problem into local elemental subproblems and determined a suitable approxi-

mation λ_h for our continuity multiplier $\tilde{\lambda}$, we can write the lower bounding minimization induced by the Lagrange saddle point property as

$$\inf_{\hat{w}\in\hat{\mathcal{U}}} \mathcal{L}(\hat{w};\tilde{\lambda}) = \sum_{T\in\mathcal{T}_h} \inf_{w\in\mathcal{U}(T)} J_T(w)$$

for

$$J_T(w) \equiv \frac{1}{2}\int_T \nabla w\cdot\nabla w\,\mathrm{d}\Omega - \int_T f\,w\,\mathrm{d}\Omega - \int_{\partial T}\sigma_T\,\tilde{\lambda}\,w\,\mathrm{d}\Gamma, \tag{2.8}$$

and consider a representative minimization subproblem. The minimization subproblem simply corresponds to a Poisson problem of the type represented in equation (1.1) with Neumann boundary conditions posed on a single subdomain. We have done nothing to change the nature of the original problem, but have only acted to decompose the global problem into a sequence of independent local problems.

We do not require, and in general cannot compute, the exact minimum of the infinite-dimensional local subproblem, but we do require a lower bound for it and we proceed now to introduce the primary ingredient for obtaining this local lower bound.

PROPOSITION 2.2. *If we define the positive functional*

$$J_T^c(\mathbf{q}) \equiv \frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\,\mathrm{d}\Omega, \tag{2.9}$$

where $\mathbf{q}\in\mathcal{H}(\mathrm{div};T)$ *and* $\mathcal{H}(\mathrm{div};T)\equiv\left\{\mathbf{q}\,|\,\mathbf{q}\in(L^2(T))^d,\ \nabla\cdot\mathbf{q}\in L^2(T)\right\}$ *for a problem posed in d spacial dimensions, then we have*

$$J_T(w) \geq -J_T^c(\mathbf{q}) \quad \forall w\in\mathcal{H}^1(T),\ \forall\mathbf{q}\in\mathcal{Q}(T;\tilde{\lambda}), \tag{2.10}$$

for the set of functions

$$\mathcal{Q}(T;\tilde{\lambda}) \equiv \left\{\mathbf{q}\in\mathcal{H}(\mathrm{div};T)\;\middle|\; \int_T \nabla\cdot\mathbf{q}\,v\,\mathrm{d}\Omega - \int_{\partial T}\mathbf{q}\cdot\mathbf{n}\,v\,\mathrm{d}\Gamma = -\int_T f\,v\,\mathrm{d}\Omega - \int_{\partial T}\sigma_T\tilde{\lambda}\,v\,\mathrm{d}\Gamma \quad \forall v\in\mathcal{H}^1(T)\right\}. \tag{2.11}$$

Proof. We begin by appealing to the following positive expression:

$$\frac{1}{2}\int_T (\mathbf{q}-\nabla w)^2\,\mathrm{d}\Omega \geq 0$$

for any $w\in\mathcal{H}^1(T)$ and any $\mathbf{q}\in\mathcal{Q}(T;\tilde{\lambda})$. This expression expands to

$$\frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\,\mathrm{d}\Omega + \frac{1}{2}\int_T \nabla w\cdot\nabla w\,\mathrm{d}\Omega - \int_T \mathbf{q}\cdot\nabla w\,\mathrm{d}\Omega \geq 0,$$

in which we apply the Green's identity $-\int_T \mathbf{q}\cdot\nabla w\,\mathrm{d}\Omega = \int_T \nabla\cdot\mathbf{q}\,w\,\mathrm{d}\Omega - \int_{\partial T}\mathbf{q}\cdot\mathbf{n}\,w\,\mathrm{d}\Gamma$ to obtain

$$\frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\,\mathrm{d}\Omega + \frac{1}{2}\int_T \nabla w\cdot\nabla w\,\mathrm{d}\Omega + \int_T \nabla\cdot\mathbf{q}\,w\,\mathrm{d}\Omega - \int_{\partial T}\mathbf{q}\cdot\mathbf{n}\,w\,\mathrm{d}\Gamma \geq 0. \tag{2.12}$$

The constraint contained in $\mathcal{Q}(T;\tilde{\lambda})$ makes this expression equivalent to

$$\frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\,\mathrm{d}\Omega + \frac{1}{2}\int_T \nabla w\cdot\nabla w\,\mathrm{d}\Omega - \int_T f\,w\,\mathrm{d}\Omega - \int_{\partial T}\sigma_T\,\tilde{\lambda}\,w\,\mathrm{d}\Gamma \geq 0. \tag{2.13}$$

Identifying $J_T(w)$ and $J_T^c(\mathbf{q})$ we arrive at the desired expression for the local lower bound. ∎

The best possible local lower bound can be obtained with the following maximization problem:

$$\sup_{\mathbf{q}\in\mathcal{Q}(T;\tilde{\lambda})} -J_T^c(\mathbf{q}) \leq \inf_{w\in\mathcal{U}(T)} J_T(w),$$

for which we will obtain equality as a result of the convexity of J_T and J_T^c. It is clear that we have derived a classic dual formulation[1] for our local elemental minimization problem and essentially transformed a primal minimization problem into a dual feasibility problem. As we have alluded to earlier, the functional $J_T^c(\mathbf{q})$ is often called the *complementary energy* functional [16], when taken over the whole domain, Ω, with a globally admissible complementary field.

2.3.1. Subproblem approximation. Significantly, we can make these subproblems computable by choosing an appropriate finite-dimensional set in which to search for $\mathbf{q}$. At the very least the set must be chosen so that the divergence of its functions contain the forcing function, f, in T and the normal traces of its functions contain the approximate continuity multiplier, λ_h, on ∂T. In multiple dimensions, however, the polynomial approximation for the continuity multiplier will nullify any components of the set with nonpolynomial normal trace. Therefore, we choose the polynomial approximation subset

$$\begin{aligned}\mathcal{Q}_h(T) \equiv \Bigg\{ \mathbf{q}\in(\mathbb{P}^q(T))^d \;\Bigg|\; & \int_T \nabla\cdot\mathbf{q}\,v\,\mathrm{d}\Omega - \int_{\partial T}\mathbf{q}\cdot\mathbf{n}\,v\,\mathrm{d}\Gamma \\ & = -\int_T f\,v\,\mathrm{d}\Omega - \int_{\partial T}\sigma_T\lambda_h\,v\,\mathrm{d}\Gamma\ \forall v\in\mathcal{H}^1(T)\Bigg\}\end{aligned} \tag{2.14}$$

with $q\geq p$. As a consequence, the method as we have presented it is limited to forcing functions, $f|_T$, that are members of the polynomial space $\mathbb{P}^r(T)$ for $q>r$ on each elemental domain. While in one dimension we gain no advantage in taking q greater than $r+1$, in multiple dimensions we can do so in an attempt to sharpen the bounds. The interior constraint data, f, and the boundary constraint data, $\sigma_T\lambda_h$, cannot be chosen independently of each other, but must satisfy a compatibility condition in order to ensure solvability as manifest by the following lemma.

LEMMA 2.3. *Suppose the forcing function $f|_T$ is a member of $\mathbb{P}^r(T)$ and that λ_h satisfies* (2.7)*, then there exists at least one dual feasible function, $\mathbf{q}$, that is a member of $\mathcal{Q}_h(T)$ for $q\geq p$ and $q>r$.*

Proof. We begin by expressing $\mathbf{q}$, a member of $(\mathbb{P}^q(T))^d$, as the combination $\mathbf{q}=\mathbf{q}_D+\mathbf{q}_0$, with $\mathbf{q}_D$ a normal boundary condition satisfying component, $\mathbf{q}_D\cdot\mathbf{n}=$

[1] The classic derivation for the dual of the Poisson problem would begin by letting $\mathbf{q}=\nabla w$ (a statement of Fourier's law in the context of heat conduction) and proceed by eliminating w from the problem.

$\sigma_T \lambda_h$ on ∂T, and $\mathbf{q}_0$ a homogeneous normal boundary condition satisfying component, $\mathbf{q}_0 \cdot \mathbf{n} = 0$ on ∂T. With this lifting, we can write the feasibility constraint as

$$-\int_T \nabla \cdot \mathbf{q}_0 \, v \, \mathrm{d}\Omega = \int_T f \, v \, \mathrm{d}\Omega + \int_T \nabla \cdot \mathbf{q}_D \, v \, \mathrm{d}\Omega.$$

Recognizing the divergence operator on the left-hand side, which maps $(\mathbb{P}^q(T))^d$ into $\mathbb{P}^{q-1}(T)$, we note that we need only test against $v \in \mathbb{P}^{q-1}(T)$. Furthermore, finite-dimensional linear equations are solvable if and only if the right-hand side data lies in the range of the operator, which is orthogonal to the null space of the adjoint operator. The adjoint operator is easily found to be $\int_T \mathbf{q}_0 \cdot \nabla v \, \mathrm{d}\Omega$ which has the null space $v \in \mathbb{P}^0(T)$, and thus the right-hand side data must be in $\mathbb{P}^{q-1}(T) \setminus \mathbb{P}^0(T)$.

To prove solvability, we need only to verify that the right-hand side data is orthogonal to the constants, since the requirements that $q \geq p$ and $q > r$ ensure that the right-hand side data is in $\mathbb{P}^{q-1}$. Choosing $v = \text{const}$ in the right-hand side of the constraint, rewritten as

$$\int_T f \, v \, \mathrm{d}\Omega + \int_T \nabla \cdot \mathbf{q}_D \, v \, \mathrm{d}\Omega = \int_T f \, v \, \mathrm{d}\Omega - \int_T \mathbf{q}_D \cdot \nabla v \, \mathrm{d}\Omega + \int_{\partial T} \sigma_T \lambda_h \, v \, \mathrm{d}\Gamma,$$

reveals the compatibility condition

$$\int_{\partial T} \sigma_T \lambda_h \, \mathrm{d}\Gamma = -\int_T f \, \mathrm{d}\Omega, \tag{2.15}$$

which is satisfied by our choice for λ_h, as can be seen by choosing $\hat{v} = \text{const}$ on T in the equilibration condition (2.7). The equilibration condition thus ensures that the constraint data is compatible and that there exists at least one $\mathbf{q}$ satisfying the constraint. ∎

2.4. Energy bound procedure. In discussing the global procedure and its properties, we denote the global aggregate of independent elemental quantities by accenting them with a diacritical hat as we did for the global broken quantities, and we denote the aggregate of local functional forms by dropping the subscript T. In particular, $\hat{\mathcal{Q}}_h$ denotes the aggregate approximate dual function space, $\prod_{T \in \mathcal{T}_h} \mathcal{Q}_h(T)$, and $J^c(\hat{\mathbf{q}})$ the aggregate dual energy functional, $\sum_{T \in \mathcal{T}_h} J^c_T(\mathbf{q}|_T)$. The complete method for the energy bounds consists of three steps.

1. *Global approximation:* Find $u_h \in \mathcal{U}_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v \, \mathrm{d}\Omega = \int_\Omega f \, v \, \mathrm{d}\Omega \quad \forall v \in \mathcal{U}_h, \tag{2.16}$$

   and calculate the upper bound $\varepsilon_h^+ = -\frac{1}{2} \int_\Omega \nabla u_h \cdot \nabla u_h \, \mathrm{d}\Omega$.
2. *Global equilibration:* Find $\lambda_h \in \Lambda_h$ such that

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} \sigma_T \, \lambda_h \, \hat{v} \, \mathrm{d}\Gamma = \int_\Omega \nabla u_h \cdot \nabla \hat{v} \, \mathrm{d}\Omega - \int_\Omega f \, \hat{v} \, \mathrm{d}\Omega \quad \forall \hat{v} \in \hat{\mathcal{U}}_h. \tag{2.17}$$

3. *Local dual approximations:* Find ε_h^- such that

$$\varepsilon_h^- = \sup_{\hat{\mathbf{q}}_h \in \hat{\mathcal{Q}}_h} -J^c(\hat{\mathbf{q}}_h). \tag{2.18}$$

The last step requires the solution of a series of finite-dimensional quadratic programming problems with convex objective functions and linear equality constraints. The per-element cost remains low due to the small size of the elemental subproblems, while the total cost of computing the lower bound is asymptotically linear in the number elements.

2.4.1. Properties of the energy bound. As previously discussed, the upper bound follows directly from the conforming nature of the finite element approximation and the lower bound follows directly from Proposition 2.2. We close our presentation of the energy bound method by showing that the lower bound converges at the same rate as the upper bound, and thus inherits the well-known a priori finite element convergence property for the energy norm of the error. We begin by proving an orthogonality result.

LEMMA 2.4. *Let $\hat{\mathbf{p}}$ be any dual feasibility correction to ∇u_h such that $\hat{\mathbf{q}} = \nabla u_h + \hat{\mathbf{p}}$ is a member of $\hat{\mathcal{Q}}(\lambda_h)$, then $\hat{\mathbf{p}}$ satisfies the orthogonality property*

$$\sum_{T\in\mathcal{T}_h} \int_T \hat{\mathbf{p}} \cdot \nabla \hat{v} \, \mathrm{d}\Omega = 0 \quad \forall \hat{v} \in \hat{\mathcal{U}}_h. \tag{2.19}$$

Proof. We begin by examining the condition that the feasibility correction $\hat{\mathbf{p}}$ must satisfy by substituting $\nabla u_h + \hat{\mathbf{p}}$ into the constraint contained in the definition of $\hat{\mathcal{Q}}(\lambda_h)$ to obtain

$$\begin{aligned} \int_\Omega \nabla \cdot \hat{\mathbf{p}} \, \hat{v} \, \mathrm{d}\Omega - \sum_{T\in\mathcal{T}_h} \int_{\partial T} \hat{\mathbf{p}} \cdot \mathbf{n} \, \hat{v} \, \mathrm{d}\Gamma = &- \int_\Omega f \, \hat{v} \, \mathrm{d}\Omega - \int_\Omega \nabla \cdot \nabla u_h \, \hat{v} \, \mathrm{d}\Omega \\ &- \sum_{T\in\mathcal{T}_h} \int_{\partial T} \sigma_T \lambda_h \, \hat{v} \, \mathrm{d}\Gamma + \sum_{T\in\mathcal{T}_h} \int_{\partial T} \nabla u_h \cdot \mathbf{n} \, \hat{v} \, \mathrm{d}\Gamma \quad \forall \hat{v} \in \hat{\mathcal{U}}_h. \end{aligned} \tag{2.20}$$

Applying Green's formula to both the $\hat{\mathbf{p}}$ and u_h terms yields the equivalent constraint

$$\int_\Omega \hat{\mathbf{p}} \cdot \nabla \hat{v} \, \mathrm{d}\Omega = \int_\Omega f \, \hat{v} \, \mathrm{d}\Omega - \int_\Omega \nabla u_h \cdot \nabla \hat{v} \, \mathrm{d}\Omega + \sum_{T\in\mathcal{T}_h} \int_{\partial T} \sigma_T \lambda_h \, \hat{v} \, \mathrm{d}\Gamma \quad \forall \hat{v} \in \hat{\mathcal{U}}_h. \tag{2.21}$$

Restricting $\hat{v}$ to $\hat{\mathcal{U}}_h$ produces the sought orthogonality property as a consequence of equilibration (2.17). ∎

LEMMA 2.5. *Let $\hat{\mathbf{p}}_h^*$ be the dual feasibility correction to ∇u_h that maximizes $-J^c(\hat{\mathbf{p}}_h)$ such that $\nabla u_h + \hat{\mathbf{p}}_h^*$ is a member of $\hat{\mathcal{Q}}_h$, then $J^c(\hat{\mathbf{p}}_h^*)$ is bounded from above by*

$$J^c(\hat{\mathbf{p}}_h^*) \leq C|u - u_h|_1^2, \tag{2.22}$$

for the seminorm $|v|_1^2 \equiv \int_\Omega \nabla v \cdot \nabla v \, \mathrm{d}\Omega$, if the approximate continuity multiplier λ_h computed in (2.17) *has the bound*

$$\sum_{T\in\mathcal{T}_h} h^{\frac{1}{2}} \|\lambda - \lambda_h\|_{\partial T} \leq C|u - u_h|_1, \tag{2.23}$$

where $\lambda|_{\partial T} \equiv \sigma_T \frac{\partial u}{\partial \mathbf{n}}$ is the exact continuity multiplier and $\|v\|_{\partial T}^2 \equiv \int_{\partial T} v^2 \, \mathrm{d}\Gamma$. Everywhere, C is a generic constant independent of $h = \mathrm{diam}(T)$.

Proof. Using (2.21), the constrained maximization for the continuous dual feasible correction $\hat{\mathbf{p}}^*$ can be written as $\sup_{\hat{\mathbf{p}}\in\hat{\mathcal{Q}}(\lambda_h)} -\frac{1}{2}\int_\Omega \hat{\mathbf{p}}\cdot\hat{\mathbf{p}}\,\mathrm{d}\Omega$. Explicitly, the constraint for this maximization problem is written as

$$-\int_\Omega \hat{\mathbf{p}}\cdot\nabla\hat{\phi}\,\mathrm{d}\Omega = \int_\Omega \nabla u_h\cdot\nabla\hat{\phi}\,\mathrm{d}\Omega - \int_\Omega f\,\hat{\phi}\,\mathrm{d}\Omega - \sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T\lambda_h\,\hat{\phi}\,\mathrm{d}\Gamma \quad \forall\hat{\phi}\in\hat{\mathcal{U}}. \tag{2.24}$$

The gradient condition then informs us that $\hat{\mathbf{p}}^* = \nabla\hat{\phi}^*$.

The approximate solution u_h has an associated approximate continuity multiplier λ_h satisfying (2.17), while the exact solution u also has an associated exact continuity multiplier λ satisfying

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T\,\lambda\,\hat{v}\,\mathrm{d}\Gamma = \int_\Omega \nabla u\cdot\nabla\hat{v}\,\mathrm{d}\Omega - \int_\Omega f\,\hat{v}\,\mathrm{d}\Omega \quad \forall\hat{v}\in\hat{\mathcal{U}}, \tag{2.25}$$

as can be verified by integration by parts. Adding (2.25) to the constraint of (2.24) with $\hat{\mathbf{p}} = \hat{\mathbf{p}}^*$ and $\hat{v} = \hat{\phi}^*$ we find for $\|\hat{v}\|^2 = \sum_{T\in\mathcal{T}_h}\int_T v^2\,\mathrm{d}\Omega$ that

$$\begin{aligned}\int_\Omega \hat{\mathbf{p}}^*\cdot\nabla\hat{\phi}^*\,\mathrm{d}\Omega &= \int_\Omega \nabla(u-u_h)\cdot\nabla\hat{\phi}^*\,\mathrm{d}\Omega - \sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T(\lambda-\lambda_h)\hat{\phi}^*\,\mathrm{d}\Gamma\\ &\le C|u-u_h|_1\|\nabla\hat{\phi}^*\| + \sum_{T\in\mathcal{T}_h} C\|\lambda-\lambda_h\|_{\partial T}\|\hat{\phi}^*\|_{\partial T}\\ &\le C|u-u_h|_1\|\nabla\hat{\phi}^*\| + \sum_{T\in\mathcal{T}_h} Ch^{\frac{1}{2}}\|\lambda-\lambda_h\|_{\partial T}\|\nabla\hat{\phi}^*\|,\end{aligned}$$

in which we applied the inequality $\|w\|_{\partial T}\le Ch^{\frac{1}{2}}|w|_{1,T}$, valid for any $w\in\mathcal{H}^1(T)$ that has zero mean [11].

We then invoke the bound (2.23) and substitute $\nabla\hat{\phi}^* = \hat{\mathbf{p}}^*$ before dividing both sides by $\|\hat{\mathbf{p}}^*\|$ and recognizing that $\|\hat{\mathbf{p}}^*\|^2 = 2J^c(\hat{\mathbf{p}}^*)$ to obtain

$$J^c(\hat{\mathbf{p}}^*)\le C|u-u_h|_1^2.$$

The proof is completed by showing that $J^c(\hat{\mathbf{p}}_h^*)\le\tilde{C}J^c(\hat{\mathbf{p}}^*)$. This can be done by noting that $\hat{\mathbf{p}}_h^*$ can be obtained from $\hat{\mathbf{p}}^*$ by a bounded projection. A standard scaling argument can then be used to show that the constant $\tilde{C}$ is independent of h; see [4, 17]. ∎

Ainsworth and Oden prove in [1] that under certain assumptions the flux average of the finite element solution across the edges is bounded by (2.23) so that, by way of the triangle inequality, the burden rests in showing that the nonunique equilibrating corrections required to satisfy (2.17) decrease at the requisite rate. Maday and Patera give in [11] a basic method for computing approximate continuity multipliers that has been proven a priori to satisfy (2.23).

PROPOSITION 2.6. *Suppose that λ_h is the solution of the equilibration problem* (2.17) *for u_h the solution of the finite element approximation problem* (2.16)*, then*

$$\varepsilon-\varepsilon^-\le C|u-u_h|_1^2. \tag{2.26}$$

Proof. Let $\hat{\mathbf{p}}_h^*$ be chosen according to Lemma 2.5, then

$$-J^c(\nabla u_h + \hat{\mathbf{p}}_h^*) \leq \sup_{\hat{\mathbf{q}}_h \in \hat{\mathcal{Q}}_h} -J^c(\hat{\mathbf{q}}_h) = -J^c(\hat{\mathbf{q}}_h^*)$$

for $\hat{\mathbf{q}}_h^* = \arg\sup_{\hat{\mathbf{q}}_h \in \hat{\mathcal{Q}}_h} -J^c(\hat{\mathbf{q}}_h)$. From this relationship and from the definition of $\hat{\mathbf{p}}_h^*$ we know that $J^c(\hat{\mathbf{q}}_h^*) \leq J^c(\nabla u_h) + J^c(\hat{\mathbf{p}}_h^*)$, because $\sum_{T \in \mathcal{T}_h} \int_T \hat{\mathbf{p}}_h^* \cdot \nabla u_h \, \mathrm{d}\Omega = 0$ from Lemma 2.4 and the fact that u_h is a member of $\hat{\mathcal{U}}_h$. Adding the exact energy $\varepsilon = -J^c(\nabla u)$ to each side and recalling that $\varepsilon_h^+ = -J^c(\nabla u_h)$ and $\varepsilon_h^- = -J^c(\hat{\mathbf{q}}_h^*)$, we have our desired result

$$\varepsilon - \varepsilon_h^- \leq \varepsilon - \varepsilon_h^+ + J^c(\hat{\mathbf{p}}_h^*) \leq C|u - u_h|_1^2,$$

where we have again evoked Lemma 2.5 in addition to the well-known finite element energy error bound. ∎

3. Computing output bounds. We will continue to keep the presentation simple by considering only simple linear functional interior outputs. In particular, we will develop upper and lower bounds, $s^{\pm}$, on the output quantity

$$s \equiv \int_\Omega f^{\mathcal{O}} \, u \, \mathrm{d}\Omega, \tag{3.1}$$

where u is the exact solution of Poisson's equation (1.1) and $f^{\mathcal{O}}|_T$ is a member of $\mathbb{P}^r(T)$ for all elements T in $\mathcal{T}_h$. We stress, however, that more interesting outputs, such as boundary fluxes, can also be treated using techniques previously employed in the context of two-level methods (see, for example, the treatment of the normal force output for linear elasticity in [14]).

3.1. Weak continuity reformulation. To begin, we must formulate a generalized analogue to the minimization statement (2.3). There are two parts to this task. First, we must replace the intrinsic energy of the variational problem with an energy reformulation of the linear output functional. Second, now that the minimization of the objective functional no longer corresponds to the solution of our original equation, we must explicitly ensure that the minimizer is the solution to our problem by including it as a constraint. Furthermore, to obtain both upper and lower bounds, we consider two cases which vary by the sign of the original output. The resulting pair of constrained minimization statements for the homogeneous[2] Dirichlet boundary problem under consideration are

(3.2)

$$\begin{aligned} \mp s = \inf_{\hat{w}^{\pm} \in \hat{\mathcal{U}}} \; & \mp \int_\Omega f^{\mathcal{O}} \, \hat{w}^{\pm} \, \mathrm{d}\Omega + \frac{\kappa}{2} \left\{ \int_\Omega \nabla \hat{w}^{\pm} \cdot \nabla (\hat{w}^{\pm} - \bar{u}) \, \mathrm{d}\Omega - \int_\Omega f \, (\hat{w}^{\pm} - \bar{u}) \, \mathrm{d}\Omega \right\} \\ \text{s.t.} \quad & \int_\Omega \nabla \hat{w}^{\pm} \cdot \nabla \psi \, \mathrm{d}\Omega = \int_\Omega f \, \psi \, \mathrm{d}\Omega \quad \forall \psi \in \mathcal{U}, \\ & \sum_{T \in \mathcal{T}_h} \int_{\partial T} \sigma_T \, \lambda \, \hat{w}^{\pm} \, \mathrm{d}\Gamma = 0 \qquad\qquad \forall \lambda \in \Lambda, \end{aligned}$$

[2]The extension to nonhomogeneous Dirichlet boundaries requires choosing $\bar{u}$ from the set of admissible functions and weakly enforcing the Dirichlet boundary data, u_D, by replacing the continuity constraint with $\sum_{T \in \mathcal{T}_h} \int_{\partial T} \sigma_T \, \lambda \, \hat{w}^{\pm} \, \mathrm{d}\Gamma = \sum_{\gamma \in \partial \mathcal{T}_h} \int_\gamma \sigma_{T(\gamma)} \, \lambda \, u_D \, \mathrm{d}\Gamma \; \forall \lambda \in \Lambda$.

where $\bar{u}$ is any element of space $\mathcal{U}$, and κ is a positive real scaling parameter which serves both as a coefficient providing dimensional consistency in the engineering context and as an additional degree of freedom which we will use to tighten the bounds. The quadratic objective functional has been constructed so that all terms but the desired output functional vanish when $\hat{w}^{\pm}$ is the exact solution, u, while the constraints enforce equilibrium and interelement continuity.

Paraschivoiu, Peraire, and Patera [14] and Paraschivoiu and Patera [13] originally proposed this reformulation in the context of two-level output bounding methods which appeal to a second refined but localized finite element approximation, and therefore provided bounds only against a refined finite element approximation instead of the exact infinite-dimensional solution. With this constrained minimization reformulation, we can proceed more or less mechanically to apply the ideas from the energy bound to this more general context. The development of the output bound is very close to that for the energy bound, but with the extra burden of carrying an additional Lagrange multiplier for the equilibrium constraint and of managing the concurrent development of both upper and lower bounds on the output, as neither arise implicitly from the finite element discretization.

3.2. Localization by continuity relaxation. Considering the Lagrangian of problem (3.2),

$$
\begin{aligned}
(3.3)\quad & \mathcal{L}^{\pm}(\hat{w}^{\pm};\psi^{\pm},\lambda^{\pm}) \\
& \equiv \mp\int_{\Omega} f^{\mathcal{O}}\,\hat{w}^{\pm}\,\mathrm{d}\Omega + \frac{\kappa}{2}\left\{\int_{\Omega}\nabla\hat{w}^{\pm}\cdot\nabla(\hat{w}^{\pm}-\bar{u})\,\mathrm{d}\Omega - \int_{\Omega} f\,(\hat{w}^{\pm}-\bar{u})\,\mathrm{d}\Omega\right\} \\
& \quad + \int_{\Omega} f\,\psi^{\pm}\,\mathrm{d}\Omega - \int_{\Omega}\nabla\hat{w}^{\pm}\cdot\nabla\psi^{\pm}\,\mathrm{d}\Omega - \sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T\,\lambda^{\pm}\,\hat{w}^{\pm}\,\mathrm{d}\Gamma,
\end{aligned}
$$

we know, as we did for the energy bound, from the saddle point property of Lagrange multipliers and from the strong duality of convex minimizations that

$$
\inf_{\hat{w}^{\pm}\in\hat{\mathcal{U}}}\mathcal{L}^{\pm}(\hat{w}^{\pm};\tilde{\psi}^{\pm},\tilde{\lambda}^{\pm}) \le \sup_{\substack{\psi^{\pm}\in\mathcal{U}\\ \lambda\in\Lambda}}\inf_{\hat{w}^{\pm}\in\hat{\mathcal{U}}}\mathcal{L}^{\pm}(\hat{w}^{\pm};\psi^{\pm},\lambda^{\pm}) = \mp s
$$

$\forall(\tilde{\psi}^{\pm},\tilde{\lambda}^{\pm})\in\mathcal{U}\times\Lambda$. The lower bounding minimization for a given $\tilde{\lambda}^{\pm}$ and $\tilde{\psi}^{\pm}$ is separable and, for an appropriate choice for $\tilde{\lambda}^{\pm}$, provides nontrivial upper and lower bounds on the exact output s.

3.2.1. Lagrange multiplier approximation. We proceed, as we did for the energy bound, to obtain approximate Lagrange multipliers with a finite element discretization of the gradient condition of (3.3). Let $\psi_h^{\pm} = \pm\psi_h$, $\lambda_h^{\pm} = \frac{\kappa}{2}\lambda_h^u \pm \lambda_h^{\psi}$, and $\bar{u} = u_h$, all of which we find by solving the following discrete problems:

1. Find $u_h\in\mathcal{U}_h$ such that

$$
(3.4)\qquad \int_{\Omega}\nabla u_h\cdot\nabla v\,\mathrm{d}\Omega = \int_{\Omega} f\,v\,\mathrm{d}\Omega \quad \forall v\in\mathcal{U}_h.
$$

2. Find $\psi_h\in\mathcal{U}_h$ such that

$$
(3.5)\qquad \int_{\Omega}\nabla v\cdot\nabla\psi_h\,\mathrm{d}\Omega = -\int_{\Omega} f^{\mathcal{O}}\,v\,\mathrm{d}\Omega \quad \forall v\in\mathcal{U}_h.
$$

3. Find $\lambda_h^u \in \Lambda_h$ such that

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T\,\lambda_h^u\,\hat{v}\,\mathrm{d}\Gamma=\int_\Omega\nabla u_h\cdot\nabla\hat{v}\,\mathrm{d}\Omega-\int_\Omega f\,\hat{v}\,\mathrm{d}\Omega\quad\forall\hat{v}\in\hat{\mathcal{U}}_h. \tag{3.6}$$

4. Find $\lambda_h^\psi \in \Lambda_h$ such that

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T\,\lambda_h^\psi\,\hat{v}\,\mathrm{d}\Gamma=-\int_\Omega f^{\mathcal{O}}\,\hat{v}\,\mathrm{d}\Omega-\int_\Omega\nabla\hat{v}\cdot\nabla\psi_h\,\mathrm{d}\Omega\quad\forall\hat{v}\in\hat{\mathcal{U}}_h. \tag{3.7}$$

The first two problems comprise the well-known primal-adjoint pair which occur often in output oriented a posteriori error estimation techniques [3, 14, 13] as well as in computational approaches to design optimization [8], while the last two problems are their independent equilibrations. The first and third problems are identical to the global approximation problems required for the energy bound. These particular choices for the Lagrange multipliers ensure a finite lower bound in the saddle point property.

LEMMA 3.1. *If the Lagrange multipliers* $\psi_h^\pm=\pm\psi_h$ *and* $\lambda_h^\pm=\frac{\kappa}{2}\lambda_h^u\pm\lambda_h^\psi$ *satisfy the equilibration conditions* (3.6) *and* (3.7), *then the minimums* $\inf_{\hat{w}^\pm\in\hat{\mathcal{U}}}\mathcal{L}(\hat{w}^\pm;\psi_h^\pm,\lambda_h^\pm)$ *are bounded from below.*

Proof. This is true for essentially the same reason that it is true for Lemma 2.1. The only algebraic difference being that in the present output bounding case the property $\mathcal{L}^\pm(\hat{w}^{\pm'}+\hat{c};\psi_h^\pm,\lambda_h^\pm)=\mathcal{L}(\hat{w}^{\pm'};\psi_h^\pm,\lambda_h)$ results from the combined action of both equilibration conditions. ∎

3.3. Local dual subproblem. Restricting our attention to a single elemental subproblem, $T\in\mathcal{T}_h$, we first rewrite our local Lagrangian functional in a form suitable for applying the ideas developed for the energy bound. Every term other than the dissipative energy term, $\frac{\kappa}{2}\int_T\nabla w\cdot\nabla w\,\mathrm{d}\Omega$, must not involve derivatives of $\hat{w}^\pm$, which we can do in the present case by application of the Green's identity, $-\int_T\nabla u\cdot\nabla w\,\mathrm{d}\Omega=\int_T\Delta u\,w\,\mathrm{d}\Omega-\int_{\partial T}\nabla u\cdot\mathbf{n}\,w\,\mathrm{d}\Gamma$, to obtain the equivalent local Lagrangian functional

$$\begin{aligned}\mathcal{L}_T^\pm(w^\pm;\pm\tilde{\psi},\frac{\kappa}{2}\tilde{\lambda}^u\pm\tilde{\lambda}^\psi)&\equiv\frac{\kappa}{2}\int_T\nabla w^\pm\cdot\nabla w^\pm\,\mathrm{d}\Omega\\&-\frac{\kappa}{2}\left\{\int_T(f-\Delta\bar{u})\;w^\pm\,\mathrm{d}\Omega+\int_{\partial T}\left(\sigma_T\tilde{\lambda}^u+\nabla\bar{u}\cdot\mathbf{n}\right)\;w^\pm\,\mathrm{d}\Gamma+\int_T f\,\bar{u}\,\mathrm{d}\Omega\right\}\\&\mp\left\{\int_T\left(f^{\mathcal{O}}-\Delta\tilde{\psi}\right)\;w^\pm\,\mathrm{d}\Omega+\int_{\partial T}\left(\sigma_T\tilde{\lambda}^\psi+\nabla\tilde{\psi}\cdot\mathbf{n}\right)\;w^\pm\,\mathrm{d}\Gamma+\int_T f\,\tilde{\psi}\,\mathrm{d}\Omega\right\}.\end{aligned} \tag{3.8}$$

The functional we wish to minimize over $w^\pm$ can now be defined as

$$J_T^\pm(w^\pm)\equiv\frac{\kappa}{2}\int_T\nabla w^\pm\cdot\nabla w^\pm\,\mathrm{d}\Omega-\int_T f^\pm\,w^\pm\,\mathrm{d}\Omega-\int_{\partial T}g^\pm\,w^\pm\,\mathrm{d}\Gamma \tag{3.9}$$

for $f^\pm\equiv\frac{\kappa}{2}\{f-\Delta\bar{u}\}\pm\{f^{\mathcal{O}}-\Delta\tilde{\psi}\}$ and $g^\pm\equiv\frac{\kappa}{2}\{\sigma_T\tilde{\lambda}^u+\nabla\bar{u}\cdot\mathbf{n}\}\pm\{\sigma_T\tilde{\lambda}^\psi+\nabla\tilde{\psi}\cdot\mathbf{n}\}$. Thus, the local relaxed primal minimization once again corresponds to a Poisson problem of the type represented in (1.1) with Neumann boundary conditions posed on a single element.

As was the case for the energy bound, we do not require, and in general cannot compute, the exact minimum of this local infinite-dimensional primal subproblem,

but we can apply the same technique of dualizing this minimization problem in order to procure a computable lower bounding approximate to it.

PROPOSITION 3.2. *If we define the positive functional*

$$J_T^c(\mathbf{q}) \equiv \frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\,\mathrm{d}\Omega, \tag{3.10}$$

where $\mathbf{q}\in\mathcal{H}(\mathrm{div};T)$*, then we have*

$$J_T^{\pm}(w^{\pm}) \geq -\frac{1}{\kappa}J_T^c(\mathbf{q}^{\pm}) \quad \forall w^{\pm}\in\mathcal{H}^1(T),\ \forall \mathbf{q}^{\pm}\in\mathcal{Q}^{\pm}(T), \tag{3.11}$$

for the set of functions

$$\mathcal{Q}^{\pm}(T) \equiv \left\{ \mathbf{q}\in\mathcal{H}(\mathrm{div};T) \;\middle|\; \int_T \nabla\cdot\mathbf{q}\,v\,\mathrm{d}\Omega - \int_{\partial T}\mathbf{q}\cdot\mathbf{n}\,v\,\mathrm{d}\Gamma = -\int_T f^{\pm}\,v\,\mathrm{d}\Omega - \int_{\partial T} g^{\pm}\,v\,\mathrm{d}\Gamma,\ \forall v\in\mathcal{H}^1(T)\right\}. \tag{3.12}$$

Proof. The local dual problem is derived as it was for the energy bound, but with modified data and the addition of the scaling parameter, κ. After expanding the positive expression for $\mathbf{q}\in\mathcal{Q}^{\pm}(T)$,

$$\frac{1}{2\kappa}\int_T (\mathbf{q}^{\pm}-\kappa\nabla w)^2\,\mathrm{d}\Omega \geq 0, \tag{3.13}$$

applying Green's formula, and substituting the constraint from $\mathcal{Q}^{\pm}(T)$, we obtain the expression

$$\frac{1}{2\kappa}\int_T \mathbf{q}^{\pm}\cdot\mathbf{q}^{\pm}\,\mathrm{d}\Omega + \frac{\kappa}{2}\int_T \nabla w^{\pm}\cdot\nabla w^{\pm}\,\mathrm{d}\Omega - \int_T f^{\pm}\,w^{\pm}\,\mathrm{d}\Omega - \int_{\partial T} g^{\pm}\,w^{\pm}\,\mathrm{d}\Gamma \geq 0. \tag{3.14}$$

Identifying $J_T^{\pm}(w^{\pm})$ and $J_T^c(\mathbf{q}^{\pm})$ we arrive at the desired expression for the local lower bound. ∎

As the functional $J_T^{\pm}(w^{\pm})$ contains only the terms from the Lagrangian that depended on $w^{\pm}$, we must reintroduce the constant terms to secure the complete contributions from the local dual subproblems:

$$\mp s_T^{\pm} = \int_T f\left(\frac{\kappa}{2}u_h \pm \psi_h\right)\mathrm{d}\Omega + \sup_{\mathbf{q}^{\pm}\in\mathcal{Q}^{\pm}(T)} -\frac{1}{\kappa}J_T^c(\mathbf{q}^{\pm}). \tag{3.15}$$

3.3.1. Subproblem approximation. Consider the splitting implied by the definition $\mathbf{q}_h = \kappa\nabla\bar{u} + \frac{\kappa}{2}\mathbf{q}_h^u \pm \mathbf{q}_h^{\psi}$. Propagation of this definition into the elemental subproblem reveals through the linearity of the gradient condition that indeed $\mathbf{q}_h^u$ and $\mathbf{q}_h^{\psi}$ can be computed independently. The resulting subproblems are

$$\begin{aligned} \mathbf{q}_h^u &= \arg\inf_{\mathbf{q}_h\in\mathcal{Q}_h^u(T)} J^c(\mathbf{q}_h), \\ \mathbf{q}_h^{\psi} &= \arg\inf_{\mathbf{q}_h\in\mathcal{Q}_h^{\psi}(T)} J^c(\mathbf{q}_h), \end{aligned} \tag{3.16}$$

for the dual feasible approximation sets

(3.17)
$$\mathcal{Q}_h^u(T) \equiv \left\{ \mathbf{q} \in (\mathbb{P}^q(T))^d \;\middle|\; \int_T \nabla \cdot \mathbf{q}\, v \,\mathrm{d}\Omega - \int_{\partial T} \mathbf{q}\cdot\mathbf{n}\, v \,\mathrm{d}\Gamma = -\int_T (f + \Delta u_h)\, v \,\mathrm{d}\Omega \right.$$
$$\left. - \int_{\partial T} (\sigma_T \lambda_h^u - \nabla u_h \cdot \mathbf{n})\, v \,\mathrm{d}\Gamma \;\forall v \in \mathcal{H}^1(T) \right\},$$
$$\mathcal{Q}_h^\psi(T) \equiv \left\{ \mathbf{q} \in (\mathbb{P}^q(T))^d \;\middle|\; \int_T \nabla \cdot \mathbf{q}\, v \,\mathrm{d}\Omega - \int_{\partial T} \mathbf{q}\cdot\mathbf{n}\, v \,\mathrm{d}\Gamma = -\int_T (f^{\mathcal{O}} - \Delta \psi_h)\, v \,\mathrm{d}\Omega \right.$$
$$\left. - \int_{\partial T} (\sigma_T \lambda_h^\psi + \nabla \psi_h \cdot \mathbf{n})\, v \,\mathrm{d}\Gamma \;\forall v \in \mathcal{H}^1(T) \right\},$$

in which we have again chosen $\bar{u} = u_h$ commensurate with our choice for the approximate multipliers. As the additional terms in the data of the dual feasibility constraint are just polynomial functions in the local finite element basis, there are no difficulties in choosing our dual approximation sets in this manner. The solvability of these subproblems is addressed by the following result.

LEMMA 3.3. *Suppose the forcing function $f|_T$ and output function $f^{\mathcal{O}}|_T$ are members of $\mathbb{P}^r(T)$, that λ_h^u satisfies* (3.6)*, and that λ_h^ψ satisfies* (3.7)*, then there exists at least one dual feasible function $\mathbf{q}_h^u$ that is a member of $\mathcal{Q}_h^u(T)$ and one dual feasible function $\mathbf{q}_h^\psi$ that is a member of $\mathcal{Q}_h^\psi(T)$ for $q \geq p$ and $q > r$.*

Proof. Applying Green's formula to the u_h Laplacian term in the constraint data for $\mathcal{Q}_h^u(T)$ of (3.17) and duplicating the proof of Lemma 2.3 with the resulting constraint data reveals the compatibility condition

$$\int_{\partial T} \sigma_T \lambda_h^u \,\mathrm{d}\Gamma = -\int_T f \,\mathrm{d}\Omega, \tag{3.18}$$

which is satisfied by our choice for λ_h^u as can be seen by choosing $\hat{v} = \text{const}$ on T in the equilibration condition (3.6). The same argument holds for the adjoint dual subproblem, yielding the analogous compatibility condition

$$\int_{\partial T} \sigma_T \lambda_h^\psi \,\mathrm{d}\Gamma = -\int_T f^{\mathcal{O}} \,\mathrm{d}\Omega \tag{3.19}$$

for $f^{\mathcal{O}}$ and λ_h^ψ. ∎

With the subproblem splitting just defined, the aggregated contributions to the upper and lower bounds become

$$\begin{aligned} s_h^{\pm} &= \mp \int_\Omega f \left(\frac{\kappa}{2} u_h \pm \psi_h\right) \mathrm{d}\Omega \pm \frac{1}{\kappa} J^c(\kappa \nabla u_h + \frac{\kappa}{2}\hat{\mathbf{q}}_h^u \pm \hat{\mathbf{q}}_h^\psi) \\ &= \mp \int_\Omega f \left(\frac{\kappa}{2} u_h \pm \psi_h\right) \mathrm{d}\Omega \pm \frac{\kappa}{2}\int_\Omega \nabla u_h \cdot \nabla u_h \,\mathrm{d}\Omega + \int_\Omega (\frac{\kappa}{2}\hat{\mathbf{q}}_h^u \pm \hat{\mathbf{q}}_h^\psi)\cdot \nabla u_h \,\mathrm{d}\Omega \\ &\quad + \frac{1}{2}\int_\Omega \hat{\mathbf{q}}_h^u \cdot \hat{\mathbf{q}}_h^\psi \,\mathrm{d}\Omega \pm \frac{\kappa}{4} J^c(\hat{\mathbf{q}}_h^u) \pm \frac{1}{\kappa} J^c(\hat{\mathbf{q}}_h^\psi) \\ &= -\int_\Omega f\, \psi_h \,\mathrm{d}\Omega + \frac{1}{2}\int_\Omega \hat{\mathbf{q}}_h^u \cdot \hat{\mathbf{q}}_h^\psi \,\mathrm{d}\Omega \pm \frac{\kappa}{4} J^c(\hat{\mathbf{q}}_h^u) \pm \frac{1}{\kappa} J^c(\hat{\mathbf{q}}_h^\psi), \end{aligned}$$

in which we have invoked (3.4) with $v = u_h$ and we have used orthogonality relationships analogous to that proved in Lemma 2.4 as well.

3.4. Output bound procedure. The introduction of the scaling parameter κ allows us to optimize the sharpness of the computed bounds in addition to providing dimensional consistency. From the previous section we have the expression for the upper and lower output bounds

$$s_h^{\pm} = \bar{s}_h \pm \kappa z_h^u \pm \frac{1}{\kappa} z_h^{\psi},$$

where

$$\bar{s}_h = \frac{1}{2}\int_\Omega \hat{\mathbf{q}}_h^u \cdot \hat{\mathbf{q}}_h^{\psi}\,\mathrm{d}\Omega - \int_\Omega f\,\psi_h\,\mathrm{d}\Omega, \qquad z_h^u = \frac{1}{4} J^c(\hat{\mathbf{q}}_h^u), \qquad z_h^{\psi} = J^c(\hat{\mathbf{q}}_h^{\psi}). \tag{3.20}$$

Maximizing the lower bound and minimizing the upper bound with respect to κ yield the optimal value $\kappa^2 = z_h^{\psi}/z_h^u$.

The complete method with optimal scaling for upper and lower bounds on linear functional outputs can now be written as three steps.

1. *Global approximation:* Find $u_h \in \mathcal{U}_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v\,\mathrm{d}\Omega = \int_\Omega f\,v\,\mathrm{d}\Omega \quad \forall v \in \mathcal{U}_h, \tag{3.21}$$

and find $\psi_h \in \mathcal{U}_h$ such that

$$\int_\Omega \nabla v \cdot \nabla \psi_h\,\mathrm{d}\Omega = -\int_\Omega f^{\mathcal{O}}\,v\,\mathrm{d}\Omega \quad \forall v \in \mathcal{U}_h. \tag{3.22}$$

2. *Global equilibration:* Find $\lambda_h^u \in \Lambda_h$ such that

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T} \sigma_T\,\lambda_h^u\,\hat{v}\,\mathrm{d}\Gamma = \int_\Omega \nabla u_h \cdot \nabla \hat{v}\,\mathrm{d}\Omega - \int_\Omega f\,\hat{v}\,\mathrm{d}\Omega \quad \forall \hat{v} \in \hat{\mathcal{U}}_h, \tag{3.23}$$

and find $\lambda_h^{\psi} \in \Lambda_h$ such that

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T} \sigma_T\,\lambda_h^{\psi}\,\hat{v}\,\mathrm{d}\Gamma = -\int_\Omega f\,\hat{v}\,\mathrm{d}\Omega - \int_\Omega \nabla \hat{v} \cdot \nabla \psi_h\,\mathrm{d}\Omega \quad \forall \hat{v} \in \hat{\mathcal{U}}_h. \tag{3.24}$$

3. *Local dual subproblems:* Find $\hat{\mathbf{q}}_h^u$ such that

$$\hat{\mathbf{q}}_h^u = \arg \inf_{\hat{\mathbf{q}}_h \in \hat{\mathcal{Q}}_h^u} J^c(\hat{\mathbf{q}}_h), \tag{3.25}$$

find $\hat{\mathbf{q}}_h^{\psi}$ such that

$$\hat{\mathbf{q}}_h^{\psi} = \arg \inf_{\hat{\mathbf{q}}_h \in \hat{\mathcal{Q}}_h^{\psi}} J^c(\hat{\mathbf{q}}_h), \tag{3.26}$$

and, from (3.20) and the optimal κ, calculate

$$s_h^{\pm} = \bar{s}_h \pm 2\sqrt{z_h^u z_h^{\psi}}. \tag{3.27}$$

The local dual subproblems for the output bounds can be solved in the same manner as the local energy dual subproblems. The important point being that once the finite element approximations u_h and ψ_h have been computed, the solutions can be equilibrated and quantitative bounds computed on the exact output to the infinite-dimensional continuum equation with asymptotically linear cost in the size of the finite element discretization and in parallel. In addition, the elemental contribution to the bound gap, $\frac{\kappa}{4} J_T^c(\mathbf{q}_h^u) + \frac{1}{\kappa} J_T^c(\mathbf{q}_h^{\psi})$, can serve as an informative mesh adaptivity indicator for controlling the error in the *output*, as was done in [15] for a two-level error bound method and in [3] for an asymptotic error estimation method.

3.4.1. Properties of the output bounds. The upper and lower bounding properties are direct consequences of the saddle point property of the relaxed constrained minimization reformulation (3.2) and the local dual property of Proposition 3.2. The following proposition addresses the accuracy of the computed bounds by showing that the bounds will converge at the optimal rate when both the primal and adjoint finite element approximations are in the asymptotic convergence regime.

PROPOSITION 3.4. *Suppose that u_h, ψ_h, λ_h^u, and λ_h^ψ are solutions of the above finite element approximation problems and equilibration problems, then*

$$\begin{aligned} s - s_h^- &\le C|u - u_h|_1|\psi - \psi_h|_1, \\ s_h^+ - s &\le C|u - u_h|_1|\psi - \psi_h|_1. \end{aligned} \tag{3.28}$$

Proof. Applying the definitions from the procedure, we know that the lower a posteriori bound, for instance, itself has the bound $s - s_h^- \le s_h^+ - s_h^- = 2\sqrt{z_h^u z_h^\psi}$. The arguments of Lemma 2.5 can be applied to the z_h^u and z_h^ψ factors to show that they are bounded by $C|u - u_h|_1^2$ and $C|\psi - \psi_h|_1^2$, respectively. ∎

4. Numerical results. We verify the method numerically for three cases: constant forcing on the unit square, linear forcing on the unit square, and zero forcing on an L-shaped domain with a corner singularity. Linear finite elements, $p = 1$, and quadratic subproblems, $q = 2$, are employed with the domain average output $s = \int_\Omega f^{\mathcal{O}} u \, \mathrm{d}\Omega$, where $f^{\mathcal{O}} = \text{const}$, for all cases.

All three cases have analytically exact solutions with which we are able to verify the method and calculate the effectivities of the bounds,

$$\theta^\pm = \frac{|s - s_h^\pm|}{|s - s_h|}, \tag{4.1}$$

which indicate the sharpness by comparing the error in the bounds to the error in the finite element approximation.

4.1. Uniformly forced square domain. The first case is a uniformly forced unit square domain with $f = f^{\mathcal{O}} = \sqrt{10}$. The analytical solution is given by

$$u(x,y) = \frac{16\sqrt{10}}{\pi^4} \sum_{\text{odd } i,j=1}^{\infty} \frac{(-1)^{(i+j)/2-1}}{ij(i^2+j^2)} \cos\left(i\frac{\pi}{2}x\right)\cos\left(j\frac{\pi}{2}y\right).$$

This case is special in that the forcing and output are identical and the boundary data is homogeneous, leading to primal and adjoint problem data which differ by only a sign. It is well known that for this special case, called compliance, the finite element approximation for the output is a lower bound. The numerical results given in Table 4.1 and Figure 4.1 demonstrate that our method, while more expensive, does no worse than the inherent bound for this special case. The results for both the finite element approximation and the output bounds asymptotically approach the optimal finite element convergence rate of $O(h^2)$. This example also evinces that the bound average, $\bar{s}_h$, can sometimes be a more accurate output approximation than that from the finite element approximation.

4.2. Linearly forced square domain. The second case is a linearly forced square domain with $f^{\mathcal{O}} = 1$, and the forcing and nonhomogeneous boundary conditions chosen to produce the exact solution

$$u(x,y) = \frac{3}{2}y^2(1-y) + 4xy.$$

TABLE 4.1
Numerical results for the uniformly forced square domain test case for which $s = 0.351$.

h	s_h	s_-	$\bar{s}_h$	s^+	θ^-	θ^+
$\frac{1}{2}$	0.156	0.156	0.394	0.632	1.0	1.4
$\frac{1}{4}$	0.288	0.288	0.367	0.446	1.0	1.5
$\frac{1}{8}$	0.334	0.334	0.356	0.377	1.0	1.5
$\frac{1}{16}$	0.347	0.347	0.353	0.358	1.0	1.5

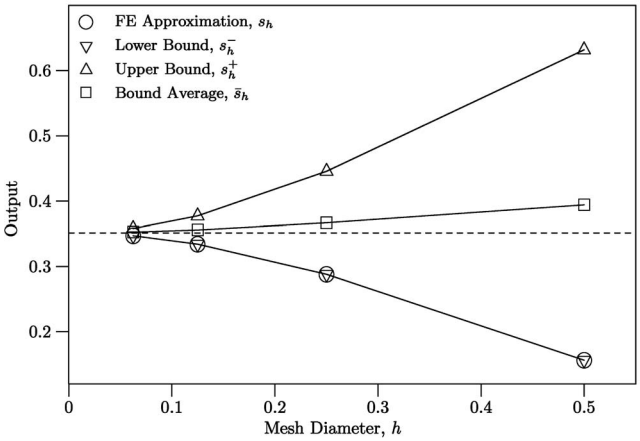


FIG. 4.1. *Uniformly forced square domain.*

As this test case is not a special case, the convergence histories of Table 4.2 and Figure 4.2 depict the more general situation in which none of the computed quantities coincide. Whereas in the first example we saw that the bound average can possibly be a more accurate output approximation than the finite element approximation, in this example we see that this is definitely not always true since the finite element approximation for the output is 0.5% better. As for the first example, the results for both the finite element approximation and the output bounds asymptotically approach the optimal finite element convergence rate of $O(h^2)$.

4.3. Unforced corner domain. Last, we consider the Laplace equation on a nonconvex domain with $f^{\mathcal{O}} = 1$. The domain is the standard L-shaped domain with a reentrant corner that results from removing the lower-right quadrant of the unit square. The Dirichlet boundary conditions were chosen to produce the solution

$$u(r, \phi) = r^{\frac{2}{3}} \sin \frac{2}{3}\phi,$$

where the distance from the corner point is $r(x, y) = \{(x - 1/2)^2 + (y - 1/2)^2\}^{\frac{1}{2}}$ and the angle from the upper surface of the corner is $\phi(x, y) = \arctan(\frac{y-1/2}{x-1/2})$.

In this example we demonstrate that the bounds are valid even for problems with singularities. As can be seen from Table 4.3 and Figure 4.3, the results for both the finite element approximation and the output bounds asymptotically approach the optimal finite element convergence rate of $O(h^{\frac{4}{3}})$ for elliptic problems posed on a domain with right-angled reentrant corner [19]. Once again we see that the bound average has the potential to be a better output approximation than the finite element method.

TABLE 4.2
Numerical results for the linearly forced square domain test case for which $s = 1.125$.

h	s_h	s_-	$\bar{s}_h$	s^+	θ^-	θ^+
$\frac{1}{2}$	1.177	0.860	1.068	1.276	5.1	2.9
$\frac{1}{4}$	1.138	1.050	1.111	1.171	5.7	3.5
$\frac{1}{8}$	1.128	1.106	1.212	1.137	5.9	3.8
$\frac{1}{16}$	1.126	1.120	1.124	1.128	6.0	3.8

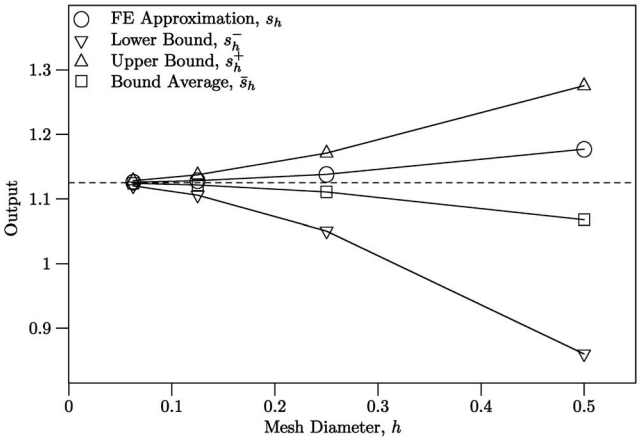


FIG. 4.2. *Linearly forced square domain.*

TABLE 4.3
Numerical results for the unforced corner domain test case for which $s = 0.792$.

h	s_h	s_-	$\bar{s}_h$	s^+	θ^-	θ^+
$\frac{1}{2}$	0.775	0.702	0.799	0.897	5.1	6.0
$\frac{1}{4}$	0.785	0.761	0.795	0.829	4.3	5.1
$\frac{1}{8}$	0.789	0.781	0.793	0.805	3.6	4.4
$\frac{1}{16}$	0.791	0.788	0.792	0.797	3.1	3.9

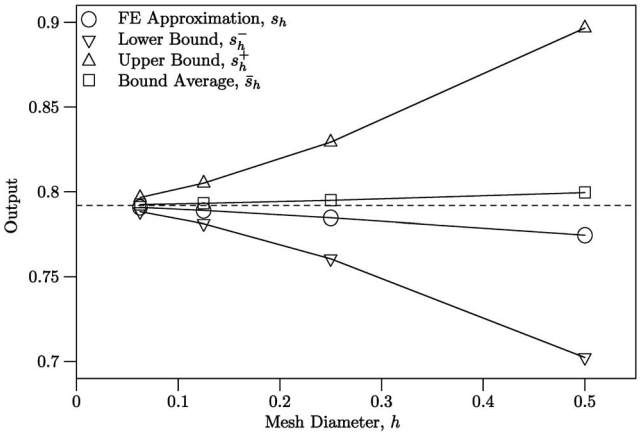
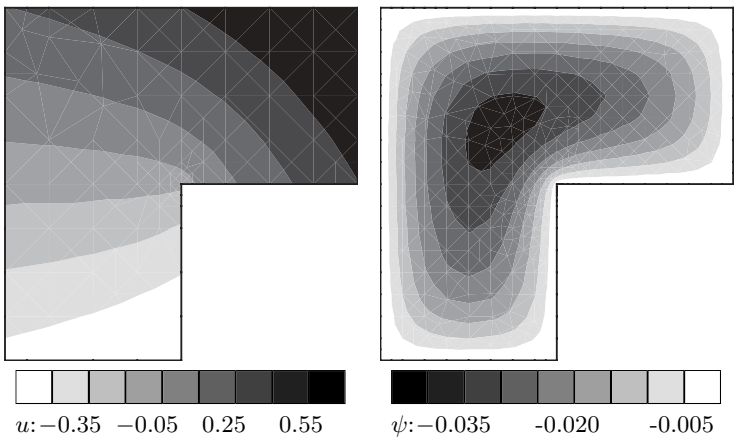
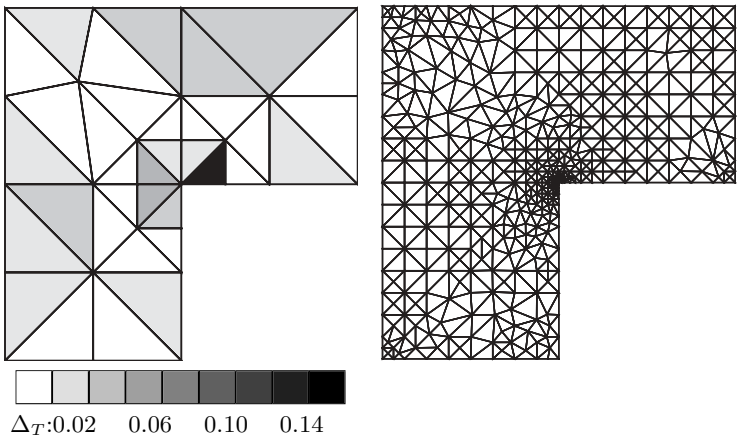


FIG. 4.3. *Unforced corner domain.*



(a) Adaptive solution field.

(b) Adaptive adjoint field.



(c) Intermediate local error indicators.

(d) Final adaptive mesh.

FIG. 4.4. *Unforced corner domain adaptive solutions, local indicators and meshes.*

We close by demonstrating the use of the previously introduced elemental contributions to the bound gap, $\Delta_T \equiv \frac{\kappa}{4} J_T^c(\mathbf{q}_h^u) + \frac{1}{\kappa} J_T^c(\mathbf{q}_h^\psi)$, as a mesh adaptivity indicator for controlling the error in the output. Figure 4.4 displays the solution field, adjoint field, intermediate local error indicators, and final adaptive mesh that results from preferentially refining elements with relatively high contributions to the bound gap. We initiated the adaptive process on a uniform mesh of 6 elements and adaptively refined until the output uncertainty was less than 0.005, that is, until we had a certificate of precision at least as good as $s = s_h \pm 0.005$. The adaptive refinement process met this target by producing the certificate of precision $s = 0.791 \pm 0.00456$ using 1167 elements. Achieving the same certainty with uniform refinement requires 6144 elements.

Acknowledgments. We would like to acknowledge our long-standing and fecund collaboration with Professor A. T. Patera at MIT. We would also like to thank Nuria Pares-Marine for her insightful comments on the convergence analysis provided while revising the manuscript.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A unified approach to a posteriori error estimation using element residual methods*, Numer. Math., (1993), pp. 23–50.
[2] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Comput. Methods Appl. Mech. Engrg., 142 (1997), pp. 1–88.
[3] R. Becker and R. Rannacher, *A feedback approach to error control in finite element method: Basic analysis and examples*, East-West J. Numer. Math., 4 (1996), pp. 237–264.
[4] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.
[5] T. Cao, D. W. Kelly, and I. H. Sloan, *Local error bounds for post-processed finite element calculations*, Internat. J. Numer. Methods Engrg., 45 (1999), pp. 1085–1098.
[6] P. Destuynder and B. Métivet, *Explicit error bounds in a conforming finite element method*, Math. Comput., 68 (1999), pp. 1379–1396.
[7] B. Fraeijs de Veubeke, *Displacement and equilibrium models in the finite element method*, in B. M. Fraeijs de Veubeke Memorial Volume of Selected Papers, M. Geradin, ed., Sijthoff and Noordhoff International, The Netherlands, 1980.
[8] A. Jameson, *Aerodynamic design via control theory*, J. Sci. Comput., 3 (1988), pp. 233–260.
[9] P. Ladeveze and D. Leguillon, *Error estimate procedure in the finite element method and applications*, SIAM J. Numer. Anal., 20 (1983), pp. 485–509.
[10] P. Ladevèze and P. Rougeot, *New advances on a posteriori error on constitutive relation in f.e. analysis*, Comput. Methods Appl. Mech. Engrg., 150 (1997), pp. 239–249.
[11] Y. Maday and A. T. Patera, *Numerical analysis of a posteriori finite element bounds for linear functional outputs*, Math. Models Methods Appl. Sci., 10 (2000), pp. 785–799.
[12] P. Morin, R. H. Nochetto, and K. G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.
[13] M. Paraschivoiu and A. T. Patera, *A hierarchical duality approach to bounds for the outputs of partial differential equations*, Comp. Methods Appl. Mech. Engrg., 158 (1998), pp. 389–407.
[14] M. Paraschivoiu, J. Peraire, and A. T. Patera, *A posteriori finite element bounds for linear-functional outputs of elliptic partial differential equations*, Comp. Methods Appl. Mech. Engrg., 150 (1997), pp. 289–312.
[15] J. Peraire and A. Patera, *Bounds for linear-functional outputs of coercive partial differential equations: Local indicators and adaptive refinement*, in Proceedings of the Workshop on New Advances in Adaptive Computational Methods in Mechanics, P. Ladeveze and J. Oden, eds., Elsevier, Amsterdam, 1998.
[16] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, New York, 1997.
[17] J. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, Finite Element Methods (Part 1), P. Ciarlet and J. Lions, eds., North–Holland, Amsterdam, 1991, pp. 523–639.
[18] A. M. Sauer-Budge and J. Peraire, *Computing bounds for linear functional outputs of exact weak solutions to the advection-diffusion-reaction equation*, SIAM J. Sci. Comput., 26 (2004), pp. 636–652.
[19] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1631–1647

A COMPARISON OF DEFLATION AND COARSE GRID CORRECTION APPLIED TO POROUS MEDIA FLOW*

R. NABBEN† AND C. VUIK‡

Abstract. In this paper we compare various preconditioners for the numerical solution of partial differential equations. We compare a coarse grid correction preconditioner used in domain decomposition methods with a so-called deflation preconditioner. We prove that the effective condition number of the deflated preconditioned system is always, for all deflation vectors and all restrictions and prolongations, below the condition number of the system preconditioned by the coarse grid correction. This implies that the conjugate gradient method applied to the deflated preconditioned system is expected always to converge faster than the conjugate gradient method applied to the system preconditioned by the coarse grid correction. Numerical results for porous media flows emphasize the theoretical results.

Key words. deflation, coarse grid correction, preconditioners, conjugate gradients, porous media flow, scalable parallel preconditioner

AMS subject classifications. 65F10, 65F50, 65N22

DOI. 10.1137/S0036142903430451

1. Introduction.

It is well known that the convergence rate of the conjugate gradient (CG) method is bounded as a function of the condition number of the system matrix to which it is applied. Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite. We assume that the vector $b \in \mathbb{R}^n$ represents a discrete function on a grid Ω and that we are searching for the vector $x \in \mathbb{R}^n$ on Ω which solves the linear system

$$Ax = b.$$

Such systems are encountered, for example, when a finite volume/difference/element method is used to discretize an elliptic PDE defined on the continuous analogue of Ω.

Let us denote the ith eigenvalue in nondecreasing order by $\lambda_i(A)$ or simply by λ_i when it is clear to which matrix we are referring. After k iterations of the CG method, the error is bounded by (cf. [9, Thm. 10.2.6])

$$\|x - x_k\|_A \leq 2\,\|x - x_0\|_A \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \tag{1.1}$$

where $\kappa = \kappa(A) = \lambda_n/\lambda_1$ is the spectral condition number of A and the A-norm of x is given by $\|x\|_A = (x^T Ax)^{1/2}$. The convergence may be significantly faster if the eigenvalues of A are clustered (see [24]).

If the condition number of A is large it is advisable to solve, instead, a preconditioned system $M^{-1}Ax = M^{-1}b$, where the symmetric positive definite preconditioner M is chosen such that $M^{-1}A$ has a more clustered spectrum or a smaller condition number than that of A. Furthermore, M must be cheap to solve relative to the

*Received by the editors June 26, 2003; accepted for publication (in revised form) January 28, 2004; published electronically December 27, 2004.
http://www.siam.org/journals/sinum/42-4/43045.html
†Institut für Mathematik, Technische Universität Berlin, MA 3-3, Strasse des 17. Juni 136, 10623 Berlin, Germany (nabben@math.tu-berlin.de).
‡Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Applied Mathematical Analysis, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands (c.vuik@math.tudelft.nl).

improvement it provides in convergence rate. A final desirable property in a preconditioner is that it should parallelize well, especially on distributed memory computers. Probably one of the most effective preconditioning strategies in common use is to take $M = LL^T$ to be an incomplete Cholesky (IC) factorization of A (see [16]). We denote the preconditioned conjugate gradient method as PCG.

With respect to the known preconditioners, at least two problems remain:

- If there are large jumps in the coefficients of the discretized PDE, the convergence of PCG becomes very slow, and
- if a block preconditioner is used in a domain decomposition algorithm the condition number of the preconditioned matrix deteriorates if the number of blocks increases.

Both problems can be solved by a deflation technique or a suitable coarse grid correction. In this section we describe both methods, which are compared in the next sections. To describe the deflation method we define the projection P_D by

$$P_D = I - AZ(Z^T AZ)^{-1}Z^T, \quad Z \in \mathbb{R}^{n\times r}, \tag{1.2}$$

where the column space of Z is the deflation subspace, i.e., the space to be projected out of the residual, and I is the identity matrix of appropriate size. We assume that $r \ll n$ and that Z has rank r. Under this assumption $E \equiv Z^T AZ$ may be easily computed and factored and is symmetric positive definite. Since $x = (I - P_D^T)x + P_D^T x$ and because

$$(I - P_D^T)x = Z(Z^T AZ)^{-1}Z^T Ax = ZE^{-1}Z^T b \tag{1.3}$$

can be immediately computed, we only need to compute $P_D^T x$. In light of the identity $AP_D^T = P_D A$, we can solve the deflated system

$$P_D A\tilde{x} = P_D b \tag{1.4}$$

for $\tilde{x}$ using the CG method, premultiply this by P_D^T, and add it to (1.3).

Obviously (1.4) is singular. What consequences does the singularity of (1.4) imply for the CG method? Kaasschieter [12] notes that a positive semidefinite system can be solved as long as the right-hand side is consistent (i.e., as long as $b = Ax$ for some x). This is certainly true for (1.4), where the same projection is applied to both sides of the nonsingular system. Furthermore, he notes (with reference to [24]) that because the null space never enters the iteration, the corresponding zero eigenvalues do not influence the convergence. Motivated by this fact, we define the *effective condition number* of a positive semidefinite matrix $C \in \mathbb{R}^{n\times n}$ with r zero eigenvalues to be the ratio of its largest to smallest *positive* eigenvalues:

$$\kappa_{\text{eff}}(C) = \frac{\lambda_n}{\lambda_{r+1}}.$$

It is possible to combine both a standard preconditioning and preconditioning by deflation (for details, see [8]). The convergence is then described by the effective condition number of $M^{-1}P_D A$.

The deflation technique has been exploited by several authors. For nonsymmetric systems, approximate eigenvectors can be extracted from the Krylov subspace produced by GMRES. Morgan [17] uses this approach to improve the convergence after a restart. In this case, deflation is not applied as a preconditioner, but the deflation

vectors are augmented with the Krylov subspace and the minimization property of GMRES ensures that the deflation subspace is projected out of the residual (for related references, we refer the reader to [8] and [7]). A comparable approach for the CG method is described in [22]. Mansfield [14] shows how Schur complement-type domain decomposition methods can be seen as a series of deflations. Nicolaides [19] chooses Z to be a piecewise constant interpolation from a set of r subdomains and points out that deflation might be effectively used with a conventional preconditioner. Mansfield [15] uses the same "subdomain deflation" in combination with damped Jacobi smoothing, obtaining a preconditioner which is related to the two-grid method. In [13] Kolotilina uses a twofold deflation technique for simultaneously deflating the r largest and the r smallest eigenvalues using an appropriate deflating subspace of dimension r. Other authors have attempted to choose a subspace a priori that effectively represents the slowest modes. In [27] deflation is used to remove a few stubborn but *known* modes from the spectrum. This method is used in [3] to solve electromagnetic problems with large jumps in the coefficients. Thereafter this method has been generalized to other choices of the deflation vectors (see [26, 28]). Finally, an analysis of the effective condition number and a parallel implementation is given in [8, 25].

We compare the deflation preconditioner with a well-known coarse grid correction preconditioner of the form

$$P_C = I + ZE^{-1}Z^T \tag{1.5}$$

and in the preconditioned case

$$P_{CM^{-1}} = M^{-1} + ZE^{-1}Z^T. \tag{1.6}$$

In the multigrid or domain decomposition language the matrices Z and Z^T are known as restriction and prolongation or interpolation operator. Moreover, the matrix $E = Z^T AZ$ is the Galerkin operator.

The above coarse grid correction preconditioner belongs to the class of additive Schwarz preconditioner. It is called the two-level additive Schwarz preconditioner. If used in domain decomposition methods, typically M^{-1} is the sum of the local (exact or inexact) solves in each domain. To speed up convergence a coarse grid correction $ZE^{-1}Z^T$ is added.

These methods are introduced by Bramble, Pasciak, and Schatz [2], Dryja and Widlund [5, 6], and Dryja [4]. They show under mild conditions that the convergence rate of the PCG method is independent of the grid sizes.

For more details about this preconditioner we refer the reader to the books of Quarteroni and Valli [21] and Smith, Bjørstad, and Gropp [23]. A more abstract analysis of this preconditioner is given by Padiy, Axelsson, and Polman [20]. To make the condition number of $P_{CM^{-1}}A$ smaller, Padiy, Axelsson, and Polman used a parameter $\sigma > 0$ and considered

$$P_C = I + \sigma ZE^{-1}Z^T \tag{1.7}$$

and

$$P_{CM^{-1}} = M^{-1} + \sigma ZE^{-1}Z^T. \tag{1.8}$$

If $M = I$, Z consists of eigenvectors, and λ_{max} is known, then a good choice is $\sigma = \lambda_{max}$, which implies that $\kappa(P_C A) \leq \frac{2\lambda_{max}}{\lambda_{r+1}}$ (see [20]). If $M \neq I$ and/or Z consists of general vectors, and λ_{max} is not known, it is not clear how to choose σ.

More abstract results about Schwarz methods applied to nonsymmetric problems are given by Benzi et al. [1] and Nabben [18].

In this paper we prove that the effective condition number of the deflated preconditioned system $M^{-1}P_DA$ is always below the condition number of the system preconditioned by the coarse grid correction $P_{CM^{-1}}A$. This implies that for all matrices $Z \in \mathbb{R}^{n\times r}$ and all positive definite preconditioners M^{-1} the CG method applied to the deflated preconditioned system is expected always to converge faster than the CG method applied to the system preconditioned by the coarse grid correction. These results are stated in section 2. In section 3 we compare other properties of the deflation and coarse grid preconditioner. These properties are scaling, approximation of E^{-1}, and an estimate of the smallest eigenvalue. Section 4 contains our numerical results for porous media flows and parallel problems.

2. Spectral properties. In this section we compare the effective condition number for the deflation and coarse grid correction preconditioned matrices. In section 2.1 we give some definitions and preliminary results. Thereafter a comparison is made if the projection vectors are equal to eigenvectors in section 2.2 and for general projection vectors in section 2.3.

2.1. Notations and preliminary results. In the following we denote by $\lambda_i(M)$ the eigenvalues of a matrix M. If the eigenvalues are real, the $\lambda_i(M)$'s are ordered increasingly.

For two Hermitian $n \times n$ matrices A and B we write $A \succeq B$, if $A - B$ is positive semidefinite.

Next we mention well-known properties of the eigenvalues of Hermitian matrices.

LEMMA 2.1. *Let $A, B \in \mathbb{C}^{n\times n}$ be Hermitian. For each $k = 1, 2, \ldots, n$ we have*

$$\lambda_k(A) + \lambda_1(B) \leq \lambda_k(A + B) \leq \lambda_k(A) + \lambda_n(B).$$

From the above lemma we easily obtain the next lemma.

LEMMA 2.2. *If $A, B \in \mathbb{C}^{n\times n}$ are positive semidefinite with $A \succeq B$, then $\lambda_i(A) \geq \lambda_i(B)$.*

Moreover, we will use the following lemma.

LEMMA 2.3. *Let $A, B \in \mathbb{C}^{n\times n}$ be Hermitian, and suppose that B has rank at most r. Then*

- $\lambda_k(A + B) \leq \lambda_{k+r}(A), \quad k = 1, 2, \cdots n - r,$
- $\lambda_k(A) \leq \lambda_{k+r}(A + B), \quad k = 1, 2, \cdots n - r.$

Lemmas 2.1, 2.2, and 2.3 can be found, e.g., as Theorem 4.3.1, Corollary 7.7.4., and Theorem 4.3.6, respectively, in [10].

2.2. Projection vectors chosen as eigenvectors. In this section we compare the effective condition number of P_DA and P_CA if the projection vectors are equal to eigenvectors of A.

DEFINITION 2.4. *Choose the eigenvectors v_k of A such that $v_k^T v_j = \delta_{kj}$, and define $Z = [v_1 \ldots v_r]$.*

THEOREM 2.5. *Using Z as given in Definition 2.4, the spectra of P_DA and P_CA given in (1.2) and (1.7) are*

$$\text{spectrum}(P_DA) = \{0, \ldots, 0, \lambda_{r+1}, \ldots, \lambda_n\} \text{ and}$$

$$\text{spectrum}(P_CA) = \{\sigma + \lambda_1, \ldots, \sigma + \lambda_r, \lambda_{r+1}, \ldots, \lambda_n\}.$$

Proof. For this choice of Z we have that

$$E = Z^T AZ = \operatorname{diag}(\lambda_1, \ldots, \lambda_r). \tag{2.1}$$

To prove the first part we note that (2.1) implies $P_D = I - AZE^{-1}Z^T = I - ZZ^T$. Consider $P_D A v_k = (I - ZZ^T)\lambda_k v_k$ for $k = 1, \ldots, n$. Since $ZZ^T v_k = v_k$, for $k = 1, \ldots, r$ and $ZZ^T v_k = 0$ for $k = r+1, \ldots, n$ it is easy to show that

$$P_D A v_k = 0, \text{ for } k = 1, \ldots, r, \text{ and } P_D A v_k = \lambda_k v_k, \text{ for } k = r+1, \ldots, n,$$

which proves the first part.

Second, we consider $P_C A v_k$. For $k = 1, \ldots, r$ we obtain

$$P_C A v_k = \left(I + \sigma Z \operatorname{diag}\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_r} \right) Z^T \right) \lambda_k v_k = (\sigma + \lambda_k) v_k,$$

whereas for $k = r+1, \ldots, n$ it appears that

$$P_C A v_k = \left(I + \sigma Z \operatorname{diag}\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_r} \right) Z^T \right) \lambda_k v_k = \lambda_k v_k$$

since $Z^T v_k = 0$ for $k = r+1, \ldots, n$. This proves the second part (cf. Theorem 2.6 in [20]). ∎

In order to compare both approaches we note that

$$\kappa_{eff}(P_D A) = \frac{\lambda_n}{\lambda_{r+1}} \tag{2.2}$$

and

$$\kappa(P_C A) = \frac{\max\{\sigma + \lambda_r, \lambda_n\}}{\min\{\sigma + \lambda_1, \lambda_{r+1}\}}. \tag{2.3}$$

From (2.2) and (2.3) it follows that $\kappa(P_C A) \geq \kappa_{eff}(P_D A)$, so the convergence bound based on the effective condition number implies that deflated CG converges faster than CG combined with coarse grid correction if both methods use the eigenvectors corresponding to the r smallest eigenvalues as projection vectors.

2.3. Projection vectors chosen as general vectors. In the last section we showed that the deflation technique is better than a coarse grid correction, if eigenvectors are used. However, computing the r smallest eigenvalues is, in general, very expensive. Moreover, in multigrid methods and domain decomposition methods special interpolation and prolongation matrices are used to obtain grid independent convergence rates. So a comparison only for eigenvectors is not enough. But in this section we generalize the results of section 2.2. We prove that the effective condition number of the deflated preconditioned system is always, for all matrices $Z \in \mathbb{R}^{n \times r}$, below the condition number of the system preconditioned by the coarse grid correction.

THEOREM 2.6. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric positive definite. Let $Z \in \mathbb{R}^{n \times r}$ with* rank $Z = r$. *Then the preconditioner defined in* (1.2) *and* (1.7) *satisfies*

$$\lambda_1(P_D A) = \cdots = \lambda_r(P_D A) = 0, \tag{2.4}$$

$$\lambda_n(P_D A) \leq \lambda_n(P_C A), \tag{2.5}$$

$$\lambda_{r+1}(P_D A) \geq \lambda_1(P_C A). \tag{2.6}$$

Proof. Obviously all eigenvalues of $P_C A$ are real and positive. By Lemma 2.1 of [8], $P_D A$ is positive semidefinite. Thus, all eigenvalues of $P_D A$ are real and nonnegative. Since $P_D AZ = 0$, statement (2.4) holds.

We obtain

$$A^{\frac{1}{2}} P_C A^{\frac{1}{2}} - P_D A = AZE^{-1}Z^T A + \sigma A^{\frac{1}{2}} ZE^{-1}Z^T A^{\frac{1}{2}}.$$

The right-hand side is positive semidefinite. Thus, we have with Lemma 2.2 that

$$\lambda_i(P_C A) = \lambda_i(A^{\frac{1}{2}} P_C A^{\frac{1}{2}}) \geq \lambda_i(P_D A).$$

Hence, (2.5) holds. Next consider

$$\begin{aligned} P_C A P_C - P_D A &= A + \sigma ZE^{-1}Z^T A + \sigma AZE^{-1}Z^T + \sigma^2 ZE^{-1}Z^T AZE^{-1}Z^T \\ &\quad -A + AZE^{-1}Z^T A \\ &= \sigma ZE^{-1}Z^T A + \sigma AZE^{-1}Z^T + \sigma^2 ZE^{-1}Z^T + AZE^{-1}Z^T A \\ &= (A + \sigma I)ZE^{-1}Z^T(A + \sigma I). \end{aligned}$$

Thus, $P_C A P_C - P_D A$ is symmetric and of rank r. Using Lemma 2.3 we obtain

$$\lambda_{r+1}(P_D A) \geq \lambda_1(P_C A P_C) = \lambda_1(P_C^2 A).$$

But since $P_C - I$ is positive semidefinite, $P_C^2 - P_C$ and $A^{\frac{1}{2}} P_C^2 A^{\frac{1}{2}} - A^{\frac{1}{2}} P_C A^{\frac{1}{2}}$ are positive semidefinite also. Hence,

$$\lambda_i(P_C^2 A) = \lambda_i(A^{\frac{1}{2}} P_C^2 A^{\frac{1}{2}}) \geq \lambda_i(A^{\frac{1}{2}} P_C A^{\frac{1}{2}}) = \lambda_i(P_C A).$$

Thus,

$$\lambda_{r+1}(P_D A) \geq \lambda_1(P_C^2 A) \geq \lambda_1(P_C A). \qquad \square$$

It follows from Theorem 2.6 that

$$\kappa(P_C A) \geq \kappa_{eff}(P_D A),$$

so the convergence bound based on the effective condition number implies that deflated CG converges faster than CG combined with coarse grid correction for arbitrary matrices $Z \in \mathbb{R}^{n \times r}$.

In Theorem 2.11 we will extend this result to the preconditioned versions of the deflation and coarse grid correction preconditioners.

Before that, we will show how the deflated preconditioner behaves if we increase the number of deflation vectors. In detail we will show that the effective condition number decreases if we use a matrix Z_2 in (1.2) satisfying $ImZ \subseteq ImZ_2$ rather than Z. To do so we need several lemmas.

The first lemma is probably well known, but for completeness we give the proof here.

LEMMA 2.7. *Let $A \in \mathbb{R}^{n \times n}$ be nonsingular and be partitioned as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11} \in M_r(R)$ and $A_{22} \in M_{n-r}(R)$. Assume that A_{11} is nonsingular. Define

$$\tilde{A}_{11}^{-1} := \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Then, rank $(A^{-1} - \tilde{A}_{11}^{-1}) = n - r$.

Proof. The inverse of A is given by

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix},$$

where $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$. Hence

$$\begin{aligned} A^{-1} - \tilde{A}_{11}^{-1} &= \begin{bmatrix} A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1} \end{bmatrix} \left[A_{21}A_{11}^{-1}, -I\right]. \end{aligned}$$

Since S and the $n-r \times n-r$ identity matrix I have rank $n-r$, we get rank $(A^{-1} - \tilde{A}_{11}^{-1}) = n - r$. ∎

In the next lemma we compare the preconditioned matrices if a different number of deflation vectors is used.

LEMMA 2.8. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Let $Z_1 \in \mathbb{R}^{n\times r}$ and $Z_2 \in \mathbb{R}^{n\times s}$ with* rank $Z_1 = r$ *and* rank $Z_2 = s$. *Define $E_1 := Z_1^T A Z_1$ and $E_2 := Z_2^T A Z_2$. If $ImZ_1 \subseteq ImZ_2$, then*

$$(I - AZ_1E_1^{-1}Z_1^T)A \succeq (I - AZ_2E_2^{-1}Z_2^T)A.$$

Proof. It suffices to prove that

$$Z_2E_2^{-1}Z_2^T \succeq Z_1E_1^{-1}Z_1^T.$$

Since $ImZ_1 \subseteq ImZ_2$, there exists a matrix $T \in M_{s\times r}(R)$ such that

$$Z_1 = Z_2T.$$

Therefore,

$$\begin{aligned} Z_2E_2^{-1}Z_2^T - Z_1E_1^{-1}Z_1^T &= Z_2(E_2^{-1} - TE_1^{-1}T^T)Z_2^T \\ &= Z_2E_2^{-\frac{1}{2}}(I - E_2^{\frac{1}{2}}TE_1^{-1}T^TE_2^{\frac{1}{2}})E_2^{-\frac{1}{2}}Z_2^T. \end{aligned}$$

Moreover, we have

$$\begin{aligned} (E_2^{\frac{1}{2}}TE_1^{-1}T^TE_2^{\frac{1}{2}})^2 &= E_2^{\frac{1}{2}}TE_1^{-1}T^TE_2TE_1^{-1}T^TE_2^{\frac{1}{2}} \\ &= E_2^{\frac{1}{2}}TE_1^{-1}T^TZ_2^TAZ_2TE_1^{-1}T^TE_2^{\frac{1}{2}} \\ &= E_2^{\frac{1}{2}}TE_1^{-1}Z_1^TAZ_1E_1^{-1}T^TE_2^{\frac{1}{2}} \\ &= E_2^{\frac{1}{2}}TE_1^{-1}E_1E_1^{-1}T^TE_2^{\frac{1}{2}} \\ &= E_2^{\frac{1}{2}}TE_1^{-1}T^TE_2^{\frac{1}{2}}. \end{aligned}$$

Hence, $E_2^{\frac{1}{2}}TE_1^{-1}T^TE_2^{\frac{1}{2}}$ is an orthogonal projection. Thus $E_2^{\frac{1}{2}}TE_1^{-1}T^TE_2^{\frac{1}{2}}$ has only the eigenvalues 0 and 1. Hence, $I - E_2^{\frac{1}{2}}TE_1^{-1}T^TE_2^{\frac{1}{2}}$ is positive semidefinite. Therefore,

$$Z_2E_2^{-1}Z_2^T \succeq Z_1E_1^{-1}Z_1^T.$$ ∎

In the next lemma we show that $P_{D_1}A = P_{D_2}A$, if $ImZ_1 = ImZ_2$.

LEMMA 2.9. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Let $Z_1 \in \mathbb{R}^{n\times r}$ and $Z_2 \in \mathbb{R}^{n\times r}$ with* $\text{rank}Z_1 = \text{rank}Z_2 = r$. *Define $E_1 := Z_1^T AZ_1$ and $E_2 := Z_2^T AZ_2$. If $ImZ_1 = ImZ_2$, then*

$$(I - AZ_1E_1^{-1}Z_1^T)A = (I - AZ_2E_2^{-1}Z_2^T)A.$$

Proof. We can follow the proof of Lemma 2.8. Since $ImZ_1 = ImZ_2$, the matrix T is nonsingular. Hence,

$$\begin{aligned} Z_2E_2^{-1}Z_2^T - Z_1E_1^{-1}Z_1^T &= Z_2(E_2^{-1} - TE_1^{-1}T^T)Z_2^T \\ &= Z_2E_2^{-\frac{1}{2}}(I - E_2^{\frac{1}{2}}TE_1^{-1}T^TE_2^{\frac{1}{2}})E_2^{-\frac{1}{2}}Z_2^T. \\ &= Z_2E_2^{-\frac{1}{2}}(I - E_2^{\frac{1}{2}}T(T^TE_2T)^{-1}T^TE_2^{\frac{1}{2}})E_2^{-\frac{1}{2}}Z_2^T \\ &= Z_2E_2^{-\frac{1}{2}}(I - E_2^{\frac{1}{2}}TT^{-1}E_2^{-1}T^{-T}T^TE_2^{\frac{1}{2}})E_2^{-\frac{1}{2}}Z_2^T \\ &= 0. \qquad \square \end{aligned}$$

Using the above lemmas, we can prove the following theorem.

THEOREM 2.10. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Let $Z_1 \in \mathbb{R}^{n\times r}$ and $Z_2 \in \mathbb{R}^{n\times s}$ with* $\text{rank}Z_1 = r$ *and* $\text{rank}Z_2 = s$. *Let $E_1 := Z_1^T AZ_1$ and $E_2 := Z_2^T AZ_2$. If $ImZ_1 \subseteq ImZ_2$, then*

$$\lambda_n((I - AZ_1E_1^{-1}Z_1^T)A) \geq \lambda_n((I - AZ_2E_2^{-1}Z_2^T)A), \tag{2.7}$$

$$\lambda_{r+1}((I - AZ_1E_1^{-1}Z_1^T)A) \leq \lambda_{s+1}((I - AZ_2E_2^{-1}Z_2^T)A). \tag{2.8}$$

Proof. With Lemmas 2.2 and 2.8 we obtain inequality (2.7).

Next, we will prove (2.8). Observe that $Z_1E_1^{-1}Z_1^T$ and $Z_2E_2^{-1}Z_2^T$ are invariant under permutations of the columns of Z_1 and Z_2, respectively.

Thus, using Lemma 2.9, we can assume without loss of generality that $Z_2 = [Z_1, D]$ with $D \in \mathbb{R}^{n\times s-r}$.

Moreover, define the $s \times s$ matrix

$$\tilde{E}_1^{-1} = \begin{bmatrix} E_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Obviously, we then obtain

$$Z_1E_1^{-1}Z_1^T = Z_2\tilde{E}_1^{-1}Z_2^T.$$

Thus,

$$\begin{aligned} (I - AZ_2E_2^{-1}Z_2^T)A - (I - AZ_1E_1^{-1}Z_1^T)A &= A(Z_1E_1^{-1}Z_1^T - Z_2E_2^{-1}Z_2^T)A \\ &= A(Z_2\tilde{E}_1^{-1}Z_2 - Z_2E_2^{-1}Z_2^T)A \\ &= AZ_2(\tilde{E}_1^{-1} - E_2^{-1})Z_2^TA. \end{aligned}$$

But since E_1 is the leading principal $r \times r$ submatrix of E_2, we can apply Lemma 2.7. Thus $(I - AZ_2E_2^{-1}Z_2^T)A - (I - AZ_1E_1^{-1}Z_1^T)A$ is of rank $s - r$. Hence, with Lemma 2.3,

$$\lambda_{r+1}((I - AZ_1E_1^{-1}Z_1^T)A) \leq \lambda_{s+1}((I - AZ_2E_2^{-1}Z_2^T)A). \qquad \square$$

Theorem 2.10 states that the effective condition number decreases if we increase the number of deflation vectors. However, the dimension of the system Z^TAZ which has to be solved increases also.

Next, we include an additional symmetric positive definite preconditioner M^{-1}. Then we consider the coarse grid preconditioner

$$P_{CM^{-1}} := M^{-1} + \sigma ZE^{-1}Z^T. \tag{2.9}$$

This type of preconditioner includes many well-known preconditioners. It belongs to the class of additive Schwarz preconditioners and is called the two-level additive Schwarz preconditioner. If used in domain decomposition methods, typically M^{-1} is the sum of the local (exact or inexact) solves in each domain. To speed up convergence a coarse grid correction $ZE^{-1}Z^T$ is added. Notice that the Bramble–Pasciak–Schatz (BPS) preconditioner introduced in [2] and by Dryja and Widlund [5, 6] and Dryja [4] are of the same type. They show under mild conditions that the convergence rate of the PCG method is independent of the grid sizes.

We compare the preconditioner (2.9) with the corresponding deflated preconditioner

$$M^{-1}P_D. \tag{2.10}$$

We obtain the following theorem.

THEOREM 2.11. *Let $A \in \mathbb{R}^{n\times n}$ and $M \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Let $Z \in \mathbb{R}^{n\times r}$ with* $\mathrm{rank}Z = r$. *Then*

$$\lambda_n(M^{-1}P_DA) \le \lambda_n(P_{CM^{-1}}A), \tag{2.11}$$

$$\lambda_{r+1}(M^{-1}P_DA) \ge \lambda_1(P_{CM^{-1}}A). \tag{2.12}$$

Proof. First observe that Theorem 2.6 still holds if we replace A everywhere by $L^{-1}AL^{-T}$ with an arbitrary nonsingular matrix L. Here, we will consider $M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$. The idea is to transform P_D and P_C to this form. We start with

$$M^{-1}P_DA = M^{-1}(A - AZE^{-1}Z^TA).$$

The eigenvalues of this matrix are the same as the eigenvalues of

$$M^{-\frac{1}{2}}P_DAM^{-\frac{1}{2}} = M^{-\frac{1}{2}}(A - AZE^{-1}Z^TA)M^{-\frac{1}{2}}.$$

Define the matrix G such that $G = M^{\frac{1}{2}}Z$ and thus $Z = M^{-\frac{1}{2}}G$. Substituting this in the previous matrix leads to $E = Z^TAZ = G^TM^{-\frac{1}{2}}AM^{-\frac{1}{2}}G$ and

$$\begin{aligned} M^{-\frac{1}{2}}P_DAM^{-\frac{1}{2}} &= M^{-\frac{1}{2}}(A - AM^{-\frac{1}{2}}GE^{-1}G^TM^{-\frac{1}{2}}A)M^{-\frac{1}{2}} \\ &= (I - M^{-\frac{1}{2}}AM^{-\frac{1}{2}}GE^{-1}G^T)M^{-\frac{1}{2}}AM^{-\frac{1}{2}}, \end{aligned}$$

which is in the required form.

In the same way we can transform $P_{CM^{-1}}A = (M^{-1} + \sigma ZE^{-1}Z^T)A$ to

$$P_{CM^{-1}}A = M^{-1}A + \sigma M^{-\frac{1}{2}}GE^{-1}G^TM^{-\frac{1}{2}}A,$$

which has the same eigenvalues as

$$M^{-\frac{1}{2}}AM^{-\frac{1}{2}} + \sigma GE^{-1}G^TM^{-\frac{1}{2}}AM^{-\frac{1}{2}} = (I + \sigma GE^{-1}G^T)M^{-\frac{1}{2}}AM^{-\frac{1}{2}},$$

which is also in the required form.

Thus, Theorem 2.6 gives the desired result. ∎

For the case $L^{-1}AL^{-T}$ the same result can be proved if one chooses $G = L^T Z$.

Theorem 2.11 describes the most general case. Arbitrary vectors or matrices $Z \in \mathbb{R}^{n\times r}$ combined with arbitrary preconditioners are considered. The effective condition number of the deflated CG method is always below the condition number of the CG method preconditioned by the coarse grid correction. Thus, the interpolation or prolongation matrices Z used, for example, in the BPS method give a better preconditioner if used in a deflation technique.

At the end of this section we generalize Theorem 2.10.

THEOREM 2.12. *Let $A, M \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Let $Z_1 \in \mathbb{R}^{n\times r}$ and $Z_2 \in \mathbb{R}^{n\times s}$ with* $\text{rank}Z_1 = r$ *and* $\text{rank}Z_2 = s$. *Let $E_1 := Z_1^T A Z_1$ and $E_2 := Z_2^T A Z_2$. If $ImZ_1 \subseteq ImZ_2$, then*

$$\lambda_n(M^{-1}(I - AZ_1E_1^{-1}Z_1^T)A) \geq \lambda_n(M^{-1}(I - AZ_2E_2^{-1}Z_2^T)A), \tag{2.13}$$

$$\lambda_{r+1}(M^{-1}(I - AZ_1E_1^{-1}Z_1^T)A) \leq \lambda_{s+1}(M^{-1}(I - AZ_2E_2^{-1}Z_2^T)A). \tag{2.14}$$

Proof. The proof is almost the same as the proof of Theorem 2.10.

3. Other properties of deflation and coarse grid correction. In this section we compare other properties of deflation and coarse grid correction. These properties are scaling, inaccurate solution, and an estimate of the smallest eigenvalue.

Scaling. Note that $P_D A$ is scaling invariant, whereas $P_C A$ is not scaling invariant. This means that if deflation is applied to a system $\gamma Ax = \gamma b$, the effective condition number of $P_{D\gamma A}\gamma A = (I - \gamma AZ(Z^T\gamma AZ)^{-1}Z^T)\gamma A$ is independent of the scalar γ, i.e.,

$$\kappa_{eff}(P_{D\gamma A}\gamma A) = \frac{\gamma\lambda_n(P_{DA}A)}{\gamma\lambda_{r+1}(P_{DA}A)} = \kappa_{eff}(P_{DA}A).$$

Whereas the condition number of $P_C\gamma A$ depends on the choice of γ,

$$\kappa(P_{C\gamma A}\gamma A) \neq \kappa(P_{CA}A).$$

Inaccurate solution. If the dimension matrix E becomes large (i.e., many projection vectors are used), it seems to be good to compute E^{-1} approximately (by an iterative method or by doing the procedure recursively). It appears that the coarse grid correction operator is insensitive to the accuracy of the approximation of E^{-1}, while the deflation is sensitive to it. A proof of this property if the projection vectors are eigenvectors is given in the next lemma.

LEMMA 3.1. *Use Z as given in Definition* 2.4*, and assume that*

$$\tilde{E}^{-1} = \text{diag}\left(\frac{1}{\lambda_1}(1-\epsilon_1), \ldots, \frac{1}{\lambda_r}(1-\epsilon_r)\right)$$

is an approximation of E^{-1}, where $|\epsilon_i|$ is small. The spectra of $\tilde{P}_D A$ and $\tilde{P}_C A$ given in (1.2) *and* (1.5)*, where E^{-1} is replaced by $\tilde{E}^{-1}$, are*

$$\text{spectrum}(\tilde{P}_D A) = \{\lambda_1\epsilon_1, \ldots, \lambda_r\epsilon_r, \lambda_{r+1}, \ldots, \lambda_n\} \text{ and}$$

$$\text{spectrum}(\tilde{P}_C A) = \{\lambda_1 + \sigma(1-\epsilon_1), \ldots, \lambda_r + \sigma(1-\epsilon_r), \lambda_{r+1}, \ldots, \lambda_n\}.$$

Proof. The proof of this lemma is almost the same as the proof of Theorem 2.5. ∎

For general vectors a similar situation appears. Assume that $\tilde{E}^{-1} = (I-F)E^{-1}(I-F)$ is a symmetric approximation ($F = F^T$) of E^{-1}. Let $H := -FE^{-1} - E^{-1}F + FE^{-1}F$. Then we have

$$\tilde{P}_D A = P_D A + AZHZ^T A.$$

Hence, using Lemma 2.1 we obtain

$$\lambda_k(P_D A) + \lambda_1(AZHZ^T A) \leq \lambda_k(\tilde{P}_D A) \leq \lambda_k(P_D A) + \lambda_n(AZHZ^T A).$$

Since the first r eigenvalues of $\lambda_k(P_D A)$ are 0, we get for $i = 1, \ldots, r$,

$$\lambda_1(AZHZ^T A) \leq \lambda_i(\tilde{P}_D A) \leq \lambda_n(AZHZ^T A).$$

If all eigenvalues of $AZHZ^T A$ are small, the first r eigenvalues $\lambda_i(\tilde{P}_D A)$ also are very small. Observe that $\lambda_1(\tilde{P}_D A)$ can be negative if the perturbation H is negative definite.

For the coarse grid correction

$$\tilde{P}_C A = P_C A + ZHZ^T A,$$

we obtain

$$\lambda_k(P_C A) + \lambda_1(ZHZ^T A) \leq \lambda_k(\tilde{P}_C A) \leq \lambda_k(P_C A) + \lambda_n(ZHZ^T A).$$

Thus, if all eigenvalues of $ZHZ^T A$ are small, the perturbation does not have much effect.

Hence, the coarse grid correction operator is insensitive for the accuracy of the approximation, whereas deflation is sensitive.

To illustrate this we consider two problems. The first one is motivated by a porous media flow with large contrasts in the coefficients (ratio 10^{-6}; see the seven-layer problem in section 4), and the second one is a Poisson problem. In both examples $r = 7$ *algebraic* projection vectors are used (see [28, Def. 4]). We replace E^{-1} by $\tilde{E}^{-1} = (I + \epsilon R)E^{-1}(I + \epsilon R)$, where R is a symmetric $r \times r$ matrix with random elements chosen from the interval $[-\frac{1}{2}, \frac{1}{2}]$. From Figure 3.1 (porous media flow) it follows that the convergence of the error remains good for $|\epsilon| < 10^{-12}$. For larger values of $|\epsilon|$ we see that the convergence stagnates. For the Poisson problem it appears that the convergence is good as long as $|\epsilon| < 10^{-6}$ (see Figure 3.2). For the coarse grid correction operator, there is no difference in the convergence behavior. Using the coarse grid correction operator we need 75 iterations for the porous media flow problem and 70 iterations for the Poisson problem.

We also have investigated the convergence behavior of deflation if a perturbed Cholesky decomposition of E is used. For this experiment we compute the Cholesky factor L of E and use in the deflation method the matrix $\tilde{L}$ which is such that $\tilde{L}_{ij} = L_{ij}(1 + \epsilon_{ij})$ and $|\epsilon_{ij}| < \epsilon$. In Figure 3.3 the results are given. We observe again that the convergence stagnates if ϵ is too large.

Estimate of smallest eigenvalue. In this paragraph we restrict ourselves to the case that the deflation vectors approximate the eigenvectors corresponding to the smallest eigenvalues. In practice it is very important to have a reliable stopping criterion,

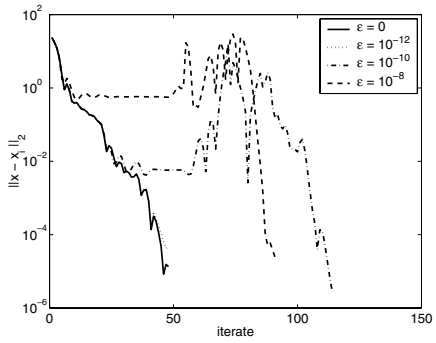


FIG. 3.1. *Convergence behavior of DICCG for the straight layers problem.*

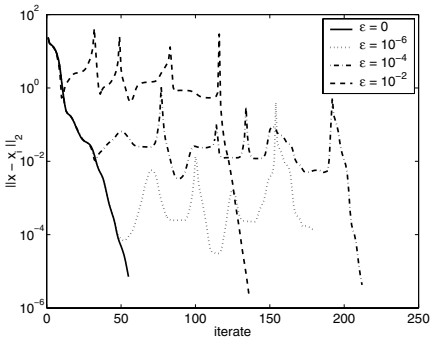


FIG. 3.2. *Convergence behavior of DICCG for the Poisson problem.*

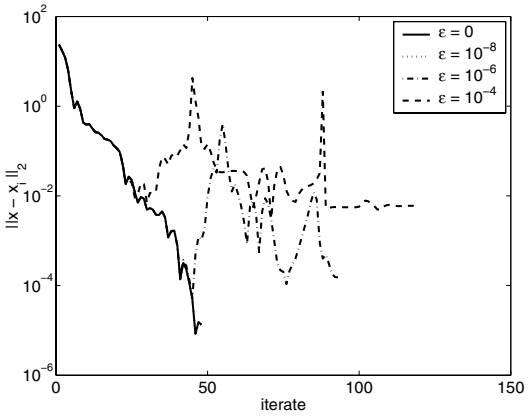


FIG. 3.3. *DICCG for the straight layers problem with a perturbed Cholesky decomposition.*

especially for a porous media flow problem, because for such a problem the linear system is ill conditioned. The following stopping criterion

$$\|r_k\|_2 \leq \lambda_1 \|x_k\|_2 \epsilon \tag{3.1}$$

gives that

$$\frac{\|x - x_k\|_2}{\|x_k\|_2} \leq \epsilon,$$

which implies that the relative error is small. To use this criterion, an estimate of the smallest eigenvalue should be available. From the CG method an approximation of the extreme eigenvalues can be obtained from the Ritz values (see [11]). However, for the deflated operator $P_D A$ this leads to an estimate of λ_{r+1} instead of λ_1. The same holds for the preconditioned system. In order to estimate λ_1 we note that

$$\lambda_1(M^{-\frac{1}{2}} A M^{-\frac{1}{2}}) \leq \min_{y \in \mathbb{R}^r} \frac{y^T G^T M^{-\frac{1}{2}} A M^{-\frac{1}{2}} G y}{y^T G^T G y} = \min_{y \in \mathbb{R}^r} \frac{y^T Z^T A Z y}{y^T Z^T M Z y}.$$

This means that the smallest eigenvalue μ_{min} of the generalized eigenvalue problem

$$Ey = \mu Z^T M Z y$$

is an upper bound for the smallest eigenvalue of $M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$, whereas the smallest eigenvalue μ_{min} of the generalized eigenvalue problem

$$Ey = \mu Z^T Zy$$

is an upper bound for the smallest eigenvalue of A. From experiments for the porous media flow problem, it appears that the estimates are reasonably sharp (see Table 3.1), so they can be used in stopping criterion (3.1).

TABLE 3.1
The estimated smallest eigenvalue using matrix E.

Matrix	λ_1	$\lambda_1(estimated)$
$M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$	$0.7 \cdot 10^{-8}$	$3.1 \cdot 10^{-8}$
A	$3.3 \cdot 10^{-9}$	$9.9 \cdot 10^{-9}$

4. Numerical experiments. All numerical experiments are done by using the SEPRAN FEM package developed at Delft University of Technology. The multiplication $y = E^{-1}b$ is always done by solving y from $Ey = b$, where E is decomposed in its Cholesky factor. The choice of the boundary conditions is such that all problems have as exact solution the vector with components equal to 1. In order to make the convergence behavior representative for general problems, we chose a random vector as starting solution, instead of the zero start vector.

4.1. Porous media flows. In this section we consider problems motivated by porous media flow (see [27]). Our first problem is a simple two-dimensional model problem, whereas our second problem mimics the flow of oil in a reservoir. In both problems *physical* projection vectors are used (see [28, Def. 2]), which approximate the eigenvectors corresponding to the small eigenvalues.

Seven-layer problem. We solve the equation

$$\text{div } (\sigma \nabla p) = 0$$

with p the fluid pressure and σ the permeability. At the earth's surface the fluid pressure is prescribed. At the other boundaries we use homogeneous Neumann conditions. In this two-dimensional problem we consider seven horizontal layers. We use linear triangular elements, and the number of grid points is equal to 22,680. The top layer is sandstone, then a shale layer, etc. We assume that σ in sandstone is equal to 1 and σ in shale is equal to 10^{-7}. From [26] it follows that the IC preconditioned matrix has three eigenvalues of order 10^{-7}, whereas the remaining eigenvalues are of order 1. Computing the solution with three projection vectors, we observe that in every iteration the norm of the residual using deflation or coarse grid correction is comparable. In Figure 4.1 the norm of the error for both methods is given. Note that the error using deflation stagnates at a lower level than that of coarse grid correction. This surprises us because the results presented in section 3 suggested that deflation can be more sensitive to rounding errors than coarse grid correction.

An oil flow problem. In this paragraph we simulate a porous media flow in a three-dimensional layered geometry, where the layers vary in thickness and orientation (see Figures 4.2 and 4.3 for a four-layer problem). Figure 4.2 shows a part of the earth's crust. The depth of this part varies between three and six kilometers, whereas horizontally its dimensions are 40 x 60 kilometers. The upper layer is a mixture of sandstone and shale and has a permeability of 10^{-4}. Below this layer, shale and

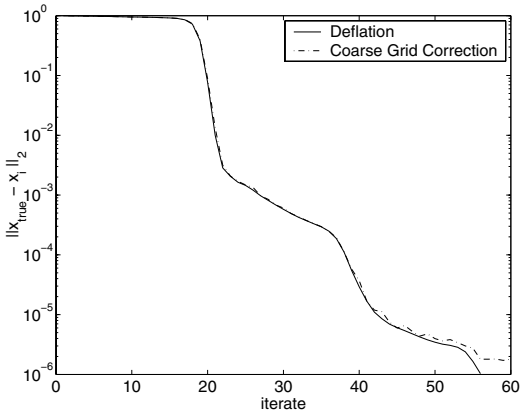


FIG. 4.1. *The norm of the error for projected ICCG for the seven-layer problem.*

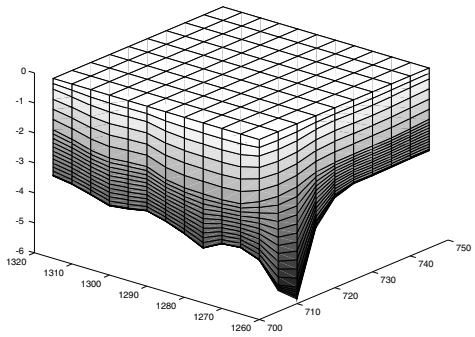


FIG. 4.2. *The geometry of an oil flow problem.*

sandstone layers are present with permeabilities of 10^{-7} and 10, respectively. We consider a problem with nine layers. Five sandstone layers are separated by four shale layers. At the top of the first sandstone/shale layer a Dirichlet boundary condition is posed, so the IC preconditioned matrix has four small eigenvalues. We use four *physical* projection vectors and stop if $\|r_k\|_2 \leq 10^{-5}$. Trilinear hexahedral elements are used, and the total number of gridpoints is equal to 148,185. The results are given in Table 4.1 and correspond well with our theoretical results.

4.2. Parallel problems. In this section we consider a Poisson equation on a two-dimensional rectangular domain. On top a Dirichlet boundary condition is posed, whereas at the other boundaries a homogeneous Neumann condition is used. We use linear triangular elements. We stop the iteration if $\|r_k\|_2 \leq 10^{-8}$.

As a first test we solve a problem, in which the grid is decomposed into seven layers with various gridsizes per layer. The results are given in Table 4.2. In this table the symbol "No" means that there is no projection method used. Note that in the parallel case we use a block IC preconditioner. Deflation again needs fewer iterations than coarse grid correction. However, both projection methods lead to a considerable gain in the number of iterations. Note that the number of iterations increases if the gridsize per layer increases.

Second, we consider the parallel performance for an increasing number of layers

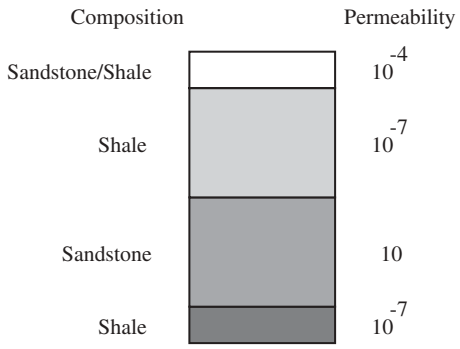


FIG. 4.3. *Permeabilities for each layer.*

TABLE 4.1
The results for the oil flow problem.

Method	Deflation	CGC
Iterations	36	47
CPU time	5.9	8.2

TABLE 4.2
The effect of the gridsize per layer.

	Sequential			Parallel		
Grid points	Deflation	CGC	No	Deflation	CGC	No
10×10	21	29	35	25	38	50
20×20	36	48	65	42	61	90
40×40	62	82	125	80	103	168
80×80	106	131	244	128	161	321

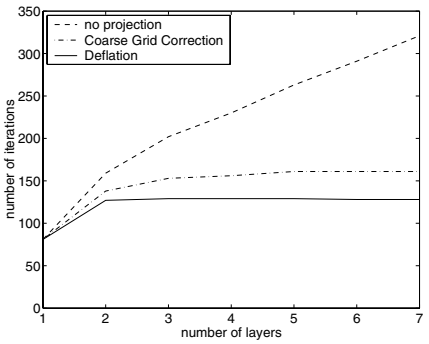


FIG. 4.4. *The number of iterations for a layered domain decomposition.*

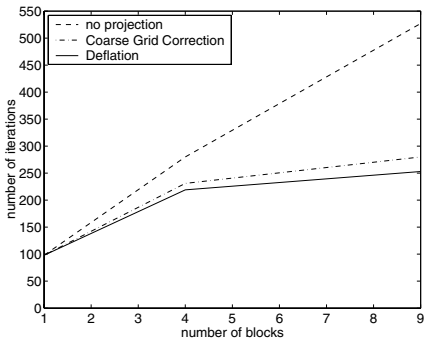


FIG. 4.5. *The number of iterations for a block domain decomposition.*

or blocks. The gridsize per layer is 80×80 and per block is 100×100. This implies that the total number of grid points increases proportionally to the number of layers/blocks. In Figures 4.4 and 4.5 the results are given. Note that initially both projection methods show a small increase in the number of iterations if the number of layers/blocks increases but thereafter the number of iterations is constant (scalable). If no projection method is used, the number of iterations keep increasing.

5. Conclusions. We have compared various preconditioners used in the numerical solution of partial differential equations. On one hand we considered a coarse grid correction preconditioner. On the other hand a so-called deflation preconditioner was studied. It turned out that the effective condition number of the deflated preconditioned system is always, for all deflation vectors and all restrictions and prolongations, below the condition number of the system preconditioned by the coarse grid correction. This implies that the CG method applied to the deflated preconditioned system converges always faster than the CG method applied to the system preconditioned by the coarse grid correction. Numerical results for porous media flows and parallel preconditioners emphasized the theoretical results.

REFERENCES

[1] M. Benzi, A. Frommer, R. Nabben, and D. Szyld, *Algebraic theory of multiplicative Schwarz methods*, Numer. Math., 89 (2001), pp. 606–639.

[2] J. Bramble, J. Pasciak, and A. Schatz, *The construction of preconditioners for elliptic problems by substructuring.* I, Math. Comput., 47 (1986), pp. 103–134.

[3] H. De Gersem and K. Hameyer, *A deflated iterative solver for magnetostatic finite element models with large differences in permeability*, Eur. Phys. J. Appl. Phys., 13 (2000), pp. 45–49.

[4] M. Dryja, *An additive Schwarz algorithm for two- and three-dimensional finite element elliptic problems*, in Domain Decomposition Methods, T. Chan et al., eds., SIAM, Philadelphia, 1989, pp. 168–172.

[5] M. Dryja and O. Widlund, *Towards a unified theory of domain decomposition algorithms for elliptic problems*, in Proceedings of the Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1990, pp. 3–21.

[6] M. Dryja and O. Widlund, *Multilevel additive methods for elliptic finite element problems in three dimensions*, in Parallel Algorithms for PDEs, Vieweg, Braunschweig, Germany, 1991, pp. 58–69.

[7] M. Eiermann, O. Ernst, and O. Schneider, *Analysis of acceleration strategies for restarted minimal residual methods*, J. Comput. Appl. Math., 123 (2000), pp. 261–292.

[8] J. Frank and C. Vuik, *On the construction of deflation-based preconditioners*, SIAM J. Sci. Comput., 23 (2001), pp. 442–462.

[9] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[10] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[11] E. Kaasschieter, *A practical termination criterion for the conjugate gradient method*, BIT, 28 (1988), pp. 308–322.

[12] E. F. Kaasschieter, *Preconditioned conjugate gradients for solving singular systems*, J. Comput. Appl. Math., 24 (1988), pp. 265–275.

[13] L. Kolotilina, *Preconditioning of systems of linear algebraic equations by means of twofold deflation.* I. *Theory*, J. Math. Sci. (New York), 89 (1998), pp. 1652–1689.

[14] L. Mansfield, *On the conjugate gradient solution of the Schur complement system obtained from domain decomposition*, SIAM J. Numer. Anal., 27 (1990), pp. 1612–1620.

[15] L. Mansfield, *Damped Jacobi preconditioning and coarse grid deflation for conjugate gradient iteration on parallel computers*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1314–1323.

[16] J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.

[17] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.

[18] R. Nabben, *Comparisons between additive and multiplicative Schwarz iterations in domain decomposition methods*, Numer. Math., 95 (2003), pp. 145–162.

[19] R. A. Nicolaides, *Deflation of conjugate gradients with applications to boundary value problems*, SIAM J. Numer. Anal., 24 (1987), pp. 355–365.

[20] A. Padiy, O. Axelsson, and B. Polman, *Generalized augmented matrix preconditioning approach and its application to iterative solution of ill-conditioned algebraic systems*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 793–818.

[21] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, Oxford, 1999.

[22] Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc'h, *A deflated version of the conjugate gradient algorithm*, SIAM J. Sci. Comput., 21 (2000), pp. 1909–1926.

[23] B. Smith, P. Bjørstad, and W. Gropp, *Domain Decomposition*, Cambridge University Press, Cambridge, UK, 1996.

[24] A. van der Sluis and H. A. van der Vorst, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.

[25] F. Vermolen, C. Vuik, and A. Segal, *Deflation in preconditioned conjugate gradient methods for finite element problems*, in Conjugate Gradient and Finite Element Methods, Springer-Verlag, Berlin, 2004, pp. 103–129.

[26] C. Vuik, A. Segal, J. Meijerink, and G. Wijma, *The construction of projection vectors for a deflated ICCG method applied to problems with extreme contrasts in the coefficients*, J. Comput. Phys., 172 (2001), pp. 426–450.

[27] C. Vuik, A. Segal, and J. A. Meijerink, *An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts in the coefficients*, J. Comput. Phys., 152 (1999), pp. 385–403.

[28] C. Vuik, A. Segal, L. E. Yaakoubi, and E. Dufour, *A comparison of various deflation vectors applied to elliptic problems with discontinuous coefficients*, Appl. Numer. Math., 41 (2002), pp. 219–233.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1648–1668

UNIFORM CONVERGENCE OF AN EXPONENTIALLY FITTED SCHEME FOR THE QUANTUM DRIFT DIFFUSION MODEL*

RENÉ PINNAU†

Abstract. We analyze an exponentially fitted finite element scheme for the unipolar quantum drift diffusion model in one-dimensional space. The existence of discrete solutions is shown under very mild assumptions, and convergence of a subsequence is proved by compactness arguments. The scheme is constructed in such a way that it reduces in the semiclassical limit to the well-known Scharfetter–Gummel discretization for the classical drift diffusion model. We derive uniform error bounds which allow for the semiclassical limit on the discrete level. Numerical tests underlining the analytical results are presented.

Key words. quantum drift diffusion, generalized Scharfetter–Gummel discretization, mixed finite elements, exponential fitting, uniform convergence, semiclassical limit, semiconductor

AMS subject classifications. 35J60, 35J70, 65N12, 65N15, 65N30, 76Y05

DOI. 10.1137/S0036142903429961

1. Introduction.

From the earliest days of semiconductor industry there has been a never-ending drive towards increased miniaturization. The original aim was to produce more devices per unit area, but now scientists and engineers are exploiting quantum size effects to introduce new electronic properties into existing materials. Many devices, like MOSFETs or resonant tunneling structures, already reached the decanano length scale [19]. The Semiconductor Industry Association (SIA) projects that by 2009 the leading edge MOS device will employ a 0.05 μm length scale and an oxide thickness of 1.5 nm or less. But already today quantum mechanical effects, like confinement in barrier structures or inversion layers as well as direct tunneling through the oxide causing gate leakage in MOS structures are no longer negligible [18]. Hence, scientists are in charge to develop "correct" models which can be easily incorporated into existing modern simulation tools.

During the last years much effort has been spent on the derivation and analysis of *macroscopic* quantum models, which allow for an accurate description of the underlying physics of the devices by reasonable numerical costs. Nowadays, there exists a whole hierarchy of macroscopic models leading from the quantum hydrodynamic (QHD) models [25, 20] over the quantum energy transport (QET) model to the quantum drift diffusion (QDD) model, which can be derived from a moment expansion of the Wigner–Poisson system (see [26, 31] and the references therein for a comprehensive overview). Recently, extensions of these models were derived, which are better suited to deal with quantum tunneling and coherence effects [21, 22, 13, 17].

In this work we analyze a new numerical scheme for the QDD model. The mathematical analysis and numerical understanding of this model is in a rather mature state [31]. Essentially, this model is a dispersive regularization of the classical drift diffusion (DD) model of Van Roosbroeck [29], which accounted for the immense success of the macroscopic theory of charge transport in semiconductors and is commonly

*Received by the editors June 20, 2003; accepted for publication (in revised form) May 18, 2004; published electronically December 27, 2004. This work was supported by European network HYKE, funded by the EC under contract HPRN-CT-2002-00282.

http://www.siam.org/journals/sinum/42-4/42996.html

†Fachbereich Mathematik, Technische Universität Darmstadt, D-64289 Darmstadt, Germany (pinnau@mathematik.tu-darmstadt.de).

used (with all its enhancements) in modern simulation tools. First, Ancona [8], Ancona and Tiersten [5], and Ancona and Iafrate [10] proposed a quantum correction of this well-understood system. This *density-gradient theory* is impressively capable of describing the correct device behavior in the vicinity of strong inversion layers in MOS structures when compared to one-electron quantum mechanic simulations [8]. But already the shrinking device size poses severe numerical problems, since the local field strength increases and interior layers in the solution become more abrupt [14]. The QDD model was employed for the simulation of many quantum semiconductor devices and has proved its numerical efficiency, especially in several space dimensions [7, 11, 37]. Due to its numerical robustness it is already programmed into the 2d/3d PROPHET simulation code from *Lucent Technologies* as well as into various commercial device simulators, e.g., those from *ISE* and *Silvaco*. Encouraging comparisons with Schrödinger–Poisson simulations can be found in [3, 37].

The unscaled QDD model equations stated on a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2$, or 3, read as

$$\frac{\partial n}{\partial t} + \frac{1}{q} \operatorname{div} J = 0, \tag{1.1a}$$

$$\frac{q\, k_B\, T_0}{m} \nabla n + \frac{q^2}{m}\, n\, \nabla V - \frac{q\, \hbar^2}{2\, m^2}\, n\, \nabla \left(\frac{\Delta \sqrt{n}}{\sqrt{n}} \right) = -\frac{J}{\tau_0}, \tag{1.1b}$$

which are self-consistently coupled with the Poisson equation for the electrostatic potential

$$-\epsilon\, \Delta V = q\, (n - C_{dop})\,. \tag{1.1c}$$

The variables are the electron density $n = n(x,t)$, the current density $J = J(x,t)$, and the electrostatic potential $V = V(x,t)$. The physical constants are the elementary charge q, the Boltzmann constant k_B, the effective electron mass m, and the reduced Planck constant $\hbar$. For the values of these constants we refer the reader to [29]. Physical parameters are the permittivity ϵ, the relaxation time τ_0, and the lattice temperature T_0. The time-independent doping profile $C_{dop} = C_{dop}(x)$ represents the distribution of charged background ions.

In this paper we consider the stationary QDD model for unipolar devices in one-dimensional space. We introduce the diffusion scaling, where the new dimensionless quantities are marked by a tilde:

$$n \to C_m\, \tilde{n}, \qquad C_{dop} \to C_m\, \tilde{C}_{dop}, \qquad x \to L\, \tilde{x},$$
$$t \to \frac{m\, L^2}{k_B\, T_0\, \tau_0} \tilde{t}, \qquad V \to \frac{k_B\, T_0}{q} \tilde{V}, \qquad J \to \frac{q\, k_B\, T_0\, C_m\, \tau_0}{L\, m} \tilde{J}.$$

Here, C_m denotes the maximal absolute value of the doping profile C_{dop} and L is a characteristic device length, e.g., the diameter. Defining the scaled Planck constant ε and the scaled Debye length λ,

$$\varepsilon^2 = \frac{\hbar^2}{2\, m\, k_B\, T_0\, L^2}, \quad \lambda^2 = \frac{\epsilon\, k_B\, T_0}{q^2\, C_m\, L^2},$$

and introducing the quantum quasi-Fermi level F via $J = n\, \partial_x F$, we can divide equation (1.1b) by n and integrate once. This yields the scaled QDD model stated

on the bounded domain $\Omega = (0,1)$:

$$\partial_x (n\, \partial_x F) = 0, \tag{1.2a}$$

$$-\varepsilon^2 \frac{\partial_{xx}\sqrt{n}}{\sqrt{n}} + \log(n) + V = F, \tag{1.2b}$$

$$-\lambda^2 \partial_{xx} V = n - C_{dop}. \tag{1.2c}$$

Throughout the paper we assume that $C_{dop} \in L^\infty(\Omega)$. In (1.2) the electron density $n = n(x) \geq 0$, the quantum quasi-Fermi level $F = F(x)$, and the electrostatic potential $V = V(x)$ are unknown.

The model equations (1.2) are supplemented with Dirichlet boundary conditions modeling the Ohmic contacts of the device:

$$n = n_D > 0, \quad V = V_D \overset{\text{def}}{=} V_{eq} + V_{ext}, \quad F = F_D \overset{\text{def}}{=} F_{eq} + V_{ext} \quad \text{on } \partial\Omega, \tag{1.3}$$

where V_{ext} is the applied biasing voltage. This set of boundary conditions is motivated by its analogy to the classical DD model [30, 27, 28]. Nevertheless, the correct choice of the Dirichlet data is still an open problem. A recent discussion can be found in [6]. Clearly, the thermal equilibrium density n_{eq} is a possible candidate for n_D. The built-in potential is given by V_{eq}, and F_{eq} is chosen accordingly.

Remark 1. The restriction to the unipolar case is just to keep the notation simple. In fact, Ben Abdallah and Unterreiter [1] proved existence of solutions and considered the semiclassical limit for the bipolar case. The results of this paper are easily extendable to the bipolar setting.

So far, only standard discretization schemes were employed, which require very fine meshes to ensure an adequate resolution of the desired quantities. To account for this problems we want to generalize the classical Scharfetter–Gummel (SG) discretization for the DD equations [35] to this quantum model. A first step in this direction can be found in [9, 4] where a nonlinear discretization scheme is suggested. Here, we follow a different approach [33] since we are moreover interested in a scheme which is stable in the semiclassical limit $\varepsilon \to 0$ recovering the classical SG scheme. However, the SG method relies on the introduction of the so-called Slotboom variable which allows for the symmetrization of the continuity equation [29]. This is impossible in the formulation (1.2) of the QDD model, since we have the additional quantum Bohm potential. Nevertheless, we can deal with this problem by interpreting the Bohm potential as a correction of the classical electrostatic potential V and introducing the *corrected potential* G via

$$G = -\varepsilon^2 \frac{\partial_{xx}\sqrt{n}}{\sqrt{n}} + V$$

which yields for the current density $J = \partial_x n + n\partial_x G$; i.e., the drift is now given by G. Then, system (1.2) can be written as

$$\partial_x J = 0, \quad J = \partial_x n + n\partial_x G, \tag{1.4a}$$

$$-\varepsilon^2 \frac{\partial_{xx}\sqrt{n}}{\sqrt{n}} + V = G, \tag{1.4b}$$

$$-\lambda^2 \partial_{xx} V = n - C_{dop}. \tag{1.4c}$$

We introduce the generalized Slotboom variable $u = e^G n$, which yields for the current density $J = e^{-G}\partial_x u$. Assuming vanishing quantum effects and vanishing quantum

current at the boundary, we get

$$G = V_D, \quad u = e^{V_D} n_D \quad \text{on } \partial\Omega. \tag{1.5}$$

This nonlinear system is discretized using a mixed finite element discretization for the current density J and the Slotboom variable u, and standard linear elements for n and V. For an overview of stabilized discretization schemes for the classical DD model see [23] and the references therein.

We prove under very mild assumptions that the resulting nonlinear discrete system possesses a solution and that at least a subsequence of the sequence of discrete solutions converges to a continuous solution. Since we have no uniqueness for the QDD model in general, we cannot expect convergence of the full sequence. The proof is based on a variational argument similar to the one used in [1] and the derivation of appropriate a priori bounds. Especially, we can show that a discrete solution fulfills the same maximum principle as a continuous solution.

Our mixed finite element scheme is chosen in such a way that in the case of vanishing quantum effects ($\varepsilon = 0$) one recovers the classical SG discretization of the DD model. By deriving a priori bounds on the discrete solutions which are independent of ε, we can even perform the semiclassical limit $\varepsilon \to 0$ on the discrete level and derive estimates on the convergence rate, which are uniform in ε.

We present simulations of a ballistic diode and a resonant tunneling structure which exactly reproduce the predicted accuracy results underlining the feasibility of our approach. Moreover, these simulations show that the asymptotic constant in the error estimate for the current density seems to be almost independent of the size of the scaled Planck constant ε, which is essential from the engineering point of view, since it allows for an accurate computation of the current density also in the semiclassical limit.

The paper is organized as follows. In section 2 we introduce our nonlinear discrete scheme and section 3 is devoted to the proof of the existence and convergence of discrete solutions. The semiclassical limit is performed in section 4, where also uniform convergence rates are given. Finally, simulations of a ballistic diode and a resonant tunneling structure are presented in section 5. Concluding remarks are given in section 6.

2. A generalized SG discretization. In this section we present a discretization of system (1.4) in the spirit of the well-known SG discretization for the classical DD model [35]. Here the drift is given by the generalized potential G, such that we have to take additional care about (1.4b) which involves the quantum Bohm potential.

First, we write (1.4) in a weak form. For notational convenience we define the spaces

$$X = H_0^1(\Omega), \quad \Sigma = L^2(\Omega),$$

and testing appropriately we get the following:

Find $n \in n_D + X$, $V \in V_D + X$, $G \in G_D + X$, and $J \in \Sigma$ such that

$$\int_\Omega e^G J \cdot \tau \, dx - \int_\Omega \partial_x \left(e^G n\right) \cdot \tau \, dx = 0, \tag{2.1a}$$

$$\int_\Omega J \cdot \partial_x \phi \, dx = 0, \tag{2.1b}$$

$$\varepsilon^2 \int_\Omega \partial_x \sqrt{n} \partial_x \left(\frac{\phi}{\sqrt{n}} \right) \, dx = \int_\Omega (G - V)\phi \, dx, \tag{2.1c}$$

$$\lambda^2 \int_\Omega \partial_x V \partial_x \phi \, dx = \int_\Omega (n - C_{dop})\phi \, dx \tag{2.1d}$$

for all $\phi \in X$ *and* $\tau \in \Sigma$.

We discretize (2.1) on the possibly nonuniform grid $0 = x_0 < x_1 < \cdots < x_N = 1$ defining the subintervals and the grid spacing by

$$I_i = (x_{i-1}, x_i], \quad h_i = x_i - x_{i-1}, \quad h = \max_i h_i.$$

We employ finite-dimensional spaces of linear and constant finite elements:

$$X_h = \left\{ w \in H_0^1(\Omega) : w|_{I_i} \in P_1, i = 1, \ldots, N \right\},$$
$$\Sigma_h = \left\{ w \in L^2(\Omega) : w|_{I_i} \in P_0, i = 1, \ldots, N \right\}.$$

For every function $w \in C^0(\bar{\Omega})$ let w^I denote the linear interpolant verifying $w^I(x_i) = w(x_i)$ for $i = 0, \ldots, N$. We will make frequent use of the following interpolation result [23].

PROPOSITION 2.1. *There exists a constant* $c > 0$, *independent of* h, *such that*

$$\left| w - w^I \right|_{H^1(\Omega)} \le c\, h \left(\sum_i |w|^2_{H^2(I_i)} \right)^{1/2},$$
$$\left| w - w^I \right|_{L^2(\Omega)} \le c\, h^2 \left(\sum_i |w|^2_{H^2(I_i)} \right)^{1/2}$$

for all $w \in H^1(\Omega) \cap \left\{ H^2(I_i), \text{ for all } i = 1, \ldots, N \right\}$.

Let $(\cdot, \cdot)$ denote the standard inner product on $L^2(\Omega)$. We define its discrete analogue using the trapezoidal rule

$$(u, v)_h \stackrel{\text{def}}{=} \int_\Omega (uv)^I \, dx = \sum_{i=0}^N \omega_i \, u(x_i) v(x_i),$$

where $\omega_i > 0$ denotes the corresponding weights of the quadrature formula. We have the following consistency result for the discrete inner product.

LEMMA 2.2. *Let* $f, g \in X_h$. *Then there exists a constant* $c > 0$, *independent of* h, *such that*

$$|(f, g) - (f, g)_h| \le c h^2 \left\| \partial_x f \right\|_{L^2(\Omega)} \left\| \partial_x g \right\|_{L^2(\Omega)}.$$

The corresponding discretization of (2.1) reads as follows:

Find $n_h \in n_D + X_h$, $V_h \in V_D + X_h$, $G_h \in G_D + X_h$, *and* $J_h \in \Sigma_h$ *such that*

$$\left(e^{G_h} J_h, \tau_h \right) - \left(\partial_x \left(e^{G_h} n_h \right)^I, \tau_h \right) = 0, \tag{2.2a}$$

$$(J_h, \partial_x \phi_h) = 0, \tag{2.2b}$$

$$\varepsilon^2 \left(\partial_x (\sqrt{n_h})^I, \partial_x \left(\frac{\phi_h}{\sqrt{n_h}} \right)^I \right) = (G_h - V_h, \phi_h)_h, \tag{2.2c}$$

$$\lambda^2 \left(\partial_x V_h, \partial_x \phi_h \right) = \left(n_h - C_{dop}, \phi_h \right)_h \tag{2.2d}$$

for all $\phi_h \in X_h$ *and* $\tau_h \in \Sigma_h$.

The discretization of the generalized Slotboom variable u is given by $u_h = \left(e^{G_h}\, n_h\right)^I$.

We define the piecewise constant function $\bar{G}_h$ by

$$e^{\bar{G}_h}|_{I_i} = \frac{1}{h_i} \int_{I_i} e^{G_h}\, dx.$$

Using that the discrete current density J_h is constant on each element and the identity

$$G_h = \log\left(\frac{u_h}{n_h}\right)^I = \log(u_h)^I - \log(n_h)^I,$$

we can rewrite (2.2) equivalently as follows:

Find $n_h \in n_D + X_h$, $V_h \in V_D + X_h$, $u_h \in u_D + X_h$, *and* $G_h \in G_D + X_h$ *such that*

$$\left(e^{-\bar{G}_h} \partial_x u_h, \partial_x \phi_h\right) = 0, \tag{2.3a}$$

$$\varepsilon^2 \left(\partial_x (\sqrt{n_h})^I, \partial_x \left(\frac{\phi_h}{\sqrt{n_h}}\right)^I\right) + \left(\log(n_h)^I, \phi_h\right)_h = \left(\log(u_h)^I - V_h, \phi_h\right)_h, \tag{2.3b}$$

$$\lambda^2 \left(\partial_x V_h, \partial_x \phi_h\right) = \left(n_h - C_{dop}, \phi_h\right)_h. \tag{2.3c}$$

Remark 2. Formally, we deduce from (2.2c) that for $\varepsilon = 0$ it holds that $G_h \equiv V_h$ and the mixed finite element scheme reduces to the classical one. Note that in contrast to the classical SG scheme, (2.3a) determines the unknown corrected potential G_h, while (2.3b) is now the one for the electron density n_h.

Remark 3. The *nonlinear discretization scheme* developed in [9, 4] is based on finite differences and differs in the discretization of (2.3b). There, additionally some kind of exponential fitting is used for this equation; i.e., the electron density is approximated by an exponential function on each element. The scheme performs extremely well especially for large grid-spacings, but so far no numerical analysis is available. It is worth noting that an exponential transformation ($n = \exp(w)$) was also employed in the study of the transient problem [27, 28], but a numerical analysis for the fully discrete tranformed system is left for future research.

3. Existence and convergence of discrete solutions. We show that the nonlinear discrete system (2.3) possesses at least one solution and we derive a priori bounds on the sequence of discrete solution which ensure that there exists a subsequence converging to the continuous solution. The existence proof is based on a variational argument, which also allows to derive the desired a priori bounds.

We state the main theorem of this section establishing existence of a discrete solution and its convergence.

THEOREM 3.1. *For each* $h > 0$ *there exists a discrete solution* $(n_h, V_h, G_h, u_h) \in (n_D, V_D, G_D, u_D) + X_h^4$ *of* (2.3). *Further, there exists a subsequence, again denoted by* (n_h, V_h, G_h, u_h), *such that*

$$\left((\sqrt{n_h})^I, V_h, G_h, u_h\right) \to \left(\sqrt{n}, V, G, u\right) \quad \text{in } \left[H^1(\Omega)\right]^4,$$

for $h \to 0$, *where* $(n, V, G, u) \in (n_D, V_D, G_D, u_D) + X^4$ *solves the continuous problem* (2.1).

COROLLARY 3.2. *The sequence of discrete current densities* (J_h) *possesses a subsequence such that* $J_h \to J$ *in* $L^2(\Omega)$ *for* $h \to 0$.

Remark 4. Generally, we cannot expect convergence of the whole sequence, since the continuous as well as the discrete QDD model may admit for multiple solutions. Uniqueness can only be proven near to the thermal equilibrium state, i.e., for small applied biasing voltages V_{ext} [34].

For the existence proof we employ Brouwer's fixed point theorem. We define the closed, bounded, and convex set

$$\mathcal{U}_h \stackrel{\text{def}}{=} \{u_h \in u_D + X_h : \underline{u} \le u_h \le \overline{u}\},$$

where the lower and upper bound are given by

$$\underline{u} \stackrel{\text{def}}{=} \min_{\partial\Omega} e^{V_D} n_D, \quad \overline{u} \stackrel{\text{def}}{=} \max_{\partial\Omega} e^{V_D} n_D.$$

On this set we define the fixed point mapping $T_h : \mathcal{U}_h \to \mathcal{U}_h$, where $u_h = T_h(w_h)$ is calculated via the following iteration:

1. *Find* $(n_h, V_h) \in (n_D, V_D) + X_h^2$ *as the solution of*

(3.1a)
$$\varepsilon^2 \left(\partial_x (\sqrt{n_h})^I, \partial_x \left(\frac{\phi_h}{\sqrt{n_h}} \right)^I \right) + \left(\log(n_h)^I, \phi_h \right)_h = \left(\log(w_h)^I - V_h, \phi_h \right)_h,$$

(3.1b)
$$\lambda^2 \left(\partial_x V_h, \partial_x \phi_h \right) = \left(n_h - C_{dop}, \phi_h \right)_h$$

for all $\phi_h \in X_h$.

2. *Set* $G_h = \log(w_h)^I - \log(n_h)^I$.
3. *Find* $u_h \in u_D + X_h$ *as the solution of*

(3.2)
$$\left(e^{-\bar{G}_h} \partial_x u_h, \partial_x \phi_h \right) = 0$$

for all $\phi_h \in X_h$.

3.1. Well-posedness of the fixed point mapping. The well-posedness of the first step is the content of the following result, which also provides uniform bounds on the electron density n_h and the potential V_h.

LEMMA 3.3. *Let* $w_h \in \mathcal{U}_h$ *be given. Then there exists a unique solution* $(n_h, V_h) \in (n_D, V_D) + X_h^2$ *of the nonlinear system* (3.1). *Further, there exists a constant* $\theta \in (0,1)$, *independent of* h, *such that*

$$\theta \le n_h \le 1/\theta, \quad \left\| (\sqrt{n_h})^I \right\|_{H^1(\Omega)} \le 1/\theta, \quad \|V_h\|_{H^1(\Omega)} \le 1/\theta. \tag{3.3}$$

Proof. For the proof we employ a variational argument following the ideas in [36]. Let $H(s)$ be a primitive of $\log(s)^I$ with $H \ge 0$. We introduce the auxiliary variable $\rho_h \stackrel{\text{def}}{=} \sqrt{n_h}$. On the closed set

$$\mathcal{R}_h \stackrel{\text{def}}{=} \left\{ \rho_h : \rho_h^2 \in n_D + X_h, \rho_h \ge 0 \right\},$$

we define the functional

$$E(\rho) \stackrel{\text{def}}{=} \varepsilon^2 \int_\Omega \left| \partial_x \rho^I \right|^2 \, dx + \sum_{i=0}^N \omega_i H(\rho_i^2) + \frac{\lambda^2}{2} \int_\Omega \left| \partial_x V_h [\rho^2 - C_{dop}] \right|^2 \, dx - \sum_{i=0}^N \omega_i \log(w_i) \, \rho_i^2,$$

where $V_h \stackrel{\text{def}}{=} V_h[\rho^2 - C_{dop}] \in V_D + X_h$ is the unique discrete solution of Poisson's equation $\lambda^2(\partial_x V_h, \partial_x \phi_h) = (\rho^2 - C_{dop}, \phi_h)_h$ for all $\phi_h \in X_h$. Identifying ρ with its vector of nodal values in $\mathbb{R}^{N+1}$ one easily verifies that E possesses a unique minimizer $\rho_h \in \mathcal{R}_h$, since E is bounded from below, continuous, and convex with respect to ρ_h^2 [34]. The minimizer also satisfies the Euler–Lagrange equation

$$\varepsilon^2 \left(\partial_x \rho_h^I, \partial_x \phi_h\right) + \left(\rho_h \log(\rho_h^2)^I, \phi_h\right)_h = \left(\rho_h(\log(w_h)^I - V_h[\rho^2 - C_{dop}]), \phi_h\right)_h \tag{3.4}$$

for all $\phi_h \in X_h$.

Now we derive uniform estimates on the solution. In the following let $\theta \in (0,1)$ denote not necessarily identical constants, which are assumed to be independent of h. Choosing $\rho_D \stackrel{\text{def}}{=} \sqrt{n_D} \in \mathcal{R}_h$ as a comparison function we clearly have $E(\rho_h) \le E(\rho_D)$, from which we deduce

$$\varepsilon^2 \int_\Omega \left|\partial_x \rho_h^I\right|^2 \, dx + \frac{\lambda^2}{2} \int_\Omega \left|\partial_x V_h[\rho_h^2 - C_{dop}]\right|^2 \, dx + \sum_{i=0}^N \omega_i H(\rho_i^2) \le E(\rho_D) + |\log(\overline{u})| \, (\rho_h, \rho_h)_h \, .$$

This yields immediately the existence of a constant $\theta \in (0,1)$, independent of h, such that

$$\left\|\partial_x \rho_h^I\right\|_{L^2(\Omega)} \le 1/\theta \quad \text{and} \quad \|\partial_x V_h\|_{L^2(\Omega)} \le 1/\theta.$$

Employing Sobolev's embedding theorem in one-dimensional space [2], i.e., $H^1(\Omega) \hookrightarrow C^{0,\beta}(\bar{\Omega})$, $\beta \in [0, 1/2)$, we find a uniform constant $\theta \in (0,1)$ with

$$\left\|\rho_h^I\right\|_{C^{0,\beta}(\bar{\Omega})} \le 1/\theta \quad \text{and} \quad \|V_h\|_{C^{0,\beta}(\bar{\Omega})} \le 1/\theta.$$

A direct calculation finally yields $\|n_h\|_{L^\infty(\Omega)} \le 1/\theta$, for some uniform constant $\theta \in (0,1)$.

Next we prove the uniform positivity of ρ_h. Let $[\phi]^-$ denote the linear interpolant of $\phi^- \stackrel{\text{def}}{=} \min(0, \phi)$. Testing (3.4) with $\phi_h = [\rho_h - \underline{\rho}]^-$ for $\underline{\rho} > 0$ yields

$$\varepsilon^2 \left(\partial_x \rho_h^I, \partial_x [\rho_h - \underline{\rho}]^-\right) = \left(\left[-\log(\rho_h^2)^I + \log(w_h)^I - V_h[\rho_h^2 - C_{dop}]\right] \rho_h, [\rho_h - \underline{\rho}]^-\right)_h ,$$

which can be estimated as follows:

$$\varepsilon^2 \left(\partial_x \rho_h^I, \partial_x [\rho_h - \underline{\rho}]^-\right) \le \sum_{i=0}^N \left[\omega_i(-\log(\rho_i^2) + \log(\underline{u}) - 1/\theta)\rho_i(\rho_i - \underline{\rho})^-\right] \le 0$$

if we choose $\underline{\rho}^2 = e^{-1/\theta}\underline{u}$. Further we calculate

$$\left(\partial_x \rho_h^I, \partial_x [\rho_h - \underline{\rho}]^-\right) = \sum_{i=0}^N h_i(\rho_{i+1} - \rho_i)\left[(\rho_{i+1} - \underline{\rho})^- - (\rho_i - \underline{\rho})^-\right] \ge \sum_{i=0}^N h_i \left|(\rho_{i+1} - \underline{\rho})^- - (\rho_i - \underline{\rho})^-\right|^2 .$$

Hence,

$$\sum_{i=0}^{N} h_i \left|(\rho_{i+1}-\underline{\rho})^- - (\rho_i-\underline{\rho})^-\right|^2 \le 0,$$

which implies $(\rho_{i+1}-\underline{\rho})^- = (\rho_i-\underline{\rho})^-, i \in \{0,\dots,N-1\}$, and due to the positivity of ρ_D we have

$$(\rho_i-\underline{\rho})^- \equiv 0 \quad \text{for all } i \in \{0,\dots,N\}.$$

Thus, it holds that, $\rho_h \ge \underline{\rho}$ and $n_h \ge \underline{\rho}^2$, respectively. ∎

An easy consequence of Lemma 3.3 is the following result, which establishes uniform $L^\infty(\Omega)$-bounds on the discrete generalized potential G_h.

COROLLARY 3.4. *Let $w_h \in \mathcal{U}_h$ be given. Then there exist uniform bounds $\underline{G}, \overline{G} > 0$, independent of h, such that*

$$\underline{G} \le G_h \le \overline{G}. \tag{3.5}$$

Further, Corollary 3.4 and standard results from elliptic theory [16, 24] yield the unique solvability of (3.2).

LEMMA 3.5. *Let $G_h \in G_D + X_h$ be given with $G_h > \underline{G}$ uniformly in h. Then there exists a unique solution $u_h \in \mathcal{U}_h$ of (3.2). Further, there exists a constant $\theta \in (0,1)$, independent of h, such that*

$$\|u_h\|_{H^1(\Omega)} \le 1/\theta. \tag{3.6}$$

3.2. Proof of the existence theorem. The results derived so far ensure the well-posedness of the fixed point mapping and we are now in the position to prove the convergence theorem.

Proof of Theorem 3.1. First we note that the fixed point mapping T_h is well defined and continuous due to Lemma 3.3, Corollary 3.4, and Lemma 3.5.

Identifying u_h with its vector of nodal values in $\mathbb{R}^{N+1}$ we deduce the existence of a fixed point $u_h \in \mathcal{U}_h$ from Brouwer's fixed point theorem [38], since $\mathcal{U}_h$ is a closed, convex and compact subset of $\mathbb{R}^{N+1}$.

The uniform bounds given in Lemma 3.3, Corollary 3.4, and Lemma 3.5 imply the existence of a subsequence $(\rho_{h_k}, V_{h_k}, G_{h_k}, u_{h_k})$ such that

$$\left(\rho^I_{h_k}, V_{h_k}, G_{h_k}, u_{h_k}\right) \rightharpoonup (\rho, V, G, u) \quad \text{weakly in } \left[H^1(\Omega)\right]^4,$$

for $h_k \to 0$, where $(\rho^2, V, G, u) \in (n_D, V_D, G_D, u_D) + X_h^2$.

These convergences are by far sufficient to pass to the limit in (2.3): due to Sobolev's embedding theorem we have $\rho^I_{h_k} \to \rho$ and also $\rho_{h_k} \to \rho$ in $L^\infty(\Omega)$, such that we can deduce the strong $H^1(\Omega)$ convergence of $(\rho^I_{h_k})$ from (3.4). Standard results from elliptic theory [24] yield

$$\begin{aligned} V_{h_k} &\to V \quad \text{in } H^1(\Omega),\\ u_{h_k} &\to u \quad \text{in } H^1(\Omega), \end{aligned}$$

and finally $G_{h_k} \to G$ in $H^1(\Omega)$ for $h_k \to 0$.

Hence, (ρ^2, V, G, u) is in fact a solution of (2.3), which ends the proof. ∎

3.3. Convergence rates. For completeness we also state a result establishing convergence rates for the finite element discretization (2.2). Since we consider a fully nonlinear system of equations which may admit multiple solutions, we have to impose an additional assumption on the isolatedness of the continuous solution.

THEOREM 3.6. *Let $(n, V, G, u) \in \left[H^2(\Omega)\right]^4$ be a solution of the continuous problem and assume that the Fréchet derivative $(I-DT)(u) \in \mathcal{L}(H^1(\Omega), H^1(\Omega))$ of $I-T : H^1(\Omega) \to H^1(\Omega)$ at u is boundedly invertible. Then there exists a constant $h_0 > 0$ such that for $h < h_0$ there exists a solution (n_h, V_h, G_h, u_h) of the discrete problem (2.3). Further, there exists a constant $c > 0$, independent of h, such that*

$$\|n - n_h\|_{H^1(\Omega)} + \|V - V_h\|_{H^1(\Omega)} + \|u - u_h\|_{H^1(\Omega)} \leq c\, h. \tag{3.7}$$

The proof can be found in [32]. Note that here the constant c generally depends on ε, such that the performance of the semiclassical limit is not possible on this level. In the next section we will use different techniques to overcome this problem.

4. Semiclassical limit—Uniform convergence. In this section we provide estimates independent of the parameter ε, which allow to perform the semiclassical limit in the numerical scheme recovering the classical SG discretization.

We need estimates on the discrete solution, which are independent of ε and h, generalizing the estimates given in section 3. Examining carefully the proofs of that section, we conclude that they crucially depend on ε, since we exploited the monotonicity of the quantum Bohm potential. But in fact we can introduce a different fixed point mapping. The key idea is to reinterpret the equation for n as the one for G and vice versa. This yields the fixed point map $N : n_D + X \to n_D + X$ with $N(m) = n$, where given $m \in n_D + X$ the solution n is calculated via the following iteration:

1. *Find $V \in V_D + X$ such that*

$$\lambda^2 \int_\Omega \partial_x V\, \partial_x \phi\, dx = \int_\Omega (m - C_{dop})\phi\, dx \tag{4.1a}$$

for all $\phi \in X$.

2. *Find $(n, G) \in (n_D, G_D) + X^2$ such that*

$$-\int_\Omega n\, \partial_x G\, \partial_x \phi\, dx - \int_\Omega \partial_x n\, \partial_x \phi\, dx = 0, \tag{4.1b}$$

$$\varepsilon^2 \int_\Omega \partial_x \sqrt{n}\, \partial_x \left(\frac{\phi}{\sqrt{n}}\right) dx = \int_\Omega (G - V)\phi\, dx \tag{4.1c}$$

for all $\phi \in X$.

Remark 5. The reader easily verifies that the fixed point mapping N is well defined and possesses a fixed point exploiting the relation $G = F - \log(n)$. Hence, (4.1) is just a reformulation of (1.2).

Also the discrete system (2.2) can be reformulated in the former manner:

Find $n_h \in n_D + X_h$, $V_h \in V_D + X_h$, and $G_h \in G_D + X_h$ such that,

$$-\int_\Omega \partial_x n_h\, \partial_x \phi_h\, dx - \int_\Omega n_h\, \partial_x G_h \partial_x \phi_h\, dx = \langle f_h, \phi_h \rangle\,, \tag{4.2a}$$

$$\varepsilon^2 \int_\Omega \partial_x (\sqrt{n_h})^I\, \partial_x \left(\frac{\phi_h}{\sqrt{n_h}}\right)^I dx = (G_h - V_h, \phi_h)_h, \tag{4.2b}$$

$$\lambda^2 \int_\Omega \partial_x V_h\, \partial_x \phi_h\, dx = (n_h - C_{dop}, \phi_h)_h \tag{4.2c}$$

for all $\phi_h \in X_h$.

Here, the right-hand side f_h is given by

$$\langle f_h, \phi_h \rangle \stackrel{\text{def}}{=} \int_\Omega e^{-G_h} \partial_x \left(e^{G_h} n_h \right) \partial_x \phi_h \, dx - \int_\Omega e^{-\bar{G}_h} \partial_x \left(e^{G_h} n_h \right)^I \partial_x \phi_h \, dx.$$

The main result of this section reads as follows.

THEOREM 4.1. *Let* $(n, V, G, u) \in \left[H^2(\Omega)\right]^4$ *be a solution of the continuous problem* (1.2) *and assume that the Fréchet derivative* $(I - DN)(n) \in \mathcal{L}(H^1(\Omega), H^1(\Omega))$ *of the mapping* $I - N : H^1(\Omega) \to H^1(\Omega)$ *at a solution* n *is uniformly bounded invertible, i.e.,*

$$\left\| (I - DN)^{-1} \right\|_{\mathcal{L}(H^1(\Omega), H^1(\Omega))} \leq M,$$

where $M > 0$ *is independent of* ε.

Then there exists a constant $c > 0$ *such that for each* $0 < \varepsilon < \varepsilon_0$ *with* $\varepsilon_0 = \varepsilon_0(h) > 0$ *we have the uniform estimates*

$$\|n - n_h\|_{H^1(\Omega)} \leq c\, h, \tag{4.3a}$$

$$\|V - V_h\|_{H^1(\Omega)} \leq c\, h, \tag{4.3b}$$

$$\|G - G_h\|_{H^1(\Omega)} \leq c\, h, \tag{4.3c}$$

$$\|u - u_h\|_{H^1(\Omega)} \leq c\, h, \tag{4.3d}$$

$$\|J - J_h\|_{L^2(\Omega)} \leq c\, h. \tag{4.3e}$$

The proof of Theorem 4.1 is done in several steps.

4.1. Regularity of weak solutions. First, we show that a weak solution of system (2.1) is in fact in $\left[H^2(\Omega)\right]^4$ and that the stronger norm is also uniformly bounded in ε. This generalizes the results given in [1].

THEOREM 4.2. *Let* (n, V, G, J) *be a weak solution of* (2.1). *Then it holds that* $(n, V, G) \in \left[H^2(\Omega)\right]^3$ *and there exists a constant* $K > 0$, *independent of* ε, *such that*

$$\|n\|_{H^2(\Omega)} + \|V\|_{H^2(\Omega)} + \|G\|_{H^2(\Omega)} \leq K.$$

Proof. We eliminate G and J in (1.4) and introduce the auxiliary $\rho = \sqrt{n}$, which yields the fourth-order equation

$$\varepsilon^2 \rho_{xxxx} - \varepsilon^2 \frac{\rho_{xx}^2}{\rho} - 2\rho_{xx} - 2\frac{\rho_x^2}{\rho} - 2\rho_x V_x - \rho V_{xx} = 0$$

supplemented with boundary conditions

$$\rho = \rho_D, \quad \rho_{xx} = 0 \quad \text{ on } \partial\Omega.$$

This equation possesses a unique weak solution $\rho \in H^4(\Omega)$ (see [27]). Testing the fourth-order equation with $\phi = -\rho_{xx}$ we get

$$\varepsilon^2 \int_\Omega \rho_{xxx}^2 \, dx + \varepsilon^2 \int_\Omega \frac{\rho_{xx}^3}{\rho} \, dx + 2 \int_\Omega \rho_{xx}^2 \, dx + 2 \int_\Omega \frac{\rho_x^2}{\rho} \rho_{xx} \, dx + 2 \int_\Omega \rho_x V_x \rho_{xx} \, dx + \int_\Omega \rho V_{xx} \rho_{xx} \, dx = 0.$$

From the Gagliardo–Nierenberg inequality [24] we derive

$$\|\rho_x\|_{L^4(\Omega)} \le c_1 \|\rho\|^{1/4}_{H^2(\Omega)} \|\rho\|^{3/4}_{H^1(\Omega)},$$

with $c_1 = c_1(\Omega) > 0$, which yields, using Poincaré's inequality,

$$\begin{aligned} 2\int_\Omega \frac{\rho_x^2}{\rho}\rho_{xx} &\le \frac{2}{\underline{\rho}} \|\rho_x\|^2_{L^4(\Omega)} \|\rho_{xx}\|_{L^2(\Omega)} \\ &\le 2\frac{c_2(\Omega)}{\underline{\rho}} \|\rho\|^{3/2}_{H^2(\Omega)} \|\rho\|^{3/2}_{H^1(\Omega)} \\ &\le \frac{1}{2}\|\rho\|^2_{H^2(\Omega)} + 2\frac{c_2^2}{\underline{\rho}^2}\|\rho\|^6_{H^1(\Omega)}. \end{aligned}$$

Note that the upper and lower bounds $\underline{\rho} \le \rho \le \overline{\rho}$ are also uniform in ε (see [1]). Further, the Gagliardo–Nierenberg inequality gives

$$\|\rho_{xx}\|_{L^3(\Omega)} \le c_3(\Omega) \|\rho\|^{7/12}_{H^3(\Omega)} \|\rho\|^{5/12}_{H^1(\Omega)},$$

which yields

$$\begin{aligned} \varepsilon^2 \int_\Omega \frac{\rho_{xx}^3}{\rho} &\le \frac{\varepsilon^2}{\underline{\rho}} \|\rho_{xx}\|^3_{L^3(\Omega)} \\ &\le \frac{\varepsilon^2 c_3}{\underline{\rho}} \|\rho\|^{7/4}_{H^3(\Omega)} \|\rho\|^{5/4}_{H^1(\Omega)} \\ &\le \frac{\varepsilon^2}{2}\|\rho\|^2_{H^3(\Omega)} + \frac{\varepsilon^2 c_3^2}{2\underline{\rho}}\|\rho\|^{10}_{H^1(\Omega)}. \end{aligned}$$

Finally, we derive from Sobolev's embedding theorem and standard regularity results [24]

$$\begin{aligned} 2\int_\Omega \rho_x V_x \rho_{xx} &\le 2\|\rho_x\|_{L^2(\Omega)} \|\rho_{xx}\|_{L^2(\Omega)} \|V_x\|_{L^\infty(\Omega)} \\ &\le \frac{1}{4}\|\rho_{xx}\|^2_{L^2(\Omega)} + c_4 \|\rho\|^2_{H^1(\Omega)} \|V\|^2_{H^2(\Omega)} \\ &\le \frac{1}{4}\|\rho_{xx}\|^2_{L^2(\Omega)} + c_4 \|\rho\|^4_{H^1(\Omega)} \end{aligned}$$

and

$$\begin{aligned} \int_\Omega \rho V_{xx}\rho_{xx} &= -\frac{1}{\lambda^2}\int_\Omega \rho\rho_{xx}(\rho^2 - C_{dop}) \\ &\le \frac{1}{\lambda^2}\|\rho\|_{L^\infty(\Omega)} \left(\|\rho\|^2_{L^\infty(\Omega)} + \|C_{dop}\|_{L^\infty(\Omega)}\right) \|\rho_{xx}\|_{L^2(\Omega)} \\ &\le \frac{1}{8}\|\rho_{xx}\|^2_{L^2(\Omega)} + c_5(\lambda, \overline{\rho}, C_{dop}). \end{aligned}$$

Combining these estimates and using Poincaré's inequality we get

$$\frac{\varepsilon^2}{2}\int_\Omega \rho_{xxx}^2 + \frac{1}{4}\int_\Omega \rho_{xx}^2 \le c_6(\Omega, \lambda, \underline{\rho}, \|\rho\|_{H^1(\Omega)}),$$

where c_6 is independent of ε.

Hence, we established $\|\rho\|_{H^2(\Omega)} \leq \tilde{K}$ for some $\tilde{K} > 0$ uniformly in ε and due to the uniform upper and lower bounds, it even holds that $\|n\|_{H^2(\Omega)} \leq K$. The uniform boundedness of $\|V\|_{H^2(\Omega)}$ and $\|G\|_{H^2(\Omega)}$ follows now from the standard elliptic theory. ∎

4.2. Uniform bounds on the discrete solution. Secondly, we derive uniform bounds for the discrete solution of system (2.2).

LEMMA 4.3. *There exist constants $K > 0$ and $\theta \in (0,1)$, independent of ε and h, such that*

$$\left\|(\sqrt{n_h})^I\right\|_{H^1(\Omega)} + \|V_h\|_{H^1(\Omega)} + \|u_h\|_{H^1(\Omega)} + \|G_h\|_{H^1(\Omega)} \leq K,$$
$$\theta \leq n_h, u_h \leq 1/\theta.$$

Proof. By construction we have $\underline{u} \leq u_h \leq \overline{u}$. Further, $\inf_{\mathcal{R}_h} E$ is also uniformly bounded in ε, such that each term of $E(\rho_h)$ is uniformly bounded. This implies $\|V_h\|_{H^1(\Omega)} \leq K$, where the constant $K > 0$ is independent of ε and h. Mimicking the proof of Lemma 3.3 we get $\theta \leq n_h$, where $\theta \in (0,1)$ is independent of ε and h. This yields $G_h \leq 1/\theta$, which in turn implies $\|u_h\|_{H^1(\Omega)} \leq K$.

Now let $\xi_h \in \mathcal{R}_h$ be the unique minimizer of the classical energy functional

$$E_{class}(\xi) = \int_\Omega H(\xi^2)\, dx + \frac{\lambda^2}{2} \int_\Omega \left|\partial_x V_h[\xi^2 - C_{dop}]\right|^2\, dx - \int_\Omega \log(u_h)^I\, \xi^2\, dx.$$

For $\rho_h = \sqrt{n_h}$ it holds that $E(\rho_h) \leq E(\xi_h)$ and $E_{class}(\xi_h) \leq E_{class}(\rho_h)$, which implies

$$\int_\Omega \left|\partial_x \rho_h^I\right|^2\, dx \leq \int_\Omega \left|\partial_x \xi_h^I\right|^2\, dx.$$

In fact, we can calculate ξ_h explicitly from $\xi_h^2 = (u_h\, e^{-V_h})^I$. The bounds derived so far ensure the uniform boundedness of ξ_h in $H^1(\Omega)$ with respect to ε and h. Hence, we finally get $\left\|(\sqrt{n_h})^I\right\|_{H^1(\Omega)} \leq K$ and $n_h \leq 1/\theta$ as well as $-1/\theta \leq G_h$. From $G_h = \log(u_h)^I - \log(n_h)^I$ we deduce that $\|G_h\|_{H^1(\Omega)} \leq K$ by a direct calculation. ∎

4.3. Consistency of the discrete fixed point operator. Third, we introduce some auxiliary problems, which allow to derive the consistency of the two steps of the fixed point mapping N. Let $\hat{n} \overset{\text{def}}{=} N(n_h) \in n_D + X$ and define $\hat{n}_h \in n_D + X$ as the solution of

$$-\int_\Omega \partial_x \hat{n}_h\, \partial_x \phi\, dx - \int_\Omega \hat{n}_h\, \partial_x G_h\, \partial_x \phi\, dx = \langle f_h, \phi \rangle$$

for all $\phi \in X$. The functions $\hat{V}$ and $\hat{G}$ as well as $\hat{V}_h$ and $\hat{G}_h$ are defined in analogy.

Standard results for finite element approximations of linear elliptic equations directly yield the following result [16].

LEMMA 4.4. *Let $(\hat{n}_h, \hat{V}_h)$ be defined as above and (n_h, V_h) is a discrete solution of (4.2). Then there exists a constant $c > 0$, independent of ε and h, such that*

$$\|\hat{n}_h - n_h\|_{H^1(\Omega)} + \left\|\hat{V}_h - V_h\right\|_{H^1(\Omega)} \leq ch.$$

Further, we introduce the auxiliary variables $\hat{u} \stackrel{\text{def}}{=} e^{\hat{G}}\hat{n} \in u_D + X$ and $\hat{u}_h \stackrel{\text{def}}{=} e^{G_h}\hat{n}_h \in u_D + X$, which fulfill

$$-\int_\Omega e^{-\hat{G}}\partial_x\hat{u}\,\partial_x\phi\,dx = 0 \quad \text{and} \quad -\int_\Omega e^{-G_h}\partial_x\hat{u}_h\,\partial_x\phi\,dx = \langle f_h, \phi\rangle$$

for all $\phi \in X$.

Taking the difference of these two equations and testing with $\phi = \hat{u} - \hat{u}_h \in X$ yield

$$e^{-\hat{G}}\,\|\partial_x(\hat{u}-\hat{u}_h)\|_{L^2(\Omega)} \le \left\|e^{-\hat{G}} - e^{-G_h}\right\|_{L^\infty(\Omega)} \|\partial_x\hat{u}_h\|_{L^2(\Omega)} + \|f_h\|_{L^2(\Omega)}\,,$$

from which we deduce

$$\|\hat{u}-\hat{u}_h\|_{H^1(\Omega)} \le c\left\{\left\|\hat{G}-G_h\right\|_{H^1(\Omega)} + h\right\}. \tag{4.4}$$

Further, we have by a direct calculation

$$\|\partial_x(\hat{n}-\hat{n}_h)\|_{L^2(\Omega)} \le c\left(\|\partial_x(\hat{u}-\hat{u}_h)\|_{L^2(\Omega)} + \left\|\hat{G}-G_h\right\|_{H^1(\Omega)}\right) \tag{4.5}$$

for some constant $c > 0$.

We need the following consistency result, which can be easily derived by cumbersome calculations and thus are omitted here for the sake of a compact presentation.

LEMMA 4.5. *There exists a constant $c > 0$, independent of ε and h, such that*

$$\sup_{\|\phi\|_{H^1(\Omega)}=1} |\langle f_h, \phi\rangle| \le ch.$$

Next, we prove the key estimate, which will finally allow for the derivation of the uniform convergence rates.

LEMMA 4.6. *Let $\hat{G}$ be defined as above and G_h a solution of (4.2). Then there exists a constant $c = c(\hat{n}, \hat{V}, \hat{G}) > 0$, independent of ε and h, such that*

$$\left\|G_h - \hat{G}^I\right\|_{L^2(\Omega)} \le c\left(h^2 + h^{-1}\varepsilon^2\,\|\hat{n}-n_h\|_{H^1(\Omega)} + \varepsilon^2\right).$$

Proof. We define

$$\langle A(n),\phi\rangle \stackrel{\text{def}}{=} \varepsilon^2\int_\Omega \partial_x(\sqrt{n})\,\partial_x\left(\frac{\phi}{\sqrt{n}}\right)\,dx,$$
$$\langle A_h(n_h),\phi\rangle \stackrel{\text{def}}{=} \varepsilon^2\int_\Omega \partial_x(\sqrt{n_h})^I\,\partial_x\left(\frac{\phi}{\sqrt{n_h}}\right)^I\,dx.$$

First, we estimate the difference $G_h - \hat{G}^I$. Due to (4.1) it holds that

$$\langle A_h(n_h),\phi_h\rangle = (G_h - V_h, \phi_h)_h \quad \text{for all } \phi_h \in X_h,$$
$$\langle A(\hat{n}),\phi\rangle = \int_\Omega (\hat{G}-\hat{V})\phi\,dx \quad \text{for all } \phi \in X.$$

Testing the difference of these two equations with $\phi = G_h - \hat{G}^I \in X_h$ yields

$$\begin{aligned}
&\langle A_h(n_h) - A(\hat{n}), \phi\rangle \\
&= \left(G_h - \hat{G} - (V_h - \hat{V}), \phi\right) + (G_h - V_h, \phi) - (G_h - V_h, \phi)_h \\
&= \left(G_h - \hat{G}^I, \phi\right) + \left(\hat{G}^I - \hat{G}, \phi\right) - \left(V_h - \hat{V}, \phi\right) + (G_h - V_h, \phi) - (G_h - V_h, \phi)_h ,
\end{aligned}$$

which implies, due to Lemma 2.2,

$$\begin{aligned}
\left\|G_h - \hat{G}^I\right\|^2_{L^2(\Omega)} \le{}& \left\|\hat{G}^I - \hat{G}\right\|_{L^2(\Omega)} \|\phi\|_{L^2(\Omega)} + \left\|V_h - \hat{V}\right\|_{L^2(\Omega)} \|\phi\|_{L^2(\Omega)} \\
&+ c_1 h^2 \|\partial_x (G_h - V_h)\|_{L^2(\Omega)} \|\partial_x \phi\|_{L^2(\Omega)} + \langle A_h(n_h) - A_h(\hat{n}), \phi\rangle \\
&+ \langle A_h(\hat{n}) - A(\hat{n}), \phi\rangle .
\end{aligned}$$

We estimate termwise. First,

$$\begin{aligned}
|\langle A_h(n_h) - A_h(\hat{n}), \phi\rangle| &= \varepsilon^2 \int_\Omega \partial_x \left(\sqrt{n_h} - \sqrt{\hat{n}}\right)^I \partial_x \left(\frac{\phi}{\sqrt{\hat{n}}}\right)^I dx \\
&\quad + \varepsilon^2 \int_\Omega \partial_x \left(\sqrt{n_h}\right)^I \partial_x \left(\frac{\phi}{\sqrt{n_h}} - \frac{\phi}{\sqrt{\hat{n}}}\right)^I dx \\
&\le \varepsilon^2 \left\|\partial_x \left(\sqrt{n_h} - \sqrt{\hat{n}}\right)^I\right\|_{L^2(\Omega)} \left\|\partial_x \left(\frac{\phi}{\sqrt{\hat{n}}}\right)^I\right\|_{L^2(\Omega)} \\
&\quad + \varepsilon^2 \left\|\partial_x \left(\sqrt{n_h}\right)^I\right\|_{L^2(\Omega)} \left\|\partial_x \left(\frac{\phi}{\sqrt{n_h}} - \frac{\phi}{\sqrt{\hat{n}}}\right)^I\right\|_{L^2(\Omega)}
\end{aligned}$$

which can be estimated using Proposition 2.1 by

$$\begin{aligned}
|\langle A_h(n_h) - A_h(\hat{n}), \phi\rangle| &\le \varepsilon^2 c_2 \left\|\partial_x \left(\sqrt{n_h} - \sqrt{\hat{n}}\right)\right\|_{L^2(\Omega)} \left\|\partial_x \left(\frac{\phi}{\sqrt{\hat{n}}}\right)\right\|_{L^2(\Omega)} \\
&\quad + \varepsilon^2 c_3 \|\partial_x \left(\sqrt{n_h}\right)\|_{L^2(\Omega)} \left\|\partial_x \left(\frac{\phi}{\sqrt{n_h}} - \frac{\phi}{\sqrt{\hat{n}}}\right)\right\|_{L^2(\Omega)}
\end{aligned}$$

and employing the uniform bounds derived so far

$$\begin{aligned}
|\langle A_h(n_h) - A_h(\hat{n}), \phi\rangle| &\le \|n_h - \hat{n}\|_{H^1(\Omega)} \left\{\varepsilon^2 c_4 \|\phi\|_{H^1(\Omega)} + \varepsilon^2 c_5 \|\partial_x \phi\|_{L^2(\Omega)}\right\} \\
&\le \varepsilon^2 c_6 \|n_h - \hat{n}\|_{H^1(\Omega)} \|\phi\|_{H^1(\Omega)}
\end{aligned}$$

for some uniform constants $c_i > 0$, $i = 1, \ldots, 6$.

Second, employing successively Proposition 2.1 we get

$$\begin{aligned}
|\langle A_h(\hat{n}) - A(\hat{n}), \phi\rangle| &= \varepsilon^2 \int_\Omega \partial_x \left(\left(\sqrt{\hat{n}}\right)^I - \sqrt{\hat{n}} \right) \partial_x \left(\frac{\phi}{\sqrt{\hat{n}}} \right) \, dx \\
&\quad + \varepsilon^2 \int_\Omega \partial_x \left(\sqrt{\hat{n}}\right)^I \partial_x \left[\left(\frac{\phi}{\sqrt{\hat{n}}} \right)^I - \left(\frac{\phi}{\sqrt{\hat{n}}} \right) \right] \, dx \\
&\le \varepsilon^2 \left\| \partial_x \left(\left(\sqrt{\hat{n}}\right)^I - \sqrt{\hat{n}} \right) \right\|_{L^2(\Omega)} \left\| \partial_x \left(\frac{\phi}{\sqrt{\hat{n}}} \right) \right\|_{L^2(\Omega)} \\
&\quad + \varepsilon^2 \left\| \partial_x \left(\sqrt{\hat{n}}\right)^I \right\|_{L^2(\Omega)} \left\| \partial_x \left[\left(\frac{\phi}{\sqrt{\hat{n}}} \right)^I - \left(\frac{\phi}{\sqrt{\hat{n}}} \right) \right] \right\|_{L^2(\Omega)} \\
&\le \varepsilon^2 c_7 h \left\| \sqrt{\hat{n}} \right\|_{H^2(\Omega)} \|\phi\|_{H^1(\Omega)} + \varepsilon^2 c_8 h \left\| \frac{\phi}{\sqrt{\hat{n}}} \right\|_{H^2(\Omega)} \\
&\le c_9 \, \varepsilon^2 h \, \|\hat{n}\|_{H^2(\Omega)} \, \|\phi\|_{H^1(\Omega)}
\end{aligned}$$

for some uniform constants $c_i > 0$, $i = 7, 8, 9$.

Combining the estimates derived so far we have

$$\begin{aligned}
\left\| G_h - \hat{G}^I \right\|_{L^2(\Omega)} &\le \varepsilon^2 c_6 h^{-1} \|n_h - \hat{n}\|_{H^1(\Omega)} + c_9 \, \varepsilon^2 \\
&\quad + \left\| \hat{G}^I - \hat{G} \right\|_{L^2(\Omega)} + \left\| V_h - \hat{V} \right\|_{L^2(\Omega)} + c_1 h \, \|\partial_x (G_h - V_h)\|_{L^2(\Omega)} \,,
\end{aligned}$$

where we employed the inverse estimate $\|\phi_h\|_{H^1(\Omega)} \le c h^{-1} \|\phi_h\|_{L^2(\Omega)}$ for all $\phi_h \in X_h$. Finally, we use again Proposition 2.1 to end with

$$\left\| G_h - \hat{G}^I \right\|_{L^2(\Omega)} \le c_{10} \left\{ \varepsilon^2 h^{-1} \|n_h - \hat{n}\|_{H^1(\Omega)} + \varepsilon^2 + h^2 + h \right\}. \qquad \square$$

Next, we estimate the remaining difference $n_h - \hat{n}$.

LEMMA 4.7. *Let $\hat{n}$ be defined as above and n_h a solution of (4.2). Then there exist constants $c = c(\hat{n}, \hat{V}, \hat{G}, \hat{u}) > 0$, independent of ε and h, and $\varepsilon_0 = \varepsilon_0(h) > 0$ such that for $\varepsilon < \varepsilon_0$ it holds that*

$$\|n_h - \hat{n}\|_{H^1(\Omega)} \le ch.$$

Proof. We estimate

$$\|n_h - \hat{n}\|_{H^1(\Omega)} \le \|n_h - \hat{n}_h\|_{H^1(\Omega)} + \|\hat{n}_h - \hat{n}\|_{H^1(\Omega)}$$

and using Lemma 4.4 as well as (4.4) and (4.5)

$$\begin{aligned}
\|n_h - \hat{n}\|_{H^1(\Omega)} &\le c_1 \left\{ \|\hat{u} - \hat{u}_h\|_{H^1(\Omega)} + \left\| G_h - \hat{G} \right\|_{H^1(\Omega)} + h \right\} \\
&\le c_2 \left\{ \left\| G_h - \hat{G} \right\|_{H^1(\Omega)} + h \right\} \\
&\le c_2 \left\{ \left\| G_h - \hat{G}^I \right\|_{H^1(\Omega)} + \left\| \hat{G}^I - \hat{G} \right\|_{H^1(\Omega)} + h \right\}
\end{aligned}$$

and employing Proposition 2.1 and the inverse estimate $\|\phi_h\|_{H^1(\Omega)} \leq ch^{-1} \|\phi_h\|_{L^2(\Omega)}$ for all $\phi_h \in X_h$, we have using Lemma 4.6

$$\begin{aligned}\|n_h - \hat{n}\|_{H^1(\Omega)} &\leq c_3 \left\{ h \left\| \hat{G} \right\|_{H^2(\Omega)} + h^{-1} \left\| G_h - \hat{G}^I \right\|_{L^2(\Omega)} + h \right\} \\ &\leq c_4 \left\{ h + h^{-2}\varepsilon^2 \|n_h - \hat{n}\|_{H^1(\Omega)} + h^{-1}\varepsilon^2 \right\}\end{aligned}$$

for some uniform constants $c_i > 0$, $i = 1, \ldots, 4$. Now we assume

$$\varepsilon^2 \leq \varepsilon_0^2 \stackrel{\text{def}}{=} \frac{h^2}{2c_4},$$

which yields the desired estimate $\|n_h - \hat{n}\|_{H^1(\Omega)} \leq ch$, with $c = 2c_4 + 1$. ∎

4.4. Proof of the uniform convergence result. Now we are in the position to prove the main theorem of this section.

Proof of Theorem 4.1. We have the identity

$$n_h - n + N(n) - N(n_h) = (I - DN(\xi))(n_h - n) = n_h - \hat{n}_h + \hat{n}_h - \hat{n},$$

which yields

$$\|n - n_h\|_{H^1(\Omega)} \leq \left\| (I - DN)^{-1} \right\|_{\mathcal{L}(H^1(\Omega), H^1(\Omega))} \|n_h - \hat{n}\|_{H^1(\Omega)}.$$

Hence, we have due to Lemma 4.7

$$\|n - n_h\|_{H^1(\Omega)} \leq ch$$

for some constant $c > 0$, independent of ε and h. Finally, the other uniform estimates follow from standard results for the approximation of elliptic equations. ∎

5. Numerical results. In this section we present numerical simulations underlining the feasibility of the previously analyzed extended SG discretization. We study a ballistic n^+–n–n^+ diode fabricated of GaAs and a resonant tunneling structure. Both devices consist of a channel and source and drain contact regions, which are assumed to be equally long. The channel is moderately doped with a doping density of $5 \cdot 10^{21}\,\mathrm{m}^{-3}$, while the drain and source are highly doped with $10^{24}\,\mathrm{m}^{-3}$. The resonant tunneling diode has the same underlying structure, but the channel is replaced by a quantum well sandwiched between two barriers. This resonant barrier structure is itself sandwiched between two spacer layers (see Figure 5.1). The physical effect of the barriers is a shift in the Fermi level, which can be modeled by an additional step function B added to the electrostatic potential; i.e., V is replaced by $V + B$. Since, we need more smoothness of B for the numerical analysis, we used instead a smoothed function B, which is depicted together with the doping profile in Figure 5.2. We choose a scaled Debye length of $\lambda^2 = 10^{-2}$ and set the scaled biasing voltage to $V_{ext} = 5$. To emphasize the large gradients occurring in the electron density in the case of the resonant tunneling diode, we show in Figure 5.2 also the computed densities for the different values of the scaled Planck constant ε. These are in fact a consequence of the barrier function and can only be smoothed for large ε. Note that for $\varepsilon^2 = 10^{-5}$ there is already no visible difference to the classical solution.

Remark 6. There is numerical evidence that the QDD model shows negative differential resistance for some resonant tunneling diodes [34, 15], but one has to admit

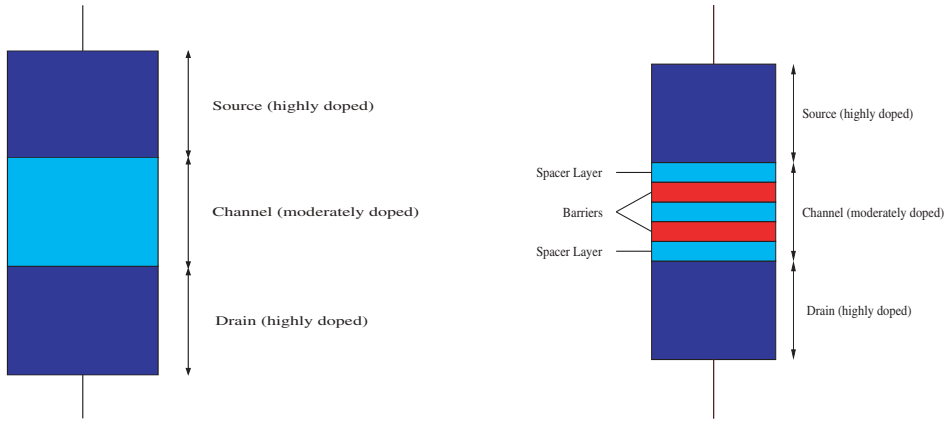


FIG. 5.1. *Diode structure (left: ballistic, right: RTD).*

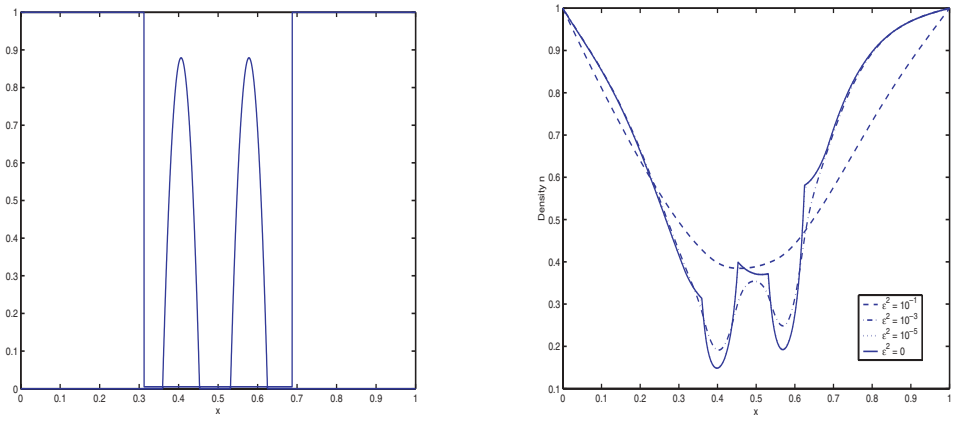


FIG. 5.2. *RTD (left: doping profile, barriers, right: electron densities).*

that the model is far from giving accurate quantitative results for this application. This stems from the fact that the model is mainly designed for scattering dominated transport and no quantum coherence is included. Nevertheless, this device is a good test example due to the large gradients in the electron density near to the barriers.

For the computations we used several uniform grids of variable size and the discretization (2.2). To investigate the semiclassical limit numerically we decreased the scaled squared Planck constant ε^2 from 10^{-1} to 0. The discrete nonlinear system is solved with a damped Newton iteration, which proved to be stable. In Figure 5.3 to Figure 5.5 we present the errors in the electron densities n, the electrostatic potentials V, and the generalized potentials G measured in the $H^1(\Omega)$-seminorm. Further, we depict in Figure 5.6 the error of the current densities J in the $L^2(\Omega)$-norm. The left picture always corresponds to the ballistic diode, while the right one shows the error for the resonant tunneling structure. Since in both cases an analytical solution is not available, we take the solution on the finest grid ($h = 1/4096$) as the "exact" solution.

The numerical results show in both cases that the error behaves like $\mathcal{O}(h)$ for all independent variables. Further, the scheme is stable in the semiclassical limit since the convergence rate is not affected by the size of ε. Note that the error in

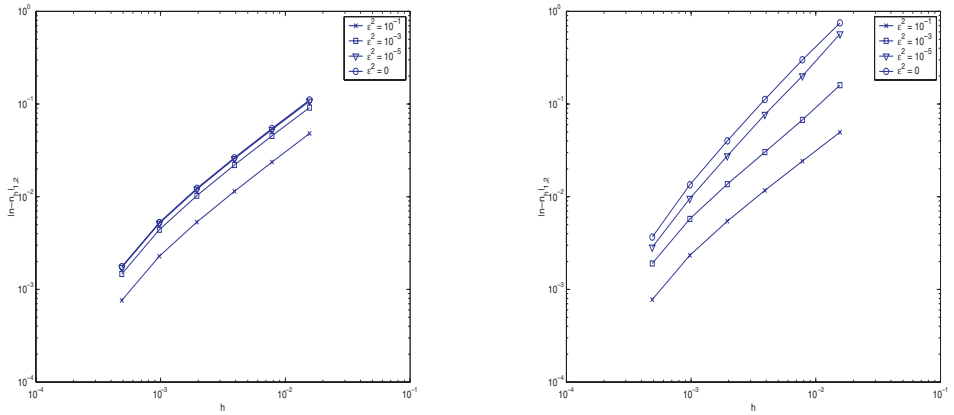


FIG. 5.3. *Error of the electron density (left: ballistic, right: RTD).*

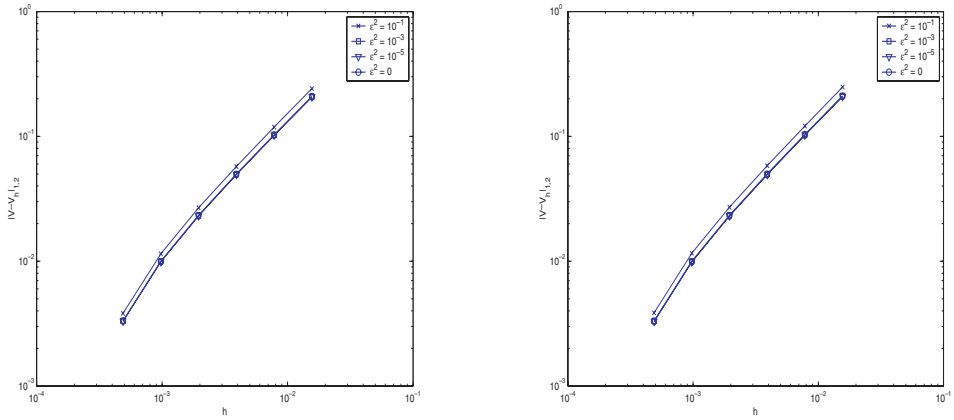


FIG. 5.4. *Error of the potential (left: ballistic, right: RTD).*

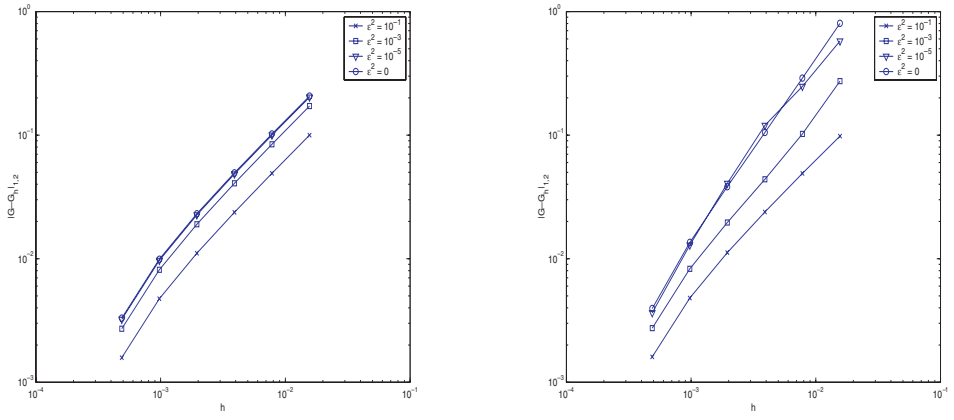


FIG. 5.5. *Error of the generalized potential (left: ballistic, right: RTD).*

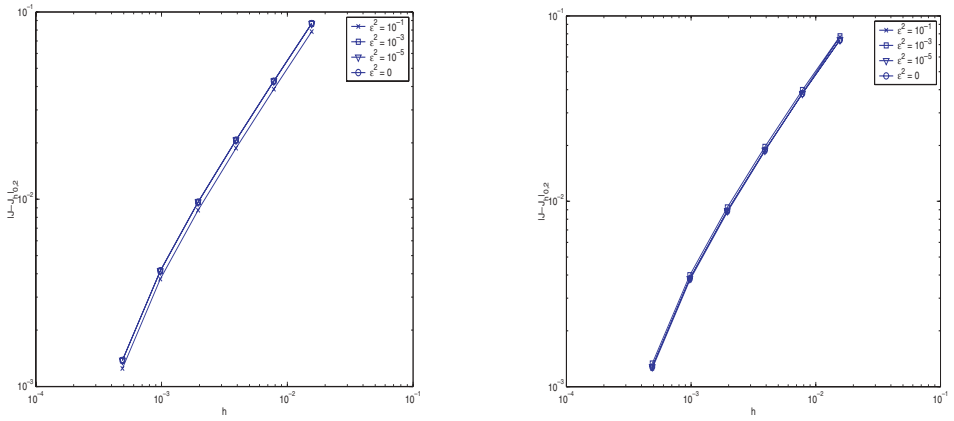


FIG. 5.6. *Error of the current density (left: ballistic, right: RTD).*

the electron density and the generalized potential of the RTD is much more affected by the size of ε, which is a consequence of the additional barrier function B. Most interestingly, in both cases the error in the current density even does not depend on ε. This observation is essential from the engineering point of view, since it allows for an accurate computation of current voltage characteristics also in the semiclassical limit.

6. Conclusions. We presented and analyzed a new stabilized finite element discretization for the quantum drift diffusion model, which is a generalization of the well-known Scharfetter–Gummel discretization for the classical drift diffusion model. The scheme yields the expected approximation errors and allows for the performance of the semiclassical limit on the discrete level, in such a way that the error estimates hold uniformly. The extension to bipolar devices is straightforward. Further, the numerical scheme can be easily extended to space dimensions larger than one using, e.g., the finite element spaces described in [12]. However, the proofs in this paper are not directly extendible, since we employed embedding theorems and inverse estimates, which crucially depend on the space dimension.

REFERENCES

[1] N. Ben ABDALLAH AND A. UNTERREITER, *On the stationary quantum drift diffusion model*, Z. Angew. Math. Phys., 49 (1998), pp. 251–275.
[2] R. A. ADAMS, *Sobolev Spaces*, 1st ed., Academic Press, New York, 1975.
[3] M. ANCONA, *Equations of state for silicon inversion layers*, IEEE Trans. Elect. Devices, 47 (2000), pp. 1449–1456.
[4] M. ANCONA, *Finite-difference schemes for the density-gradient equations*, J. Comp. Elect., 1 (2002), pp. 435–443.
[5] M. ANCONA AND H. TIERSTEN, *Macroscopic physics of the silicon inversion layer*, Phys. Rev., 35 (1987), pp. 7959–7965.
[6] M. ANCONA, D. YERGEAU, Z. YU, AND B. BIEGEL, *On Ohmic boundary conditions for density-gradient theory*, J. Comp. Elect., 1 (2002), pp. 103–107.
[7] M. ANCONA, Z. YU, R. DUTTON, P. VOORDE, M. CAO, AND D. VOOK, *Density-gradient analysis of MOS tunneling*, IEEE Trans. Elect. Devices, 47 (2000), pp. 2310–2318.
[8] M. G. ANCONA, *Diffusion-drift modelling of strong inversion layers*, COMPEL, 6 (1987), pp. 11–18.
[9] M. G. ANCONA AND B. BIEGEL, *Nonlinear discretization scheme for the density-gradient equations*, in Proceedings of the International Conference on Simulation of Semiconductor Pro-

cesses and Devices (SISPAD), Seattle, WA, 2000, pp. 196–199.
[10] M. G. Ancona and G. J. Iafrate, *Quantum correction of the equation of state of an electron gas in a semiconductor*, Phys. Rev. B, 39 (1989), pp. 9536–9540.
[11] A. Asenov, S. Kaya, J. Davies, and S. Saini, *Oxide thickness variation induced threshold voltage fluctuations in decanano MOSFET: A 3d density gradient simulation study*, Superlattices and Microstructures, 28 (2000), pp. 507–515.
[12] F. Brezzi, L. D. Marini, and P. Pietra, *Two-dimensional exponential fitting and applications to drift-diffusion models*, SIAM J. Numer. Anal., 26 (1989), pp. 1342–1355.
[13] F. A. Buot, Y. Ying, and A. I. Fedoseyec, *Quantum hydrodynamic equations and quantum-hierarchy decoupling scheme*, Phys. Rev. E, 66 (2002), 066119.
[14] G. F. Carey, A. L. Pardhanani, and S. W. Bowa, *Advanced numerical methods and software approaches for semiconductor device simulation*, VLSI Design, 10 (2000), pp. 391–414.
[15] P. Caussignac, J. Descloux, and A. Yamnahakki, *Simulation of some quantum models for semiconductors*, Math. Models Methods Appl. Sci., 12 (2002), pp. 1049–1074.
[16] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, 1st ed., North–Holland, Amsterdam, 1978.
[17] P. Degond and C. Ringhofer, *Quantum moment hydrodynamics and the entropy principle*, J. Statist. Phys., 112 (2003), pp. 587–628.
[18] D. Ferry, *Effective potentials and the onset of quantization in ultrasmall MOSFETs*, Superlattices and Microstructures, 28 (2000), pp. 420–423.
[19] D. Ferry and J.-R. Zhou, *Form of the quantum potential for use in hydrodynamic equations for semiconductor device modeling*, Phys. Rev. B, 48 (1993), pp. 7944–7950.
[20] C. L. Gardner, *The quantum hydrodynamic model for semiconductor devices*, SIAM J. Appl. Math., 54 (1994), pp. 409–427.
[21] C. L. Gardner and C. Ringhofer, *Smooth quantum potential for the hydrodynamic model*, Phys. Rev. E, 53 (1996), pp. 157–167.
[22] C. L. Gardner and C. Ringhofer, *Approximation of thermal equilibrium for quantum gases with discontinuous potentials and applications to semiconductor devices*, SIAM J. Appl. Math., 58 (1998), pp. 780–805.
[23] E. Gatti, S. Micheletti, and R. Sacco, *A new Galerkin framework for the drift-diffusion equation in semiconductors*, East-West J. Numer. Math., 6 (1998), pp. 101–135.
[24] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 1st ed., Springer-Verlag, Berlin, 1983.
[25] H. L. Grubin and J. P. Kreskovsky, *Quantum moment balance equations and resonant tunneling structures*, Solid State Electr., 32 (1989), pp. 1071–1075.
[26] A. Jüngel, *Quasi hydrodynamic Semiconductor Equations*, (-) Birkhäuser, Progr. Nonlinear Differential Equations Appl. 41, Basel, 2001.
[27] A. Jüngel and R. Pinnau, *Global nonnegative solutions of a nonlinear fourth-order parabolic equation for quantum systems*, SIAM J. Math. Anal., 32 (2000), pp. 760–777.
[28] A. Jüngel and R. Pinnau, *A positivity-preserving numerical scheme for a nonlinear fourth order parabolic system*, SIAM J. Numer. Anal., 39 (2001), pp. 385–406.
[29] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, 1st ed., Springer-Verlag, Wien, 1990.
[30] R. Pinnau, *A note on boundary conditions for quantum hydrodynamic models*, Appl. Math. Lett., 12 (1999), pp. 77–82.
[31] R. Pinnau, *A review on the quantum drift diffusion model*, Transport Theory Statist. Phys., 31 (2002), pp. 367–395.
[32] R. Pinnau, *Convergence of a generalized Scharfetter–Gummel discretization for the quantum drift diffusion model*, submitted.
[33] R. Pinnau, *A Scharfetter–Gummel type discretization of the quantum drift diffusion model*, Proc. Appl. Math. Mech., 2 (2003), pp. 37–40.
[34] R. Pinnau and A. Unterreiter, *The stationary current-voltage characteristics of the quantum drift-diffusion model*, SIAM J. Numer. Anal., 37 (1999), pp. 211–245.
[35] D. Scharfetter and H. Gummel, *Large signal analysis of a silicon read diode oscillator*, IEEE Trans. Elect. Devices, 15 (1969), pp. 64–77.
[36] A. Unterreiter, *The thermal equilibrium solution of a generic bipolar quantum hydrodynamic model*, Comm. Math. Phys., 188 (1997), pp. 69–88.
[37] A. Wettstein, A. Schenk, and W. Fichtner, *Quantum device-simulation with the density-gradient model on unstructured grids*, IEEE Trans. Elect. Devices, 48 (2001), pp. 279–284.
[38] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, 1st ed., Vol. II/A and II/B, Springer-Verlag, Berlin, 1990.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1669–1693

ON GENERALIZING THE ALGEBRAIC MULTIGRID FRAMEWORK*

ROBERT D. FALGOUT† AND PANAYOT S. VASSILEVSKI†

Abstract. We present a theory for algebraic multigrid (AMG) methods that allows for general smoothing processes and general coarsening approaches. The goal of the theory is to provide guidance in the development of new, more robust, AMG algorithms. In particular, we introduce several compatible relaxation methods and give theoretical justification for their use as tools for measuring the quality of coarse grids.

Key words. algebraic multigrid, compatible relaxation

AMS subject classifications. 65N30, 65F10

DOI. 10.1137/S0036142903429742

1. Introduction.

The algebraic multigrid (AMG) method was originally developed to solve general matrix equations using multigrid principles [6, 17, 3, 18]. The fact that it used only information in the underlying matrix made it attractive as a potential black box solver, a notion that has largely been abandoned. Instead, a wide variety of AMG algorithms have been developed that target different problem classes and have different robustness and efficiency properties.

In recent years, much work has been done to increase the robustness of AMG methods. The classical AMG method of Ruge and Stüben [18] was built upon heuristics based on properties of M-matrices. Although this algorithm works remarkably well for a wide variety of problems [10], the M-matrix assumption still limits its applicability. To address this, a new class of algorithms was developed based on multigrid theory: AMGe (element-based AMG) [7, 14], element-free AMGe [12], and spectral AMGe [9]. All of these algorithms (including Ruge–Stüben AMG) assume a basic framework in their construction: they assume that relaxation is a simple pointwise method, then they build the coarse-grid correction step to eliminate the so-called *algebraically smooth* error left over by the relaxation process. In the AMGe methods, this is done with the help of a *measure* and an associated approximation property that, if satisfied, implies uniform multigrid convergence. The approximation property induces a new heuristic that relates the accuracy of interpolation to the spectrum of the system matrix, namely, that eigenmodes with small associated eigenvalue must be interpolated well.

In this paper, we present a theory that generalizes the AMG framework to address even broader classes of problems. For example, the eddy current formulation of Maxwell's equations (when discretized using the common Nédélec finite elements) has a particularly large (near) null space. In the previous framework, it would be necessary to take all $O(N)$ of the null-space components to the coarse grid, yielding a nonoptimal method. This difficulty can be overcome by using non-pointwise

*Received by the editors June 6, 2003; accepted for publication (in revised form) May 19, 2004; published electronically December 27, 2004. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

http://www.siam.org/journals/sinum/42-4/42974.html

†Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, P.O. Box 808, L-561, Livermore, CA 94551 (rfalgout@llnl.gov, panayot@llnl.gov).

smoothers that damp some of the null space components on the fine grid, leaving a more manageable system to solve on the coarse grid. Examples include overlapping block relaxation [1] and a form of Brandt's distributive relaxation [5, 19] described by Hiptmair in [13].

The theory presented here allows for more general smoothing processes, and changes the above AMGe heuristic in a subtle but important way. It also allows for general coarsening approaches, including vertex-based, cell-based, and agglomeration-based approaches. Yet another aspect of the new theory and framework is *compatible relaxation*, an idea originally proposed by Brandt [4]. We introduce several variants of compatible relaxation and give theoretical justification for its use. The hope is that this work will provide guidance in the development of new AMG methods able to handle difficult problems such as Maxwell's equations.

We assume that the reader is somewhat familiar with AMG research, as numerous comparisons will be made to AMGe and other methods such as smoothed aggregation [20]. In section 2, we introduce two new measures and provide two-level convergence theory. In section 3, we analyze the min-max problem for the new measures. In section 4, we discuss the process of building interpolation, and provide additional theory to support this approach. In section 5, we show how to use compatible relaxation to evaluate the measure and select coarse grids. In section 6, we present two examples illustrating the application of the theory to real problems.

2. New measures and convergence theory. We begin with some notation. Capital italic Roman letters (A, M, P, R) denote matrices and bold lowercase Roman and Greek letters denote vectors $(\mathbf{u}, \mathbf{v}, \boldsymbol{\varepsilon})$. Other lowercase letters denote scalars, while capital calligraphic letters denote sets and spaces $(\mathcal{C}, \mathcal{F}, \mathcal{S})$. We represent the standard Euclidean inner product by $\langle \cdot, \cdot \rangle$ with associated norm $\| \cdot \| := \langle \cdot, \cdot \rangle^{1/2}$. The A-norm (also called the energy norm) is defined by $\| \cdot \|_A := \langle A\cdot, \cdot \rangle^{1/2}$.

Consider solving via AMG the linear system

$$A\mathbf{u} = \mathbf{f}, \tag{2.1}$$

where A is a real symmetric positive definite (SPD) matrix, with $\mathbf{u}, \mathbf{f} \in \mathbb{R}^n$. We consider smoothers (relaxation methods) of the form

$$\mathbf{u}_{k+1} = \mathbf{u}_k + M^{-1}\mathbf{r}_k, \tag{2.2}$$

where $\mathbf{r}_k = \mathbf{f} - A\mathbf{u}_k$ is the residual at the kth iteration. The error propagation for this iteration is given by

$$\mathbf{e}_{k+1} = (I - M^{-1}A)\mathbf{e}_k. \tag{2.3}$$

We also assume that $(M + M^T - A)$ is SPD. It is easy to see that this is a necessary and sufficient condition for convergence (e.g., see the first line in the proof of Theorem 2.2), and hence a reasonable assumption.

Let $P : \mathbb{R}^{n_c} \to \mathbb{R}^n$ be the *interpolation* (or *prolongation*) operator, where $\mathbb{R}^{n_c}$ is a lower-dimensional (*coarse*) vector space, and define $Q : \mathbb{R}^n \to \mathbb{R}^n$ to be a projection onto range (P),

$$Q = PR, \tag{2.4}$$

for some restriction operator $R : \mathbb{R}^n \to \mathbb{R}^{n_c}$ such that $RP = I_c$, the identity on $\mathbb{R}^{n_c}$. Note that R is not the multigrid restriction operator (we will use P^T and the Galerkin

procedure). Also note that the form of R will be important in the remaining sections of the paper.

Define the following measure (we will introduce a second, simpler measure later):

$$\mu(Q,\mathbf{e}) := \frac{\langle M(M+M^T-A)^{-1}M^T(I-Q)\mathbf{e},(I-Q)\mathbf{e}\rangle}{\langle A\mathbf{e},\mathbf{e}\rangle}. \tag{2.5}$$

This measure differs from the AMGe measure in [7] by the inclusion of the term $M(M+M^T-A)^{-1}M^T$ in the numerator. The additional term takes into account the general relaxation process in (2.2). It also provides a natural scaling that eliminates the need to pre-scale A to have diagonal equal one, as in the theory for AMGe.

We now prove that if the measure in (2.5) is bounded by a constant, then two-level multigrid converges uniformly. Furthermore, a smaller measure yields faster convergence. Denote the A-orthogonal projector onto range (P) by

$$Q_A := P(P^TAP)^{-1}P^TA, \tag{2.6}$$

so that $I-Q_A$ represents the error propagation matrix for the coarse-grid correction step. We first prove the following lemma.

LEMMA 2.1. *Let Q be any projection onto* range (P). *Assume that the following approximation property is satisfied for some constant K:*

$$\mu(Q,\mathbf{e}) \le K \quad \forall \mathbf{e}\in\mathbb{R}^n\setminus\{0\}. \tag{2.7}$$

If $\langle A\mathbf{e},\mathbf{v}\rangle = 0\ \forall \mathbf{v}\in$ range (P), then

$$\|(M+M^T-A)^{1/2}M^{-1}A\mathbf{e}\|^2 \ge \frac{1}{K}\langle A\mathbf{e},\mathbf{e}\rangle. \tag{2.8}$$

Proof. Note that range (Q) = range (P), hence

$$\langle A\mathbf{e},Q\mathbf{v}\rangle = 0 \quad \forall \mathbf{v}\in\mathbb{R}^n. \tag{2.9}$$

Assume that (2.7) holds. From (2.9) and the Cauchy–Schwartz inequality, we have

$$\begin{aligned}\langle A\mathbf{e},\mathbf{e}\rangle &= \langle A\mathbf{e},(I-Q)\mathbf{e}\rangle\\ &= \langle (M+M^T-A)^{1/2}M^{-1}A\mathbf{e},(M+M^T-A)^{-1/2}M^T(I-Q)\mathbf{e}\rangle\\ &\le \|(M+M^T-A)^{1/2}M^{-1}A\mathbf{e}\|\|(M+M^T-A)^{-1/2}M^T(I-Q)\mathbf{e}\|\\ &\le \|(M+M^T-A)^{1/2}M^{-1}A\mathbf{e}\|K^{1/2}\langle A\mathbf{e},\mathbf{e}\rangle^{1/2}.\end{aligned}$$

The result (2.8) now follows by dividing through by $\langle A\mathbf{e},\mathbf{e}\rangle K^{1/2}$ and squaring the result. ∎

THEOREM 2.2. *Assume that approximation property* (2.7) *is satisfied for some constant K. Then $K\ge 1$ and*

$$\|(I-M^{-1}A)(I-Q_A)\mathbf{e}\|_A \le \left(1-\frac{1}{K}\right)^{1/2}\|\mathbf{e}\|_A. \tag{2.10}$$

Proof. We have the following identity:

$$\begin{aligned}\|(I-M^{-1}A)\mathbf{e}\|_A^2 &= \langle A\mathbf{e},\mathbf{e}\rangle - \langle A\mathbf{e},M^{-1}A\mathbf{e}\rangle\\ &\quad - \langle M^{-1}A\mathbf{e},A\mathbf{e}\rangle + \langle AM^{-1}A\mathbf{e},M^{-1}A\mathbf{e}\rangle\\ &= \langle A\mathbf{e},\mathbf{e}\rangle - \langle (M+M^T-A)(M^{-1}A)\mathbf{e},(M^{-1}A)\mathbf{e}\rangle.\end{aligned}$$

Replacing $\mathbf{e}$ with $(I - Q_A)\mathbf{e}$ and applying the result in Lemma 2.1 yield

$$\begin{aligned}\|(I - M^{-1}A)(I - Q_A)\mathbf{e}\|_A^2 &\leq \left(1 - \frac{1}{K}\right)\|(I - Q_A)\mathbf{e}\|_A^2 \\ &\leq \left(1 - \frac{1}{K}\right)\|\mathbf{e}\|_A^2 .\end{aligned}$$

To show that $K \geq 1$, note that the identity at the beginning of the proof implies (since norms are nonnegative)

$$\|(M + M^T - A)^{1/2}M^{-1}A\mathbf{e}\|^2 \leq \langle A\mathbf{e}, \mathbf{e}\rangle .$$

The result follows by restricting $\mathbf{e} \neq 0$ to be A-orthogonal to range (P) and applying Lemma 2.1. ∎

The result in Theorem 2.2 is similar to the AMGe result in [7], but applies to more general relaxation methods (than Richardson relaxation). As in AMGe, the bound on the convergence factor approaches 1 as K becomes large, while a smaller K yields a smaller bound on the convergence factor. Note, however, that neither the new measure μ nor the corresponding convergence result reduces to the AMGe measure or convergence result in the case of Richardson relaxation. To complete the connection between the two theories, we now introduce a second, simpler measure,

$$\mu_\sigma(Q, \mathbf{e}) := \frac{\langle \sigma(M)(I - Q)\mathbf{e}, (I - Q)\mathbf{e}\rangle}{\langle A\mathbf{e}, \mathbf{e}\rangle}, \tag{2.11}$$

where $\sigma(M) := \frac{1}{2}(M + M^T)$ is the symmetric part of M. Note that the term $\sigma(M)$ can be replaced equivalently by M, but the symmetric form of this measure is more natural in the theory that follows. The relationship between the measures μ and μ_σ is given in the next lemma.

LEMMA 2.3. *Assume that $(M + M^T - A)$ is SPD. Then,*

$$\mu(Q, \mathbf{e}) \leq \frac{\Delta^2}{2 - \omega}\mu_\sigma(Q, \mathbf{e}), \tag{2.12}$$

where $\Delta \geq 1$ measures the deviation of M from its symmetric part in the sense that

$$\langle M\mathbf{v}, \mathbf{w}\rangle \leq \Delta\langle \sigma(M)\mathbf{v}, \mathbf{v}\rangle^{1/2}\langle \sigma(M)\mathbf{w}, \mathbf{w}\rangle^{1/2}, \tag{2.13}$$

and where

$$0 < \omega := \lambda_{\max}(\sigma(M)^{-1}A) < 2. \tag{2.14}$$

Proof. Note that since $(M + M^T - A)$ is SPD, then both $\sigma(M)$ and $\sigma(M^{-1})$ are also SPD. From (2.13), letting $\mathbf{v} = M^{-1}\mathbf{x}$ and $\mathbf{w} = \sigma(M)^{-1}\mathbf{x}$, we have that

$$\langle \sigma(M)^{-1}\mathbf{x}, \mathbf{x}\rangle^2 \leq \Delta^2\langle \sigma(M)M^{-1}\mathbf{x}, M^{-1}\mathbf{x}\rangle\langle \sigma(M)^{-1}\mathbf{x}, \mathbf{x}\rangle.$$

Dividing both sides by $\langle \sigma(M)^{-1}\mathbf{x}, \mathbf{x}\rangle$ yields

$$\begin{aligned}\langle \sigma(M)^{-1}\mathbf{x}, \mathbf{x}\rangle &\leq \Delta^2\langle MM^{-1}\mathbf{x}, M^{-1}\mathbf{x}\rangle \\ &= \Delta^2\langle M^{-T}\mathbf{x}, \mathbf{x}\rangle \\ &= \Delta^2\langle \sigma(M^{-1})\mathbf{x}, \mathbf{x}\rangle.\end{aligned}$$

From this and (2.14), we then have

$$\begin{aligned}
\mu(Q,\mathbf{e}) &= \frac{\langle M(M+M^T-A)^{-1}M^T(I-Q)\mathbf{e},(I-Q)\mathbf{e}\rangle}{\langle A\mathbf{e},\mathbf{e}\rangle} \\
&\le \max_{\mathbf{x}} \frac{\langle M(M+M^T-A)^{-1}M^T\mathbf{x},\mathbf{x}\rangle}{\langle \sigma(M)\mathbf{x},\mathbf{x}\rangle}\mu_\sigma(Q,\mathbf{e}) \\
&\le \left(\min_{\mathbf{x}} \frac{\langle (M(M+M^T-A)^{-1}M^T)^{-1}\mathbf{x},\mathbf{x}\rangle}{\langle \sigma(M^{-1})\mathbf{x},\mathbf{x}\rangle}\right)^{-1} \Delta^2\mu_\sigma(Q,\mathbf{e}) \\
&= \frac{\Delta^2}{\lambda_{\min}(\sigma(M^{-1})^{-1}(2\sigma(M^{-1})-M^{-T}AM^{-1}))}\mu_\sigma(Q,\mathbf{e}) \\
&= \frac{\Delta^2}{2-\omega}\mu_\sigma(Q,\mathbf{e}). \qquad \square
\end{aligned}$$

Lemma 2.3 provides an obvious corollary to Theorem 2.2 for measure μ_σ. This corollary is the analogue to the AMGe two-level convergence theory in [7]. To see this, note that for a weighted Richardson iteration with weight ω_r, we have that $M^{-1} = \omega_r\|A\|^{-1}I$. If we assume that the AMGe measure is bounded by some constant K_r, then the lemma implies that $\Delta = 1$ and $\omega = \omega_r$, and hence

$$\mu(Q,\mathbf{e}) \le \omega_r^{-1}(2-\omega_r)^{-1}\,\|A\|\,K_r.$$

Applying Theorem 2.2 then yields the AMGe convergence result.

In order for μ_σ to be a useful measure in practice, we need the constants ω and Δ to be "good" constants. In particular, we want both constants to be mesh independent, and we want ω to be bounded away from two. Bounding ω away from two is always possible by using appropriate weighting factors in the relaxation method. In the classical setting, this requirement is equivalent to satisfying a smoothing property; in general, it means that the smoother must damp large eigenmodes of A. Note that this does not preclude the smoother from also damping small eigenmodes (e.g., as required for Maxwell's equations).

To further elaborate on the constants ω and Δ, consider the discrete Laplacian on a uniform grid in one, two, or three dimensions, i.e., the standard 3-pt, 5-pt, and 7-pt operators that arise from finite difference discretizations. First, define $m := \lambda_{\max}(D^{-1}A)$, where D is the diagonal of A. For weighted Jacobi relaxation with weighting factor 2/3, we have that $\omega = (2/3)m$. Since $m \le 2$ for the Laplacian, then $\omega \le 4/3$. For Gauss–Seidel relaxation, let $A = D + L + L^T$, where L is the strictly lower-triangular part of A. Then, $M = D + L$ implies that $\sigma(M) = \frac{1}{2}(D + A)$, and hence

$$\omega = \lambda_{\max}[2(D+A)^{-1}A] = \frac{2}{1+m^{-1}}. \tag{2.15}$$

For the Laplacian, this again implies that $\omega \le 4/3$. We can also use (2.15) to estimate ω in more general settings. For example, in the case of finite elements, one can show that m is not larger than the maximum number of element degrees of freedom. Likewise for any sparse matrix A, one can show that m is not larger than the maximum number of nonzeros per row (column) of A.

The constant Δ is equal to 1 when M is symmetric. As an example of a nonsymmetric M, again, consider Gauss–Seidel. With m equal to the maximum number of

nonzeros in a row (column) of A, and letting $\mathbf{v} = (v_i)$, $\mathbf{w} = (w_i)$, and $A = (a_{ij})$, we have

$$\begin{aligned}
\langle M\mathbf{v}, \mathbf{w}\rangle &\leq \sum_{j\leq i: a_{ij}\neq 0} |a_{ij}|\,|v_j|\,|w_i| \\
&\leq \sum_{j\leq i: a_{ij}\neq 0} \sqrt{a_{jj}}\sqrt{a_{ii}}|v_j|\,|w_i| \\
&\leq \left[\sum_{j\leq i: a_{ij}\neq 0} a_{jj}(v_j)^2\right]^{1/2} \left[\sum_{j\leq i: a_{ij}\neq 0} a_{ii}(w_i)^2\right]^{1/2} \\
&\leq 1/2(m+1)\langle D\mathbf{v}, \mathbf{v}\rangle^{1/2}\langle D\mathbf{w}, \mathbf{w}\rangle^{1/2} \\
&\leq 1/2(m+1)\langle (D+A)\mathbf{v}, \mathbf{v}\rangle^{1/2}\langle (D+A)\mathbf{w}, \mathbf{w}\rangle^{1/2} \\
&= (m+1)\langle \sigma(M)\mathbf{v}, \mathbf{v}\rangle^{1/2}\langle \sigma(M)\mathbf{w}, \mathbf{w}\rangle^{1/2}.
\end{aligned}$$

3. The min-max problem. In this section, we analyze the optimal min-max solution of the measures (2.5) and (2.11), and use the results as a discussion point for relating and comparing the new theory to existing methods such as AMGe, spectral AMGe, and smoothed aggregation [20]. We also introduce generalized notions of the C-pt (coarse point) and F-pt (fine point) terminology used in the classical Ruge–Stüben AMG algorithm. The material in this section serves as a launching pad for the ideas and results in the remainder of the paper.

To analyze the min-max solution of (2.5) and (2.11), we analyze the following base measure:

$$\mu_{_X}(Q, \mathbf{e}) := \frac{\langle X(I-Q)\mathbf{e}, (I-Q)\mathbf{e}\rangle}{\langle A\mathbf{e}, \mathbf{e}\rangle}, \tag{3.1}$$

where, here again, Q has the form $Q = PR$ for some restriction operator $R : \mathbb{R}^n \to \mathbb{R}^{n_c}$ such that $RP = I_c$, and where X represents any given SPD matrix. In the remainder of the paper, it will be important that we fix R so that it does not depend on P (as in spectral AMGe). This operator defines the *coarse-grid variables* [4], $\mathbf{u}_c = R\mathbf{u}$, and specifies, for example, whether they are a subset of the fine-grid variables (vertex-centered), averages of fine-grid variables (cell-centered), or coefficients of fine-grid basis functions (agglomeration, e.g., as in spectral AMGe or smoothed aggregation). The coarse-grid variables, $R\mathbf{u}$, are analogous to C-pts in Ruge–Stüben AMG.

Now, define $S : \mathbb{R}^{n_s} \to \mathbb{R}^n$, where $n_s = n - n_c$, such that $RS = 0$. Think of range(S) as the "smoother space", i.e., the space on which the smoother must be effective. Note that S is not unique (but range(S) is). The variables $S^T\mathbf{u}$ are analogous to F-pts. Note also that S and R^T define an orthogonal decomposition of $\mathbb{R}^n$. That is, any vector $\mathbf{e}$ can be written as $\mathbf{e} = S\mathbf{e}_s + R^T\mathbf{e}_c$ for some $\mathbf{e}_s$ and $\mathbf{e}_c$. We will see in Theorem 3.1 below that the min-max problem of this section also induces an A-orthogonal decomposition of $\mathbb{R}^n$ involving the operator S.

THEOREM 3.1. *Assume we are given a coarse grid Ω_c, and define*

$$\mu_{_X}^\star := \min_P \max_{\mathbf{e}\neq 0} \mu_{_X}(PR, \mathbf{e}). \tag{3.2}$$

The arg min *of* (3.2)*, $P_\star$, satisfies*

$$P_\star^T AS = 0. \tag{3.3}$$

The minimum is given by

$$\mu_X^\star = \frac{1}{\lambda_{\min}((S^T X S)^{-1}(S^T A S))}. \tag{3.4}$$

Proof. Note that since $Q = PR$, $RP = I_c$, and $RS = 0$, we have

$$(I - Q)P = 0; \qquad (I - Q)S = S. \tag{3.5}$$

Also note that $\mathbf{e} - PR\mathbf{e} = (I - Q)\mathbf{e} \in \text{range}\,(S)$ since $R(I - Q) = 0$. Hence $\mathbf{e} = S\mathbf{e}_s + P\mathbf{e}_c$ for some $\mathbf{e}_s$ and $\mathbf{e}_c = R\mathbf{e}$. From (3.2), using (3.5), we then have that

$$\mu_X^\star = \min_P \max_{\mathbf{e}_c, \mathbf{e}_s} \frac{\langle XS\mathbf{e}_s, S\mathbf{e}_s\rangle}{\langle AS\mathbf{e}_s, S\mathbf{e}_s\rangle + 2\langle AS\mathbf{e}_s, P\mathbf{e}_c\rangle + \langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle} \tag{3.6}$$

$$= \min_P \max_{\mathbf{e}_s} \frac{\langle XS\mathbf{e}_s, S\mathbf{e}_s\rangle}{\min_{\mathbf{e}_c}\left(\langle S^T AS\mathbf{e}_s, \mathbf{e}_s\rangle + 2\langle P^T AS\mathbf{e}_s, \mathbf{e}_c\rangle + \langle P^T AP\mathbf{e}_c, \mathbf{e}_c\rangle\right)}. \tag{3.7}$$

The denominator in (3.7) is a quadratic form in the variable $\mathbf{e}_c$ with solution

$$\mathbf{e}_c = -(P^T AP)^{-1} P^T AS\mathbf{e}_s. \tag{3.8}$$

Plugging (3.8) back into (3.7) gives

$$\mu_X^\star = \min_P \max_{\mathbf{e}_s \neq 0} \frac{\langle XS\mathbf{e}_s, S\mathbf{e}_s\rangle}{\langle S^T AS\mathbf{e}_s, \mathbf{e}_s\rangle - \langle (P^T AP)^{-1} P^T AS\mathbf{e}_s, P^T AS\mathbf{e}_s\rangle}. \tag{3.9}$$

Since the second term in the denominator of (3.9) is nonnegative for any $\mathbf{e}_s$, the arg min must satisfy $P_\star^T AS = 0$. Hence,

$$\mu_X^\star = \max_{\mathbf{e}_s \neq 0} \frac{\langle S^T XS\mathbf{e}_s, \mathbf{e}_s\rangle}{\langle S^T AS\mathbf{e}_s, \mathbf{e}_s\rangle} = \frac{1}{\lambda_{\min}((S^T XS)^{-1}(S^T AS))}. \qquad \square$$

Theorem 3.1 is used to motivate the main result in section 4. It will also be used to prove many of the results in sections 4 and 5. Note that $P_\star$ is unique, even though S is not, since R is fixed and $RP = I_c$. An interesting corollary to the theorem is the following.

COROLLARY 3.2. *The optimal $P_\star$ in Theorem* 3.1 *is given by the formula*

$$P_\star = \begin{bmatrix} S & R^T \end{bmatrix} \begin{bmatrix} -(S^T AS)^{-1}(S^T AR^T) \\ I \end{bmatrix} = (I - S(S^T AS)^{-1} S^T A)R^T. \tag{3.10}$$

Proof. This is obtained by solving the equation $S^T AP_\star = 0$. For any $\mathbf{v}$ consider $\mathbf{w} = P_\star \mathbf{v}_c$ and use its decomposition $\mathbf{w} = S\mathbf{w}_s + R^T \mathbf{w}_c$. We have $\mathbf{v}_c = RP_* \mathbf{v}_c = R\mathbf{w} = RR^T \mathbf{w}_c = \mathbf{w}_c$. On the other hand, since $S^T A\mathbf{w} = 0$ one arrives at

$$S^T AS\mathbf{w}_s + S^T AR^T \mathbf{w}_c = 0.$$

That is, $\mathbf{w}_s = -(S^T AS)^{-1} S^T AR^T \mathbf{w}_c = -(S^T AS)^{-1} S^T AR^T \mathbf{v}_c$. Thus

$$P_\star \mathbf{v}_c = (-S(S^T AS)^{-1} S^T A + I)R^T \mathbf{v}_c,$$

which completes the proof. $\square$

Remark 3.3. The first expression in (3.10) can be viewed as a generalization of the optimal interpolation for the AMGe measure (see Corollary 3.4 below). Alternatively, the second expression in (3.10) can be viewed as a kind of smoothed aggregation method. That is, the operator R^T is a type of tentative prolongator, and the term $(I - S(S^TAS)^{-1}S^TA)$ is a type of smoother (because it removes error components in the "smoother space" spanned by S). The interpolation operator in the smoothed aggregation method is formed similarly by smoothing a tentative prolongation operator, except that a simpler, local smoother is used. Another similarity is that the smoothed aggregation smoother is designed to leave unchanged the kernel components in range (R^T) (those kernel components that are representable on the coarse grid). In (3.10), the fact that range (S) is A-orthogonal to range $(P_\star)$ also insures this.

The following two corollaries specialize the results in Theorem 3.1 and Corollary 3.2 to the particular cases of AMGe and spectral AMGe. These results are useful primarily because of the insight and guidance they provide for developing algorithms in these settings.

COROLLARY 3.4. *Assume that P and R are as in AMGe and have the specific forms*

$$P = \begin{bmatrix} W \\ I \end{bmatrix}, \qquad R = \begin{bmatrix} 0 & I \end{bmatrix}, \tag{3.11}$$

where we have reordered the equations so that

$$A = \begin{bmatrix} A_{ff} & A_{fc} \\ A_{cf} & A_{cc} \end{bmatrix}. \tag{3.12}$$

Let $X = \|A\|I$ in (3.1). *Then, the* arg min *and minimum of* (3.2) *are given by*

$$P_\star = \begin{bmatrix} -A_{ff}^{-1}A_{fc} \\ I \end{bmatrix}, \qquad \mu_X^\star = \frac{\|A\|}{\lambda_{\min}(A_{ff})}. \tag{3.13}$$

Proof. Let $S = [I\ 0]^T$. Then $RS = 0$ and $S^TAS = A_{ff}$. The result then follows trivially from (3.10) and (3.4). ∎

COROLLARY 3.5. *Assume that R has the form*

$$R^T = [\mathbf{p}_1, \ldots, \mathbf{p}_c], \tag{3.14}$$

where the $\mathbf{p}_i$, $1 \le i \le n$, are the orthonormal eigenvectors of A with corresponding eigenvalues $\lambda_1 \le \cdots \le \lambda_c \le \cdots \le \lambda_n$. Let $X = \|A\|I$ in (3.1). *Then, the* arg min *and minimum of* (3.2) *are given by*

$$P_\star = R^T, \qquad \mu_X^\star = \frac{\|A\|}{\lambda_{c+1}} = \frac{\lambda_n}{\lambda_{c+1}}. \tag{3.15}$$

Proof. Let $S = [\mathbf{p}_{c+1}, \ldots, \mathbf{p}_n]$. Then $RS = 0$ and $S^TAS = \operatorname{diag}(\lambda_{c+1}, \ldots, \lambda_n)$. The result then follows trivially from (3.10) and (3.4). ∎

Now, consider tailoring the base min-max problem (3.2) to the case of the new measures in (2.5) and (2.11). Assume again that Q has the form $Q = PR$ for some fixed restriction operator $R : \mathbb{R}^n \to \mathbb{R}^{n_c}$ such that $RP = I_c$. As before, define $S : \mathbb{R}^{n_s} \to \mathbb{R}^n$ such that $RS = 0$, and assume we are given a coarse grid Ω_c. Define, based on (2.5) and (2.11),

$$\mu^\star := \min_P \max_{\mathbf{e}\neq 0} \mu(PR, \mathbf{e}), \tag{3.16}$$

$$\mu_\sigma^\star := \min_P \max_{\mathbf{e}\neq 0} \mu_\sigma(PR, \mathbf{e}). \tag{3.17}$$

The quantities $\mu^\star$ and $\mu_\sigma^\star$ measure the ability of the coarse grid to represent algebraically smooth error, where algebraically smooth error is defined to be error components that are not being effectively damped by the more general relaxation process in (2.2). Strictly speaking, this interpretation of $\mu^\star$ and $\mu_\sigma^\star$ assumes that the interpolation operator is the optimal one; i.e., that $P = P_\star$. Hence, given a coarse grid, small quantities indicate that there exists *some* interpolation operator that can interpolate smooth error. Whether or not there exists a practical (e.g., local) interpolation operator is an important research question that is not addressed in this paper. However, empirical evidence so far indicates that $\mu^\star$ and $\mu_\sigma^\star$ are useful measures in practice, particularly for PDE problems.

4. Building interpolation. In the previous section, we defined the quantities $\mu^\star$ and $\mu_\sigma^\star$ as indicators of the ability of the coarse grid to represent smooth error. Assuming that either of these quantities is "small" (we will present an efficient approach for estimating $\mu^\star$ and $\mu_\sigma^\star$ in the next section), we then need to build an interpolation operator. In practice, this means that we must somehow localize the new measure. However, note that the result (3.3) in Theorem 3.1 does not depend on the X in (3.1). This suggests the possibility that, once an adequate coarse grid has been chosen, the procedure for building an interpolation operator can be done without knowledge of the relaxation process. This is quantified in the next lemma and theorem.

LEMMA 4.1. *The following statements are equivalent, where $Q = PR$, P, R, and S are as before, and where $\eta \geq 1$ is some constant:*

$$\langle AQ\mathbf{e}, Q\mathbf{e}\rangle \leq \eta \langle A\mathbf{e}, \mathbf{e}\rangle \quad \forall \mathbf{e}; \tag{4.1}$$

$$\langle A(I-Q)\mathbf{e}, (I-Q)\mathbf{e}\rangle \leq \eta \langle A\mathbf{e}, \mathbf{e}\rangle \quad \forall \mathbf{e}; \tag{4.2}$$

$$\langle AP\mathbf{e}_c, S\mathbf{e}_s\rangle^2 \leq \left(1 - \frac{1}{\eta}\right) \langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle \langle AS\mathbf{e}_s, S\mathbf{e}_s\rangle \quad \forall \mathbf{e}_c, \mathbf{e}_s. \tag{4.3}$$

Proof. We first show that the approximate harmonic property of P, (4.1), implies the strengthened Cauchy–Schwarz inequality (4.3). Letting $\mathbf{e} = tS\mathbf{e}_s + P\mathbf{e}_c$ for any $\mathbf{e}_c$, $\mathbf{e}_s$ and any real t, and noting that $Q\mathbf{e} = P\mathbf{e}_c$, then (4.1) leads to the following quadratic inequality for t:

$$t^2\langle AS\mathbf{e}_s, S\mathbf{e}_s\rangle + 2t\langle AP\mathbf{e}_c, S\mathbf{e}_s\rangle + \left(1 - \frac{1}{\eta}\right)\langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle \geq 0.$$

This implies that the discriminant of the above quadratic form is nonpositive, which is exactly the strengthened Cauchy–Schwarz inequality (4.3). In the same way, we can also show that (4.2) implies (4.3) by noting that $(I-Q)\mathbf{e} = tS\mathbf{e}_s$.

To show that the strengthened Cauchy–Schwarz inequality (4.3) implies the approximate harmonic property (4.1), let $\mathbf{e} = S\mathbf{e}_s + R^T\mathbf{e}_c$ and note that $R(I-Q)\mathbf{e} = 0$. Therefore, there is an $\hat{\mathbf{e}}_s$ such that $(I-Q)\mathbf{e} = S\hat{\mathbf{e}}_s$. That is, $\mathbf{e} = S\hat{\mathbf{e}}_s + P\mathbf{e}_c$, and one has

$$\langle A\mathbf{e}, \mathbf{e}\rangle = \langle AS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle + 2\langle AS\hat{\mathbf{e}}_s, P\mathbf{e}_c\rangle + \langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle.$$

Using (4.3) implies

$$\begin{aligned}
\langle A\mathbf{e}, \mathbf{e}\rangle &\geq \langle AS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle - 2\sqrt{1-\frac{1}{\eta}}\langle AS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle^{1/2}\langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle^{1/2} + \langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle \\
&= \frac{1}{\eta}\langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle + \left[\langle AS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle^{1/2} - \sqrt{1-\frac{1}{\eta}}\langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle^{1/2}\right]^2 \\
&\geq \frac{1}{\eta}\langle AQ\mathbf{e}, Q\mathbf{e}\rangle.
\end{aligned}$$

In the same way, we can also show that (4.3) implies (4.2). ∎

THEOREM 4.2. *Define μ_X and $\mu_X^\star$ as in* (3.1) *and* (3.2) *for any SPD matrix X. Assume that a coarse grid has been chosen such that the following condition holds:*

C1: *$\mu_X^\star \leq K$ for some constant K.*

Assume also that an interpolation operator P has been defined, and satisfies the following additional condition:

C2: (4.1), (4.2), *or* (4.3) *holds for some constant $\eta \geq 1$.*

Then, the following weak approximation property holds:

$$\mu_X(Q,\mathbf{e}) \leq \eta K \quad \forall \mathbf{e} \in \mathbb{R}^n \setminus \{0\}. \tag{4.4}$$

Proof. From Lemma 4.1, we can assume the strengthened Cauchy–Schwarz inequality (4.3). Now, consider the left-hand side of the desired inequality (4.4) and decompose $\mathbf{e} = S\mathbf{e}_s + R^T\mathbf{e}_c$. Note that $R(I-Q)\mathbf{e} = 0$, which implies there is an $\hat{\mathbf{e}}_s$ such that $(I-Q)\mathbf{e} = S\hat{\mathbf{e}}_s$. Hence, using (4.3) and Theorem 3.1, we have

$$\begin{aligned}
\max_{\mathbf{e}} \mu_X(Q,\mathbf{e}) &= \max_{\hat{\mathbf{e}}_s} \max_{\mathbf{e}_c} \frac{\langle XS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle}{\langle A(S\hat{\mathbf{e}}_s + P\mathbf{e}_c), (S\hat{\mathbf{e}}_s + P\mathbf{e}_c)\rangle} \\
&= \max_{\hat{\mathbf{e}}_s} \max_{\mathbf{e}_c} \max_{t\in\mathbb{R}} \frac{\langle XS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle}{\langle A(S\hat{\mathbf{e}}_s + tP\mathbf{e}_c), (S\hat{\mathbf{e}}_s + tP\mathbf{e}_c)\rangle} \\
&= \max_{\hat{\mathbf{e}}_s} \max_{\mathbf{e}_c} \frac{\langle XS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle}{\min_{t\in\mathbb{R}}\langle A(S\hat{\mathbf{e}}_s + tP\mathbf{e}_c), (S\hat{\mathbf{e}}_s + tP\mathbf{e}_c)\rangle} \\
&= \max_{\hat{\mathbf{e}}_s} \max_{\mathbf{e}_c} \frac{\langle XS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle}{\langle AS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle - \frac{\langle AP\mathbf{e}_c, S\hat{\mathbf{e}}_s\rangle^2}{\langle AP\mathbf{e}_c, P\mathbf{e}_c\rangle}} \\
&\leq \max_{\hat{\mathbf{e}}_s} \frac{\langle XS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle}{\langle AS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle - (1-\frac{1}{\eta})\langle AS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle} \\
&= \eta \max_{\hat{\mathbf{e}}_s} \frac{\langle XS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle}{\langle AS\hat{\mathbf{e}}_s, S\hat{\mathbf{e}}_s\rangle} \\
&= \eta\mu_X^\star \\
&\leq \eta K. \qquad \square
\end{aligned}$$

The corollaries to Theorem 4.2 for measures μ and μ_σ separate coarse-grid correction into two distinct parts: **C1** insures the quality of the coarse grid, i.e., its ability to represent algebraically smooth error components; and **C2** insures that these smooth components are adequately interpolated. Hence, once an adequate coarse grid is chosen, it is sufficient to build interpolation based on any one of the three statements in **C2**. In fact, the following result holds.

COROLLARY 4.3. *The statements in* **C2** *are necessary conditions for obtaining a uniformly convergent method.*

Proof. To see this in the case of measure μ_σ, note that our assumption that $(M + M^T - A)$ is SPD implies that $2\langle\sigma(M)\mathbf{e},\mathbf{e}\rangle \geq \langle A\mathbf{e},\mathbf{e}\rangle$. Hence, an approximation property that bounds the measure μ_σ (with constant K_σ) also implies (4.2) (with $\eta = 2K_\sigma$). $\square$

The significance of the above result is that the statements in **C2** nowhere involve the relaxation process. This implies that we can construct interpolation coefficients (again, assuming a coarse grid has already been chosen) using previously developed methods, even those methods that assumed a pointwise smoother. For example, in

AMGe, a local procedure is used for constructing interpolation that produces an approximation property of the form

$$\|A\|\|(I-Q)\mathbf{e}\|^2 \leq \eta \langle A\mathbf{e}, \mathbf{e}\rangle . \tag{4.5}$$

But, it is obvious that this also implies (4.2), which in turn implies the more general result in Theorem 4.2. Note that, even if the constant η is sharp in (4.5), this may be an extremely pessimistic constant for (4.2). See section 6.1 for an example.

5. Compatible relaxation. In this section, we introduce the idea of *compatible relaxation* and show how its convergence rate may be used to estimate the quantities $\mu^\star$ and $\mu^\star_\sigma$ in (3.16) and (3.17). That is, we will show how compatible relaxation may be used to insure **C1** of Theorem 4.2. We will present four variants of compatible relaxation, each having its own advantages and disadvantages, and suggest a simple algorithm for using these techniques to choose coarse grids in AMG methods.

Compatible relaxation, as defined by Brandt [4], is *a modified relaxation scheme that keeps the coarse-level variables invariant.* Consider the following compatible relaxation iteration (represented here by its corresponding error propagation):

$$\mathbf{e}_{k+1} = (I - S(S^TMS)^{-1}S^TA)\mathbf{e}_k, \tag{5.1}$$

where $S : \mathbb{R}^{n_s} \to \mathbb{R}^n$ is defined, as before, in terms of some restriction operator R. Recall that the coarse-grid variables are defined by $\mathbf{u}_c = R\mathbf{u}$. Since $RS = 0$, we see from (5.1) that $R\mathbf{e}_{k+1} = R\mathbf{e}_k$; that is, the coarse-grid variables are invariant under this iteration. Hence, we need only consider compatible relaxation in the complementary space via the following iteration:

$$\mathbf{e}_{k+1} = (I - (S^TMS)^{-1}(S^TAS))\mathbf{e}_k. \tag{5.2}$$

Brandt states that *a general measure for the quality of the set of coarse variables is the convergence rate of compatible relaxation.* In the next theorem, we will make this statement rigorous by relating the convergence of the compatible relaxation process in (5.2) to the measure $\mu^\star$ in (3.16) (equivalently, $\mu^\star_\sigma$ in (3.17)).

THEOREM 5.1. *Assume that* $(M + M^T - A)$ *is SPD. Then,*

$$\mu^\star \leq \frac{\Delta^2}{2-\omega} \cdot \frac{1}{1-\rho_s}, \tag{5.3}$$

where constants Δ *and* ω *are as in Lemma* 2.3*, and where*

$$\rho_s = \|(I - M_s^{-1}A_s)\|_{A_s} \tag{5.4}$$

with $M_s = (S^TMS)$ *and* $A_s = (S^TAS)$*. Note that, although we use* ρ *to represent the spectral radius of a matrix, the quantity* ρ_s *is in general only an upper bound for the spectral radius of compatible relaxation; it is equal to the spectral radius when* M *is symmetric.*

Proof. From (3.16), (3.17), and Lemma 2.3, we have that

$$\mu^\star \leq \frac{\Delta^2}{2-\omega}\mu^\star_\sigma.$$

But, from (2.11) and Theorem 3.1,

$$\begin{aligned}\mu_\sigma^\star &= \frac{1}{\lambda_{\min}(\sigma(M_s)^{-1}A_s)}\\ &= \max_{\mathbf{v}_s}\frac{\langle M_s\mathbf{v}_s,\mathbf{v}_s\rangle}{\langle A_s\mathbf{v}_s,\mathbf{v}_s\rangle}\\ &\le \|A_s^{-1/2}M_sA_s^{-1/2}\|.\end{aligned}$$

Hence, we have

$$\mu^\star \le \frac{\Delta^2}{2-\omega}\|A_s^{-1/2}M_sA_s^{-1/2}\|,$$

and it remains to show that

$$\|A_s^{-1/2}M_sA_s^{-1/2}\| \le (1-\rho_s)^{-1}. \tag{5.5}$$

Consider the following symmetric compatible relaxation matrix:

$$H_{ss} = (I - M_s^{-1}A_s)(I - M_s^{-T}A_s).$$

We have that

$$\begin{aligned}\rho(H_{ss}) &= \rho(A_s^{1/2}H_{ss}A_s^{-1/2})\\ &= \rho((I - A_s^{1/2}M_s^{-1}A_s^{1/2})^T(I - A_s^{1/2}M_s^{-1}A_s^{1/2}))\\ &= \|(I - A_s^{1/2}M_s^{-1}A_s^{1/2})\|^2\\ &= \|(I - M_s^{-1}A_s)\|_{A_s}^2\\ &= \rho_s^2.\end{aligned}$$

Noting that H_{ss} can also be written as $I - M_{ss}^{-1}A_s$, where

$$M_{ss}^{-1} = (M_s^{-1} + M_s^{-T} - M_s^{-1}A_sM_s^{-T}),$$

we have that

$$\rho_s^2 = \rho(H_{ss}) = \max_\lambda |1 - \lambda(M_{ss}^{-1}A_s)| \ge 1 - \lambda_{\min}(M_{ss}^{-1}A_s).$$

Letting $Y_s^{-1} = A_s^{1/2}M_s^{-1}A_s^{1/2}$, one arrives at the coercivity estimate

$$\begin{aligned}(1-\rho_s^2)\langle\mathbf{v}_s,\mathbf{v}_s\rangle &\le \langle M_{ss}^{-1}A_s^{1/2}\mathbf{v}_s, A_s^{1/2}\mathbf{v}_s\rangle\\ &= \langle(Y_s^{-T} + Y_s^{-1} - Y_s^{-1}Y_s^{-T})\mathbf{v}_s,\mathbf{v}_s\rangle\\ &= 2\langle Y_s^{-T}\mathbf{v}_s,\mathbf{v}_s\rangle - \langle Y_s^{-T}\mathbf{v}_s, Y_s^{-T}\mathbf{v}_s\rangle.\end{aligned} \tag{5.6}$$

Using the Cauchy–Schwarz inequality,

$$\langle Y_s^{-T}\mathbf{v}_s,\mathbf{v}_s\rangle \le \langle\mathbf{v}_s,\mathbf{v}_s\rangle^{1/2}\langle Y_s^{-T}\mathbf{v}_s, Y_s^{-T}\mathbf{v}_s\rangle^{1/2},$$

in (5.6), we arrive at

$$(\langle\mathbf{v}_s,\mathbf{v}_s\rangle^{1/2} - \langle Y_s^{-T}\mathbf{v}_s, Y_s^{-T}\mathbf{v}_s\rangle^{1/2})^2 \le \rho_s^2\langle\mathbf{v}_s,\mathbf{v}_s\rangle.$$

That is,

$$(1-\rho_s)^2\langle \mathbf{v}_s, \mathbf{v}_s\rangle \le \langle Y_s^{-T}\mathbf{v}_s, Y_s^{-T}\mathbf{v}_s\rangle.$$

Adding the left- and right-hand sides of the last estimate and estimate (5.6), one gets

$$(1-\rho_s)\langle \mathbf{v}_s, \mathbf{v}_s\rangle \le \langle Y_s^{-T}\mathbf{v}_s, \mathbf{v}_s\rangle.$$

This implies, letting $\mathbf{v}_s := Y_s\mathbf{v}_s$, that

$$\begin{aligned}\|Y_s\mathbf{v}_s\|^2 &= \langle Y_s\mathbf{v}_s, Y_s\mathbf{v}_s\rangle \\ &\le (1-\rho_s)^{-1}\langle \mathbf{v}_s, Y_s\mathbf{v}_s\rangle \\ &\le (1-\rho_s)^{-1}\|Y_s\mathbf{v}_s\|\|\mathbf{v}_s\|.\end{aligned}$$

Therefore, $\|Y_s\mathbf{v}_s\| \le (1-\rho_s)^{-1}\|\mathbf{v}_s\|$, which implies (5.5), and hence, the result. ∎

Theorem 5.1 shows that if compatible relaxation is fast to converge (i.e., ρ_s is small), then $\mu^\star$ is small (similarly for $\mu_\sigma^\star$). To use this result in practice as a means of measuring the quality of a given coarse grid, we must be able to efficiently estimate the value of ρ_s in (5.4). One obvious approach for doing this is to run the compatible relaxation iteration in (5.2) and monitor its convergence. In some cases, this may not be feasible. However, in the case where M is derived from a matrix splitting, $A = M - N$, such that M is explicitly available, the iteration in (5.2) is at least computable.

5.1. Compatible relaxation via subspace correction. Another practical form of compatible relaxation is based on the general subspace correction method framework [21], which encompasses both additive and multiplicative Schwarz. Of particular interest is the question of how to define a compatible relaxation variant of overlapping Schwarz. The iteration in (5.2) does not readily admit how to achieve this. In fact, the question of how to define compatible relaxation variants of general subspace correction methods requires some care.

Consider the following additive method:

$$I - M^{-1}A; \quad M^{-1} = \sum_i I_i(I_i^T A I_i)^{-1} I_i^T, \tag{5.7}$$

where $I_i : \mathbb{R}^{n_i} \to \mathbb{R}^n$ has full rank, $n_i < n$, and $\mathbb{R}^n = \bigcup_i \text{range}\,(I_i)$. Define full rank normalized operators S_i and R_i^T such that

$$\text{range}\,(S_i) = \text{range}\,(I_i^T S), \tag{5.8}$$

$$\text{range}\,(R_i^T) = \text{range}\,(I_i^T R^T). \tag{5.9}$$

In order to define a usable additive version of compatible relaxation, the I_i must be chosen so that the local spaces S_i and R_i^T are orthogonal, i.e., $R_i S_i = 0$. Compatible relaxation is then defined as follows:

$$I - M_{cr}^{-1}A_s; \quad M_{cr}^{-1} = \sum_i S^T I_{s,i}(I_{s,i}^T A I_{s,i})^{-1} I_{s,i}^T S; \quad I_{s,i} = I_i S_i. \tag{5.10}$$

One natural relaxation method that is represented by (5.10) is additive Schwarz. We will discuss this method in more detail below. First, we prove the following lemma and theorem.

LEMMA 5.2. *Assume that we are given the decomposition*

$$\mathbf{v} = S\mathbf{v}_s + R^T\mathbf{v}_c = \begin{bmatrix} S & R^T \end{bmatrix} \begin{bmatrix} \mathbf{v}_s \\ \mathbf{v}_c \end{bmatrix},$$

such that $RS = 0$ *and* $S^TS = I$. *For any matrix* M, *we have that*

$$(S^TM^{-1}S)^{-1} = \overline{M}_{\text{Schur}} := S^TMS - S^TMR^T(RMR^T)^{-1}RMS.$$

If M *is SPD, then the following also holds:*

$$\langle (S^TM^{-1}S)^{-1}\mathbf{v}_s, \mathbf{v}_s\rangle = \min_{\mathbf{v}_c} \langle M(S\mathbf{v}_s + R^T\mathbf{v}_c), (S\mathbf{v}_s + R^T\mathbf{v}_c)\rangle.$$

Proof. Define the hierarchical basis matrix

$$\overline{M} := \begin{bmatrix} S & R^T \end{bmatrix}^T M \begin{bmatrix} S & R^T \end{bmatrix} = \begin{bmatrix} \overline{M}_{ss} & \overline{M}_{sc} \\ \overline{M}_{cs} & \overline{M}_{cc} \end{bmatrix}. \tag{5.11}$$

One has $S^TMS = \overline{M}_{ss}$ by definition. Again, from the definition of $\overline{M}$,

$$\begin{bmatrix} S & R^T \end{bmatrix} \overline{M}^{-1} \begin{bmatrix} S & R^T \end{bmatrix}^T = M^{-1}.$$

Hence,

$$S^TM^{-1}S = S^T \begin{bmatrix} S & R^T \end{bmatrix} \overline{M}^{-1} \begin{bmatrix} S & R^T \end{bmatrix}^T S.$$

Now, using the fact that $RS = 0$ and $S^TS = I$, one gets

$$S^TM^{-1}S = \begin{bmatrix} I & 0 \end{bmatrix} \overline{M}^{-1} \begin{bmatrix} I & 0 \end{bmatrix}^T.$$

Finally, since

$$\overline{M}^{-1} = \begin{bmatrix} (\overline{M}_{\text{Schur}})^{-1} & \star \\ \star & \star \end{bmatrix},$$

one gets

$$S^TM^{-1}S = \begin{bmatrix} I & 0 \end{bmatrix} \overline{M}^{-1} \begin{bmatrix} I & 0 \end{bmatrix}^T = (\overline{M}_{\text{Schur}})^{-1},$$

which implies the first result. The second result follows trivially by noting that

$$\min_{\mathbf{v}_c} \langle M(S\mathbf{v}_s + R^T\mathbf{v}_c), (S\mathbf{v}_s + R^T\mathbf{v}_c)\rangle$$

is a quadratic form in the variable $\mathbf{v}_c$. The minimum is $\langle \overline{M}_{\text{Schur}}\mathbf{v}_s, \mathbf{v}_s\rangle$. ∎

THEOREM 5.3. *Let* M^{-1} *and* M_{cr}^{-1} *be as in* (5.7) *and* (5.10), *respectively. Define* ω *as in Theorem* 5.1, *and define*

$$\rho_{cr} = \|(I - M_{cr}^{-1}A_s)\|_{A_s} = \rho(I - M_{cr}^{-1}A_s). \tag{5.12}$$

Then,

$$\mu^\star \le \frac{1}{2-\omega} \cdot \frac{1}{1-\rho_{cr}}. \tag{5.13}$$

Proof. As before, we can write any vector $\mathbf{e}$ as $\mathbf{e} = S\mathbf{e}_s + R^T\mathbf{e}_c$ for some $\mathbf{e}_s$ and $\mathbf{e}_c$. From (5.8) and (5.9), there exist vectors $\mathbf{e}_{s,i}$ and $\mathbf{e}_{c,i}$ such that $S_i\mathbf{e}_{s,i} = I_i^T S\mathbf{e}_s$ and $R_i^T\mathbf{e}_{c,i} = I_i^T R^T\mathbf{e}_c$. Using this, together with the result in Lemma 5.2 (replacing M^{-1} by $(I_i^T A I_i)$, and R and S by R_i and S_i, respectively), we have

$$\begin{aligned}
\langle M_{cr}^{-1}\mathbf{e}_s, \mathbf{e}_s\rangle &= \sum_i \langle (S_i^T I_i^T A I_i S_i)^{-1} S_i^T I_i^T S\mathbf{e}_s, S_i^T I_i^T S\mathbf{e}_s\rangle \\
&= \sum_i \langle (S_i^T I_i^T A I_i S_i)^{-1}\mathbf{e}_{s,i}, \mathbf{e}_{s,i}\rangle \\
&\le \sum_i \left\langle (I_i^T A I_i)^{-1}\begin{bmatrix} S_i & R_i^T\end{bmatrix}\begin{bmatrix}\mathbf{e}_{s,i}\\ \mathbf{e}_{c,i}\end{bmatrix}, \begin{bmatrix} S_i & R_i^T\end{bmatrix}\begin{bmatrix}\mathbf{e}_{s,i}\\ \mathbf{e}_{c,i}\end{bmatrix}\right\rangle \\
&= \sum_i \left\langle (I_i^T A I_i)^{-1} I_i^T\begin{bmatrix} S & R^T\end{bmatrix}\begin{bmatrix}\mathbf{e}_{s}\\ \mathbf{e}_{c}\end{bmatrix}, I_i^T\begin{bmatrix} S & R^T\end{bmatrix}\begin{bmatrix}\mathbf{e}_{s}\\ \mathbf{e}_{c}\end{bmatrix}\right\rangle \\
&= \sum_i \langle I_i (I_i^T A I_i)^{-1} I_i^T\mathbf{e}, \mathbf{e}\rangle \\
&= \langle M^{-1}\mathbf{e}, \mathbf{e}\rangle.
\end{aligned}$$

Since $\mathbf{e}_c$ was arbitrary, this implies (again, using Lemma 5.2) that

$$\langle M_{cr}^{-1}\mathbf{e}_s, \mathbf{e}_s\rangle \le \min_{\mathbf{e}_c}\langle M^{-1}\mathbf{e}, \mathbf{e}\rangle = \langle (S^T M S)^{-1}\mathbf{e}_s, \mathbf{e}_s\rangle = \langle M_s^{-1}\mathbf{e}_s, \mathbf{e}_s\rangle.$$

Hence, from (3.16), (3.17), and Lemma 2.3, we have that

$$\begin{aligned}
\mu^\star &\le (2-\omega)^{-1}\mu_\sigma^\star \\
&= (2-\omega)^{-1}\frac{1}{\lambda_{\min}(M_s^{-1}A_s)} \\
&\le (2-\omega)^{-1}\frac{1}{\lambda_{\min}(M_{cr}^{-1}A_s)} \\
&\le (2-\omega)^{-1}(1-\rho_{cr})^{-1}. \qquad \square
\end{aligned}$$

When the coarse-grid variables are a subset of the fine-grid variables, then we have that $R = [0\ I]$ and $S = [I\ 0]^T$, and the additive Schwarz method satisfies the criteria for the compatible relaxation in (5.10). To see this, note that, for additive Schwarz, each I_i is a characteristic function over some local subdomain Ω_i. That is, $I_i\mathbf{w} = \mathbf{w}_i$ on Ω_i and zero outside of Ω_i. From the construction of S_i and R_i^T in (5.8) and (5.9), it is clear that they are also just characteristic functions: R_i^T over the C-pts in Ω_i; and S_i over the F-pts in Ω_i. Hence, $R_i S_i = 0\ \forall i$.

Multiplicative versions of compatible relaxation are also possible but more difficult to construct, and may not be necessary anyway. Standard Gauss–Seidel and block Gauss–Seidel methods have straightforward compatible relaxation variants, but a general form for multiplicative subspace correction or multiplicative Schwarz (with overlap) is not apparent.

Multiplicative methods are not as practical in the parallel setting, but have better smoothing properties in the sense that ω is usually bounded away from two without the need for additional damping factors. In practice, a good smoother to use is the natural generalization of F-C relaxation. That is, (post) smoothing should consist of the above additive compatible relaxation process followed by the analogous additive compatible relaxation process on the R^T space. Since $S^T A S$ and RAR^T are well-conditioned in some sense, the additive compatible relaxation methods should work well.

5.2. A more general form of compatible relaxation. Although the compatible relaxation methods presented so far cover many of the traditional relaxation methods, there are still some that may not be represented. In particular, the iteration in (5.2) requires that the matrix M is available *and* that the matrix S^TMS is easily inverted. This may not always be feasible. Additive Schwarz is one such example, albeit one that fortunately has a remedy as described in (5.10). In general, the action of M^{-1} is always available, and motivates us to consider the following compatible relaxation process:

$$\mathbf{e}_{k+1} = (I - (S^TM^{-1}S)(S^TAS))\mathbf{e}_k, \tag{5.14}$$

where, here, S must be normalized so that $S^TS = I_s$, the identity on $\mathbb{R}^{n_s}$. This method is always computable, but must be used with care, as we describe below. First, we state the following result.

THEOREM 5.4. *Assume that the smoother (SPD) M is stable w.r.t. the decomposition $\mathbf{v} = S\mathbf{v}_s + R^T\mathbf{v}_c$ in the sense that for some constant $\gamma \in [0,1)$ the following strengthened Cauchy–Schwarz inequality holds:*

$$\langle MS\mathbf{v}_s, R^T\mathbf{v}_c\rangle \le \gamma\langle MS\mathbf{v}_s, S\mathbf{v}_s\rangle^{1/2}\langle MR^T\mathbf{v}_c, R^T\mathbf{v}_c\rangle^{1/2} \quad \forall \mathbf{v}_s, \mathbf{v}_c. \tag{5.15}$$

Then, the following estimates hold for all $\mathbf{v}_s$:

$$\langle (S^TMS)^{-1}\mathbf{v}_s, \mathbf{v}_s\rangle \le \langle S^TM^{-1}S\mathbf{v}_s, \mathbf{v}_s\rangle \le \frac{1}{1-\gamma^2}\langle (S^TMS)^{-1}\mathbf{v}_s, \mathbf{v}_s\rangle.$$

In other words, the modified compatible relaxation matrix, $(S^TM^{-1}S)$, is spectrally equivalent to the true one, $(S^TMS)^{-1}$.

Proof. Define $\overline{M}$ as in (5.11) in the proof of Lemma 5.2. From the lemma, one trivially has

$$\langle (S^TM^{-1}S)^{-1}\mathbf{v}_s, \mathbf{v}_s\rangle = \langle \overline{M}_{\text{Schur}}\mathbf{v}_s, \mathbf{v}_s\rangle \le \langle S^TMS\mathbf{v}_s, \mathbf{v}_s\rangle.$$

Replacing M by M^{-1} yields the first inequality. The second inequality follows from the corollary to the strengthened Schwarz inequality,

$$\langle \overline{M}_{ss}\mathbf{v}_s, \mathbf{v}_s\rangle \le \frac{1}{1-\gamma^2}\min_{\mathbf{v}_c}\left\langle \overline{M}\begin{bmatrix}\mathbf{v}_s\\ \mathbf{v}_c\end{bmatrix}, \begin{bmatrix}\mathbf{v}_s\\ \mathbf{v}_c\end{bmatrix}\right\rangle = \frac{1}{1-\gamma^2}\langle \overline{M}_{\text{Schur}}\mathbf{v}_s, \mathbf{v}_s\rangle.$$

Here again, replace M by M^{-1} to get the result. ∎

The above theorem implies the following about the eigenvalues of the corresponding iteration matrix (5.14) and the original compatible relaxation matrix in (5.2):

$$\begin{aligned}\lambda(I - (S^TM^{-1}S)A_s) &\le \lambda(I - (S^TMS)^{-1}A_s)\\ &\le \gamma^2 + (1-\gamma^2)\lambda(I - (S^TM^{-1}S)A_s).\end{aligned}$$

Hence, if ρ_g is the spectral radius of $(I - (S^TM^{-1}S)A_s)$, we arrive at the following result, analogous to the results of Theorems 5.1 and 5.3:

$$\mu^\star \le \frac{1}{2-\omega}\cdot\frac{1}{1-\gamma^2}\cdot\frac{1}{1-\rho_g}. \tag{5.16}$$

From this, we see that in order to use the compatible relaxation method in (5.14), we must first have an estimate for the size of γ.

In practice, γ can often be estimated locally. This is the case, for example, when M is assembled from small matrices. That is, let $\langle M\mathbf{v},\mathbf{v}\rangle = \sum_e \langle M_e\mathbf{v}_e,\mathbf{v}_e\rangle = \sum_e \langle M_e(I_e)^T\mathbf{v},(I_e)^T\mathbf{v}\rangle$. Here, $\mathbf{v}_e := \mathbf{v}|_e$. Similarly, for a given $\mathbf{v}_e$ on e, $I_e\mathbf{v}_e$ is the extension of $\mathbf{v}_e$ as zero outside e. Let also $(I_e)^T S = S_e(I_{s,e})^T$ and $(I_e)^T R = R_e(I_{c,e})^T$ for S_e, $I_{s,e}$, R_e, and $I_{c,e}$ supported in e. Then,

$$\langle S^T MS\mathbf{v}_s,\mathbf{v}_s\rangle = \sum_e \langle (S_e)^T M_e S_e \mathbf{v}_{s,e},\mathbf{v}_{s,e}\rangle.$$

If one can say something about the local matrices $(S_e)^T M_e S_e$ and the local Schur complement $\overline{M}_{e,\mathrm{Schur}}$ of $\overline{M}_e = [S_e\ R_e^T]^T M_e [S_e\ R_e^T]$, the maximum of all local γ_e's gives an upper bound for the global γ. This technique is well known in the two-level hierarchical basis literature; cf., e.g., Bank [2].

A similar approach can be used to estimate γ in the case where M^{-1} is obtained by assembling local matrices. As an example, for additive Schwarz, we have that

$$M^{-1} = \sum_i I_i (I_i^T A I_i)^{-1} I_i^T,$$

where, as described near the end of the previous section, I_i is the characteristic function over some local subdomain Ω_i. If we have a local estimate of the form

$$\langle S_i^T (I_i^T A I_i)^{-1} S_i \mathbf{e}_{s,i},\mathbf{e}_{s,i}\rangle \le \frac{1}{1-\gamma_i^2}\langle (S_i^T I_i^T A I_i S_i)^{-1}\mathbf{e}_{s,i},\mathbf{e}_{s,i}\rangle,$$

then, using the proof of Theorem 5.3 for the last inequality below, we can show that

$$\begin{aligned}
\left\langle M^{-1}S\mathbf{e}_s, S\mathbf{e}_s\right\rangle &= \sum_i \langle I_i (I_i^T A I_i)^{-1} I_i^T S\mathbf{e}_s, S\mathbf{e}_s\rangle \\
&= \sum_i \langle S_i^T (I_i^T A I_i)^{-1} S_i \mathbf{e}_{s,i},\mathbf{e}_{s,i}\rangle \\
&\le \frac{1}{1-\max_i \gamma_i^2} \sum_i \langle (S_i^T I_i^T A I_i S_i)^{-1}\mathbf{e}_{s,i},\mathbf{e}_{s,i}\rangle \\
&\le \frac{1}{1-\max_i \gamma_i^2} \langle (S^T MS)^{-1}\mathbf{e}_s,\mathbf{e}_s\rangle.
\end{aligned}$$

The compatible relaxation method in (5.14) is similar to the habituated compatible relaxation scheme in [15]. The latter has the error propagation

$$\mathbf{e}_{k+1} = (I - S^T M^{-1} A S)\mathbf{e}_k. \tag{5.17}$$

The theoretical result is similar to (5.16). We have the following theorem.

THEOREM 5.5. *Assume that the smoother (SPD) M is stable w.r.t. the decomposition $\mathbf{v} = S\mathbf{v}_s + R^T\mathbf{v}_c$ in the sense that for some constant $\gamma \in [0,1)$ the strengthened Cauchy–Schwarz inequality in* (5.15) *holds. Assume that for some constant $\rho_h < 1$ the following convergence estimate holds:*

$$\langle A_s\mathbf{e}_{k+1},\mathbf{e}_{k+1}\rangle \le \rho_h^2 \langle A_s\mathbf{e}_k,\mathbf{e}_k\rangle.$$

Then, the following coercivity estimate holds:

$$\delta\langle M_g\mathbf{e}_s,\mathbf{e}_s\rangle \le \langle A_s\mathbf{e}_s,\mathbf{e}_s\rangle,$$

where $M_g = (S^T M^{-1} S)^{-1}$ *and* $\delta \geq \frac{1}{2}(1-\rho_h)^2$. *The latter coercivity estimate implies convergence of the compatible relaxation method in* (5.14) *with convergence factor* $\rho_g = (1-\delta)$.

Proof. Given $\mathbf{e}_s$, consider the solution $\mathbf{x}$ of the problem

$$M\mathbf{x} = AS\mathbf{e}_s.$$

The following inequality then follows:

$$\begin{aligned}\langle M\mathbf{x}, \mathbf{x}\rangle &= \langle M^{-1/2} AS\mathbf{e}_s, M^{1/2}\mathbf{x}\rangle \\ &\leq \langle M^{-1} AS\mathbf{e}_s, AS\mathbf{e}_s\rangle^{1/2} \langle M\mathbf{x}, \mathbf{x}\rangle^{1/2}.\end{aligned}$$

This implies that $\langle M\mathbf{x}, \mathbf{x}\rangle \leq \langle M^{-1} AS\mathbf{e}_s, AS\mathbf{e}_s\rangle$ which, from Lemma 5.2 and the fact that $2M - A$ is SPD, leads to

$$\begin{aligned}\langle M_g \mathbf{x}_s, \mathbf{x}_s\rangle &= \min_{\mathbf{x}_c} \langle M(S\mathbf{x}_s + R^T\mathbf{x}_c), (S\mathbf{x}_s + R^T\mathbf{x}_c)\rangle \\ &\leq \langle M\mathbf{x}, \mathbf{x}\rangle \leq \langle M^{-1} AS\mathbf{e}_s, AS\mathbf{e}_s\rangle \leq 2\langle A_s\mathbf{e}_s, \mathbf{e}_s\rangle.\end{aligned} \tag{5.18}$$

Now, using Cauchy–Schwarz and the fact that the habituated compatible relaxation is convergent, one has

$$\begin{aligned}\langle M_g\mathbf{e}_s, \mathbf{e}_s - \mathbf{x}_s\rangle &= \langle A_s^{-1/2} M_g\mathbf{e}_s, A_s^{1/2}(I - S^T M^{-1} AS)\mathbf{e}_s\rangle \\ &\leq \rho_h \langle A_s^{-1} M_g\mathbf{e}_s, M_g\mathbf{e}_s\rangle^{1/2} \langle A_s\mathbf{e}_s, \mathbf{e}_s\rangle^{1/2}.\end{aligned}$$

This inequality, using Cauchy–Schwarz and estimate (5.18), implies

$$\begin{aligned}\langle M_g\mathbf{e}_s, \mathbf{e}_s\rangle &\leq \langle \mathbf{x}_s, M_g\mathbf{e}_s\rangle + \rho_h \langle A_s^{-1} M_g\mathbf{e}_s, M_g\mathbf{e}_s\rangle^{1/2} \langle A_s\mathbf{e}_s, \mathbf{e}_s\rangle^{1/2} \\ &\leq \sqrt{2}\langle A_s\mathbf{e}_s, \mathbf{e}_s\rangle^{1/2} \langle M_g\mathbf{e}_s, \mathbf{e}_s\rangle^{1/2} \\ &\quad + \rho_h \langle A_s^{-1} M_g\mathbf{e}_s, M_g\mathbf{e}_s\rangle^{1/2} \langle A_s\mathbf{e}_s, \mathbf{e}_s\rangle^{1/2}.\end{aligned}$$

Dividing through by $\langle A_s\mathbf{e}_s, \mathbf{e}_s\rangle^{1/2} \langle M_g\mathbf{e}_s, \mathbf{e}_s\rangle^{1/2}$, one ends up with the inequality

$$\sqrt{\frac{\langle M_g\mathbf{e}_s, \mathbf{e}_s\rangle}{\langle A_s\mathbf{e}_s, \mathbf{e}_s\rangle}} \leq \sqrt{2} + \rho_h \sqrt{\frac{\langle A_s^{-1} M_g\mathbf{e}_s, M_g\mathbf{e}_s\rangle}{\langle M_g\mathbf{e}_s, \mathbf{e}_s\rangle}}.$$

Now, let

$$\frac{1}{\delta} = \sup_{\mathbf{e}_s} \frac{\langle M_g\mathbf{e}_s, \mathbf{e}_s\rangle}{\langle A_s\mathbf{e}_s, \mathbf{e}_s\rangle} = \sup_{\mathbf{e}_s} \frac{\langle A_s^{-1}\mathbf{e}_s, \mathbf{e}_s\rangle}{\langle M_g^{-1}\mathbf{e}_s, \mathbf{e}_s\rangle}.$$

Then, the following inequality is obtained:

$$\frac{1}{\sqrt{\delta}} \leq \sqrt{2} + \rho_h \frac{1}{\sqrt{\delta}}.$$

That is,

$$\frac{1}{\delta} \leq \frac{2}{(1-\rho_h)^2}. \qquad \square$$

From the theorem and (5.16), we have the following result for habituated compatible relaxation:

$$\mu^\star \le \frac{1}{2-\omega} \cdot \frac{1}{1-\gamma^2} \cdot \frac{2}{(1-\rho_h)^2}. \tag{5.19}$$

This result is weaker than the previous results for the other compatible relaxation methods. However, as with the method in (5.14), habituated compatible relaxation is always computable. In fact, it is the easiest to implement in practice because it directly involves the global smoother $I - M^{-1}A$. To see this, note that since S is normalized, the S^T and S in (5.17) can be pulled outside of the parentheses.

5.3. A coarsening algorithm. The above results suggest that compatible relaxation may serve as a useful tool for selecting coarse grids in AMG methods. We now present a prototype for such a coarsening algorithm in the case where the coarse grid is a subset of the fine grid. That is, consider the case where $R = [0\ I]$ and $S = [I\ 0]^T$. In the coarsening algorithm, one may apply any of the compatible relaxation methods above, i.e., either (5.2), (5.10), (5.14), or (5.17) to the homogeneous equations

$$(S^T A S)\mathbf{x} = 0 \tag{5.20}$$

with some initial guess, say $\mathbf{x}^0 = (x_i^0)$, where $x_i^0 = 1$ or random positive numbers.

(5.21a) Initialize $\mathcal{U} = \Omega$; $\quad \mathcal{C} = \emptyset$

(5.21b) While $\mathcal{U} \neq \emptyset$

(5.21c) Do ν compatible relaxation sweeps

(5.21d) $\mathcal{U} = \{i : (|x_i^\nu|/|x_i^{\nu-1}|) > \theta\}$

(5.21e) $\mathcal{C} = \mathcal{C} \cup \{\text{independent set of } \mathcal{U}\}$

This algorithm is similar to what Livne [15] and Brandt [4] use. Note that the pointwise convergence factor in step (5.21d) is not a meaningful measure when ν is large, and the question of how to choose the candidate set $\mathcal{C}$ is an active area of research.

6. Examples. In this section, we present two examples illustrating the theoretical results of the paper. The first example is a simple anisotropic diffusion problem that demonstrates the ability of the theory (and compatible relaxation) to account for a more general relaxation process; in this case, line relaxation. The example also demonstrates the use of previously developed methods (here, AMGe) for defining adequate interpolation operators in the sense of satisfying **C2** in Theorem 4.2. The second example illustrates how a nontrivial geometric multigrid method for $\mathbf{H}(\mathrm{div})$ fits into the new framework.

6.1. Compatible line relaxation for anisotropic diffusion. Consider the grid-aligned anisotropic problem

$$-\epsilon u_{xx} - u_{yy} = f, \quad (x, y) \in \Omega = (0, 1)^2,$$

with Dirichlet boundary conditions, discretized on a uniform rectangular grid with mesh size $h_x = h_y = h = 2^{-\ell}$ as in Figure 6.1. Using piecewise linear elements on

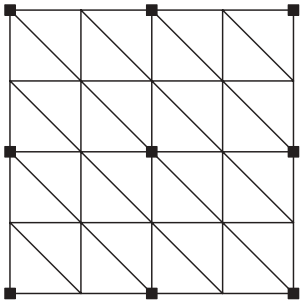


FIG. 6.1. *Uniform grid with triangular elements and standard coarse grid.*

triangles, the resulting macroelement matrix for each rectangle is given by

$$A_e = \epsilon \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

The vertices (nodes) of every rectangle are assumed to have the ordering (x_i, y_j), (x_{i+1}, y_j), (x_i, y_{j+1}), (x_{i+1}, y_{j+1}); where $x_i = ih_x$, $y_j = jh_y$, $i, j = 0, 1, \ldots, 2^\ell$.

Consider a block smoother, where the blocks are given by vertical lines of nodes in the grid. That is, consider a line smoother, where the lines are in the "strong" vertical direction. We note that M can be assembled from the same element matrices as A by zeroing some couplings in A_e (namely, the ones in the x-direction), yielding

$$M_e = \epsilon \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

Assume standard coarsening, so that $S = [I\ 0]^T$, where the zero block corresponds to the coarse nodes. We now analyze the convergence rate of the compatible relaxation process in (5.2). Note that $S^T M S$ and $S^T A S$ can also be assembled from local matrices $M_{s,e}$ and $A_{s,e}$; namely, those obtained from the above matrices in which a row and a column are deleted corresponding to the only coarse node in each rectangular element. Due to symmetry, we delete the last row and last column to get

$$A_{s,e} = \epsilon \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \tag{6.1}$$

and

$$M_{s,e} = \epsilon \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

It is sufficient to compute the eigenvalues of the generalized eigenvalue problem

$$A_{s,e}\mathbf{x} = \lambda M_{s,e}\mathbf{x}.$$

This leads to the following cubic equation for λ:

$$\begin{vmatrix} (1-\lambda)(1+\epsilon) & -\epsilon & -(1-\lambda) \\ -\epsilon & (1-\lambda)(1+\epsilon) & 0 \\ -(1-\lambda) & 0 & (1-\lambda)(1+\epsilon) \end{vmatrix} = 0.$$

The roots are

$$\lambda = 1, 1 \pm \sqrt{\frac{\epsilon}{2+\epsilon}}.$$

Hence, the spectrum of the compatible relaxation iteration matrix $(I - M_s^{-1}A_s)$ is contained in the interval

$$\left[-\sqrt{\frac{\epsilon}{2+\epsilon}}, \sqrt{\frac{\epsilon}{2+\epsilon}}\right].$$

For $\epsilon \in (0, 1]$, this implies that $\rho_s \leq 1/\sqrt{3}$. It is well known that linear interpolation is bounded in energy; i.e., it satisfies (4.1) for some constant η independent of ϵ. In fact, for right-angled triangles, one has $\eta = \frac{1}{1-\gamma^2}$ with $\gamma^2 = \frac{1}{2}$; cf. [16]. Hence, from Theorems 4.2 and 2.2, we can conclude that the two-grid method with the above line smoother converges with a rate bounded independent of ϵ (also a well-known fact).

Now, consider the AMGe measure η in (4.5). We know from Corollary 3.4 that

$$\eta \geq \|A\| \frac{1}{\lambda_{\min}(A_{ff})} \tag{6.2}$$

for any interpolation operator P. Again, because of symmetry, we can bound the minimum eigenvalue of A_{ff} by considering the eigenvalues of the local stiffness matrix with the first and last rows deleted. That is, we can look at the eigenvalues of $A_{s,e}$ in (6.1), which satisfy the following cubic equation for λ:

$$\begin{vmatrix} (1+\epsilon-\lambda) & -\epsilon & -1 \\ -\epsilon & (1+\epsilon-\lambda) & 0 \\ -1 & 0 & (1+\epsilon-\lambda) \end{vmatrix} = 0.$$

The roots are

$$\lambda = (1+\epsilon), (1+\epsilon) \pm \sqrt{1+\epsilon^2}. \tag{6.3}$$

Hence,

$$\lambda_{\min}(A_{ff}) = (1+\epsilon) - \sqrt{1+\epsilon^2} \leq \epsilon, \tag{6.4}$$

which implies that

$$\eta \geq \|A\| \frac{1}{\epsilon} \tag{6.5}$$

for any interpolation operator P. But, as mentioned earlier in this example, linear interpolation satisfies (4.1) for a constant η independent of ϵ. Hence, although the AMGe measure η in (4.5) also implies (4.1), it is clearly a poor estimate for the latter. Note, however, that we may still use (4.5) to construct good interpolation operators. In particular, the AMGe method can produce linear interpolation for this example; the method is just unable to judge the quality of this interpolation operator when the smoother is line relaxation.

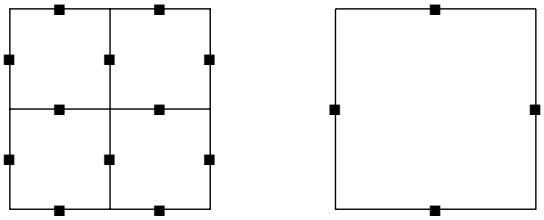


FIG. 6.2. *Coarse rectangle and its refinement. The DOFs of the respective Raviart–Thomas elements are associated with the midpoints of the edges of the elements.*

6.2. Geometric two-grid method for H(div). The space $\mathbf{H}(\mathrm{div})$ is spanned by vector functions $\underline{\chi}$ in $(L_2(\Omega))^d$ ($d = 2$ in the present example) whose divergence is also in $L_2(\Omega)$. Consider the Raviart–Thomas finite element discretization [11] of the $\mathbf{H}(\mathrm{div})$ bilinear form

$$(k^{-1}\underline{\chi}, \underline{\theta}) + (\nabla \cdot \underline{\chi}, \underline{\theta}). \tag{6.6}$$

Here, $k = k(x)$ is a given positive coefficient and $(\cdot, \cdot)$ stands for the $L_2(\Omega)$ inner product. The two-dimensional domain Ω is formed from rectangular fine-grid elements of mesh size h. The elements are obtained by successive steps of uniform refinement of an initial rectangular coarse mesh. The Raviart–Thomas finite element space of lowest order is spanned locally on every fine-grid rectangle by vector polynomials of the form

$$\begin{bmatrix} ax + b \\ cy + d \end{bmatrix}. \tag{6.7}$$

It is clear that by specifying $\underline{\chi} \cdot \mathbf{n}$ on every edge of the rectangles, then every rectangle has four degrees of freedom, and hence the four coefficients a, b, c, and d are uniquely determined. One also notices that $\underline{\chi} \cdot \mathbf{n}$ on every edge is constant. Hence, $\underline{\chi} \cdot \mathbf{n}$ is continuous across every edge of the fine-grid elements and the vector function $\underline{\chi}$ is globally contained in $\mathbf{H}(\mathrm{div})$.

Consider now two triangulations: fine-grid rectangles of mesh size h and coarse-grid rectangles of mesh size $H = 2h$. The degrees of freedom are shown in Figure 6.2. A standard "Lagrangian" basis of V_h is constructed by choosing, for every fine-grid edge, a function $\underline{\phi}$ which has normal component equal to 1 and zero normal components at the remaining edges. Let T be a coarse rectangle formed by four fine-grid ones. The degrees of freedom (DOFs) of a fine-grid vector $\mathbf{v}$ (w.r.t. the chosen Lagrangian basis) restricted to T can be partitioned into two groups: interior (to T) DOFs and boundary DOFs. The boundary DOFs on every edge of T are given by

$$\begin{bmatrix} \mathbf{v} \cdot \mathbf{n}_1 \\ \mathbf{v} \cdot \mathbf{n}_2 \end{bmatrix} \begin{matrix} \} \text{ first fine-grid edge} \\ \} \text{ second fine-grid edge,} \end{matrix}$$

and can be decomposed as follows:

$$\begin{bmatrix} \mathbf{v} \cdot \mathbf{n}_1 \\ \mathbf{v} \cdot \mathbf{n}_2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \mathbf{v} \cdot \mathbf{n}_1 - \mathbf{v} \cdot \mathbf{n}_2 \\ \mathbf{v} \cdot \mathbf{n}_2 - \mathbf{v} \cdot \mathbf{n}_1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \mathbf{v} \cdot \mathbf{n}_1 + \mathbf{v} \cdot \mathbf{n}_2 \\ \mathbf{v} \cdot \mathbf{n}_1 + \mathbf{v} \cdot \mathbf{n}_2 \end{bmatrix}.$$

Introduce now the operators acting on vectors spanned by the boundary DOFs,

$$R_B = \frac{1}{\sqrt{2}} \begin{bmatrix} I & I \end{bmatrix} \quad \text{and} \quad S_B = \frac{1}{\sqrt{2}} \begin{bmatrix} I & -I \end{bmatrix}^T. \tag{6.8}$$

Next, partition the stiffness matrix A into a 2×2 block form with blocks corresponding to the interior and boundary DOFs. That is,

$$A = \begin{bmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{BB} \end{bmatrix} \begin{matrix} \} \text{ interior fine-grid edges w.r.t. to coarse elements} \\ \} \text{ boundary fine-grid edges w.r.t. to coarse elements.} \end{matrix}$$

Note that A_{II} is block-diagonal with blocks of size 4×4. Denote the reduced matrix (obtained by "static condensation") $A_B = A_{BB} - A_{BI}(A_{II})^{-1}A_{IB}$. Note that A_B is sparse and explicitly available. For every coarse element edge, fix an ordering of the underlying fine-grid edges. This induces a natural partitioning of the boundary DOFs into two groups, corresponding to the above block structure (6.8) of R_B and S_B. Finally, introduce the global decomposition operators

$$S = \begin{bmatrix} I & -A_{II}^{-1}A_{IB}S_B \\ 0 & S_B \end{bmatrix} \tag{6.9}$$

and

$$R = \begin{bmatrix} 0 & R_B \end{bmatrix}. \tag{6.10}$$

Clearly, $RS = 0$ and $RR^T = R_B(R_B)^T = I$.

We now choose the following smoother:

$$\begin{aligned} M &= \begin{bmatrix} A_{II} & 0 \\ A_{BI} & \operatorname{diag}(A_B) \end{bmatrix} \begin{bmatrix} I & -A_{II}^{-1}A_{IB} \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} A_{II} & A_{IB} \\ A_{BI} & \operatorname{diag}(A_B) + A_{BI}A_{II}^{-1}A_{IB} \end{bmatrix}. \end{aligned} \tag{6.11}$$

Since M is in a factored form, it is straightforward to implement its inverse action; it involves two actions of the block-diagonal matrix $(A_{II})^{-1}$ and one solves with the scalar diagonal matrix $\operatorname{diag}(A_B)$.

One can see that

$$S^T M S = \begin{bmatrix} A_{II} & 0 \\ 0 & S_B^T \operatorname{diag}(A_B) S_B \end{bmatrix}.$$

Similarly,

$$S^T A S = \begin{bmatrix} A_{II} & 0 \\ 0 & S_B^T A_B S_B \end{bmatrix}.$$

The compatible relaxation in (5.2) tells us to look at the matrix

$$(S^T M S)^{-1}(S^T A S) = \begin{bmatrix} I & 0 \\ 0 & (S_B^T \operatorname{diag}(A_B) S_B)^{-1} S_B^T A_B S_B \end{bmatrix}.$$

Based on a result by Cai, Goldstein, and Pasciak [8], one can show that $S_B^T A_B S_B$ is spectrally equivalent to a diagonal matrix. In particular, it is spectrally equivalent to the matrix $\text{const} \cdot I_B$, where const is piecewise constant w.r.t. the coarse element edges. This verifies that the respective compatible relaxation gives rise to a well-conditioned matrix $(S^T M S)^{-1}(S^T A S)$.

It remains to construct a bounded in energy ("approximate harmonic") interpolation operator P. We choose here the P which is naturally defined from the embedding

$V_H \subset V_h$. In operator form, P is the identity. However, in matrix form, its action is computed as follows. Given $\mathbf{v}_H$, consider its four DOFs of the form $\mathbf{v}_H \cdot \mathbf{n}$ for the four edges of every coarse element. These are four constants. Based on these DOFs, one finds the polynomial representation of $\mathbf{v}_H$ on every coarse element. It has the form (6.7). That is, one determines the four constants a, b, c, and d. Then one computes $\mathbf{v}_H \cdot \mathbf{n}$ for all interior fine-grid edges. These, as mentioned above, are also constants (four). Then on every fine-grid edge we have specified the fine-grid DOFs $\mathbf{v} \cdot \mathbf{n}$ which are used in the computation.

To prove the energy boundedness of P we proceed as follows. Given $\mathbf{v} \in V_h$. Compute $R\mathbf{v}$. This takes into account only the DOFs which correspond to the boundary (w.r.t. the coarse elements) fine-grid edges. Using function notation it means that we have computed the coarse edge integrals $\int_F \mathbf{v} \cdot \mathbf{n}\, d\varrho$ for every coarse edge F. Based on the four values for every coarse element we construct the unique coarse vector $\mathbf{v}_H = P(R\mathbf{v})$. It has the property that $\int_F \mathbf{v}_H \cdot \mathbf{n}\, d\varrho = \sqrt{2} \int_F \mathbf{v} \cdot \mathbf{n}\, d\varrho$. In other words, for any constant function w on a given coarse element T, we get

$$\int_{\partial T} w\mathbf{v}_H \cdot \mathbf{n}\, d\varrho = \sqrt{2} \int_{\partial T} w\mathbf{v} \cdot \mathbf{n}\, d\varrho.$$

Using the fact that $\nabla w = 0$ on T and the divergence theorem, we get

$$\int_T w \nabla \cdot \mathbf{v}_H\, dx\, dy = \sqrt{2} \int_T w \nabla \cdot \mathbf{v}\, dx\, dy.$$

If one introduces the elementwise L_2-projection Q_H onto the space of piecewise constant functions (w.r.t. the coarse elements), the above identity shows that $\nabla \cdot (PR\mathbf{v}) = \sqrt{2} Q_H \nabla \cdot \mathbf{v}$. This immediately implies the inequality

$$(\nabla \cdot (PR\mathbf{v}), \nabla \cdot (PR\mathbf{v})) = 2(Q_H \nabla \cdot \mathbf{v}, Q_H \nabla \cdot \mathbf{v}) \le 2(\nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{v}).$$

It remains to bound the L_2-norm of $(PR\mathbf{v})$,

$$(PR\mathbf{v}, PR\mathbf{v}) \le \eta(\mathbf{v}, \mathbf{v}),$$

for a mesh-independent constant η. We note that

$$\int_T \mathbf{v} \cdot \mathbf{v}\, dxdy \simeq h^2 \sum_{f\text{: edge of fine-grid element } \tau \subset T} (\mathbf{v} \cdot \mathbf{n}_f)^2.$$

Similarly,

$$\int_T \mathbf{v}_H \cdot \mathbf{v}_H\, dxdy \simeq H^2 \sum_{F\text{: edge of } T} (\mathbf{v}_H \cdot \mathbf{n}_F)^2.$$

Let $F = f_1 \cup f_2$. Since $\mathbf{v}_H \cdot \mathbf{n}_F = \frac{1}{\sqrt{2}}(\mathbf{v} \cdot \mathbf{n}_{f_1} + \mathbf{v} \cdot \mathbf{n}_{f_2})$, hence

$$(\mathbf{v}_H \cdot \mathbf{n}_F)^2 = \frac{1}{2}(\mathbf{v} \cdot \mathbf{n}_{f_1} + \mathbf{v} \cdot \mathbf{n}_{f_2})^2 \le (\mathbf{v} \cdot \mathbf{n}_{f_1})^2 + (\mathbf{v} \cdot \mathbf{n}_{f_2})^2.$$

This shows that there exists a mesh-independent constant η such that

$$\int_T \mathbf{v}_H \cdot \mathbf{v}_H\, dxdy \le \eta \int_T \mathbf{v} \cdot \mathbf{v}\, dxdy,$$

which after summation over all coarse elements leads to the required L_2-boundedness of PR. Thus, we get the desired result that PR is bounded in $\mathbf{H}(\mathrm{div})$-norm.

REFERENCES

[1] D. N. Arnold, R. S. Falk, and R. Winther, *Multigrid in H(div) and H(curl)*, Numer. Math., 85 (2000), pp. 197–217.

[2] R. E. Bank, *Hierarchical Bases and the Finite Element Method*, Acta Numer. 5, Cambridge University Press, Cambridge, 1996, pp. 1–43.

[3] A. Brandt, *Algebraic multigrid theory: The symmetric case*, Appl. Math. Comput., 19 (1986), pp. 23–56.

[4] A. Brandt, *General highly accurate algebraic coarsening*, Electron. Trans. Numer. Anal., 10 (2000), pp. 1–20.

[5] A. Brandt and N. Dinar, *Multigrid solutions to elliptic flow problems*, in Numerical Methods for Partial Differential Equations, S. Parter, ed., Academic Press, New York, 1979, pp. 53–147.

[6] A. Brandt, S. F. McCormick, and J. W. Ruge, *Algebraic multigrid (AMG) for sparse matrix equations*, in Sparsity and Its Applications, D. J. Evans, ed., Cambridge University Press, Cambridge, 1984.

[7] M. Brezina, A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge, *Algebraic multigrid based on element interpolation (AMGe)*, SIAM J. Sci. Comput., 22 (2000), pp. 1570–1592.

[8] Z. Cai, C. I. Goldstein, and J. E. Pasciak, *Multilevel iteration for mixed finite element systems with penalty*, SIAM J. Sci. Comput., 14 (1993), pp. 1073–1088.

[9] T. Chartier, R. D. Falgout, V. E. Henson, J. Jones, T. Manteuffel, S. McCormick, J. Ruge, and P. S. Vassilevski, *Spectral AMGe (ρAMGe)*, SIAM J. Sci. Comput., 25 (2003), pp. 1–26.

[10] A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, G. N. Miranda, and J. W. Ruge, *Robustness and scalability of algebraic multigrid*, SIAM J. Sci. Comput., 21 (2000), pp. 1886–1908.

[11] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations, Theory and Algorithms*, Springer-Verlag, New York, 1986.

[12] V. E. Henson and P. S. Vassilevski, *Element-free AMGe: General algorithms for computing interpolation weights in AMG*, SIAM J. Sci. Comput., 23 (2001), pp. 629–650.

[13] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.

[14] J. E. Jones and P. S. Vassilevski, *AMGe based on element agglomeration*, SIAM J. Sci. Comput., 23 (2001), pp. 109–133.

[15] O. E. Livne, *Coarsening by compatible relaxation*, Numer. Linear Algebra Appl., 11 (2004), pp. 205–227.

[16] J. F. Maitre and F. Musy, *The contraction number of a class of two-level method; an exact evaluation for some finite element subspaces and model problems*, in Multigrid Methods, W. Hackbusch and U. Trottenberg, eds., Lecture Notes in Math. 960, Springer-Verlag, Berlin, 1982, pp. 534–544.

[17] J. W. Ruge and K. Stüben, *Efficient solution of finite difference and finite element equations by algebraic multigrid (AMG)*, in Multigrid Methods for Integral and Differential Equations, D. J. Paddon and H. Holstein, eds., Clarendon Press, Oxford, 1985, pp. 169–212.

[18] J. W. Ruge and K. Stüben, *Algebraic multigrid (AMG)*, in Multigrid Methods, S. F. McCormick, ed., Frontiers Appl. Math. 3, SIAM, Philadelphia, 1987, pp. 73–130.

[19] U. Trottenberg, C. W. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, London, 2001.

[20] P. Vaněk, J. Mandel, and M. Brezina, *Algebraic multigrid by smoothed aggregation for second and fourth order problems*, Computing, 56 (1996), pp. 179–196.

[21] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1694–1728

CONSTRAINT-PRESERVING UPWIND METHODS FOR MULTIDIMENSIONAL ADVECTION EQUATIONS*

M. TORRILHON† AND M. FEY†

Abstract. A general framework for constructing constraint-preserving numerical methods is presented and applied to a multidimensional divergence-constrained advection equation. This equation is part of a set of hyperbolic equations that evolve a vector field while locally preserving either its divergence or its curl. We discuss the properties of these equations and their relation to ordinary advection. Due to the constraint, such equations form model equations for general evolution equations with intrinsic constraints which appear frequently in physics.

The general framework allows the construction of numerical methods that preserve *exactly* the discretized constraint by special flux distribution. Assuming a rectangular, two-dimensional grid as a first approach, application of this framework leads to a locally constraint-preserving multidimensional upwind method. We prove consistency and stability of the new method and present several numerical experiments. Finally, extensions of the method to the three-dimensional case are described.

Key words. multidimensional hyperbolic equations, advection, constraints, finite-volume method, stability

AMS subject classifications. 65M06, 65M12

DOI. 10.1137/S0036142903425033

1. Introduction.

Many evolution equations in physics and engineering come with intrinsic constraints, i.e., local differential constraints that follow directly from the evolution operator. Such evolutions will be called *constraint-preserving*. The most popular example is the evolution of the magnetic flux density $\mathbf{B}$ in electrodynamics: The divergence of $\mathbf{B}$ must be zero for the initial conditions; afterwards the analytic evolution will keep the divergence of the field untouched. The same property is present in the system of magnetohydrodynamics of plasma flows (e.g., [6]), and a similar operator arises in the vorticity equation of incompressible flow (e.g., [12]). Vorticity-preserving equations are used, e.g., in meteorological flows [19], while in [20] a vorticity-preserving system is investigated, which may be related to the linearized Euler equations. Furthermore, the evolution equations of general relativity possess constraints whose properties are lively discussed; see, e.g., [22].

Intrinsic constraints are also expected to hold in numerical calculations of the corresponding evolution, at least in a discrete manner. The discrete approximation of the evolution operator should mimic the analytic properties as far as possible in order to obtain a most physical discrete solution. Nevertheless, the construction of commonly available numerical methods ignores constraints and indeed those methods generally introduce disturbances to the constraints. These disturbances may be argued to be small due to consistency: Since the constraint is an analytic property of the evolution operator it will be respected in a converged solution (see [3] in the context of general relativity). But this argument holds only for smooth solutions, where the disturbances of the constraint are of the order of the truncation error. For discontinuous solutions the error of the constraint becomes so large that computations completely fail (see, e.g., [6] in the context of magnetohydrodynamics). In

*Received by the editors March 26, 2003; accepted for publication (in revised form) November 26, 2003; published electronically December 27, 2004.
http://www.siam.org/journals/sinum/42-4/42503.html
†Seminar for Applied Mathematics, ETH Zurich, ETH-Zentrum, CH-8092 Zürich, Switzerland (manuel@math.ethz.ch, fey@math.ethz.ch).

general, relativity errors in the constraints can excite instabilities [22]. It becomes obvious that controlling intrinsic constraints in numerical methods is required in the construction of accurate and reliable schemes. Even if the constraint is not preserved by the complete evolution operator but only by a part of it, a corresponding partial constraint-preserving discretization is most desirable. This yields that the constraint is numerically only affected by those causes which arose from the discretization of the nonpreserving part in the equations. A similar statement may also be found in [27].

The literature provides many works which deal with the divergence-constraint in the equations of magnetohydrodynamics. Global approaches like in [2] solve elliptic equations each time step in order to correct the solution. A popular approach uses a local correction procedure with the help of a staggered grid (see [1], [5], and [8]) which is applied after a time step of a usual finite-volume method. A third approach modifies the evolution equation (see [21], [6]) so that the error in the constraint is advected or diffused away. The common idea of these approaches is to *correct an existing error* of the constraint. See also [26] for a collection and comparison of methods and [7] for an approach on unstructured grids.

The staggered approach is equivalent to the mimetic discretizations as presented in [13] and [14] if applied to a rectangular grid. These schemes store different variables at different locations in the grid, like edges and vertices. The starting point is to derive discrete vector-analytic identities using special div-, curl-, and grad-definitions. These identities are responsible for discrete constraint preservation. The results of [13] are used in computational electrodynamics; see [14]. The application in finite-volume schemes is complicated since the usage of cell averages for the variables is then mandatory. Examples for a staggered grid scheme in meteorological flows and in vorticity methods are given in [19] and [12], respectively.

This paper will present a general framework for constructing genuine locally constraint-preserving finite-volume methods. We aim at explicit methods that use only a primary finite-volume grid. All variables will be stored in the cell centers and considered as cell mean values. The constraints as well as the flux operator will be discretized with this single grid using the cell averaged values. As an example we will consider so-called constraint-preserving advection equations. These advection equations must be seen as model equations for general evolution equations with constraints. Besides this they also provide interesting aspects of the advection of vector fields. The application of the presented general framework to constraint-preserving advection leads to an upwind method which exactly preserves the local values of the discrete constraint. This is the discrete imitation of the analytic property. The main idea is the usage of a special discrete operator for the constraint. Since the constraint and its preservation are relevant only in more than one dimension the resulting scheme is necessarily multidimensional. We obtain a method that is second order in time and space and is stable for Courant numbers $|c| \leq 1$. Consistency and stability are proven. Several numerical experiments with smooth and discontinuous solutions demonstrate the performance of the scheme. Within the framework we could also re-derive two methods that are known in the literature but which are inappropriate for solving constraint-preserving advection due to instabilities. The main part of the paper considers a two-dimensional setting on a rectangular grid. The presented framework also applies to general grids at the cost of more involved calculations. Methods on unstructured grids are the subject of future work. In the last section we give a sketch of the method in three dimensions.

The paper is organized as follows: In the next section we introduce constrained advection equations for vector fields and discuss their properties and their relation to

ordinary advection as well as to real physical models. In section 3 the general framework is presented that describes how numerical constraint-preserving methods may be constructed. The application of this framework to constraint-preserving advection follows in section 4. In the beginning of that section we discuss discrete constraint operators and deduce some instructive methods, while the final upwind method and its properties are presented in section 4.3. Section 5 is devoted to the numerical experiments and considers smooth as well as discontinuous solutions. Finally, we give details of the three-dimensional case in section 6 and draw conclusions in the last section.

2. Constrained advection equations. We consider a given velocity field $\mathbf{v}$ in a domain Ω of the three-dimensional space

$$\mathbf{v} : \Omega \subseteq \mathbb{R}^3 \to \mathbb{R}^3 \tag{1}$$

which remains constant in time. A second vector field $\mathbf{u} : \Omega \to \mathbb{R}^3$ is said to be advected in the velocity field $\mathbf{v}$ if it obeys the evolution equation

$$\partial_t \mathbf{u} + \operatorname{div}(\mathbf{u} \otimes \mathbf{v}) = 0 \quad \text{in } \Omega, \tag{2}$$

where the divergence acts on the rows of the tensorial product $\mathbf{u} \otimes \mathbf{v}$, i.e., in the components of $\mathbf{v}$. Hence, (2) represents scalar advection equations for each component of $\mathbf{u}$. In components the vector field $\mathbf{u}$ and the advection velocity $\mathbf{v}$ are written as

$$\mathbf{u} = (u^{(x)}, u^{(y)}, u^{(z)})^T, \quad \mathbf{v} = (v^{(x)}, v^{(y)}, v^{(z)})^T. \tag{3}$$

An evolution like (2) represents a raw model for virtually any physical transport process. Correspondingly there exists a vast amount of work concerning analytical and numerical aspects of (2) in the literature. Note that advection of type (2) decouples the components of the vector field $\mathbf{u}$ and each component is advected separately. We will call this *ordinary advection*.

2.1. div / curl-preserving advection. There are two more evolution equations which we shall show to be closely related to ordinary advection. They follow formally from (2) by replacing the differential operator and the tensorial product. We write

$$\begin{aligned} \partial_t \mathbf{u} + \operatorname{grad}(\mathbf{u} \cdot \mathbf{v}) &= 0, \\ \partial_t \mathbf{u} + \operatorname{curl}(\mathbf{u} \times \mathbf{v}) &= 0, \end{aligned} \tag{4}$$

where $\mathbf{u} \cdot \mathbf{v}$ and $\mathbf{u} \times \mathbf{v}$ denote the scalar product and the cross product, respectively. Note that the components of $\mathbf{u}$ are now coupled in the equations (4).

Since for any function ψ we have $\operatorname{curl}\operatorname{grad}\psi \equiv 0$ and $\operatorname{div}\operatorname{curl}\psi \equiv 0$ we can deduce an accompanying equation for both types of evolutions in (4) which may be integrated. We obtain for the considered domain Ω

$$\begin{aligned} &\text{for } (4)_{\text{grad}} \to && \operatorname{curl}\mathbf{u} = \text{const in time}, \\ &\text{for } (4)_{\text{curl}} \to && \operatorname{div}\mathbf{u} = \text{const in time} \end{aligned} \tag{5}$$

as additional equations. These equations state that the curl of the vector field in the case of $(4)_{\text{grad}}$ or its divergence in the case of $(4)_{\text{curl}}$ stays locally (hence globally) unaffected from the evolution. The initial fields of curl or divergence in the particular cases are frozen and their values stay locally the same. We therefore denote the evolution equation $(4)_{\text{grad}}$ by *curl-preserving advection* and $(4)_{\text{curl}}$ by *div-preserving advection*.

The equations (5) may be viewed as *intrinsic or inherent constraints* to the evolution equations in (4). In the language of [4] these constraints form involutions of the equations (4). We stress that these constraints are intrinsic to the evolution equations since they must not be additionally imposed to the solution. They are an inherent property of the transport operator. Any analytic solution of (4) fulfills the constraints of (5) automatically. However, this might not be true in a numerical setting where the equations are discretized. Furthermore, the apparently elliptic character of the constraints do not influence the character of the evolution. We will show later that the equations in (4) are purely hyperbolic.

2.2. Physical examples. Though less frequently than ordinary advection, the constraint-preserving evolution equations in (4) may be found in physical models as well. Furthermore, both equations should be viewed as special cases of more general models where the differential evolution operators act on general functions of $\mathbf{u}$.

A well-known example is the Maxwell equations of electrodynamics

$$\begin{aligned} \partial_t \mathbf{B} + \operatorname{curl} \mathbf{E} &= 0, \\ \partial_t \mathbf{D} - \operatorname{curl} \mathbf{H} &= \mathbf{j}, \end{aligned} \tag{6}$$

where $\mathbf{B}$ is the magnetic flux density and $\mathbf{D}$ the electric displacement. Both evolutions have the structure of $(4)_{\mathrm{curl}}$. Since $\operatorname{div} \mathbf{B} = 0$ is stated in the third Maxwell equation, the intrinsic constraint of $(6)_1$ establishes the solenoidality of the $\mathbf{B}$-field during the entire evolution. The second equation $(6)_2$ together with the fourth Maxwell equation $\operatorname{div} \mathbf{D} = \rho$ yields the conservation law for the charge density ρ. This additional law must be viewed as a constraint to the evolution $(6)_2$.

In ideal magnetohydrodynamics of plasma flows only the first Maxwell equation $(6)_1$ plays a role and $\mathbf{E}$ is given by $\mathbf{E} = -\mathbf{v} \times \mathbf{B}$, where $\mathbf{v}$ is the plasma velocity. Thus we have

$$\partial_t \mathbf{B} + \operatorname{curl}(\mathbf{B} \times \mathbf{v}) = 0 \tag{7}$$

as evolution equation for the $\mathbf{B}$-field which is identical to $(4)_{\mathrm{curl}}$. Due to the intrinsic constraint of (7), the divergence of $\mathbf{B}$ remains zero if it is zero initially. Since this property is spoiled in an ordinary numerical calculation, the preservation of the divergence is a major issue in computational magnetohydrodynamics; see, e.g., [6].

The Navier–Stokes equations for incompressible flow read as

$$\begin{aligned} \partial_t \mathbf{v} + \mathbf{v} \cdot \operatorname{grad} \mathbf{v} + \operatorname{grad} p &= \Delta \mathbf{v}, \\ \operatorname{div} \mathbf{v} &= 0, \end{aligned} \tag{8}$$

where $\mathbf{v}$ is the flow velocity and p is the pressure. Note that the second equation *is not an intrinsic constraint*. It does not follow from the evolution equation for $\mathbf{v}$; instead it is an equation to determine the pressure. In some approaches (see, e.g., [12]) the system of Navier–Stokes is rewritten in terms of the vorticity $\Omega = \operatorname{curl} \mathbf{v}$. The evolution equation for the vorticity may then be found from $(8)_1$ using the identity $\mathbf{v} \cdot \operatorname{grad} \mathbf{v} = \operatorname{grad} \frac{1}{2}\mathbf{v}^2 - \mathbf{v} \times \operatorname{curl} \mathbf{v}$ and is given by

$$\partial_t \Omega + \operatorname{curl}(\Omega \times \mathbf{v}) = \Delta \Omega. \tag{9}$$

This represents again a div-preserving evolution like $(4)_{\mathrm{curl}}$.

An evolution like $(4)_{\mathrm{grad}}$ appears to be less frequent. It is encountered, for example, in meteorological models where it originates from the system for shallow water

flows, and the preservation of $\operatorname{curl}\mathbf{v}$ is a concern in numerical meteorology; see, e.g., [19]. The shallow water system is usually written in conservation laws

$$\begin{aligned} \partial_t h + \operatorname{div}(h\mathbf{v}) &= 0, \\ \partial_t h\mathbf{v} + \operatorname{div}\left(h\mathbf{v}\otimes\mathbf{v} + (\tfrac{1}{2}g\,h^2)\mathbf{I}\right) &= 0 \end{aligned} \tag{10}$$

for the water height h and the flow velocity $\mathbf{v}$. The gravitational constant is g. In meteorology the flow is assumed to be smooth and the momentum balance $(10)_2$ is reduced to an equation for $\mathbf{v}$. Using the first equation and again the identity $\mathbf{v}\cdot\operatorname{grad}\mathbf{v} = \operatorname{grad}\frac{1}{2}\mathbf{v}^2 - \mathbf{v}\times\operatorname{curl}\mathbf{v}$ we obtain

$$\partial_t\mathbf{v} + \operatorname{grad}(\tfrac{1}{2}\mathbf{v}^2 + g\,h) = \mathbf{v}\times\Omega, \tag{11}$$

where again $\Omega = \operatorname{curl}\mathbf{v}$ is introduced. In this equation the curl-preserving operator of $(4)_{\text{grad}}$ is present. Here, the vector field $\mathbf{u}$ coincides with the advection velocity $\mathbf{v}$. The shallow water system is a two-dimensional model ($\partial_z \equiv 0$) with vanishing z-component of $\mathbf{v}$. Hence, the vorticity Ω has only one nonvanishing component $\Omega = \partial_x v^{(y)} - \partial_y v^{(x)}$ and the right-hand side of (11) has the form $\left(\Omega\, v^{(y)}, -\Omega\, v^{(x)}\right)^T$.

2.3. Identification as degenerated advection. We return to the equations in (4) to discuss more of its properties. So far it is not obvious that these equations are related to a kind of advection. Clearly, they state processes different from ordinary advection. We proceed to uncover the relation. A first inspection leads to the fact that both equations in (4) may be transformed into the form of a conservation law $\partial_t\mathbf{u} + \operatorname{div}\mathbf{f}(\mathbf{u}) = 0$ with appropriate definition of the matrix $\mathbf{f}(\mathbf{u})$. We obtain

$$\begin{aligned} &\text{for } (4)_{\text{grad}} \rightarrow && \partial_t\mathbf{u} + \operatorname{div}\left((\mathbf{u}\cdot\mathbf{v})\,\mathbf{I}\right) = 0, \\ &\text{for } (4)_{\text{curl}} \rightarrow && \partial_t\mathbf{u} + \operatorname{div}(\mathbf{u}\otimes\mathbf{v} - \mathbf{v}\otimes\mathbf{u}) = 0, \end{aligned} \tag{12}$$

where $\mathbf{I}$ represents the identity matrix. Thus, in both processes each component of $\mathbf{u}$ is conserved and the evolution equations are conservation laws. Now, to investigate (12) the matrix $\mathbf{A}(\mathbf{n})$ of linear combinations of the directional Jacobians of the flux function $\mathbf{f}$ is formed. We have

$$\mathbf{A}(\mathbf{n}) = D\mathbf{f}(\mathbf{u})\,\mathbf{n}, \tag{13}$$

where $\mathbf{n}$ is a space direction to be chosen. This matrix is used to classify a conservation law; see, e.g., [23]. An equation is hyperbolic if $\mathbf{A}(\mathbf{n})$ has real eigenvalues and a complete set of eigenvectors for any direction of $\mathbf{n}$. The eigenvalues are then interpreted as characteristic velocities. The eigenvectors represent the part of the conserved vector $\mathbf{u}$ which is transported with the corresponding velocity.

We recall that in case of ordinary advection we have $\mathbf{f}(\mathbf{u}) = \mathbf{u}\otimes\mathbf{v}$ and

$$\mathbf{A}^{(\text{ordinary})}(\mathbf{n}) = (\mathbf{n}\cdot\mathbf{v})\,\mathbf{I} \quad\Rightarrow\quad \lambda_{1,2,3} = \mathbf{n}\cdot\mathbf{v}, \quad V_{1,2,3} = \mathbb{R}^3 \tag{14}$$

with eigenvalues λ_i and corresponding eigenspaces V_i. The eigenvalue $\mathbf{n}\cdot\mathbf{v}$ is real and threefold and the complete 3-space is the eigenspace for this eigenvalue. The process of advection may be defined by the presence of an eigenvalue $\mathbf{n}\cdot\mathbf{v}$. This follows from the Friedrichs diagram (see, e.g., [16]), which displays the propagation of a point disturbance associated to a certain characteristic velocity. In case of advection, the point disturbance remains a point and is simply propagated with the advection velocity $\mathbf{v}$. Due to the eigenspace in this case any vector from 3-space can be advected which corresponds to the decoupling of the advection equations in (2).

For the evolution equation of type $(4)_{\text{grad}}$ we obtain

$$\mathbf{A}^{(\text{grad})}(\mathbf{n}) = \mathbf{n} \otimes \mathbf{v} \quad \Rightarrow \quad \begin{cases} \lambda_1 = \mathbf{n} \cdot \mathbf{v}, & V_1 = [\mathbf{n}], \\ \lambda_{2,3} = 0, & V_{2,3} = [\mathbf{v}]^{\perp} \end{cases} \tag{15}$$

for the eigenvalues and eigenspaces. One eigenvalue is given by $\mathbf{n} \cdot \mathbf{v}$ which identifies the process as advection. However, there exists a second eigenvalue which is zero, and this leads to a splitting of the 3-space into two eigenspaces. That is, not all components of a vector $\mathbf{u}$ are advected in the Friedrichs diagram. Indeed, according to $V_{2,3}$ any vector *orthogonal* to the advection velocity will stay in place. Any other vector is simply advected. This behavior must be viewed against the background of the constraint-preserving property: The eigenvalue $\lambda_{2,3}$ represents the constraint mode (see also [22]) which keeps the curl of $\mathbf{u}$ locally untouched.

The evolution of type (4) shows the analogous behavior. Eigenvalues and eigenspaces are given by

$$\mathbf{A}^{(\text{curl})}(\mathbf{n}) = (\mathbf{n} \cdot \mathbf{v})\,\mathbf{I} - \mathbf{v} \otimes \mathbf{n} \quad \Rightarrow \quad \begin{cases} \lambda_1 = 0, & V_1 = [\mathbf{v}], \\ \lambda_{2,3} = \mathbf{n} \cdot \mathbf{v}, & V_{2,3} = [\mathbf{n}]^{\perp}. \end{cases} \tag{16}$$

Again we can identify the process as advection due to the eigenvalue $\lambda_{2,3}$. The first eigenvalue is zero and represents the constraint mode. In this case vectors *parallel* to $\mathbf{v}$ remain untouched due to the eigenspace V_1. This corresponds to the preservation of the divergence of $\mathbf{u}$.

Note that it is not possible to decouple the equations in (12) since the spatial derivatives do not diagonalize simultaneously. Furthermore, though only real eigenvalues exists, the hyperbolicity of the evolution equations in (12) degenerates due to the lack of eigenvectors in certain cases. Indeed, in both cases of the evolutions (4), directions $\mathbf{n}$ orthogonal to the advection velocity give only $V_1 \subset V_{2,3} \subset \mathbb{R}^3$.

2.4. Special cases. It is instructive to consider some special cases of curl-preserving and div-preserving advection. They will emphasize the advection character of the equations.

If the value of the curl or of the divergence of $\mathbf{u}$ is assumed to vanish initially in Ω, i.e.,

$$\begin{aligned} &\text{for } (4)_{\text{grad}} \rightarrow && \operatorname{curl} \mathbf{u} \equiv 0, \\ &\text{for } (4)_{\text{curl}} \rightarrow && \operatorname{div} \mathbf{u} \equiv 0, \end{aligned} \tag{17}$$

then their value will stay zero for all times. For div-preserving advection in case of $\mathbf{u}$ being a magnetic flux this is the physically relevant case. If we additionally assume a constant advection velocity

$$\operatorname{grad} \mathbf{v} = 0, \tag{18}$$

all evolution equations (2), $(4)_{\text{grad}}$, and $(4)_{\text{curl}}$ are reduced to the form

$$\partial_t \mathbf{u} + \mathbf{v} \cdot \operatorname{grad} \mathbf{u} = 0. \tag{19}$$

Hence, all types of advection become indistinguishable from constant advection.

If, still under the assumption of vanishing constraints, the velocity field is purely rotational, we have

$$\operatorname{grad} \mathbf{v} = -(\operatorname{grad} \mathbf{v})^T. \tag{20}$$

In such a case the advection given in (4) differs from ordinary advection. However, if we consider the 2-norm of $\mathbf{u}$, we obtain the ordinary advection equation

$$\partial_t \|\mathbf{u}\|^2 + \mathbf{v} \cdot \operatorname{grad} \|\mathbf{u}\|^2 = 0 \tag{21}$$

for both curl-preserving and div-preserving advection. Hence, $\|\mathbf{u}\|^2$ is rotated as scalar quantity. The components of $\mathbf{u}$, however, are not advected as scalar quantities. In fact, the vector $\mathbf{u}$ is advected as a whole preserving its position relative to the rotating velocity.

2.5. Two-dimensional equations. The numerical methods in the next sections are mainly developed for the two-dimensional case, that is, $\partial_z \to 0$. We proceed to display the two-dimensional equations.

For the div-preserving advection $(4)_{\text{curl}}$ the equation for the component $u^{(z)}$ decouples from the first two equations for $(u^{(x)}, u^{(y)})$. Furthermore, the constraint $\operatorname{div} \mathbf{u}$ is no longer influenced by $u^{(z)}$. Hence, we will discard the equation for $u^{(z)}$ in the following. The remaining equations are given by

$$\begin{aligned} &\partial_t u^{(x)} + \partial_y (u^{(x)} v^{(y)} - v^{(x)} u^{(y)}) = 0, \\ &\partial_t u^{(y)} - \partial_x (u^{(x)} v^{(y)} - v^{(x)} u^{(y)}) = 0 \end{aligned} \tag{22}$$

for the components $(u^{(x)}, u^{(y)})$. Note that $u^{(z)}$ is not zero, nor is its evolution trivial. The component $u^{(z)}$ and its evolution simply does not play a role in the following construction of div-preserving methods.

In the two-dimensional case of $(4)_{\text{grad}}$ it follows $u^{(z)} = \text{const}$ in time; however, the equations for $(u^{(x)}, u^{(y)})$ still depend on $u^{(z)}$. In the most important application of curl-preserving advection—the shallow water system—the additional condition $v^{(z)} = 0$ holds, which yields decoupled equations. Having in mind this kind of application, the remaining equations for the components $(u^{(x)}, u^{(y)})$ are written as

$$\begin{aligned} &\partial_t u^{(x)} + \partial_x (u^{(x)} v^{(x)} + v^{(y)} u^{(y)}) = 0, \\ &\partial_t u^{(y)} + \partial_y (u^{(x)} v^{(x)} + v^{(y)} u^{(y)}) = 0 \end{aligned} \tag{23}$$

for the two-dimensional version of $(4)_{\text{grad}}$. As in the case of (22) the component $u^{(z)}$ is not further considered.

Correspondingly the constraints are written

$$\begin{aligned} &\text{for } (4)_{\text{grad}} \to \quad && \partial_x u^{(y)} - \partial_y u^{(x)} = \text{const}, \\ &\text{for } (4)_{\text{curl}} \to \quad && \partial_x u^{(x)} + \partial_y u^{(y)} = \text{const} \end{aligned} \tag{24}$$

in two dimensions. Note that the constraint of type $(4)_{\text{grad}}$, which was a vectorial quantity in (5), became a scalar equation.

The dual behavior of curl-preserving and div-preserving advection already observable in the previous section becomes perfect in the two-dimensional case. By substituting the two-dimensional vector $\mathbf{u}$ by its orthogonal complement

$$\begin{pmatrix} u^{(x)} \\ u^{(y)} \end{pmatrix} \leftrightarrow \begin{pmatrix} u^{(y)} \\ -u^{(x)} \end{pmatrix} \quad \Longleftrightarrow \quad (23) \leftrightarrow (22), \tag{25}$$

we can transform curl-preserving and div-preserving advection into each other. Hence, in what follows, any statement or numerical method for the system (23) can be transformed into a statement or numerical method for (22) with the same properties and vice versa.

3. General framework. A numerical solution of equations like curl-preserving and div-preserving advection should respect the intrinsic constraints in a numerical way. That is, a discrete version of the constraint should follow directly from the numerical discretization of the evolutions. Ordinary numerical schemes, however, do not care about the constraints which leads to well-known problems, e.g., in calculating magnetohydrodynamical flows [6]. We propose that a numerical scheme has to be constructed on the basis given by a discretization of the constraints. Since the equations of interest are hyperbolic with local domain of dependency we expect that the constraint can be controlled locally as well. In this section we set up a general framework for locally constrained transport schemes.

We consider $\mathbf{u} \in \Omega \subseteq \mathbb{R}^D$ (D: space-dimension) and a generic evolution

$$\partial_t \mathbf{u} + \mathcal{F}(\mathbf{u}; \mathbf{x}) = 0 \tag{26}$$

with a transport operator $\mathcal{F}$ depending explicitly on the space variable $\mathbf{x}$. The generic constraint $\mathcal{C}$ is assumed to be linear and intrinsic for (26), that is, the relation

$$\mathcal{C}(\mathcal{F}(\mathbf{u}; \mathbf{x})) \equiv 0 \tag{27}$$

holds independently of $\mathbf{u}$ and $\mathbf{x}$, which directly implies

$$\mathcal{C}(\mathbf{u}) = \text{const in time} \tag{28}$$

for any solution of (26); see also [4]. The computational domain Ω is covered by a grid $\mathcal{T} = \{K_i\}_{i=1,2,\ldots}$ with nonoverlapping polygonal cells K_i whose inner diameter is bounded by h. Two cells are called neighbors if they have a common edge *or* if they only share a vertex. The set $\mathcal{N}(K)$ gives all neighbors of the cell K. A time discretization by Δt leads to a cell-wise constant grid function $\tilde{\mathbf{u}}^m : \mathcal{T} \to \mathbb{R}^D$ which approximates $\mathbf{u}$ after m time steps by cell mean values.

3.1. Flux distribution formulation. The central quantity of this paper is the so-called flux distribution. It will be the structure of the flux distribution that determines whether a certain scheme is *constraint-preserving*.

DEFINITION 3.1 (flux distribution). *Given the space of vector-valued grid functions denoted by* $V = \{g : \mathcal{T} \to \mathbb{R}^D\}$, *we define a "flux distribution"* $\Phi_K : V \to V$ *which is attached to a grid cell and maps the grid function* $\tilde{\mathbf{u}}$ *into another grid function, that is,*

$$\Phi_K(\tilde{\mathbf{u}}) : \mathcal{T} \to \mathbb{R}^D \tag{29}$$

with $\operatorname{supp}(\Phi_K(\tilde{\mathbf{u}})) = K \cup \bigcup_{\hat{K} \in \mathcal{N}(K)} \hat{K}$. *The evaluation* $\Phi_K(\tilde{\mathbf{u}})|_{\hat{K}}$ *gives the change of* $\tilde{\mathbf{u}}$ *at cell* $\hat{K}$ *caused by cell* K *during a time step, that is, the flux.*

A flux distribution is assigned to each cell of the grid and may depend on the value of $\tilde{\mathbf{u}}$ in this particular cell but also on that in other cells. The definition is more general than that of usual intercell fluxes, since it admits fluxes to any neighboring cell, especially across corners. This incorporates multidimensionality from the very beginning. Conservation of $\tilde{\mathbf{u}}$ may be expressed by the statement that the integral of $\Phi_K(\tilde{\mathbf{u}})$ vanishes.

A certain form of the flux distribution and its dependency on $\tilde{\mathbf{u}}$ is usually constructed from consistency with the transport equation. Once the flux distribution is defined, an explicit evolution scheme follows by simply collecting contributions of all

flux distributions. Written in complete grid functions we have

$$\tilde{\mathbf{u}}^{m+1} = \tilde{\mathbf{u}}^m + \sum\nolimits_{\hat{K}} \Phi_{\hat{K}}\left(\tilde{\mathbf{u}}^m\right) \tag{30}$$

as an update for the complete grid. The restriction to a certain cell yields a local formulation, viz.,

$$\tilde{\mathbf{u}}^{m+1}|_K = \tilde{\mathbf{u}}^m\big|_K + \sum_{\hat{K}\in\{K\}\cup\mathcal{N}(K)} \Phi_{\hat{K}}\left(\tilde{\mathbf{u}}^m\right)|_K. \tag{31}$$

Here the value of $\tilde{\mathbf{u}}$ in a cell K is updated by contributions of all neighboring cells which are given by evaluations of flux distributions. Note that virtually any finite-volume scheme can be written in the form (31), and the flux distribution may then be identified.

To approximate the transport equation given in (26) consistency is required in the form of cell mean values, viz.,

$$\mathcal{F}\left(\mathbf{u};\mathbf{x}\right)|_K = -\lim_{\Delta t,h\to 0} \frac{1}{\Delta t} \sum_{\hat{K}\in\{K\}\cup\mathcal{N}(K)} \Phi_{\hat{K}}\left(\tilde{\mathbf{u}}\right)|_K, \tag{32}$$

where $\tilde{\mathbf{u}}$ is the projection of $\mathbf{u}$ onto the grid.

In order to clarify the notion of a flux distribution, we briefly present a possible flux distribution for one-dimensional constant advection $u_t + a\,u_x = 0$ on a uniform mesh. The flux distribution given by

$$\Phi_i\left(\tilde{u}^m\right) = \begin{cases} \frac{\max(0,a)\Delta t}{h}\tilde{u}_i^m & \text{for cell } i+1, \\ -\left(\frac{\max(0,a)\Delta t}{h} + \frac{\min(0,a)\Delta t}{h}\right)\tilde{u}_i^m & \text{for cell } i, \\ \frac{\min(0,a)\Delta t}{h}\tilde{u}_i^m & \text{for cell } i-1 \end{cases} \tag{33}$$

has entries in cell i and its neighbors $i \pm 1$. The evolution (31) may then be written

$$\tilde{u}_i^{m+1} = \tilde{u}_i^m - \frac{\max\left(0,a\right)\Delta t}{h}(\tilde{u}_i^m - \tilde{u}_{i-1}^m) + \frac{\min\left(0,a\right)\Delta t}{h}(\tilde{u}_{i+1}^m - \tilde{u}_i^m), \tag{34}$$

which represents the donor cell scheme for constant advection.

3.2. Constraint preservation. Since the constraint is linear we expect a discretization which may be written as matrix operation

$$\mathcal{C}\left(\mathbf{u}\right)\big|_K = \tilde{C}_K\,\tilde{\mathbf{u}} + O\left(h^n\right) \tag{35}$$

on the grid function $\tilde{\mathbf{u}}$ which is obtained from the function $\mathbf{u}$ by cell-wise constant projection. If preservation of the constraint should be achieved for the scheme (30), the following quite obvious statement leads the way.

LEMMA 3.2 (constraint preservation). *If the condition*

$$\tilde{C}_K\,\Phi_{\hat{K}}(\tilde{\mathbf{u}}) = 0 \quad \forall K, \hat{K}, \tilde{\mathbf{u}} \tag{36}$$

holds for a specific discrete constraint and a flux distribution, it follows for the evolution scheme given in (30)

$$\tilde{C}_K\,\tilde{\mathbf{u}}^{m+1} = \tilde{C}_K\,\tilde{\mathbf{u}}^m, \tag{37}$$

i.e., the discrete operator is preserved locally by this scheme.

Note that condition (36) is sufficient only for constraint preservation, since the flux distribution is completely unspecified. Contributions of different flux distributions in (30) could interact in such a way that the discrete constraint is preserved even if (36) is not valid. However, we will not consider such schemes.

Since the condition is difficult to control for any grid function $\tilde{\mathbf{u}}$, we assume a decomposition

$$\Phi_K(\tilde{\mathbf{u}}) = \varphi_K(\tilde{\mathbf{u}})\ \hat{\Phi}_K \tag{38}$$

into a factor $\varphi_K(\tilde{\mathbf{u}}) \in \mathbb{R}$ and a skeleton or shape function $\hat{\Phi}_K$. As indicated, only the factor depends on the field $\tilde{\mathbf{u}}$. Due to the linearity of the constraint this factor drops out of the condition in (36) and we obtain

$$\tilde{\mathcal{C}}_K\, \hat{\Phi}_{\hat{K}} = 0 \quad \forall K, \hat{K} \tag{39}$$

as a purely geometric condition. To some extent this is the discrete analogon to (27) which states that the constraint is intrinsic. Indeed, for the case of div-preserving advection the curl in $(4)_{\text{curl}}$ must be discretized in an update $\hat{\Phi}_{\hat{K}}$ such that a discrete divergence $\tilde{\mathcal{C}}_K$ gives exactly zero. This is also the approach in [13], [14], where discrete analogons of vector-analytic relations are considered and used to discretize Maxwell's equations. The work of [13], [14], however, relies on using different locations, i.e., cell-center, face, edge, and vertex, to discretize vector functions and define the differential operators. The operators div and curl are then defined on different grids and for differently stored variables. For a finite-volume approach with exclusive use of cell mean values this is unsatisfactory. The condition in (39) aims at discrete operators and updates that use only cell centered variables. However, at least one of the resulting schemes may be translated in a "mimetic" scheme described in [14] by appropriate averaging. This will be demonstrated at the end of section 4.2.2.

If a generic cell $\hat{K}$ is fixed, (39) gives a homogeneous linear system of equations and the flux distributions are elements of its kernel. The system will be finite if the discrete constraint has a finite stencil since then evaluations of (39) for cells K far off the support of $\Phi_{\hat{K}}$ will vanish identically. The crucial task of designing constrained transport schemes is to find nontrivial solutions of (39) for a given discrete constraint operator. A nontrivial solution of (39) is expected to exist if functions with compact support exist for which the analytic constraint vanishes. The structure of the solutions depends strongly on the discretization used of the constraint operator.

The system (39) for a fixed cell $\hat{K}$ is homogeneous and possesses more equations than unknowns since the evaluation of $\tilde{\mathcal{C}}_K$ on cells neighboring the support of $\hat{\Phi}_{\hat{K}}$ will yield nontrivial expressions. Experience with concrete examples showed that due to symmetry most equations are linear dependent and the entire system has a rank less than the number of unknowns. However, a proof of the general statement that the system (39) always has a rank less than the number of equations is not yet available. We expect a solution space for (39) from which we only consider an appropriate basis set of flux distributions $\{\hat{\Phi}_K^{(g)}\}$ with $g = 1, 2, \ldots$ which are all constraint-preserving. The final flux distribution has to be assembled from these solutions,

$$\Phi_K(\tilde{\mathbf{u}}) = \sum\nolimits_g \varphi_K^{(g)}(\tilde{\mathbf{u}})\, \hat{\Phi}_K^{(g)} \tag{40}$$

with unknown coefficients $\varphi_K^{(g)}$. Note that the choice of $\varphi_K^{(g)}$ does not affect the preservation of the constraint which is already established by the skeletons of the

flux distributions. The expression for Φ_K enters the scheme (31), and the remaining coefficients $\varphi_K^{(g)}$ have to follow from consistency (32) and stability as well.

The local character of the constraint is crucial at this point. If the constraint has a global influence, like the divergence condition in the elliptic Stokes problem, it will not be possible to find a flux distribution which is consistent *and* locally constraint-preserving. In the case of the elliptic Stokes problem either the constraint condition (39) for a consistent flux distribution or the consistency condition (32) for a preserving one would result in a global problem accounting for the ellipticity.

4. Rectangular grid in two dimensions. We proceed with applying the general framework of the preceding section to the system (22), thus concentrating on div-preserving advection. Both equations in (22) are governed by a single flux function which we denote by

$$F(\mathbf{u}, \mathbf{v}) = u^{(x)} v^{(y)} - v^{(x)} u^{(y)}. \tag{41}$$

As indicated in section 2.3, a numerical scheme for (22) can be directly transformed into a scheme for (23) by duality. Further investigations will be presented in the case of a rectangular grid with cells $K = (i, j)$ at positions (x_i, y_j) and size $\Delta x \times \Delta y$. The geometry factor of the grid $\alpha = \frac{\Delta x}{\Delta y}$ shall be bounded from above and below. In cases of accuracy considerations we refer to $h = \max(\Delta x, \Delta y)$.

Note that the general framework is valid for any kind of polygonal grid. However, the construction and investigation of discrete constraint operators on triangular or quadrilateral grids become complicated. The extension to more general grids is subject to future work.

4.1. Discrete constraints. Since the discrete version of the constraint operator influences the structure of the flux distribution, we present a certain class of discrete divergence operators. Each operator is obtained from a discretization of the first derivative in a finite difference approach. We require a symmetric 3×3 stencil and an approximation of second order. The following lemma gives all possible approximations of this type.

LEMMA 4.1 (discrete first derivatives). *Any second order approximation of the first derivative of a smooth function ψ in the center cell (i, j) of a symmetric 3×3 Cartesian stencil has the form*

$$\left.\frac{\partial \psi}{\partial x}\right|_{i,j} = \tilde{\mathcal{D}}_{i,j}(\alpha, \beta, \gamma)\ \tilde{\psi} + O\left(h^2\right) \tag{42}$$

with arbitrary values for α, β, γ and

$$\begin{aligned}\tilde{\mathcal{D}}_{i,j}(\alpha, \beta, \gamma) =& \frac{1}{2\Delta x} \begin{array}{|c|c|c|}\hline 0 & 0 & 0 \\ \hline -1 & 0 & 1 \\ \hline 0 & 0 & 0 \\ \hline \end{array} + \frac{\alpha}{\Delta x} \begin{array}{|c|c|c|}\hline -1 & 0 & 1 \\ \hline 2 & 0 & -2 \\ \hline -1 & 0 & 1 \\ \hline \end{array} \\ &+ \frac{\beta}{\Delta x} \begin{array}{|c|c|c|}\hline 1 & -2 & 1 \\ \hline 0 & 0 & 0 \\ \hline -1 & 2 & -1 \\ \hline \end{array} + \frac{\gamma}{\Delta x} \begin{array}{|c|c|c|}\hline 1 & -2 & 1 \\ \hline -2 & 4 & -2 \\ \hline 1 & -2 & 1 \\ \hline \end{array}\end{aligned} \tag{43}$$

as weights in the grid. In the tables the middle cell corresponds to the cell (i, j).

Proof. In the case $\alpha = \beta = \gamma = 0$ the operator (43) reduces to the classical symmetric finite differences which are visible in the first block of (43). Assuming sufficient smoothness of ψ, this gives a second order approximation to the first derivative of ψ. We need only to show that the additional blocks in (43) do contribute only terms of $O\left(h^2\right)$. Indeed, since these blocks are discretizations of higher order cross derivatives, evaluation yields

$$\begin{aligned}\tilde{\mathcal{D}}\left(\alpha,\beta,\gamma\right)\,\psi &= \frac{\partial\psi}{\partial x} + O\left(\Delta x^2\right) + \frac{\alpha}{\Delta x}\left(2\Delta x\Delta y^2\frac{\partial^3\psi}{\partial x\partial y^2} + O\left(h^3\right)\right)\\ &\quad + \frac{\beta}{\Delta x}\left(2\Delta x^2\Delta y\frac{\partial^3\psi}{\partial x^2\partial y} + O\left(h^3\right)\right)\\ &\quad + \frac{\gamma}{\Delta x}\left(\Delta x^2\Delta y^2\frac{\partial^4\psi}{\partial x^2\partial y^2} + O\left(h^4\right)\right)\\ &= \frac{\partial\psi}{\partial x} + O\left(h^2\right).\end{aligned}$$

Finally, we observe that it is not possible to include more discrete cross derivatives. They would be built from at least third order derivatives with respect to x or y which have no representation with a 3×3 stencil. ∎

In the following, we will evaluate discrete derivatives via (43) by using cell mean values instead of point values, which introduces only an error of second order.

The constraint for $(4)_{\text{curl}}$ in (24) is now to be replaced with a discrete formulation. By the lemma on discrete first derivatives given above a discrete divergence operator $\tilde{\mathcal{C}}_{i,j}$ has the form

$$\begin{aligned}\operatorname{div}\mathbf{u}|_{i,j} &= \tilde{\mathcal{D}}_{i,j}\left(\alpha,\beta,\gamma\right)\,\tilde{u}^{(x)} - \tilde{\mathcal{D}}^T_{i,j}\left(\alpha,\beta,\gamma\right)\,\tilde{u}^{(y)} + O\left(h^2\right)\\ &\equiv \tilde{\mathcal{C}}_{i,j}\left(\alpha,\beta,\gamma\right)\,\tilde{\mathbf{u}} + O\left(h^2\right),\end{aligned} \tag{44}$$

which leads to a three parameter family of operators. We mention that if any operator taken from (44) by specifying α, β, γ vanishes, all other operators give a result of $O\left(h^2\right)$ for smooth functions. Hence, if a numerical scheme respects one operator exactly, all others will give only a discretization error. Even in the discontinuous case the control of a single operator is sufficient to avoid nonphysical solutions. See, e.g., [5], where a staggered operator is controlled.

4.2. Flux distributions. In order to derive divergence-preserving flux distributions we have to look for nontrivial solutions of (39) for a specific operator chosen from (44). The flux distribution skeleton $\hat{\Phi}_{i,j}$ in the two-dimensional rectangular case covers a region of 3×3 cells and gives a two-vector in each cell, thus consisting of $2\times 9 = 18$ unknown entries. If we fix $\hat{K}$ in (39) and evaluate the divergence operator around the flux distribution in the cells of a 5×5 area with center $\hat{K}$, we obtain a system of 25 equations which describes the skeleton entries. The divergence of more remote cells is not influenced by the flux distribution at cell $\hat{K}$ and need not to be considered. The resulting system, of course, depends on the chosen operator.

4.2.1. Classical operator. The classical discrete divergence operator $\tilde{\mathcal{C}}^{(0)}_{i,j}$ is obtained from (44) by setting $\alpha = \beta = \gamma = 0$, which leads to

$$\tilde{\mathcal{C}}^{(0)}_{i,j}\,\tilde{\mathbf{u}} = \tilde{\mathcal{C}}_{i,j}\left(0,0,0\right)\,\tilde{\mathbf{u}} = \frac{\tilde{u}^{(x)}_{i+1,j} - \tilde{u}^{(x)}_{i-1,j}}{2\Delta x} + \frac{\tilde{u}^{(y)}_{i,j+1} - \tilde{u}^{(y)}_{i,j-1}}{2\Delta y}. \tag{45}$$

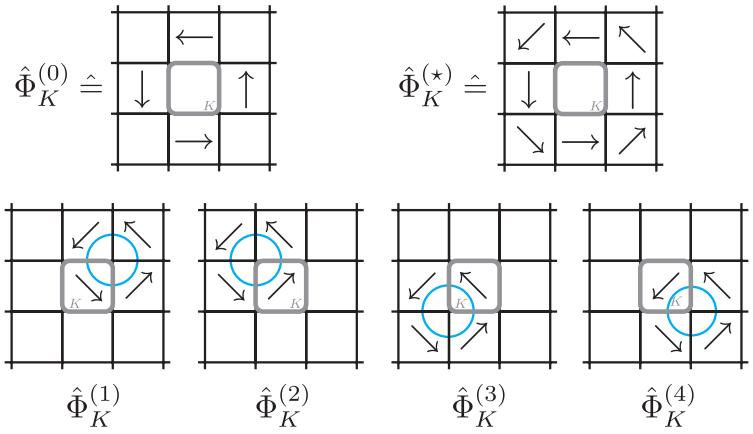


FIG. 1. *Several shape functions for flux distributions that are divergence-preserving. All of them share the same physical interpretation: They have to approximate closed curves in order to avoid introducing sources into the vector field* $\mathbf{u}$ *and thus they preserve the divergence. In terms of differential forms, these flux distributions are minimal co-cycles of the discrete outer derivative.*

In general, this operator represents the best second order approximation to $\operatorname{div} \mathbf{u}$ in the sense that the constant hidden in $O\left(h^2\right)$, i.e., the residual of the second order Taylor expansion, is minimal. The linear system (39) for this operator has rank 17, thus it has a one-dimensional null space. We choose a representative of this kernel and denote it by $\hat{\Phi}_{i,j}^{(0)}$. The nonvanishing entries are given by

$$\begin{aligned} \hat{\Phi}_{i,j}^{(0)}\Big|_{i+1,j} &= (0, \Delta y), \quad & \hat{\Phi}_{i,j}^{(0)}\Big|_{i,j+1} &= (-\Delta x, 0), \\ \hat{\Phi}_{i,j}^{(0)}\Big|_{i,j-1} &= (\Delta x, 0), \quad & \hat{\Phi}_{i,j}^{(0)}\Big|_{i-1,j} &= (0, -\Delta y), \end{aligned} \tag{46}$$

and all other elements of the kernel follow by multiplication with a constant factor. This flux distribution is sketched in the upper left corner of Figure 1. The picture has to be interpreted as follows: A flux originating in K may change the value of $\tilde{\mathbf{u}}$ in the right neighboring cell only in the y-direction. Furthermore, if this neighbor is changed in that way, all other neighboring cells have to be changed correspondingly as depicted in the figure in order to obey the constraint. Of course, the value $\tilde{u}^{(x)}$ in the right neighbor of K does not remain constant, since it may be changed by flux distributions originating from other cells. This kind of flux distribution results in a coupling of fluxes into neighboring cells, and it is this coupling that is responsible for the local preservation of the divergence.

The divergence-preserving flux distributions also have interpretations in the theory of differential forms. They represent minimal discrete co-cycles of the corresponding discrete outer derivatives; see [7], [15], [24].

In order to construct the final scheme we assemble the shape function $\hat{\Phi}_{i,j}$ of the flux distribution according to

$$\Phi_{i,j}^{(0)}\left(\tilde{\mathbf{u}}\right) = \varphi_{i,j}^{(0)}(\tilde{\mathbf{u}})\, \hat{\Phi}_{i,j}^{(0)} \tag{47}$$

with an unknown function φ. Note that $\hat{\Phi}_{i,j} = O(h)$ and, since $\Phi_{i,j}$ has to be $O(1)$, it follows that $\varphi_{i,j} = O(h^{-1})$. The final scheme is obtained by following (31) with (46)

and reads as

$$\begin{pmatrix} \tilde{u}^{(x)} \\ \tilde{u}^{(y)} \end{pmatrix}_{i,j}^{m+1} = \begin{pmatrix} \tilde{u}^{(x)} \\ \tilde{u}^{(y)} \end{pmatrix}_{i,j}^{m} + \begin{pmatrix} (\varphi_{i,j-1}^{(0)}(\tilde{\mathbf{u}}) - \varphi_{i,j+1}^{(0)}(\tilde{\mathbf{u}}))\Delta x \\ (\varphi_{i+1,j}^{(0)}(\tilde{\mathbf{u}}) - \varphi_{i-1,j}^{(0)}(\tilde{\mathbf{u}}))\Delta y \end{pmatrix}^{m}. \tag{48}$$

By Taylor expansion and comparison with the original equation (22), we deduce

$$\varphi_{i,j}^{(0)}(\tilde{\mathbf{u}}) = -\frac{\Delta t}{2\Delta x \Delta y} F\left(\tilde{\mathbf{u}}_{i,j}, \mathbf{v}_{i,j}\right), \tag{49}$$

which makes (48) consistent up to second order. This scheme solves for div-preserving advection while *exactly* preserving the value of the classical divergence operator (45). The scheme introduces central differences for the derivatives in (22) and was proposed ad hoc by Toth [26] in a magnetohydrodynamic setting. Since $\hat{\Phi}_{i,j}^{(0)}$ is the only flux distribution respecting condition (39) with the classical operator, we conclude that this scheme is the only second order scheme which preserves the divergence via the classical operator (45).

However, the scheme (48) is unconditionally unstable due to central differences. For the investigation of stability we have to look for the maximal spectral radius of the amplifier matrix

$$\rho_{\max} = \max_{\xi,\eta \in (-\pi.\pi)} \rho\left(\mathbf{T}_{\xi,\eta}\right) \tag{50}$$

(see section 4.3.2 for more details). Assuming a constant advection velocity and defining the Courant numbers

$$a = \frac{\Delta t\, v^{(x)}}{\Delta x}, \quad b = \frac{\Delta t\, v^{(y)}}{\Delta y} \tag{51}$$

we obtain the result

$$\rho_{\max}^{(0)} = \max_{\xi,\eta \in (-\pi,\pi)} |1 - \mathbf{i}\,(a \sin\xi + b \sin\eta)| > 1 \tag{52}$$

unless $a = b = 0$ for the case of (48). The imaginary unit is denoted by $\mathbf{i} = \sqrt{-1}$. In spite of this instability the scheme (48) could be used in [26] in the context of magnetohydrodynamics due to the use of predictor values for $\mathbf{u}$.

4.2.2. Extended operator. In order to design a more flexible scheme we look for a different divergence operator which leads to a larger null space of (39) and thus provides more nontrivial solutions. Empirical evaluations of the family given in (43) by computer algebra software suggest that the choice $\alpha = \frac{1}{8}$, $\gamma = 0$ admits a four-dimensional kernel for any value of β; i.e., the system (39) has rank 14. The best approximation is then given by $\beta = 0$, and the resulting operator is called *extended* operator $\tilde{\mathcal{C}}_{i,j}^{(\star)}$. Operators with larger kernels could not be found. The extended operator is defined by

$$\begin{aligned} \tilde{\mathcal{C}}_{i,j}^{(\star)}\,\tilde{\mathbf{u}} &= \tilde{\mathcal{C}}_{i,j}\left(\tfrac{1}{8}, 0, 0\right)\,\tilde{\mathbf{u}} \\ &= \frac{\{\tilde{u}_{i+1,j}^{(x)}\}_y - \{\tilde{u}_{i-1,j}^{(x)}\}_y}{2\Delta x} + \frac{\{\tilde{u}_{i,j+1}^{(y)}\}_x - \{\tilde{u}_{i,j-1}^{(y)}\}_x}{2\Delta y}, \end{aligned} \tag{53}$$

where curled brackets stand for

$$\begin{aligned} \{\psi_{i,j}\}_y &= \tfrac{1}{4}\left(\psi_{i,j+1} + 2\psi_{i,j} + \psi_{i,j-1}\right), \\ \{\psi_{i,j}\}_x &= \tfrac{1}{4}\left(\psi_{i+1,j} + 2\psi_{i,j} + \psi_{i-1,j}\right), \end{aligned} \tag{54}$$

i.e., averaging in x- and y-direction. The four admissible skeletons of flux distributions are displayed in the lower row of Figure 1. In detail, the nonvanishing entries of the first one are given by

$$\begin{aligned} \hat{\Phi}^{(1)}_{i,j}\Big|_{i+1,j+1} &= (-\Delta x, \Delta y), \qquad \hat{\Phi}^{(1)}_{i,j}\Big|_{i,j+1} = (-\Delta x, -\Delta y), \\ \hat{\Phi}^{(1)}_{i,j}\Big|_{i,j} &= (\Delta x, -\Delta y), \qquad\qquad \hat{\Phi}^{(1)}_{i,j}\Big|_{i+1,j} = (\Delta x, \Delta y) \end{aligned} \tag{55}$$

and the remaining three flux distributions follow by translation. Note that the classical operator applied to these flux distributions will not vanish. We remark further that any scheme built upon these flux distribution skeletons will be conservative, since the cell-wise sum of all flux distribution components, i.e., the integral, gives zero.

As first choice for a flux distribution, we choose the symmetric distribution $\hat{\Phi}^{(\star)}_{i,j}$ which is given by

$$\hat{\Phi}^{(\star)}_{i,j} = \hat{\Phi}^{(1)}_{i,j} + \hat{\Phi}^{(2)}_{i,j} + \hat{\Phi}^{(3)}_{i,j} + \hat{\Phi}^{(4)}_{i,j} \tag{56}$$

and shown in the upper right corner of Figure 1. Like in the preceding section this flux distribution is assembled with an unknown function φ to give

$$\Phi^{(\star)}_{i,j}\left(\tilde{\mathbf{u}}\right) = \varphi^{(\star)}_{i,j}(\tilde{\mathbf{u}})\,\hat{\Phi}^{(\star)}_{i,j}. \tag{57}$$

For the resulting scheme we obtain

$$\begin{pmatrix} \tilde{u}^{(x)} \\ \tilde{u}^{(y)} \end{pmatrix}^{m+1}_{i,j} = \begin{pmatrix} \tilde{u}^{(x)} \\ \tilde{u}^{(y)} \end{pmatrix}^{m}_{i,j} + \begin{pmatrix} (\{\varphi^{(\star)}_{i,j-1}(\tilde{\mathbf{u}})\}_x - \{\varphi^{(\star)}_{i,j+1}(\tilde{\mathbf{u}})\}_x)\Delta x \\ (\{\varphi^{(\star)}_{i+1,j}(\tilde{\mathbf{u}})\}_y - \{\varphi^{(\star)}_{i-1,j}(\tilde{\mathbf{u}})\}_y)\Delta y \end{pmatrix}^{m}, \tag{58}$$

where again the curled brackets denote the averaging of (54). The demand for second order consistency with (22) leads to

$$\varphi^{(\star)}_{i,j}(\tilde{\mathbf{u}}) = -\frac{\Delta t}{2\Delta x \Delta y} F\left(\tilde{\mathbf{u}}_{i,j}, \mathbf{v}_{i,j}\right). \tag{59}$$

This scheme *exactly* preserves the value of the extended divergence operator (53). However, as in the case of the $\Phi^{(0)}$-scheme, this scheme is unconditionally unstable. For the maximal spectral radius of the amplifier matrix we calculate with constant advection and Courant numbers from (51)

$$\rho^{(\star)}_{\max} = \max_{\xi,\eta\in(-\pi,\pi)} \left|1 - \mathbf{i}\left(a\sin\xi\tfrac{1+\cos\eta}{2} + b\sin\eta\tfrac{1+\cos\xi}{2}\right)\right| \geq 1, \tag{60}$$

where equality holds only if $a = b = 0$. The instability of schemes (58) as well as (48) could also be observed in our numerical experiments.

Equivalence with staggered approach. In the context of magnetohydrodynamics one approach of controlling the divergence is to store the components of $\mathbf{u}$ in the edges of the cells, the so-called staggered grid. This idea was proposed in [8] and

further developed by [5], [1]. We follow the presentation in [26]. In our notation the resulting staggered grid scheme reads

$$\begin{pmatrix} \tilde{u}^{(x)}_{i+\frac{1}{2},j} \\ \tilde{u}^{(y)}_{i,j+\frac{1}{2}} \end{pmatrix}^{m+1} = \begin{pmatrix} \tilde{u}^{(x)}_{i+\frac{1}{2},j} \\ \tilde{u}^{(y)}_{i,j+\frac{1}{2}} \end{pmatrix}^{m} + \begin{pmatrix} \frac{\Delta t}{\Delta y}(F_{i+\frac{1}{2},j+\frac{1}{2}} - F_{i+\frac{1}{2},j-\frac{1}{2}}) \\ \frac{\Delta t}{\Delta x}(F_{i+\frac{1}{2},j+\frac{1}{2}} - F_{i-\frac{1}{2},j+\frac{1}{2}}) \end{pmatrix}^{m} \tag{61}$$

for the normal components $\tilde{u}^{(x)}_{i+\frac{1}{2},j}$ and $\tilde{u}^{(y)}_{i,j+\frac{1}{2}}$ on each edge and the function $F_{i+\frac{1}{2},j+\frac{1}{2}}$ evaluated at the vertices. This scheme corresponds to the so-called mimetic discretization of [13], [14] for the equation $(4)_{\mathrm{curl}}$ if it is applied in the case of a rectangular grid.

In the context of computational magnetohydrodynamics and in this paper all variables are represented as mean values located in the cell centers. Hence, the staggered variables in (61) have to be substituted; see [26]. The flux function at the vertices is obtained by averaging

$$F_{i+\frac{1}{2},j+\frac{1}{2}} = \frac{1}{4}\left(F_{i,j} + F_{i+1,j} + F_{i,j+1} + F_{j+1,i+1}\right), \tag{62}$$

where the expression $F_{i,j}$ corresponds to the evaluation of the flux function (41) in the cell (i,j). Finally, the edge values of $u^{(x)}$ and $u^{(y)}$ are averaged via

$$\tilde{u}^{(x)}_{i,j} = \frac{1}{2}\left(\tilde{u}^{(x)}_{i+\frac{1}{2},j} + \tilde{u}^{(x)}_{i-\frac{1}{2},j}\right), \quad \tilde{u}^{(y)}_{i,j} = \frac{1}{2}\left(\tilde{u}^{(y)}_{i,j+\frac{1}{2}} + \tilde{u}^{(y)}_{i,j-\frac{1}{2}}\right) \tag{63}$$

after each time step (61). In [26] it was noted that this averaging procedure can be done explicitly with the scheme (61), which leads to a scheme where staggered values of $\mathbf{u}$ are eliminated. The resulting scheme (formerly staggered) is equivalent to the symmetric $\Phi^{(\star)}$-scheme in (58).

Furthermore, this shows that the extended operator $\tilde{\mathcal{C}}^{(\star)}_{i,j}$ is exactly preserved on the primary grid cells in a staggered grid calculation. This also suggests a close relation between the present extended divergence operator $\tilde{\mathcal{C}}^{(\star)}_{i,j}$ and the DIV-operator which is preserved in the mimetic schemes of [14]. The DIV-operator for the normal components on the edges is written

$$DIV\,\mathbf{u}|_{i,j} = \frac{\tilde{u}^{(x)}_{i+\frac{1}{2},j} - \tilde{u}^{(x)}_{i-\frac{1}{2},j}}{\Delta x} - \frac{\tilde{u}^{(y)}_{i,j+\frac{1}{2}} - \tilde{u}^{(y)}_{i,j-\frac{1}{2}}}{\Delta y} \tag{64}$$

in the rectangular case; see [13]. This is exactly the $\tilde{\mathcal{C}}^{(\star)}_{i,j}$-operator if the edge values are obtained from the cell centered variables by averaging

$$\tilde{u}^{(x)}_{i+\frac{1}{2},j} = \frac{\tilde{u}^{(x)}_{i+1,j+1} + 2\tilde{u}^{(x)}_{i+1,j} + \tilde{u}^{(x)}_{i+1,j-1} + \tilde{u}^{(x)}_{i,j+1} + 2\tilde{u}^{(x)}_{i,j} + \tilde{u}^{(x)}_{i,j-1}}{8}, \tag{65}$$

and analogously for $\tilde{u}^{(y)}_{i,j+\frac{1}{2}}$. This is also the same averaging formula which is used in [19] to switch from cell centered variables to normal edge components.

The symmetric scheme in (58) as well as the staggered approach is unstable, since these schemes do not take upwind directions into account. As in the case of (48) the staggered grid scheme is stabilized in magnetohydrodynamics calculations by use of predictors; see [1], [5], and [26].

4.3. Upwind scheme. The symmetric flux distribution (56) uses the same coefficient $\varphi^{(\star)}$ for all basis elements $\Phi^{(g)}$. This results in central differences and instability of the final scheme. To construct an upwind scheme we propose

$$\Phi^{(\mathrm{up})}_{i,j}\left(\tilde{\mathbf{u}}\right) = \sum_{g=1}^{4} \varphi^{(g)}_{i,j}(\tilde{\mathbf{u}})\, \hat{\Phi}^{(g)}_{i,j} \tag{66}$$

as flux distribution with four coefficients $\varphi^{(g)}$ yet to be specified. The final scheme reads as

(67)
$$\begin{aligned}\begin{pmatrix}\tilde{u}^{(x)}\\ \tilde{u}^{(y)}\end{pmatrix}^{m+1}_{i,j} &= \begin{pmatrix}\tilde{u}^{(x)}\\ \tilde{u}^{(y)}\end{pmatrix}^{m}_{i,j} + \begin{pmatrix}\delta^{(y)}_{i-\frac{1}{2},j-\frac{1}{2}}(\varphi^{(1)})\Delta x\\ -\delta^{(x)}_{i-\frac{1}{2},j-\frac{1}{2}}(\varphi^{(1)})\Delta y\end{pmatrix} + \begin{pmatrix}\delta^{(y)}_{i+\frac{1}{2},j-\frac{1}{2}}(\varphi^{(2)})\Delta x\\ -\delta^{(x)}_{i+\frac{1}{2},j-\frac{1}{2}}(\varphi^{(2)})\Delta y\end{pmatrix}\\ &\quad + \begin{pmatrix}\delta^{(y)}_{i+\frac{1}{2},j+\frac{1}{2}}(\varphi^{(3)})\Delta x\\ -\delta^{(x)}_{i+\frac{1}{2},j+\frac{1}{2}}(\varphi^{(3)})\Delta y\end{pmatrix} + \begin{pmatrix}\delta^{(y)}_{i-\frac{1}{2},j+\frac{1}{2}}(\varphi^{(4)})\Delta x\\ -\delta^{(x)}_{i-\frac{1}{2},j+\frac{1}{2}}(\varphi^{(4)})\Delta y\end{pmatrix},\end{aligned}$$

where we used the abbreviations

$$\begin{aligned}\delta^{(x)}_{i+\frac{1}{2},j+\frac{1}{2}}(\varphi) &= \varphi_{i+1,j+1} + \varphi_{i+1,j} - \varphi_{i,j+1} - \varphi_{i,j},\\ \delta^{(y)}_{i+\frac{1}{2},j+\frac{1}{2}}(\varphi) &= \varphi_{i+1,j+1} + \varphi_{i,j+1} - \varphi_{i+1,j} - \varphi_{i,j},\end{aligned} \tag{68}$$

which represent finite differences. The next two subsections specify the coefficients $\varphi^{(g)}$ by requiring consistency and stability of the scheme (67).

4.3.1. Consistency. The following lemma gives expressions for $\varphi^{(g)}$ such that a first or second order method is obtained. A weight function $\omega^{(g)}$ which controls the influence of the different basis flux distributions remains unspecified.

THEOREM 4.2 (consistency). *Let the values $\omega^{(g)}$ $(g = 1, 2, 3, 4)$ be weights such that $\sum_{g=1}^{4} \omega^{(g)} = 1$. In general these weights depend on $\tilde{\mathbf{u}}$ and $\mathbf{v}$. Furthermore let the expressions $\frac{\Delta t}{\Delta x}$ and $\frac{\Delta t}{\Delta y}$ be $O(1)$, the geometry factor $\alpha = \frac{\Delta x}{\Delta y}$ be bounded from above and below, and $h = \max(\Delta x, \Delta y)$. Then the scheme displayed in* (67) *is consistent with the constrained advection equation* (22) *in smooth regions of the solution up to*

(i) *first order in space and time, if the flux distribution factor in* (66) *is given by*

$$\varphi^{(g)}_{i,j}\left(\tilde{\mathbf{u}}\right) = -\frac{\Delta t}{2\Delta x \Delta y}\omega^{(g)}(\tilde{\mathbf{u}}_{i,j}, \mathbf{v}_{i,j}) F\left(\tilde{\mathbf{u}}_{i,j}, \mathbf{v}_{i,j}\right); \tag{69}$$

(ii) *second order in space and time, if the flux distribution factor in* (66) *is given by*

$$\varphi^{(g)}_{i,j}\left(\tilde{\mathbf{u}}\right) = -\frac{\Delta t}{2\Delta x \Delta y}\omega^{(g)}(\tilde{\mathbf{u}}_{i,j}, \mathbf{v}_{i,j}) \left(F - \frac{\Delta t}{2}(v^{(x)}\partial_x F + v^{(y)}\partial_y F) + \Lambda\right)_{i,j} \tag{70}$$

with $\Lambda = \sum_{g=1}^{4}(\frac{\Delta x}{2}\, r_g\, \partial_x(\omega^{(g)}F) + \frac{\Delta y}{2}\, l_g\, \partial_x(\omega^{(g)}F))$ and coefficients r_g and l_g as given in the table (71).

We remark that the second order result (70) uses derivatives of the weight function ω. Hence, ω considered as a function in the domain Ω needs to have at least one continuous derivative in order to obtain second order accuracy. We will present numerical experiments where second order is not recovered due to a nonsmooth weight function.

Proof. We consider the flux distribution factor $\varphi_{i,j}$ as function $\varphi_{i,j} \equiv \varphi(x_i, y_j)$ whose evaluations in different grid cells can be expanded in a Taylor series. Second order expansion of the scheme (67) gives

$$\begin{aligned}\begin{pmatrix}\tilde{u}^{(x)}\\ \tilde{u}^{(y)}\end{pmatrix}_{i,j}^{m+1} &= \begin{pmatrix}\tilde{u}^{(x)}\\ \tilde{u}^{(y)}\end{pmatrix}_{i,j}^{m}\\ &+ \begin{pmatrix} \frac{\partial}{\partial y}\sum_{g=1}^{4}\left(2\varphi^{(g)} - \Delta x\, r_g\, \partial_x\varphi^{(g)} - \Delta y\, l_g\, \partial_y\varphi^{(g)}\right)\Delta x\Delta y \\ -\frac{\partial}{\partial x}\sum_{g=1}^{4}\left(2\varphi^{(g)} - \Delta y\, l_g\, \partial_y\varphi^{(g)} - \Delta x\, r_g\, \partial_x\varphi^{(g)}\right)\Delta y\Delta x \end{pmatrix}_{i,j}\\ &+ O\left(h^3\right),\end{aligned}$$

where l_g and r_g are defined by the table

(71)

	$g=1$	$g=2$	$g=3$	$g=4$
l_g	1	1	–1	–1
r_g	1	–1	–1	1

and φ_x, φ_y are derivatives of φ. Note that we changed the interpretation of $\tilde{\mathbf{u}}_{i,j}$ from cell mean value to point value in the middle of the cell. This switch introduces only an error of $O(h^2)$ on both the left-hand and the right-hand sides of the equation. However, the leading expression within $O(h^2)$ cancels on both sides and the remaining error is $O(h^3)$.

By use of (22) with the definition of the flux function F in (41) we obtain the expansion of the exact solution

$$\begin{aligned}\mathbf{u}_{i,j}\left(t+\Delta t\right) &= \mathbf{u}_{i,j}\left(t\right) - \Delta t \begin{pmatrix}\partial_y F\\ -\partial_x F\end{pmatrix}_{i,j}\\ &+ \frac{\Delta t^2}{2}\begin{pmatrix}\partial_y\left(v^{(x)}\partial_x F + v^{(y)}\partial_y F\right)\\ -\partial_x\left(v^{(x)}\partial_x F + v^{(y)}\partial_y F\right)\end{pmatrix}_{i,j} + O\left(\Delta t^3\right).\end{aligned}$$

Since the Courant numbers are bounded we have $O(h^3) = O(\Delta t^3)$. Hence, the direct comparison of exact and numerical increments of $\mathbf{u}_{i,j}$ yields the *consistency condition*

$$\begin{aligned}&\sum_{g=1}^{4}\left(2\varphi^{(g)} - \Delta x\, r_g\, \partial_x\varphi^{(g)} - \Delta y\, l_g\, \partial_y\varphi^{(g)}\right)\Delta x\Delta y\\ &\quad = -\Delta t\, F + \frac{\Delta t^2}{2}\left(v^{(x)}\partial_x F + v^{(y)}\partial_y F\right).\end{aligned}$$

This relation can be solved for $\varphi^{(g)}$ by means of an ansatz with a first order and a second order contribution to $\varphi^{(g)}$, viz.,

$$\varphi^{(g)} = \frac{\Delta t}{\Delta x\Delta y}\varphi^{(g,1)} + \frac{\Delta t^2}{\Delta x\Delta y}\varphi^{(g,2)}.$$

Introducing this into the consistency condition and comparison of coefficients of Δt-expressions leads us to

$$\sum_{g=1}^{4}\varphi^{(g,1)} = -\frac{1}{2}F,$$

which, of course, can be satisfied in many ways. We propose weights $\omega^{(g)}$, yet unspecified, which sum up to unity and write

$$\varphi^{(g,1)} = -\omega^{(g)} \frac{1}{2} F$$

as first order flux distribution factor. By using this in our ansatz and further comparison of Δt^2-coefficients in the consistency condition, the second order factor

$$\varphi^{(g,2)} = \omega^{(g)} \frac{1}{4} \left(v^{(x)} \partial_x F + v^{(y)} \partial_y F - \sum_{g=1}^{4} \left(\frac{\Delta x}{\Delta t} r_g \, \partial_x (\omega^{(g)} F) + \frac{\Delta y}{\Delta t} l_g \, \partial_y (\omega^{(g)} F) \right) \right)$$

is obtained. □

In implementations we substitute the x- and y-derivatives in (70) by appropriate finite differences. Following the TVD analysis of one-dimensional methods, these finite differences need a limiting procedure in order to obtain nonoscillatory solutions. In the numerical experiments with discontinuous solutions we used the so-called WENO-limiter (see [17]), which is given by

$$WENO(d_1, d_2) = \frac{\frac{d_1}{\sqrt{d_1^2+\varepsilon}} + \frac{d_2}{\sqrt{d_2^2+\varepsilon}}}{\frac{1}{\sqrt{d_1^2+\varepsilon}} + \frac{1}{\sqrt{d_2^2+\varepsilon}}}, \tag{72}$$

where d_1 and d_2 are left- and right-hand side finite differences and ε is a small number ($\varepsilon \approx 10^{-8}$). With use of this limiter we have

$$\left. \frac{\partial \psi}{\partial x} \right|_i = \frac{WENO(\psi_i - \psi_{i-1}, \psi_{i+1} - \psi_i)}{\Delta x} \tag{73}$$

for the limited derivative of a grid function ψ.

4.3.2. Stability. The weights which have been introduced during the proof of consistency control the activation of the different basis flux distributions shown in Figure 1 (lower row). Their value should be chosen according to the direction of the advection. Clearly, one would not activate the first flux distribution $\hat{\Phi}^{(1)}$ which is oriented towards the upper right if the wind is pointing in opposite direction. This would yield an unstable scheme. Indeed, stability is the issue that will specify the right choice of weights.

To investigate the stability we consider the *one-sided* scheme which uses only the first and the fourth flux distribution. Hence, it follows for the weights

$$\omega^{(1)} \equiv \omega, \quad \omega^{(2,3)} = 0, \quad \omega^{(4)} = 1 - \omega \tag{74}$$

with unknown ω. An impression of how different choices of ω influences the stability of the scheme is given in Figure 2. It shows the contours of the maximal eigenvalue of the amplifier matrix $\rho_{\max}$ for different choices of ω and different directions of the flow. The contour values and their shape have been obtained numerically for the first order scheme. We can see that a flow pointing exactly in the direction of a flux distribution ($\theta_1 = 0$ or $\theta_4 = 0$) requires the activation of only the corresponding flux distribution ($\omega_1 = 1$ or $\omega_4 = 1$) to yield stability. Furthermore $\omega_1 = \omega_4 = 1/2$ gives a stable scheme only for flows in the x-direction. This corresponds to the intuitive choices in these cases. In between these extreme cases Figure 2 indicates the existence of a single stable choice $\bar{\omega}(\theta)$ for the weight.

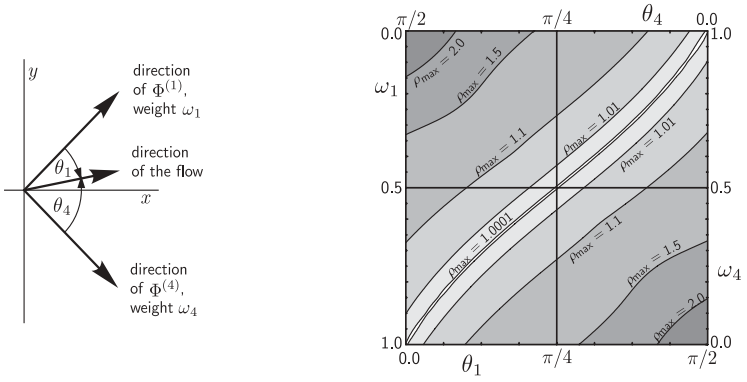


FIG. 2. *Left: A sketch of the elements of the one-sided scheme to clarify the notation used. Right: Numerical evaluations of the maximal spectral radius of the amplifier matrix in the* (θ_1, ω_1)*-plane for the first order one-sided scheme. The plot exhibits a distinct function* $\bar{\omega}(\theta)$ *which yields stability.*

The following lemma specifies this weight and the stability conditions of the one-sided scheme.

THEOREM 4.3 (stability). *Assume the advection velocity to be constant and* $\left|v^{(x)}\right| \neq 0$. *Then, the one-sided, first order scheme consisting of flux distribution* $\Phi^{(1)}$ *and* $\Phi^{(4)}$ *with single weight* ω, *time step* Δt, *and cells* $\Delta x \times \Delta y$ *is stable in the sense of a von Neumann analysis under the conditions*

$$\left(\frac{\Delta t\, v^{(y)}}{\Delta y}\right)^2 \leq \frac{\Delta t\, v^{(x)}}{\Delta x} \leq 1 \quad \textit{and} \quad \omega \equiv \bar{\omega} = \frac{1}{2}\left(1 + \frac{\Delta x}{\Delta y}\frac{v^{(y)}}{v^{(x)}}\right). \tag{75}$$

Under these stability conditions we have, furthermore, for the maximal spectral radius $\rho_{\max}$ *of the amplifier matrix*

$$\rho_{\max}(\bar{\omega} + \delta\omega) = 1 + c\,\delta\omega^2 + O(\delta\omega^3) \quad \textit{with } c > 0, \tag{76}$$

i.e., the weight $\bar{\omega}$ *is a local minimum of* $\rho_{\max}$.

Proof. We follow the stability analysis of von Neumann (see, e.g., [11]). The Fourier transform of the grid function $\mathbf{u}_{i,j}$ is denoted by

$$\hat{\mathbf{u}}_{i,j} = \hat{\mathbf{u}}_0\, e^{\mathbf{i}(i\,\xi + j\,\eta)}$$

and introduced into the scheme (67) with (69) and (74), which leads to

$$\hat{\mathbf{u}}_0^{m+1} = \mathbf{T}_{\xi,\eta}\,\hat{\mathbf{u}}_0^m.$$

Here, $\mathbf{T}_{\xi,\eta}$ is the amplifier matrix of the scheme. The imaginary unit is denoted by $\mathbf{i} = \sqrt{-1}$. Since the advection velocity is assumed to be constant the amplifier matrix has the form

$$\mathbf{T}_{\xi,\eta} = \begin{pmatrix} 1 - b\,t^{(x)}(\xi,\eta,\omega) & a\,\alpha\,t^{(x)}(\xi,\eta,\omega) \\ -b\,\frac{1}{\alpha}\,t^{(y)}(\xi,\eta,\omega) & 1 + a\,t^{(y)}(\xi,\eta,\omega) \end{pmatrix}$$

with the Courant numbers $a = \frac{\Delta t\, v^{(x)}}{\Delta x}$ and $b = \frac{\Delta t\, v^{(y)}}{\Delta y}$ as well as $\alpha = \frac{\Delta x}{\Delta y}$. The functions $t^{(x)}$ and $t^{(y)}$ depend on ξ, η, and ω and follow from the scheme. For stability

the maximal spectral radius

$$\rho_{\max}(a,b,\omega) = \max_{\xi,\eta\in(-\pi,\pi)} \rho\left(\mathbf{T}_{\xi,\eta}\right)$$

has to be smaller or equal to unity. The geometry factor α drops out during the calculation. Obviously, $\mathbf{T}_{\xi,\eta}$ has the eigenvector $(v^{(x)}, v^{(y)})^T$ with eigenvalue $\lambda_1 = 1$, which corresponds to the first eigenvalue and eigenvector of the Jacobian given in (16) if the identity matrix is added. The second eigenvalue of $\mathbf{T}_{\xi,\eta}$ varies with ξ and η. It has the form $\lambda_2 = \tau_1(\eta) + \tau_2(\eta)\, e^{\mathbf{i}\,\xi}$ with $\tau_{1,2}(\eta) \in \mathbb{R}$, thus its maximal absolute value $|\lambda_2| = |\tau_1(\eta)| + |\tau_2(\eta)|$ depends only on η. A straightforward calculation leads to

$$\rho_{\max}(a,\beta,\omega) = \max_{\eta\in(-\pi,\pi)} \left(\sqrt{\left(1 - a + a(1+\beta-2\,\omega\,\beta)\tfrac{1-\cos\eta}{2}\right)^2 + \tfrac{a^2}{4}\left(1+\beta-2\,\omega\right)^2 \sin^2\eta} + \sqrt{\left(a - a(1-\beta+2\,\omega\,\beta)\tfrac{1-\cos\eta}{2}\right)^2 + \tfrac{a^2}{4}\left(1-\beta-2\,\omega\right)^2 \sin^2\eta} \right),$$

where we assumed $|a| > 0$ and defined the ratio $\beta = \frac{b}{a}$.

The stable weight as defined in (75) can be written as

$$\bar{\omega} = \frac{a+b}{2a} = \frac{1+\beta}{2},$$

which may be introduced into $\rho_{\max}$, yielding

$$\rho_{\max}(a,\beta,\bar{\omega}) = \max_{\eta\in(-\pi,\pi)} \left(\left| 1 - a\left(1 - (1-\beta^2)\tfrac{1-\cos\eta}{2}\right)\right| + |a| \left(1 - (1-\beta^2)\tfrac{1-\cos\eta}{2}\right)\right)$$

after some calculation. The expression in large brackets is a positive quantity with the bounds

$$0 \le \min\left(1,\beta^2\right) \le 1 - (1-\beta^2)\tfrac{1-\cos\eta}{2} \le \max\left(1,\beta^2\right).$$

Especially, for a given value of β there exists an η such that this expression is non-vanishing. From this fact we conclude for $\rho_{\max}\left(a,\beta,\bar{\omega}\right)$

$$a < 0 \quad \Rightarrow \quad \rho_{\max} > 1;$$

hence $a \ge 0$ is necessary for stability. The conditions for $\rho_{\max} \le 1$ now follow from the condition that the modulus expression in $\rho_{\max}$ should be nonnegative. Thus we obtain

$$a\left(1 - (1-\beta^2)\tfrac{1-\cos\eta}{2}\right) \le 1 \quad \Rightarrow \quad a \max\left(1,\beta^2\right) \le 1,$$

which, since $\beta = b/a$, finally gives $b^2 \le a \le 1$ as stated in the lemma.

For the second part of the lemma we consider the Taylor expansion

$$\rho_{\max}(a,\beta,\bar{\omega}+\delta\omega) = \rho_{\max}|_{\omega=\bar{\omega}} + \left.\frac{\partial\rho_{\max}}{\partial\omega}\right|_{\omega=\bar{\omega}} \delta\omega + \left.\frac{\partial^2\rho_{\max}}{\partial\omega^2}\right|_{\omega=\bar{\omega}} \delta\omega^2 + \mathcal{O}(\delta\omega^3)$$

for the maximal spectral radius. Under the conditions $b^2 \le a \le 1$ we have shown $\rho_{\max}|_{\omega=\bar{\omega}} = 1$. Starting with the general formula for $\rho_{\max}(a,b,\omega)$ as given above, computer algebra software easily gives

$$\frac{\partial\rho_{\max}}{\partial\omega}(a,\beta,\bar{\omega}) = 0$$

and

$$\frac{\partial^2 \rho_{\max}}{\partial \omega^2}(a,\beta,\bar{\omega}) = \frac{2\,a\sin^2\eta}{|1+\beta^2+(1-\beta^2)\cos\eta|} + \frac{2\,a^2\sin^2\eta}{|2-a(1+\beta^2)-(1-\beta^2)\cos\eta|}$$

for the derivatives, which justifies the expansion with positive constant c. □

Note that stability of the one-sided scheme is given also for specific flows with $\theta_1 < 0$ or $\theta_4 < 0$ (according to Figure 2), which point outside the range given by the two flux distributions. Intuitively, we would expect $|b| \le a$ for the Courant numbers, but the lemma states only $|b| \le \sqrt{a}$. This condition becomes more and more restrictive if the angles $\theta_{1,4}$ approach $-\pi/4$. For the extreme cases $\theta_{1,4} = -\pi/4$ we would obtain a flow in a negative (respectively, positive) y-direction and the condition $|b| \le \sqrt{a} = 0$. Furthermore, one of the weights becomes negative if $\theta_1 < 0$ or $\theta_4 < 0$ holds.

Lemma 4.1 investigates the first order scheme. The analysis for the second order scheme becomes much more involved and hardly solvable by hand. But numerical experiments suggest that the second order scheme appears to remain stable under the same conditions. Furthermore the numerical exploration of the amplifier matrix results in a picture very similar to the right-hand side of Figure 2. A detailed investigation of the second order scheme remains for future work.

Finally, we generalize the result for the one-sided scheme to the full upwind scheme with four flux distributions. The one-sided scheme may easily be formulated for all four possible coordinate directions. For a general scheme, we propose a superposition of these four one-sided schemes in order to obtain a full upwind scheme. Hence, for any flow the weights are chosen such that two flux distributions are activated according to the appropriate one-sided case. The resulting weights may be constructed from the direction vector each skeleton is associated with. These vectors are given by

$$\mathbf{n}_1 = (1,1)\,, \quad \mathbf{n}_2 = (-1,1), \quad \mathbf{n}_3 = (-1,-1), \quad \mathbf{n}_4 = (1,-1), \tag{77}$$

following the numbering of the sketch in Figure 1. Based on these vectors the general local weights $\omega^{(1,2,3,4)}$ have the representation

$$\omega^{(g)}\left(\mathbf{v}_{i,j}\right) = \frac{\max(\mathbf{n}_g \cdot \mathbf{v}_{i,j}, 0)}{\sum_{\gamma=1}^{4} \max(\mathbf{n}_\gamma \cdot \mathbf{v}_{i,j}, 0)}, \tag{78}$$

which may be verified to coincide with the appropriate one-sided case depending on the direction of $\mathbf{v}$. In addition we define $\omega^{(g)}(\mathbf{0}) = 0$. By extrapolation of Lemma 4 we may draw the conclusion that the scheme (67) with the weights (78) will be stable provided we have

$$\max_{x,y\in\Omega}(|a_{i,j}|) \le 1 \quad \text{and} \quad \max_{x,y\in\Omega}(|b_{i,j}|) \le 1, \tag{79}$$

where $a_{i,j}$ and $b_{i,j}$ are local Courant numbers. One of the weights of (78) is displayed in Figure 3 as the dark curve. Note the correspondence of the shapes between the curve in this figure and the contour in Figure 2. Unfortunately, the weight given in (78) is not differentiable at points where $\mathbf{v}$ is orthogonal to any of the $\mathbf{n}_g$ due to the function $\max(\cdot,0)$. However, at least one continuous derivative is needed for second order accuracy as stated in the remark following Lemma 3.2. As regularization of $\max(\cdot,0)$, we propose

$$\max_\varepsilon(x,0) = \frac{1}{2}(x + \sqrt{x^2+4\varepsilon}) \tag{80}$$

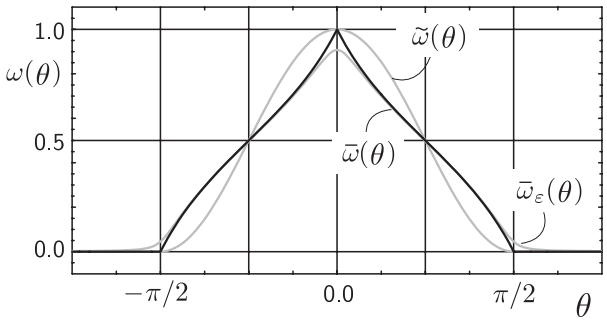


FIG. 3. *The nonsmooth weight function $\bar{\omega}$ obtained from Lemma 4 (dark curve) and two possible regularizations ω_ε and $\tilde{\omega}$. The angle θ is the angle between the flow and the direction of the flux distribution. Compare this plot with the stability contours in Figure 2.*

for use in (78) resulting in a regularized weight $\bar{\omega}_\varepsilon$. The curve of $\bar{\omega}_\varepsilon$ is also shown in Figure 3. Note that this weight gives a deviation from the weight $\bar{\omega}$ obtained in Lemma 4.1. However, if we choose $\varepsilon = h = \max(\Delta x, \Delta y)$ we have

$$\rho_{\max}(\bar{\omega}_\varepsilon) \approx \rho_{\max}(\bar{\omega} + \sqrt{h}) \approx 1 + c\,h \tag{81}$$

according to the second statement in Lemma 4.1. This increases the error constant of the scheme but still gives stability; see, e.g., [11].

Another possible regularization is given by

$$\tilde{\omega}^{(g)}(\mathbf{v}) = \frac{\max(\mathbf{n}_g \cdot \mathbf{v}, 0)^2}{2\,\|\mathbf{v}\|^2}, \tag{82}$$

which is also depicted in Figure 3. This weight deviates considerably from $\bar{\omega}$ and stability is not assured by Lemma 4.1. However, we want to remark that in our numerical calculations with this weight instabilities did not occur. This fact needs further investigation. It could be possible that error modes considered by the von Neumann analysis are not excited in the numerical evaluations due to the constraint-preserving property.

5. Numerical experiments. We proceed to present numerical experiments for two-dimensional div-preserving advection (22) calculated with the upwind scheme given in (67). This scheme exactly preserves the extended divergence operator $\tilde{\mathcal{C}}^{(\star)}$. The symmetric schemes (48), which preserves the classical operator and (58) are not considered due to their instability. For the scheme (67) we write FD, which is an abbreviation of "flux distribution." $FD^{(2)}$ stands for the second order scheme (70) with weight $\bar{\omega}$, while $FD^{(2)}_\varepsilon$ uses the regularized weight $\bar{\omega}_\varepsilon$. Analogously, $FD^{(1)}$ denotes the first order scheme (69) with weight $\bar{\omega}$. For the smooth test cases we used central finite differences to approximate the derivatives in the second order flux distribution coefficient given by (70).

5.1. Smooth test cases. In order to obtain empirical orders of convergence we considered smooth initial conditions

$$\mathbf{u}_0(x, y) = \begin{pmatrix} -1 + \frac{1}{2}\sin(\pi x) + \frac{1}{4}\cos(\pi y) \\ 1 + \frac{1}{2}\cos(\pi x) + \frac{1}{4}\sin(\pi y) \end{pmatrix} \tag{83}$$

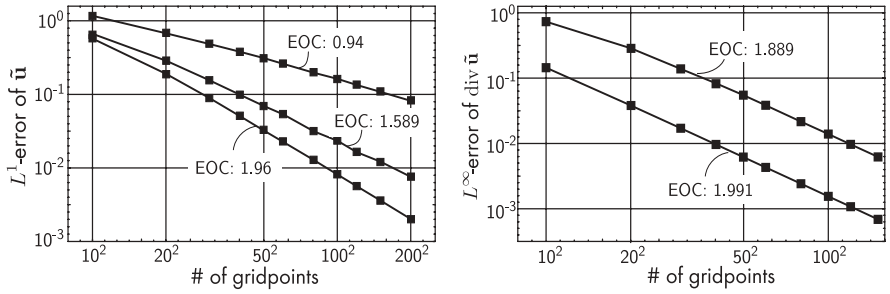


FIG. 4. *Empirical order of convergence for a smooth solution and the divergence of the solution; for problem specification see section* 5.1. *From above, the three curves in the left plot refer to the first order upwind scheme, the second order upwind scheme with nonregularized weight, and the second order upwind scheme with regularized weight. On the right, the upper (lower) curve shows evaluations of the classical (extended) discrete divergence operator. Beside the error curves, averaged empirical orders of convergence are displayed.*

in the computational domain $\Omega = [-1, 1]^2$. To eliminate the influence of boundary conditions periodic boundaries were furnished in both dimensions. The initial vector field $\mathbf{u}_0$ has a nonvanishing divergence which will be frozen under div-preserving advection. The vector field is advected by the velocity field

$$\mathbf{v}(x, y) = \begin{pmatrix} 1 + \frac{1}{4}\cos(\pi x) + \frac{1}{2}\sin(\pi y) \\ 1 + \frac{1}{4}\sin(\pi x) + \frac{1}{2}\cos(\pi y) \end{pmatrix}, \tag{84}$$

which is periodic as $\mathbf{u}_0$. As end time, $t = 0.5$ was chosen. Since an analytic solution is not available for this case, a reference solution has been calculated on a uniform grid with 1200×1200 points and 540 constant time steps. The maximal Courant number

$$c_{\max} = \max_{x,y\in\Omega} \left(\frac{v^{(x)}\Delta t}{\Delta x}, \frac{v^{(y)}\Delta t}{\Delta y} \right) \tag{85}$$

for this solution was approximately 0.97.

The reference solution is used to calculate empirical orders of convergence for calculations on $N \times N$ grids with $N = 10, 20, 30, 40, 50, 60, 80, 100, 120, 150, 200$. All these calculations were performed with constant time steps such that $c_{\max} \approx 0.875$. For the coarsest grid with 10×10 this results in five time steps. The left-hand side of Figure 4 shows the L^1-errors of second order schemes with regularized and nonregularized weight as well as the L^1-errors of the first order scheme. As predicted in the preceding section, the $FD^{(2)}$-scheme does not achieve full second order. Only due to the regularization (80) full second order is obtained with the $FD^{(2)}_\varepsilon$-scheme. The errors and the order of convergence depend slightly on the regularization parameter ε. In Figure 4, $\varepsilon = 5\Delta x$ was chosen. Higher values give a slightly improved order of convergence. $FD^{(1)}$ exhibits first order independently of the regularity of the weight.

The right-hand side of Figure 4 displays the L^∞-error of discrete divergences of the $FD^{(2)}_\varepsilon$-solution at $t = 0.5$. The curves refer to evaluations with the classical and the extended operator, $\tilde{\mathcal{C}}^{(0)}$ and $\tilde{\mathcal{C}}^{(\star)}$ as given in (45) and (53), respectively. Due to the constraint preservation of all FD-schemes, the evaluation of the extended operator $\tilde{\mathcal{C}}^{(\star)}$ yields the *same numerical value for the divergence during the entire calculation*. This value is given by the discrete initial conditions. Hence, the lower curve in Figure 4 (right) simply represents the increasing resolution of the initial

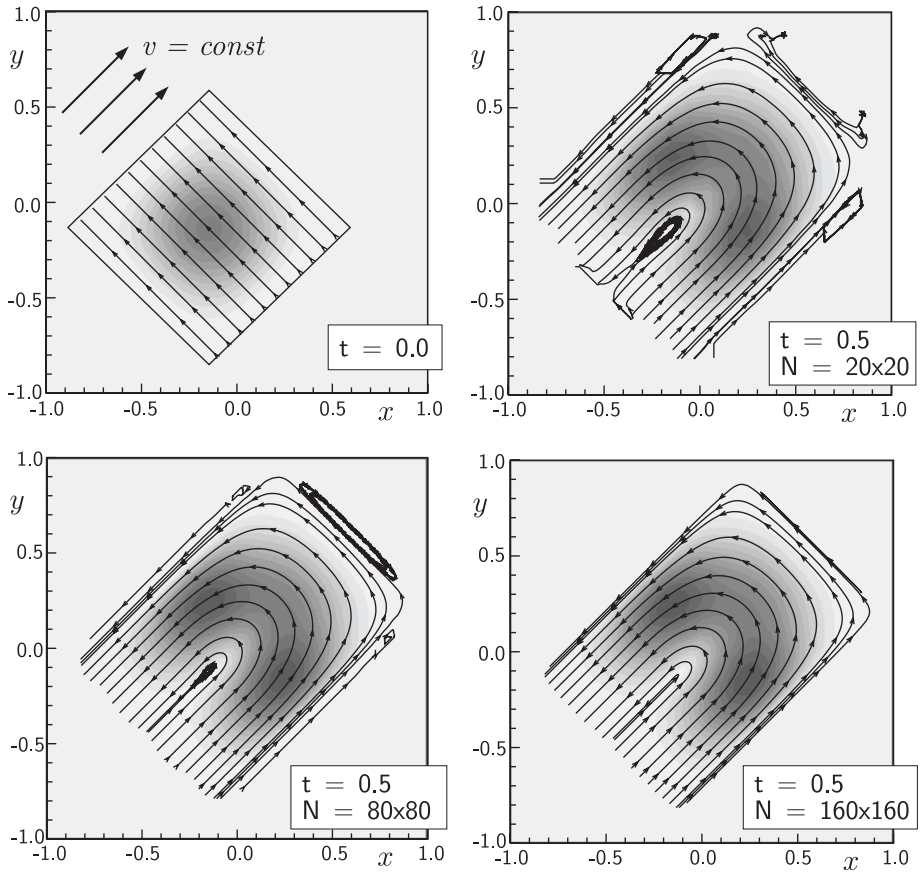


FIG. 5. *Numerical results for the box test case (see section* 5.1.1*) for three different grids. The initial conditions and the direction of the constant advection velocity are shown in the upper left corner.*

conditions and demonstrates second order for the extended operator. In contrast, the value of the classical operator is affected during the numerical calculation (not shown). However, since the solution is smooth the evaluation of $\tilde{\mathcal{C}}^{(0)}$ at $t = 0.5$ is converging, which is visible in the right plot (upper curve) of Figure 4.

5.1.1. Box test case. If initial conditions for the div-preserving advection (22) are given in the form

$$\mathbf{u}_0(x,y) = (0, g'(x)\, g'(y)) \tag{86}$$

with derivatives of a function g and the velocity field by $\mathbf{v}(x,y) = (1,0)$, i.e., pointing constantly in the x-direction, the exact solution has the form

$$\mathbf{u}(x,y,t) = (g''(y)\,(g(x) - g(x-t))\,, g'(x-t)\, g'(y))\,. \tag{87}$$

As an example we choose

$$g(x) = \begin{cases} \frac{16}{5}s^5 - \frac{8}{3}s^3 + s, & -\frac{1}{2} < s < \frac{1}{2}, \\ \pm\frac{4}{15} & \text{else}, \end{cases} \tag{88}$$

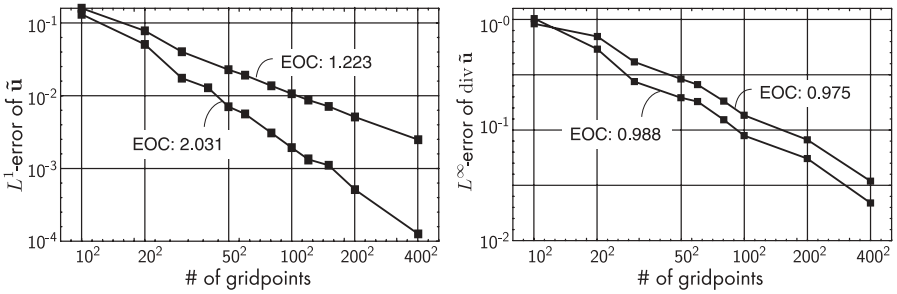


FIG. 6. *Numerical errors of the solution of the box test case at $t = 0.5$ with different grids. Left: L^1-error of $\mathbf{u}$. The upper curve is obtained with the first order scheme $FD^{(1)}$, the lower curve refers to $FD^{(2)}$. Note that regularity of the weight does not play a role in this example due to constant advection velocity. Right: L^∞-error for the extended (lower curve) and classical (upper curve) divergence operator. Beside the error curves averaged empirical orders of convergence are displayed.*

so that the initial field $\mathbf{u}_0$ is nonvanishing only inside the box $(-1/2, 1/2)^2$. The initial condition has a nonvanishing divergence; hence, the advection will differ from ordinary advection, though the advection velocity is constant.

In the numerical test the system given by these initial conditions is rotated by 45° such that advection takes place in a diagonal grid direction. The initial conditions are displayed in the upper left corner of Figure 5. The center of the box is moved to $(-0.1, -0.1)$. The contour shading represents values of $\|\tilde{\mathbf{u}}\|$, which ranges form zero to 1.36, while the lines in the plots represent field lines of $\tilde{\mathbf{u}}$ (the flow which is induced by $\tilde{\mathbf{u}}$). The calculation is conducted in the domain $[-1, 1]^2$ with constant extrapolation in the boundary cells. Besides the initial conditions, Figure 5 displays numerical results at $t = 0.5$ for three uniform meshes with different resolutions. All results were obtained with the $FD^{(2)}$ scheme with constant time steps such that the maximal courant number $c_{\max} \approx 0.884$. Note that the weight is constant in this example since the advection velocity is constant. In the exact solutions the field lines are bent outside the box due to the advection. Since the divergence of $\tilde{\mathbf{u}}$ does not vanish initially, the field lines fill up the way and their starting and ending points stay inside the initial box. In other words, the nonvanishing divergence inside the box acts as source and sink for the field lines.

On the coarse grid the solution is spoiled at the sides by artificial field lines. These field lines correspond to values of $\tilde{\mathbf{u}}$ in the magnitude of the truncation error which are introduced by the finite stencil of the scheme at the boundary of the box. Note that outside the initial box erroneous field lines appear as closed lines which indicates the solenoidal character of the scheme. On the fine mesh the solution is well resolved.

In Figure 6 we display the L^1-error of the variable $\mathbf{u}$ and the L^∞-error of the divergence for the box test case together with averaged empirical orders of convergence. Second order is well obtained, while the $FD^{(1)}$-scheme shows a slight superconvergence for this example. The irregularities in the second order error curve might be due to the nonsmooth gradient of the solution (87) with (88) along the lines $y = \pm\frac{1}{2}$. This is also the reason that the convergence of the divergence on the right-hand side of the figure is reduced to first order. Like in the smooth test case, the divergence error for the extended operator gives the same value during the entire simulation since this value is locally preserved by the scheme. The method freezes the discrete divergence of the initial conditions like the analytical system does.

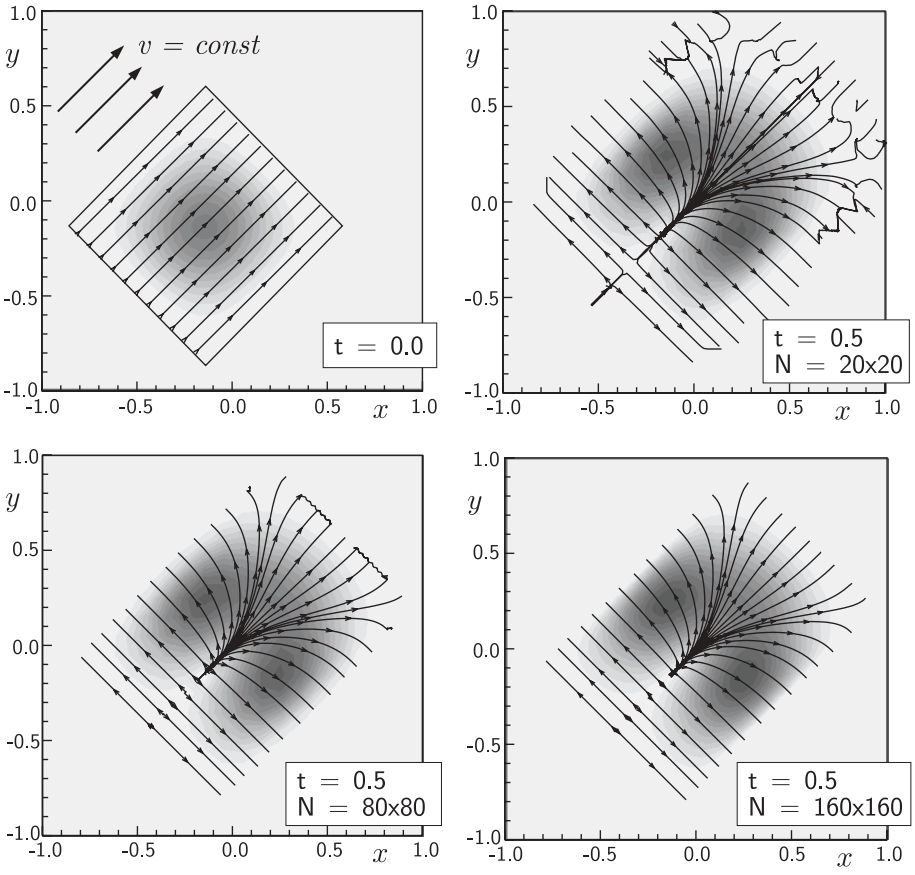


FIG. 7. *Initial conditions and grid convergence study for the box test case of* curl*-free advection. The initial conditions and the result are dual to that of* div*-free advection and should be compared to the corresponding results in Figure* 5.

Box test case for curl*-preserving advection.* It is interesting to ask for the dual solution of the box test case in the sense of the duality of curl-preserving and div-preserving advection as indicated in (25). The solution is depicted in Figure 7, which should be directly compared with Figure 5. The dual solution is obtained by taking the orthogonal complement of the initial conditions (86) as well as of the solution (87). The result is a solution of the curl-preserving advection (23). Accordingly, the field lines in Figures 5 and 7 are orthogonal.

The plots in Figure 7 can be obtained equivalently in two ways: either by taking the orthogonal vector in each cell of the result of the scheme for div-preserving advection or by constructing the corresponding flux distribution scheme for curl-preserving advection and applying it to the dual initial conditions. In fact, this scheme would differ from the div-preserving scheme only in the structure of the flux distribution shape functions in (55). These shape functions are substituted by their orthogonal complement, yielding outward pointing arrows instead of the approximate loops in Figure 1. The resulting scheme preserves perfectly the discrete value of the curl but has the same properties in consistency and stability as its dual counterpart, which was constructed in the preceding sections. Indeed the plot of errors and the empirical

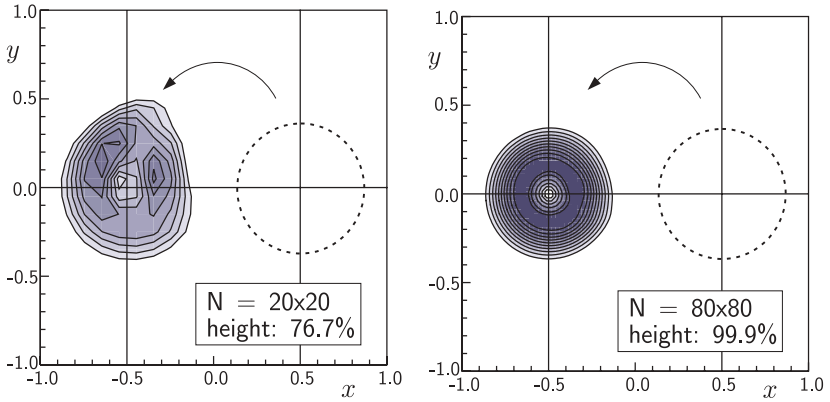


FIG. 8. *Calculation of a smooth hump rotating around the origin on two different grids. Both results were obtained with the second order method* $FD_\varepsilon^{(2)}$. *The plots show the calculation for* $t = \pi$. *The loss of height compared to the exact solution is also displayed.*

order of convergence in Figure 6 also hold identically for the curl-preserving scheme.

5.1.2. Rotating hump. The advection velocity $\mathbf{v}(x, y) = (-y, x)^T$ results in a rotational flow around the origin. As mentioned in section 2.4, the components of $\mathbf{u}$ are not ordinarily advected in such a flow if div-preserving advection is considered. Indeed, the exact solution for divergence-preserving advection given in (22) with a rotational flow is given by

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{R}(t)^{-1}\, \mathbf{u}_0(\mathbf{R}(t)\,\mathbf{x}), \tag{89}$$

where $\mathbf{R}(t)$ is a orthogonal matrix which rotates a vector by the angle t and $\mathbf{u}_0$ is the initial condition. In the case of ordinary advection the inverse $\mathbf{R}^{-1}$ would be missing in the solution. However, if $\mathbf{u}$ has initially vanishing divergence, the 2-norm $\|\mathbf{u}\|$ satisfies an ordinary advection equation, which follows directly from (89).

We consider the initial condition

$$\mathbf{u}_0(x, y) = \frac{1}{5\varepsilon} \begin{pmatrix} -y \\ x - \frac{1}{2} \end{pmatrix} \exp\left(-\frac{(x - \frac{1}{2})^2 + y^2}{\varepsilon}\right) \tag{90}$$

with $\varepsilon = \frac{1}{20}$. This vector field is easily verified to be solenoidal. It produces field lines circling around $(\frac{1}{2}, 0)$ and a smooth but distinct hump in $\|\mathbf{u}\|$ with an essential radius of approximately $\frac{1}{2}$. Note that in the center of this hump, $\|\mathbf{u}\|$ is zero. The computations are conducted in the domain $[-1, 1]^2$, where the exact solution is prescribed in the ghost cells of the boundary.

Two numerical solutions of the problem calculated with $FD_\varepsilon^{(2)}$ are depicted in Figure 8 at time $t = \pi$. The Courant number in these calculations was 0.963. The figure shows contours and contour lines of the absolute value $\|\tilde{\mathbf{u}}\|$ for results obtained with two different grids, 20×20 and 80×80 cells. It also displays the loss of height of the hump compared to the exact solution. The fine grid calculation exhibits a good preservation of symmetry and height.

5.2. Calculating discontinuities. Finally, we present numerical experiments with discontinuous solutions. Discontinuities are most challenging for divergence-

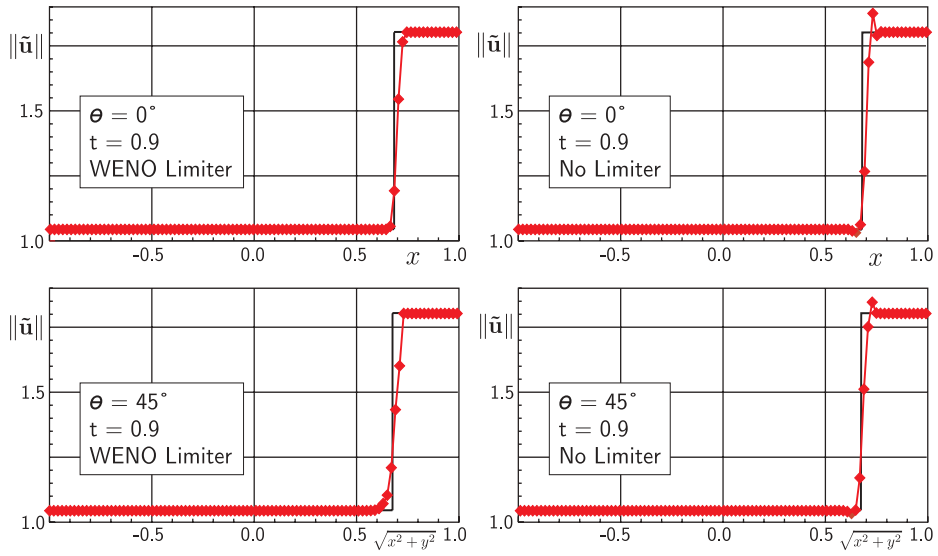


FIG. 9. *Cuts through two-dimensional calculations of advected discontinuities. The upper and lower rows show the results for horizontal and diagonal advection, respectively. Both problems have been computed with and without limited finite differences in the second order scheme.*

preserving methods in the context of magnetohydrodynamics where the magnetic field jumps across shock waves; see, e.g., [6], [26].

5.2.1. Horizontal and diagonal direction. We consider constant advection in the x-direction, i.e., $\mathbf{v}(x,y) = (\frac{3}{4}, 0)^T$. The initial vector field is given by

$$\mathbf{u}_0(x,y) = \begin{pmatrix} 1.0 \\ 0.3 + 1.2\,h(x) \end{pmatrix} \tag{91}$$

with the Heaviside function $h(x)$, which is zero for $x \leq 0$ and unity if x is positive. In both half spaces $x \lessgtr 0$ the vector field is smooth and its divergence is zero. Furthermore, across the discontinuity the normal component of $\mathbf{u}_0$ remains constant, which leads to zero divergence in the weak sense. The vector field $\mathbf{u}_0$ mimics the behavior of the magnetic field in a magnetohydrodynamic shock wave. As the divergence vanishes and the advection is constant, the discontinuity will be linearly advected. The setting will be varied by rotation with an angle θ. Horizontal advection corresponds to $\theta = 0$.

The problem at hand is calculated in $\Omega = [-1,1]^2$ with the second order scheme $FD^{(2)}$ on a grid with 100×100 cells up to time $t = 0.9$. Ghost cells at the boundary are filled by constant extrapolation and adjustment according to the angle θ. In Figure 9 we display the results for horizontal ($\theta = 0$) and diagonal ($\theta = 45^\circ$) advection. Both problems have been computed either with central finite difference or limited differences (73). The time step for both horizontal and diagonal advection was chosen after a Courant number of approximately 0.96; hence the horizontal advection took more time steps due to a more restrictive stability condition. The figure shows the absolute value $\|\tilde{\mathbf{u}}\|$ by following cuts of the solutions normal to the discontinuities. The solutions with central finite differences exhibit familiar oscillations which are eliminated by the use of the limiter. The discontinuities are well resolved. Note the slight asymmetry in the profiles of the discontinuities compared to the exact solution,

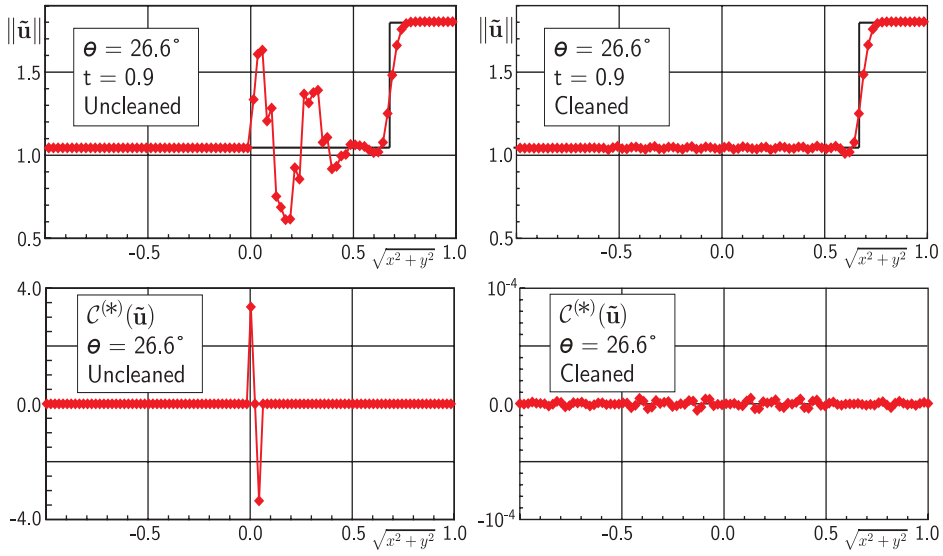


FIG. 10. *Divergence-cleaned and uncleaned solutions of the discontinuous test example in the case of a rotation with* $\theta = 26.6$ *(upper row). Below, the discrete initial divergence of both cases are displayed. The strong deviation from zero leads to a misfit of the computed solution. A divergence cleaning procedure applied to the* initial *conditions removes the disagreement. Both computations are conducted with use of WENO limitation.*

which is drawn as a thin line in Figure 9. This is due to displaying $\|\tilde{\mathbf{u}}\|$ instead of the components $\tilde{u}^{(x)}$ or $\tilde{u}^{(y)}$.

5.2.2. Oblique directions and initial divergence cleaning. The evaluations of both the extended and the classical divergence operator give exactly zero for the initial conditions of the horizontal or diagonal discontinuity. This comes due to symmetry. For discontinuities in all other noncoordinate and nondiagonal directions this is no longer true. Though the analytic initial condition is divergence-free, the discrete evaluation of the divergence in the vicinity of the discontinuity leads to significant deviations from zero.

The left-hand side of Figure 10 shows the result of the computation with $\theta \approx 26.6,^\circ$ which corresponds to $\tan\theta = \frac{1}{2}$. The upper right plot displays the run of $\|\tilde{\mathbf{u}}\|$ along a normal cut and exhibits a complete disagreement with the exact solution (thin line). The plot below shows the evaluation of the extended divergence operator $\tilde{\mathcal{C}}^{(\star)}$ along the same cut of the initial conditions. Due to the constraint preservation this curve stays the same for all time steps. The strong deviations of the divergence from zero are responsible for the disagreement of the computed with the exact solution. Hence, the upper right plot does not represent a failure of the method, but rather indicates the high quality of the constraint preservation. In fact, the computed result belongs to a solution for analytic initial conditions whose divergence is disturbed according to the curve in the lower right plot.

In order to get rid of the divergence in the initial conditions, the discrete field has to be corrected as proposed, e.g., in [2]. We stress that this cleaning procedure is *only* needed for *initial* conditions with nonvanishing divergence due to discontinuities. Hence, the procedure is only applied once in the beginning of the calculation. If the discrete initial divergence is zero, it stays zero due to the properties of our scheme.

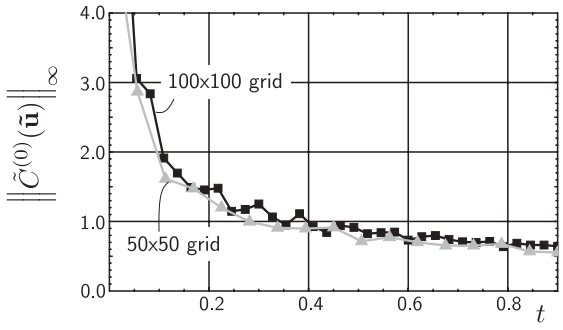


FIG. 11. *Maximal value of the discrete divergence obtained with the classical operator during the two calculations of the discontinuous test example. Independent of the grid size, the classical operator gives decreasing value due to the smoothing of the numerical method.*

The cleaning procedure solves the elliptic equations

$$\begin{aligned} \operatorname{div}\operatorname{grad}\psi &= \operatorname{div}\tilde{\mathbf{u}} \quad \text{in } \Omega, \\ \psi &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{92}$$

for the auxiliary discrete field ψ. The discrete initial field $\tilde{\mathbf{u}}$ is afterwards corrected by $\tilde{\mathbf{u}} \to \tilde{\mathbf{u}} - \operatorname{grad}\psi$ which gives a discrete solenoidal field. This procedure represents the projection onto the divergence-free space (Hodge projection). The differential operator $divgrad \equiv \triangle$ has to be built from the extended divergence operator (53) since the result should give a divergence-free field according the extended operator. The use of the traditional discretization of Laplace operator will not lead to this property. The construction of the Laplace operator by applying an appropriate discrete gradient and afterwards $\tilde{\mathcal{C}}^{(\star)}$ to the field ψ results in a special discrete Laplace operator which assures that the evaluation of $\tilde{\mathcal{C}}^{(\star)}$ on the corrected solution will be zero. The discretized form of (92) may be solved by using iterative linear solvers.

The discrete divergence of the corrected initial condition (91) in the case of $\theta = 26.6°$ is shown in the lower right plot of Figure 10. Note the scale of the ordinate. Finally, the approximation ability of the scheme is fully revealed as can be seen in the upper right plot of Figure 10. The small fluctuations visible in the solution are introduced by the initial cleaning procedure, they vanish with grid refinement.

Note that during the cleaning procedure based on the extended operator $\tilde{\mathcal{C}}^{(\star)}$ we have no control over the value of the *classical* operator (45). Correspondingly, the value of the divergence obtained with this operator is not vanishing if evaluated for the initial conditions. It will also vary during the time steps of the flux distribution scheme. However, the maximal value remains finite during the calculation independent of the grid size as is visible in Figure 11. Moreover, the value of the classical operator decreases due to the numerical smoothing of the discontinuous solution.

6. Sketch of the method in 3 dimensions. We will shortly give a sketch how to extend the constraint-preserving method to the three-dimensional case. The presentation will not be exhaustive but will provide evidence that three-dimensional methods may be constructed from the presented framework as well.

We restrict ourself to div-preserving advection, given in $(4)_{\text{curl}}$. In three dimensions, methods for curl-preserving advection cannot be obtained by duality but need extra considerations. Furthermore, the most important application of curl-preserving advection is the shallow water system which is restricted to two dimensions.

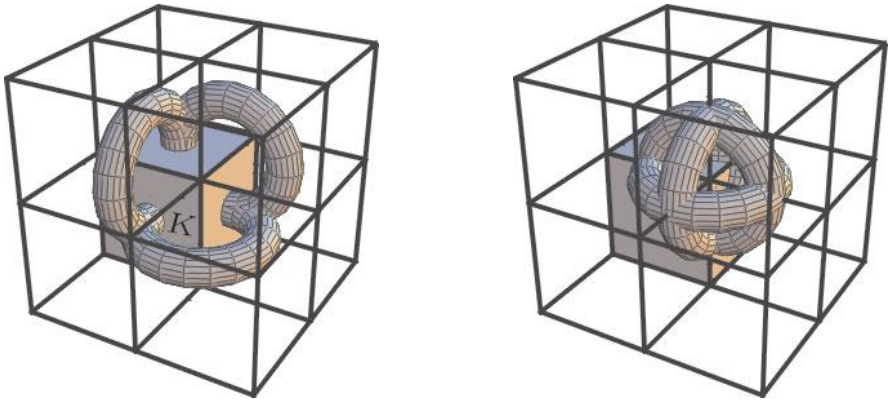


FIG. 12. *Possible sets of shape functions for flux distributions in three dimensions. This figure is to be compared with Figure* 1. *Again the flux distributions has to approximate closed lines which are overstated as tubes in the figure. The right-hand side shape functions result from appropriate averaging.*

6.1. Flux distributions. For discrete divergence operators in three dimensions a representation similar to that of Lemma 2 may be found. However, in this case there exists a family of operators with 17 parameters which is quite involved. Inspired by the two-dimensional results, we proceed by generalizing the extended operator (53) directly to three dimensions, obtaining

$$\begin{aligned}\tilde{\mathcal{C}}_K^{(\star)}\tilde{\mathbf{u}} = {} & \frac{\{\tilde{u}^{(x)}_{i+1,j,k}\}_{y,z} - \{\tilde{u}^{(x)}_{i-1,j,k}\}_{y,z}}{2\Delta x} \\ & + \frac{\{\tilde{u}^{(y)}_{i,j+1,k}\}_{x,z} - \{\tilde{u}^{(y)}_{i,j-1,k}\}_{x,z}}{2\Delta y} + \frac{\{\tilde{u}^{(z)}_{i,j,k+1}\}_{x,y} - \{\tilde{u}^{(z)}_{i,j,k-1}\}_{x,y}}{2\Delta z},\end{aligned} \tag{93}$$

where curled brackets this time stand for

$$\begin{aligned}\{\psi_{i,j,k}\}_{y,z} &= \tfrac{1}{16}(4\psi_{i,j,k} + 2\psi_{i,j+1,k} + 2\psi_{i,j-1,k} + 2\psi_{i,j,k-1} + 2\psi_{i,j,k+1} \\ &\quad +\psi_{i,j+1,k+1} + \psi_{i,j+1,k-1} + \psi_{i,j-1,k+1} + \psi_{i,j-1,k-1}), \\ \{\psi_{i,j,k}\}_{x,z} &= \tfrac{1}{16}(4\psi_{i,j,k} + 2\psi_{i+1,j,k} + 2\psi_{i-1,j,k} + 2\psi_{i,j,k-1} + 2\psi_{i,j,k+1} \\ &\quad +\psi_{i+1,j,k+1} + \psi_{i+1,j,k-1} + \psi_{i-1,j,k+1} + \psi_{i-1,j,k-1}), \\ \{\psi_{i,j,k}\}_{x,y} &= \tfrac{1}{16}(4\psi_{i,j,k} + 2\psi_{i+1,j,k} + 2\psi_{i-1,j,k} + 2\psi_{i,j-1,k} + 2\psi_{i,j+1,k} \\ &\quad +\psi_{i+1,j+1,k} + \psi_{i+1,j-1,k} + \psi_{i-1,j+1,k} + \psi_{i-1,j-1,k}),\end{aligned} \tag{94}$$

i.e., plane-wise averaging. Solving the linear system (39) now gives possible shape functions for flux distributions. All the resulting skeletons have essentially the two-dimensional shape given in (55) and depicted in Figure 1, except they now come with three possible orientations, approximating a circle either in the (x, y)-plane, the (x, z)-plane, or the (y, z)-plane. Hence, there are 36 possible flux distributions altogether, four circles in each cut of the $3 \times 3 \times 3$ grid box. Three of them are sketched on the left-hand side of Figure 12.

Note that it is necessary to take at least three flux distributions to construct a three-dimensional method, since the flux $\mathbf{F} = \mathbf{u} \times \mathbf{v}$ in $(4)_{\text{curl}}$ now has three independent components.

6.2. Possible methods. The dimensionally split character of the three-dimensional flux distributions leads to a method which uses directly the two-dimensional

results. Indeed, the evolution equation $(4)_{\text{curl}}$ may be split into three two-dimensional operators as well. We write

$$\frac{\partial}{\partial t}\begin{pmatrix} u^{(x)} \\ u^{(y)} \\ u^{(z)} \end{pmatrix} + \begin{pmatrix} \partial_y F^{(z)} \\ -\partial_x F^{(z)} \\ 0 \end{pmatrix} + \begin{pmatrix} -\partial_z F^{(y)} \\ 0 \\ \partial_x F^{(y)} \end{pmatrix} + \begin{pmatrix} 0 \\ \partial_z F^{(x)} \\ -\partial_y F^{(x)} \end{pmatrix} = 0, \tag{95}$$

where $\mathbf{F} = (F^{(x)}, F^{(y)}, F^{(z)})^T = \mathbf{u} \times \mathbf{v}$ represents the flux function. It becomes obvious that each bracket can be discretized by the two-dimensional method (67). The resulting flux across a corner is represented by the left picture in Figure 12. The procedure is similar to the operator splitting approach where each flux in a multidimensional conservation law is discretized in a one-dimensional manner (see e.g., [11]), except here we use two-dimensional methods for the single operators. Nevertheless, we expect loss of stability since the cell directly across the corner (see Figure 12, left) is not affected in a single time step. Possible and straightforward help would be to use a fractional time step method, e.g., with Strang splitting, which updates the brackets in (95) successively.

To circumvent the use of splitting it is possible to construct a fully three-dimensional flux distribution as sketched in Figure 12 (right). These flux distributions result from averaging each flux distribution on the left-hand side of the figure with its counterpart in the neighboring parallel grid plane (not shown). A scheme using this single set of flux distributions has the form

$$\begin{aligned} \begin{pmatrix} \tilde{u}^{(x)} \\ \tilde{u}^{(y)} \\ \tilde{u}^{(z)} \end{pmatrix}_{i,j}^{m+1} &= \begin{pmatrix} \tilde{u}^{(x)} \\ \tilde{u}^{(y)} \\ \tilde{u}^{(z)} \end{pmatrix}_{i,j}^{m} \\ &+ \begin{pmatrix} \delta^{(y)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi^{(z,1)})\Delta x - \delta^{(z)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi^{(y,1)})\Delta x \\ -\delta^{(x)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi^{(z,1)})\Delta y + \delta^{(z)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi^{(x,1)})\Delta y \\ \delta^{(x)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi^{(y,1)})\Delta z - \delta^{(y)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi^{(x,1)})\Delta z \end{pmatrix}, \end{aligned} \tag{96}$$

where we used the abbreviations

$$\begin{aligned} \delta^{(x)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi) &= \varphi_{i+1,j+1,k} + \varphi_{i+1,j,k} + \varphi_{i+1,j+1,k+1} + \varphi_{i+1,j,k+1} \\ &\quad -\varphi_{i,j+1,k} - \varphi_{i,j,k} - \varphi_{i,j+1,k+1} - \varphi_{i,j,k+1}, \\ \delta^{(y)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi) &= \varphi_{i+1,j+1,k} + \varphi_{i,j+1,k} + \varphi_{i+1,j+1,k+1} + \varphi_{i,j+1,k+1} \\ &\quad -\varphi_{i+1,j,k} - \varphi_{i,j,k} - \varphi_{i+1,j,k+1} - \varphi_{i,j,k+1}, \\ \delta^{(z)}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}(\varphi) &= \varphi_{i,j+1,k+1} + \varphi_{i+1,j,k+1} + \varphi_{i+1,j+1,k+1} + \varphi_{i,j,k+1} \\ &\quad -\varphi_{i,j+1,k} - \varphi_{i+1,j,k} - \varphi_{i+1,j+1,k} - \varphi_{i,j,k}\,. \end{aligned} \tag{97}$$

Analogously one has to incorporate the flux distributions for the rest of the corners of the grid cell. First order consistency for the single set of flux distributions in (96) leads to

$$\begin{pmatrix} \varphi^{(x,1)} \\ \varphi^{(y,1)} \\ \varphi^{(z,1)} \end{pmatrix} = -\frac{\Delta t}{4\Delta x \Delta y \Delta z}\begin{pmatrix} \Delta x\, F^{(x)} \\ \Delta y\, F^{(y)} \\ \Delta z\, F^{(z)} \end{pmatrix}, \tag{98}$$

which gives a method for div-free advection in three dimensions which exactly preserves the discrete value of the divergence evaluated by (93). The flux distributions

of the rest of the corners could be incorporated by weights as in the two-dimensional case. The elaboration of the details of the three-dimensional method remains for future work.

7. Conclusions. In this paper we drew attention to constraint-preserving advection equations. These equations are characterized by the existence of an intrinsic differential constraint which holds locally during the evolution. They form models for general evolution equations with constraints which can be found in various fields of physics and engineering.

Starting from the hypothesis that numerical methods should respect the constraints, we proposed a general framework for constructing constraint-preserving schemes. Based on this framework a multidimensional upwind method was developed. Consistency and stability were proven, and various numerical experiments demonstrated the ability and reliability of the new scheme. We also re-derived former numerical schemes within our framework. The new method relies on special flux distribution and does not require staggered grids, time-step-wise global correction procedures, or modified evolution equations as proposed in former approaches, e.g., [8], [2], [6].

In [18] a precursor of the present method was used in the context of the method of transport [9], [10] to solve the magnetic evolution part of a magnetohydrodynamic computation. In [25] the results of this paper are used to derive general divergence-preserving finite-volume schemes for magnetohydrodynamics. Future work will also include applications to electrodynamics, meteorological flows, and Einstein equations.

In this work we considered a rectangular mesh as a first approach. The treatment of more general grids, e.g., triangular or quadrilateral, is a major issue for future work. The framework given in this paper allows for constraint-preserving methods on such grids. The main problem is to find an appropriate discretization of the constraint on the given grid. In [25] divergence-preserving methods on triangular grids are derived using the framework of this paper. In [7] an approach to triangular grids is presented based on staggered grids.

The discrete constraint preservations also requires further investigations on discrete data treatment. Implementations of boundary conditions as well as restriction and prolongation in an adaptive grid (see [27]) should be revised from the angle of constraint preservation.

Acknowledgment. The authors would like to thank Prof. S. Noelle (RWTH Aachen, Germany) and his group for valuable discussions.

REFERENCES

[1] D. S. Balsara and D. S. Spicer, *A staggered mesh algorithm using high order Godunov fluxes to ensure solenoidal magnetic fields in magnetohydrodynamic simulations*, J. Comput. Phys., 149 (1999), pp. 270–292.

[2] J. Brackbill and D. C. Barnes, *The effect of nonzero $\nabla \cdot B$ on the numerical solution of the magnetohydrodynamic equations*, J. Comput. Phys., 35 (1980), pp. 426–430.

[3] M. W. Choptuik, *Consistency of finite-difference solutions of Einstein's equation*, Phys. Rev. D, 44 (1991), pp. 3124–3135.

[4] C. M. Dafermos, *Quasilinear hyperbolic systems with involutions*, Arch. Rational Mech. Anal., 94 (1986), pp. 373–389.

[5] W. Dai and P. R. Woodward, *On the divergence-free condition and conservation laws in numerical simulations for supersonic magnetohydrodynamic flows*, Astrophys. J., 494 (1998), pp. 317–335.

[6] A. Dedner, F. Kemm, D. Kröner, C.-D. Munz, T. Schnitzer, and M. Wesenberg, *Hyperbolic divergence cleaning for the MHD equations*, J. Comput. Phys., 175 (2002), pp. 645–673.
[7] H. DeSterck, *Multi-Dimensional Upwind Constrained Transport on Unstructured Grids for Shallow Water Magnetohydrodynamics*, AIAA Paper 2001–2623, 2001.
[8] C. R. Evans and J. F. Hawley, *Simulation of magnetohydrodynamic flows: A constrained transport method*, Astrophys. J., 332 (1988), pp. 659–677.
[9] M. Fey, *Multidimensional upwinding. Part* I. *The method of transport for solving the Euler equations,* J. Comput. Phys., 143 (1998), pp. 159–180.
[10] M. Fey, *Multidimensional upwinding. Part* II. *Decomposition of the Euler equations into advection equations,* J. Comput. Phys., 143 (1998), pp. 181–199.
[11] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer-Verlag, New York, 1996.
[12] G. Guj and F. Stella, *A vorticity-velocity method for the numerical solution of* 3*D incompressible flow,* J. Comput. Phys., 106 (1993), pp. 286–298.
[13] J. M. Hyman and M. Shashkov, *Natural discretizations for the divergence, gradient, and curl on logically rectangular grids*, Comput. Math. Appl., 33 (1997), pp. 81–104.
[14] J. M. Hyman and M. Shashkov, *Mimetic discretizations for Maxwell's equations*, J. Comput. Phys., 151 (1999), pp. 881–909.
[15] K. Jänich, *Vector Analysis*, Springer-Verlag, New York, 2001.
[16] A. Jeffrey and T. Taniuti, *Non-linear Wave Propagation*, Academy Press, New York, 1964.
[17] G.-S. Jiang and C.-W. Shu, *Efficient Implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.
[18] R. Limacher, *Simulation der MHD-Gleichungen mit der Transportmethode*, Diploma Thesis, ETH Zurich, 2000.
[19] S.-J. Lin and R. B. Rood, *An explicit flux-form semi-Lagrangian shallow-water on the sphere*, Q. J. Roy. Meteor. Soc., 123 (1997), p. 2477.
[20] K. W. Morton and P. L. Roe, *Vorticity-preserving Lax–Wendroff-type schemes for the system wave equation*, SIAM J. Sci. Comput., 23 (2001), pp. 170–192.
[21] K. G. Powell, *An Approximate Riemann Solver for Magnetohydrodynamics (That Works in More Than One Dimension)*, ICASE Report 94-24, 1994.
[22] M. Alcubierre, G. Allen, B. Brügmann, E. Seidel, and W.-M. Suen, *Towards an understanding of the stability properties of the* $3+1$ *evolution equations in general relativity*, Phys. Rev. D, 62 (2000), 124011.
[23] D. Serre, *Systems of Conservation Laws*, Vol. 1, Cambridge University Press, Cambridge, UK, 1999.
[24] A. S. Schwarz, *Topology for Physicists*, Fundamental Principles of Mathematical Sciences 308, Springer-Verlag, Berlin, 1994.
[25] M. Torrilhon, *Locally divergence-preserving upwind finite volume schemes for magnetohydrodynamic equations*, SIAM J. Sci. Comput., to appear.
[26] G. Toth, *The* $\nabla \cdot B$ *constraint in shock-capturing magnetohydrodynamics codes*, J. Comput. Phys., 161 (2000), pp. 605–652.
[27] G. Toth and P. L. Roe, *Divergence- and curl-preserving prolongation and restriction formulas*, J. Comput. Phys., 180 (2002), pp. 736–750.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1729–1744

GLOBAL SUPERCONVERGENCE AND A POSTERIORI ERROR ESTIMATORS OF THE FINITE ELEMENT METHOD FOR A QUASI-LINEAR ELLIPTIC BOUNDARY VALUE PROBLEM OF NONMONOTONE TYPE*

LIPING LIU†, TANG LIU‡, MICHAL KŘÍŽEK§, TAO LIN¶, AND SHUHUA ZHANG‡

Abstract. In this paper we are concerned with finite element approximations to a nonlinear elliptic partial differential equation with homogeneous Dirichlet boundary conditions. This kind of problems arises for example from modeling a stationary heat conduction in nonlinear inhomogeneous and anisotropic media. For finite elements of degree $k \geq 1$ in each variable, by means of an interpolation postprocessing technique, we obtain the global superconvergence of $O(h^{k+1})$ in the H^1-norm and $O(h^{k+2})$ in the L^2-norm provided the weak solution is sufficiently smooth. As by-products, the global superconvergence results can be used to generate efficient a posteriori error estimators. Representative numerical examples are also given to illustrate our theoretical analysis.

Key words. nonlinear boundary problem, finite elements, supercloseness, global superconvergence, a posteriori error estimators, anisotropic heat conduction

AMS subject classifications. 65N15, 65N30

DOI. 10.1137/S0036142903428402

1. Introduction.

Our purpose in this paper is to study the Galerkin finite element method for a quasi-linear elliptic problem whose classical formulation is as follows: Find $u \in C(\overline{\Omega})$ such that $u|_\Omega \in C^2(\Omega)$ and

$$\begin{aligned} -\nabla \cdot (A(x,u)\nabla u) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, is an open bounded polyhedral domain with a Lipschitz boundary, $f \in L^2(\Omega)$, and $A = (a_{ij})_{i,j=1}^d$ is a bounded uniformly positive definite matrix, i.e.,

$$\max_{x\in\Omega} \max_{\xi\in\mathbb{R}} |a_{ij}(x,\xi)| \leq C \quad \forall i,j \in \{1,\dots,d\}, \tag{1.2}$$

$$C_0 \eta^T \eta \leq \eta^T A(x,\xi)\eta \quad \forall \eta \in \mathbb{R}^d, \quad \forall x \in \Omega, \quad \forall \xi \in \mathbb{R}, \tag{1.3}$$

*Received by the editors May 22, 2003; accepted for publication (in revised form) May 27, 2004; published electronically December 27, 2004. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada, the Academy of Sciences of the Czech Republic (A1019201), the National Natural Science Foundation of China (NSF10471103), and the Liu Hui Center for Applied Mathematics of Nankai University and Tianjin University, Tianjin University of Finance and Economics, and Tianjin Education Committee.
http://www.siam.org/journals/sinum/42-4/42840.html

†Department of Mathematical Sciences, Lakehead University, Thunder Bay, ON P7B 5E1, Canada (liping.liu@lakeheadu.ca).

‡Department of Mathematics, Tianjin University of Finance and Economics, Tianjin 300222, China, and Liu Hui Center for Applied Mathematics of Nankai University and Tianjin University, China (tangliu@eyou.com, shuhua@eyou.com).

§Mathematical Institute, Academy of Sciences, Žitná 25, CZ-115 67, Prague 1, Czech Republic (krizek@math.cas.cz).

¶Department of Mathematics, Virginia Technology University, Blackburg, VA 24061 (tlin@math.vt.edu).

where C and C_0 are positive constants. In addition, we assume that the derivatives $\partial a_{ij}/\partial\xi$, $\partial^2 a_{ij}/\partial\xi^2$, and $\partial^2 a_{ij}/\partial\xi\partial x_\ell$ are bounded and continuous on $\overline{\Omega}\times\mathbb{R}$ for all $i,j,\ell\in\{1,\dots,d\}$. However, the matrix A is not necessarily symmetric.

Unlike nonlinear problems of other types, for $d>1$ problem (1.1) cannot be converted, in general, by the well-known Kirchhoff transformation (see [14]) to a linear problem even if A would be independent of x, since A is a matrix function. This makes its theoretical and numerical analysis much more difficult.

As mentioned in [19], since an analogue of the well-known Céa's lemma holds for those nonlinear elliptic problems whose associated operators are strongly monotone and Lipschitz continuous (see, for example, [4, 14]), it is easy for these problems to derive the rate of convergence $O(h^k)$ in the H^1-norm for the Lagrange elements of degree k. Thus, their finite element approximations have been extensively studied. See, for instance, [7, 8, 24]. In [10, 13] some one-dimensional examples are presented to illustrate that problem (1.1) is of a nonmonotone and nonpotential type.

There are some results available for problem (1.1). For example, the uniqueness of the classical and weak solutions was proven in [12, 13], respectively. In [13], the existence of the weak solution was derived as a weak limit of Galerkin approximations. Other existence results for various kinds of boundary conditions are presented in [9, 10, 13, 21]. As to numerical methods, Douglas and Dupont [6] proved an optimal rate of convergence of the finite element approximation for problem (1.1) in the case that

$$A(x,u)=\lambda(x,u)I, \tag{1.4}$$

where I is the identity matrix and λ is a smooth scalar function. Nitsche [22] derived an asymptotic error estimate of the finite element method in the L^∞-norm for the case (1.4). In [20], a mixed finite element method was studied which also has an optimal rate of convergence in the L^p-norm for the case (1.4). Similar results were also obtained in [1, 2].

In [19], the result from [6] was generalized to any smooth uniformly positive definite matrix $A(x,u)$ satisfying (1.2) and (1.3), which represents a practically interesting case, since problem (1.1) describes a steady-state heat conduction in nonlinear inhomogeneous anisotropic media (e.g., in magnetic cores of large transformers [14]), where the unknown function u stands for the temperature, A is the matrix of heat conductivities, and f is the density of volume heat sources (in that case A is symmetric). In fact, for finite elements of degree $k\geq 1$, [19] shows that optimal rates of convergence of $O(h^k)$ in the H^1-norm and $O(h^{k+1})$ in the L^2-norm, provided that the weak solution of (1.1) is sufficiently smooth.

There are several papers dealing with superconvergence of finite element methods for nonlinear elliptic boundary value problems, see survey paper [15, pp. 319–320]. They usually require that the operator associated with a given nonlinear elliptic problem be (strongly) monotone, which is not our case. Also, Wahlbin [23, Chapter 9] presents a linearization technique to examine superconvergence phenomena for nonlinear elliptic problems. However, our technique introduced in this paper is totally different from his. In [3], a superconvergence at the Gauss points of rectangular biquadratic elements is proved for a class of semilinear problems. This class contains nonlinearities only in the absolute term, whereas the nonlinearities of problem (1.1) stand at the second derivatives.

Our aim in this paper is to investigate the global superconvergence and a posteriori error estimators of finite element methods for problem (1.1). By means of an

optimal partition and an optimal interpolation technique, we first establish the superapproximation between the finite element solution and an interpolation function of the weak solution of problem (1.1). Then, we obtain global superconvergent approximations by virtue of a class of interpolation postprocessing methods. On the basis of global superconvergent approximations, efficient a posteriori error estimators are provided to assess the accuracy of finite element solutions in applications.

2. Weak formulation and finite element approximation. In this paper we employ the standard Sobolev space notation. The norm in the product Sobolev space $(W_p^k(\Omega))^n$, $k \in \{0, 1, \ldots\}$, $p \in [1, \infty]$, $n \in \{1, 2, \ldots\}$, is denoted by $\|\cdot\|_{k,p}$. In particular, if $p = 2$, then we set $H^k(\Omega) = W_2^k(\Omega)$ and $\|\cdot\|_k = \|\cdot\|_{k,2}$. By $H_0^1(\Omega)$ we mean the space of functions from $H^1(\Omega)$ whose traces vanish on $\partial\Omega$. The symbol $(\cdot, \cdot)$ stands for the usual scalar product in $L^2(\Omega)$.

The weak formulation of problem (1.1) consists of finding $u \in H_0^1(\Omega)$ such that

$$a(u; u, v) = (f, v) \quad \forall v \in H_0^1(\Omega), \tag{2.1}$$

where

$$a(z; v, w) = \int_\Omega (\nabla w)^T A(x, z) \nabla v \, dx, \quad v, w \in H^1(\Omega),\ z \in L^2(\Omega).$$

The variable x will be sometimes omitted from now on. It follows from (1.3), (1.2), and Friedrichs' inequality that there exist positive constants C_0 and C_1 such that

$$a(z; v, v) \geq C_0 \|v\|_1^2 \quad \forall z \in L^2(\Omega),\ \forall v \in H_0^1(\Omega),$$

and

$$|a(z; w, v)| \leq C_1 \|v\|_1 \|w\|_1 \quad \forall z \in L^2(\Omega),\ \forall w, v \in H^1(\Omega).$$

This means that $a(\cdot; \cdot, \cdot)$ is uniformly $H_0^1(\Omega)$-elliptic and continuous. In [13], it has been proven that the weak solution of (2.1) exists uniquely.

Let $\{\mathcal{T}_h\}$ be a regular family of finite element partitions of $\overline{\Omega}$ (see [4]) and let $V_h \subset H_0^1(\Omega)$ be associated finite element spaces. The finite element solution of (1.1) or (2.1) is then defined as the function $u_h \in V_h$ such that

$$a(u_h; u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h. \tag{2.2}$$

The space V_h can consist of, for example, the Lagrange elements.

The existence of at least one solution u_h of (2.2) can be proven by the Brouwer fixed-point theorem (see [13]). Sufficient conditions guaranteeing the uniqueness of u_h have been presented in [11, 13]. However, the uniqueness of u_h, in general, remains an open problem.

3. Global superconvergence. From now on, all generic constants C will possibly depend on the weak solution u, but they will be independent of the discretization parameter h.

In [13], the convergence of finite element solutions u_h to the weak solution u of (1.1) in the H^1-norm is proven. However, no attempt to derive any rate of convergence is made there. Optimal convergence rates in the H^1-norm and in the L^2-norm are derived in [19]. They are based on the following adjoint problem: Find $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$L^*\varphi \equiv -\nabla \cdot (A^T(x, u)\nabla\varphi) + (\nabla u)^T A_u^T(x, u)\nabla\varphi = \theta \quad \text{in } \Omega, \tag{3.1}$$

where u is the unique solution of (2.1), $\theta \in L^2(\Omega)$, $A_u = ((a_{ij})_u)_{i,j=1}^d$, and the subscript u means the differentiation with respect to the last variable, i.e., $(a_{ij})_u = \partial a_{ij}(x,u)/\partial u$. In [19], a sufficient condition guaranteeing the existence and uniqueness of the weak (generalized) solution of linear problem (3.1) is given. Moreover, we assume the following elliptic regularity:

$$\|\varphi\|_2 \le C\|\theta\|_0. \tag{3.2}$$

Using this, we have the optimal error estimates [19] for problem (1.1).

THEOREM 3.1. *Let $u \in H^{k+1}(\Omega)$ for $k \ge 1$ be the weak solution of* (1.1)*, let the second derivatives $\partial^2 a_{ij}/\partial u^2$ be bounded and continuous on $\overline{\Omega} \times \mathbb{R}$, and let* (3.2) *hold. Let V_h contain polynomials up to degree k. If u_h is a solution of* (2.2)*, then there exist C, $h_0 > 0$ such that for any $h \in (0, h_0)$ we have*

$$\|u - u_h\|_0 + h\|u - u_h\|_1 \le Ch^{k+1},$$

where $C = \overline{C}\|u\|_{k+1}(1 + \|u\|_{k+1})$ and $\overline{C}$ is independent of u (see [18, p. 48]*).*

Throughout the paper generic constants C may depend on derivatives of u up to order $k+1$. In this section, we shall investigate the global superconvergence of the finite element approximation to problem (1.1). To this end, we employ a class of projection interpolation operators and integral identities proposed in [16, 17] to establish the supercloseness between the finite element solution and an interpolant of the exact solution of (1.1). On the basis of this supercloseness, we utilize an interpolation postprocessing technique to prove the global superconvergence.

For simplicity, we assume from now on that the domain Ω is an open rectangle and $\mathcal{T}_h$ is a rectangular partition over $\overline{\Omega}$. In fact, our analysis here is also true for $d = 1$ and $d = 3$. The finite element space V_h will be constructed by means of spaces $Q_k(e)$, $e \in \mathcal{T}_h$; i.e., for $k = 1, 2, 3, \ldots$ we shall use bilinear, biquadratic, bicubic, ... elements, respectively. For $k > 1$ we shall define a special type of projection interpolation operators i_h^k as we have mentioned above, rather than the usual nodal Lagrange interpolation operators, of degree not exceeding k (≥ 2) in x_1 and x_2. Let $e \in \mathcal{T}_h$ be any rectangular element, and let l_i and p_i ($i = 1, 2, 3, 4$) be its edges and vertices, respectively. Then, the bi-kth interpolation operator i_h^k is defined for $u \in H^2(\Omega)$ according to the following so-called vertex-edge-element conditions (see [16, p. 28]):

$$\begin{cases} i_h^k u \in Q_k(e), \\ i_h^k u(p_i) = u(p_i), \qquad i = 1, 2, 3, 4, \\ \displaystyle\int_{l_i} (i_h^k u - u)v\, ds = 0, \quad \forall v \in P_{k-2}(l_i),\ i = 1, 2, 3, 4, \\ \displaystyle\int_e (i_h^k u - u)v\, dx = 0, \quad \forall v \in Q_{k-2}(e), \end{cases} \tag{3.3}$$

where $P_{k-2}(l_i)$ are the polynomial spaces of degree no more than $k-2$ on l_i. In our further analysis, the notation i_h^1 stands for the usual nodal Lagrange bilinear interpolation operator.

By the standard interpolation theory (based on the Bramble–Hilbert lemma) we find that

$$\begin{aligned} \|u - i_h^k u\|_0 &\le Ch^{k+1}\|u\|_{k+1}, \\ \|u - i_h^k u\|_{0,\infty} &\le Ch^{k+1}\|u\|_{k+1,\infty}. \end{aligned} \tag{3.4}$$

Employing an integral identity technique, we obtain the following lemma [16] (for a special case see also [17]).

LEMMA 3.2. *In* (1.1), *assume that entries of the uniformly positive definite matrix* A *are in* $W^{1,\infty}(\Omega)$. *Then there exists a constant* $C > 0$ *such that for all* $v_h \in V_h$ *we have the following estimates:*

$$a(u; u - i_h^k u, v_h) \le \begin{cases} Ch^{k+1}\|u\|_{k+2}\|v_h\|_1, & k \ge 1, \\ Ch^{k+2}\|u\|_{k+2}|||v_h|||_{2,h}, & k \ge 2, \end{cases}$$

where $|||v_h|||_{2,h} := (\sum_e \|v_h\|_{2,e}^2)^{1/2}$.

Now we are in a position to get our main theorems on supercloseness of this section.

THEOREM 3.3. *Let* $u \in H^{k+2}(\Omega)$ *be the weak solution of* (1.1) *for* $k \ge 1$ *and let* u_h *be the corresponding finite element solution. Then we have under the assumptions of Theorem* 3.1 *and Lemma* 3.2 *that*

$$\|u_h - i_h^k u\|_1 \le Ch^{k+1}\|u\|_{k+2},$$

where C *depends on* $\|u\|_{k+1}$.

Proof. It follows from the uniform $H_0^1(\Omega)$-ellipticity of $a(\cdot;\cdot,\cdot)$, (2.1), and (2.2) that

$$\begin{aligned} C_0\|u_h - i_h^k u\|_1^2 &\le a(u_h; u_h - i_h^k u, u_h - i_h^k u) \\ &= a(u_h; u_h, u_h - i_h^k u) - a(u_h; i_h^k u, u_h - i_h^k u) \\ &= a(u; u, u_h - i_h^k u) - a(u_h; i_h^k u, u_h - i_h^k u) \\ &= a(u; u - i_h^k u, u_h - i_h^k u) \\ &\quad + [a(u; i_h^k u, u_h - i_h^k u) - a(u_h; i_h^k u, u_h - i_h^k u)] \\ &=: I_1 + I_2. \end{aligned} \tag{3.5}$$

From Lemma 3.2 we obtain that

$$\begin{aligned} |I_1| &= |a(u; u - i_h^k u, u_h - i_h^k u)| \\ &\le Ch^{k+1}\|u\|_{k+2}\|u_h - i_h^k u\|_1. \end{aligned} \tag{3.6}$$

For any $x \in \Omega$ we have by the mean value theorem that

$$\begin{aligned} A(x,u) - A(x,u_h) &= (u - u_h)\int_0^1 A_u(x, u + t(u_h - u))\,dt \\ &= (u - u_h)\overline{A}_u(x), \end{aligned} \tag{3.7}$$

where $\overline{A}_u = ((\overline{a}_{ij})_u)_{i,j=1}^2$ and $(\overline{a}_{ij})_u = (a_{ij})_u(u + \theta_{ij}(u_h - u))$ for some $\theta_{ij} = \theta_{ij}^h(x) \in [0,1]$. This, together with Theorem 3.1, leads to

$$\|A(x,u) - A(x,u_h)\|_0 \le C\|u - u_h\|_0 \le Ch^{k+1}. \tag{3.8}$$

By the interpolation theorem (see, for example, [4]) and the Sobolev imbedding theorem,

$$\|i_h^k u\|_{1,\infty} \le \|u - i_h^k u\|_{1,\infty} + \|u\|_{1,\infty} \le C\|u\|_{k+2}.$$

From this, by means of the Cauchy–Schwarz inequality and (3.8) we have

$$
\begin{aligned}
|I_2| &= |a(u; i_h^k u, u_h - i_h^k u) - a(u_h; i_h^k u, u_h - i_h^k u)| \\
&= \left| \int_\Omega (\nabla(u_h - i_h^k u))^T [A(x,u) - A(x,u_h)] \nabla i_h^k u \, dx \right| \\
&\le \|i_h^k u\|_{1,\infty} \|A(x,u) - A(x,u_h)\|_0 \|u_h - i_h^k u\|_1 \\
&\le C h^{k+1} \|u\|_{k+2} \|u_h - i_h^k u\|_1.
\end{aligned} \tag{3.9}
$$

Combining (3.6) and (3.9) with (3.5) yields

$$C_0 \|u_h - i_h^k u\|_1^2 \le C h^{k+1} \|u\|_{k+2} \|u_h - i_h^k u\|_1,$$

and hence

$$\|u_h - i_h^k u\|_1 \le C h^{k+1} \|u\|_{k+2}.$$

Thus, the proof of the theorem is complete. ∎

THEOREM 3.4. *Under the assumptions of Theorem* 3.3 *we have for* $k \ge 2$ *that*

$$\|u_h - i_h^k u\|_0 \le C h^{k+2} \|u\|_{k+2},$$

where C *depends on* $\|u\|_{k+1}$.

Proof. We use a duality argument based on the Aubin–Nitsche trick to prove the theorem. Let $\varphi \in H_0^1(\Omega) \cap H^2(\Omega)$ be the weak solution of the linear adjoint problem (3.1) with

$$\theta = u_h - i_h^k u.$$

Then, we have by the Green formula that

$$
\begin{aligned}
\|\theta\|_0^2 &= \int_\Omega \theta^2 \, dx = \int_\Omega \theta (L^* \varphi) \, dx \\
&= \int_\Omega [(\nabla\theta)^T A^T(u) \nabla\varphi + \theta (\nabla u)^T A_u^T(u) \nabla\varphi] \, dx \\
&= \int_\Omega [(\nabla\varphi)^T A(u) \nabla\theta + \theta (\nabla u)^T A_u^T(u) \nabla\varphi] \, dx \\
&= \int_\Omega [(\nabla(\varphi - v_h))^T A(u) \nabla\theta + (\nabla v_h)^T A(u) \nabla\theta + \theta (\nabla u)^T A_u^T(u) \nabla\varphi] \, dx
\end{aligned}
$$

for some $v_h \in V_h$. Therefore,

$$
\begin{aligned}
\|\theta\|_0^2 = & \int_\Omega (\nabla(\varphi - v_h))^T A(u) \nabla\theta \, dx + \int_\Omega (\nabla v_h)^T A(u) \nabla(u - i_h^k u) \, dx \\
& + \int_\Omega [(\nabla v_h)^T A(u) \nabla(u_h - u) + \theta (\nabla\varphi)^T A_u(u) \nabla u] \, dx \\
= & \int_\Omega (\nabla(\varphi - v_h))^T A(u) \nabla\theta \, dx + a(u; u - i_h^k u, v_h) \\
& + \int_\Omega [(\nabla v_h)^T A(u) \nabla(u_h - u) + \theta (\nabla\varphi)^T A_u(u) \nabla u] \, dx \\
=: & \, II_1 + II_2 + II_3,
\end{aligned} \tag{3.10}
$$

where $v_h \in \hat{V}_h = \{\psi \in H_0^1(\Omega) \cap H^2(\Omega) : \psi|_e \in Q_k(e),\ e \in \mathcal{T}_h\} \subset V_h$. The space $\hat{V}_h$ for a fixed k possesses the following approximation properties:

$$\min_{v_h \in \hat{V}_h} \{\|\varphi - v_h\|_0 + h\|\varphi - v_h\|_1 + h^2\|\varphi - v_h\|_2\} \le Ch^l\|\varphi\|_l, \quad \varphi \in H_0^1(\Omega) \cap H^l(\Omega), \quad 2 \le l \le k+1, \tag{3.11}$$

which leads to

$$\|v_h\|_2 \le \|v_h - \varphi\|_2 + \|\varphi\|_2 \le C\|\varphi\|_2 \tag{3.12}$$

for v_h being the best approximation to φ in $\hat{V}_h$ in the sense of (3.11). Therefore, from (3.11) and Theorem 3.3 we know that

$$|II_1| \le C\|\varphi - v_h\|_1\|\theta\|_1 \le Ch^{k+2}\|\varphi\|_2\|u\|_{k+2}. \tag{3.13}$$

From Lemma 3.2 and (3.12) we find that

$$|II_2| \le Ch^{k+2}\|v_h\|_2\|u\|_{k+2} \le Ch^{k+2}\|\varphi\|_2\|u\|_{k+2}. \tag{3.14}$$

Setting $A_u = A_u(u)$, it follows from (3.10), (2.1), (2.2), and (3.7) that

$$\begin{aligned}
II_3 &= \int_\Omega [(\nabla v_h)^T A(u)\nabla u_h - (\nabla v_h)^T A(u)\nabla u + \theta(\nabla\varphi)^T A_u \nabla u]\, dx \\
&= \int_\Omega [(\nabla v_h)^T (A(u) - A(u_h))\nabla u_h + \theta(\nabla\varphi)^T A_u \nabla u]\, dx \\
&= \int_\Omega [(\nabla v_h)^T (u - u_h)\overline{A}_u \nabla u_h + \theta(\nabla\varphi)^T A_u \nabla u]\, dx \\
&= \int_\Omega [(\nabla v_h)^T (u - i_h^k u)\overline{A}_u \nabla u_h - \theta(\nabla v_h)^T \overline{A}_u \nabla u_h + \theta(\nabla\varphi)^T A_u \nabla u]\, dx \\
&= \int_\Omega [(\nabla v_h)^T (u - i_h^k u)\overline{A}_u \nabla (u_h - u) + (\nabla v_h)^T (u - i_h^k u)\overline{A}_u \nabla u]\, dx \\
&\quad + \int_\Omega [-\theta(\nabla v_h)^T \overline{A}_u \nabla u_h + \theta(\nabla\varphi)^T A_u \nabla u]\, dx,
\end{aligned}$$

that is,

$$\begin{aligned}
II_3 &= \int_\Omega (\nabla v_h)^T (u - i_h^k u)\overline{A}_u \nabla (u_h - u)\, dx \\
&\quad + \int_\Omega (\nabla (v_h - \varphi))^T (u - i_h^k u)\overline{A}_u \nabla u\, dx \\
&\quad + \int_\Omega (\nabla\varphi)^T (u - i_h^k u)\overline{A}_u \nabla u\, dx + \int_\Omega \theta(\nabla v_h)^T \overline{A}_u \nabla (u - u_h)\, dx \\
&\quad + \int_\Omega \theta(\nabla(\varphi - v_h))^T \overline{A}_u \nabla u\, dx + \int_\Omega \theta(\nabla\varphi)^T (A_u - \overline{A}_u)\nabla u\, dx \\
&=: \sum_{i=1}^6 II_3^i.
\end{aligned} \tag{3.15}$$

According to (3.12), (3.4), Theorem 3.1, and the imbedding $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$, we obtain

$$\begin{aligned} |II_3^1| &= \left| \int_\Omega (\nabla v_h)^T (u - i_h^k u) \overline{A}_u \nabla (u_h - u)\, dx \right| \\ &\le C \|u - i_h^k u\|_{0,\infty} \|v_h\|_1 \|u - u_h\|_1 \\ &\le C h^2 \|u\|_{2,\infty} \|\varphi\|_2 \|u - u_h\|_1 \\ &\le C h^2 \|u\|_4 \|\varphi\|_2\, h^k \|u\|_{k+1} \\ &\le C h^{k+2} \|\varphi\|_2 \|u\|_{k+2}. \end{aligned} \tag{3.16}$$

From (3.11) and the interpolation theorem again, we get

$$\begin{aligned} |II_3^2| &= \left| \int_\Omega (u - i_h^k u)(\nabla (v_h - \varphi))^T \overline{A}_u \nabla u\, dx \right| \\ &\le C \|u - i_h^k u\|_0 \|v_h - \varphi\|_1 \\ &\le C h^{k+2} \|\varphi\|_2 \|u\|_{k+2}. \end{aligned} \tag{3.17}$$

Let

$$\alpha(x) = (\nabla \varphi)^T \overline{A}_u \nabla u \qquad \text{and} \qquad \overline{\alpha}|_e = \frac{1}{|e|} \int_e \alpha(x)\, dx,$$

where $|e|$ is the area of the element e.

It is well known (see [5]) that

$$\|\alpha - \overline{\alpha}\|_0 \le C h \|\alpha\|_1. \tag{3.18}$$

Using the fact that the derivatives $\frac{\partial^2 a_{ij}}{\partial u \partial x_1}$, $\frac{\partial^2 a_{ij}}{\partial u \partial x_2}$, and $\frac{\partial^2 a_{ij}}{\partial u^2}$ are bounded, we shall prove now that

$$\|\alpha\|_1 \le C \|\varphi\|_2.$$

In fact, it follows from

$$\begin{aligned} \alpha(x) &= (\nabla \varphi)^T \overline{A}_u \nabla u = (\varphi_{x_1} \;\; \varphi_{x_2}) \int_0^1 A_u(x, u + t(u_h - u)) dt \begin{pmatrix} u_{x_1} \\ u_{x_2} \end{pmatrix} \\ &= \varphi_{x_1} u_{x_1} \int_0^1 (a_{11})_u (x, u + t(u_h - u)) dt + \varphi_{x_2} u_{x_1} \int_0^1 (a_{21})_u (x, u + t(u_h - u)) dt \\ &\quad + \varphi_{x_1} u_{x_2} \int_0^1 (a_{12})_u (x, u + t(u_h - u)) dt + \varphi_{x_2} u_{x_2} \int_0^1 (a_{22})_u (x, u + t(u_h - u)) dt \\ &= \sum_{i=1}^4 Y_i \end{aligned}$$

that

$$\alpha_{x_1}(x) = \sum_{i=1}^4 \frac{\partial Y_i}{\partial x_1}.$$

Moreover, we have

$$\begin{aligned}\frac{\partial Y_1}{\partial x_1} &= (\varphi_{x_1x_1}u_{x_1} + \varphi_{x_1}u_{x_1x_1})\int_0^1 (a_{11})_u(x, u + t(u_h - u))dt \\ &\quad + \varphi_{x_1}u_{x_1}\int_0^1 (a_{11})_{ux_1}(x, u + t(u_h - u))dt \\ &\quad + \varphi_{x_1}u_{x_1}\int_0^1 (a_{11})_{uu}(x, u + t(u_h - u)) \cdot (u_{x_1} + t(u_h - u)_{x_1})dt,\end{aligned}$$

which, together with the boundedness of $\frac{\partial^2 a_{11}}{\partial u \partial x_1}$ and $\frac{\partial^2 a_{11}}{\partial u^2}$ as well as the estimate $\|u - u_h\|_1 \le Ch\|u\|_2$ in Theorem 3.1, leads to

$$\int_\Omega \left(\frac{\partial Y_1}{\partial x_1}\right)^2 dx \le C\int_\Omega (\varphi^2_{x_1x_1} + \varphi^2_{x_1})dx.$$

By the same argument, we can obtain the similar estimates for $\frac{\partial Y_i}{\partial x_1}$ $(i = 2, 3, 4)$. Thus, we have

$$\int_\Omega \left(\frac{\partial \alpha}{\partial x_1}\right)^2 dx \le C\|\varphi\|_2^2,$$

and analogously we have

$$\int_\Omega \left(\frac{\partial \alpha}{\partial x_2}\right)^2 dx \le C\|\varphi\|_2^2,$$

which yields

$$\|\alpha\|_1 \le C\|\varphi\|_2.$$

From the definition of the interpolation operator i_h^k given by (3.3) we know that

$$\int_\Omega (u - i_h^k u)\overline{\alpha}\, dx = \sum_e \int_e (u - i_h^k u)\overline{\alpha}\, dx = 0,$$

which yields by (3.18) that

$$\begin{aligned}|II_3^3| &= \left|\int_\Omega (u - i_h^k u)\alpha(x)\, dx\right| \\ &= \left|\int_\Omega (u - i_h^k u)\overline{\alpha}\, dx + \int_\Omega (u - i_h^k u)(\alpha - \overline{\alpha})\, dx\right| \\ &\le \|u - i_h^k u\|_0 \|\alpha - \overline{\alpha}\|_0 \\ &\le Ch^{k+2}\|\alpha\|_1\|u\|_{k+2} \\ &\le Ch^{k+2}\|\varphi\|_2\|u\|_{k+2}.\end{aligned} \tag{3.19}$$

As to II_3^4 we have by means of the Hölder inequality that

$$\begin{aligned}|II_3^4| &= \left|\int_\Omega \theta(\nabla v_h)^T \overline{A}_u \nabla(u - u_h)\, dx\right| \\ &\le C\|\theta\|_{0,3}\|\nabla v_h\|_{0,6}\|\nabla(u - u_h)\|_0.\end{aligned}$$

Since

$$\|\theta\|_{0,3} \le C\|\theta\|_1 \qquad \text{and} \qquad H^2(\Omega) \hookrightarrow W_6^1(\Omega),$$

we further have according to (3.12) and Theorems 3.1 and 3.3 that

$$|II_3^4| \le Ch\|\theta\|_1\|v_h\|_2 \le Ch^{k+2}\|\varphi\|_2\|u\|_{k+2}. \tag{3.20}$$

It follows from the Hölder inequality again, the fact that $\nabla u \in (H^3(\Omega))^2 \hookrightarrow (C(\overline{\Omega}))^2$, and estimates (3.11) and (3.12) that

$$\begin{aligned} |II_3^5| &= \left| \int_\Omega \theta(\nabla(\varphi - v_h))^T \overline{A}_u \nabla u \, dx \right| \\ &\le C\|\theta\|_{0,3}\|\nabla(\varphi - v_h)\|_0\|\nabla u\|_{0,6} \\ &\le C\|\theta\|_1\|\varphi - v_h\|_1 \\ &\le Ch\|\theta\|_1\|\varphi\|_2 \le Ch^{k+2}\|\varphi\|_2\|u\|_{k+2}. \end{aligned} \tag{3.21}$$

Using similar arguments as before and the differentiability of A with respect to u up to order two and the substitution $z = st$, we find for any $x \in \Omega$ that

$$\begin{aligned} A_u(x,u) - \overline{A}_u(x) &= \int_0^1 [A_u(x,u) - A_u(x, u + t(u_h - u))]\, dt \\ &= \int_0^1 \left(\int_0^1 A_{uu}(x, u + st(u_h - u))t(u - u_h)\, ds \right) dt \\ &= (u_h - u)\int_0^1 \left(\int_0^t A_{uu}(x, u + z(u_h - u))\, dz \right) dt \\ &= (u_h - u)\int_0^1 \left(\int_z^1 A_{uu}(x, u + z(u_h - u))\, dt \right) dz \\ &= (u_h - u)\int_0^1 (1 - z) A_{uu}(x, u + z(u_h - u))\, dz \\ &=: (u_h - u)\overline{A}_{uu}(x). \end{aligned}$$

Therefore, we have by the Hölder inequality, the relation $\nabla u \in (H^3(\Omega))^2 \hookrightarrow (C(\overline{\Omega}))^2$, and Theorems 3.1 and 3.3 that

$$\begin{aligned} |II_3^6| &= \left| \int_\Omega \theta(\nabla\varphi)^T (A_u - \overline{A}_u)\nabla u \, dx \right| \\ &= \left| \int_\Omega \theta(u_h - u)(\nabla\varphi)^T \overline{A}_{uu} \nabla u \, dx \right| \\ &\le C\|\theta\|_{0,3}\|u_h - u\|_0\|\nabla\varphi\|_{0,6} \\ &\le Ch\|\theta\|_1\|\varphi\|_2 \le Ch^{k+2}\|\varphi\|_2\|u\|_{k+2} \end{aligned}$$

which, together with (3.15)–(3.21), leads to

$$|II_3| \le Ch^{k+2}\|\varphi\|_2\|u\|_{k+2}. \tag{3.22}$$

Combining (3.13), (3.14), and (3.22) with (3.10) implies

$$\|\theta\|_0^2 \le Ch^{k+2}\|\varphi\|_2\|u\|_{k+2}$$

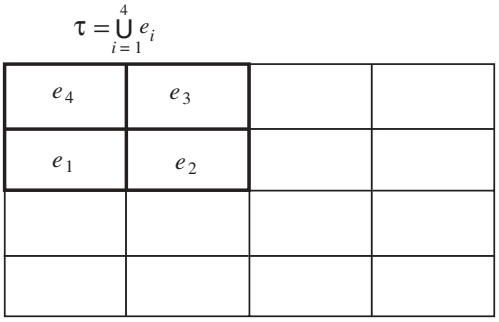


FIG. 3.1.

which, together with (3.2), leads to

$$\|\theta\|_0 \le Ch^{k+2}\|u\|_{k+2}.$$

Thus, the proof of the theorem is complete. □

In order to improve the accuracy of the finite element approximation in the whole domain, a simple postprocessing method is proposed [16, 17]. To this end, we need to define a postprocessing interpolation operator I_{2h}^{k+1} of degree at most $k+1$ in x_1- and x_2-direction. Thus, we assume that $\mathcal{T}_h$ has been obtained from $\mathcal{T}_{2h}$ with mesh size $2h$ by subdividing each element of $\mathcal{T}_{2h}$ into four congruent rectangles (see Figure 3.1). Let $\tau := \bigcup_{i=1}^{4} e_i \in \mathcal{T}_{2h}$ with $e_i \in \mathcal{T}_h$.

To express this idea clearly, we first consider the one-dimensional case, where I_{2h}^{k+1} $(k \ge 3)$ is determined by the following "vertex-interval" conditions [16]:

$$\begin{cases} I_{2h}^{k+1}u(p_i) = u(p_i), \quad i = 1,2,3, \\ \displaystyle\int_{l_i} I_{2h}^{k+1}u\,ds = \int_{l_i} u\,ds, \quad i = 1,2, \\ \displaystyle\int_{L} I_{2h}^{k+1}uv\,ds = \int_{L} uv\,ds \quad \forall v \in P_{k-3}(L)/P_0(L). \end{cases}$$

Here, $L := l_1 \cup l_2 \in \mathcal{T}_{2h}$, $l_i \in \mathcal{T}_h$, and p_i $(i = 1,2,3)$ are the vertices of l_1 and l_2. If $k < 3$, then I_{2h}^{2} is defined by $I_{2h}^{2}u(p_i) = u(p_i)$ $(i = 1,2,3)$, and I_{2h}^{3} by $I_{2h}^{3}u(p_i) = u(p_i)$ $(i = 1,2,3)$ and $\int_{l_1} I_{2h}^{3}u\,ds = \int_{l_1} u\,ds$ (or $\int_{l_2} I_{2h}^{3}u\,ds = \int_{l_2} u\,ds$). The operator I_{2h}^{k+1} in the two-dimensional case is now constructed by the tensor product of the two one-dimensional interpolation operators $I_{2h}^{k+1}(x_1)$ and $I_{2h}^{k+1}(x_2)$ of degree not exceeding $k+1$ in x_1- and x_2-direction, respectively, as follows:

$$I_{2h}^{k+1}(x_1,x_2) := I_{2h}^{k+1}(x_1) \otimes I_{2h}^{k+1}(x_2).$$

Moreover, the following properties can be checked [16, 17]:

$$\begin{cases} I_{2h}^{k+1}i_h^k = I_{2h}^{k+1}, \\ \|I_{2h}^{k+1}v_h\|_q \le C\|v_h\|_q \quad \forall v_h \in V_h,\ q = 0,1, \\ \|I_{2h}^{k+1}u - u\|_q \le Ch^{k+2-q}\|u\|_{k+2} \quad \forall u \in H^{k+2}(\Omega),\ q = 0,1. \end{cases} \tag{3.23}$$

Then, on the basis of Theorems 3.3 and 3.4 we obtain the following global superconvergence theorem.

THEOREM 3.5. *Let the assumptions of Theorem* 3.3 *hold. Then we have*

$$\|I_{2h}^{k+1}u_h - u\|_1 \le Ch^{k+1}\|u\|_{k+2}, \quad k \ge 1, \tag{3.24}$$

$$\|I_{2h}^{k+1}u_h - u\|_0 \le Ch^{k+2}\|u\|_{k+2}, \quad k \ge 2. \tag{3.25}$$

Proof. We find from the first property of the operator I_{2h}^{k+1} in (3.23) that

$$I_{2h}^{k+1}u_h - u = I_{2h}^{k+1}(u_h - i_h^k u) + (I_{2h}^{k+1}u - u).$$

Therefore, it follows from (3.23) and Theorem 3.3 that

$$\begin{aligned}\|I_{2h}^{k+1}u_h - u\|_1 &\le C\|u_h - i_h^k u\|_1 + Ch^{k+1}\|u\|_{k+2}\\ &\le Ch^{k+1}\|u\|_{k+2},\end{aligned}$$

and thus (3.24) is proven.

Estimate (3.25) can also be derived from Theorem 3.4 and the same arguments as those for deriving (3.24). ∎

4. A posteriori error estimators. It is of great importance for a finite element method to have a computable a posteriori error estimator by which we can assess the accuracy of the finite element solution in applications. One way to construct error estimators is to employ certain superconvergent approximations properties of finite element solutions. In fact, the following theorem holds.

THEOREM 4.1. *Under the assumptions of Theorem* 3.3 *we have*

$$\|u - u_h\|_1 = \|I_{2h}^{k+1}u_h - u_h\|_1 + O(h^{k+1}), \quad k \ge 1, \tag{4.1}$$

$$\|u - u_h\|_0 = \|I_{2h}^{k+1}u_h - u_h\|_0 + O(h^{k+2}), \quad k \ge 2. \tag{4.2}$$

In addition, if there exist positive constants C_1, C_2 *and* $\varepsilon_1, \varepsilon_2 \in (0,1)$ *such that*

$$\|u - u_h\|_1 \ge C_1 h^{k+1-\varepsilon_1}, \tag{4.3}$$

$$\|u - u_h\|_0 \ge C_2 h^{k+2-\varepsilon_2}, \tag{4.4}$$

then for the effectivity index we have

$$\lim_{h\to 0} \frac{\|I_{2h}^{k+1}u_h - u_h\|_1}{\|u - u_h\|_1} = 1, \tag{4.5}$$

$$\lim_{h\to 0} \frac{\|I_{2h}^{k+1}u_h - u_h\|_0}{\|u - u_h\|_0} = 1. \tag{4.6}$$

Proof. It follows from Theorem 3.5 and the equality

$$u - u_h = (I_{2h}^{k+1}u_h - u_h) + (u - I_{2h}^{k+1}u_h)$$

that

$$\|u - u_h\|_1 = \|I_{2h}^{k+1}u_h - u_h\|_1 + O(h^{k+1}).$$

Thus, by (4.3) we have

$$\frac{\|I_{2h}^{k+1}u_h - u_h\|_1}{\|u - u_h\|_1} + Ch^{\varepsilon_1} \ge 1$$

or

$$\lim_{h \to 0} \frac{\|I_{2h}^{k+1} u_h - u_h\|_1}{\|u - u_h\|_1} \geq 1. \tag{4.7}$$

Similarly, it follows from (4.3) and

$$\|I_{2h}^{k+1} u_h - u_h\|_1 = \|u - u_h\|_1 + O(h^{k+1})$$

that

$$\overline{\lim_{h \to 0}} \frac{\|I_{2h}^{k+1} u_h - u_h\|_1}{\|u - u_h\|_1} \leq 1,$$

which, together with (4.7), leads to (4.5).

Analogously, we can obtain (4.2) from Theorem 3.5 and (4.6) from condition (4.4). □

Remark 4.2. We know from (4.1) that the computable error estimate $\|I_{2h}^{k+1} u_h - u_h\|_1$ is the principal part of the finite element error $\|u - u_h\|_1$, and can be used as an a posteriori error estimator to assess the accuracy of the finite element solution. Estimate (4.3) seems to be a reasonable assumption, because $O(h^k)$ is the optimal convergence rate of the finite element solution in H^1-norm by Theorem 3.1, and from (4.5) we can further see that $\|I_{2h}^{k+1} u_h - u_h\|_1$ is a quite reliable a posteriori error estimator. The same comments are valid also for (4.2), (4.4), and (4.6).

5. Numerical experiments. In this section, we present some typical results obtained with numerical experiments carried out for the following boundary value problem:

$$\begin{aligned} -\nabla \cdot (A(x,u)\nabla u) &= f, \quad x = (x_1, x_2) \in \Omega, \\ u|_{\partial\Omega} &= 0, \end{aligned}$$

where $\Omega = (0,1) \times (0,1)$,

$$A(x_1, x_2, u) = \left(1 + x_1^2 + x_2^4 + \frac{1}{2}\sin u\right) I,$$

(cf. (1.4)) and f is chosen such that

$$u(x_1, x_2) = \sin(\pi x_1)\sin(2\pi x_2)$$

is the exact solution. The nonlinear problem is obtained with the aid of the standard Gauss quadrature formulae exact for all quintic polynomials and it is solved by Kačanov's method [14]. In the postprocessing we use again use the Gauss quadrature formulae. In the tables below listing the numerical results, we will use

$$E_0(u,v) = \sqrt{\int_\Omega (u-v)^2 \, dx},$$

$$E_1(u,v) = \sqrt{\int_\Omega (\nabla(u-v))^T \nabla(u-v) \, dx}$$

to denote the difference between two functions u and v. Obviously, $E_0(u,v)$ measures the difference in the usual L^2-norm, while $E_1(u,v)$ measures the difference in the usual H^1-seminorm.

TABLE 5.1
H^1-errors of the bilinear FE solutions and the corresponding postprocessed FE solutions.

h	$E_1(u_h, u)$	Reduction factor	$E_1(I_{2h}^2 u_h, u)$	Reduction factor
1/10	6.50e-1	None	2.30e-1	None
1/20	3.09e-1	2.1036	5.34e-2	4.3071
1/40	1.51e-1	2.0464	1.27e-2	4.2047
1/80	7.43e-2	2.0323	3.10e-3	4.0968
1/160	3.69e-2	2.0136	7.66e-4	4.0470

TABLE 5.2
L^2-errors of the biquadratic FE solutions and the corresponding postprocessed FE solutions.

h	$E_0(u_h, u)$	Reduction factor	$E_0(I_{2h}^3 u_h, u)$	Reduction factor
1/10	9.76e-4	None	5.40e-4	None
1/20	1.05e-4	9.2952	1.64e-5	32.927
1/40	1.21e-5	8.6777	7.81e-7	20.999
1/80	1.46e-6	8.2877	4.48e-8	17.433
1/160	1.79e-7	8.1564	2.95e-9	15.186

TABLE 5.3
H^1-errors of the biquadratic FE solutions and the corresponding postprocessed FE solutions.

h	$E_1(u_h, u)$	Reduction factor	$E_1(I_{2h}^3 u_h, u)$	Reduction factor
1/10	5.72e-2	None	1.60e-2	None
1/20	1.28e-2	4.4341	1.61e-3	9.9379
1/40	3.06e-3	4.2157	1.86e-4	8.6559
1/80	7.46e-4	4.1019	2.23e-5	8.3408
1/160	1.84e-4	4.0543	2.74e-6	8.1387

Table 5.1 lists the results for the approximations generated by the bilinear finite element (FE) solutions. We also calculated the problem for $h = 1/10$, $h = 1/20$, $h = 1/30$, ..., $h = 1/150$, $h = 1/160$. Using the linear regression, we find that the obtained data satisfy the following relations:

$$E_1(u_h, u) \approx 6.7634\ h^{1.0284},$$
$$E_1(I_{2h}^2 u_h, u) \approx 24.8825\ h^{2.0498}.$$

These approximations and the reduction factors in Table 5.1 confirm the theoretical results of Theorem 3.5 for the case in which $k = 1$.

Tables 5.2 and 5.3 list the results for the approximations generated by the biquadratic finite element basis functions ($k = 2$). Applying linear regression to the obtained data for $h = 1/10$, $h = 1/20$, $h = 1/30$, ... , $h = 1/150$, $h = 1/160$, we find that

$$E_0(u_h, u) \approx 1.1007\ h^{3.0858},$$
$$E_0(I_{2h}^3 u_h, u) \approx 6.8671\ h^{4.2865},$$
$$E_1(u_h, u) \approx 6.2018\ h^{2.0577},$$
$$E_1(I_{2h}^3 u_h, u) \approx 18.0236\ h^{3.1004},$$

These and the reduction factors in Tables 5.2 and 5.3 corroborate the predictions of Theorem 3.5 for the case in which $k = 2$.

TABLE 5.4
H^1-errors of the bilinear FE solutions and their estimates by the postprocessed FE solutions.

h	$E_1(u_h, u)$	$E_1(u_h, I_{2h}^2 u_h)$	Effectivity index
1/10	6.499e-1	6.704e-1	1.0315
1/20	3.089e-1	3.098e-1	1.0029
1/40	1.506e-1	1.506e-1	1.0000
1/80	7.434e-2	7.434e-2	1.0000
1/160	3.694e-2	3.694e-2	1.0000

TABLE 5.5
Errors of the biquadratic FE solutions and their estimates by the postprocessed FE solutions.

h	$E_0(u_h, u)$	$E_0(u_h, I_{2h}^3 u_h)$	Eff. index	$E_1(u_h, u)$	$E_1(u_h, I_{2h}^3 u_h)$	Eff. index
1/10	9.762e-4	1.043e-3	1.0684	5.724e-2	5.477e-2	0.95685
1/20	1.046e-4	1.039e-4	0.99331	1.289e-2	1.273e-2	0.98759
1/40	1.211e-5	1.208e-5	0.99752	3.062e-3	3.053e-3	0.99706
1/80	1.458e-6	1.456e-6	0.99863	7.463e-4	7.457e-4	0.99920
1/160	1.788e-7	1.788e-7	1.0000	1.842e-4	1.842e-4	1.0000

Table 5.4 compares the actual errors in the bilinear finite element solutions with their estimates by the postprocessed finite element solutions. Not only we can see that the computable quantity $E_1(u_h, I_{2h}^2 u_h)$ yields an accurate assessment of the actual error, but also, by applying linear regression to the data for $h = 1/10$, $h = 1/20$, $h = 1/30, \ldots, h = 1/150$, $h = 1/160$, we can see that the difference between the actual error and the estimated one satisfies the following relation:

$$|E_1(u_h, u) - E_1(u_h, I_{2h}^2 u_h)| \approx 6.3438\ h^{3.2954},$$

which is within the prediction of Theorem 4.1.

Table 5.5 compares the actual errors in the biquadratic finite element solutions and their estimates by the postprocessed finite element solutions. Again, we can see that both computable quantities $E_0(u_h, I_{2h}^3 u_h)$ and $E_1(u_h, I_{2h}^3 u_h)$ yield accurate assessments of the actual errors, and by applying linear regression to the data for $h = 1/10$, $h = 1/20$, $h = 1/30, \ldots, h = 1/150$, $h = 1/160$, we can see that the difference between the actual errors and those estimated satisfies the following relations:

$$|E_0(u_h, u) - E_0(u_h, I_{2h}^3 u_h)| \approx 3.0081\ h^{4.9526},$$
$$|E_1(u_h, u) - E_1(u_h, I_{2h}^3 u_h)| \approx 13.0755\ h^{3.8155},$$

which are again within the prediction of Theorem 4.1.

Acknowledgment. The authors would like to thank Professor Miloslav Feistauer, Karel Kolman, and Tomáš Vejchodský for their valuable suggestions.

REFERENCES

[1] J. BRANDTS, *Superconvergence and a posteriori error estimation in triangular mixed finite elements*, Numer. Math., 68 (1994), pp. 311–324.
[2] Z. CHEN, *On the existence, uniqueness and convergence of nonlinear mixed finite element methods*, Mat. Apl. Comput., 8 (1989), pp. 241–258.
[3] S. S. CHOW, G. F. CAREY, AND R. D. LAZAROV, *Natural and postprocessed superconvergence in semilinear problems*, Numer. Methods Partial Differential Equations, 7 (1991), pp. 245–259.

[4] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[5] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 2 (1975), pp. 77–84.
[6] J. Douglas and T. Dupont, *A Galerkin method for a nonlinear Dirichlet problem*, Math. Comp., 29 (1975), pp. 689–696.
[7] M. Feistauer and V. Sobotíková, *Finite element approximation of nonlinear elliptic problems with discontinuous coefficients*, RAIRO Modél. Math. Anal. Numér., 24 (1990), pp. 457–500.
[8] M. Feistauer and A. Ženíšek, *Compactness method in finite element theory of nonlinear elliptic problems*, Numer. Math., 52 (1988), pp. 147–163.
[9] J. Franců, *Weakly continuous operators. Applications to differential equations*, Appl. Math., 39 (1994), pp. 45–56.
[10] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.
[11] I. Hlaváček, *Reliable solution of a quasilinear nonpotential elliptic problem of a nonmonotone type with respect to the uncertainty in coefficients*, J. Math. Anal. Appl., 212 (1997), pp. 452–466.
[12] I. Hlaváček and M. Křížek, *On a nonpotential and nonmonotone second order elliptic problem with mixed boundary conditions*, Stability Appl. Anal. Contin. Media, 3 (1993), pp. 85–97.
[13] I. Hlaváček, M. Křížek, and J. Malý, *On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type*, J. Math. Anal. Appl., 184 (1994), pp. 168–189.
[14] M. Křížek and P. Neittaanmäki, *Mathematical and Numerical Modelling in Electrical Engineering: Theory and Applications*, Kluwer, Dordrecht, The Netherlands, 1996.
[15] M. Křížek and P. Neittaanmäki, *Bibliography on superconvergence*, in Proc. Conf. Finite Element Methods: Superconvergence, Post-processing and A Posteriori Estimates, Lecture Notes in Pure and Appl. Math. 196, Marcel Dekker, New York, 1998.
[16] Q. Lin and N. Yan, *The Construction and Analysis of Efficient Finite Elements*, Hebei University Publishers, Baoding, China, 1996.
[17] Q. Lin and S. Zhang, *An immediate analysis for global superconvergence for integrodifferential equations*, Appl. Math., 42 (1997), pp. 1–22.
[18] L. Liu, *Finite Element Analysis of Nonlinear Heat Conduction Problems*, Report 75, Department of Mathematics, University of Jyväskylä, Jyväskylä, Finland, 1997.
[19] L. Liu, M. Křížek, and P. Neittaanmäki, *Higher order finite element approximation of a quasilinear elliptic boundary value problem of a non-monotone type*, Appl. Math., 41 (1996), pp. 467–478.
[20] F. A. Milner, *Mixed finite element methods for quasilinear second-order elliptic problems*, Math. Comp., 44 (1985), pp. 303–320.
[21] J. Nečas, *Introduction to the Theory of Nonlinear Elliptic Equations*, Teubner, Leipzig, 1983.
[22] J. A. Nitsche, *On L_∞-convergence of finite element approximations to the solution of nonlinear boundary value problem*, in Proc. of Numer. Anal. Conf., J. H. Miller, ed., Academic Press, New York, 1977, pp. 317–325.
[23] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Lecture Notes in Math. 1605, Springer, Berlin, 1995.
[24] A. Ženíšek, *The finite element method for nonlinear elliptic equations with discontinuous coefficients*, Numer. Math., 58 (1990), pp. 51–77.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1745–1759

A LOCAL MINIMAX-NEWTON METHOD FOR FINDING MULTIPLE SADDLE POINTS WITH SYMMETRIES*

ZHI-QIANG WANG† AND JIANXIN ZHOU‡

Abstract. In this paper, a local minimax-Newton method is developed to solve for multiple saddle points. The local minimax method [*SIAM J. Sci. Comput.*, 23 (2001), pp. 840–865]. is used to locate an initial guess and a version of the generalized Newton method is used to speed up convergence. When a problem possesses a symmetry, the local minimax method is invariant to the symmetry. Thus the symmetry can be used to greatly enhance the efficiency and stability of the local minimax method. But such an invariance is sensitive to numerical error and the Haar projection has been used to enforce the symmetry [*SIAM J. Numer. Anal.*, submitted]. In this paper, we prove that the Newton method is invariant to symmetries and that such an invariance is insensitive to numerical error. When a symmetric degeneracy takes place, it is proved that the Newton direction can be easily solved in an invariant subspace. Thus the Newton method can be used not only to speed up convergence but also to avoid using the Haar projection if the symmetric degeneracy is removable by a discretization. Finally, numerical examples are presented to illustrate the theory.

Key words. saddle point, symmetry, group action, invariance, Newton's method

AMS subject classifications. 35A40, 35A15, 58E05, 58E30

DOI. 10.1137/S0036142903431675

1. Introduction.

Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and $J \in C^2(H, \mathbb{R})$, $J' : H \to H^*$ be its Frechet derivative and $\nabla J : H \to H$ be the gradient, and $J'' : H \to L(H, H^*)$ its second Fréchet derivative. Since there is a canonical identification between H^* and H, $\nabla J(u)$ is the identification of $J'(u)$. We may also use the identification of $J''(u)$ so $J''(u)$ is seen as in $L(H, H)$. A point $u \in H$ is a critical point of J if u solves the Euler–Lagrange equation $J'(u) = 0$. Many boundary value problems in nonlinear elliptic PDEs can be converted to solving its Euler–Lagrange equation for a critical point. A critical point u is nondegenerate if $J''(u)$ is invertible. The first candidates for a critical point are the local extrema which are well studied in the classical calculus of variation. Traditional numerical (variational) methods focus on finding such stable solutions. Critical points that are not local extrema are *unstable* and are called *saddle points*. In physical systems, saddle points appear as unstable equilibria or transient excited states. A huge number of papers exist in the literature on the existence of multiple saddle points in various nonlinear problems [1, 5, 6, 8, 9, 10, 17, 21, 24, 25, 26, 28].

To theoretical and computational physics and chemistry, saddle points between two stable states on the potential hypersurface are of great interests and lie in the theme of the so-called Transition State Theory or Activated Complex Theory, as they correspond to the transition states or the minimum energy paths between reactant molecules and product molecules [13]. A large literature can be found in this area.

*Received by the editors July 16, 2003; accepted for publication (in revised form) May 6, 2004; published electronically December 27, 2004.

http://www.siam.org/journals/sinum/42-4/43167.html

†Department of Mathematics and Statistics, Utah State University, Logan, UT 84322 (wang@math.usu.edu).

‡Department of Mathematics, Texas A&M University, College Station, TX 77843 (jzhou@math.tamu.edu). The research of this author was supported in part by NSF grant DMS-0311905.

Solitons arise in many fields, such as condensed matter physics, dynamics of biomolecules, nonlinear optics, etc. Among them, solutions which are not ground states, are the so-called excited states. In the study of self-guided light waves in nonlinear optics [11, 12, 19], excited states are of great interests. All those solitons are saddle points, thus unstable solutions.

On the other hand, symmetries exist in many natural phenomena, such as in crystals, elementary particle physics, symmetry of the Schrödinger equation for the atomic nucleus and the electron shell with respect to permutations and rotations, energy conservation law for systems which are invariant with respect to time translation, etc. Symmetries described by compact group actions in variational problems have been used in the literature to prove the existence of *multiple critical points*, typically, in the Ljusternik–Schnirelman theory (see, e.g., [14] and others). It is known that symmetries in a nonlinear variational problem can lead to the existence of many solutions of saddle type and can also cause (symmetric) degeneracy.

Due to the unstable nature, finding multiple saddle points numerically *in a stable way* is very challenging. There is virtually no theory in the literature to devise such a numerical algorithm until recently a local minimax method was developed in [15, 16] to find multiple saddle points in a sequential order of their Morse indices and its convergence was established in [16]. Techniques to enhance efficiency and stability of this method for computing saddle points with symmetries by using the Haar projection are developed in [27].

Since the local minimax method [15, 16] is a gradient type, first-order algorithm, to speed up convergence, it is quite natural to consider a Newton's method. Due to the instability and multiplicity nature of our problems, we consider a Newton's method of the form

$$u_{k+1} = u_k - s_k \nu_k \quad \text{with } \nu_k = (J''(u_k))^{-1} J'(u_k),$$

where ν_k is the Newton direction and $s_k > 0$ is a stepsize to enhance the stability of the algorithm; e.g., in Armijo's rule, $s_k > 0$ is chosen such that

$$\|\nabla J(u_{k+1})\| - \|\nabla J(u_k)\| < -\frac{1}{2} s_k \|\nabla J(u_k)\|. \tag{1.1}$$

For the algorithm to converge to a desirable critical point, two basic conditions are assumed:

(a) a good initial guess to start with, otherwise it can be extremely slow or divergent, or can lead to an unwanted trivial or known critical point;

(b) the problem has to be nondegenerate; i.e., $J''(u_k)$ is invertible along the trajectory of a Newton's method.

When $J''(u_k)$ is not invertible, a generalized Newton's method is suggested in the literature by using the generalized (Moore–Penrose) inverse $J''(u_k)^\dagger$, where the Newton direction $\nu_k = J''(u_k)^\dagger J'(u_k)$ is the least-norm solution to the minimization problem

$$\min_{\nu \in H} \|J''(u_k)\nu - J'(u_k)\|. \tag{1.2}$$

Under standard conditions and $s_k \equiv 1$, the generalized Newton method converges locally and quadratically [20]. This approach seems to be very general but also too complicated to apply to solve an infinite-dimensional problem. Therefore people tend to avoid using the generalized Newton method in solving variational problems. It is

also very difficult for us to examine its response to the effects of a symmetry in a problem.

Although attempts have been made by several researchers, e.g., [18, 22], to use a Newton's method to find multiple saddle points in various problems, the answer to how to deal with those two basic issues (a) and (b) remains largely unsatisfactory. Locating a good initial guess in an infinite-dimensional space is itself a challenging problem, in particular, when multiple solutions are involved. By using the local minimax method [15, 16], a good initial guess can be provided. However, degeneracy exists in every multiple saddle point problem due to a sign change of the eigenvalues of $J''(u)$. Either a solution to be found is degenerate or $J''(u)$ is not invertible at a point u along the Newton trajectory. How to handle such a case within the framework of a Newton's method remains a very interesting problem. On the other hand, when the problems possess some symmetries, they may create symmetric degeneracy; see Example 2.5. How a Newton's method responds to symmetries of the problems is, in general, still unknown. In this paper we shall try to address these questions. To do so, we use an approach somewhat between the standard and the generalized Newton method. When J is C^2 and $J''(u)$ has a closed range, for given $u \in H$, we consider a solution ν to

$$J''(u)\nu = J'(u). \tag{1.3}$$

In the following we assume that $J''(u)$ is a Fredholm operator with index zero. Since $J''(u)$ is self-adjoint, it has a finite-dimensional kernel, $\ker(J''(u))$, and a closed range. Then it is known that (1.3) may have none, unique, or infinitely many solutions, and (1.3) has a solution if and only if $\nabla J(u) \perp \ker(J''(u))$. In this case, *the Newton direction is just the least-norm solution* to the linear system (1.3). Note that in general, with the Armijo rule (1.1), the Newton method may approximate a critical point u^* of the function $g(u) = \|\nabla J(u)\|$, i.e.,

$$\langle g'(u^*), v\rangle = \frac{\langle J''(u^*)v, \nabla J(u^*)\rangle}{\|\nabla J(u^*)\|} = 0 \quad \forall v \in H.$$

If we choose $v = \nu$, a solution to (1.3), we have $J''(u^*)\nu = \nabla J(u^*)$ and $\langle g'(u^*), \nu\rangle = \|\nabla J(u^*)\| = 0$. Thus a critical point u^* of $g(u) = \|\nabla J(u)\|$ where (1.3) is solvable must be a critical point of J.

In this paper, we assume that a solution u^* to be found possesses certain symmetry and that the degeneracy of u^* is created *only* by the symmetry. Our method will be particularly useful in situations where there are multiple-saddle-point-type solutions due to symmetries. Our analysis uncovers the effects of symmetries in the problems on the Newton method. In summary, we shall undertake the following steps towards giving a theoretical strategy and implementing a numerical algorithm for computing multiple-saddle-point-type solutions when symmetries are present:

(1) prove the invariance of the Newton direction under symmetries;
(2) prove the solvability of (1.3) under symmetric degeneracies;
(3) show that the invariance of the Newton direction to symmetries is insensitive to numerical error, which contrasts to the fact that the invariance of the local minimax method to symmetries is sensitive to numerical error [27].

Due to the invariance of the local minimax method to symmetries, symmetries can be used to greatly enhance the efficiency and stability of the method [27]. However, such an invariance is sensitive to numerical error. Thus the Haar projection has to be used to enforce the symmetry. When a symmetry is associated with a continuous

group of actions, the corresponding Haar projection is an integral over the group. It becomes very difficult to compute. On the other hand, in many applications such as those examples in this paper, such a symmetric degeneracy is removable when a discretization of the problem is used. After the analysis in this paper we realize that with a least-norm solution linear solver, the Newton method can be used, following the local minimax method, to not only speed up convergence but also avoid using the Haar projection when the symmetric degeneracy is removable by a discretization. This is the *local minimax-Newton method* we shall describe in this paper. In the last section, we present several numerical examples to illustrate the theory.

2. The Newton method. We will need some preliminaries from transformation groups and invariant functionals [2]. Let H be a Hilbert space, $\mathcal{G}$ be a compact Lie group that acts isometrically on H, and $J \in C^2(H, \mathbb{R})$ be $\mathcal{G}$-invariant, i.e., $J(gu) = J(u)$, $\forall g \in \mathcal{G}$ and $u \in H$, and $J''(u)$ have a closed range for each $u \in H$. For a subgroup G of $\mathcal{G}$, let $H_G = \{u \in H \mid gu = u \ \forall g \in G\}$ be the invariant subspace of H under the group actions of G. For $u \in H$, the G-orbit of u is the set $Gu = \{gu : g \in G\}$ and the isotropy subgroup of u is $\mathcal{G}_u = \{g \in \mathcal{G} : gu = u\}$. When Gu is differentiable at u, we denote by $T_u(Gu)$ the tangent space of Gu at u.

2.1. Invariance and solvability of the Newton direction.

LEMMA 2.1.

(a) *∇J is $\mathcal{G}$-equivariant, i.e., $\nabla J(gu) = g^{-1}\nabla J(u)$ $\forall u \in H$, $g \in \mathcal{G}$;*

(b) *$\nabla J(u) \in H_G$ for any subgroup $G \subset \mathcal{G}$ and $u \in H_G$;*

(c) *$\langle J''(u)w, v\rangle = \langle J''(gu)gw, gv\rangle$ $\forall u, v, w \in H$, $g \in \mathcal{G}$, and in particular, we have $J''(u)(H_G) \subset H_G$ for any subgroup $G \subset \mathcal{G}$ and $u \in H_G$.*

Proof. By using the invariance of J, $\langle J'(gu), v\rangle = \langle J'(u), gv\rangle = \langle g^{-1}J'(u), v\rangle$. This shows $\nabla J(gu) = g^{-1}\nabla J(u)$, which implies (a) and (b). Let G be a subgroup of $\mathcal{G}$. To prove (c), differentiating again we have $\langle J''(u)w, v\rangle = \langle J''(gu)gw, gv\rangle$ $\forall u, v, w \in H$. For $u \in H_G$ and $w \in H_G$, we obtain $\langle J''(u)w, v\rangle = \langle g^{-1}J''(u)w, v\rangle$. Thus $J''(u)w = g^{-1}J''(u)w$ $\forall g \in G$, and we conclude $J''(u)w \in H_G$, i.e, $J''(u)(H_G) \subset H_G$. ∎

Lemma 2.1(a) states that if u^* is a critical point, i.e., $\nabla J(u^*) = 0$, then $\nabla J(gu^*) = 0$ $\forall g \in \mathcal{G}$. This implies that when $\mathcal{G}$ has a continuous subgroup G and $u^* \notin H_G$, the continuous orbit G_{u^*} is a critical point set continuous at u^*. Thus u^* is not isolated and therefore degenerate, i.e., $\ker(J''(u^*)) \neq \{0\}$. If G is differentiable subgroup of $\mathcal{G}$, we have the following lemma.

LEMMA 2.2. *Let G be a differentiable subgroup of $\mathcal{G}$ and $u \notin H_G$ be a critical point of J. Then $T_u(Gu) \subset \ker(J''(u))$. Here $T_u(Gu)$ is the tangent space of Gu at u.*

Proof. Let $v \in T_u(Gu)$ and consider a one-parameter curve $\gamma : (-\epsilon, \epsilon) \to Gu$ such that $\gamma(0) = u$ and $\gamma'(0) = v$. Then $J'(\gamma(t)) = 0$ $\forall t \in (-\epsilon, \epsilon)$. For any fixed w let $g : (-\epsilon, \epsilon) \to \mathbb{R}$ be defined by $g(t) = (J'(\gamma(t)), w)$. Then $g'(0) = 0$, but $g'(0) = (J''(u)v, w)$. Since w is arbitrary we have $J''(u)v = 0$. ∎

If u^* is a nondegenerate critical point of J, that is, $\ker(J''(u^*)) = \{0\}$, then $\ker(J''(u)) = \{0\}$ for u close to u^*. When the degeneracy of a critical point u^* is caused only by differentiable group actions of G, i.e., $\ker(J''(u^*)) = T_{u^*}(Gu^*)$, we must have $u^* \notin H_G$. Thus it is reasonable to assume that for u close to u^* and $u \in H \setminus H_G$, $\ker(J''(u)) \subset T_u(Gu)$ holds. Then we have the following lemma.

LEMMA 2.3. *Let G be a differentiable subgroup of $\mathcal{G}$ and $u \in H \setminus H_G$. If $\ker(J''(u)) \subset T_u(Gu)$, then (1.3) is always solvable.*

Proof. If $\nabla J(u) = 0$, the lemma is obvious. Let $v \in T_u(Gu)$ and consider a one-parameter curve $\gamma : (-\epsilon, \epsilon) \to Gu$ such that $\gamma(0) = u$ and $\gamma'(0) = v$. Let $g(t) = J(\gamma(t))$. Then $g'(0) = 0$ due to the invariance of the functional J. Since

$g'(0) = (\nabla J(u), v)$, we have $\nabla J(u) \perp T_u(Gu)$ and therefore $\nabla J(u) \perp \ker(J''(u))$ when $\ker(J''(u)) \subset T_u(Gu)$. So (1.3) is always solvable. ∎

The above result implies that the Newton direction ν of J at u can be solved from (1.3) instead of the much more complicated problem (1.2) when u is close to a critical point u^* whose degeneracy is caused only by differentiable subgroup actions of G. Can (1.3) be uniquely solved? Will the Newton direction ν of J at u have the same symmetry as that of u? These two uniqueness and invariance problems are actually closely related.

LEMMA 2.4. *Let G be a subgroup of $\mathcal{G}$ and $u \in H_G$. If $w \in H$ is a solution to* (1.3)*, then $w_G \in H_G$ is a solution of* (1.3) *where $w_G = \int_G g(w)dg$ is the Haar projection of w onto H_G. Thus the Newton direction is always in H_G. Furthermore, if* (1.3) *is uniquely solvable in H_G, then w_G is the Newton direction.*

Proof. Since $\nabla J(u) \in H_G$ by Lemma 2.1(b) and $w_G \in H_G$ by the Haar projection, we only have to prove that w_G is a solution to (1.3). By Lemma 2.1(c), we have

$$\langle \nabla J(u), v\rangle = \langle J''(u)w, v\rangle = \langle J''(gu)gw, gv\rangle \quad \forall v \in H,\ g \in \mathcal{G}.$$

Taking $v \in H_G$ and $g \in G$, we obtain $\langle J''(gu)gw, gv\rangle = \langle J''(u)gw, v\rangle$. Thus

$$\langle J''(u)gw - \nabla J(u), v\rangle = 0 \quad \text{or} \quad (J''(u)gw - \nabla J(u)) \perp H_G \quad \forall g \in G.$$

Since the Haar integral is linear and normalized, and $\nabla J(u) \in H_G$, it follows that

$$\int_G (J''(u)gw - \nabla J(u))\, dg = (J''(u)w_G - \nabla J(u)) \perp H_G$$

as well. Then by Lemma 2.1(c), $w_G \in H_G$ implies $J''(u)w_G - \nabla J(u) \in H_G$. We must have $J''(u)w_G - \nabla J(u) = 0$. When (1.3) is solvable, the Newton direction ν must be a solution to (1.3). It has been shown in [27] that the Haar projection ν_G of ν is the orthogonal projection of ν onto H_G and ν_G is also a solution to (1.3) by the previous part. We have $\|\nu_G\| \le \|\nu\|$ and the equality holds if and only if $\nu \in H_G$.

If (1.3) is uniquely solvable in H_G, which means for all solutions w of (1.3), their orthogonal projections w_G onto H_G are the same, then w_G is the Newton direction. ∎

We conclude here that finding w_G by the Haar projection is equivalent to solving the least-norm solution to the linear system (1.3).

2.2. Implementation of the Newton method. Let G be a differentiable subgroup of $\mathcal{G}$ and $u^* \in H \setminus H_G$ be a critical point to be found whose degeneracy is created only by the group actions of G. Thus $u^* \in H_{\mathcal{G}_{u^*}}$. Assume that each $u \in H_{\mathcal{G}_{u^*}}$ is an isolated point in $H_{\mathcal{G}_{u^*}} \cap Gu$.[1] Thus the degeneracy caused by the group actions of G does not take place in $H_{\mathcal{G}_{u^*}}$. It follows that the equation $J''(u)\nu = v$ has a unique solution ν in $H_{\mathcal{G}_{u^*}}$ $\forall u, v \in H_{\mathcal{G}_{u^*}}$. Therefore the uniqueness and invariance problems can be solved by confining our problem to the subspace $H_{\mathcal{G}_{u^*}}$. This implies that we have to enforce the symmetries defined by the isotropy subgroup $\mathcal{G}_{u^*}$. For numerical implementation, it can be easily done as follows.

Choose an initial guess $u_0 \in H_{\mathcal{G}_{u^*}}$ close to u^* (such that u_0 has the same symmetry as u^*). This can be done by the local minimax method due to its invariance to symmetries (see [27]). Then by Lemma 2.1, $J'(u_0) \in H_{\mathcal{G}_{u^*}}$ and (1.3) or $J''(u_0)\nu = J'(u_0)$ has a unique solution $\nu_0 \in H_{\mathcal{G}_{u^*}}$ which can be found through solving (1.3) for

[1] For most applications this assumption will be satisfied.

the least-norm solution. The updated solution $u_1 = u_0 - s_0 v_0 \in H_{\mathcal{G}_{u^*}}$ where $s_0 > 0$ is a stepsize determined by, e.g., Armijo's rule, has the same symmetry as that of u_0. Thus the symmetry of u_0 is preserved and passed to u_1 and we can continue this way to obtain the uniqueness and invariance of the Newton direction in $H_{\mathcal{G}_{u^*}}$. The local convergence of the generalized Newton method is then applied. When numerical error is considered, to overcome the symmetric degeneracy problem, in general, the Haar projection is needed to ensure the solvability of (1.3). The following example is instructional.

Example 2.5. Let $J(x,y) = \frac{1}{2}r^2 - \frac{1}{4}r^4$ where $r^2 = x^2 + y^2$. Then

$$J'(x,y) = \begin{bmatrix} x(1-r^2) \\ y(1-r^2) \end{bmatrix}, \qquad J''(x,y) = \begin{bmatrix} 1-2x^2-r^2 & -2xy \\ -2xy & 1-2y^2-r^2 \end{bmatrix},$$
$$\det(J''(x,y)) = (1-r^2)(1-3r^2).$$

Thus $(0,0)$ is the local minimum-type critical point and (x_s, y_s) with $x_s^2 + y_s^2 = 1$ are the saddle points. Let $\mathcal{G} = \mathbb{O}(2) = \mathbb{Z}_2 \times \mathbb{S}^1$ where $\mathbb{O}(2)$ is the group of all 2×2 orthogonal matrices, $\mathbb{Z}_2$ is generated by the matrix $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ and $\mathbb{S}^1$ is the group of all matrices $\left[\begin{smallmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{smallmatrix}\right]$, $0 \le \theta < 2\pi$. Thus the subgroup $\mathbb{Z}_2$ represents the reflection about the line $x = y$ and the subgroup $\mathbb{S}^1$ represents all rotations. The subgroup $G = \mathbb{S}^1$ is differentiable and creates degeneracy of a critical point not in H_G. We have $(0,0) \in H_G$ and $(x_s, y_s) \notin H_G$. It is clear that $(0,0)$ is a nondegenerate critical point with $\det(J''(0,0)) = 1$ and all the saddle points (x_s, y_s) are degenerate. For each $u = (x_s, y_s)$, $Gu = \{(x,y) : x^2 + y^2 = 1\}$ and $T_u(Gu) = \{(x,y) : x_s x + y_s y = 0\} = \{(x, -\frac{x_s x}{y_s}) : x \in \mathbb{R}\}$. By Lemma 2.2, $T_u(Gu) \subset \ker(J''(x_s, y_s))$. Indeed we have $J''(x_s, y_s)(x, -\frac{x_s x}{y_s})^T = (0,0)^T$.

Although for all (x,y) with $x^2 + y^2 \neq 1, \frac{1}{3}$, $J''(x,y)$ is invertible, the condition number of the matrix $J''(x,y)$ gets worse as $(x,y) \to (x_s, y_s)$. The usual Newton method will fail to provide any reliable solution. If we consider the saddle point $u^* = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$, the isotropy subgroup at u^* is $\mathbb{Z}_2$. The corresponding invariant subspace is $H_{\mathbb{Z}_2} = \{(x,y)^T\}$ such that $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$, i.e., $H_{\mathbb{Z}_2} = \{(x,x)^T\}$. By confining the problem in the subspace $H_{\mathbb{Z}_2}$, we have $J(x) = x^2 - x^4$, $J'(x) = 2x(1-2x^2)$, and $J''(x) = 2(1-6x^2)$. At the saddle point $x = \frac{\sqrt{2}}{2}$, $J''(\frac{\sqrt{2}}{2}) = -4$ is invertible. In implementation, for each $u = (z,z)^T \in H_{\mathbb{Z}_2}$, we have $H_{\mathbb{Z}_2} \cap Gu = \{u, -u\}$. Thus G will not cause any degeneracy in $H_{\mathbb{Z}_2}$. With $J'(z,z) = (z(1-2z^2), z(1-2z^2))^T \in H_{\mathbb{Z}_2}$, the equation $J''(z,z)(x,y)^T = J'(z,z)$ has a unique solution $(x,y) \in H_{\mathbb{Z}_2}$ where $x = y = \frac{z(1-2z^2)}{1-6z^2}$ and $-\frac{\sqrt{6}}{6} < z < \frac{\sqrt{6}}{6}$.

2.3. Insensitivity of invariance of Newton's method to numerical error. In [27], the invariance of the local minimax method to a symmetry is proved; i.e., if an initial guess u_0 is chosen in an invariant subspace H_G under a subgroup $G \subset \mathcal{G}$, then the sequence generated by the algorithm will remain in H_G. However, such an invariance is sensitive to numerical error in computing saddle points, because it searches a saddle point through a min-max method. The minimization process keeps J strictly descending along the sequence $\{u_k\}$ generated by the algorithm. To see the significant differences, let $u_k \in H_G$ be a point close to a saddle point $u^* \in H_G$. Thus $\nabla J(u_k)$ is small, the numerical errors in computing $\nabla J(u_k)$ dominate the symmetry of $\nabla J(u_k)$. This leads to $u_{k+1} \in H \setminus H_G$. For the minimax method, since u^* is a saddle point, the minimization search finds a slider (a descent direction) outside H_G away from u^*. Then $\|\nabla J(u_{k+1})\|$ increases and the asymmetric part of $\nabla J(u_{k+1})$

gets larger. Consequently the invariance of the sequence $\{u_k\}$ collapses and the search fails to reach u^*. The Haar projection has to be used (see [7, 27]) to preserve the symmetry of $\nabla J(u_k)$. In contrast to the local minimax method, the Newton method does not assume or use a variational structure. It finds a local minimum point u^*, not a saddle point, of $\|\nabla J(u)\|$. Once u_k is in a local basin of $\|\nabla J(u)\|$ around u^*, due to Armijo's rule, it keeps $\|\nabla J(u_k)\|$ strictly descending. Although $u_{k+1} \in H \setminus H_G$, $\|\nabla J(u_{k+1})\|$ is closer to 0. Thus u_{k+1} is still in the local basin around u^* and is a better approximation to $u^* \in H_G$. The asymmetric part of u_k will be kept within the norm of the numerical errors. In conclusion, the invariance of the Newton method is insensitive to numerical errors, therefore the Haar projection (an averaging formula) as suggested and used for the local minimax method in [27] is not necessary for the Newton method to preserve a symmetry.

The insensitivity of the invariance of the Newton method to numerical errors is double edged. If one knows the symmetry of a solution u^* to be found, then it is advantageous to use. One can choose an initial guess u_0 with the same symmetry of u^* to obtain an easy implementation for finding the Newton direction and preserve its invariance. Otherwise, it becomes a trap, when an initial guess u_0 has a symmetry different from that of u^*, the whole sequence generated by the Newton method will be trapped in the invariant subspace defined by the symmetry of u_0 and fails to reach u^*.

When a symmetry is associated with a continuous group G of actions, it causes degeneracy and the corresponding Haar projection is an integral over G and very difficult to compute. To overcome the symmetric degeneracy problem with numerical error, the Haar projection is needed in general. However, in many applications such as the examples in section 3, such a symmetric degeneracy is removable when a discretization is used, because after a discretization, G is approximated by a finite group. The truncation error is unpredictable, but it is in a much high order than the discretization error, which actually makes (1.3) more solvable. The above analysis suggests that in this case, the Haar projection is not needed to overcome the symmetric degeneracy problem with numerical error. Thus the Newton method can be used not only to speed up convergence but also to avoid using the Haar projection. This leads to the following *local minimax-Newton algorithm*.

2.4. A local minimax-Newton algorithm.

Step 1: Given $\varepsilon_M > \varepsilon_N > 0$ and $n-1$ previously found critical points $w_1, \dots, w_{n-1}$, of which w_{n-1} has the highest critical value. Set the support space $L = \text{span}\{w_1, \dots, w_{n-1}\}$. Let $v^1 \in L^\perp$ be an ascent direction at w_{n-1}. Let $t_0^0 = 1$, $v_L^0 = w_{n-1}$ and set $k = 0$;

Step 2: Using the initial guess $w = t_0^k v^k + v_L^k$, solve for $w^k = \arg\max_{u \in [L, v^k]} J(u)$ and denote $w^k = t_0^k v^k + v_L^k$ where t_0^k, v_L^k have been updated;

Step 3: Compute the negative gradient $d^k = -\nabla J(w^k)$;

Step 4: If $\|d^k\| \le \varepsilon_M$, then set $w^0 = w^k$, $k = 0$ and goto Step 7; else goto Step 5;

Step 5: Set $v^k(s^k) = \frac{v^k + s^k d^k}{\|v^k + s^k d^k\|}$ where s^k satisfies certain stepsize rule (see [15, 16]);

Step 6: Set $v^{k+1} = v^k(s^k)$ and update $k = k + 1$ then goto Step 2;

Step 7: Solve $J''(w^k)\nu = J'(w^k)$ for the least-norm solution ν^k;

Step 8: Set $w^{k+1} = w^k - s^k \nu^k$ where s^k satisfies, e.g., the Armijo's rule (1.1);

Step 9: Compute the gradient $\nabla J(w^{k+1})$;

Step 10: If $\|\nabla J(w^{k+1})\| < \varepsilon_N$, then output w^{k+1} and stop; else set $k = k + 1$, goto Step 7.

Steps 1–6 represent the local minimax method [15, 16] to locate an initial guess

that is sufficiently close to a desirable saddle point and Steps 7–10 represent the Newton method described in this paper to speed up the convergence.

When a symmetry is involved in a saddle point u^* to be found, we

(1) identify the symmetry of u^* by defining an invariant subspace H_G. Let $L_G = L \cap H_G$ and replace L by L_G in the algorithm; In many cases, such as those examples in section 3, we have $L_G = \{0\}$;

(2) choose an initial guess $v^1 \in H_G$;

(3) do iterations from Step 2 to Step 6.

Case 1. If we do not want to enforce the symmetry, we should choose $\varepsilon_M = 10\varepsilon$ where ε represents the order of the numerical error in computing $\nabla J(w^k)$, e.g., $\varepsilon = 10^{-2}$. Since the minimax method is invariant to a symmetry, when $\|\nabla J(w^k)\| > \varepsilon_M$, the symmetry of $\nabla J(w^k)$ still dominates the numerical error in $\nabla J(w^k)$. Usually the numerical error starts to dominate the symmetry of $\nabla J(w^k)$ when $\|\nabla J(w^k)\|$ is close to ε.

Case 2. If we want to enforce the symmetry, we only have to change Step 3 as $d^k = -\mathcal{H}(\nabla J(w^k))$ where $\mathcal{H}$ is the Haar projection defined in Lemma 2.4. In this case, we can choose $\varepsilon_M = 10\varepsilon$ or smaller.

(4) For Steps 7–10, if a degeneracy caused by a continuous group G of actions is removable by a discretization, then no Haar projection is needed, otherwise do the Haar projection.

3. Applications to semilinear elliptic equations.

3.1. Problems and setting up. The model equation we look at is the following semilinear elliptic equation:

$$\begin{cases} -\Delta u(x) = f(x, u(x)) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

where $\Omega \subset \mathbb{R}^N$ is bounded, f is a C^1 function satisfying certain growth and regularity conditions [23] and we seek weak solutions in $H = W_0^{1,2}(\Omega)$. The energy functional is

$$J(u) = \int_\Omega \left\{ \frac{1}{2}|\nabla u(x)|^2 - F(x, u(x)) \right\} dx, \quad \text{where } F(x,t) = \int_0^t f(x,\tau)d\tau. \tag{3.2}$$

Then critical points of $J(u)$ correspond to weak solutions of (3.1). Problems of this type appear as models in many applied areas. Mathematically, people have been interested in understanding the solution structures in terms of existence and nonexistence, the number of solutions as well as in obtaining qualitative property of solutions such as the geometric, symmetric, and nodal properties. Though great progress has been made, still many important open questions remain unsettled. Here, we are mainly concerned in uncovering new phenomena by numerically examining the qualitative behavior of both positive solutions and nodal solutions of this type of elliptic boundary value problem. For $u, w \in H$, we have

$$\langle J'(u), w\rangle_{H^*\times H} = \frac{d}{dt}\bigg|_{t=0} J(u + tw) = \int_\Omega \nabla u \nabla w - f(x, u(x))w\, dx.$$

Thus $d = \nabla J(u) = u - (-\Delta)^{-1} f(x,u) \in H$. Taking the second derivative, we have

$$\langle J''(u)\nu, w\rangle = \frac{d}{dt}\bigg|_{t=0} \langle J'(u + t\nu), w\rangle = \int_\Omega \nabla\nu\nabla w - f'_u(x, u(x))\nu w\, dx \quad \forall \nu \in H,$$

which implies that $J''(u) = I - (-\Delta)^{-1} f'_u(\cdot, u)$. Under standard conditions [23] on f, $(-\Delta)^{-1} f'_u(\cdot, u)$ is a compact operator and $J''(u)$ is a Fredholm operator with index zero. By setting $\langle J'(u), w\rangle = \langle J''(u)\nu, w\rangle \ \forall w \in H$, the Newton direction ν as defined in (1.3) can be obtained from weakly solving

$$\begin{cases} -\Delta\nu(x) - f'_u(x, u(x))\nu(x) = -\Delta u(x) - f(x, u(x)), & x \in \Omega, \\ \nu(x) = 0, & x \in \partial\Omega. \end{cases} \tag{3.3}$$

Remark 3.1. (a) Newton's method has been applied to variational problems in the literature usually by solving a discretized Euler–Lagrange equation. This approach requires to solve for $J'(u)$ and $J''(u)$, then compute $\nu = (J''(u))^{-1} J'(u)$, or, $\nu = (J''(u))^{\dagger} J'(u)$ when $J''(u)$ is not invertible, and therefore is much more computationally expensive and difficult. While solving the Newton direction ν directly from (3.3) is much simpler and less expensive. In many cases when $J''(u)$ is not invertible, ν is still solvable from (3.3), such as the case where the singularity of $J''(u)$ is caused only by a continuous group of actions.

(b) When an initial guess u_0 and its Laplacian Δu_0 are given, the Newton direction ν_0 is solved from (3.3) and s_0 is determined by, e.g., the Armijo rule. Then $u_1 = u_0 - s_0\nu_0$ and $\Delta u_1 = \Delta u_0 - s_0\Delta\nu_0$ where $\Delta\nu_0(x) = \Delta u_0(x) + f(x, u_0(x)) - f'_u(x, u_0(x))\nu_0(x)$ is known. Thus no computation of the Laplacian of the updated numerical solution u_1 is required.

3.2. Numerical examples. In this section, we apply the local minimax method (MM), the Newton method (NM) and the local minimax-Newton method (MM+NM) to numerically solve the Henon equation

$$\begin{cases} -\Delta u(x) = |x|^q u^3(x) & \text{in } \Omega, \\ u(x) = 0 & \text{on } \partial\Omega \end{cases} \tag{3.4}$$

for multiple solutions in $H = H_0^1(\Omega)$ where Ω is either the unit disk or an annulus. We are interested in finding new phenomena in symmetry breaking and nodal property of solution structure. The symmetries of the problem can be described by the group actions $\mathcal{G} = \mathbb{O}(2) = \mathbb{Z}_2 \times \mathbb{S}^1$ where $\mathbb{O}(2)$ is the set of all 2×2 orthogonal matrices, $\mathbb{Z}_2$ and $\mathbb{S}^1$ represent, respectively, the reflection about the x-axis and all the rotations. For $u \in H$, $g \in \mathbb{S}^1$, and the generator $\bar{h} \in \mathbb{Z}_2$, we define $g(u)(x) = u(gx)$ and $h(u)(x) = \pm u(\bar{h}x)$, where $+1$ and -1 represent, respectively, the even and the odd reflections, and the odd reflection is applicable if an even n-rotationally symmetry is considered. Then $\mathcal{G}$ becomes a compact Lie group that acts isometrically on H and $G = \mathbb{S}^1$ is a differentiable subgroup that creates degeneracy for a critical point $u^* \notin H_G$, i.e., u^* is radially asymmetric (or nonradial).

For a radially asymmetric but n-rotationally symmetric solution u^*, the isotropy subgroup of $\mathcal{G}$ at u^* is $\mathcal{G}_{u^*} = \{h^i g_i,\ i = 0, 1, \dots, n-1\}$ where $g_i = \left[\begin{smallmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{smallmatrix}\right]$ and $\theta_i = i\frac{2\pi}{n}$, $i = 0, 1, \dots, n-1$, and for each $u \in H_{\mathcal{G}_{u^*}}$, $H_{\mathcal{G}_{u^*}} \cap Gu = \{g_i u,\ i = 0, 1, \dots, n-1\}$. Thus the differentiable subgroup G causes no degeneracy in $H_{\mathcal{G}_{u^*}}$. By confining the problem in $H_{\mathcal{G}_{u^*}}$, the Newton direction can be uniquely solved from (3.3) in $H_{\mathcal{G}_{u^*}}$. For implementation, this means that we need only to take an initial guess u_0 in $H_{\mathcal{G}_{u^*}}$ and close to u^*. In the following numerical examples, $\varepsilon = \|\nabla J(u_k)\|$

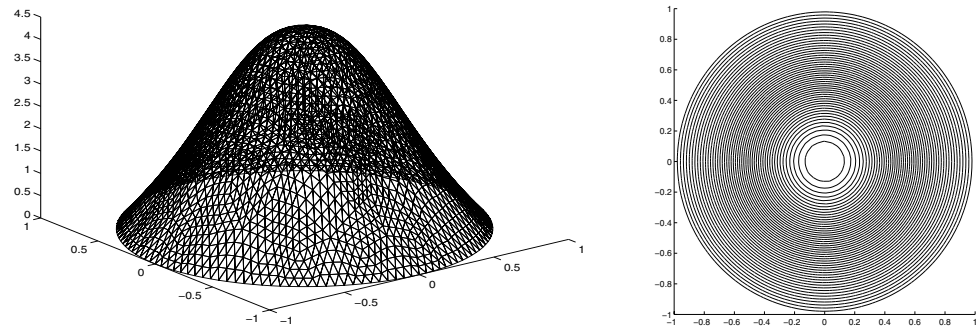


FIG. 1. $q = 0.5$. *The radially symmetric ground state with* $J = 21.5347$.

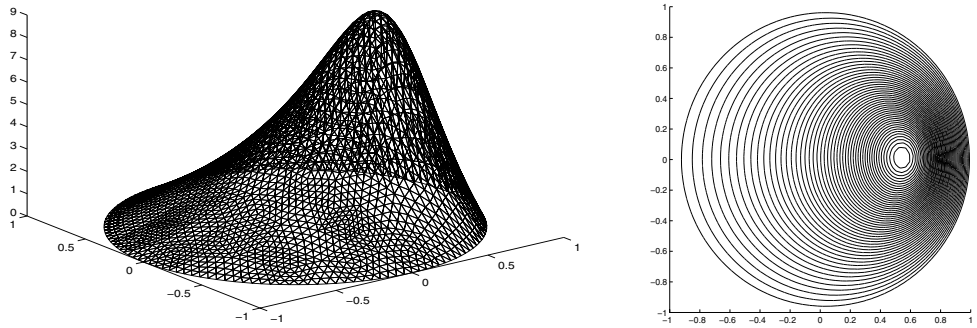


FIG. 2. $q = 2$. *A radially asymmetric ground state with* $J = 70.9280$.

and u_0 is computed from solving the linear equation

(3.5)

$$\begin{cases} -\Delta u_0(x) = c(x), & x \in \Omega, \\ u_0(x) = 0, & x \in \partial\Omega, \end{cases} \qquad \text{where } c(x) = \begin{cases} +1 & \text{if } u_0 \text{ is concave down at } x, \\ -1 & \text{if } u_0 \text{ is concave up a } x, \\ 0 & \text{otherwise.} \end{cases}$$

Case 1: $\Omega = \{(x_1, x_2) : x_1^2 + x_2^2 < 1\}$.

(1) Let $q = 0.5$ in (3.4). It is known that the equation has a unique positive solution which is radially symmetric as shown in Figure 1.

(a) Using an initial guess u_0 which is radially asymmetric but symmetric about the x-axis with $c(x_1, x_2) = -1$ if $|(x_1, x_2) - (0.5, 0)| \le 0.5$ and $c(x_1, x_2) = 0$ otherwise. Then NM failed to converge in 120 iterations and 35 MM iterations yield $\varepsilon < 10^{-4}$. While 6 MM iterations give $\varepsilon < 10^{-1}$ and then followed by 5 NM iterations, it yields $\varepsilon < 10^{-8}$.

(b) Using a radially symmetric initial guess u_0 from $c(x_1, x_2) = -1$. Then 5 NM iterations yield $\varepsilon < 10^{-12}$ and 8 MM iterations reach $\varepsilon < 10^{-4}$.

(2) Next let $q = 2$ in (3.4). Then the equation has a radially symmetric positive solution and other radially asymmetric positive solutions (see [3]). Rotating a radially

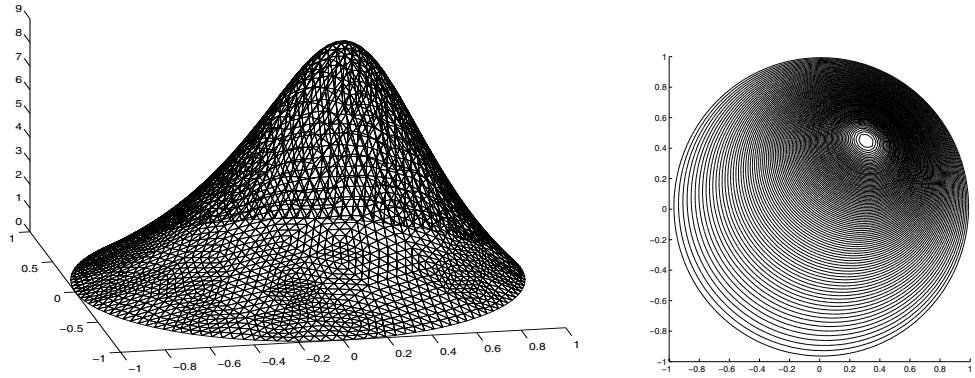


FIG. 3. $q = 2$. *Another radially asymmetric ground state with* $J = 70.8941$.

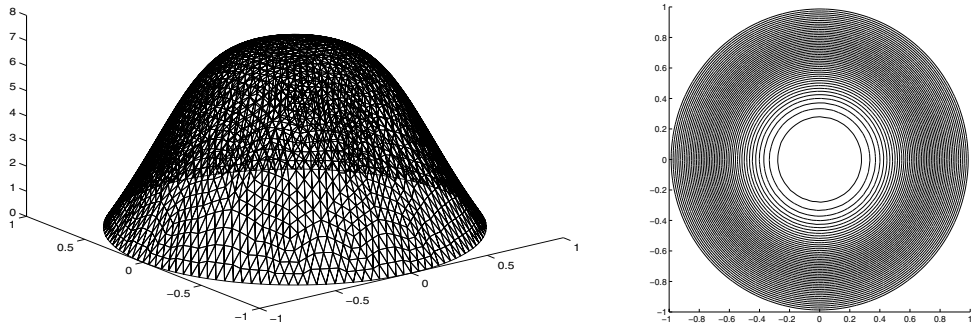


FIG. 4. $q = 2$. *The radially symmetric solution with* $J = 88.1740$.

asymmetric solution for any angle gives a radially asymmetric solution as well. Thus such a solution is degenerate. The radially symmetric positive solution has the highest energy among all the positive solutions. Without using the symmetry, such a solution is extremely elusive to capture.

(a) Using a radially asymmetric initial guess u_0 from $c(x_1, x_2) = -1$ if $|(x_1, x_2) - (0.5, 0)| \leq 0.5$ and $c(x_1, x_2) = 0$ otherwise. Then 11 MM iterations get the solution as in Figure 2 with $\varepsilon < 5 * 10^{-3}$ and 7 NM iterations find the same solution with $\varepsilon < 10^{-7}$.

(b) Using a radially symmetric initial guess u_0 from $c(x_1, x_2) = -1$. Then 21 MM iterations obtain the solution as in Figure 3 with $\varepsilon < 3 * 10^{-3}$, which is a rotation of the solution in Figure 2 and 4 NM iterations find the radially symmetric solution as in Figure 4 with $\varepsilon < 10^{-7}$. Such a solution cannot be captured by MM without enforcing the symmetry.

(c) Using an initial guess u_0 from $c(x_1, x_2) = -\text{sign}(x_1)$. u_0 is odd 2-rotation y-axis symmetric. NM failed to converge. Then first 2 MM iterations followed by 8 NM iterations yield a sign-changing solution as in Figure 5 with $\varepsilon < 10^{-7}$. Note that the solution in Figure 5 has the same symmetries as that of the initial guess u_0.

(d) To show that the invariance of NM is very insensitive to numerical error, using an initial guess u_0 from $c(x_1, x_2) = +1$ if $-\frac{1}{4}\pi < \tan^{-1}(\frac{x_2}{x_1}) < \frac{1}{4}\pi$ or $\frac{3}{4}\pi <$

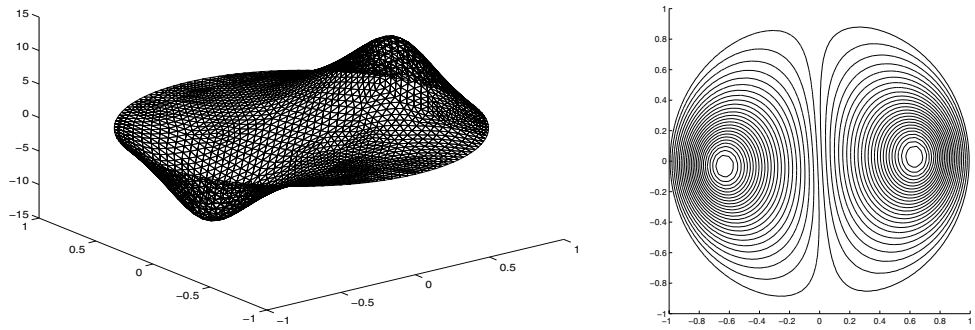


FIG. 5. $q = 2$. *An odd* 2*-rotationally symmetric sign-changing solution with* $J = 182.9987$.

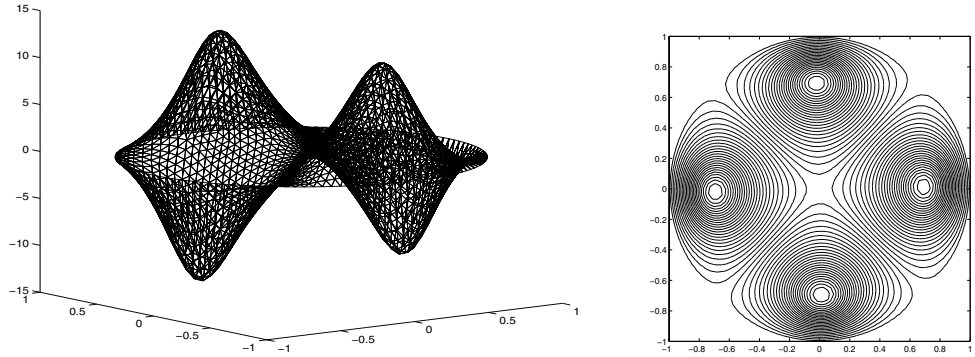


FIG. 6. $q = 2$. *An odd* 4*-rotationally symmetric sign-changing solution with* $J = 489.2240$.

$\tan^{-1}(\frac{x_2}{x_1}) < \frac{5}{4}\pi$ and $g(x_1, x_2) = -1$ otherwise. u_0 is odd 4-rotationally symmetric. The corresponding invariant subspace is much smaller. Again NM failed to converge. First 2 MM iterations followed by 9 NM iterations yield a solution as in Figure 6 with $\varepsilon < 10^{-11}$.

Case 2: $\Omega = \{(x_1, x_2) : 0.4 < x_1^2 + x_2^2 < 1\}$ and $q = 2$ in (3.4).

The equation has a radially symmetric and other radially asymmetric positive solutions (see [4]). Rotating a radially asymmetric solution for any angle is still a radially asymmetric solution. Thus such a solution is degenerate. The radially symmetric positive solution has the highest energy among all the positive solutions. Without using the symmetry, such a solution is extremely elusive to capture.

(a) Using a radially asymmetric initial guess u_0 from $c(x_1, x_2) = -1$ if $|(x_1, x_2) - (0.7, 0)| \le 0.3$ and $c(x_1, x_2) = 0$ otherwise. Then 17 MM iterations get the solution as in Figure 7 with $\varepsilon < 3 * 10^{-3}$ and 9 NM iterations yield the same solution with $\varepsilon < 10^{-9}$.

(b) Using an initial guess u_0 from $c(x_1, x_2) = -1$ if $-\frac{1}{4}\pi < \tan^{-1}(\frac{x_2}{x_1}) < \frac{1}{4}\pi$ or $\frac{3}{4}\pi < \tan^{-1}(\frac{x_2}{x_1}) < \frac{5}{4}\pi$ and $c(x_1, x_2) = 0$ otherwise. u_0 is even 2-rotationally symmetric. First 2 MM iterations followed by 6 NM iterations yield a solution as in Figure 8 with $\varepsilon < 10^{-11}$.

(c) Using an initial guess u_0 from $c(x_1, x_2) = -1$ if $-\frac{1}{6}\pi < \tan^{-1}(\frac{x_2}{x_1}) < \frac{1}{6}\pi$,

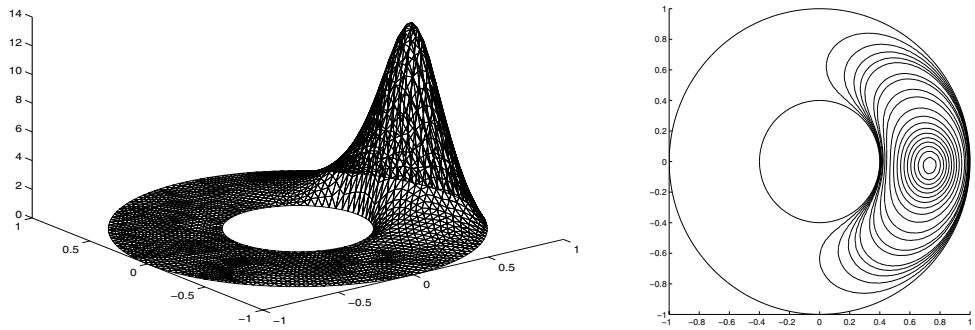


FIG. 7. $q = 2$. *A radially asymmetric ground state with* $J = 143.9674$.

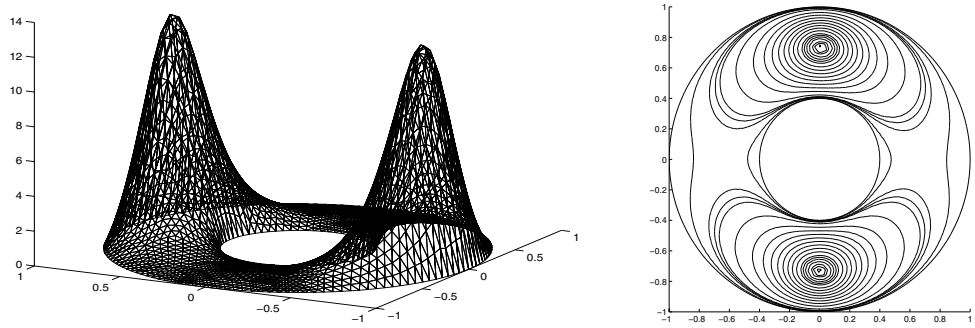


FIG. 8. $q = 2$. *An even* 2*-rotationally symmetric solution with* $J = 288.5556$.

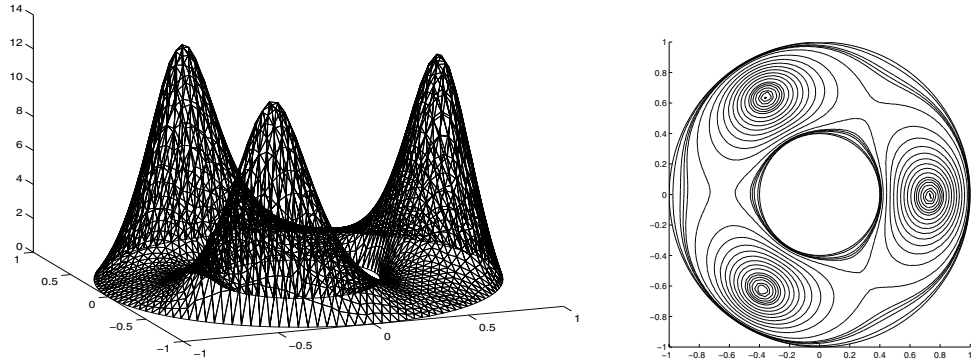


FIG. 9. $q = 2$. *A* 3*-rotationally symmetric solution with* $J = 429.9529$.

$\frac{1}{2}\pi < \tan^{-1}(\frac{x_2}{x_1}) < \frac{5}{6}\pi$ or $-\frac{5}{6}\pi < \tan^{-1}(\frac{x_2}{x_1}) < -\frac{1}{2}\pi$ and $c(x_1, x_2) = 0$ otherwise. u_0 is even 3-rotationally symmetric. First 2 MM iterations followed by 7 NM iterations yield a solution as in Figure 9 with $\varepsilon < 10^{-10}$.

(d) Using a radially symmetric initial guess u_0 by setting $c(x_1, x_2) \equiv -1$. Then

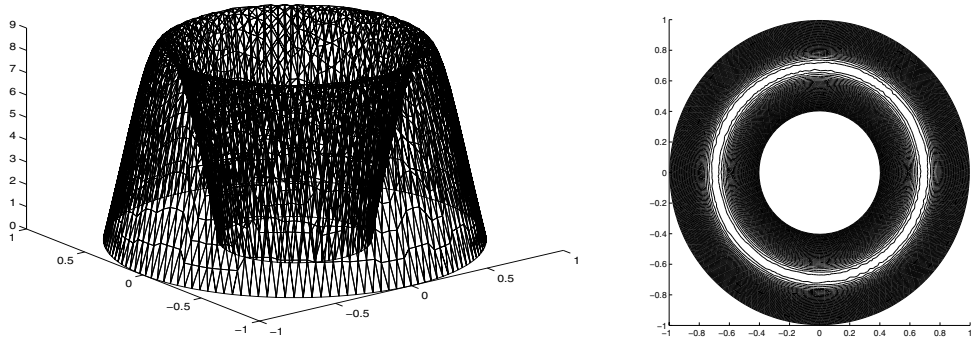


FIG. 10. $q = 2$. *The radially symmetric solution with* $J = 631.9575$.

4 NM iterations yield the radially symmetric solution as in Figure 10 with $\varepsilon < 10^{-7}$. But MM fails to find the solution without enforcing the symmetry.

For all numerical examples in this section, the Matlab PDE Toolbox is used to generate the domains and finite-element meshes and do computations. The Matlab function **assempde** is used to solve (3.3) for the Newton direction. Note that the degeneracy caused by symmetries in the examples is removable when a discretization is used. Since when the disk or annulus is discretized into finite-element grids, the continuous subgroup S^1 is approximated by a finite subgroup $S^1_n = \{g_i,\ i = 0, 1, \dots, n-1\}$ and the radial symmetry of the problem is approximated by the n-rotationally symmetry. With this approximation, the symmetric degeneracy of the problem is removed. Without a degeneracy, (3.3) is uniquely solvable and yields the Newton direction ν. By our analysis in section 2, (3.3) is solvable without numerical error and now it is also solvable with numerical error, therefore such an approximation or a refinement of discretization (finite-element grids) should be stable. Thus no Haar projection is needed.

With the local minimax-Newton algorithm, we are able to carry out many numerical investigations for examining the qualitative behavior and finding new phenomena of both positive and nodal solutions of nonlinear elliptic boundary value problems, e.g., the symmetry breaking and bifurcation phenomena, and the dependency of solutions on boundary approximation. We will address those new findings in subsequential papers.

Acknowledgment. The authors wish to thank two anonymous reviewers for their comments which stimulate us for this revision.

REFERENCES

[1] A. AMBROSETTI AND P. RABINOWITZ, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349–381.

[2] G. E. BREDON, *Introduction to Compact Transformation Groups*, Academic Press, New York, 1972.

[3] J. BYEON AND Z.-Q. WANG, *On the Hénon equation: Asymptotic profile of ground states*, Ann. Inst. H. Poincaré Anal. Non Linéaire, submitted.

[4] F. CATRINA AND Z.-Q. WANG, *Nonlinear elliptic equations on expanding symmetric domains*, J. Differential Equations, 156 (1999), pp. 153–181.

[5] K. C. Chang, *Infinite Dimensional Morse Theory and Multiple Solution Problems*, Birkhäuser, Boston, 1993.
[6] Y. S. Choi and P. J. McKenna, *A mountain pass method for the numerical solution of semilinear elliptic problems*, Nonlinear Anal., 20 (1993), pp. 417–437.
[7] D. Costa, Z. Ding, and J. Neuberger, *A numerical investigation of sign-changing solutions to superlinear elliptic equations on symmetric domains*, J. Comput. Appl. Math., 131 (2001), pp. 299–319.
[8] W. Y. Ding and W. M. Ni, *On the existence of positive entire solutions of a semilinear elliptic equation*, Arch. Ration. Mech. Anal., 91 (1986), pp. 238–308.
[9] Z. Ding, D. Costa, and G. Chen, *A high linking method for sign changing solutions for semilinear elliptic equations*, Nonlinear Anal., 38 (1999), pp. 151–172.
[10] I. Ekeland and N. Ghoussoub, *Selected new aspects of the calculus of variations in the large*, Bull. Amer. Math. Soc. (N.S.), 39 (2002), pp. 207–265.
[11] J. J. Garcia-Ripoll, V. M. Perez-Garcia, E. A. Ostrovskaya, and Y. S. Kivshar, *Dipole-mode vector solitons*, Phys. Rev. Lett., 85 (2000), pp. 82–85.
[12] J. J. García-Ripoll and V. M. Pérez-García, *Optimizing Schrödinger functionals using Sobolev gradients: Applications to quantum mechanics and nonlinear optics*, SIAM J. Sci. Comput., 23 (2001), pp. 1316–1334.
[13] G. Henkelman, G. Johannesson, and H. Jonsson, *Methods for Finding Saddle Points and Minimum Energy Paths*, Vol. 5, Comput. Chemistry, D. Schwartz, ed., Kluwer, Dordrecht, 2000.
[14] S. Li and Z.-Q. Wang, *Ljusternik-Schnirelman theory in partially orderes Hilbert spaces*, Trans. Amer. Math. Soc., 354 (2002), pp. 3207–3227.
[15] Y. Li and J. Zhou, *A minimax method for finding multiple critical points and its applications to semilinear PDEs*, SIAM J. Sci. Comput., 23 (2001), pp. 840–865
[16] Y. Li and J. Zhou, *Convergence results of a minimax method for finding multiple critical points*, SIAM J. Sci. Comput., 24 (2002), pp. 865–885.
[17] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, New York, 1989.
[18] J. More and T. Munson, *Computing mountain passes and transition states*, Math. Program., 100 (2004), pp. 151–182.
[19] Z. H. Musslimani, M. Segev, D. N. Christodoulides, and M. Soljacic, *Composite multi-hump vector solitons carrying topological charge*, Phys. Rev. Lett., 84 (2000), pp. 1164–1167.
[20] M. Z. Nashed and X. Chen, *Convergence of Newton like method for singular operator equations using outer inverse*, Numer. Math., 66 (1993), pp. 235–257.
[21] Z. Nehari, *On a class of nonlinear second-order differential equations*, Trans. Amer. Math. Soc., 95 (1960), pp. 101–123.
[22] J. Neuberger and J. Swift, *Newton's method and Morse index for semilinear elliptic PDEs*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 11 (2001), pp. 801–820.
[23] P. Rabinowitz, *Minimax Method in Critical Point Theory with Applications to Differential Equations*, CBMS Reg. Conf. Ser. Math. 65, AMS, Providence, RI, 1986.
[24] M. Schechter, *Linking Methods in Critical Point Theory*, Birkhäuser, Boston, 1999.
[25] M. Struwe, *Variational Methods*, Springer-Verlag, Berlin, 1996.
[26] Z.-Q. Wang, *On a superlinear elliptic equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 43–57.
[27] Z.-Q. Wang and J. Zhou, *An efficient and stable method for computing multiple saddle points with symmetries*, SIAM J. Numer. Anal., submitted.
[28] M. Willem, *Minimax Theorems*, Birkhäuser, Boston, 1996.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1760–1779

ON KOROBOV LATTICE RULES IN WEIGHTED SPACES*

XIAOQUN WANG†‡, IAN H. SLOAN‡, AND JOSEF DICK‡

Abstract. This paper studies the error bounds of multivariate integration in weighted function spaces using lattice rules of the Korobov form, in which the generating vector for an n-point rule with n prime has the form $(1, a, \ldots, a^{d-1}) (\mathrm{mod}\, n)$. With the parameter a chosen optimally, we establish new error bounds for Korobov lattice rules in weighted Korobov spaces. In particular, we prove that if the weights decay sufficiently fast, the optimal Korobov lattice rule has an error bound of order $O(n^{-\alpha/2+\delta})$ (for arbitrary $\delta > 0$), with the implied constant depending at worst polynomially on the dimension. Here $\alpha > 1$ is the smoothness parameter of the weighted Korobov spaces. We generalize the construction to the case where n is a product of arbitrary distinct prime numbers, with the purpose of reducing the construction cost without sacrificing much of the quality of the lattice rules. A corresponding result is deduced for weighted Sobolev spaces of nonperiodic functions, using randomly shifted optimal Korobov lattice rules. A comparison of the worst-case errors for Korobov lattice rules and the recent component-by-component constructions is presented. The investigation establishes the usefulness of (shifted) optimal Korobov lattice rules for integration, even in high dimensions, if the weights which characterize the weighted spaces are suitably chosen.

Key words. quasi-Monte Carlo methods, good lattice rules, multivariate integration

AMS subject classifications. 65C05, 65D30, 65D32

DOI. 10.1137/S0036142903425021

1. Introduction.

The computation of high-dimensional integrals plays a central role in many applications. Consider an integral over the d-dimensional unit cube

$$I_d(f) = \int_{[0,1)^d} f(\mathbf{x}) d\mathbf{x}.$$

We are mostly interested in cases where the dimension d is large. For large d, classical methods based on the Cartesian product of a one-dimensional integration rule (trapezoidal rule, Simpson's rule, Gaussian rule, etc.) are not practical because of the *curse of dimensionality*: the computational cost grows exponentially with the dimension. High-dimensional integrals are usually approximated by Monte Carlo or quasi-Monte Carlo (QMC) algorithms of the form

$$Q_{n,d}(f; \mathbf{P}_n) = \frac{1}{n} \sum_{k=0}^{n-1} f(\mathbf{x}_k),$$

where $\mathbf{P}_n = \{\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}\} \subset [0,1)^d$ is a set of random points in Monte Carlo, or a set of points chosen in some deterministic way in QMC.

In the present work we consider only the special class of QMC algorithms known as rank-1 lattice rules (see [15, 19]) having the form

$$Q_{n,d}(f; \mathbf{P}_n) = \frac{1}{n} \sum_{k=0}^{n-1} f\left(\left\{ \frac{k\mathbf{z}}{n} \right\} \right), \tag{1}$$

*Received by the editors March 25, 2003; accepted for publication (in revised form) March 26, 2004; published electronically December 27, 2004. This research was supported by the National Science Foundation of China and the Australian Research Council.

http://www.siam.org/journals/sinum/42-4/42502.html

†Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China (xwang@math.tsinghua.edu.cn).

‡School of Mathematics, University of New South Wales, Sydney 2052, Australia (i.sloan@unsw.edu.au, josi@maths.unsw.edu.au).

where $\mathbf{z} = (z_1, \ldots, z_d)$, called the "generating vector," is an integer vector with no factor in common with n, and the notation $\{\mathbf{x}\}$ means that each component of $\mathbf{x}$ is replaced by its fractional part. The accuracy of a lattice rule depends on the generating vector $\mathbf{z}$ and on the functions f.

How should the vector $\mathbf{z}$ in (1) be chosen? In this paper we present new algorithms for choosing $\mathbf{z}$, in the setting of the "weighted" spaces of functions introduced by Sloan and Woźniakowski in [22, 23], and with the choice of $\mathbf{z}$ restricted to the very classical form introduced by Korobov [12, 13]: if n is prime,

$$\mathbf{z} := (1, a, \ldots, a^{d-1}) \,(\mathrm{mod}\, n), \tag{2}$$

where a is a suitable integer from the set $\{1, \ldots, n-1\}$. The role of the corresponding algorithm is then to make a "good" choice of a. The theory to be developed here guarantees that the worst-case error of the resulting algorithm in an appropriately weighted space grows only slowly with the dimension d.

The results to be obtained here will be compared to those for the recently proposed "component-by-component" (or CBC) constructions for the vector $\mathbf{z}$, investigated in [14, 21]. The present theoretical results are almost, but not quite, as good as those for the CBC constructions. (For the CBC construction the bound for the worst-case error in an appropriately weighted space does not grow at all with d, whereas for the present construction a polynomial growth with d is the best we can prove.)

On the other hand, the present algorithm is both simpler and faster if the algorithm is needed only for a single value of dimension d. (The cost of the present algorithm is $O(dn^2)$, whereas the cost of the CBC algorithm is $O(d^2n^2)$.) Moreover, the observed performance in the two cases (see section 6) turns out to be quite similar, in spite of the difference between the theoretical bounds.

It should be emphasized that the resulting vector $\mathbf{z}$, though of the classical Korobov form (2), will almost always be quite different from that produced by the classical algorithm. In the classical theory of lattice rules (see [15, 19]) the function f is assumed to be a periodic function in the Korobov class $E_{\alpha,d}(c)$, for $c > 0$ and $\alpha > 1$, which is the class of functions with Fourier coefficients $\widehat{f}(\mathbf{h})$ satisfying

$$|\widehat{f}(\mathbf{h})| \le \frac{c}{(\bar{h}_1 \cdots \bar{h}_d)^\alpha},$$

where $\mathbf{h} = (h_1, \ldots, h_d)$ with integers h_j,

$$\hat{f}(\mathbf{h}) = \int_{[0,1]^d} \exp(-2\pi i \mathbf{h} \cdot \mathbf{x}) f(\mathbf{x}) d\mathbf{x},$$

$\mathbf{h} \cdot \mathbf{x} = h_1 x_1 + \cdots + h_d x_d$, and $\bar{h}_j = \max(1, |h_j|)$. The classical quality measure for rank-1 lattice rules is the worst-case quadrature error P_α in the Korobov class $E_{\alpha,d}(1)$:

$$P_\alpha := \sup\{|I_d(f) - Q_{n,d}(f; \mathbf{P}_n)| : f \in E_{\alpha,d}(1)\} = \sum_{\mathbf{h}\cdot\mathbf{z} \equiv 0 \,(\mathrm{mod}\, n)}{}' \frac{1}{(\bar{h}_1 \cdots \bar{h}_d)^\alpha},$$

where the prime on the sum indicates that the term $\mathbf{h} = \mathbf{0}$ is omitted. For a function $f \in E_{\alpha,d}(c)$, a simple error bound for a rank-1 lattice rule is (see [15, 19])

$$\left| I_d(f) - \frac{1}{n} \sum_{k=0}^{n-1} f\left(\left\{ \frac{k\mathbf{z}}{n} \right\} \right) \right| \le c\, P_\alpha.$$

It has been known for many years [15, 19] that there exists a "good" generating vector $\mathbf{z}$ such that

$$P_\alpha = O(n^{-\alpha}(\log n)^\beta)$$

for some β of the order of d (this is nearly the best possible, since Sarygin [18] proved a general lower bound which shows that P_α is at least of the order of magnitude $O(n^{-\alpha}(\log n)^{d-1})$). For fixed dimension d, this convergence is asymptotically much faster than the Monte Carlo convergence $O(n^{-1/2})$. The problem for larger values of d is that the error bound for functions from the classical Korobov classes, even for the optimal lattice rules, is *exponentially* dependent on d. Integration problems where the dimension d is large occur commonly in practice, for example, in finance [1, 17]. For large d (say, $d > 10$), the advantage of the classical Korobov lattice rules disappears in this setting. In fact, in high dimensions the optimal Korobov lattice rules constructed in the classical way (using P_α as the measure) may lead to error bounds that are worse, or much worse, than those for the Monte Carlo algorithm. The possible poor quality of P_α as a criterion was already pointed out in [6]. Generalizations of P_α were studied in [8].

In many applications the importance of successive variables decreases (see [27] for examples in finance, where it was shown that the sensitivity indices of successive variables decrease). It is also common that the lower-order "interactions" among variables are more important than higher-order "interactions." Such properties of functions often lead to small *effective dimension* in truncation or in superposition sense (see [1, 16, 26]). The problem of why high-dimensional problems in finance are often of low effective dimension was investigated in [28]. Intuitively, functions of small effective dimension might be easier to integrate by QMC methods.

To quantify the different importance of variables, Sloan and Woźniakowski [22, 23] introduced *weighted* spaces, where a sequence of weights is introduced to characterize the relative importance of successive variables, with the products of the weights for a group of variables relating to the importance of that group. They proved nonconstructively that there exists a lattice rule for which the worst-case error is bounded in d, or grows only slowly with d, provided the weights decay sufficiently rapidly. In this setting the CBC constructions (see above) achieve the same bound, which is known (see [23]) to be optimal.

This paper is organized as follows. In the next section, we introduce weighted Korobov spaces. In section 3, we study the error bounds in those spaces of the QMC algorithms based on the optimal Korobov lattice rules when n is prime. In section 4, we generalize the results to the case of n being a product of several distinct primes. A brief investigation on weighted Sobolev spaces of nonperiodic functions is presented in section 5. A numerical comparison of the worst-case errors is given in section 6. Concluding remarks are presented in the last section.

2. Weighted Korobov spaces. Let H_d be a Hilbert space of functions defined on $[0,1)^d$ with norm $||\cdot||_{H_d}$. Define the worst-case error of the algorithm $Q_{n,d}(f;\mathbf{P}_n)$ as the worst-case error over the unit ball of H_d:

$$e(\mathbf{P}_n; H_d) = \sup\{|I_d(f) - Q_{n,d}(f;\mathbf{P}_n)| : f \in H_d, ||f||_{H_d} \le 1\}.$$

The worst-case error $e(\mathbf{P}_n; H_d)$ is a natural quality measure of an algorithm.

We now briefly describe the "weighted Korobov spaces" introduced in [23]. They are (weighted) L_2 versions of the classical Korobov spaces.

Let $\mathbf{Z} = \{\ldots, -1, 0, 1, \ldots\}$ be the set of integers and let $\mathbf{Z}^d$ be the Cartesian product of d copies of $\mathbf{Z}$. We define a *weighted* Korobov space $H(K_{d,\alpha,\gamma})$ with the following reproducing kernel:

$$K_{d,\alpha,\gamma}(\mathbf{x},\mathbf{y}) = \prod_{j=1}^{d}\left(1+\gamma_j \sum_{h\in\mathbf{Z}}{}' \frac{e^{2\pi i h(x_j-y_j)}}{|h|^\alpha}\right), \qquad \alpha>1,\ i=\sqrt{-1}, \tag{3}$$

where $\gamma := \{\gamma_j\}$ is a sequence of positive numbers, which quantify the relative importance of successive variables. These function spaces are spaces of periodic functions with absolutely convergent Fourier series. The smoothness parameter $\alpha > 1$ characterizes the rate of decay of the Fourier coefficients. If $\gamma_j = 1$ for all j, then the space $H(K_{d,\alpha,\gamma})$ is the L_2 version of the classical (unweighted) Korobov space. The space $H(K_{d,\alpha,\gamma})$ is equipped with the inner product

$$\langle f, g\rangle = \sum_{\mathbf{h}\in\mathbf{Z}^d}\left[\prod_{j=1}^{d} r_\alpha(\gamma_j,h_j)\right]\ \hat{f}(\mathbf{h})\,\overline{\hat{g}(\mathbf{h})},$$

where

$$r_\alpha(\gamma,h) = \begin{cases} 1 & \text{if } h=0,\\ \gamma^{-1}|h|^\alpha & \text{if } h\neq 0.\end{cases}$$

Note that if a weight γ_j is small, then a function $f(\mathbf{x})$ with a norm at most 1 can depend only weakly on the jth variable. In this sense, the weights characterize the relative importance of variables.

For the lattice points (1) with the generating vector $\mathbf{z}$, the corresponding squared worst-case error can be expressed as (see [23])

$$e^2_{n,d}(\mathbf{z}) := e^2(\mathbf{P}_n; H(K_{d,\alpha,\gamma})) = -1+\frac{1}{n}\sum_{k=0}^{n-1}\prod_{j=1}^{d}\left(1+\gamma_j\sum_{h\in\mathbf{Z}}{}'\frac{\exp(2\pi i h k z_j/n)}{|h|^\alpha}\right). \tag{4}$$

For the case where α is an even integer, it is known that

$$\sum_{h\in\mathbf{Z}}{}'\frac{e^{2\pi i h x}}{|h|^\alpha} = \frac{(2\pi)^\alpha}{(-1)^{\frac{\alpha}{2}+1}\alpha!}\,B_\alpha(x), \qquad x\in[0,1], \tag{5}$$

where $B_\alpha(x)$ is the Bernoulli polynomial of degree α. In particular, $B_2(x) = x^2 - x + 1/6$. Thus the infinite sum in the formula of the worst-case error (4) can be computed easily when α is even.

3. The error bounds for optimal Korobov lattice rules. Suppose initially that n is prime. Consider the generators of Korobov form:

$$\mathbf{z}_1(a) := (1, a, \ldots, a^{d-1}) \,(\mathrm{mod}\, n) \ \text{ with } \ a\in\{1,2,\ldots,n-1\}. \tag{6}$$

Such generators have been used in the classical unweighted Korobov spaces, as mentioned in the introduction. Here we use these generators in the weighted Korobov spaces. The quality measure used is the worst-case error instead of the classical quantity P_α. The algorithm to find the optimal Korobov lattice rules in the weighted Korobov spaces is as follows.

ALGORITHM 1 (for prime n). *For each fixed dimension d and given weights $\{\gamma_j\}$, the optimal Korobov generator is found by minimizing the squared worst-case error*

$$e^2_{n,d}(\mathbf{z}_1(a)) = -1 + \frac{1}{n}\sum_{k=0}^{n-1}\prod_{j=1}^{d}\left(1+\gamma_j \sum_{h\in\mathbf{Z}}{}' \frac{\exp\left(2\pi i k h a^{j-1}/n\right)}{|h|^\alpha}\right)$$

with respect to $a \in \{1,2,\dots,n-1\}$, where $\mathbf{z}_1(a)$ is a vector in Korobov form (6).

We emphasize that the optimal Korobov generator depends on the weights γ_j. The number of operations needed to find the optimal Korobov lattice rule for even α and a single dimension d is $O(dn^2)$.

Let a_* be a minimizer of $e^2_{n,d}(\mathbf{z}_1(a))$. To see how small $e_{n,d}(\mathbf{z}_1(a_*))$ might be, we define

$$M_{n,d}(\alpha) := \frac{1}{n-1}\sum_{a=1}^{n-1} e^2_{n,d}\left(\mathbf{z}_1(a)\right), \tag{7}$$

the average of the squared worst-case error over all lattice rules of the Korobov form (6). We have the following results.

THEOREM 1. *Suppose n is a prime number and $M_{n,d}(\alpha)$ is defined by* (7)*; then*

$$M_{n,d}(\alpha) \le \frac{d}{n-1}\exp\left(2\zeta(\alpha)\sum_{j=1}^{d}\gamma_j\right),$$

where ζ is the Riemann zeta function, $\zeta(\alpha) = \sum_{h=1}^{\infty} h^{-\alpha}$ (for $\alpha > 1$). Hence, the optimal Korobov generator $\mathbf{z}_1(a_*) = (1, a_*, \dots, a_*^{d-1}) \pmod n$ obtained from Algorithm 1 satisfies

$$e^2_{n,d}(\mathbf{z}_1(a_*)) \le \frac{d}{n-1}\exp\left(2\zeta(\alpha)\sum_{j=1}^{d}\gamma_j\right). \tag{8}$$

For the proof it is convenient to make use of the following two lemmas. The first lemma gives an equivalent expression for the worst-case error $e_{n,d}(\mathbf{z})$. Similar expressions appeared in [19, 23].

LEMMA 2. *For the rank-1 lattice points* (1)*, the squared worst-case error in the weighted Korobov spaces $H(K_{d,\alpha,\gamma})$ can be written as*

$$e^2_{n,d}(\mathbf{z}) = \sum_{\mathbf{h}\in\mathbf{Z}^d}{}' \frac{\delta_n(\mathbf{h}\cdot\mathbf{z})}{\prod_{j=1}^{d} r_\alpha(\gamma_j, h_j)},$$

where

$$\delta_n(m) = \begin{cases} 1 & \text{if } m \equiv 0 \pmod n, \\ 0 & \text{if } m \not\equiv 0 \pmod n. \end{cases}$$

It is clear that if $\gamma_j = 1$ for all j, then $e^2_{n,d}(\mathbf{z}) = P_\alpha$. Thus the worst-case error is a natural generalization of the classical quality measure P_α.

The next lemma is well known in number theory (see [5]). It is significant when n is much larger than d (which is usually true in numerical integration) and is important for establishing the error bounds.

LEMMA 3. *Suppose that n is prime. Let $g(x) := h_1 + h_2 x + \cdots + h_d x^{d-1} = \mathbf{h} \cdot (1, x, \ldots, x^{d-1})$ be a polynomial with integer coefficients $h_1, \ldots, h_d$. Let $A_n(\mathbf{h})$ be the number of integers x with $0 \le x \le n-1$ satisfying $g(x) \equiv 0 \pmod n$. Let D be the greatest common divisor of $h_1, \ldots, h_d$. Then*

$$A_n(\mathbf{h}) \begin{cases} = n & \text{if } D \equiv 0 \pmod n, \\ \le (d-1) & \text{if } D \not\equiv 0 \pmod n. \end{cases}$$

Proof of Theorem 1. From Lemma 2 and the definition of $M_{n,d}(\alpha)$, we have

$$\begin{aligned} M_{n,d}(\alpha) &= \frac{1}{n-1} \sum_{a=1}^{n-1} {\sum_{\mathbf{h} \in \mathbf{Z}^d}}' \frac{\delta_n(\mathbf{h} \cdot \mathbf{z}_1(a))}{\prod_{j=1}^d r_\alpha(\gamma_j, h_j)} \\ &= \frac{1}{n-1} {\sum_{\mathbf{h} \in \mathbf{Z}^d}}' \prod_{j=1}^d [r_\alpha(\gamma_j, h_j)]^{-1} \sum_{a=1}^{n-1} \delta_n\left(\mathbf{h} \cdot (1, a, \ldots, a^{d-1})\right) \\ &=: S_1 + S_2, \end{aligned}$$

where in the last step the sum over $\mathbf{h}$ is split into two sums S_1 and S_2, with S_1 being the sum over all those $\mathbf{h}$ such that at least one component of $\mathbf{h}$ is not a multiple of n, and S_2 being the sum over all those nonzero $\mathbf{h}$ such that each component is a multiple of n.

For the first sum S_1, since at least one of the components of $\mathbf{h}$ is not a multiple of n, and since n is prime, we have $\gcd(h_1, \ldots, h_d) \not\equiv 0 \pmod n$. From Lemma 3, it follows that

$$\sum_{a=1}^{n-1} \delta_n\left(\mathbf{h} \cdot (1, a, \ldots, a^{d-1})\right) \le \sum_{a=0}^{n-1} \delta_n\left(\mathbf{h} \cdot (1, a, \ldots, a^{d-1})\right) = A_n(\mathbf{h}) \le d-1,$$

where $A_n(\mathbf{h})$ is defined in Lemma 3. Thus we have

$$\begin{aligned} S_1 &\le \frac{d-1}{n-1} {\sum_{\mathbf{h} \in \mathbf{Z}^d}}' \prod_{j=1}^d [r_\alpha(\gamma_j, h_j)]^{-1} \\ &< \frac{d-1}{n-1} \prod_{j=1}^d \left(1 + \gamma_j {\sum_{h \in \mathbf{Z}}}' |h|^{-\alpha}\right) \\ &= \frac{d-1}{n-1} \prod_{j=1}^d (1 + 2\zeta(\alpha)\gamma_j) \\ &= \frac{d-1}{n-1} \exp\left(\sum_{j=1}^d \log(1 + 2\zeta(\alpha)\gamma_j)\right) \\ &\le \frac{d-1}{n-1} \exp\left(2\zeta(\alpha) \sum_{j=1}^d \gamma_j\right). \end{aligned}$$

For the second sum S_2, since each component of the corresponding $\mathbf{h}$ is a multiple of n, i.e., $\mathbf{h} = (nm_1, \ldots, nm_d)$ for some nonzero integer vector $\mathbf{m} = (m_1, \ldots, m_d) \in \mathbf{Z}^d$, we have for arbitrary $a = 1, \ldots, n-1$ that

$$\delta_n\left(\mathbf{h} \cdot (1, a, \ldots, a^{d-1})\right) = \delta_n\left(n\mathbf{m} \cdot (1, a, \ldots, a^{d-1})\right) = 1.$$

Therefore,

$$\begin{aligned}
S_2 &= \frac{1}{n-1} {\sum_{\mathbf{m}\in\mathbf{Z}^d}}' \prod_{j=1}^{d} [r_\alpha(\gamma_j, nm_j)]^{-1} \sum_{a=1}^{n-1} \delta_n\left(n\mathbf{m}\cdot(1,a,\dots,a^{d-1})\right) \\
&= {\sum_{\mathbf{m}\in\mathbf{Z}^d}}' \prod_{j=1}^{d} [r_\alpha(\gamma_j, nm_j)]^{-1} \\
&= \prod_{j=1}^{d} \left(1 + \gamma_j {\sum_{m\in\mathbf{Z}}}' n^{-\alpha}|m|^{-\alpha}\right) - 1 \\
&= \prod_{j=1}^{d} \left(1 + \frac{2\zeta(\alpha)\gamma_j}{n^\alpha}\right) - 1 \\
&\le \frac{1}{n^\alpha} \prod_{j=1}^{d} (1 + 2\zeta(\alpha)\gamma_j) \\
&\le \frac{1}{n^\alpha} \exp\left(2\zeta(\alpha) \sum_{j=1}^{d} \gamma_j\right) \\
&< \frac{1}{n-1} \exp\left(2\zeta(\alpha) \sum_{j=1}^{d} \gamma_j\right).
\end{aligned}$$

Finally, we obtain

$$M_{n,d}(\alpha) = S_1 + S_2 \le \frac{d}{n-1} \exp\left(2\zeta(\alpha) \sum_{j=1}^{d} \gamma_j\right).$$

This completes the proof of the first part.

The second assertion is obvious: the squared worst-case error corresponding to the optimal Korobov generator $\mathbf{z}_1(a_*)$ is no larger than the average $M_{n,d}(\alpha)$; thus the bound (8) is satisfied. □

Note that the bound (8) is significant only when n is much larger than d (which is usually the case in numerical integration). If $\gamma_j = 1$ for all j as in the classical case, the upper bound (8) depends on d *exponentially*, whereas if $\sum_{j=1}^{\infty} \gamma_j < \infty$, then it depends only *linearly* on d. The factor of d is not present in the error bound for the global *optimal* rank-1 lattice rule in weighted Korobov spaces (see [23]). The convergence order shown in Theorem 1 is only $O(n^{-1/2})$ for fixed d, the same as that of the Monte Carlo algorithm. A higher convergence order can be proved if the weights decay sufficiently fast. In the optimal case, the convergence order can be $O(n^{-\alpha/2+\delta})$ (for any $\delta > 0$), with the implied constant depending at worst polynomially on d.

Before we state and prove the next theorem, we recall Jensen's inequality:

$$\sum_k A_k \le \left(\sum A_k^\beta\right)^{1/\beta} \quad \text{for } 0 < \beta \le 1,$$

where A_k are nonnegative numbers.

THEOREM 4. *Let n be a prime and $\mathbf{z}_1(a_*)$ be the optimal Korobov generator found by Algorithm* 1.

(i) *For arbitrary* $\tau \in [1, \alpha)$, *we have*

$$e_{n,d}(\mathbf{z}_1(a_*)) \le C_d(\alpha,\tau)\left(\frac{d}{n-1}\right)^{\tau/2}, \tag{9}$$

where

$$C_d(\alpha,\tau) = \exp\left(\tau\zeta(\alpha/\tau)\sum_{j=1}^{d}\gamma_j^{1/\tau}\right). \tag{10}$$

(ii) *Suppose the weight sequence* $\{\gamma_j\}$ *satisfies* $\sum_{j=1}^{\infty}\gamma_j < \infty$. *Let* τ_0 *be defined by*

$$\tau_0 := \sup\left\{\tau : \sum_{j=1}^{\infty}\gamma_j^{1/\tau} < \infty\right\}.$$

Then for any $\tau \in [1, \min(\tau_0, \alpha))$ (*or if* $\tau_0 = 1$, *put* $\tau = 1$), *we have*

$$C_d(\alpha,\tau) \le C_\infty(\alpha,\tau) := \lim_{d\to\infty} C_d(\alpha,\tau) < \infty;$$

i.e., $C_d(\alpha,\tau)$ *is uniformly bounded in* d *and the bound* (9) *is polynomial in dimension* d.

(iii) *Suppose that*

$$B^* := \limsup_{d\to\infty}\frac{\sum_{j=1}^{d}\gamma_j}{\log d} < \infty.$$

Then for any $\delta > 0$, *there exists a constant* C_δ *independent of* d *and* n, *such that*

$$e_{n,d}(\mathbf{z}_1(a_*)) \le C_\delta\, d^{1/2+\zeta(\alpha)(B^*+\delta)}(n-1)^{-1/2}.$$

Proof. We introduce temporarily an alternative notation $e_{n,d}(\mathbf{z},\alpha,\gamma)$ for the worst-case error $e_{n,d}(\mathbf{z})$ to stress its dependence on the parameter α and the weight sequence $\{\gamma_j\}$. According to Lemma 2, we have

$$e^2_{n,d}(\mathbf{z},\alpha,\gamma) = \sum_{\mathbf{h}\in\mathbf{Z}^d}{}' \frac{\delta_n(\mathbf{h}\cdot\mathbf{z})}{\prod_{j=1}^{d} r_\alpha(\gamma_j,h_j)} = \sum_{\mathbf{h}\cdot\mathbf{z}\equiv 0 \,(\mathrm{mod}\, n)}{}' \prod_{j=1}^{d}[r_\alpha(\gamma_j,h_j)]^{-1}.$$

By applying the Jensen inequality to the sum on the right-hand side, we have

$$e^2_{n,d}(\mathbf{z},\alpha,\gamma) \le \left(\sum_{\mathbf{h}\cdot\mathbf{z}\equiv 0 \,(\mathrm{mod}\, n)}{}' \prod_{j=1}^{d}[r_\alpha(\gamma_j,h_j)]^{-\beta}\right)^{1/\beta} = \left(e^2_{n,d}(\mathbf{z},\alpha\beta,\gamma^\beta)\right)^{1/\beta} \tag{11}$$

for $\frac{1}{\alpha} < \beta \le 1$, where γ^β means the weight sequence with values γ_j^β, and we use the relation that $[r_\alpha(\gamma,h)]^\beta = r_{\alpha\beta}(\gamma^\beta,h)$.

We see from Theorem 1, with α replaced by $\alpha\beta$ and γ_j replaced by γ_j^β, that there exists a generator of Korobov form $\overline{\mathbf{z}}_\beta = (1,\overline{a},\ldots,\overline{a}^{d-1}) \,(\mathrm{mod}\, n)$ such that

$$e^2_{n,d}(\overline{\mathbf{z}}_\beta,\alpha\beta,\gamma^\beta) \le \frac{d}{n-1}\exp\left(2\zeta(\alpha\beta)\sum_{j=1}^{d}\gamma_j^\beta\right).$$

(Note that since $\frac{1}{\alpha} < \beta$, we have $\alpha\beta > 1$ and thus $\zeta(\alpha\beta)$ is finite.) Thus for this $\overline{\mathbf{z}}_\beta$, we have from (11)

$$e^2_{n,d}(\overline{\mathbf{z}}_\beta, \alpha, \gamma) \le \left(\frac{d}{n-1}\right)^{1/\beta} \exp\left(\frac{2}{\beta}\zeta(\alpha\beta)\sum_{j=1}^{d}\gamma_j^\beta\right).$$

Therefore, for the optimal Korobov form generator $\mathbf{z}_1(a_*)$ we have for any $\frac{1}{\alpha} < \beta \le 1$ that

$$e^2_{n,d}(\mathbf{z}_1(a_*), \alpha, \gamma) \le \left(\frac{d}{n-1}\right)^{1/\beta} \exp\left(\frac{2}{\beta}\zeta(\alpha\beta)\sum_{j=1}^{d}\gamma_j^\beta\right). \tag{12}$$

For any $\tau \in [1, \alpha)$, putting $\beta = 1/\tau$ in (12), the error bound (9) follows immediately.

Now we prove (ii). Since $\sum_{j=1}^{\infty}\gamma_j < \infty$, we have $\tau_0 \ge 1$. If $\tau_0 = 1$, the result is trivial. If $\tau_0 > 1$, then for any $\tau \in [1, \min(\tau_0, \alpha))$, we have

$$C_d(\alpha, \tau) \le \exp\left(\tau\zeta(\alpha/\tau)\sum_{j=1}^{\infty}\gamma_j^{1/\tau}\right) := C_\infty(\alpha, \tau).$$

Clearly, $C_\infty(\alpha, \tau)$ is a constant independent of d and n. It is well defined: $\zeta(\alpha/\tau)$ is finite since $\tau < \alpha$, and $\sum_{j=1}^{\infty}\gamma_j^{1/\tau}$ converges since $\tau < \tau_0$. Thus the constant $C_d(\alpha, \tau)$ is uniformly bounded in d, and the bound (9) is polynomial in d.

Now we prove (iii). Since B^* is finite, we have that for any $\delta > 0$ there exists d_δ such that

$$\sum_{j=1}^{d}\gamma_j \le (B^* + \delta)\log d \quad \forall d \ge d_\delta.$$

Hence, from Theorem 1 we have for $d \ge d_\delta$ that

$$M_{n,d}(\alpha) \le \frac{d}{n-1}\exp\left(2\zeta(\alpha)(B^* + \delta)\log d\right) = \frac{1}{n-1}\, d^{1+2\zeta(\alpha)(B^*+\delta)}.$$

Thus there exists a constant C_δ^2 such that for all $d \ge 1$

$$M_{n,d}(\alpha) \le C_\delta^2\, d^{1+2\zeta(\alpha)(B^*+\delta)}(n-1)^{-1}.$$

Therefore, the optimal Korobov lattice rule satisfies

$$e^2_{n,d}(\mathbf{z}_1(a_*)) \le C_\delta^2\, d^{1+2\zeta(\alpha)(B^*+\delta)}(n-1)^{-1}.$$

This concludes the proof. ∎

4. Korobov lattice rules with nonprime number of points. The number of operations to find an optimal Korobov lattice rule needed in Algorithm 1 is $O(dn^2)$. For large n, this is expensive. The search cost can be substantially reduced if we allow the number of points to be a nonprime number. In this section we study the cases where the number of points n is large and is the product of distinct primes. We first consider the case of two distinct prime numbers, then we generalize to the case of an arbitrary number of distinct primes.

4.1. The case of two distinct primes. Consider the case where $n = pq$, with p and q being two distinct primes. The Korobov construction [13] can also be generalized to the weighted Korobov spaces. The algorithm is as follows.

ALGORITHM 2 (for $n = pq$).

(i) *Find the optimal* $a_* \in \{1, 2, \ldots, p-1\}$ *by using Algorithm* 1 *in section* 3, *but with* n *replaced by* p.

(ii) *Let the generating vector* $\mathbf{z}_2(b)$ *be of the form*

(13)
$$\mathbf{z}_2(b) := \left(p\,(1, b, \ldots, b^{d-1}) + q\,(1, a_*, \ldots, a_*^{d-1})\right) (\mathrm{mod}\ pq), \quad b \in \{1, 2, \ldots, q-1\},$$

where $a_* \in \{1, 2, \ldots, p-1\}$ *is the number obtained in step* (i). *Find the optimal* $b_* \in \{1, \ldots, q-1\}$ *such that the squared worst-case error*

$$e^2_{pq,d}(\mathbf{z}_2(b)) = -1 + \frac{1}{pq} \sum_{k=0}^{pq-1} \prod_{j=1}^{d} \left(1 + \gamma_j \sum_{h \in \mathbf{Z}}{}' \frac{\exp[2\pi i h k (p b^{j-1} + q a_*^{j-1})/(pq)]}{|h|^\alpha}\right)$$

is minimized with respect to b. *We obtain the optimal Korobov generator* $\mathbf{z}_2(b_*)$.

In order to study the worst-case error corresponding to the optimal Korobov generator $\mathbf{z}_2(b_*)$, we define

$$\widetilde{M}_{n,d}(\alpha) := \frac{1}{q-1} \sum_{b=1}^{q-1} e^2_{n,d}\left(\mathbf{z}_2(b)\right), \tag{14}$$

which is the average (over all b) of the squared worst-case error of the lattice rules with the generators of the form (13). We have the following bound.

THEOREM 5. *Suppose* $n = pq$, *with* p *and* q *being two distinct primes. Let* $\widetilde{M}_{n,d}(\alpha)$ *be defined by* (14); *then*

$$\widetilde{M}_{n,d}(\alpha) \le \frac{d^2}{(p-1)(q-1)} \exp\left(2\zeta(\alpha) \sum_{j=1}^{d} \gamma_j\right).$$

The squared worst-case error $e^2_{n,d}(\mathbf{z}_2(b_*))$ *corresponding to the optimal Korobov generator* $\mathbf{z}_2(b_*)$ *satisfies the same upper bound.*

Proof. From Lemma 2 we have

$$e^2_{n,d}(\mathbf{z}_2(b)) = \sum_{\mathbf{h} \in \mathbf{Z}^d}{}' \frac{\delta_{pq}(\mathbf{h} \cdot \mathbf{z}_2(b))}{\prod_{j=1}^{d} r_\alpha(\gamma_j, h_j)}. \tag{15}$$

Note that for integers n_1 and n_2, if $\gcd(p, q) = 1$ (which is certainly satisfied if p and q are distinct prime numbers), then $n_1 p + n_2 q \equiv 0 \pmod{pq}$ is equivalent to $n_1 \equiv 0 \pmod{q}$ and $n_2 \equiv 0 \pmod{p}$. Thus

$$\begin{aligned}\delta_{pq}(\mathbf{h} \cdot \mathbf{z}_2(b)) &= \delta_{pq}\left(p\,\mathbf{h} \cdot (1, b, \ldots, b^{d-1}) + q\,\mathbf{h} \cdot (1, a_*, \ldots, a_*^{d-1})\right) \\ &= \delta_p\left(\mathbf{h} \cdot (1, a_*, \ldots, a_*^{d-1})\right)\ \delta_q\left(\mathbf{h} \cdot (1, b, \ldots, b^{d-1})\right).\end{aligned}$$

Based on this equality, from (14) and (15) we have

$$\widetilde{M}_{n,d}(\alpha) = \frac{1}{q-1} \sum_{b=1}^{q-1} \sum_{\mathbf{h} \in \mathbf{Z}^d}{}' \frac{\delta_n(\mathbf{h} \cdot \mathbf{z}_2(b))}{\prod_{j=1}^{d} r_\alpha(\gamma_j, h_j)}$$

$$= \frac{1}{q-1} \sum_{\mathbf{h}\in\mathbf{Z}^d}{}' \frac{\delta_p\left(\mathbf{h}\cdot(1,a_*,\ldots,a_*^{d-1})\right)}{\prod_{j=1}^d r_\alpha(\gamma_j,h_j)} \sum_{b=1}^{q-1} \delta_q\left(\mathbf{h}\cdot(1,b,\ldots,b^{d-1})\right)$$
$$=: R_1 + R_2,$$

where in the last step the sum over $\mathbf{h}$ is split into R_1 and R_2, with R_1 being the sum over all those $\mathbf{h}\in\mathbf{Z}^d$ such that at least one component of $\mathbf{h}$ is not a multiple of q, and R_2 being the sum over all those nonzero $\mathbf{h}\in\mathbf{Z}^d$ such that each component is a multiple of q.

Consider the first sum R_1. Since q is a prime number, according to Lemma 3 we have

$$\sum_{b=1}^{q-1} \delta_q\left(\mathbf{h}\cdot(1,b,\ldots,b^{d-1})\right) \le A_q(\mathbf{h}) \le d-1.$$

Thus

$$R_1 \le \frac{d-1}{q-1} \sum_{\mathbf{h}\in\mathbf{Z}^d}{}' \frac{\delta_p\left(\mathbf{h}\cdot(1,a_*,\ldots,a_*^{d-1})\right)}{\prod_{j=1}^d r_\alpha(\gamma_j,h_j)} = \frac{d-1}{q-1}\, e_{p,d}^2(\mathbf{z}_1(a_*)),$$

where $\mathbf{z}_1(a_*) = (1,a_*,\ldots,a_*^{d-1}) \pmod p$ is the optimal generator found in step (i) of Algorithm 2.

For the second sum R_2, each component of the corresponding $\mathbf{h}$ is a multiple of q; i.e., $\mathbf{h} = (qm_1,\ldots,qm_d)$ for some nonzero integer vector $\mathbf{m} = (m_1,\ldots,m_d)\in\mathbf{Z}^d$. So for arbitrary $b = 1,\ldots,q-1$, we have

$$\delta_q\left(\mathbf{h}\cdot(1,b,\ldots,b^{d-1})\right) = 1.$$

Thus

$$R_2 = \sum_{\mathbf{m}\in\mathbf{Z}^d}{}' \frac{\delta_p\left(\mathbf{m}\cdot(1,a_*,\ldots,a_*^{d-1})\right)}{\prod_{j=1}^d r_\alpha(\gamma_j,qm_j)} \le \frac{1}{q^\alpha} \sum_{\mathbf{m}\in\mathbf{Z}^d}{}' \frac{\delta_p\left(\mathbf{m}\cdot(1,a_*,\ldots,a_*^{d-1})\right)}{\prod_{j=1}^d r_\alpha(\gamma_j,m_j)} \le \frac{1}{q-1}\, e_{p,d}^2(\mathbf{z}_1(a_*)).$$

Finally, we have

$$\widetilde{M}_{n,d}(\alpha) = R_1 + R_2 \le \frac{d}{q-1}\, e_{p,d}^2(\mathbf{z}_1(a_*)), \tag{16}$$

and therefore,

$$e_{n,d}^2(\mathbf{z}_2(b_*)) \le \frac{d}{q-1}\, e_{p,d}^2(\mathbf{z}_1(a_*)). \tag{17}$$

Since $\mathbf{z}_1(a_*)$ is the optimal Korobov generator found in step (i) of Algorithm 2 with the number of points p (which is a prime), from Theorem 1 it follows that

$$e_{p,d}^2(\mathbf{z}_1(a_*)) \le \frac{d}{(p-1)} \exp\left(2\zeta(\alpha)\sum_{j=1}^d \gamma_j\right).$$

Combining this inequality with (16) and (17), the results follow immediately. ∎

From (17) one can see that there is a guarantee that $e_{pq,d}(\mathbf{z}_2(b_*))$ is less than $e_{p,d}(\mathbf{z}_1(a_*))$ only if $q > d+1$.

Remark 1. From the nature of the proof, the relations (16) and (17) can be generalized as follows: Let p be an integer and q be a prime. If $\gcd(p,q) = 1$, then for an arbitrary vector $\mathbf{z} = (z_1, \ldots, z_d)$ with $z_j \in \{1, \ldots, p-1\}$, there exists $\overline{b} \in \{1, \ldots, q-1\}$ such that

$$e^2_{pq,d}\left(\left(p\ \left(1, \overline{b}, \ldots, \overline{b}^{d-1}\right) + q\mathbf{z}\right)\ (\mathrm{mod}\ pq)\right) \le \frac{d}{q-1}\ e^2_{p,d}(\mathbf{z}).$$

This relation will be useful in the next subsection.

Based on Theorem 5, similar error bounds to those in Theorem 4 can be proved for the optimal Korobov lattice rules constructed by Algorithm 2.

THEOREM 6. *Let $n = pq$, with p and q being two distinct primes, and let $\mathbf{z}_2(b_*)$ be the optimal Korobov generator found by Algorithm* 2.

(i) *For arbitrary $\tau \in [1, \alpha)$, we have*

$$e_{n,d}(\mathbf{z}_2(b_*)) \le C_d(\alpha,\tau) d^{1+\tau/2}(p-1)^{-\tau/2}(q-1)^{-1/2}, \tag{18}$$

where $C_d(\alpha,\tau)$ is defined in (10). *The quantity $C_d(\alpha,\tau)$ is uniformly bounded in d under the same conditions as in Theorem* 4(ii).

(ii) *Let B^* be the same as in Theorem* 4(iii). *Then the optimal Korobov generator $\mathbf{z}_2(b_*)$ satisfies*

$$e_{n,d}(\mathbf{z}_2(b_*)) \le C_\delta\ d^{1+\zeta(\alpha)(B^*+\delta)}(p-1)^{-1/2}(q-1)^{-1/2} \quad \forall \delta > 0,$$

where C_δ is a constant independent of d and n.

Proof. From Theorem 4(i), for any $\tau \in [1, \alpha)$ we have (with n replaced by p in (9))

$$e_{p,d}(\mathbf{z}_1(a_*)) \le C_d(\alpha,\tau) d^{\tau/2}(p-1)^{-\tau/2}, \tag{19}$$

where the constant $C_d(\alpha,\tau)$ is given in (10). Combining this inequality with (17), the error bound (18) follows immediately. The conclusion (ii) can be proved by combining Theorem 4(iii) with the relation (17). ∎

Remark 2. From Theorem 6, for fast decaying weights the improved convergence (or even nearly the optimal convergence) for the case of $n = pq$ can only be achieved by the first factor p. For the other factor q, the low convergence order $O(q^{-1/2})$ remains the same as in Theorem 5. The same happens in the CBC construction of lattice rules [3].

4.2. The case of multiple distinct primes. Now we consider the case of $n = \prod_{j=1}^t p_j$, with $p_1, p_2, \ldots, p_t$ being distinct primes ($t \ge 2$). Algorithm 2 in the previous subsection can be generalized as follows.

ALGORITHM 3 (for $n = \prod_{j=1}^t p_j$).

(i) *Find the optimal $a_1 \in \{1, \ldots, p_1 - 1\}$ using Algorithm* 1 *in section* 3, *but with n replaced by p_1. We obtain a vector $\mathbf{z}_1(a_1) := (1, a_1, \ldots, a_1^{d-1})\ (\mathrm{mod}\, p_1)$.*

(ii) *For fixed $\ell = 2, \ldots, t$, let $p := \prod_{j=1}^{\ell-1} p_j$ and $q := p_\ell$. Let the generating vector $\mathbf{z}_\ell(a)$ be of the form*

$$\mathbf{z}_\ell(a) := \left(p\ (1, a, \ldots, a^{d-1}) + q\, \mathbf{z}_{\ell-1}(a_{\ell-1})\right)\ (\mathrm{mod}\, pq),$$

where $\mathbf{z}_{\ell-1}(a_{\ell-1})$ *is the vector found in the previous step. Find the optimal* $a_\ell \in \{1,\dots,p_\ell-1\}$*, such that the following squared worst-case error is minimized with respect to* a*:*

$$e^2_{pq,d}(\mathbf{z}_\ell(a)) = -1 + \frac{1}{pq}\sum_{k=0}^{pq-1}\prod_{j=1}^{d}\left(1+\gamma_j\sum_{h\in\mathbf{Z}}{}'\,\frac{\exp\left(2\pi i hk(pa^{j-1}+qa_{\ell-1}^{j-1})/(pq)\right)}{|h|^\alpha}\right).$$

At the last step, we obtain the optimal generator $\mathbf{z}_t(a_t)$. We shall still call the resulting lattice rule an optimal Korobov lattice rule (though it is not of the original Korobov form any more).

THEOREM 7. *Suppose* $n=\prod_{j=1}^t p_j$*, with* $p_1,\dots,p_t$ *being distinct primes* ($t\ge 2$)*. Let* $\mathbf{z}_t(a_t)$ *be the final optimal generator found by Algorithm* 3*. Then*

$$e^2_{n,d}(\mathbf{z}_t(a_t)) \le \frac{d^t}{\prod_{j=1}^t(p_j-1)}\exp\left(2\zeta(\alpha)\sum_{j=1}^{d}\gamma_j\right).$$

Proof. Let $a_1,a_2,\dots,a_t$ be the integers found in successive steps of Algorithm 3 and let $\mathbf{z}_1(a_1),\mathbf{z}_2(a_2),\dots,\mathbf{z}_t(a_t)$ be the corresponding vectors. For any fixed ℓ with $2\le\ell\le t$, let $p=\prod_{j=1}^{\ell-1}p_j$ and $q=p_\ell$. Since $p_1,\dots,p_\ell$ are distinct primes, we have $\gcd(p,q)=1$ and q is prime. Similar arguments as those presented in Theorem 5, which led to relation (17), are applicable here (see Remark 1 in the previous subsection). Thus we have a recursive relation

$$e^2_{pq,d}(\mathbf{z}_\ell(a_\ell)) \le \frac{d}{q-1}\,e^2_{p,d}(\mathbf{z}_{\ell-1}(a_{\ell-1})),\quad \ell=2,\dots,t.$$

Using this recursive relation, we have

$$\begin{aligned} e^2_{p_1\cdots p_t,d}(\mathbf{z}_t(a_t)) &\le \frac{d^{t-1}}{\prod_{j=2}^t(p_j-1)}\,e^2_{p_1,d}(\mathbf{z}_1(a_1)) \\ &\le \frac{d^t}{\prod_{j=1}^t(p_j-1)}\exp\left(2\zeta(\alpha)\sum_{j=1}^{d}\gamma_j\right). \end{aligned}$$

In the second step we used Theorem 1. ∎

Based on this theorem, similar error bounds to those in Theorems 4 and 6 can be established for the optimal Korobov lattice rules constructed by Algorithm 3. For example, for arbitrary $\tau\in[1,\alpha)$, we have

$$e_{n,d}(\mathbf{z}_t(a_t)) \le C_d(\alpha,\tau)\,d^{t-1+\tau/2}(p_1-1)^{-\tau/2}\prod_{j=2}^{t}(p_j-1)^{-1/2}, \tag{20}$$

where $C_d(\alpha,\tau)$ is given in (10), which is uniformly bounded in d under the same conditions as Theorem 4(ii).

The advantage of Algorithm 3 over Algorithm 1 is that the construction cost of Algorithm 3 is much cheaper (for approximately the same n). Indeed, in dimension d the total number of operations needed to find the final optimal generator $\mathbf{z}_t(a_t)$ for even α by Algorithm 3 is

$$O\left(d(p_1^2+p_1p_2^2+p_1\cdots p_{t-1}p_t^2)\right).$$

By choosing

$$p_{t-1} \approx p_t^2,\ p_{t-2} \approx p_{t-1}^2 \approx p_t^4, \ldots, p_1 \approx p_2^2 \approx p_t^{2^{t-1}},$$

the number of operations needed to find the final generator is $O(dn^{1+(2^t-1)^{-1}})$. For example, for $t = 2$ (corresponding to Algorithm 2) or $t = 3$, the number of operations needed is $O(dn^{4/3})$ or $O(dn^{8/7})$, respectively. Note that even one evaluation of the worst-case error $e_{n,d}(\mathbf{z})$ requires $O(dn)$ operations. Thus the number of operations needed in Algorithms 2 or 3 is much smaller than that in Algorithm 1 (for prime n), for which the number of operations is $O(dn^2)$.

In general, the larger the value of t, the cheaper the construction. However, there is a trade-off—for larger values of t the convergence order is worse, and the dependence on the dimension is stronger, which implies that the quality of the resulting lattice rules might be worse. For example, by choosing $p_1, \ldots, p_t$ as above, if $\alpha = 2$ and $\tau_0 \geq 2$, then from (20) the convergence order is approximately $O\left(n^{-(2^t+2^{t-1}-1)/(2^{t+1}-2)}\right)$, with the implied constant depending at worst polynomially on d. For a large value t, it is approximately $O(n^{-3/4})$. For $t = 2$ or $t = 3$, the convergence order is approximately $O(n^{-5/6})$ or $O(n^{-11/14})$, respectively. In general, if the required number of points n is relatively small, one may use Algorithm 1; if n is very large, then one may use Algorithm 2 or 3. It is not recommended to use Algorithm 2 or 3 if n is relatively small and d is large.

Remark 3. The theoretical bounds on the worst-case errors for Korobov lattice rules in sections 3 and 4 grow polynomially with the dimension d even for rapidly decaying weights, so the theoretical properties of such lattice rules are worse than those of the CBC lattice rules, for which the corresponding errors can be bounded uniformly in d for fast decaying weights [14]. However, the bounds are established by "averaging arguments" and are in general quite conservative. The numerical comparison of the worst-case errors in section 6 suggests that the true dependence of the worst-case error of the optimal Korobov lattice rules on the dimension is almost the same as that of the CBC lattice rules.

5. Weighted Sobolev spaces. In this section, we briefly study multivariate integration in weighted Sobolev spaces of nonperiodic functions using shifted Korobov lattice rules. We use a similar technique to that in [23]. However, we consider a more general class of weighted Sobolev spaces $H(K_{d,\gamma}^{\mathrm{sob}})$ with the following reproducing kernels (see [7, 24]):

$$K_{d,\gamma}^{\mathrm{sob}}(\mathbf{x}, \mathbf{y}) = \prod_{j=1}^{d} [1 + \gamma_j \eta(x_j, y_j)],$$

where

$$\eta(x, y) = \frac{1}{2} B_2(\{x - y\}) + (x - 1/2)(y - 1/2) + \mu(x) + \mu(y) + M.$$

Here $B_2(x)$ is the Bernoulli polynomial of degree 2, $\mu(x)$ is a function with bounded derivative in $[0, 1)$ such that $\int_0^1 \mu(x)dx = 0$, and the constant M is equal to $\int_0^1 (\mu'(x))^2 dx$. Several common choices of $\mu(x)$ are

$$(A)\ \ \mu(x) = \frac{1}{6} - \frac{x^2}{2};\ \ (B)\ \ \mu(x) = -\frac{1}{2} B_2(\{x - 1/2\});\ \ (C)\ \ \mu(x) = 0.$$

For example, for the first choice of $\mu(x)$ the kernel is

$$K_{d,\gamma}^{\rm sob}(\mathbf{x},\mathbf{y}) = \prod_{j=1}^{d}[1+\gamma_j \min(1-x_j, 1-y_j)].$$

The tractability of multivariate integration in the corresponding Hilbert space is studied in a nonconstructive way in [22], and in constructive ways in [10, 20, 25]. For the third choice of $\mu(x)$ the reproducing kernel is

$$K_{d,\gamma}^{\rm sob}(\mathbf{x},\mathbf{y}) = \prod_{j=1}^{d}\left[1+\gamma_j\left(\frac{1}{2}B_2(\{x-y\})+(x-1/2)(y-1/2)\right)\right],$$

which has been studied in [4, 24]. In this third case there is a good property for the functions in the corresponding Hilbert space: the Hilbert space decomposition of a function coincides with the ANOVA (analysis of variance) decomposition [4].

The shift-invariant kernel $K_{d,\gamma}^{\rm shift}(\mathbf{x},\mathbf{y})$ associated with the kernel $K_{d,\gamma}^{\rm sob}(\mathbf{x},\mathbf{y})$ is

$$\begin{aligned} K_{d,\gamma}^{\rm shift}(\mathbf{x},\mathbf{y}) &:= \int_{[0,1)^d} K_{d,\gamma}^{\rm sob}(\{\mathbf{x}+\Delta\},\{\mathbf{y}+\Delta\})d\Delta \\ &= \prod_{j=1}^{d}[1+\gamma_j(B_2(\{x-y\})+M)] \\ &= \prod_{j=1}^{d}(1+M\gamma_j)\prod_{j=1}^{d}[1+2\pi^2\widehat{\gamma}_j B_2(\{x-y\})] \end{aligned}$$

with

$$\widehat{\gamma}_j = \frac{\gamma_j}{2\pi^2(1+M\gamma_j)}. \tag{21}$$

Different choices of the function $\mu(x)$ correspond to similar shift-invariant kernels; only the constant M and the weights $\widehat{\gamma}_j$ are different. Note that the Bernoulli polynomial of degree 2 can be expressed as (see (5))

$$B_2(x) = \frac{1}{2\pi^2}\sum_{h\in\mathbf{Z}}{}' \frac{e^{2\pi i h x}}{h^2}, \quad x\in[0,1).$$

Therefore, apart from the factor $\prod_{j=1}^{d}(1+M\gamma_j)$, the kernel $K_{d,\gamma}^{\rm shift}(\mathbf{x},\mathbf{y})$ is just the Korobov reproducing kernel $K_{d,2,\widehat{\gamma}}(\mathbf{x},\mathbf{y})$ in (3) with $\alpha=2$ and the weight sequence $\widehat{\gamma} := \{\widehat{\gamma}_j\}$, i.e.,

$$K_{d,\gamma}^{\rm shift}(\mathbf{x},\mathbf{y}) = \prod_{j=1}^{d}(1+M\gamma_j)\, K_{d,2,\widehat{\gamma}}(\mathbf{x},\mathbf{y}).$$

It is shown in [8, 11] that for a point set $\mathbf{P}_n = \{\mathbf{x}_0,\mathbf{x}_1,\dots,\mathbf{x}_{n-1}\}\subset[0,1)^d$,

$$\begin{aligned} \int_{[0,1)^d} e^2(\mathbf{P}_n+\Delta; H(K_{d,\gamma}^{\rm sob}))d\Delta &= e^2(\mathbf{P}_n; H(K_{d,\gamma}^{\rm shift})) \\ &= \prod_{j=1}^{d}(1+M\gamma_j)\, e^2(\mathbf{P}_n; H(K_{d,2,\widehat{\gamma}})), \end{aligned}$$

where $\mathbf{P}_n + \Delta$ denotes the shifted point set $\{\{\mathbf{x}_0 + \Delta\}, \ldots, \{\mathbf{x}_{n-1} + \Delta\}\}$. Clearly, the square of the initial integration error in the space $H(K^{\rm sob}_{d,\gamma})$ is

$$e^2(0; H(K^{\rm sob}_{d,\gamma})) := \prod_{j=1}^{d}(1 + M\gamma_j).$$

Thus there exists a shift $\Delta \in [0,1)^d$ such that

$$e(\mathbf{P}_n + \Delta; H(K^{\rm sob}_{d,\gamma})) \le e(0; H(K^{\rm sob}_{d,\gamma}))\ \ e(\mathbf{P}_n; H(K_{d,2,\widehat{\gamma}})).$$

Therefore, in order to reduce the worst-case error in the weighted Sobolev space $H(K^{\rm sob}_{d,\gamma})$ by a factor of ε from its initial error, it is sufficient to reduce the worst-case error in the weighted Korobov space $H(K_{d,2,\widehat{\gamma}})$ to ε, where the weights $\widehat{\gamma}_j$ are given by (21). The error bounds in the space $H(K_{d,2,\widehat{\gamma}})$ have been studied in the previous sections (corresponding to $\alpha = 2$ and the weights $\widehat{\gamma}_j$). Thus similar error bounds for randomly shifted optimal Korobov lattice rules can be established in weighted Sobolev spaces. We only state the case for prime n and omit the proof.

THEOREM 8. *Suppose n is a prime. Let $\mathbf{P}_n$ be the lattice point set corresponding to the optimal Korobov lattice rule found by Algorithm 1 (with γ_j replaced by $\hat{\gamma}_j$).*

(i) *There exists a shift $\Delta \in [0,1)^d$ such that*

$$e^2(\mathbf{P}_n + \Delta; H(K^{\rm sob}_{d,\gamma})) \le \frac{d}{n-1}\ \exp\left(\frac{1}{6}\sum_{j=1}^{d}\gamma_j\right) e(0; H(K^{\rm sob}_{d,\gamma})).$$

(ii) *There exists a shift $\Delta \in [0,1)^d$ such that for arbitrary $\tau \in [1,2)$*

$$e(\mathbf{P}_n + \Delta; H(K^{\rm sob}_{d,\gamma})) \le C'_d(\tau)\left(\frac{d}{n-1}\right)^{\tau/2} e(0; H(K^{\rm sob}_{d,\gamma})),$$

where

$$C'_d(\tau) = \exp\left(\frac{\tau\zeta(2/\tau)}{\sqrt{2}\,\pi}\sum_{j=1}^{d}\gamma_j^{1/\tau}\right).$$

(iii) *Suppose that the weights γ_j satisfy $\sum_{j=1}^{\infty}\gamma_j < \infty$. Let τ_0 be the same as in Theorem 4. Then for any $\tau \in [1, \min(\tau_0, 2))$ (or if $\tau_0 = 1$, then put $\tau = 1$), we have*

$$C'_d(\tau) \le C'_\infty(\tau) := \lim_{d\to\infty} C'_d(\tau) < \infty;$$

i.e., $C'_d(\tau)$ is uniformly bounded in d.

(iv) *Let B^* be the same as in Theorem 4. Then there exists a shift $\Delta \in [0,1)^d$ such that the shifted optimal Korobov lattice point set $\mathbf{P}_n + \Delta$ satisfies*

$$e(\mathbf{P}_n + \Delta; H(K^{\rm sob}_{d,\gamma})) \le \frac{C_\delta}{(n-1)^{1/2}}\ d^{\frac{1}{2}+\frac{1}{12}(B^*+2\pi^2\delta)} e(0; H(K^{\rm sob}_{d,\gamma}))$$

for any $\delta > 0$, where C_δ is a constant independent of d and n.

Note that it is not necessary to determine the optimal "shift" for practical use. Instead, we may choose it randomly [2, 21], since for fixed $\mathbf{P}_n$, the family of QMC algorithms with points $\{\mathbf{P}_n + \Delta\}$ (where Δ is uniformly distributed on $\in [0,1]^d$) is an unbiased family. This approach allows a probabilistic error estimation.

6. Numerical comparisons of the worst-case errors. We perform a comparison of the worst-case errors in weighted Korobov spaces. Two kinds of lattice rules are considered: the optimal Korobov lattice rules and the CBC lattice rules. The root mean square worst-case errors over random points and the root mean square worst-case errors over all lattice rules are also included as benchmarks. We are mainly interested in the effect of the dimension and the convergence order. We focus our attention only on the case where $\alpha = 2$ and n is prime. The mean square worst-case error averaged over random points is defined and calculated as

$$(E^{\mathrm{av}}_{n,d})^2 := \int_{[0,1)^{nd}} e^2(\{\mathbf{x}_k\}; H(K_{d,\alpha,\gamma}))d\mathbf{x}_0 \cdots d\mathbf{x}_{n-1} = \frac{1}{n}\left(\prod_{j=1}^{d}(1+2\zeta(\alpha)\gamma_j) - 1\right).$$

For prime n, the mean square worst-case error over all rank-1 lattice rules is (see [23])

$$\begin{aligned} M^{(\mathrm{LR})}_{n,d}(\alpha) &:= \frac{1}{(n-1)^d}\sum_{z_1=1}^{n-1}\cdots\sum_{z_d=1}^{n-1} e^2_{n,d}((z_1,\ldots,z_d)) \\ &= -1 + \frac{1}{n}\prod_{j=1}^{d}(1+2\zeta(\alpha)\gamma_j) + (1-n^{-1})\prod_{j=1}^{d}\left(1 - \frac{2\gamma_j\zeta(\alpha)(1-n^{1-\alpha})}{n-1}\right). \end{aligned}$$

We consider weights of the form

$$\gamma_j = a\theta^j, \quad j = 1,\ldots,d,$$

where a and θ are parameters. We intend to use a to mainly characterize the relative importance of different orders of "interaction" and use θ to reflect the relative importance of successive variables. We consider the following three choices:

$$(A)\ a = 1, \quad \theta = 1; \qquad (B)\ a = 0.098, \quad \theta = 0.98; \qquad (C)\ a = 0.25, \quad \theta = 0.75.$$

The first case corresponds to the classical unweighted case. The latter two choices are found to be "suitable" for some problems in finance (see [27]).

Computational results are given in Tables 1–3 in the appendix. The rate of convergence of the worst-case error (i.e., the value r in an expression of the form $O(n^{-r})$ for convergence, estimated from linear regression on the empirical data) is also given. Note that the root mean square worst-case error over random points converges as $O(n^{-1/2})$. The root mean square worst-case error over all rank-1 lattice rules has convergence $O(n^{-1})$ for $d=1$ and $O(n^{-1/2})$ for $d>1$. These serve as benchmarks.

Table 1, for the unweighted case (A), draws attention to the importance of incorporating weights into our function spaces: in this unweighted case the worst-case errors for both the Korobov and CBC lattice rules grow exponentially in d. This is, of course, expected from the known intractability of the problem in this case (see [23]). For large d, neither method is significantly better than the two root mean square values (which are essentially the same). For all values of d, there is little difference between the two constructions. Tables 2 and 3 show a dramatic reduction in the worst-case errors produced by decaying weights. This is especially true for the latter case, where the rapid decay of the weights leads to essentially no change in the worst-case errors beyond $d = 32$, and a high convergence is achieved by the Korobov and CBC constructions even in high dimensions. The worst-case errors of the two kinds of lattice rules are here smaller or much smaller than the mean values.

7. Concluding remarks. The striking conclusion from all the numerical results in section 6 is that the worst-case errors for the Korobov construction are always comparable with those for CBC construction. In particular, the worst-case errors for the Korobov construction do not seem to show the faster growth with d predicted by the theoretical error bounds in Theorem 4. Thus there remains a gap between theory and observation. We mentioned that the theoretical bounds are established by averaging arguments and are in general conservative.

The Korobov construction has the advantage of speed and simplicity over the CBC construction, in that Algorithm 1 requires a time of order only $O(dn^2)$ if n is prime and only a single n-point lattice rule is needed. (And the time is further reduced if n is a product of primes as in Algorithms 2 and 3). On the other hand, the CBC algorithm is extensible in dimension, whereas the Korobov construction is not. The existence of good rank-1 lattice rules that are extensible both in d and n is proved in [9], but no construction is given.

The weighted Korobov spaces are characterized by the weights, which are supposed for our theoretical analysis to be given. The weights have a strong influence on the properties (e.g., effective dimensions) of the functions in the weighted spaces (see [26, 27]). An important problem for practical application is: *What weights should be used?* Some attempts have been made in [4, 27] to choose the weights to reflect the characteristics of the functions to be integrated. In [27] it is shown that the performance of lattice rules strongly depends on the weights, and that blind use of lattice rules based on classical weights may lead to estimates that are worse than Monte Carlo estimates, especially in high dimensions (it is shown in [27] that the classical weights $\gamma_j = 1$ are "too large" for some typical high-dimensional problems in finance). Thus the problem of choosing the weights must be considered hand-in-hand with the problem addressed here, that of choosing the algorithm.

Appendix. Comparisons of the worst-case errors.

TABLE 1

The comparison of the worst-case errors in weighted Korobov spaces and the apparent convergence order in dimensions up to 64 *for the case* $(A) : \gamma_j = 1$. *(The abbreviations "Mean" and "Mean LR" denote root mean square worst-case errors over all random points and root mean square worst-case errors over all rank-*1 *lattice rules, respectively.)*

n	Method	$d=1$	$d=2$	$d=4$	$d=8$	$d=16$	$d=32$	$d=64$
251	Mean	1.14e-1	2.63e-1	1.16e00	2.14e+1	7.24e+3	8.30e+8	1.09e+19
	Mean LR	7.23e-3	2.08e-1	1.14e00	2.14e+1	7.24e+3	8.30e+8	1.09e+19
	Korobov	7.23e-3	5.10e-2	7.93e-1	2.08e+1	7.24e+3	8.30e+8	1.09e+19
	CBC	7.23e-3	5.10e-2	7.75e-1	2.04e+1	7.09e+3	8.21e+8	1.09e+19
1009	Mean	5.71e-2	1.31e-1	5.79e-1	1.07e+1	3.61e+3	4.14e+8	5.45e+18
	Mean LR	1.80e-3	1.04e-1	5.67e-1	1.07e+1	3.61e+3	4.14e+8	5.45e+18
	Korobov	1.80e-3	1.41e-2	2.94e-1	9.98e00	3.61e+3	4.14e+8	5.45e+18
	CBC	1.80e-3	1.41e-2	3.04e-1	1.01e+1	3.49e+3	4.03e+8	5.35e+18
4001	Mean	2.87e-2	6.60e-2	2.91e-1	5.35e00	1.81e+3	2.08e+8	2.74e+18
	Mean LR	4.53e-4	5.20e-2	2.85e-1	5.35e00	1.81e+3	2.08e+8	2.74e+18
	Korobov	4.53e-4	3.80e-3	1.06e-1	4.75e00	1.81e+3	2.08e+8	2.74e+18
	CBC	4.53e-4	3.80e-3	1.07e-1	4.89e00	1.73e+3	2.03e+8	2.71e+18
r	Korobov	1.00	0.94	0.73	0.53	0.50	0.50	.50
	CBC	1.00	0.94	0.71	0.52	0.51	0.50	.50

TABLE 2
The same as Table 1 *but for the case* $(B) : \gamma_j = a\theta^j$ *with* $a = 0.098, \theta = 0.98$.

n	Method	$d=1$	$d=2$	$d=4$	$d=8$	$d=16$	$d=32$	$d=64$
251	Mean	3.55e-2	5.37e-2	8.73e-2	1.66e-1	4.29e-1	1.83e00	1.17e+1
	Mean LR	2.24e-3	2.00e-2	5.26e-2	1.35e-1	4.09e-1	1.82e00	1.17e+1
	Korobov	2.24e-3	5.70e-3	2.21e-2	9.15e-2	3.63e-1	1.80e00	1.17e+1
	CBC	2.24e-3	5.70e-3	2.10e-2	8.91e-2	3.64e-1	1.80e00	1.17e+1
1009	Mean	1.77e-2	2.68e-2	4.36e-2	8.26e-2	2.14e-1	9.13e-1	5.84e00
	Mean LR	5.57e-4	9.88e-3	2.61e-2	6.71e-2	2.04e-1	9.09e-1	5.84e00
	Korobov	5.57e-4	1.53e-3	6.89e-3	3.60e-2	1.71e-1	8.88e-1	5.83e00
	CBC	5.57e-4	1.53e-3	6.93e-3	3.50e-2	1.68e-1	8.87e-1	5.83e00
4001	Mean	8.89e-3	1.34e-2	2.19e-2	4.15e-2	1.08e-1	4.58e-1	2.93e00
	Mean LR	1.40e-4	4.95e-3	1.31e-2	3.67e-2	1.02e-1	4.56e-1	2.93e00
	Korobov	1.40e-4	4.08e-4	2.38e-3	1.41e-2	7.81e-2	4.39e-1	2.92e00
	CBC	1.40e-4	4.08e-4	2.21e-3	1.38e-2	7.66e-2	4.39e-1	2.92e00
r	Korobov	1.00	0.95	0.80	0.68	0.55	0.51	0.50
	CBC	1.00	0.95	0.81	0.67	0.56	0.51	0.50

TABLE 3
The same as Table 1 *but for the case* $(C) : \gamma_j = a\theta^j$ *with* $a = 0.25, \theta = 0.75$.

n	Method	$d=1$	$d=2$	$d=4$	$d=8$	$d=16$	$d=32$	$d=64$
251	Mean	4.96e-2	7.37e-2	1.10e-1	1.49e-1	1.70e-1	1.7194e-1	1.7196e-1
	Mean LR	3.13e-3	3.40e-2	7.30e-2	1.16e-1	1.38e-1	1.4076e-1	1.4079e-1
	Korobov	3.13e-3	9.10e-3	3.25e-2	7.36e-2	9.61e-2	9.9042e-2	9.9080e-2
	CBC	3.13e-3	9.10e-3	3.02e-2	6.86e-2	9.08e-2	9.3658e-2	9.3687e-2
1009	Mean	2.47e-2	3.68e-2	5.47e-2	7.45e-2	8.46e-2	8.5755e-2	8.5767e-2
	Mean LR	7.78e-4	1.69e-2	3.63e-2	5.80e-2	6.88e-2	7.0098e-2	7.0111e-2
	Korobov	7.78e-4	2.47e-3	1.03e-2	2.77e-2	4.08e-2	4.2040e-2	4.2068e-2
	CBC	7.78e-4	2.47e-3	1.02e-2	2.62e-2	3.63e-2	3.7654e-2	3.7668e-2
4001	Mean	1.24e-2	1.85e-2	2.74e-2	3.74e-2	4.25e-2	4.3065e-2	4.3071e-2
	Mean LR	1.96e-4	8.45e-3	1.82e-2	2.91e-2	3.45e-2	3.5189e-2	3.5195e-2
	Korobov	1.96e-4	6.62e-4	3.56e-3	1.05e-2	1.58e-2	1.6624e-2	1.6634e-2
	CBC	1.96e-4	6.62e-4	3.23e-3	9.73e-3	1.45e-2	1.5063e-2	1.5070e-2
r	Korobov	1.00	0.95	0.80	0.70	0.65	0.64	0.64
	CBC	1.00	0.95	0.81	0.71	0.66	0.66	0.66

Acknowledgment. The authors would like to thank the referees for their valuable comments.

REFERENCES

[1] R. E. CAFLISCH, W. MOROKOFF, AND A. B. OWEN, *Valuation of mortgage backed securities using Brownian bridges to reduce effective dimension*, J. Comp. Finance, 1 (1997), pp. 27–46.

[2] R. CRANLEY AND T. N. L. PATTERSON, *Randomization of number theoretic methods for multiple integration*, SIAM J. Numer. Anal., 13 (1976), pp. 904–914.

[3] J. DICK AND F. Y. KUO, *Reducing the construction cost of the component-by-component construction of good lattice rules*, Math. Comp., 73 (2004), pp. 1967–1988.

[4] J. DICK, I. H. SLOAN, X. WANG, AND H. WOŹNIAKOWSKI, *Liberating the weights*, J. Complexity, 20 (2004), pp. 593–623.

[5] G. H. HARDY AND E. M. WRIGHT, *An Introduction to Number Theory*, 4th ed., Clarendon Press, Oxford, UK, 1960.

[6] F. J. HICKERNELL, *A comparison of random and quasirandom points for multidimensional quadrature*, in Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Lecture Notes in Statist. 106, H. Niederreiter and P. J.-S. Shiue, eds., Springer-Verlag, New York, 1995, pp. 213–227.

[7] F. J. Hickernell, *A generalized discrepancy and quadrature error bound*, Math. Comp., 67 (1998), pp. 299–322.
[8] F. J. Hickernell, *Lattice rules: How well do they measure up?*, in Random and Quasi-Random Point Sets, Lecture Notes in Statist. 138, P. Hellekalek and G. Larcher, eds., Springer-Verlag, New York, 1998, pp. 109–166.
[9] F. J. Hickernell and H. Niederreiter, *The existence of good extensible rank-1 lattice*, J. Complexity, 19 (2003), pp. 286–300.
[10] F. J. Hickernell and X. Wang, *The error bounds and tractability of quasi-Monte Carlo algorithms in infinite dimension*, Math. Comp., 71 (2002), pp. 1641–1661.
[11] F. J. Hickernell and H. Woźniakowski, *Integration and approximation in arbitrary dimension*, Adv. Comput. Math., 12 (2000), pp. 25–58.
[12] N. M. Korobov, *Approximate evaluation of repeated integrals*, Dokl. Akad. Nauk SSSR, 124 (1959), pp. 1207–1210 (in Russian).
[13] N. M. Korobov, *Number-Theoretic Methods in Approximate Analysis*, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1963 (in Russian).
[14] F. Y. Kuo, *Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces*, J. Complexity, 19 (2003), pp. 301–320.
[15] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 63, SIAM, Philadelphia, 1992.
[16] A. B. Owen, *The dimension distribution and quadrature test functions*, Statist. Sinica, 13 (2003), pp. 1–18.
[17] S. H. Paskov and J. F. Traub, *Faster valuation of financial derivatives*, J. Portfolio Management, 22 (1995), pp. 113–120.
[18] I. F. Sarygin, *A lower estimate for the error of quadrature formulas for certain classes of functions*, Zh. Vychisl. Mat. Mat. Fiz., 3 (1963), pp. 370–376 (in Russian).
[19] I. H. Sloan and S. Joe, *Lattice Methods for Multiple Integration*, Oxford University Press, Oxford, UK, 1994.
[20] I. H. Sloan, F. Y. Kuo, and S. Joe, *On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces*, Math. Comp., 71 (2002), pp. 1609–1640.
[21] I. H. Sloan, F. Y. Kuo, and S. Joe, *Constructing randomly shifted lattice rules in weighted Sobolev spaces*, SIAM J. Numer. Anal., 40 (2002), pp. 1650–1665.
[22] I. H. Sloan and H. Woźniakowski, *When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?*, J. Complexity, 14 (1998), pp. 1–33.
[23] I. H. Sloan and H. Woźniakowski, *Tractability of multivariate integration for weighted Korobov classes*, J. Complexity, 17 (2001), pp. 697–721.
[24] I. H. Sloan and H. Woźniakowski, *Tractability of integration in non-periodic and periodic weighted Tensor product Hilbert spaces*, J. Complexity, 18 (2002), pp. 479–499.
[25] X. Wang, *Strong tractability of multivariate integration using quasi-Monte Carlo algorithms*, Math. Comp., 72 (2003), pp. 823–838.
[26] X. Wang and K. T. Fang, *Effective dimensions and quasi-Monte Carlo integration*, J. Complexity, 19 (2003), pp. 101–124.
[27] X. Wang and I. H. Sloan, *Efficient weighted lattice rules with applications to finance*, submitted.
[28] X. Wang and I. H. Sloan, *Why are high-dimensional financial problems often of low effective dimension?*, submitted.

SIAM J. NUMER. ANAL.
Vol. 42, No. 4, pp. 1780–1800

A POSTERIORI ERROR ESTIMATES BASED ON THE POLYNOMIAL PRESERVING RECOVERY*

AHMED NAGA† AND ZHIMIN ZHANG‡

Abstract. Superconvergence of order $O(h^{1+\rho})$, for some $\rho > 0$, is established for the gradient recovered with the polynomial preserving recovery (PPR) when the mesh is mildly structured. Consequently, the PPR-recovered gradient can be used in building an asymptotically exact a posteriori error estimator.

Key words. finite element method, superconvergence patch recovery, polynomial preserving recovery, discrete least-squares best fitting, superconvergence, a posteriori error estimator

AMS subject classifications. 65N30, 65N15, 65N12, 65D10, 74S05, 41A10, 41A25

DOI. 10.1137/S0036142903413002

1. Introduction.

Adaptive control based on a posteriori error estimates have become standard in finite element methods since the pioneering work by Babuška and Rheinboldt [2]. The field of the a posteriori error estimators attracted many researchers and has become the focus of intensive investigations. For the literature, the reader is referred to recent books by Ainsworth and Oden [1] and Babuška and Strouboulis [3], a conference proceeding [8], a survey article by Bank [4], and an earlier book by Verfürth [10].

Generally speaking, error estimators can be classified under two categories. The residual type estimators (for example, see [5]) constitute the first category, while recovery based error estimators (for example, see [14]) constitute the second one. In recovery based estimators, the finite element solution (or its gradient) is postprocessed as a first step. For example, Zienkiewicz and Zhu [15] introduced the superconvergence patch recovery (SPR) that is used to recover a gradient from the gradient of the finite element solution. In another strategy, Wiberg and Li [11] and Li and Wiberg [9] used the finite element solution to build another solution. If the recovered quantity better approximates the exact one, then it can be used in building an asymptotically exact a posteriori error estimator (see [1] and [3] for some general discussion and literature).

In this work, we consider a posteriori error estimators that are based on gradient recovery. As is shown in [1], if the recovered gradient superconverges to the exact one, the corresponding a posteriori error estimator is asymptotically exact. A good example of such estimators is the Zienkiewicz–Zhu error estimator based on the SPR-recovered gradient (see [16]). The polynomial preserving recovery (PPR) is a new gradient recovery technique introduced in [13]. The PPR-recovered gradient, as we shall soon see, has superconvergence properties in mildly structured meshes. Consequently, it can be used in constructing an asymptotically exact a posteriori error estimator.

*Received by the editors May 31, 2003; accepted for publication (in revised form) August 14, 2003; published electronically December 27, 2004. This research was partially supported by National Science Foundation grants DMS-0074301 and DMS-0311807.
http://www.siam.org/journals/sinum/42-4/41300.html
†Department of Mathematics, Wayne State University, Detroit, MI 48202 (anaga@math.wayne.edu, zzhang@math.wayne.edu).

1.1. Model problem. To fix the ideas, consider the boundary value problem

$$\begin{cases} -\nabla(\mathcal{D}\nabla u + \boldsymbol{b}u) + cu = f & \text{in } \Omega, \\ \boldsymbol{n}\cdot(\mathcal{D}\nabla u + \boldsymbol{b}u) = g & \text{on } \Gamma_N, \\ u = 0 & \text{on } \Gamma_D, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain with Lipschitz boundary $\partial\Omega = \overline{\Gamma_N} \cup \overline{\Gamma_D}$, the boundary segments Γ_N and Γ_D are disjoint, $\boldsymbol{n}$ is the unit outward normal vector to $\partial\Omega$, and $\mathcal{D}$ is a 2×2 symmetric positive definite matrix. If $\Gamma_N = \partial\Omega, \boldsymbol{b} = 0,$ and $c = 0$, the compatibility condition $\int_\Omega f + \int_{\partial\Omega} g = 0$ must be satisfied and the condition $\int_\Omega u = 0$ is used to ensure the uniqueness. For simplicity, Ω is assumed to be a polygonal domain.

As usual, $W_p^m(\Omega)$ and $H^m(\Omega)$ are the classical Sobolev spaces equipped with the norms $\| \quad \|_{m,p,\Omega}$, and $\| \quad \|_{m,\Omega}$, respectively.

The variational form of this problem is to find $u \in V$ such that

$$\mathcal{B}(u,v) = L(v) \text{ for all } v \in V, \tag{1.2}$$

where

$$V = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\},$$

$$\mathcal{B}(u,v) = \int_\Omega [(\mathcal{D}\nabla u + \boldsymbol{b}u)\nabla v + cuv]dxdy,$$

and

$$L(v) = \int_\Omega fvdxdy + \int_{\Gamma_N} gvds.$$

If Γ_D is empty, we take $V = H^1(\Omega)$. We assume that the bilinear operator $\mathcal{B}$ is continuous and V-elliptic, and the linear operator L is bounded. (Of course, this requires the problem data to satisfy some conditions.) Under these assumptions, the variational problem in (1.2) has a unique weak solution $u \in V$.

Let $\mathcal{T}_h$ be a triangular partition of Ω, and let $\mathcal{N}_h$ denote the set of the mesh nodes. The area of a mesh triangle $T \in \mathcal{T}_h$ will be denoted by $|T|$. A mesh node z is called an *internal* (*boundary*) mesh node if $z \in \Omega$ ($z \in \partial\Omega$). Consider the C^0 linear finite element space $S_h \subset H^1(\Omega)$ associated with $\mathcal{T}_h$ and defined by

$$S_h = \{v \in H^1(\Omega) : v \in P_1(T) \text{ for every triangle } T, \in \mathcal{T}_h\},$$

where $P_r(\mathcal{A})$ denotes the set of all polynomials defined on $\mathcal{A} \subseteq \mathbb{R}^2$ of total degree $\leq r$. The basis functions for S_h are the standard Lagrange basis functions and I_h will denote the Lagrange interpolation operator associated with S_h. The finite element solution of (1.2) is $u_h \in S_h \cap V$ such that

$$\mathcal{B}(u_h,v) = L(v) \text{ for all } v \in S_h \cap V. \tag{1.3}$$

1.2. The SPR and PPR techniques. In general, ∇u_h is inherently discontinuous across elements boundaries, and a postprocessing operation is needed to correct this problem. Recovery techniques such as the SPR and the PPR can be used for this purpose. The recovered gradient definition in both the SPR and the PPR relies

on the following simple observation: The basis functions of S_h are the Lagrange basis functions. Hence, every function in S_h is uniquely defined by its values at the mesh nodes. Let $\{v_z : z \in \mathcal{N}_h\}$ be the Lagrange basis of S_h, and let R_h denote the gradient recovery operator associated with either the SPR or the PPR. Assuming that $R_h u_h$ is defined at every mesh node $z \in \mathcal{N}_h$, the recovered gradient $R_h u_h$ on $\overline{\Omega}$ is defined to be

$$R_h u_h = \sum_{z \in \mathcal{N}_h} R_h u_h(z) v_z.$$

According to this definition, $R_h u_h \in S_h \times S_h$, and it remains to define $R_h u_h$ at the mesh nodes. This is where the SPR and the PPR are different.

Remark 1.1. The definition of the SPR- and the PPR-recovered gradients at mesh nodes involves best fitting operations. In this paper, the best fitting is carried out in a discrete least-squares sense.

The definition of the SPR-recovered gradient at $z \in \mathcal{N}_h$ depends on the location of z.

If $z \in \Omega$, let $\mathcal{K}_z$ denote the patch consisting of the triangles attached to z as shown in Figure 1(a). Let $p_x \in P_1(\mathcal{K}_z)$ be the linear polynomial that best fits $\partial_x u_h$ at the triangle centroids in $\mathcal{K}_z$. The recovered x-derivative at z is defined to be $p_x(z)$. Similarly, we can define the recovered y-derivative at z.

If $z \in \partial\Omega$ is directly connected to no internal mesh nodes, the recovered gradient at z is defined to be $\nabla u_h(z)$.

If $z \in \partial\Omega$ is directly connected to the internal mesh nodes $z_1, z_2, \ldots, z_{N_z}$, let $\mathcal{K}_{z_i}$ be the patch associated with z_i. Again, $\mathcal{K}_{z_i}$ consists of the mesh triangles that are directly attached to z_i. Let $p_{x,z_i} \in P_1(\mathcal{K}_{z_i})$ be the linear polynomial that best fits $\partial_x u_h$ at the triangle centroids in $\mathcal{K}_{z_i}$. The recovered x-derivative at z is defined to be

$$\frac{1}{N_z} \sum_{i=1}^{N_z} p_{x,z_i}(z).$$

Similarly, we can define the recovered y-derivative at z.

Next, we turn our attention to the definition of the PPR-recovered gradient at $z \in \mathcal{N}_h$. Starting with a patch $\mathcal{K}_z$, let $p \in P_2(\mathcal{K}_z)$ be the quadratic polynomial that best fits u_h at the mesh nodes in $\mathcal{K}_z$. The PPR-recovered gradient at z is defined to be $\nabla p(z)$. The construction of $\mathcal{K}_z$ is not straightforward as in the SPR. As we will soon see, $\mathcal{K}_z$ must have at least six mesh nodes that are not on a conic section. This is to guarantee the existence and the uniqueness of p. Indeed, the construction of $\mathcal{K}_z$ starts by the patch $\mathcal{K}_{z,0}$ that consists of the mesh triangles directly attached with z. The next construction step depends on the location of z.

If $z \in \Omega$ and $\mathcal{K}_{z,0}$ has at least five mesh triangles, then $\mathcal{K}_z = \mathcal{K}_{z,0}$ as shown in Figure 1(a).

If $z \in \Omega$ and $\mathcal{K}_{z,0}$ has three or four mesh triangles, then

$$\mathcal{K}_z = \mathcal{K}_{z,0} \cup \{T \in \mathcal{T}_h : T \cap \mathcal{K}_{z,0} \text{ is an edge of } T\}$$

as shown in Figure 1(b).

If $z \in \partial\Omega$ and $\mathcal{K}_{z,0}$ has at least one internal mesh node, then

$$\mathcal{K}_z = \mathcal{K}_{z,0} \cup \{\mathcal{K}_{\hat{z}} : \hat{z} \in \mathcal{K}_{z,0} \text{ is an internal mesh node}\} \tag{1.4}$$

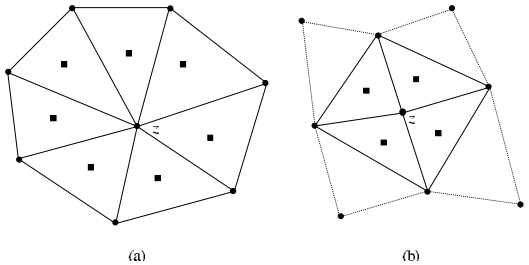


FIG. 1. *Patches used in gradient recovery at an internal mesh node z. Sampling points for the SPR are marked with ■, while those needed for the PPR are marked with ●.*

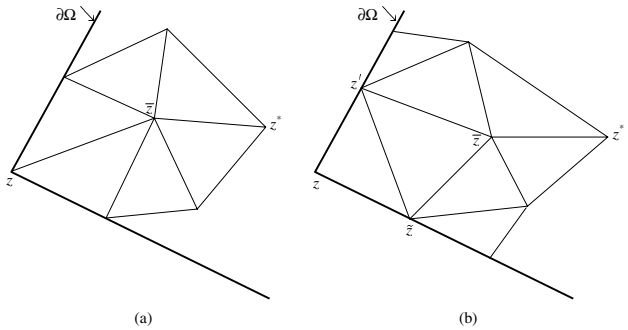


FIG. 2. *Examples for patches used in the PPR at a boundary mesh node z.*

as shown in Figure 2(a). If $\mathcal{K}_{z,0}$ has no internal mesh nodes, replace $\mathcal{K}_{z,0}$ in (1.4) by a bigger patch

$$\mathcal{K}_{z,1} = \bigcup \{\mathcal{K}_{\hat{z},0} : \hat{z} \in \mathcal{K}_{z,0} \text{ is a mesh node } \}. \tag{1.5}$$

An example of this situation is depicted in Figure 2(b). Practically, $\mathcal{K}_{z,1}$ must have at least one internal mesh node. If this is not the case, iterate the extension process in (1.5).

Having introduced the definitions of the SPR and the PPR, we have the following remarks.

1. The main difference between the SPR and the PPR is that the SPR works on ∇u_h while the PPR works on u_h.

2. The PPR has good approximation properties as it satisfies the *consistency condition*. A recovery operator R_h is said to satisfy the consistency condition if

$$R_h(I_h p) = \nabla p \quad \text{for all } p \in P_2(\Omega). \tag{1.6}$$

If R_h satisfies (1.6), the Bramble–Hilbert lemma can be used to show that

$$\|\nabla u - R_h I_h u\|_{L^\infty(\Omega)} \le C h^2 |u|_{3,\infty,\Omega} \quad \text{for all } u \in W^3_\infty(\Omega),$$

where $C > 0$ is some constant independent of u and h (see [13] for more details). The PPR satisfies (1.6) because the best fit polynomials over individual patches and the original polynomial are typically the same. On the other hand, the SPR does not satisfy the consistency condition unless $\mathcal{T}_h$ has some special structure.

3. Basically, the PPR can be viewed as a dynamic way to generate difference formulas for first order partial derivatives. The generated formulas can recover the

exact derivatives of quadratic polynomials. Babuška and Strouboulis [3] proposed a technique to generate this kind of difference formulas a priori (see Example 4.8*.4 in [3]), but this technique is not suitable for real time computations on real meshes.

4. The idea of best fitting the nodal values of u_h by a quadratic polynomial is well known in engineering applications. For example, Wiberg and Li [11] and Li and Wiberg [9] used this idea in constructing a recovered solution $\tilde{u}_h$ from u_h. According to their strategy, the true error $\|u-u_h\|_{L^2(T)}$ on $T \in \mathcal{T}_h$ is estimated using $\|\tilde{u}_h - u_h\|_{L^2(T)}$. Indeed, Wiberg and Li were mainly concerned about estimating the error and not recovering the gradient.

5. The PPR gradient recovery can be easily extended to higher order elements and to problems in $\mathbb{R}^3$. This will be the topic of a future work.

1.3. Gradient recovery and the superconvergence property. As mentioned previously, if the recovered gradient enjoys the superconvergence property, then it can be used in building an asymptotically exact a posteriori error estimator. Ainsworth and Oden [1] established a general framework that can be used in proving the superconvergence property, if it exists. Let R_h denote the recovery operator associated with a gradient recovery technique. According to this framework, there are three main requirements to show that $R_h u_h$ superconverges to ∇u:

1. R_h satisfies the consistency condition.
2. The recovery operator R_h is bounded in the following sense:

$$\|R_h v\|_{L^2(T)} \le C|v|_{1,\mathcal{K}_T} \qquad \text{for all } T \in \mathcal{T}_h \text{ and for all } v \in S_h, \tag{1.7}$$

where $\mathcal{K}_T$ is a patch of triangles containing T.

3. ∇u_h enjoys superconvergence in the following sense:

$$\|\nabla(I_h u - u_h)\|_{L^2(\Omega)} \le C h^{1+\rho} \tag{1.8}$$

for some $\rho \in (0,1]$ and some constant $C > 0$ that is independent of h.

If u_h and R_h satisfy the above requirements, then it is possible to prove that

$$\|\nabla u - R_h u_h\|_{L^2(\Omega)} \le C h^{1+\rho}. \tag{1.9}$$

With this result in hand, it is straightforward to prove that the a posteriori error estimator

$$\eta_h = \|R_h u_h - \nabla u_h\|_{L^2(\Omega)} \tag{1.10}$$

is asymptotically exact.

From this point on, we will concentrate on the PPR and its corresponding operator, which we will denote by $\mathrm{G^h}$. Our target in this paper is to show that $\mathrm{G^h}u_h$ superconverges to ∇u following the above framework. By construction, and as explained previously, $\mathrm{G^h}$ satisfies the first requirement. For the third requirement, Xu and Zhang [12] had recently established (1.8) for a wide range of meshes that are mildly structured in the sense of the following definition.

DEFINITION 1.2. *The triangulation $\mathcal{T}_h$ is said to satisfy the* condition (α, σ) *if there exist a partition $\mathcal{T}_{h,1} \cup \mathcal{T}_{h,2}$ of $\mathcal{T}_h$ and positive constants α and σ such that every two adjacent triangles in $\mathcal{T}_{h,1}$ form an $O(h^{1+\alpha})$ parallelogram and*

$$\sum_{T \in \mathcal{T}_{h,2}} |T| = O(h^{\sigma}).$$

An $O(h^{\alpha+1})$ parallelogram is a quadrilateral in which the difference between the lengths of any two opposite sides is $O(h^{\alpha+1})$. When $\alpha = \infty$, every pair of adjacent triangles in $\mathcal{T}_{h,1}$ form a parallelogram. When $\alpha = \sigma = \infty$, $\mathcal{T}_h$ is uniformly generated by lines parallel to three fixed directions. This case was handled in [7], where $u - u_h$ was expanded at mesh nodes, and the case in which $\alpha = 1$ was handled in [6]. For general α and σ, Xu and Zhang [12] proved the following theorem.

THEOREM 1.3. *Let u be the solution of* (1.2)*, let $u_h \in S_h$ be the finite element solution of* (1.3)*, and let $I_h u \in S_h$ be the linear interpolation of u. If the triangulation $\mathcal{T}_h$ satisfies the condition (α, σ) and $u \in H^3(\Omega) \cap W^2_\infty(\Omega)$, then*

$$\|u_h - I_h u\|_{1,\Omega} \le h^{1+\rho}(\|u\|_{3,\Omega} + |u|_{2,\infty,\Omega}),$$

where $\rho = \min(\alpha, \frac{1}{2}, \frac{\sigma}{2})$.

Remark 1.4. The condition(α, σ) is sufficient to guarantee the superconvergence result in (1.8), although it is not necessary, as we shall see in the numerical examples. Nevertheless, this condition is satisfied for meshes generated by many automatic mesh generators as described in [12].

The second requirement is somewhat easy to establish when the recovery technique works directly on the gradient as in weighted average recovery and the SPR. However, the situation is much harder for the PPR as it works on function values. It is not even clear how to relate $\mathrm{G^h}v$ to ∇v, where $v \in S_h$. Actually, the core of this paper is devoted to showing that $\mathrm{G^h}$ satisfies the third requirement. Having this result paves the way to show that $\mathrm{G^h}$ enjoys the superconvergence property in (1.9) and that the error estimator η_h is asymptotically exact. At the end of the paper, some numerical examples are provided to practically show that the PPR-recovered gradient superconverges to the exact gradient.

2. Definition and existence of $\mathbf{G^h}$. As mentioned previously, the construction of $\mathrm{G^h}v \in S_h \times S_h$ for a function $v \in S_h$ is complete if $(\mathrm{G^h}v)(z)$ is defined for every $z \in \mathcal{N}_h$. Therefore, it suffices to address the definition and existence questions at the level of mesh nodes.

Consider a mesh node z, and let $\mathcal{K}_z$ denote its corresponding patch. In the patch $\mathcal{K}_z$, let $T_1, T_2, \ldots, T_m$ denote the mesh triangles and let $z_0 = (x_0, y_0), z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$ denote the mesh nodes. Without loss of generality, let $z = z_0$. Let $h_z = \max\{\|z_i - z_0\| : 1 \le i \le n\}$. To avoid the computational instability associated with small h_z, the computations will be carried out on the patch

$$\omega_z = F_z(\mathcal{K}_z), \text{ where } F_z : (x, y) \to (\hat{x}, \hat{y}) = \frac{(x, y) - (x_0, y_0)}{h_z}. \tag{2.1}$$

The patch ω_z will be called *the reference patch associated with* z. For $0 \le i \le n$, let $\hat{z}_i = (\hat{x}_i, \hat{y}_i) = F(z_i)$ and set $v_i = (v \circ F_z^{-1})(\hat{z}_i) = v(z_i)$. Let $p_z \in P_2(\omega_z)$ be the quadratic polynomial that best fits the data points $\{(\hat{z}_i, v_i) : 0 \le i \le n\}$. For $(\hat{x}, \hat{y}) \in \omega_z$, $p_z(\hat{x}, \hat{y})$ can be written in the form

$$p_z(\hat{x}, \hat{y}) = \hat{\boldsymbol{x}}^T \boldsymbol{c_z},$$

where $\boldsymbol{c}_z$ is the coefficients vector $\begin{bmatrix} c_{z,1} & c_{z,2} & c_{z,3} & c_{z,4} & c_{z,5} & c_{z,6} \end{bmatrix}^T$ and $\hat{\boldsymbol{x}}$ is the monomials vector $\begin{bmatrix} 1 & \hat{x} & \hat{y} & \hat{x}^2 & \hat{x}\hat{y} & \hat{y}^2 \end{bmatrix}^T$. Since we are using discrete least-squares fitting, the coefficients vector $\boldsymbol{c}_z$ is determined by the linear system

$$A_z^T A_z \boldsymbol{c}_z = A_z^T \boldsymbol{v}_z,$$

where

$$A_z = \begin{bmatrix} \hat{\boldsymbol{x}}_0^T \\ \hat{\boldsymbol{x}}_1^T \\ \hat{\boldsymbol{x}}_2^T \\ \vdots \\ \hat{\boldsymbol{x}}_n^T \end{bmatrix}, \hat{\boldsymbol{x}}_i = \begin{bmatrix} 1 \\ \hat{x}_i \\ \hat{y}_i \\ \hat{x}_i^2 \\ \hat{x}_i\hat{y}_i \\ \hat{y}_i^2 \end{bmatrix} \text{ for } 0 \le i \le n, \text{ and } \boldsymbol{v}_z = \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}. \tag{2.2}$$

Set

$$B_z = A_z^T A_z,$$

then, assuming the existence of B_z^{-1},

$$\boldsymbol{c}_z = B_z^{-1} A_z^T \boldsymbol{v}_z.$$

By definition,

$$\mathrm{G^h} v(z) = \frac{1}{h_z} \begin{bmatrix} \partial_{\hat{x}} p_z(0,0) & \partial_{\hat{y}} p_z(0,0) \end{bmatrix}^T.$$

Therefore,

$$\mathrm{G^h} v(z) = \frac{1}{h_z} \begin{bmatrix} c_{z,2} & c_{z,3} \end{bmatrix}^T = \frac{1}{h_z} \begin{bmatrix} \boldsymbol{v}_z^T A_z B_z^{-1} e_2 & \boldsymbol{v}_z^T A_z B_z^{-1} e_3 \end{bmatrix}^T, \tag{2.3}$$

where e_2 and e_3 are the second and third columns, respectively, of the identity matrix $I_{6\times 6}$.

To this end, it is important to address the following question: Are there any sufficient conditions that guarantee the existence of B_z^{-1}? The answer to this question relies on the following simple proposition.

PROPOSITION 2.1.

1. *If B_z is not invertible, then there is a conic section passing through the mesh nodes in $\mathcal{K}_z$.*
2. *Any tangent to a branch of a hyperbola cannot intersect with the other branch.*

Proof. If B_z is not invertible, then $\boldsymbol{c}_z$ has infinitely many solutions. Therefore, there are infinitely many polynomials in $P_2(\omega_z)$ that pass through the data points $\{(\hat{z}_i, v_i) : 0 \le i \le n\}$. Let $p_{z,1}$ and $p_{z,2}$ be two such polynomials, and let $q = p_{z,1} - p_{z,2}$. Then, $q(\hat{z}_i) = 0$ for $0 \le i \le n$, and the conic section $\{(\hat{x}, \hat{y}) : q_z(\hat{x}, \hat{y}) = 0\}$ passes through $\{\hat{z}_i : 0 \le i \le n\}$. Since conic sections are invariant under affine mappings, the proof of the first part is complete. The proof of the second part is elementary. ∎

DEFINITION 2.2. *The patch $\mathcal{K}_z$ (or ω_z) is said to satisfy the* angle condition *if the sum of any two adjacent angles inside $\mathcal{K}_z$ is at most π, and is said to satisfy the* line condition *if its mesh nodes are not lying on two lines.*

Let n_1 denote the number of mesh nodes that are directly connected to z, and set $n_2 = n - n_1$. If $z \in \Omega$, then $n_1 \ge 3$. Practically, a good mesh generator can detect any node z for which $n_1 = 3$ and remove it. So, we may assume that $n_1 \ge 4$. It is obvious that for an internal mesh node z with $n_1 > 4$, $\mathcal{K}_z$ automatically satisfies the line condition. If $n_1 = 4$, $\mathcal{K}_z$ may violate this condition as shown in Figure 4.

The following theorem plays the crucial part in proving the boundedness of $\mathrm{G^h}$.

THEOREM 2.3. *Let z be an internal mesh node with $n_1 \geq 4$, and let $\mathcal{K}_z$ be its corresponding patch that satisfies the angle condition. In addition, let $\mathcal{K}_z$ satisfy the line condition when $n_1 = 4$. Then, B_z is invertible.*

Proof. By the first part of Proposition 2.1, it suffices to show that $\mathcal{K}_z$ has six distinct nodes that are not on a conic section. Since $z \in \Omega$, the sum of the angles at z is 2π. Hence, the nodes in $\mathcal{K}_z$ cannot lie on a circle, on a parabola, on an ellipse, or on one branch of a hyperbola. Since $\mathcal{K}_z$ satisfies the line condition, the nodes cannot be on two lines. The remaining possibility is to have the nodes distributed on two branches of a hyperbola. Depending on n_1, we can have one of the following two cases.

Case 1: $n_1 = 4$. In this case the triangles attached to z_0 form a quadrilateral as shown in Figure 4. Since $\mathcal{K}_z$ satisfies the angle condition, z_0 must be the intersection point of the quadrilateral diagonals. Hence, the nodes in $\mathcal{K}_z$ cannot be distributed on two branches of a hyperbola as a line intersects with a hyperbola at no more than two points.

Case 2: $n_1 > 4$. Proceed by contradiction and assume that the nodes in $\mathcal{K}_z$ are distributed on two branches of a hyperbola. Without loss of generality, assume that the real axis of the hyperbola is horizontal and that z_0 is on the right branch of the hyperbola. The left branch cannot have more than two mesh nodes. If it has three nodes as in Figure 3(a), then the sum of the angles at z_2 is more than π.

If the left branch has two nodes, then $\mathcal{K}_z$ must have nodes z_3 and z_4 on the right branch as shown in Figure 3(b). We claim that $\mathcal{K}_z$ cannot have any more nodes on the right branch. If this claim is true, $\mathcal{K}_z$ will have $n_1 = 4$ and this is a contradiction as $n_1 > 4$ by assumption. To prove the previous claim, assume that $\mathcal{K}_z$ has another node z_5 as in Figure 3(b). Then, the sum of the angles at node z_3 is greater than π unless the nodes z_1, z_3, and z_5 lie on a line that is tangent to the right branch of the hyperbola, which is impossible by the second part of Proposition 2.1.

The only remaining possibility is to have exactly one node on the left branch of the hyperbola as in Figure 3(c). Again, by an argument similar to the one used in previous case, this leads to a contradiction. ∎

COROLLARY 2.4. *Consider a boundary mesh node z, and let $\mathcal{K}_z$ be its corresponding patch. Suppose that $\mathcal{K}_z$ contains another patch $\mathcal{K}_{\hat{z}}$ corresponding to an internal mesh node $\hat{z}$. If $\mathcal{K}_{\hat{z}}$ satisfies the angle and the line conditions, then B_z is invertible.*

3. Boundedness of $\mathbf{G^h}$. Let $v \in S_h$, let $\mathcal{K}_z$ be the patch associated with $z \in \mathcal{N}_h$, and consider the mesh triangle $T_k \subset \mathcal{K}_z$ for some $1 \leq k \leq m$. Let the vertices of T_k be $(x_{k,1}, y_{k,1})$, $(x_{k,2}, y_{k,2})$, and $(x_{k,3}, y_{k,3})$, where the numbering is in a counterclockwise direction. Since $v \in P_1(T_k)$, it is easy to verify that

$$\partial_x v(x,y) = \sum_{j=1}^{3} a_{k,j} v_{k,j} \text{ and } \partial_y v(x,y) = \sum_{j=1}^{3} b_{k,j} v_{k,j} \text{ for } (x,y) \in T_k, \tag{3.1}$$

where

$$a_{k,j} = \frac{1}{2|T_k|}(y_{k,j+1} - y_{k,j+2}), b_{k,j} = \frac{1}{2|T_k|}(x_{k,j+2} - x_{k,j+1}), v_{k,j} = v(x_{k,j}, y_{k,j}),$$

and the addition in indices is mod 3. Equivalently,

$$\partial_x v(x,y) = \boldsymbol{v}_k^T \boldsymbol{a}_k \text{ and } \partial_y v(x,y) = \boldsymbol{v}_k^T \boldsymbol{b}_k \text{ for } (x,y) \in T_k,$$

where $\boldsymbol{a}_k = \begin{bmatrix} a_{k,1} & a_{k,2} & a_{k,3} \end{bmatrix}^T$, $\boldsymbol{b}_k = \begin{bmatrix} b_{k,1} & b_{k,2} & b_{k,3} \end{bmatrix}^T$, and $\boldsymbol{v}_k = \begin{bmatrix} v_{k,1} & v_{k,2} & v_{k,3} \end{bmatrix}^T$.

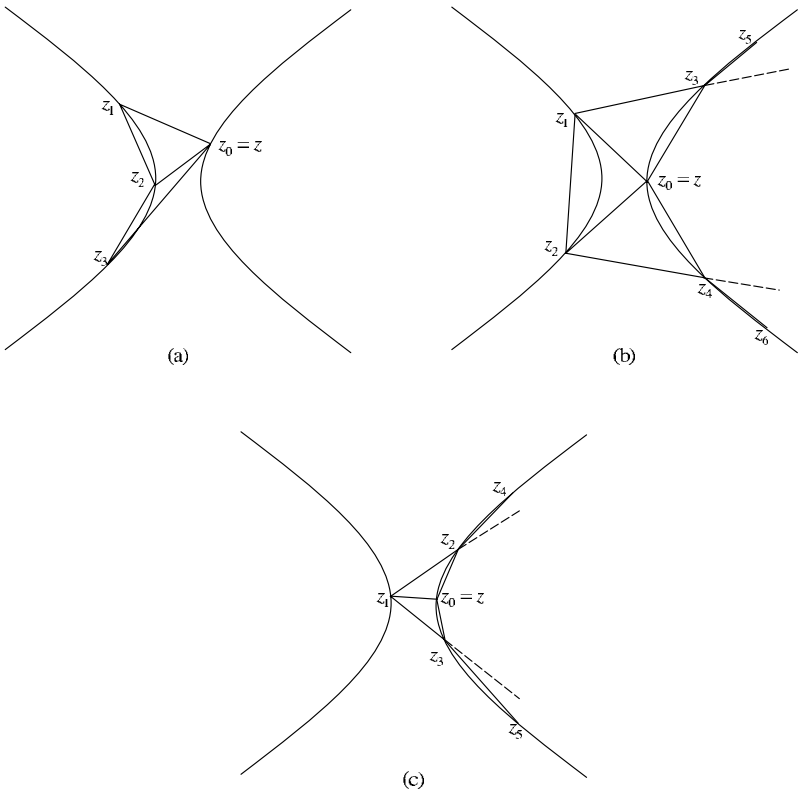


FIG. 3. *Nodes in ω_z cannot be distributed on two branches of a hyperbola when ω_z satisfies the angle condition.*

Let E_k be an $(n+1) \times 3$ Boolean matrix defined for T_k, where

$$E_k(i,j) = \begin{cases} 1 & \text{if the node } i \text{ in } \mathcal{K}_z \text{ is the vertex } j \text{ in } T_k, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\boldsymbol{v}_k = E_k^T \boldsymbol{v}_z$$

and (3.1) can be simplified to the form

$$\partial_x v(x,y) = \boldsymbol{v}_z^T E_k \boldsymbol{a}_k \text{ and } \partial_y v(x,y) = \boldsymbol{v}_z^T E_k \boldsymbol{b}_k \text{ for } (x,y) \in T_k. \tag{3.2}$$

Let ω_z be the reference patch associated with z, and let F_z be the affine mapping from $\mathcal{K}_z$ to ω_z. Let $\hat{T}_k = F_z(T_k)$, and let $(\hat{x}_{k,j}, \hat{y}_{k,j}) = F_z(x_{k,j}, y_{k,j})$ for $1 \le j \le 3$. It is easy to verify that

$$a_{k,j} = \frac{\hat{a}_{k,j}}{h_z} \text{ and } b_{k,j} = \frac{\hat{b}_{k,j}}{h_z},$$

where

$$\hat{a}_{k,j} = \frac{1}{2|\hat{T}_k|}(\hat{y}_{k,j+1} - \hat{y}_{k,j+2}) \text{ and } \hat{b}_{k,j} = \frac{1}{2|\hat{T}_k|}(\hat{x}_{k,j+2} - \hat{x}_{k,j+1}).$$

Let $\hat{\boldsymbol{a}}_k = \begin{bmatrix} \hat{a}_{k,1} & \hat{a}_{k,2} & \hat{a}_{k,3} \end{bmatrix}^T$ and $\hat{\boldsymbol{b}}_k = \begin{bmatrix} \hat{b}_{k,1} & \hat{b}_{k,2} & \hat{b}_{k,3} \end{bmatrix}^T$. Then (3.2) can be rewritten in the form

$$\partial_x v(x,y) = \frac{\boldsymbol{v}_z^T E_k \hat{\boldsymbol{a}}_k}{h_z} \text{ and } \partial_y v(x,y) = \frac{\boldsymbol{v}_z^T E_k \hat{\boldsymbol{b}}_k}{h_z} \text{ for } (x,y) \in T_k. \tag{3.3}$$

Let $\mathrm{G}^{\mathrm{h}}_1 v$ and $\mathrm{G}^{\mathrm{h}}_2 v$ stand for the recovered x- and y-derivatives, respectively. Establishing the boundedness of G^{h} in the sense of (1.7) would be easy if $\mathrm{G}^{\mathrm{h}}_l v(z)$ can be expressed as a linear combination of the first partial derivatives of v on the triangles of ω_z for $l = 1, 2$. So, we will try to find a set of *bounded values* $\alpha_{z,l,1}, \ldots, \alpha_{z,l,m}, \beta_{z,l,1}, \ldots, \beta_{z,l,m}$ such that

$$\mathrm{G}^{\mathrm{h}}_l v(z) = \sum_{k=1}^{m} [\alpha_{z,l,k} (\partial_x v)_k + \beta_{z,l,k} (\partial_y v)_k] \text{ for } l = 1, 2, \tag{3.4}$$

where $(\partial_x v)_k$ and $(\partial_y v)_k$ are the first partial derivatives of v in T_k. Using (2.3) and (3.3) in (3.4), we get

$$\boldsymbol{v}_z^T \sum_{k=1}^{m} [\alpha_{z,l,k} E_k \hat{\boldsymbol{a}}_k + \beta_{z,l,k} E_k \hat{\boldsymbol{b}}_k] = \boldsymbol{v}_z^T A_z B_z^{-1} e_{l+1} \text{ for } l = 1, 2.$$

Setting

$$M_z = \begin{bmatrix} E_1 \hat{\boldsymbol{a}}_1 & \cdots & E_m \hat{\boldsymbol{a}}_m & E_1 \hat{\boldsymbol{b}}_1 & \cdots & E_m \hat{\boldsymbol{b}}_m \end{bmatrix} \tag{3.5}$$

and

$$\boldsymbol{\gamma}_{z,l} = \begin{bmatrix} \alpha_{z,l,1} & \cdots & \alpha_{z,l,m} & \beta_{z,l,1} & \cdots & \beta_{z,l,m} \end{bmatrix}^T,$$

we get

$$\boldsymbol{v}_z^T M_z \boldsymbol{\gamma}_{z,l} = \boldsymbol{v}_z^T A_z B_z^{-1} e_{l+1} \text{ for } l = 1, 2.$$

Since this is true for all $v \in S_h$,

$$M_z \boldsymbol{\gamma}_{z,l} = A_z B_z^{-1} e_{l+1} \text{ for } l = 1, 2. \tag{3.6}$$

Note that the order of M_z is $(n+1) \times (2m)$.

LEMMA 3.1. *Consider $z \in \mathcal{N}_h$. If the patch $\mathcal{K}_z$ corresponding to z has no degenerate triangles and B_z is invertible, then* $\operatorname{Rank} M_z = n$ *and the system in* (3.6) *has infinitely many solutions.*

Proof. Since $\mathcal{K}_z$ is simply connected, then, using Euler's theorem, $(n+1) - e + m = 1$, where e is the number of edges in $\mathcal{K}_z$. Hence, $(n+1) - 2m = e - 3m + 1$. By a simple induction argument on m, we can show that $e - 3m + 1 < 0$ for $m \geq 3$. Hence, the system in (3.6) is underdetermined.

To prove that $\operatorname{Rank} M_z = n$, consider the homogeneous linear system

$$M_z^T \boldsymbol{w} = 0 \tag{3.7}$$

with $\boldsymbol{w} = \begin{bmatrix} w_0 & w_1 & \cdots & w_n \end{bmatrix}^T$. We can view $w_0, w_1, \ldots, w_n$ as the nodal values of some function $w \in S_h$ at the nodes $z_0, z_1, \ldots, z_n$ in $\mathcal{K}_z$. With this in mind, the homogeneous system in (3.7) implies that $\nabla w = 0$ in T_k for $k = 1, 2, \ldots, m$. Hence,

w is constant on $\mathcal{K}_z$, as $w \in S_h$, and $w_0 = w_1 = \cdots = w_n$ in any solution $\boldsymbol{w}$ of (3.7). Consequently, the dimension of the null space of M_z^T is 1 and $\operatorname{Rank} M_z^T = \operatorname{Rank} M_z = n$. Moreover, the only row operation on M_z that leads to a row of zeros is adding all the rows together. Since $\mathrm{G^h}$ recovers the exact gradient for any polynomial $p \in P_2(\omega_z)$, it is easy to verify that the row sum of $A_z B_z^{-1} e_{l+1}$ is 0 for $l = 1, 2$. Therefore, the homogeneous system in (3.6) is consistent for $l = 1, 2$. ∎

Among all the solutions of (3.6), we consider the one with the minimum length given by

$$\boldsymbol{\gamma}_{z,l}^* = M_z^\dagger A_z B_z^{-1} e_{l+1} \text{ for } l = 1, 2, \tag{3.8}$$

where $M_z^\dagger$ is the pseudoinverse of M_z. For every mesh triangle T, define the patch

$$\mathcal{K}_T = \bigcup \{\mathcal{K}_z : z \text{ is a vertex of } T\}.$$

For any matrix $K \in \mathbb{R}^{k_1 \times k_2}$, let $\sigma_1(K)$ and $\sigma_{\min(k1,k2)}(K)$ denote the largest and the smallest singular values of K, respectively. Recall that $\sigma_l^2(K) = \sigma_l(K^T K)$ for $l = 1, 2, \ldots, \min(k1, k2)$.

THEOREM 3.2. *Let $0 < C_1 \le \sigma_6(A_z) \le \sigma_1(A_z) \le C_2$ and $0 < C_3 \le \sigma_n(M_z)$ for every mesh node $z \in \mathcal{N}_h$ and for some constants C_1, C_2, and C_3 that are independent of h. Then, there exists a constant C, independent of h, such that*

$$\| \mathrm{G^h} v\|_{L^2(T)} \le C |v|_{1,\mathcal{K}_T} \tag{3.9}$$

for all $T \in \mathcal{T}_h$ and for all $v \in S_h$.

Proof. Consider a mesh triangle T, and let $\mathcal{K}_T$ be the patch corresponding to T. Let z be a vertex of T and consider any $v \in S_h$. Using (3.4) and (3.8) we get

$$\begin{aligned} |\, \mathrm{G^h}_l\, v(z)| &\le \|\boldsymbol{\gamma}_l^*\|_1 |v|_{1,\infty,\mathcal{K}_z} \le c_1 \|\boldsymbol{\gamma}_l^*\|_2 |v|_{1,\infty,\mathcal{K}_T} \\ &\le c_2 \|M^\dagger\|_2 \|A\|_2 \|B^{-1}\|_2 |v|_{1,\infty,\mathcal{K}_T} \\ &\le \frac{c_2 C_2}{C_3 C_1^2} |v|_{1,\infty,\mathcal{K}_T} \end{aligned}$$

for $l = 1, 2$. Since $\mathrm{G^h} v \in P_1(T) \times P_1(T)$,

$$\| \mathrm{G^h} v\|_{L^\infty(T)} \le C |v|_{1,\infty,\mathcal{K}_T}.$$

Hence,

$$\begin{aligned} \| \mathrm{G^h} v\|_{L^2(T)} &\le \sqrt{|T|}\, \| \mathrm{G^h} v\|_{L^\infty(T)} \le C \operatorname{diam}(T) |v|_{1,\infty,\mathcal{K}_T} \\ &\le C \frac{\operatorname{diam}(T)}{\operatorname{diam}(\mathcal{K}_T)} |v|_{1,\mathcal{K}_T} \le C |v|_{1,\mathcal{K}_T}. \end{aligned} \tag{3.10}$$

The inequality in (3.10) is obtained using an inverse estimate. ∎

It is obvious that the bounds assumed about the singular values of A_z and M_z in Theorem 3.2 depend on the mesh geometry as shown in the following theorem.

THEOREM 3.3. *Let $\mathcal{T}_h$ be a triangular partition of Ω that satisfies the following conditions.*

1. *There exists a finite positive integer N such that $n_1 \le N$ for all $z \in \mathcal{N}_h$, and $n_1 \ge 4$ if $z \in \Omega$.*
2. *The sum of any two adjacent angles at $z \in \mathcal{N}_h$ is π if $z \in \Omega$ with $n_1 = 4$, and is at most π if $z \in \partial\Omega$ or if $z \in \Omega$ with $n_1 > 4$.*

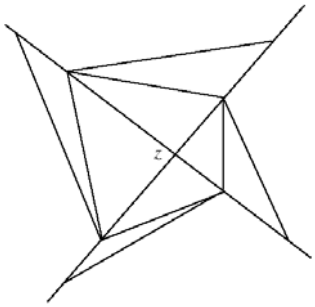


FIG. 4. *An example for patch $\mathcal{K}_z$, corresponding to an internal mesh node z, that does not satisfy the line condition.*

3. *If $z \in \Omega$ with $n_1 = 4$, then the sum of the two adjacent angles in $\mathcal{K}_z$ at one of the mesh nodes directly connected to z is at most $\pi - \phi$ for some $0 < \phi < \pi$.*
4. *If $\theta_{min,h}$ and $\theta_{max,h}$ are the smallest and the largest angles in $\mathcal{T}_h$, respectively, then there exist constants $\underline{\phi}$ and $\overline{\phi} < \pi$ such that*

$$0 < \underline{\phi} \le \theta_{min,h} \le \theta_{max,h} \le \overline{\phi} < \pi.$$

5. *Every boundary mesh node z is connected to an internal mesh node $\bar{z}$ either directly or indirectly through at most one boundary mesh node.*

Then, there exist constants C_1, C_2, and C_3, independent of h, such that

$$0 < C_1 \le \sigma_6(A_z) \le \sigma_1(A_z) \le C_2 \text{ and } 0 < C_3 \le \sigma_n(M_z) \quad \text{for all } z \in \mathcal{N}_h.$$

Remark 3.4. The third condition in Theorem 3.3 is imposed to avoid the singular situation shown in Figure 4. Also, the fifth condition can be relaxed.

Proof. Let $z \in \mathcal{N}_h$, and let D denote the closed unit disk centered at (0,0). Since the reference patch $\omega_z \subset D$, $|B_z(i,j)| \le (n+1)$ for $1 \le i, j \le 6$. Hence,

$$\sigma_1(A_z) = \sqrt{\sigma_1(B_z)} \le \sqrt{\sigma_1(|B_z|)} \le \sqrt{6(n+1)}.$$

By the first and the fifth conditions, it is easy to verify that $n \le N$ if $z \in \Omega$, $n \le N^2 - 5N + 10$ if $z \in \partial\Omega$ and z is directly attached to an internal node, and $n \le 3N - 2$ if $z \in \partial\Omega$ and z is indirectly attached to an internal node through a third boundary node. Hence, there exists $C_2 = C_2(N)$ such that $\sigma_1(A_z) \le C_2$.

To establish the existence of C_1, let us first consider the internal nodes. Let ω_z be a reference patch associated with an internal mesh node z. By definition, $\omega_z \subset D$, ω_z has a node at $(0,0)$, and ω_z has at least one node on ∂D. The first four conditions imply that ω_z has no degenerate triangles and that ω_z satisfies the line and the angle conditions. To show that $\inf\{\sigma_6(A_z) : z \in \mathcal{N}_h \cap \Omega\} \ge C_1 > 0$ for any $h > 0$, proceed by contradiction and assume that there exists a sequence of reference patches $\{\omega_i\}_{i=1}^{\infty}$ such that ω_i has all the properties of ω_z for all $i \ge 1$ and $\sigma_6(A_i) \to 0$, where A_i is the matrix defined for ω_i as in (2.2). Without loss of generality, and by the first condition, one may assume n and n_1 are the same for any patch ω_i; otherwise, we may pass to a subsequence. For $i \ge 1$, the nodes in ω_i are $\tilde{z}_{i,0} = (0,0), \tilde{z}_{i,1}, \tilde{z}_{i,2}, \ldots, \tilde{z}_{i,n}$. According to n_1, we have two cases.

Case 1: $n_1 > 4$. By compactness of D, one may assume that $\tilde{z}_{i,j} \to \tilde{z}_j \in D$ for $0 \le j \le n$; otherwise, one may pass to a subsequence. The nodes $\tilde{z}_0 = (0,0), \tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n$

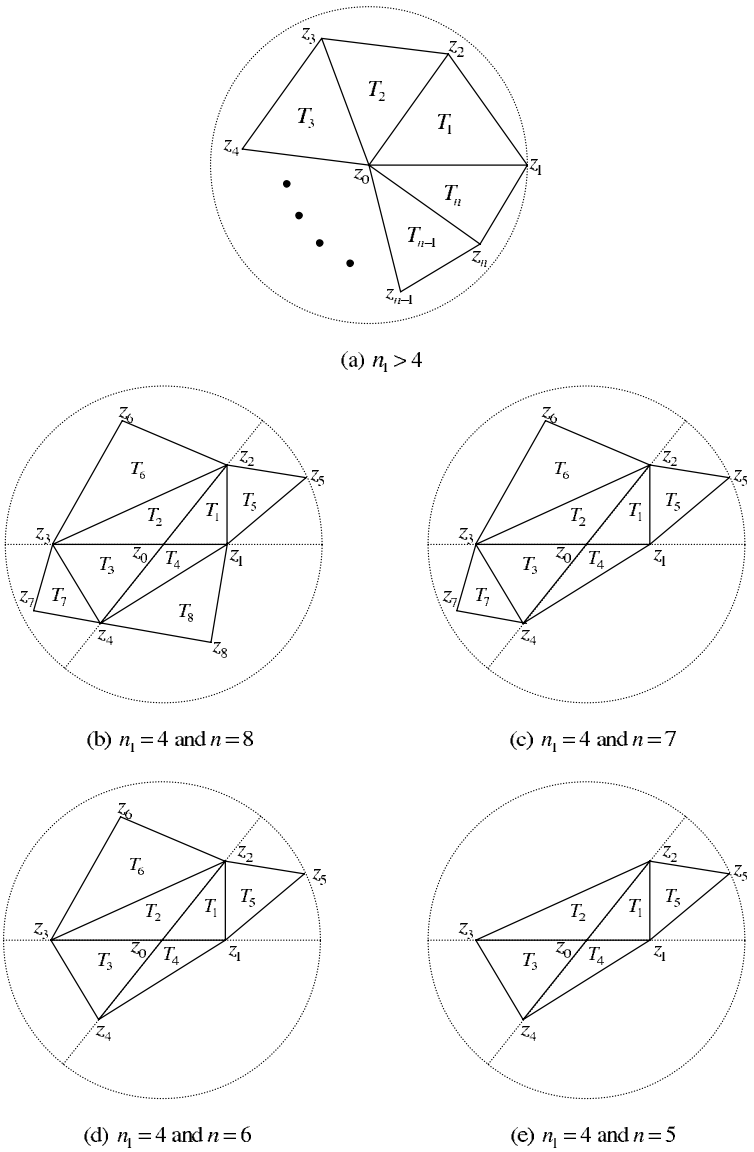


FIG. 5. *The possible patterns for a patch $\mathcal{K}_z$ corresponding to an internal mesh node z.*

can be viewed as the nodes of a patch ω whose pattern is similar to the one shown in Figure 5(a). Using the properties of ω_i, none of the triangles in ω is degenerate, ω has at least one node on ∂D, and the sum of any two adjacent angles in ω cannot exceed π. Hence, ω satisfies the angle and the line conditions. If A denotes the matrix defined for ω as in (2.2), then $\sigma_6(A) \neq 0$. Since $\tilde{z}_{i,j} \to \tilde{z}_j$ for all $0 \leq j \leq n$, $A_i \to A$ in any matrix norm. Hence, $0 \neq \sigma_6(A) = \lim_{i\to\infty} \sigma_6(A_i) = 0$, which is a contradiction.

Case 2: $n_1 = 4$. Since ω_i satisfies the angle condition, it contains four nodes that are corners of a convex quadrilateral whose diagonals intersect at $\tilde{z}_{i,0} = (0, 0)$. Denote the quadrilateral in ω_i by Q_i and denote the set of its diagonals by ℓ_i. Since one of the nodes in ω_i is on ∂D, it is easy to verify that Q_i inscribes a circle whose radius is

at least $\delta = \delta(\underline{\phi})$ for some $\delta > 0$ and for all $i \geq 1$. Consequently,

$$\max\{\text{dist}(\tilde{z}_{i,j}, \ell_i) : 0 \leq j \leq n\} \geq \delta \sin(\phi) > 0. \tag{3.11}$$

Since D is compact, assume that $\tilde{z}_{i,j} \to \tilde{z}_j$ for $0 \leq j \leq n$. The nodes $\tilde{z}_0 = (0,0), \tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n$ can be viewed as the nodes of a patch ω similar to one of the patterns shown in Figure 5(b)–(e). The patch ω has a quadrilateral denoted by Q, and the diagonals of Q are denoted by ℓ. The corners Q are the limits of the corners in Q_i. Hence, (3.11) leads to

$$\max\{\text{dist}(\tilde{z}_j, \ell) : 0 \leq j \leq n\} \geq \delta \sin(\phi),$$

and ω satisfies the line and the angle conditions. As in the previous case, this leads to a contradiction.

Let us turn our attention to the boundary mesh nodes, and let $\mathcal{K}_z$ be the patch corresponding to a boundary mesh node z. By construction, $\mathcal{K}_z$ contains, at least, one patch $\mathcal{K}_{\bar{z}}$ corresponding to an internal mesh node $\bar{z}$. By the fifth condition, $\bar{z}$ is either connected to z directly or indirectly through a third boundary mesh node $\tilde{z}$. As before, $\mathcal{K}_z$ is a subset of a disk centered at z and its radius is h_z. Hence, and without loss of generality, there exists a mesh node $z^* \in \mathcal{K}_{\bar{z}}$, as shown in Figure 2, such that

$$\|z - \bar{z}\| + \|\bar{z} - z^*\| \geq \|z - z^*\| = h_z.$$

This implies that $\max(\|z - \bar{z}\|, \|\bar{z} - z^*\|) \geq h_z/2$. If z is directly connected to $\bar{z}$ as in Figure 2(a), then $\text{diam}(\mathcal{K}_{\bar{z}}) \geq h_z$. If z is indirectly connected to $\bar{z}$ as in Figure 2(a), then we have two situations depending on whether $\|\bar{z} - z^*\| \geq h_z/2$ or $\|z - \bar{z}\| \geq h_z/2$. In the former situation $\text{diam}(\mathcal{K}_{\bar{z}}) \geq h_z$, and for the later situation consider the triangle $z\tilde{z}\bar{z}$. In this triangle the angle at z is at least $\underline{\phi}/2$ by the fourth condition, and hence $\|\bar{z} - \tilde{z}\| \geq h_z \sin(\underline{\phi}/2)/2$. If not, we can use the triangle $z\acute{z}\bar{z}$. Consequently, $\text{diam}(\mathcal{K}_{\bar{z}}) \geq h_z \sin(\underline{\phi}/2)$.

Hence, the reference patch ω_z contains a patch $\bar{\omega}$ that is a scaled translation of the reference patch $\omega_{\bar{z}}$ with $\text{diam}(\bar{\omega}) \geq \sin(\underline{\phi}/2)$. We may write $A_z = [A_{z,1}^T \, A_{z,2}^T]^T$, where $A_{1,z}$ corresponds to the nodes in $\bar{\omega}$ and $A_{z,2}$ corresponds to the rest of the nodes in ω_z. Hence, $B_z = A_{z,1}^T A_{z,1} + A_{z,2}^T A_{z,2}$. Set $B_{z,1} = A_{z,1}^T A_{z,1}$. Since $\bar{\omega}$ satisfies both the line and the angle conditions, both B_z and $B_{z,1}$ are positive definite. Moreover, $\sigma_6(B_z) \geq \sigma_6(B_{z,1})$. Hence,

$$\sigma_6(A_z) = \sqrt{\sigma_6(B_z)} \geq \sqrt{\sigma_6(B_{z,1})} = \sigma_6(A_{z,1}).$$

Using the results established for internal nodes, $\sigma_6(A_z) \geq \tilde{C}_1 > 0$ for all $z \in \partial\Omega$.

Next, we prove the second inequality about $\sigma_n(M_z)$. Again, let $z \in \mathcal{N}_h$, and let $\mathcal{K}_z$ be its corresponding patch. By the fourth condition, none of the mesh triangles in $\mathcal{K}_z$ is degenerate. Since $\sigma_6(A_z) \geq C_1 > 0$, B_z is invertible. Hence, and by Lemma 3.1, $\text{Rank}\, M_z = n$. Since the rank of a matrix can be viewed as the number of its nonzero singular values, $\sigma_n(M_z) > 0$. Using this fact and an argument similar to the one used to establish $\sigma_6(A_z) \geq C_1 > 0$, one can show that $\sigma_n(M_z) \geq C_3 > 0$. ∎

4. Superconvergence property of the PPR-recovered gradient. The following theorem establishes the superconvergence property for the PPR-recovered gradient.

THEOREM 4.1. *Let $\mathcal{T}_h$ be a triangulation of Ω that satisfies the condition (α, σ) and the assumptions in Theorem* 3.3. *If $u \in W^3_\infty(\Omega)$, then*

$$\|\nabla u - \mathrm{G^h} u_h\|_{L^2(\Omega)} \le C h^{1+\rho} \|u\|_{3,\infty,\Omega},$$

where $\rho = \min(\alpha, \frac{1}{2}, \frac{\sigma}{2})$.

Proof. Since

$$\nabla u - \mathrm{G^h} u_h = (\nabla u - \mathrm{G^h}(I_h u)) + (\mathrm{G^h}(I_h u - u_h)), \tag{4.1}$$

estimating $(\nabla u - \mathrm{G^h}(I_h u))$ and $\mathrm{G^h}(I_h u - u_h)$ establishes the proof. To estimate $(\nabla u - \mathrm{G^h}(I_h u))$, recall that $\mathrm{G^h}$ preserves polynomials in $P_2(\Omega)$. Hence, and as was shown in [13],

$$\|\nabla u - \mathrm{G^h}(I_h u)\|_{L^\infty(\Omega)} \le C h^2 |u|_{3,\infty,\Omega}. \tag{4.2}$$

Therefore,

$$\|\nabla u - \mathrm{G^h}(I_h u)\|_{L^2(\Omega)} \le C h^2 \sqrt{|\Omega|} \|u\|_{3,\infty,\Omega}. \tag{4.3}$$

To estimate $\mathrm{G^h}(I_h u - u_h)$, Theorems 3.2 and 3.3 imply the boundedness of $\mathrm{G^h}$. Thus,

$$\|\mathrm{G^h}(I_h u - u_h)\|_{L^2(T)} < C_1 \|\nabla(I_h u - u_h)\|_{L^2(\omega_T)} \text{ for all } T \in \mathcal{T}_h. \tag{4.4}$$

Consequently,

$$\begin{aligned} \|\mathrm{G^h}(I_h u - u_h)\|^2_{L^2(\Omega)} &= \sum_{T \in \mathcal{T}_h} \|\mathrm{G^h}(I_h u - u_h)\|^2_{L^2(T)} \\ &\le \sum_{T \in \mathcal{T}_h} C_1^2 \|\nabla(I_h u - u_h)\|^2_{L^2(\omega_T)} \\ &\le C \|\nabla(I_h u - u_h)\|^2_{L^2(\Omega)}. \end{aligned} \tag{4.5}$$

Since $\mathcal{T}_h$ satisfies the condition (α, σ),

$$\|\nabla(I_h u - u_h)\|_{L^2(\Omega)} \le C_2 h^{1+\rho} \|u\|_{3,\infty,\Omega}, \tag{4.6}$$

where $\rho = \min(\alpha, \frac{1}{2}, \frac{\sigma}{2})$. Using (4.6) in (4.5), we have

$$\|\mathrm{G^h}(I_h u - u_h)\|_{L^2(\Omega)} \le C h^{1+\rho} \|u\|_{3,\infty,\Omega}. \tag{4.7}$$

Using (4.3) and (4.7) in (4.1) completes the proof. ∎

Remark 4.2. The conclusion of Theorem 4.1 is true if the condition (α, σ) is replaced by any other condition that guarantees superconvergence in $\nabla(I_h u - u_h)$. But, the condition (α, σ) covers a wide range of meshes used in practice.

Consider the global a posteriori error estimator η_h defined by

$$\eta_h = \|\mathrm{G^h} u_h - \nabla u_h\|_{L^2(\Omega)}.$$

COROLLARY 4.3. *If, in addition to the assumptions in Theorem* 4.1,

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \ge c(u) h, \tag{4.8}$$

then

$$\left| \frac{\eta_h}{\|\nabla(u - u_h)\|_{L^2(\Omega)}} - 1 \right| \le C h^\rho.$$

Proof. By Theorem 4.1, and the assumption in (4.8), we have

$$\left| \frac{\eta_h}{\|\nabla(u - u_h)\|_{L^2(\Omega)}} - 1 \right| \le \frac{\|\mathrm{G^h} u_h - \nabla u_h\|_{L^2(\Omega)}}{\|\nabla(u - u_h)\|_{L^2(\Omega)}} \le \frac{C h^{1+\rho} \|u\|_{3,\infty,\Omega}}{c(u) h} = C h^\rho. \qquad ∎$$

5. Numerical results. In this section we will go over some numerical examples that demonstrate the superconvergence property of G^h and the asymptotic exactness of the G^h-based a posteriori error estimator. As it is known, the SPR is one of the best gradient recovery techniques. Moreover, the computer-based theory, developed by Babuška and Strouboulis [3], showed that the SPR-based a posteriori error estimator is the most robust one. Hence, quality of the PPR can be measured using the SPR as a reference.

As before, the gradient recovery operator associated with either the SPR or the PPR is denoted by R_h. The examples considered in this section are based on the model problem

$$-\triangle u = f \text{ in } \Omega \text{ and } u = g \text{ on } \partial\Omega.$$

In general, the quality of $R_h u_h$ deteriorates near $\partial\Omega$. Therefore, we should study separately the behavior of $R_h u_h$ inside Ω and near $\partial\Omega$. To distinguish between the regions inside Ω and the ones adjacent to $\partial\Omega$, $\mathcal{N}_h$ is partitioned into $\mathcal{N}_{h,1} \cup \mathcal{N}_{h,2}$, where

$$\mathcal{N}_{h,1} = \{z \in \mathcal{N}_h : \text{dist}(z, \partial\Omega) \geq H\}$$

for some $H > 0$. Next, $\overline{\Omega}$ is partitioned into $\Omega_1 \cup \Omega_2$, where

$$\Omega_1 = \bigcup \{T \in \mathcal{T}_h : T \text{ has all of its vertices in } \mathcal{N}_{h,1}\}.$$

Let $\mathcal{A} \subseteq \overline{\Omega}$ be the union of a set of mesh triangles in $\mathcal{T}_h$. The R_h-based a posteriori error estimator in $\mathcal{A}$ is

$$\eta_{h,\mathcal{A}} = \|R_h u_h - \nabla u_h\|_{L^2(\mathcal{A})}.$$

To measure the accuracy of $\eta_{h,\mathcal{A}}$, we use the effectivity index $\theta_{h,\mathcal{A}}$ defined by

$$\theta_{h,\mathcal{A}} = \frac{\eta_{h,\mathcal{A}}}{\|\nabla(u - u_h)\|_{L^2(\mathcal{A})}}.$$

It is customary to use colorful pictures to trace the accuracy of the a posteriori error estimator in each of the mesh triangles in $\mathcal{A}$. Instead, we will trace the mean, $\mu_{h,\mathcal{A}}$, and the standard deviation, $\sigma_{h,\mathcal{A}}$, of the effectivity indices in these triangles. If the estimator is asymptotically exact in each of the triangles in $\mathcal{A}$, then $\mu_{h,\mathcal{A}} \to 1$ and $\sigma_{h,\mathcal{A}} \to 0$ as $h \to 0$. Note that

$$\mu_{h,\mathcal{A}} = \frac{1}{N_{h,\mathcal{A}}} \sum_{T \subset \mathcal{A}} \theta_{h,T}$$

and

$$\sigma^2_{h,\mathcal{A}} = \frac{1}{N_{h,\mathcal{A}}} \sum_{T \subset \mathcal{A}} (\theta_{h,T} - \mu_{h,\mathcal{A}})^2,$$

where $N_{h,\mathcal{A}}$ is the number of mesh triangles in $\mathcal{A}$.

Example 1. In this example $\Omega = (0,1)^2$, the solution is $u = \sin(\pi x)\sin(\pi y)$, and H is $1/8$. For mesh generation we consider two cases.

In the first case, we start with an initial mesh generated by the Delaunay triangulation with $h = 0.1$ as shown in Figure 6(a). From this figure, it is clear that

(a) Initial Mesh

(b) Properties of $\mathrm{R}_h u_h$

(c) Properties of θ_h

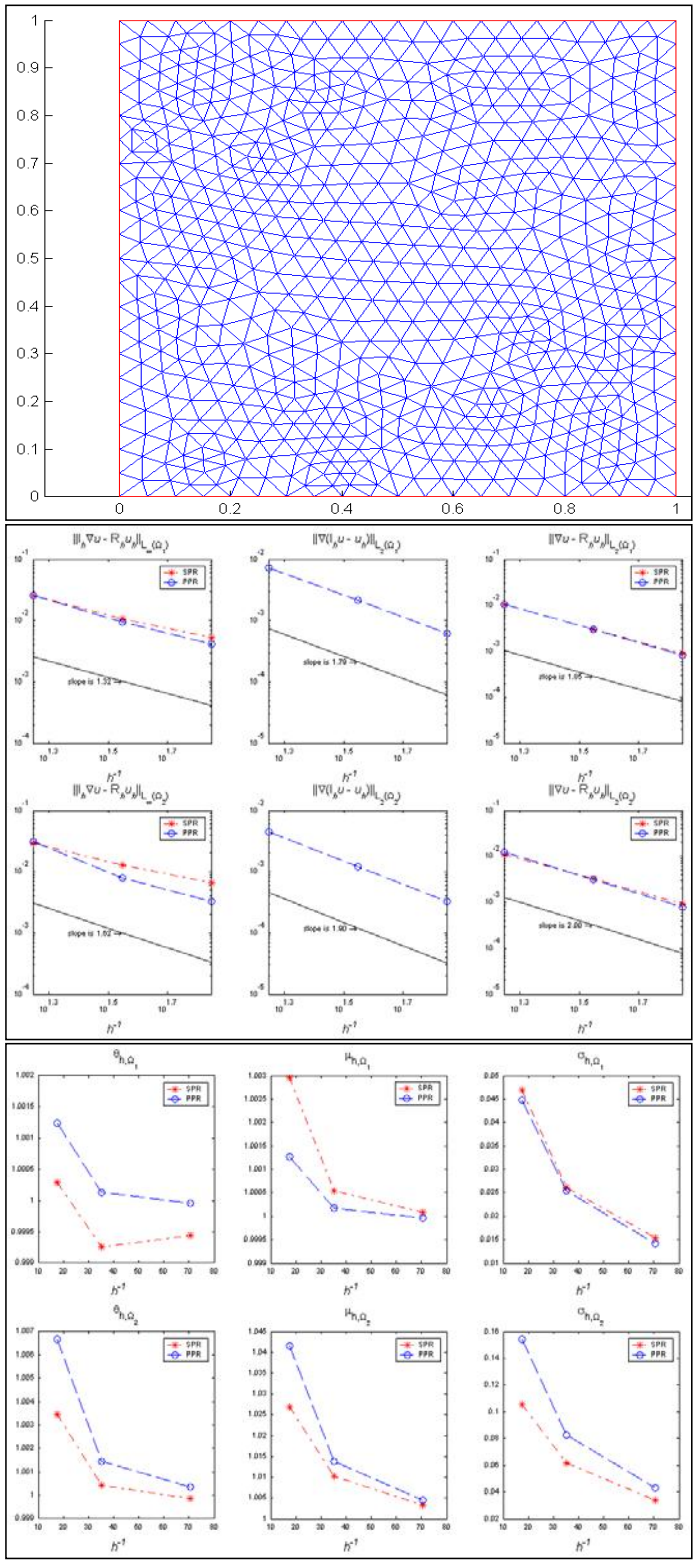


FIG. 6. *Example* 1 *(Delaunay triangulation).*

the Delaunay-generated mesh satisfies the condition (α, σ) with α close to 1 and σ relatively large.
Moreover, this mesh satisfies the conditions in Theorem 3.3. Hence, G^h is bounded, Theorem 1.3 is applicable, and the PPR-recovered gradient enjoys superconvergence. In successive iterations, the new mesh is obtained from the old one by regular refinement. The results are shown in Figure 6(b) and 6(c), where we can note two things. First, although the PPR and the SPR have almost the same global behavior in Ω_1, the statistics show that the PPR is slightly better when we consider the local behavior. Second, the global and local properties of the PPR are much better when it comes to Ω_2.

In the second case, the successive meshes are obtained by decomposing the unit square into $N \times N$ equal squares and then dividing every square into two triangles such that the mesh triangles are arranged in the Chevron pattern. This is done for $N = 16, 32$, and 64. The mesh for $N = 16$ is shown in Figure 7(a). Before we go over the results for this case, note that G^h is bounded and that Theorem 1.3 is not applicable as any pair of triangles sharing a vertical edge form a bigger triangle. As shown in Figure 7(b), we can see that $\nabla(I_h u - u_h)$ has superconvergence that enables G^h to produce superconvergent recovered gradient as mentioned in Remark 4.2. This is not the case with the SPR as it does not preserve polynomials of order 2. Consequently, the behavior of the a posteriori error estimator based on the SPR is inferior to that based on the PPR as shown in Figure 7(c). We can see that the error estimator based on the SPR is underestimating the actual error in Ω_1 and is overestimating it in Ω_2. However, the PPR error estimator is asymptotically exact in both Ω_1 and Ω_2. Moreover, the statistics in Figure 7(c) show that the PPR error estimator is asymptotically exact in each of the mesh triangles.

Example 2. In this example

$$\Omega = (-1, 1)^2 \setminus [1/2, 1)^2.$$

Using a polar coordinate system at (1/2,1/2), the solution is

$$u = r^{\frac{1}{3}} \sin\left(\frac{2\theta - \pi}{3}\right),$$

where

$$\pi/2 \leq \theta \leq 2\pi.$$

As before, H is 1/8, and we start with an initial mesh generated with the Delaunay triangulation at $h = 0.2$. Since we have singularity at the re-entrant corner (1/2,1/2), we have to refine the triangles near this point in the initial mesh so that the pollution effect is minimized. So, after getting the Delaunay triangulation we use regular refinement for the triangles that are within 0.1 from (1/2,1/2). This will serve as our initial mesh, which is shown in Figure 8(a). In the successive iterations, the mesh is regularly refined. The numerical results for this example are shown in Figure 8(b) and Figure 8(c). We should note that the mesh in this example, even after the refinement near the re-entrant corner, is not as good as the one in Figure 7(a). We can see many pairs of triangles that do not form "good" quadrilaterals, i.e., σ is relatively small. Also, because of the singularity at the re-entrant corner (1/2,1/2), we expect both the PPR and the SPR to behave badly near this point. Of course, this affects the convergence rates for the recovered gradients, especially in Ω_1, but still the PPR yields some what better results, even on individual mesh triangles.

(a) Initial Mesh

(b) Properties of $\mathrm{R}_h u_h$

(c) Properties of θ_h

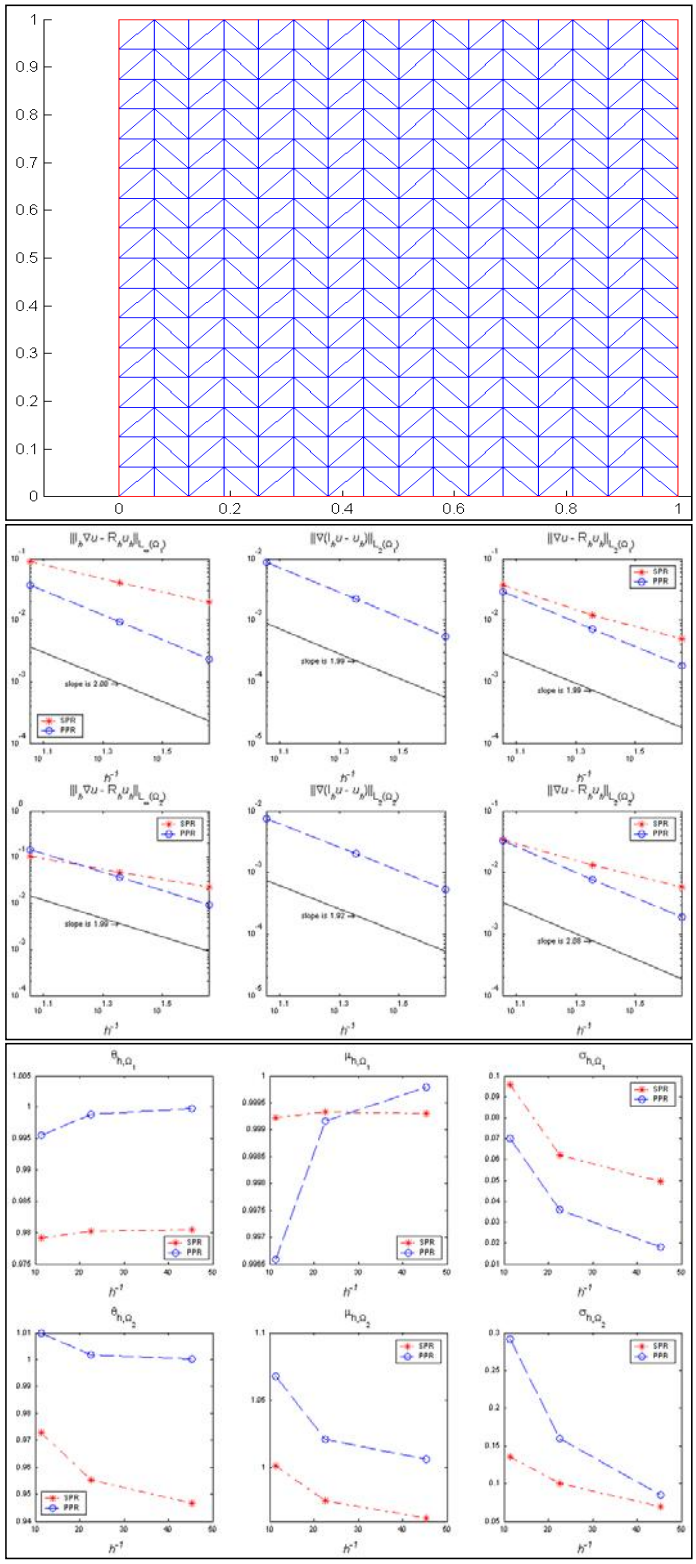


FIG. 7. *Example* 1 *(Chevron mesh).*

(a) Initial Mesh

(b) Properties of $\mathrm{R}_h u_h$

(c) Properties of θ_h

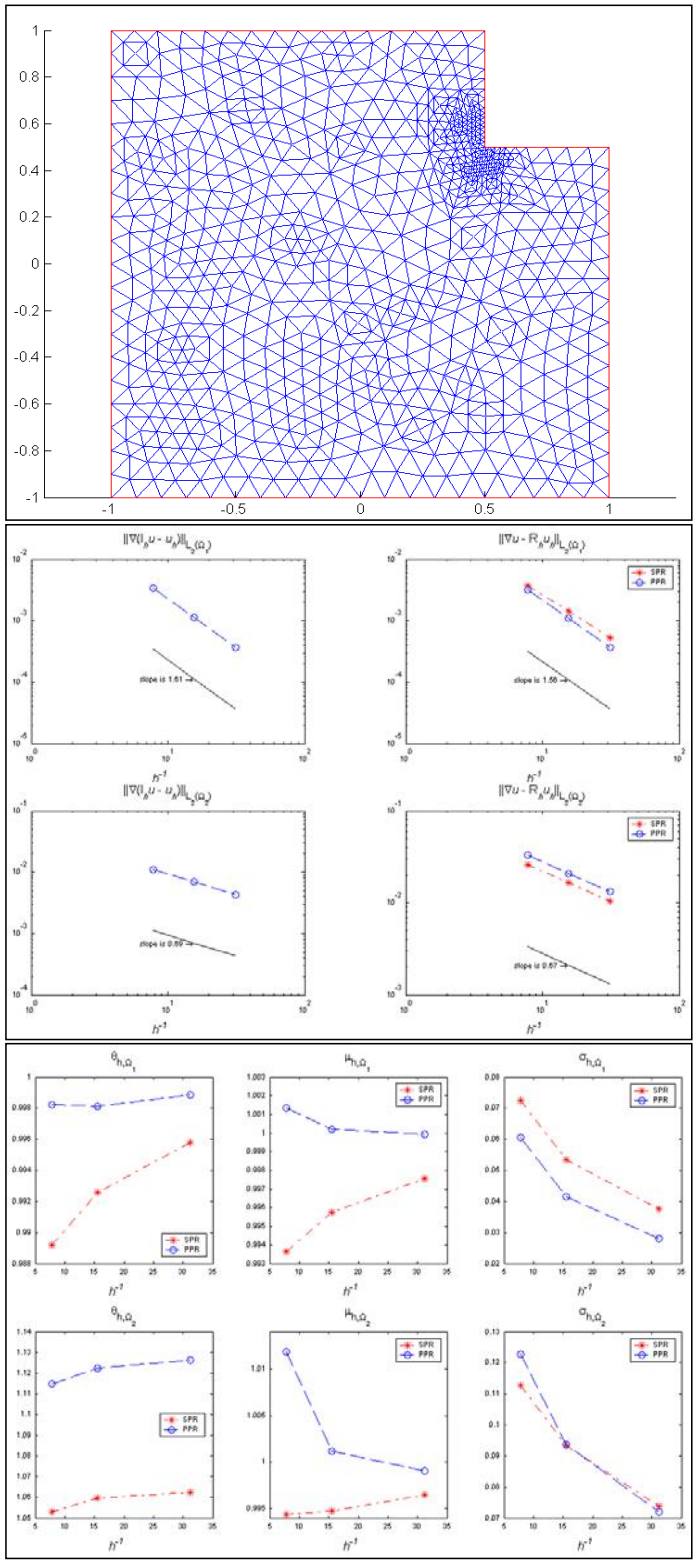


FIG. 8. *Example* 2.

In conclusion, under mild conditions, we have shown that G^h can detect any superconvergence in $\nabla(I_h u - u_h)$ and reflects it in the recovered gradient. Consequently, the PPR error estimator is asymptotically exact, at least globally. The numerical examples indicate that the PPR is, at least, as good as the SPR both inside Ω and near the $\partial\Omega$.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley-Interscience, New York, 2000.

[2] I. Babuška and W. C. Rheinboldt, *A posteriori error estimates for the finite element method*, Internat. J. Numer. Methods Engrg., 12 (1978), pp. 1597–1615.

[3] I. Babuška and T. Strouboulis, *The Finite Element Method and Its Reliability*, Oxford University Press, New York, 2001.

[4] R. E. Bank, *Hierarchical bases and the finite element method*, in Acta Numerica 1996, Acta Numer. 5, Cambridge University Press, Cambridge, UK, 1996, pp. 1–43.

[5] R. E. Bank and A. Weiser, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[6] R. E. Bank and J. Xu, *Asymptotically exact a posteriori error estimators, Part* I: *Grids with superconvergence*, SIAM J. Numer. Analy., 41 (2003), pp. 2294–2312.

[7] H. Blum, Q. Lin, and R. Rannacher, *Asymptotic error expansion and Richardson extrapolation for linear finite elements*, Numer. Math., 49 (1986), pp. 11–37.

[8] M. Křížek, P. Neittaanmäki, and R. Stenberg, eds., *Finite Element Methods: Superconvergence, Postprocessing, and A Posteriori Error Estimates*, Lecture Notes in Pure and Appl. Math. 196, Marcel Dekker, New York, 1997.

[9] X. D. Li and N.-E. Wiberg, *A posteriori error estimate by element patch postprocessing, adaptive analysis in energy norm and L_2 norms*, Comput. & Structures, 53 (1994), pp. 907–919.

[10] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley/Teubner, Stuttgart, 1996.

[11] N.-E. Wiberg and X. D. Li, *Superconvergent patch recovery of finite element solution and a posteriori error L_2 norm estimate*, Comm. Numer. Methods Engrg., 10 (1994), pp. 313–320.

[12] J. Xu and Z. Zhang, *Analysis of recovery type a posteriori error estimators for mildly structured grids*, Math. Comp., 73 (2004), pp. 1139–1152.

[13] Z. Zhang and A. Naga, *A new finite element gradient recovery method: Superconvergence property*, SIAM J. Sci. Comput., to appear.

[14] O. C. Zienkiewicz and J. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, Internat. J. Numer. Methods Engrg., 24 (1987), pp. 337–357.

[15] O. C. Zienkiewicz and J. Zhu, *The superconvergence patch recovery and a posteriori error estimates, Part* I: *The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.

[16] O. C. Zienkiewicz and J. Zhu, *The superconvergence patch recovery and a posteriori error estimates, Part* II: *Error estimates and adaptivity*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1365–1382.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1801–1817

CONVERGENCE OF THE DISCONTINUOUS GALERKIN METHOD FOR DISCONTINUOUS SOLUTIONS*

NOEL J. WALKINGTON†

Abstract. We consider linear first order scalar equations of the form $\rho_t + \mathrm{div}(\rho v) + a\rho = f$ with appropriate initial and boundary conditions. It is shown that approximate solutions computed using the discontinuous Galerkin method will converge in $L^2[0,T;L^2(\Omega)]$ when the coefficients v and a and data f satisfy the minimal assumptions required to establish existence and uniqueness of solutions. In particular, v need not be Lipschitz, so characteristics of the equation may not be defined, and the solutions being approximated my not have bounded variation.

Key words. convergence, discontinuous Galerkin method, hyperbolic equations

AMS subject classifications. 65M60, 65M12

DOI. 10.1137/S0036142902412233

1. Introduction.

1.1. Overview. In 1989 DiPerna and Lions [5] proved that weak solutions $\rho : [0,T) \times \Omega \to \mathbb{R}$ of the convection equation

$$\rho_t + \mathrm{div}(\rho v) + a\rho = f \ \text{ in } \Omega, \qquad \rho|_{t=0} = \rho_0, \tag{1.1}$$

are unique when the velocity field v is in certain Sobolev spaces; in particular, v need not be continuous. While weak solutions exist under much weaker hypotheses, the uniqueness result is subtle. They also proved the following remarkable stability result (specialized here to $L^2(\Omega)$).

THEOREM 1.1 (see DiPerna and Lions [5]). *Let $\{\rho_k\}_{k=0}^{\infty} \subset L^\infty[0,T;L^2(\Omega)]$ be weak solutions of*

$$\rho_{kt} + \mathrm{div}(\rho v_k) + a_k\rho_k = f_k \quad \text{in } \Omega$$

with initial data $\rho_k(0) = \rho_{k0}$. Assume that

- *$\{v_k\} \subset L^1[0,T;H_0^1(\Omega)]$, $v_k \to v$ in $L^1[0,T;L^2(\Omega)]$, and $\mathrm{div}(v_k) \to \mathrm{div}(v)$ in $L^1[0,T;L^\infty(\Omega)]$ with $v \in L^1[0,T;H_0^1(\Omega)]$, and*
- *$a_k \to a$ in $L^1[0,T;L^\infty(\Omega)]$, $f_k \to f$ in $L^1[0,T;L^2(\Omega)]$, and $\rho_{k0} \to \rho_0$ in $L^2(\Omega)$.*

Then $\rho_k \to \rho$ in $L^2[0,T;L^2(\Omega)]$, where ρ is the unique solution of (1.1).

In this paper we prove an analogue of this theorem for approximate solutions computed using the discontinuous method.

While the original motivation of DiPerna and Lions concerned the Bolzman equation, their results provided the tools needed to establish existence of solutions for the incompressible Navier–Stokes equations with variable density and viscosity. In modeling the flow of incompressible immiscible fluids the density is discontinuous, and both

*Received by the editors July 29, 2002; accepted for publication (in revised form) February 26, 2004; published electronically January 20, 2005. This work was supported in part by National Science Foundation grants DMS-9973285, CCR-9902091, and ITR-0086093. This work was also supported by the NSF through the Center for Nonlinear Analysis.
http://www.siam.org/journals/sinum/42-5/41223.html
†Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213 (noelw@cmu.edu).

the density and viscosity, $\mu = \mu(\rho)$, appear as coefficients multiplying the principle terms of the momentum equation. To obtain existence of the coupled system the strong convergence guaranteed by Theorem 1.1 is used in an essential fashion [14].

Similarly, to prove convergence of numerical approximations to the equations modeling incompressible, immiscible fluids, or flows containing particles, strong convergence of the approximate densities ρ_h will be required when the velocity fields are also computed approximately, and this is what is addressed here. Below we show that approximate solutions of the density equation converge strongly in $L^2[0,T;L^2(\Omega)]$ when the coefficients and data satisfy the minimal hypotheses required to obtain existence and uniqueness of the continuous problem.

While many numerical schemes have been proposed for the solution of first order hyperbolic equations, the discontinuous Galerkin method stands out as one of the best schemes in practice. In his introductory text [7] Johnson states "the discontinuous Galerkin method performs remarkably well and we know of no (linear) finite difference method that is better." Another advantage is that discontinuous Galerkin approximations of the density equation couple correctly with natural approximations of the momentum equation so that the discrete system inherits estimates on the kinetic energy $\rho(|v|^2/2)$. To recover such energy estimates it is natural to multiply the density equation by $|v|^2/2$. Since stable approximations of the momentum equation typically approximate v with piecewise quadratic functions, this suggests that piecewise quartic functions are a natural choice of discrete spaces for the discontinuous Galerkin approximation of the balance of mass. Our results below show that the discontinuous Galerkin scheme will converge when piecewise polynomial approximations of arbitrary degree in the spatial variables are used; however, for technical reasons the degree of the piecewise polynomial temporal variation is restricted to be zero or one.

1.2. Discontinuous Galerkin method. The discontinuous Galerkin method was introduced to simulate neutron transport, and in this context the coefficients v and a are constant. Most of the analysis of this method concerns rates of convergence [8, 11, 12] and requires the solution to be smooth, so is not applicable to problems involving discontinuous solutions. One exception is the work of Lin and Zhou [13], who consider equations in the form $V.\nabla\rho + a\rho = f$, which is slightly more general than the evolution equation considered here (put $V = (1, v)$ and $\nabla = (\partial_t, \nabla_x)$ to recover the evolution form). Lin and Zhou require $V \in W^{1,1}(\Omega)$ and show that if the solution is in $H^{1/2}(\Omega)$, then piecewise constant solutions of the discontinuous Galerkin method will converge to certain weak solutions (the definition of a weak solution in [13] is not standard). Below we exploit the evolution structure of the equation in an essential fashion. This allows us to avoid any smoothness assumptions on v with respect to the time variable, $v \in L^1[0,T;H_0^1(\Omega)]$. Under this assumption ρ will not be of bounded variation and typically will not belong to the fractional Sobolev space $H^{1/2}$.

Since (1.1) is a conservation law in divergence form, it is natural to consider the substantial literature concerning numerical schemes developed for nonlinear conservation laws of the form $\rho_t + \operatorname{div}(F(\rho)) = f$. Essentially all the methods (including the discontinuous Galerkin method) and theory developed in this context assume that F is independent of x and t, the exception being Kruzkov's original work [10], which allows F to depend on x and t in a C^1 fashion. In this context solutions of the conservation laws are regular in the sense that they have bounded variation and frequently stability of numerical schemes is ensured by flux limiters, which limit the variation [3, 4, 6, 9]. The best regularity one can expect for the velocity field of immiscible fluids is $v \in L^2[0,T;H_0^1(\Omega)]$ and the variation of the corresponding density will fail

to be bounded, so any scheme that limits the variation of the discrete solution could not converge strongly.

The level set method has also been used to solve certain hyperbolic equations [15]. If $a = f = \operatorname{div}(v) = 0$, then (1.1) can be written as $\rho_t/|\nabla\rho| = -v.(\nabla\rho/|\nabla\rho|)$, showing that the level sets of ρ move with normal velocity $-v$. Frequently it is possible to find a smooth function ϕ_0 having the same level sets as ρ_0 with $\rho_0 = \beta(\phi_0)$, where $\beta : \mathbb{R} \to \mathbb{R}$ is a discontinuous function. Since $\rho = \beta(\phi)$ for all subsequent times, this eliminates the need to deal with discontinuous functions. This approach has been applied to problems in fluid mechanics [2].

1.3. Notation. In this paper, $\Omega \subset \mathbb{R}^d$ is a bounded domain with unit outward normal n. We consider a regular family of finite element meshes $\{\mathcal{T}_h\}_{h>0}$ each of which we assume triangulates Ω exactly. It is assumed that the finite elements have uniformly bounded aspect ratio. The parameter $h > 0$ represents the diameter of the largest element in $\mathcal{T}_h$ and $|K|$ denotes the area (two dimensions) or volume (three dimensions) of an element $K \in \mathcal{T}_h$. Similarly, if $e \subset K$ is an edge or face, then $|e|$ denotes the length or area of e, respectively. The discontinuous Galerkin method is constructed using space-time elements of the form $K \times (t^{m-1}, t^m)$ with $K \in \mathcal{T}_h$, where $\{t^m\}_{n=0}^M$ is a partition of $[0, T]$. The space of polynomials of degree k on an element K is denoted $\mathcal{P}_k(K)$. For simplicity we assume that for each $h > 0$ a uniform partition of $[0, T]$ used with $t^m = m\tau$ where $\tau = T/M$ is assumed to converge to zero as h tends to zero. We denote the approximate solutions by ρ_h; in particular, the dependence on τ is implicit. If $a \in \mathbb{R}$, the positive and negative parts are denoted by $a^{\pm}$ with $a^+ = \max(a, 0)$ and $a^- = \min(a, 0)$.

The traces of functions ρ_h play an important role in the discontinuous Galerkin method and give rise to a lot of notation. In general, ρ_h denotes a discontinuous piecewise polynomial approximate solution of the convection equation, and annotations of the form ρ^m, ρ_-, etc., refer to various traces of ρ_h (i.e., the subscript h is omitted). We write $\rho^m = \rho_h(t^m_-) = \lim_{t \nearrow t^m} \rho_h(t)$, and the trace from above is denoted by $\rho^m_+ = \lim_{t \searrow t^m} \rho_h(t)$. The jump in ρ_h at t^m is denoted by $[\rho^m] = \rho^m_+ - \rho^m$. Integrals of the form $\int_{\partial K}(\rho_h \ldots)$ compute the trace of ρ_h from within K: $\rho_h|_{\partial K}(t, x) = \lim_{\epsilon \searrow 0} \rho_h(t, x - \epsilon n)$, where n is the unit outward normal to ∂K. An orientation of each edge, e, (face in three dimensions) between two finite elements is selected by arbitrarily selecting one of its normals, which is denoted by N (see Figure 2.1). We write $e = K_+ \cap K_-$, where N points from K_- to K_+, and write $[\rho_e] = \rho_+ - \rho_-$, where $\rho_\pm$ are the traces taken from within $K_\pm$.

Standard notation is adopted for the Lebesgue spaces, $L^p(\Omega)$, and the Sobolev spaces, $W^{m,p}(\Omega)$ or $H^m(\Omega)$. The dual exponent to p is denoted by p', $1/p + 1/p' = 1$. Solutions of the evolution equation will be functions from $[0, T]$ into these spaces, and we adopt the usual notion, $L^2[0, T; H^1(\Omega)]$, $C[0, T; H^1(\Omega)]$, etc., to indicate the temporal regularity of such functions. The space of C^∞ test functions having compact support in Ω is denoted by $\mathcal{D}(\Omega)$. For vector valued quantities, such as the velocity v, we write $v \in L^2(\Omega)$, to indicate that each component lies in the specified space. The space $H(\operatorname{div}; \Omega)$ is the set of vector valued functions in $L^2[0, T; L^2(\Omega)]$ with divergence in $L^2[0, T; L^2(\Omega)]$. Strong convergence of a sequence is indicated as $\rho_h \to \rho$ and weak convergence by $\rho_h \rightharpoonup \rho$.

2. Background. In this section we recall the essential results developed by DiPerna and Lions [5] for (1.1) and recall the discontinuous Galerkin method for approximating solutions of (1.1). Our proof of convergence is essentially a verification of the old adage "a stable consistent scheme is convergent." To make this rigorous

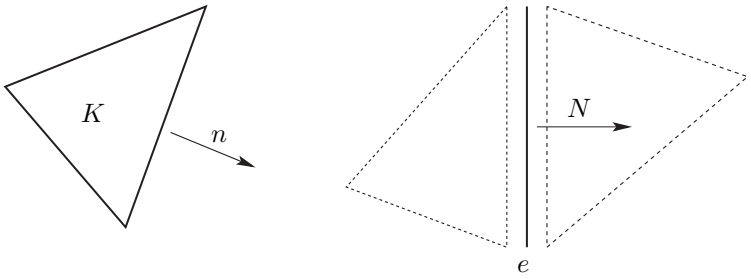


FIG. 2.1. (a) *The outward normal vector of a triangle (tetrahedra in three dimensions) is denoted by n.* (b) *A normal to each edge (face in three dimensions) is arbitrarily chosen and denoted by N.*

we use Theorem 2.2, taken from [14], in a crucial fashion. This theorem states that certain elementary relations that hold for classical solutions of (1.1) continue to hold for weak solutions. As stated above, our convergence results require the approximate solutions to be either piecewise constant or piecewise linear in time. This can be directly attributed to a lack of stability; in general, there are no estimates for the time derivative of the discrete solution. Lacking bounds on the time derivative we can't show that natural piecewise constant approximations converge weakly to the same limit as higher degree piecewise polynomial approximations; see Corollary 3.2.

2.1. DiPerna–Lions theory. For technical reasons DiPerna and Lions [5] considered velocity fields v which vanished on the boundary, and for this reason we always require $v(t) \in H_0^1(\Omega)$. In this situation no boundary conditions are required for ρ; otherwise, ρ would be specified on the inflow boundary, where $v.n < 0$. The following definition of a weak solution of (1.1) is standard and allows us to admit the possibility of discontinuous solutions.

DEFINITION 2.1. *Let $v|_{\partial\Omega} = 0$; then $\rho : [0,T) \times \Omega \to \mathbb{R}$ is a weak solution of* (1.1) *if*

$$-\int_0^T \int_\Omega \rho(\psi_t + v.\nabla\psi - a\psi) = \int_\Omega \rho_0\psi(0) + \int_0^T \int_\Omega f\psi \tag{2.1}$$

for all $\psi \in \mathcal{D}([0,T) \times \Omega)$.

If $\beta : \mathbb{R} \to \mathbb{R}$ is differentiable, then multiplying (1.1) by $\beta'(\rho)$ and formally rearranging the derivatives shows that $\beta(\rho)$ satisfies

$$\beta(\rho)_t + \operatorname{div}(\beta(\rho)v) + \big(\rho\beta'(\rho) - \beta(\rho)\big)\operatorname{div}(v) + a\rho\beta'(\rho) = f\beta'(\rho). \tag{2.2}$$

The following theorem by DiPerna and Lions [5] states that weak solutions of (1.1) will also be weak solutions of (2.2) provided each term is integrable.

THEOREM 2.2. *Let $1 \le p \le \infty$ and suppose that*

$$v \in L^1[0,T,W_0^{1,p'}(\Omega)], \qquad a, \operatorname{div}(v) \in L^1[0,T;L^\infty(\Omega)], \qquad f \in L^1[0,T;L^p(\Omega)].$$

Then for each $\rho_0 \in L^p(\Omega)$ there exists a unique weak solution $\rho \in L^\infty[0,T,L^p(\Omega)]$ of (1.1). *This solution satisfies as follows:*

- *$\rho \in C[0,T;L^p(\Omega)]$ if $p < \infty$.*

- *If $\beta \in C^1(\mathbb{R})$ satisfies $\beta'(t) \le C(1+|t|^r)$ for $C > 0$, and $r = p-1$ if $p < d/(d-1)$, $r < p-1$ if $p = d/(d-1)$, and $r = p/d$ if $p > d/(d-1)$ (r arbitrary if $p = \infty$), then (2.2) is satisfied weakly.*
- *If $\beta \in C^1(\mathbb{R})$ satisfies $\beta'(t) \le C(1+|t|^r)$ for $C > 0$ and $r \le p-1$ (β arbitrary if $p = \infty$), then*

$$\frac{d}{dt}\int_\Omega \beta(\rho) + \int_\Omega \operatorname{div}(v)(\rho\beta'(\rho) - \beta(\rho)) + a\rho\beta'(\rho) = \int_\Omega f\beta'(\rho). \tag{2.3}$$

Remark. The restrictions on r in the second statement of the theorem and the Sobolev embedding theorem guarantee that the term $\beta(\rho)v.\nabla\psi$ is integrable. Similarly, the restriction $r \le p-1$ in the third statement guarantees that each term is integrable.

2.2. Discontinuous Galerkin method. We allow for the possibility that the coefficients are computed only approximately on each mesh; $v \simeq v_h$ and $a \simeq a_h$. Since $\operatorname{div}(v_h)$ may not be bounded in $L^1[0,T;L^\infty(\Omega)]$, care is required to construct a stable approximation scheme. Let

$$\begin{aligned}\mathcal{R}_h = \{&\rho_h \in L^2[0,T;L^2(\Omega)] \mid \rho_h|_{K\times(t^{m-1},t^m)} \in \mathcal{P}_k(K)\otimes\mathcal{P}_\ell(t^{m-1},t^m),\\ &K \in \mathcal{T}_h,\ m = 1,2,\dots\}.\end{aligned}$$

The discontinuous Galerkin method requires $\rho_h \in \mathcal{R}_h$ to satisfy

$$\begin{aligned}&\int_K \left(\rho^m\psi(t_-^m) - \rho^{m-1}\psi(t_+^{m-1})\right)\\ &\quad - \int_{t^{m-1}}^{t^m}\int_K \rho_h\left(\psi_t + v_h.\nabla\psi + (1/2)(\operatorname{div}(v_h) - \operatorname{div}(v))\psi - a_h\psi\right)\\ &\quad + \int_{t^{m-1}}^{t^m}\int_{\partial K} \rho_{in}(v_h.n)\psi = \int_{t^{m-1}}^{t^m}\int_K f\psi\end{aligned} \tag{2.4}$$

for each $K \in \mathcal{T}_h$, $m = 1,2,\dots$, and $\psi \in \mathcal{R}_h$. Since functions in $\mathcal{R}_h$ are discontinuous at the boundary of each space-time element, $K\times(t^{m-1},t^m)$, we specify how the traces are to be evaluated. In all instances traces of ρ_h are taken from the upwind direction, and traces of ψ are taken from within $K\times(t^{m-1},t^m)$. That is, ρ^m and ρ^{m-1} are the traces taken from below, $\rho^m = \lim_{s\nearrow t^m}\rho_h(s)$, ρ_{in} is the inflow trace, $\rho_{in}(x) = \lim_{\epsilon\searrow 0}\rho_h(x-\epsilon v_h)$, and the subscripts $t_\pm$ are used to indicate the traces of ψ at each end of the time interval.

If $(v.n) = (v.n)^+ + (v.n)^-$ is the decomposition of $v.n|_{\partial K}$ into positive and negative parts and $e = K\cap K_-$ is an edge (face in three dimensions) common to K and K_-, then the upwind term can be written as

$$\rho_{in}(v.n) = \rho_h|_{\partial K_-}(v.n)^- + \rho_h|_{\partial K}(v.n)^+.$$

If a global orientation of e is determined by (arbitrarily) selecting one of its normals N (see Figure 2.1), then the weak statement on each element can be summed to give

$$\begin{aligned}&\int_\Omega \rho^m\psi(t_-^m) - \sum_{k=0}^{m-1}\int_\Omega \rho^k[\psi^k] - \int_0^{t^m}\int_\Omega \rho_h\left(\psi_t + v_h.\nabla\psi + (1/2)(\operatorname{div}(v_h) - \operatorname{div}(v))\psi - a_h\psi\right)\\ &\quad - \sum_e \int_0^{t^m}\int_e \left(\rho_-(v_h.N)^+ + \rho_+(v_h.N)^-\right)[\psi_e] = \int_\Omega \rho^0\psi(t_-^0) + \int_0^{t^m}\int_\Omega f\psi.\end{aligned} \tag{2.5}$$

The jump in ψ on the element boundaries is denoted by $[\psi_e] = \psi_+ - \psi_-$ with $\psi_\pm$ determined by the orientation N on an edge and the positive time direction at a temporal interface. We abused the notation by writing $\sum_K \int_K (.) = \int_\Omega (.)$ for integrands involving gradients and similarly for temporal integrals. When ψ is continuous the above expression reduces to a standard weak statement of (1.1).

Remarks. The variant of the discontinuous Galerkin scheme presented here is formulated to be convergent when the velocity field v was known only approximately but the divergence of the v was known precisely. The canonical example of this would be when v_h is an approximation of the velocity of an incompressible fluid for which $\operatorname{div}(v) = 0$. Typically $v_h \to v$ in $L^2[0,T;L^2(\Omega)]$, but $\operatorname{div}(v_h) \neq 0$ since it is difficult to construct divergence free subspaces of $H_0^1(\Omega)$; in particular, $\operatorname{div}(v_h) \not\to 0$ in $L^1[0,T;L^2(\Omega)]$.

The assumptions $v \in L^1[0,T;H_0^1(\Omega)]$ and $\operatorname{div}(v) \in L^1[0,T;L^\infty(\Omega)]$ are required for uniqueness of the solution of (1.1); however, the approximation v_h of v used in the computations need not converge to v in these spaces. For the scheme to be well defined, traces of $v_h.n$ must exist on element boundaries and the traces from each side must agree. For these reasons we require v_h to lie in the space

$$V_h = \{v_h \in L^1[0,T;H(\operatorname{div};\Omega)] \mid v_h(t)|_K \in \mathcal{P}_k(K)\}$$

for some fixed integer $k \geq 0$. Note that uniqueness of the continuous problem requires $v \in L^2[0,T;H_0^1(\Omega)]$, but we do not require $v_h \in L^2[0,T;H_0^1(\Omega)]$.

3. Stability. The following stability result is standard.

THEOREM 3.1 (stability). *Let $\rho_h \in \mathcal{R}_h$ be the approximate solution of* (1.1) *obtained with the discontinuous scheme* (2.4) *and suppose that $\rho^0 \in L^2(\Omega)$, $v_h \in V_h$,*

$$v \in L^1[0,T;H_0^1(\Omega)], \qquad a_h,\, \operatorname{div}(v) \in L^1[0,T;L^\infty(\Omega)], \qquad f \in L^1[0,T;L^2(\Omega)];$$

then

$$\begin{aligned}(1/2)\|\rho^m\|^2_{L^2(\Omega)} + (1/2)\sum_{k=0}^{m-1}\|[\rho^k]\|^2_{L^2(\Omega)} + (1/2)\sum_e \int_0^{t^m}\int_e |v_h.n|[\rho_h]^2 \\ + \int_0^{t^m}\int_\Omega (\operatorname{div}(v)/2 + a_h)\rho_h^2 = (1/2)\|\rho^0\|^2_{L^2(\Omega)} + \int_0^{t^m}\int_\Omega f\rho_h.\end{aligned} \tag{3.1}$$

(1) *If $(\operatorname{div}(v)/2 + a_h) \geq c > 0$ and $f \in L^2[0,T;L^2(\Omega)]$, then*

$$\begin{aligned}\|\rho^m\|^2_{L^2(\Omega)} + \sum_{k=0}^{m-1}\|[\rho^k]\|^2_{L^2(\Omega)} + \sum_e \int_0^{t^m}\int_e |v_h.n|[\rho_h]^2 + c\int_0^{t^m}\|\rho_h(s)\|^2_{L^2(\Omega)}\,ds \\ \leq \|\rho^0\|^2_{L^2(\Omega)} + (1/c)\int_0^{t^m}\|f(s)\|^2_{L^2(\Omega)}\,ds.\end{aligned}$$

(2) *If ρ_h is piecewise constant in time or piecewise linear in time ($\ell = 0$ or 1 in the definition of $\mathcal{R}_h$), and τ is sufficiently small, then*

$$\|\rho^m\|^2_{L^2(\Omega)} + \sum_{k=0}^{m-1}\|[\rho^k]\|^2_{L^2(\Omega)} + \sum_e \int_0^{t^m}\int_e |v_h.n|[\rho_h]^2 \leq C_1\|\rho^0\|^2_{L^2(\Omega)}\exp(C_2 t^m),$$

where C_1 and C_2 depend on the coefficients v and a_h and the data f.

Remark. When $\operatorname{div}(v)$ and a are bounded it is possible to introduce a change of variables to guarantee that $\operatorname{div}(v)/2 + a \geq c > 0$. Specifically, if $\rho = re^{\alpha t}$, then r satisfies

$$r_t + \operatorname{div}(rv) + (\alpha + a)r = e^{-\alpha t} f.$$

Proof. Selecting $\psi = \rho_h$ in (2.4) gives

$$(1/2)\int_K (\rho^m)^2 + [\rho^{m-1}]^2 - (\rho^{m-1})^2 + \int_{t^{m-1}}^{t^m}\int_K (\operatorname{div}(v)/2 + a_h)\rho_h^2 + (1/2)\int_{t^{m-1}}^{t^m}\int_{\partial K} \rho_h^2 (v_h.n)^+ + \rho_-^2 (v_h.n)^- - [\rho_h]^2 (v_h.n)^- = \int_{t^{m-1}}^{t^m}\int_K f\rho_h.$$

Summing this expression and collecting terms establishes (3.1), and if $(\operatorname{div}(v)/2 + a_h) \geq c > 0$, statement (1) follows immediately.

If ρ_h is piecewise constant in time, then

$$\int_0^{t^m}\int_\Omega f\rho_h - (\operatorname{div}(v)/2 + a_h)\rho_h^2 \leq F^m \max_{1\leq k\leq n} \|\rho^k\|_{L^2(\Omega)} + \sum_{k=1}^m \gamma_k \|\rho^k\|^2_{L^2(\Omega)},$$

where

$$\gamma_k = \int_{t^{k-1}}^{t^k} \left((1/2)\|\operatorname{div}(v(s))\|_{L^\infty(\Omega)} + \|a_h(s)\|_{L^\infty(\Omega)}\right) ds, \qquad F^m = \int_0^{t^m} \|f(s)\|_{L^2(\Omega)}\, ds.$$

If ρ_h is piecewise linear in time, then for $s \in (t^{m-1}, t^m)$

$$\begin{aligned}\|\rho_h(s)\|_{L^2(\Omega)} &= \|\rho^m(s - t^{m-1})/\tau + \rho_+^{m-1}(t^m - s)/\tau\|_{L^2(\Omega)} \\ &= \|\rho^m(s - t^{m-1})/\tau + ([\rho^{m-1}] - \rho^{m-1})(t^m - s)/\tau\|_{L^2(\Omega)} \\ &\leq \|\rho^m\|_{L^2(\Omega)}(s - t^{m-1})/\tau + \|[\rho^{m-1}] - \rho^{m-1}\|_{L^2(\Omega)}(t^m - s)/\tau.\end{aligned}$$

Then

$$\|\rho_h\|_{L^\infty[0,t^m;L^2(\Omega)]} \leq \max_{0\leq k\leq m} \|\rho^k\|_{L^2(\Omega)} + \max_{0\leq k\leq m-1} \|[\rho^k]\|_{L^2(\Omega)}$$

and

$$\|\rho_h(s)\|^2_{L^2(\Omega)} \leq \|\rho^m\|^2_{L^2(\Omega)} + 2\|\rho^{m-1}\|^2_{L^2(\Omega)} + 2\|[\rho^{m-1}]\|^2_{L^2(\Omega)}.$$

It follows that

$$\int_0^{t^m}\int_\Omega f\rho_h - (\operatorname{div}(v)/2 + a_h)\rho_h^2 \leq F^m \left(\max_{0\leq k\leq m} \|\rho^k\|_{L^2(\Omega)} + \max_{0\leq k\leq m-1} \|[\rho^k]\|_{L^2(\Omega)}\right)$$
$$+ \left(\gamma_m \|\rho^m\|^2_{L^2(\Omega)} + 2\gamma_1 \|\rho^0\|^2_{L^2(\Omega)}\right) + \sum_{k=1}^{m-1} \left((\gamma_k + 2\gamma_{k+1})\|\rho^k\|^2_{L^2(\Omega)} + 2\gamma_k \|[\rho^{k-1}]\|^2_{L^2(\Omega)}\right).$$

In each instance an estimate of the form

$$\begin{aligned}&\|\rho^m\|^2_{L^2(\Omega)} + (1 - \hat\gamma)\sum_{k=0}^{m-1} \|[\rho^k]\|^2_{L^2(\Omega)} + \sum_e \int_0^{t^m}\int_e |v_h.n|[\rho_e]^2 \\ &\quad\leq (1 + \hat\gamma_0)\|\rho^0\|^2_{L^2(\Omega)} + (C/c)(F^m)^2 + \sum_{k=1}^m \left(\hat\gamma_k \|\rho^k\|^2_{L^2(\Omega)}\right) \\ &\qquad + c\left(\max_{0\leq k\leq m} \|\rho^k\|^2_{L^2(\Omega)} + \max_{0\leq k\leq m-1} \|[\rho^k]\|^2_{L^2(\Omega)}\right)\end{aligned}$$

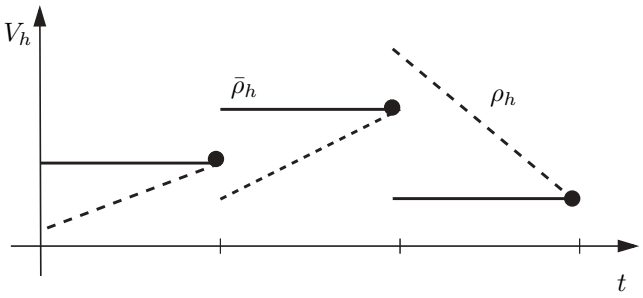


FIG. 3.1. *$\bar\rho_h$ is piecewise constant time and is equal to $\rho_h(t^m_-)$ on $(t^{m-1}, t^m]$.*

holds, where $c > 0$ is arbitrary and $\hat\gamma$ and $\hat\gamma_k$ are bounded by integrals of functions in $L^1[0,T]$ over intervals of length τ. If τ is sufficiently small, all these constants will be less than $1/2$, and statement (2) follows from the discrete Gronwall inequality. ∎

When $\operatorname{div}(v)/2 + a_h = 0$ the stability estimate only bounds ρ_h at the discrete times $\{t^m\}_{n=0}^M$. If $\bar\rho_h(t) = \rho^m$ on $(t^{m-1}, t^m]$ (see Figure 3.1), the lemma shows that $\bar\rho_h$ can be bounded in $L^\infty[0,T;L^2(\Omega)]$. Clearly $\bar\rho_h = \rho_h$ when ρ_h is piecewise constant in time ($\ell = 0$); however, when $\ell > 1$ it may happen that $\bar\rho_h$ and ρ_h have different weak limits. The following corollary shows that this will not happen when $\ell = 1$.

COROLLARY 3.2. *Let $\rho_h \in \mathcal{R}_h$ be piecewise linear in time ($\ell = 1$) and let $\bar\rho_h \in \mathcal{R}_h$ be the function piecewise constant in time equal to $\rho^m = \rho_h(t^m_-)$ on $(t^{m-1}, t^m]$.*

- *If $\psi \in L^2[0,T;L^2(\Omega)]$, then*

$$\left|\int_0^T\int_\Omega (\bar\rho_h - \rho_h)\psi\right| \le \|\bar\rho_h\|_{L^2[0,T;L^2(\Omega)]}\|\psi - \psi(.+\tau)\|_{L^2[0,T-\tau;L^2(\Omega)]}$$
$$+\sqrt{(\tau/2)}\left(\|\rho^M\|_{L^2(\Omega)} + \|\rho^0\|_{L^2(\Omega)} + \left(\sum_{m=0}^{M-1}\|[\rho^m]\|^2_{L^2(\Omega)}\right)^{1/2}\right)\|\psi\|_{L^2[0,T;L^2(\Omega)]}.$$

- $\|\rho_h\|_{L^\infty[0,T;L^2(\Omega)]} \le \|\bar\rho_h\|_{L^\infty[0,T;L^2(\Omega)]} + \max_{0\le m\le M-1}\|[\rho_h^m]\|_{L^2(\Omega)}$ *and*

$$\|\bar\rho_h - \rho_h\|^2_{L^2[0,T;L^2(\Omega)]}$$
$$\le (2/3)\left(\|\bar\rho_h - \bar\rho_h(.+\tau)\|_{L^2[0,T-\tau;L^2(\Omega)]} + \tau\sum_{m=0}^{M-1}\|[\rho^m]\|^2_{L^2(\Omega)}\right).$$

Proof. On the interval $(t^{m-1}, t^m]$ we have $\bar\rho_h - \rho_h = (\rho^m - \rho_+^{m-1})(t^m - t)/\tau$ (recall that $\rho^m = \rho^m_-$). Then

$$\int_0^T\int_\Omega(\bar\rho_h - \rho_h)\psi = \int_\Omega\sum_{m=1}^M\int_0^\tau(\rho^m - \rho_+^{m-1})(1 - s/\tau)\psi(t^{m-1}+s)\,ds$$
$$= \int_\Omega\int_0^\tau\Bigg(\rho^M\psi(t^{M-1}+s) - \rho_+^0\psi(s)$$
$$+\sum_{m=1}^{M-1}(\rho^m\psi(t^{m-1}+s) - \rho_+^m\psi(t^m+s))\Bigg)(1 - s/\tau)\,ds$$

$$= \int_\Omega \int_0^\tau \Bigg(\rho^M \psi(t^{M-1}+s) - \rho^0 \psi(s) + \sum_{m=1}^{M-1} \rho^m (\psi(t^{m-1}+s) - \psi(t^m+s)) + \sum_{m=0}^{M-1} (\rho^m - \rho_+^m)\psi(t^m+s) \Bigg) (1-s/\tau)\, ds.$$

The first statement then follows upon observing that the last term can be bounded by

$$\int_\Omega \int_0^\tau \sum_{m=0}^{M-1} (\rho^m - \rho_+^m)\psi(t^m+s)(1-s\tau) \le \sqrt{(\tau/2)} \left(\sum_{m=0}^{M-1} \|[\rho^m]\|_{L^2(\Omega)}^2 \right)^{1/2} \|\psi\|_{L^2[0,T;L^2(\Omega)]}.$$

To establish the second statement we compute

$$\begin{aligned} \|\bar\rho_h - \rho_h\|_{L^2[0,T;L^2(\Omega)]}^2 &= \sum_{m=1}^M \int_0^\tau \|\rho^m - \rho_+^{m-1}\|^2 (1-s/\tau)^2\, ds \\ &\le \sum_{m=1}^M (\tau/3) \|\rho^m - \rho^{m-1} + [\rho^{m-1}]\|^2 \\ &\le (2/3) \left(\|\bar\rho_h - \bar\rho_h(.+\tau)\|_{L^2[0,T-\tau;L^2(\Omega)]} + \tau \sum_{m=0}^{M-1} \|[\rho^m]\|_{L^2(\Omega)}^2 \right). \end{aligned}$$

∎

4. Consistency. The bounds established above show that subsequences of the approximate solutions converge weakly-star in $L^\infty[0,T;L^2(\Omega)]$. In this section we show that the limits of these sequences are weak solutions of (1.1). This is easy to show when the space $\mathcal{R}_h$ contains the continuous finite element spaces; however, when $k=0$ the only continuous functions are constants, which complicates the proof. We begin with a technical lemma.

LEMMA 4.1. *Let $K \subset \mathbb{R}^d$ be a simplex, $v \in H^1(K)^d$, and $\psi \in W^{1,p}(K)$ with $p \ge 4d/(d+4)$. Then there exists a constant C depending only on d and the aspect ratio of K such that*

$$\int_{\partial K} |v.n||\psi - \bar\psi|^2 \le C|K|^{(1/2-2/p)} h_K \|v\|_{H^1(K)} \|\psi\|_{W^{1,p}(K)}^2,$$

where $\bar\psi = (1/|K|)\int_K \psi$ is the average of ψ on K and h_K is the diameter of K.

If each component of v is a polynomial of degree k, then

$$\int_{\partial K} |v.n||\psi - \bar\psi|^2 \le C|K|^{(1/2-2/p)} h_K \|v\|_{L^2(K)} \|\psi\|_{W^{1,p}(K)}^2,$$

where C may now depend additionally on k.

Proof. Let $\hat K$ be the usual reference simplex and $\chi(\xi) = x_0 + B\xi$ be an affine mapping of $\hat K$ to K. We use a hat to denote the natural correspondence between functions defined on K and $\hat K$, $\hat\psi = \psi \circ \chi$. Writing the integral over the boundary as

the sum over the faces $e \subset \partial K$, the trace theorem with $p \geq 4d/(d+4)$ enables us to write

$$\begin{aligned}\int_{\partial K} |v.n||\psi - \bar{\psi}|^2 &= \sum_{e \subset \partial K} \int_e |v.n||\psi - \bar{\psi}|^2 \\ &= \sum_{\hat{e} \subset \partial \hat{K}} \frac{|e|}{|\hat{e}|} \int_{\hat{e}} |\hat{v}.n||\hat{\psi} - \bar{\psi}|^2 \\ &\leq C \sum_{\hat{e} \subset \partial \hat{K}} |e| \|\hat{v}\|_{H^1(\hat{K})} \|\hat{\psi} - \bar{\psi}\|^2_{W^{1,p}(\hat{K})} \\ &\leq C \sum_{\hat{e} \subset \partial \hat{K}} |e| \|\hat{v}\|_{H^1(\hat{K})} |\hat{\psi}|^2_{W^{1,p}(\hat{K})}.\end{aligned}$$

We used the fact that the average of ψ is the average of $\hat{\psi}$ and the Poincaré inequality to pass to the $W^{1,p}$ seminorm in the last line. Notice that if $v \in \mathcal{P}_k(K)^d$, then $\|\hat{v}\|_{H^1(\hat{K})}$ is equivalent to $\|\hat{v}\|_{L^2(\hat{K})}$ since $\mathcal{P}_k(\hat{K})$ is finite dimensional.

Recalling that $\|\hat{v}\|_{L^2(\hat{K})} = (|\hat{K}|/|K|)^{1/2} \|\hat{v}\|_{L^2(K))}$, $|B| \leq C h_K$,

$$|\hat{\psi}|_{W^{1,p}(\hat{K})} \leq (|\hat{K}|/|K|)^{1/p} |B| |\psi|_{W^{1,p}(K)} \leq C(|\hat{K}|/|K|)^{1/p} h_K |\psi|_{W^{1,p}(K)},$$

and $|\nabla \hat{v}|_{L^2(\hat{K})} \leq C(|\hat{K}|/|K|)^{1/2} h_K \|\nabla v\|_{L^2(K)}$, we deduce that

$$\int_{\partial K} |v.n||\psi - \bar{\psi}|^2 \leq C \sum_{e \subset \partial K} \frac{|e|}{|K|^{(1/2+2/p)}} h_K^2 \|v\|_{H^1(K)} |\psi|^2_{W^{1,p}(K)}$$

(with $\|v\|_{L^2(K)}$ replacing $\|v\|_{H^1(K)}$ if $v \in \mathcal{P}_k(K)^d$). Since

$$|K| = (1/d)|e| \times (\text{perpendicular height}) \geq c|e|h_K,$$

where c depends on the aspect ratio of K, it follows that $|e|h_K \leq C|K|$, and the proof follows. ∎

The following lemma provides sufficient conditions on the coefficients v_h and a_h that suffice to establish consistency of the discontinuous Galerkin method.

LEMMA 4.2. *Let $\{\rho_h\}$ be a (sub-) sequence of solutions of the discontinuous Galerkin scheme* (2.4) *computed on a sequence of quasi-regular meshes and suppose that $\rho_h \rightharpoonup^* \rho$ in $L^\infty[0,T;L^2(\Omega)]$. Assume that $f \in L^1[0,T;L^2(\Omega)]$, $v \in L^1[0,T;H^1_0(\Omega)]$, $v_h(t) \in V_h$, $\rho^0 \rightharpoonup \rho_0$ in $L^2(\Omega)$,*

$$v_h \to v, \quad \text{and} \quad a_h \to a, \ \text{in } L^1[0,T;L^2(\Omega)],$$

and either (1) $\operatorname{div}(v_h) \to \operatorname{div}(v)$ *in* $L^1[0,T;L^2(\Omega)]$ *or* (2) $(\operatorname{div}(v_h) - \operatorname{div}(v))|_K \perp \mathcal{P}_k(K)$ *in* $L^2(K)$ *for each* $K \in \mathcal{T}_h$. *Then ρ is a weak solution of* (1.1).

Proof. When $k, \ell > 0$ $\mathcal{R}_h$ contains the usual continuous finite element spaces, so if $\psi \in \mathcal{D}([0,T) \times \Omega)$, then the classical Lagrange interpolant $\psi_h \in \mathcal{R}_h \cap C([0,T] \times \bar{\Omega})$ converges to ψ in $W^{1,\infty}((0,T) \times \Omega)$. When ψ_h is substituted into (2.5) all the jump

terms vanish to give

$$-\int_0^T\int_\Omega \rho_h\big(\psi_{ht}+v_h.\nabla\psi_h+(1/2)(\operatorname{div}(v_h)-\operatorname{div}(v))\psi_h-a_h\psi_h\big)=\int_\Omega \rho^0\psi(0)+\int_0^T\int_\Omega f\psi_h.$$

If $\operatorname{div}(v_h)\to\operatorname{div}(v)$ in $L^1[0,T;L^2(\Omega)]$ the hypotheses suffice to pass to the limit term by term in the above equation to show that ρ is a weak solution of (1.1). If $(\operatorname{div}(v_h)-\operatorname{div}(v)|_K\perp\mathcal{P}_k(K)$ for $K\in\mathcal{T}_h$ the term involving $\operatorname{div}(v_h)-\operatorname{div}(v)$ still vanishes since

$$\begin{aligned}\int_0^T\int_\Omega \rho_h(\operatorname{div}(v_h)-\operatorname{div}(v))\psi_h &= \int_0^T\int_\Omega \rho_h(\operatorname{div}(v_h)-\operatorname{div}(v))(\psi_h-\bar\psi)\\ &\le \|\rho_h\|_{L^\infty[0,T;L^2(\Omega)]}\|\operatorname{div}(v_h)-\operatorname{div}(v)\|_{L^1[0,T;L^2(\Omega)]}Ch\\ &\to 0,\end{aligned}$$

where $\bar\psi(t,x)$ is the average of $\psi_h(t,.)$ over the element K containing x.

When $\ell=0$ or $k=0$, functions in $\mathcal{R}_h$ are piecewise constant in time or space, respectively. In this situation the terms involving ψ_{ht} or $\nabla\psi_h$ vanish and it is necessary to show that the corresponding jump terms in (2.5) approximate the missing terms: $-\sum_{k=0}^{M-1}\int_\Omega \rho^k[\psi^k]\sim-\int_0^T\int_\Omega\rho\psi_t$ and

$$-\sum_e\int_0^T\int_e\left(\rho_-(v_h.N)^+ +\rho_+(v_h.N)^-\right)[\psi_h]\sim\int_0^T\int_\Omega\rho v.\nabla\psi.$$

If $\ell=0$ and $\psi\in\mathcal{D}([0,T)\times\Omega)$, let ψ^k be a projection of $\psi(t^k)$ onto the (spatial) finite element space $\{\psi_h\in L^2(\Omega)\mid\psi_h|_K\in\mathcal{P}_k(K),\,K\in\mathcal{T}_h\}$, and let $\hat\psi_h$ denote the piecewise linear interpolant of $\{\psi^k\}$ in time. Then $\hat\psi_h$ converges to ψ in $W^{1,\infty}[0,T;L^\infty(\Omega)]$ and temporal jump terms become

$$-\sum_{k=0}^{M-1}\int_\Omega\rho^k[\psi^k]=-\sum_{k=0}^{M-1}\int_\Omega\rho^k(\psi^{k+1}-\psi^k)=-\int_0^T\int_\Omega\rho_h\hat\psi_{ht}\to-\int_0^T\int_\Omega\rho\psi_t$$

as required.

Finally consider the spatial jump terms when $k=0$. Let ψ be piecewise polynomial of degree ℓ in time with values in $\mathcal{D}(\Omega)$. Selecting $\psi_h(t)|_K$ to be the spatial average of $\psi(t)$ over each element the spatial jump terms in (2.5) become

$$\begin{aligned}&-\sum_e\int_0^T\int_e\left(\rho_-(v_h.N)^+ +\rho_+(v_h.N)^-\right)[\psi_h]\\ &=-\sum_e\int_0^T\int_e\left(\rho_-(v_h.N)^+ +\rho_+(v_h.N)^-\right)[\psi_h-\psi]\\ &=-\int_0^T\sum_K\int_{\partial K}\left(\rho_{K-}(v_h.n)^- +\rho_K(v_h.n)^+\right)(\psi_K-\psi),\end{aligned}$$

where ψ_K is the average value of ψ on K, ρ_K is the value of ρ_h on K, and ρ_{K-} is the

value of ρ_h on the upwind element K_-. Then

$$
\begin{aligned}
&-\sum_e \int_0^T \int_e \left(\rho_-(v_h.N)^+ + \rho_+(v_h.N)^-\right)[\psi_h] \\
&= -\int_0^T \sum_K \int_{\partial K} \left((\rho_{K_-} - \rho_K)(v_h.n)^- + \rho_K(v_h.n)\right)(\psi_K - \psi) \\
&= -\int_0^T \sum_K \int_K \operatorname{div}(\rho_K v_h(\psi_K - \psi)) + \int_{\partial K} [\rho_h](v_h.n)^-(\psi_K - \psi) \\
&= -\int_0^T \int_\Omega \left(\rho_h v_h.\nabla\psi + \rho_h \operatorname{div}(v_h)(\psi_h - \psi)\right) + \int_0^T \sum_K \int_{\partial K} [\rho_h](v_h.n)^-(\psi_K - \psi).
\end{aligned}
$$

Clearly the first term converges to $\int\int \rho v.\nabla\psi$; we need to show that the second term vanishes in the limit. Using Lemma 4.1, with $p = 4$, we obtain

$$
\begin{aligned}
&\int_0^T \sum_K \int_{\partial K} [\rho_h](v_h.n)^-(\psi_K - \psi) \\
&\le \left(\int_0^T \sum_K \int_{\partial K} |v_h.n|[\rho_h]^2\right)^{1/2} \left(\int_0^T \sum_K \int_{\partial K} |v_h.n|(\psi_K - \psi)^2\right)^{1/2} \\
&\le \left(\int_0^T 2\sum_e \int_e |v_h.N|[\rho_h]^2\right)^{1/2} \left(\sum_K Ch_K \int_0^T \|v_h\|_{L^2(K)} \|\psi\|^2_{W^{1,4}(K)}\right)^{1/2} \\
&\le \left(\int_0^T \sum_e \int_e |v_h.N|[\rho_h]^2\right)^{1/2} C\sqrt{h}\, \|v_h\|^{1/2}_{L^1[0,T;L^2(\Omega)]} \|\psi\|_{L^\infty[0,T;W^{1,4}(\Omega)]}.
\end{aligned}
$$

Theorem 3.1 shows that the first term is bounded so the expression above vanishes as $h \to 0$. □

Remark. The second hypotheses on the divergence in Lemma 4.2 is useful in the context of finite element approximations of incompressible fluids. If $\tilde{v}_h \in L^1[0,T; H_0^1(\Omega)]$ is a classical finite element approximation of v, then typically $\tilde{v}_h \to v$ in $L^1[0,T;L^2(\Omega)]$, but $\operatorname{div}(\tilde{v}_h)$ will only converge weakly to $\operatorname{div}(v)$ in $L^2[0,T;L^2(\Omega)]$. However, we can construct v_h satisfying the hypothesis of the Lemma as follows. At each time t let $v_h(t)$ be the $L^2(\Omega)$ projection of $\tilde{v}_h(t)$ onto the space

$$
\bar{V}_h(v) = \left\{ v_h \in RT_h^k(\Omega) \;\middle|\; \int_K \operatorname{div}(v_h)p(x) = \int_K \operatorname{div}(v)p(x),\ p \in \mathcal{P}_k(K),\ K \in \mathcal{T}_h \right\}.
$$

Here $RT_h^k(\Omega)$ is the Raviart–Thomas subspace of $H(\operatorname{div};\Omega)$ constructed using piecewise polynomials of degree k on $\mathcal{T}_h$ [1].

By construction $\operatorname{div}(v_h) - \operatorname{div}(v)$ is orthogonal to $\mathcal{P}_k(K)$ for each element K. To see that v_h also converges to v first observe that

$$
(\tilde{v}_h - v_h, w_h)_{L^2(\Omega)} = 0 \qquad \forall\ w_h \in \bar{V}_h \equiv \bar{V}_h(0).
$$

Then

$$
\begin{aligned}
\|\tilde{v}_h - v_h\|^2_{L^2(\Omega)} &= (\tilde{v}_h - v_h, \tilde{v}_h - I_h v + I_h v - v_h)_{L^2(\Omega)} \\
&= (\tilde{v}_h - v_h, \tilde{v}_h - I_h v)_{L^2(\Omega)},
\end{aligned}
$$

where $I_h v$ is the interpolant of v onto the Raviart–Thomas space [1]. The degrees of freedom of I_h are constructed to guarantee that $I_h v - v_h \in \bar{V}_h$. Then

$$\begin{aligned}\|\tilde{v}_h - v_h\|_{L^2(\Omega)} &\le \|\tilde{v}_h - v\|_{L^2(\Omega)} + \|v - I_h v\|_{L^2(\Omega)} \\ &\le \|\tilde{v}_h - v\|_{L^2(\Omega)} + Ch\|v\|_{H_0^1(\Omega)},\end{aligned}$$

so $v_h \to v$ if $\tilde{v}_h \to v$ and v is bounded in $L^1[0,T;H_0^1(\Omega)]$.

5. Convergence. When ρ_h is piecewise constant or linear in time ($\ell = 0$ or 1) the stability estimate guarantees that it is possible to pass to a subsequence for which

$$\bar{\rho}_h \rightharpoonup^* \bar{\rho} \ \text{ and } \ \rho_h \rightharpoonup^* \rho \qquad \text{in } L^\infty[0,T;L^2(\Omega)],$$

and Corollary 3.2 shows that the weak limits coincide, $\rho = \bar{\rho}$. (Recall that $\bar{\rho}_h$ is piecewise constant in time and assumes the values ρ^m in $(t^{m-1}, t^m]$ (see Figure 3.1).) If additionally $v \in L^1[0,T;H_0^1(\Omega)]$, then Theorem 2.2 shows that ρ is unique. In this situation the whole sequence $\{\rho_h\}$ converges weakly, and in the following theorem we establish strong convergence in $L^2[0,T;L^2(\Omega)]$.

THEOREM 5.1. *Let $\{\rho_h\}_{h>0}$ be the sequence of solutions of the discontinuous Galerkin scheme* (2.4) *with either $\ell = 0$ or $\ell = 1$ (ρ_h piecewise constant or linear in time) computed on a sequence of quasi-regular meshes. Assume $\rho_0 \in L^2(\Omega)$,*

$$v \in L^1[0,T;H_0^1(\Omega)], \qquad a,\ \operatorname{div}(v) \in L^1[0,T;L^\infty(\Omega)], \qquad f \in L^1[0,T;L^2(\Omega)],$$

and $\operatorname{div}(v)/2 + a \ge 0$. *If $\rho^0 \to \rho_0$ in $L^2(\Omega)$, $a_h \to a$ in $L^1[0,T;L^\infty(\Omega)]$, and $v_h \in V_h$ is an approximation of v for which $v_h \to v$ in $L^2[0,T;L^2(\Omega)]$ and either* (1) $\operatorname{div}(v_h) \to \operatorname{div}(v)$ *in $L^1[0,T;L^2(\Omega)]$ or* (2) $(\operatorname{div}(v_h) - \operatorname{div}(v))|_K \perp \mathcal{P}_k(K)$ *in $L^2(K)$ for each $K \in \mathcal{T}_h$, then ρ_h converges in $L^2[0,T;L^2(\Omega)]$ to ρ, the weak solution of* (1.1). *Moreover, the jump term*

$$J_h^M = (1/2)\sum_{k=0}^{M-1} \|[\rho^k]\|_{L^2(\Omega)}^2 + (1/2)\sum_e \int_0^T \int_e |v_h.n|[\rho_e]^2$$

converges to zero.

Proof. The idea of the proof is to show that

$$\liminf_h \|\bar{\rho}_h\|_{L^2[0,T;L^2(\Omega)]} \le \|\rho\|_{L^2[0,T;L^2(\Omega)]}. \tag{5.1}$$

If $\ell = 0$, then $\bar{\rho}_h = \rho_h$ and strong convergence of ρ_h follows. When $\ell = 1$, the first estimate in Corollary 3.2 shows that $\bar{\rho}_h$ and ρ_h have the same weak limit so $\bar{\rho}_h$ converges strongly to ρ. The second estimate in Corollary 3.2 shows that strong convergence of $\bar{\rho}_h$ implies strong convergence of ρ_h.

To establish (5.1), the key step is to observe that (2.3) in Theorem 2.2 is an *equation* instead of the usual *inequality*. The hypotheses on v allow us to select $\beta(s) = (1/2)s^2$ in Theorem 2.2 to obtain

$$\begin{aligned}(1/2)\|\rho(t)\|_{L^2(\Omega)}^2 + \int_0^t \int_\Omega ((1/2)\operatorname{div}(v) + a)\rho^2 &= (1/2)\|\rho_0\|_{L^2(\Omega)}^2 + \int_0^t \int_\Omega f\rho \\ &= \liminf_{h\to 0} \left((1/2)\|\rho^0\|_{L^2(\Omega)}^2 + \int_0^t \int_\Omega f\rho_h \right). \end{aligned} \tag{5.2}$$

Integrating both sides with respect to t and using the dominated convergence theorem to interchange the limit and integral gives

$$(1/2)\|\rho\|^2_{L^2[0,T;L^2(\Omega)]} + \int_0^T \gamma(t;\rho)\,dt = \liminf_{h\to 0} \int_0^T \left((1/2)\|\rho^0\|^2_{L^2(\Omega)} + \int_0^t \int_\Omega f\rho_h \right) dt,$$

where

$$\gamma(t;\rho) = \int_0^t \int_\Omega ((1/2)\mathrm{div}(v) + a)\rho^2.$$

Below we use (3.1) to show that

$$\begin{aligned} &\liminf_{h\to 0} \int_0^T \left((1/2)\|\rho^0\|^2_{L^2(\Omega)} + \int_0^t \int_\Omega f\rho_h \right) dt \\ &\geq \liminf_{h\to 0} \left((1/2)\|\bar\rho_h\|^2_{L^2[0,T;L^2(\Omega)]} + \int_0^T \gamma(t;\rho_h)\,dt \right), \end{aligned} \tag{5.3}$$

so that

$$(1/2)\|\rho\|^2_{L^2[0,T;L^2(\Omega)]} + \int_0^T \gamma(t;\rho)\,dt \geq \liminf_{h\to 0} \left((1/2)\|\bar\rho_h\|^2_{L^2[0,T;L^2(\Omega)]} + \int_0^T \gamma(t;\rho_h)\,dt \right).$$

Since $0 \leq (\mathrm{div}(v)/2 + a) \in L^1[0,T;L^\infty(\Omega)]$ it follows that $\gamma(t,.)$ is nonnegative and lower semicontinuous with respect to weak-star convergence in $L^\infty[0,T;L^2(\Omega)]$. Then applying Fatou's lemma to the right-hand side of the above expression shows that $\|\rho\|_{L^2[0,T;L^2(\Omega)]} \geq \liminf_h \|\bar\rho_h\|_{L^2[0,T;L^2(\Omega)]}$, which establishes (5.1).

To verify inequality (5.3), multiply (3.1) by τ and sum to obtain

$$\begin{aligned} \int_0^T \left((1/2)\|\rho^0\|^2_{L^2(\Omega)} + \int_0^t \int_\Omega f\rho_h \right) dt &= (1/2)\|\bar\rho_h\|^2_{L^2[0,T;L^2(\Omega)]} + \int_0^T \gamma(t,\rho_h) + \sum_{m=1}^M \tau J_h^m \\ &\quad + \int_0^T \int_0^t \int_\Omega f\rho_h - \sum_{m=1}^M \tau \int_0^{t^m} \int_\Omega f\rho_h \\ &\quad + \sum_{m=1}^M \tau\gamma(t^m;\rho_h) - \int_0^T \gamma(t,\rho_h)\,dt \\ &\quad + \sum_{m=1}^M \tau \int_0^{t^m} \int_\Omega (a - a_h)\rho_h^2. \end{aligned}$$

To complete the convergence proof we show that the terms in the last three lines vanish as $h \to 0$.

- The terms involving f can be combined as

$$\begin{aligned} \int_0^T \int_0^t \int_\Omega f\rho_h - \sum_{m=1}^M \tau \int_0^{t^m} \int_\Omega f\rho_h &= \sum_{m=1}^M \int_\Omega \left(\int_{t^{m-1}}^{t^m} \int_0^t f\rho_h - \tau \int_0^{t^m} f\rho_h \right) \\ &= \sum_{m=1}^M \int_\Omega \left(\int_{t^{m-1}}^{t^m} \int_{t^{m-1}}^t f\rho_h - \tau \int_{t^{m-1}}^{t^m} f\rho_h \right) \\ &= \sum_{m=1}^M \int_\Omega \int_{t^{m-1}}^{t^m} (t^m - s - \tau) f(s)\rho_h(s)\,ds, \end{aligned}$$

where the last line follows on interchanging the order of integration. It follows that

$$\left|\int_0^T\int_0^t\int_\Omega f\rho_h - \sum_{m=1}^M \tau\int_0^{t^m}\int_\Omega f\rho_h\right| \le 2\tau\|f\|_{L^1[0,T;L^2(\Omega)]}\|\rho_h\|_{L^\infty[0,T;L^2(\Omega)]}.$$

- A similar calculation is used to estimate the terms involving $\gamma(.;\rho_h)$,

$$\sum_{m=1}^M \tau\gamma(t^m;\rho_h) - \int_0^T \gamma(t,\rho_h)\,dt = \sum_{m=1}^M \int_\Omega\int_{t^{m-1}}^{t^m} (t^m - s - \tau)\left(\operatorname{div}(v)/2 + a\right)\rho_h^2\,ds,$$

so that

$$\left|\int_0^T\int_0^t \gamma(s;\rho_h)\,ds\,dt - \sum_{m=0}^M \tau\int_0^{t^m}\gamma(s;\rho_h)\,ds\right| \le 2\tau\|(1/2)\operatorname{div}(v) + a\|_{L^1[0,T;L^\infty(\Omega)]}\|\rho_h\|^2_{L^\infty[0,T;L^2(\Omega)]}.$$

- The last term is bounded by $T\|a - a_h\|_{L^1[0,T;L^\infty(\Omega)]}\|\rho_h\|^2_{L^\infty[0,T;L^2(\Omega)]}$.

To show that the jump terms converge to zero we combine (5.2) and (3.1) to get

$$\|\rho(t)\|^2_{L^2(\Omega)} + \gamma(t,\rho) = \lim_{h\to 0}\left(\|\rho^m\|_{L^2(\Omega)} + \gamma(t^m,\rho_h) + J_h^m\right),$$

where $m = m(h)$ is chosen so that $t \in (t^{m-1}, t^m]$. With this choice $\rho^m = \bar\rho_h(t)$, and since $\|\bar\rho_h(t)\|_{L^2(\Omega)}$ converges to $\|\rho(t)\|_{L^2(\Omega)}$ in $L^2[0,T]$, selecting t to be a Lebesgue point and noting that $\gamma(.,.)$ is continuous on $\mathbb{R}\times L^2[0,T;L^2(\Omega)]$ shows that $\lim_{h\to 0} J_h^m = 0$. Since $J_h^m \ge J_h^M$ for $m \ge M$, and observing that the final time can be chosen arbitrarily, we can choose a Lebesgue point $t \ge T$ to conclude that $J_h^M \to 0$. ∎

6. Monotonicity of the piecewise constant scheme. As stated above, the form of the discontinuous Galerkin method proposed in section 2 was chosen to guarantee convergence when the velocity field was known only approximately. In particular, the factor of 1/2 in the term $(1/2)(\operatorname{div}(v_h) - \operatorname{div}(v))$ guarantees stability in $L^2(\Omega)$; however, $L^p(\Omega)$ estimates would require a different weight. In particular, the piecewise constant scheme, $k = \ell = 0$, may fail to be monotone.

When $k = \ell = 0$, the discontinuous Galerkin scheme (2.4) can be written as

$$|K|(\rho_K^m - \rho_K^{m-1}) + \int_{t^{m-1}}^{t^m}\int_K \left(a_h + (1/2)(\operatorname{div}(v) + \operatorname{div}(v_h))\right)\rho_K^m + \int_{t^{m-1}}^{t^m}\int_{\partial K}(\rho_-^m - \rho^m)(v_h.n)^- = \int_{t^{m-1}}^{t^m}\int_K f.$$

Notice that if K is selected so that ρ_K^m is maximal/minimal, then the boundary term in nonnegative/positive so

$$\max(\rho^m) \le \frac{\max(\rho^{m-1}) + F^m}{1 - C^m} \qquad \text{and} \qquad \min(\rho^m) \ge \frac{\min(\rho^{m-1})}{1 + C^m} - \frac{F^m}{1 - C^m},$$

where

$$F^m = \int_{t^{m-1}}^{t^m}\|f\|_{L^\infty(\Omega)} \quad \text{and} \quad C^m = \int_{t^{m-1}}^{t^m}\|a_h + (1/2)(\operatorname{div}(v) + \operatorname{div}(v_h))\|_{L^\infty(\Omega)},$$

and we have assumed that $\min(\rho^{m-1}) \geq 0$ and τ is sufficiently small to guarantee that $C^m < 1$. Defining

$$\gamma(t) = \int_0^t \|a_h + (1/2)(\operatorname{div}(v) + \operatorname{div}(v_h))\|_{L^\infty(\Omega)}$$

then, assuming $\max(\rho^0) \geq 0$, we compute

$$(1 - o(\tau)) \max(\rho^m) \leq e^{\gamma(t^m)} \max(\rho^0) + \int_0^{t^m} e^{\gamma(t^m)-\gamma(s)} \|f(s)\|_{L^\infty(\Omega)}\, ds,$$

and if $\min(\rho^0) \geq 0$

$$(1 + o(\tau)) \min(\rho^m) \geq e^{-\gamma(t^m)} \min(\rho^0) - \int_0^{t^m} e^{\gamma(t^m)-\gamma(s)} \|f(s)\|_{L^\infty(\Omega)}\, ds,$$

whenever the right-hand side is nonnegative. The terms $1 \pm o(\tau)$ arise when products of the form $\prod_{i=1}^m 1/(1 - C^i)$ are approximated by exponentials of the form $\exp(\sum_{i=1}^m C^i) = \exp(\gamma(t^m))$. Specifically, if $0 \leq C^i \leq 1$, then

$$1 - \sum_{i=1}^m (C^i)^2 \leq \left(\prod_{i=1}^m (1 - C^i)\right) \exp\left(\sum_{i=1}^m C^i\right).$$

Since $\sum_{i=1}^m C^i = \gamma(t^m)$ is bounded and $C^i \to 0$ as $\tau \to 0$, it follows that

$$1 - \sum_{i=1}^m (C^i)^2 \geq 1 - \left(\max_{1\leq i\leq m} C^i\right) \sum_{i=1}^m C^i = 1 - o(\tau).$$

Since it is unlikely that $\operatorname{div}(v_h)$ is bounded in $L^1[0,T;L^\infty(\Omega)]$, this estimate is not particularly useful as it stands. For example, typically it will not be possible to choose τ sufficiently small to guarantee that $C^m < 1$. However, if the construction at the end of section 4 is used to guarantee that $\operatorname{div}(v_h)$ and $\operatorname{div}(v)$ have the same average on each element, then the piecewise constant discontinuous Galerkin scheme becomes

$$|K|(\rho_K^m - \rho_K^{m-1}) + \int_{t^{m-1}}^{t^m} \int_K (a_h + \operatorname{div}(v))\rho_K^m + \int_{t^{m-1}}^{t^m} \int_{\partial K} (\rho_-^m - \rho^m)(v_h.n)^- = \int_{t^{m-1}}^{t^m} \int_K f.$$

In this situation the scheme will be monotone and convergent. The following theorem summarizes these observations.

THEOREM 6.1. *Let $\{\rho_h\}_{h>0}$ be the sequence of solutions of the piecewise constant ($k = \ell = 0$) discontinuous Galerkin scheme* (2.4). *Assume that $v \in L^1[0,T;H_0^1(\Omega)]$, $v_h \in V_h$ and that the averages of* $\operatorname{div}(v_h)$ *and* $\operatorname{div}(v)$ *agree on each element $K \in \mathcal{T}_h$. If a_h,* $\operatorname{div}(v)$, *and f are bounded in $L^1[0,T;L^\infty(\Omega)]$, $0 \leq \alpha \leq \rho^0 \leq \beta$, and τ is sufficiently small, then*

$$(1 - o(\tau)) \max(\rho^m) \leq e^{\gamma(t^m)} \max(\rho^0) + \int_0^{t^m} e^{\gamma(t^m)-\gamma(s)} \|f(s)\|_{L^\infty(\Omega)}\, ds$$

and

$$(1 + o(\tau)) \min(\rho^m) \geq e^{-\gamma(t^m)} \min(\rho^0) - \int_0^{t^m} e^{\gamma(t^m)-\gamma(s)} \|f(s)\|_{L^\infty(\Omega)}\, ds,$$

provided the right-hand side is nonnegative. Here

$$\gamma(t) = \int_0^t \|a_h + \operatorname{div}(v)\|_{L^\infty(\Omega)},$$

and τ sufficiently small is interpreted to mean $\gamma(t^m+\tau)-\gamma(t^m) < 1$ for $m = 0, 1, \ldots$.

Remark. The monotonicity estimates above can be improved slightly if one-sided bounds are used instead of absolute values. For example, in the upper bound $\|a_h + \operatorname{div}(v))\|_{L^\infty(\Omega)}$ can be replaced by $\|(a_h + \operatorname{div}(v))^-\|_{L^\infty(\Omega)}$ and $\|f\|_{L^\infty(\Omega)}$ by $\|f^+\|_{L^\infty(\Omega)}$.

REFERENCES

[1] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Comput. Math. 15, Springer-Verlag, New York, 1991.

[2] Y. C. Chang, T. Y. Hou, B. Merriman, and S. Osher, *A level set formulation of Eulerian interface capturing methods for incompressible fluid flows.*, J. Comput. Phys., 124 (1996), pp. 449–464.

[3] B. Cockburn and P.-A. Gremaud, *A priori error estimates for numerical methods for scalar conservation laws.* I. *The general approach*, Math. Comp., 65 (1996), pp. 533–573.

[4] B. Cockburn, G. E. Karniadakis, and C. W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods, Springer-Verlag, Berlin, 2000, pp. 3–50.

[5] R. J. DiPerna and P. L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[6] J. Jaffré, C. Johnson, and A. Szepessy, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci., 5 (1995), pp. 367–386.

[7] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.

[8] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[9] C. Johnson and Szepessy, *Adaptive finite element methods for conservation laws based on a posteriori error estimates*, Comm. Pure Appl. Math., 48 (1995), pp. 199–234.

[10] S. N. Kruzkov, *First order quasilinear equations in several independent variables*, Mat. USSR Sb., 10 (1970), pp. 217–243.

[11] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York, 1974, pp. 89–123.

[12] Q. Lin, N. Yan, and A. Zhou, *An optimal error estimate of the discontinuous Galerkin method*, Gongcheng Shuxue Xuebao, 13 (1996), pp. 101–105.

[13] Q. Lin and A. H. Zhou, *Convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, Acta Math. Sci., 13 (1993), pp. 207–210.

[14] P. L. Lions, *Mathematical Topics in Fluid Mechanics, Volume* 1: *Incompressible Models*, Oxford University Press, Oxford, UK, 1996.

[15] S. Osher and J. Sethian, *Fronts propagating with curvature dependent speed: Algorithms based on Hamilton Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1818–1829

DISCRETE COMPACTNESS FOR EDGE ELEMENTS IN THE PRESENCE OF MIXED BOUNDARY CONDITIONS*

MIRCO RAFFETTO†

Abstract. It is shown—for the first time, to the best of the author's knowledge—that when the finite dimensional space sequence is generated by using Nedelec's edge elements of any order and of both families defined on tetrahedra, the so-called discrete compactness property holds true for Lipschitz polyhedra even in the presence of mixed boundary conditions. The family of meshes is not required to be quasi-uniform but just regular. A standard way to deal with general dielectric permittivities completes the picture.

Key words. discrete compactness property, edge elements, mixed boundary conditions, Galerkin finite element approximations, convergence, electromagnetic boundary value problems, electromagnetic eigenproblems

AMS subject classifications. 65N30, 65N12, 78M10

DOI. 10.1137/S003614290343414X

1. Introduction.

The importance of the discrete compactness property [1] has recently become evident in computational electromagnetics [2], [3], [4]. In particular, it plays a crucial role in proving the convergence of Galerkin finite element approximations for both eigenproblems [5], [2], [6], [7] and driven problems [8], [9], [4].

It is well known that the discrete compactness property was first proved for first-order Nedelec tetrahedral edge elements of the first family [10] by Kikuchi in 1989 [1]. This fundamental first proof was carried out under some regularity assumptions which are not always satisfied when material properties are discontinuous or when different boundary conditions are given on submanifolds having a common boundary [11], [12]. In other contributions [4], [2] the same property was proved for all Nedelec edge elements (of both families for tetrahedra [10], [13] and of the first family for hexahedra), but the indicated limitations of the original proof were retained. Some results in the presence of boundary conditions of different types are actually available but just for problems where the indicated lack of regularity does not arise. This is the case of [7], where eigenproblems with mixed boundary conditions are introduced to exploit the domain symmetry, or when different boundary conditions are given on different connected components of the boundary [9]. On the contrary, the technique introduced in [6] makes it possible to generalize quite easily all known proofs of discrete compactness to cover cases involving all materials of practical interest [7], [9]. Thus, thanks to this result, today "truly" mixed boundary conditions are one of the main obstacles to be overcome in order to obtain a general proof of the discrete compactness property for edge elements.

Unfortunately, in several models of practical interest different boundary conditions are given on submanifolds having a common boundary. This is the case, for example, for the classical model introduced in [14, pp. 299–301] to deal with aperture coupled rectangular waveguides. Another example is provided in [15, p. 593], where the problem of radiation by apertures is considered. In such cases $\mathbf{n} \times \mathbf{E} = \mathbf{0}$ on part

*Received by the editors September 2, 2003; accepted for publication (in revised form) May 17, 2004; published electronically January 20, 2005.
http://www.siam.org/journals/sinum/42-5/43414.html
†Department of Biophysical and Electronic Engineering, University of Genoa, Via Opera Pia 11a, I–16145, Genoa, Italy (raffetto@dibe.unige.it).

of a plane, and $\mathbf{n} \times \mathbf{H} \neq \mathbf{0}$ on the complementary part of the plane belonging to the boundary. For these and for many other simple and important models the regularity of solenoidal fields is not enough [11] for all known proofs of discrete compactness to hold true. Thus, at present, no convergence result is available for finite element approximations of these practically important problems.

The purpose of this paper is to prove that edge elements of any order and of both families defined on tetrahedra [10], [13] satisfy the discrete compactness property, provided that the family of triangulations is regular [16], independent of the presence of mixed boundary conditions (and, of course, of material inhomogeneities). Thus, as a by-product, a convergence result for edge-based finite element approximations of electromagnetic problems can be established independently of the presence of mixed boundary conditions.

This paper is organized as follows. In section 2 some assumptions on the domain, its boundary, material properties, and the family of triangulations are introduced. In section 3, a set of four conditions is proved to be sufficient for the discrete compactness property to hold true even in the presence of mixed boundary conditions. In carrying out the proof, vector fields are thought of as elements of $H(\mathrm{curl};\Omega)$ and just for the solenoidal fields of this space a standard regularity result is exploited. Then, in section 4, Nedelec's edge elements of any order and of both families defined on tetrahedra are proved to satisfy the sufficient conditions reported in section 3.

2. Assumptions and notations. Let $\Omega \subset \mathbb{R}^3$ be a simply connected Lipschitz polyhedral domain [17]. Let $\Gamma = \partial\Omega$ be its connected boundary and suppose that it splits into two disjoint open submanifolds Γ_τ and Γ_ν satisfying $\overline{\Gamma}_\tau \cup \overline{\Gamma}_\nu = \Gamma$ and $\overline{\Gamma}_\tau \cap \overline{\Gamma}_\nu \neq \emptyset$. Hence, $\overline{\Gamma}_\tau$ and $\overline{\Gamma}_\nu$ have a nonempty common boundary. Moreover, we assume that this common boundary is a piecewise straight simple closed curve. The outward unit vector normal to Γ will be denoted by $\mathbf{n}$. Notice that the limit cases where one of the above submanifolds is empty are not allowed since the result we are looking for is already known [1] in these cases. For the same reason [9], the case of an empty common boundary is also excluded. The above assumptions on Ω, Γ, Γ_τ, and Γ_ν are useful to simplify the following discussion, but not all of them are necessary; our result can be stated under very general topological assumptions [18] (see also [6] for an analogous approach).

It would be possible to introduce some notations to deal with material properties. However, as pointed out in the introduction, Proposition 2.27 of [6] provides a clear indication that material properties are no longer an obstacle. This proposition was exploited in [6] and [7] for proving the convergence of finite element approximations of electromagnetic eigenproblems and, for example, in [9] to deal with electromagnetic driven problems. Thus, we will consider only problems involving homogeneous media.

A consequence of the previous assumptions is that we can use many of the symbols introduced in [6]. For the sake of clarity, however, we report the most important spaces [18] for the next developments:

$$U = H(\mathrm{curl};\Omega), \tag{2.1}$$

$$U_0 = \{\mathbf{u} \in U \mid \mathrm{curl}\,\mathbf{u} = \mathbf{0}\}, \tag{2.2}$$

$$U_1 = \{\mathbf{u} \in U \mid (\mathbf{u},\mathbf{u}_0)_{0,\Omega} = 0 \ \forall \mathbf{u}_0 \in U_0\}, \tag{2.3}$$

$$V = \{\mathbf{v} \in U \mid \mathbf{v} \times \mathbf{n}_{|\Gamma_\tau} = \mathbf{0}\}, \tag{2.4}$$

$$V_0 = V \cap U_0, \tag{2.5}$$

$$V_1 = \{\mathbf{v} \in V \mid (\mathbf{v}, \mathbf{v}_0)_{0,\Omega} = 0 \ \forall \mathbf{v}_0 \in V_0\}, \tag{2.6}$$

where $(\mathbf{u}, \mathbf{v})_{0,\Omega} = \int_\Omega \mathbf{u} \cdot \mathbf{v}^*$ denotes the usual scalar product in $(L^2(\Omega))^3$ for complex valued vector functions. The scalar product in U is denoted by $(\ ,\ )_{curl,\Omega}$ ($(\mathbf{u}, \mathbf{v})_{curl,\Omega} = (\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v})_{0,\Omega} + (\mathbf{u}, \mathbf{v})_{0,\Omega}$), and the norms in $(L^2(\Omega))^3$ and U are denoted by $\| \ \|_{0,\Omega}$ and $\| \ \|_{curl,\Omega}$, respectively.

Let us now introduce a sequence of finite dimensional subspaces of V. In order to do so, let us introduce a family of triangulations $\{\mathcal{T}_h\}_{h\in I}$ of $\overline{\Omega}$ and a specific finite element on the triangulation $\mathcal{T}_h$, thus defining a family $\{V_h\}_{h\in I}$ (or simply $\{V_h\}$) of finite dimensional subspaces of V [16]. We will assume, as usual, that h denotes the maximum diameter of all elements of the triangulation $\mathcal{T}_h$, that I is a denumerable and bounded set of strictly positive numbers having zero as the only limit point, that the family of triangulations is regular [16], that, for every h, $\mathcal{T}_h$ exactly covers $\overline{\Omega}$ (i.e., $\overline{\Omega} = \cup_{K\in\mathcal{T}_h} K$, where K denotes as usual a generic element of the triangulation $\mathcal{T}_h$), and, finally, that, for every h, $\overline{\Gamma}_\nu$ and $\overline{\Gamma}_\tau$ are the union of faces of elements of $\mathcal{T}_h$. Due to their practical importance, in this paper we will be particularly interested in the properties of Nedelec's tetrahedral edge elements. With this aim, let us consider the following spaces [10], [13], [7] ($l \in \mathbb{N}, l > 0$, $m \in \mathbb{N}, m > 0$):

$$R_{l,hV} = \{\mathbf{u}_h \in U : \mathbf{u}_h|_K \in R_l \ \forall K \in \mathcal{T}_h\} \cap V, \tag{2.7}$$

$$Q_{l,hV} = \{\mathbf{u}_h \in U : \mathbf{u}_h|_K \in Q_l \ \forall K \in \mathcal{T}_h\} \cap V, \tag{2.8}$$

$$F_{l,hV} = \{\mathbf{u}_h \in U : \mathbf{u}_h|_K \in F_l \ \forall K \in \mathcal{T}_h\} \cap V, \tag{2.9}$$

$$E_{l+1,hV} = \{\mathbf{u}_h \in U : \mathbf{u}_h|_K \in E_{l+1} \ \forall K \in \mathcal{T}_h\} \cap V, \tag{2.10}$$

$$\begin{aligned} P_{m,1,hV} &= \{\mathbf{u}_h \in U : \mathbf{u}_h|_K \in P_{m,1} \ \forall K \in \mathcal{T}_h\} \cap V \\ &= R_{1,hV} \oplus E_{2,hV} \oplus E_{3,hV} \oplus \cdots \oplus E_{m+1,hV}, \end{aligned} \tag{2.11}$$

where R_l and Q_l are defined in [10] and [13] and are usually referred to as the first and second family of Nedelec's elements, respectively. F_l, E_{l+1}, and $P_{m,1}$ are defined in definitions 1 and 2, p. 340, for F_l and E_{l+1}, respectively, and Remark 6, p. 345, for $P_{m,1}$, of [7] as

$$F_l = \{\mathbf{u} \in Q_l : \text{ all degrees of freedom defined for } R_l \text{ vanish}\}, \tag{2.12}$$

$$E_{l+1} = \{\mathbf{u} \in R_{l+1} : \text{ all degrees of freedom defined for } Q_l \text{ vanish}\}, \tag{2.13}$$

$$P_{m,1} = R_1 \oplus E_2 \oplus E_3 \oplus \cdots \oplus E_{m+1}, \tag{2.14}$$

being the indicated degrees of freedom defined, for example, in [10], [13], [9], and [7]. These spaces will be used, as already pointed out, to define V_h. We will also make use of the following spaces:

$$P_{1,h} = \left\{p_h \in H^1(\Omega) : p_h|_K \in P_1 \ \forall K \in \mathcal{T}_h\right\}, \tag{2.15}$$

$$P_{1,hV} = P_{1,h} \cap H^1_{0,\Gamma_\tau}(\Omega), \tag{2.16}$$

where P_n is the space of polynomials of degree at most n ($n \in \mathbb{N}, n \geq 0$) [16].

Finally, by using the notation

$$V_{0h} = V_h \cap V_0, \tag{2.17}$$

$$V_{1h} = \{\mathbf{v}_h \in V_h \mid (\mathbf{v}_h, \mathbf{v}_{0h})_{0,\Omega} = 0 \ \forall \mathbf{v}_{0h} \in V_{0h}\}, \tag{2.18}$$

let us recall the conditions on $\{V_h\}$ with which we will mainly work:

(DCP) "discrete compactness property"
Any sequence $\{\mathbf{v}_h\}$ such that $\mathbf{v}_h \in V_{1h}$, $\|\mathbf{v}_h\|_{curl,\Omega} \leq C \ \forall h \in I$ contains a subsequence (still denoted by $\{\mathbf{v}_h\}$) such that $\exists \mathbf{v} \in (L^2(\Omega))^3$ such that

$$\lim_{h\to 0} \|\mathbf{v}_h - \mathbf{v}\|_{0,\Omega} = 0;$$

(CDK) "completeness of the discrete kernel"

$$\lim_{h\to 0} \inf_{\mathbf{v}_h \in V_{0h}} \|\mathbf{v} - \mathbf{v}_h\|_{curl,\Omega} = 0 \quad \forall \mathbf{v} \in V_0.$$

In the next section we prove that a sequence $\{V_h\}$ satisfies (DCP) in the presence of mixed boundary conditions, provided that a set of sufficient conditions is satisfied. In order to simplify the notation we will use the same symbols for sequences and their subsequences.

3. Sufficient conditions on $\{V_h\}$ for discrete compactness. The following proposition provides a first abstract result. The main idea is to avoid the problems due to the lack of regularity of the solenoidal fields in V by thinking of an element of V as an element of U.

PROPOSITION 3.1. *Suppose that the hypotheses reported in section* 2 *concerning the domain, its boundary, the homogeneous material, and the triangulation of the domain are satisfied. Then (DCP) is satisfied for any sequence of subspaces* $\{V_h\}$ *satisfying*

H1. $V_h \subset V \ \forall h \in I$,
H2. $\mathbf{v}_h|_K \in (H^1(K))^3 \ \forall h \in I, \forall \mathbf{v}_h \in V_h,$ and $\forall K \in \mathcal{T}_h$,
H3. $\exists C \in \mathbb{R}, C > 0$, C *independent of* h, $\mathbf{v}_h$, *and* K *such that* $\|\mathbf{v}_h|_K\|_{1,K} \leq C\|\mathbf{v}_h|_K\|_{curl,K}$,
H4. $\operatorname{grad}(P_{1,hV}) \subset V_{0h} \ \forall h \in I$.

Proof. The proof is split into five steps.

Step 1. *Definition of a new sequence.* Let us consider a sequence $\{\mathbf{v}_h\}$ such that $\mathbf{v}_h \in V_{1h}$, $\|\mathbf{v}_h\|_{curl,\Omega} \leq C_0 \ \forall h \in I$, $C_0 \in \mathbb{R}$, $C_0 > 0$.

Since $V \subset U = U_0 \oplus U_1$, by using H1 we deduce

$$\mathbf{v}_h = \mathbf{u}_{0ch} + \mathbf{u}_{1ch}, \tag{3.1}$$

where $\mathbf{u}_{0ch} \in U_0$, $\mathbf{u}_{1ch} \in U_1$, and, by the orthogonality in U of the decomposition and the boundedness of the sequence $\{\mathbf{v}_h\}$, $\|\mathbf{u}_{0ch}\|_{curl,\Omega} \leq C_0 \ \forall h \in I$ and $\|\mathbf{u}_{1ch}\|_{curl,\Omega} \leq C_0 \ \forall h \in I$.

The sequence we will mainly work with is $\{\mathbf{u}_{0ch}\}$. Since we are considering a topologically trivial situation we have that

$$\exists! p_{ch} \in H^1(\Omega) : \ \operatorname{grad} p_{ch} = \mathbf{u}_{0ch}, \quad \int_\Omega p_{ch} = 0. \tag{3.2}$$

Step 2. *Approximation of the new sequence.* Since $\mathbf{u}_{1ch} \in U_1 \subset U \cap H_0(\mathrm{div}^0;\Omega) = U \cap \{\mathbf{v} \in H(\mathrm{div};\Omega) : \mathrm{div}\,\mathbf{v} = 0,\ \mathbf{v}\cdot\mathbf{n}|_\Gamma = 0\}$, we obtain $\mathbf{u}_{1ch} \in (H^s(\Omega))^3$, $s > 1/2$, $\|\mathbf{u}_{1ch}\|_{s,\Omega} \le C_1\|\mathbf{u}_{1ch}\|_{curl,\Omega}$ [19, Proposition 3.7], being $\|\ \|_{s,\Omega}$, the natural norm in $(H^s(\Omega))^3$, and C_1, a positive constant depending only on Ω. Moreover, by using H2 we have $\mathbf{v}_h|_K \in (H^1(K))^3\ \forall K \in \mathcal{T}_h$. Then we deduce $\mathbf{u}_{0ch}|_K \in (H^s(K))^3\ \forall K \in \mathcal{T}_h$, $s > 1/2$, which implies $p_{ch}|_K \in H^{1+s}(K) \quad \forall K \in \mathcal{T}_h, \quad s > 1/2$. Thus, by Lemma 4 of [1],

$$p_{ch} \in C^0(\overline{\Omega}) \ \ \forall h \in I, \tag{3.3}$$

and if π_h denotes the standard [16] global P_1 interpolation operator for scalar fields (on tetrahedra), $\exists p_h = \pi_h p_{ch} \in P_{1,h}, \exists C_2 \in \mathbb{R}, C_2 > 0$, C_2 independent of p_{ch} and h:

$$\begin{aligned}
\|\mathrm{grad}\,p_{ch} - \mathrm{grad}\,p_h\|_{0,\Omega} &\le C_2 h^s \left(\sum_{K\in\mathcal{T}_h} \|\mathrm{grad}\,p_{ch}|_K\|^2_{s,K}\right)^{1/2} \\
&= C_2 h^s \left(\sum_{K\in\mathcal{T}_h} \|\mathbf{v}_h|_K - \mathbf{u}_{1ch}|_K\|^2_{s,K}\right)^{1/2} \\
&\le C_2 h^s \left(\sum_{K\in\mathcal{T}_h} (2\|\mathbf{v}_h|_K\|^2_{s,K} + 2\|\mathbf{u}_{1ch}|_K\|^2_{s,K})\right)^{1/2} \\
&\le 2C_2 h^s \left(\|\mathbf{u}_{1ch}\|^2_{s,\Omega} + \sum_{K\in\mathcal{T}_h} \|\mathbf{v}_h|_K\|^2_{1,K}\right)^{1/2} \\
&\le 2C_2 h^s \left(C_1^2\|\mathbf{u}_{1ch}\|^2_{curl,\Omega} + \sum_{K\in\mathcal{T}_h} \|\mathbf{v}_h|_K\|^2_{1,K}\right)^{1/2} \\
&\le 2C_2 h^s \left(C_1^2 C_0^2 + \sum_{K\in\mathcal{T}_h} \|\mathbf{v}_h|_K\|^2_{1,K}\right)^{1/2}. \qquad (3.4)
\end{aligned}$$

Further, by using H3 we obtain

$$\begin{aligned}
\|\mathrm{grad}\,p_{ch} - \mathrm{grad}\,p_h\|_{0,\Omega} &\le 2C_2 h^s \left(C_1^2 C_0^2 + C^2 \sum_{K\in\mathcal{T}_h} \|\mathbf{v}_h|_K\|^2_{curl,K}\right)^{1/2} \\
&\le 2C_2 h^s (C_1^2 C_0^2 + C^2 C_0^2)^{1/2} \le C_4 h^s. \qquad (3.5)
\end{aligned}$$

Step 3. *Behavior of* $p_{ch}|_{\Gamma_\tau}$. By using (3.2), inequality $|p_{ch}|_{1,\Omega} = \|\mathbf{u}_{0ch}\|_{0,\Omega} \le C_0$, and Poincaré's inequality for functions with vanishing mean value, we deduce $\|p_{ch}\|^2_{0,\Omega} \le C_5^2 C_0^2$. Then we immediately obtain $\|p_{ch}\|^2_{1,\Omega} \le C_6^2$ and, consequently, $\|p_{ch}|_\Gamma\|_{1/2,\Gamma} \le C_7$, which in turn implies $\|p_{ch}|_{\Gamma_\tau}\|_{1/2,\Gamma_\tau} \le C_8$. Thus we can conclude that

$$\|p_{ch}|_{\Gamma_\tau}\|_{0,\Gamma_\tau} \le C_8. \tag{3.6}$$

Since by H1 $\mathbf{v}_h \in V$, we have that on Γ_τ

$$\mathbf{n}\times\mathbf{u}_{0ch}|_{\Gamma_\tau} = \mathbf{n}\times(\mathrm{grad}\,p_{ch})|_{\Gamma_\tau} = \mathbf{n}\times\mathbf{v}_h|_{\Gamma_\tau} - \mathbf{n}\times\mathbf{u}_{1ch}|_{\Gamma_\tau} = -\mathbf{n}\times\mathbf{u}_{1ch}|_{\Gamma_\tau}. \tag{3.7}$$

Let Γ_j, $j = 1,\dots,J$, denote the faces of the polyhedron Ω and let $\Gamma_{\tau_j} = \Gamma_j \cap \Gamma_\tau$. We assume that if $j \in M \subset \{1,\dots,J\}$, then Γ_{τ_j} is not empty. On each face Γ_j the

outward unit vector $\mathbf{n}$ normal to Γ is a constant vector. Thus, by using (3.7) and the regularity of $\mathbf{u}_{1ch}$, we deduce

$$\begin{aligned}\|(\mathbf{n}\times(\operatorname{grad} p_{ch})|_{\Gamma_{\tau_j}})\times\mathbf{n}\|_{s-1/2,\Gamma_{\tau_j}} &= \|(\mathbf{n}\times\mathbf{u}_{1ch}|_{\Gamma_{\tau_j}})\times\mathbf{n}\|_{s-1/2,\Gamma_{\tau_j}}\\ &= \|\mathbf{n}\times\mathbf{u}_{1ch}|_{\Gamma_{\tau_j}}\|_{s-1/2,\Gamma_{\tau_j}} \le \|\mathbf{n}\times\mathbf{u}_{1ch}|_{\Gamma_\tau}\|_{s-1/2,\Gamma_\tau}\\ &\le \|\mathbf{n}\times\mathbf{u}_{1ch}|_{\Gamma}\|_{s-1/2,\Gamma} \le C_9\|\mathbf{u}_{1ch}\|_{s,\Omega} \le C_{10},\ j\in M. \end{aligned} \tag{3.8}$$

Finally, by using the fact that $(\mathbf{n}\times(\operatorname{grad} p_{ch})|_{\Gamma_{\tau_j}})\times\mathbf{n}$ is equal to the surface gradient of $p_{ch}|_{\Gamma_{\tau_j}}$ on Γ_{τ_j} [9], [20], by definition 1.3.2.1 of [21], and by inequalities (3.6) and (3.8), we deduce

$$p_{ch}|_{\Gamma_{\tau_j}} \in H^{1/2+s}(\Gamma_{\tau_j}),\quad \|p_{ch}|_{\Gamma_{\tau_j}}\|_{1/2+s,\Gamma_{\tau_j}} \le C_{11},\quad j\in M. \tag{3.9}$$

Let us define $e_{j_1,j_2} = (\overline{\Gamma}_{\tau_{j_1}}\cap\overline{\Gamma}_{\tau_{j_2}})^\circ, j_1, j_2 \in M$, and denote by $\gamma_{2D,j}$ the trace operator from $H^{1/2+s}(\Gamma_{\tau_j})$ to $H^s(\partial\Gamma_{\tau_j})$. By (3.3) we deduce

$$p_{ch}|_{\Gamma_\tau} \in C^0(\overline{\Gamma}_\tau)\ \ \forall h\in I. \tag{3.10}$$

Then, by (3.9) and (3.10) we obtain

$$\gamma_{2D,j_1}(p_{ch}|_{\Gamma_{\tau_{j_1}}})|_{e_{j_1,j_2}} = \gamma_{2D,j_2}(p_{ch}|_{\Gamma_{\tau_{j_2}}})|_{e_{j_1,j_2}},\quad j_1,j_2\in M, \tag{3.11}$$

in the sense of $H^s(e_{j_1,j_2})$ when e_{j_1,j_2} is not empty (if this is not the case condition (3.11) is trivial). When (3.9) and (3.11) are satisfied, in order to simplify the notation we will say that

$$p_{ch}|_{\Gamma_\tau} \in H^{1/2+s}(\Gamma_\tau) \tag{3.12}$$

and

$$\|p_{ch}|_{\Gamma_\tau}\|^2_{1/2+s,\Gamma_\tau} = \sum_{j\in M}\|p_{ch}|_{\Gamma_{\tau_j}}\|^2_{1/2+s,\Gamma_{\tau_j}}, \tag{3.13}$$

and the same notation will be adopted for spaces of functions with domain Γ [22].

By using Sobolev's imbedding theorem [23] for each $p_{ch}|_{\Gamma_{\tau_j}}$, $j\in M$, we deduce that

$$\exists p_{tj}\in H^{1+\delta}(\Gamma_{\tau_j}),\ 0<\delta<s-1/2, \tag{3.14}$$

and for a subsequence of $\{p_{ch}|_{\Gamma_{\tau_j}}\}$

$$\lim_{h\to 0}\|p_{ch}|_{\Gamma_{\tau_j}} - p_{tj}\|_{1+\delta,\Gamma_{\tau_j}} = 0,\quad j\in M. \tag{3.15}$$

By using (3.15), the continuity and the linearity of each trace operator $\gamma_{2D,j}$ from $H^{1+\delta}(\Gamma_{\tau_j})$ to $H^{1/2+\delta}(\partial\Gamma_{\tau_j})$, $j\in M$, and (3.11), we deduce

$$\begin{aligned}&\exists p_t : p_t|_{\Gamma_{\tau_j}} = p_{tj}\in H^{1+\delta}(\Gamma_{\tau_j}),\ j\in M,\\ &\gamma_{2D,j_1}(p_t|_{\Gamma_{\tau_{j_1}}})|_{e_{j_1,j_2}} = \gamma_{2D,j_2}(p_t|_{\Gamma_{\tau_{j_2}}})|_{e_{j_1,j_2}},\quad j_1,j_2\in M\end{aligned} \tag{3.16}$$

(the last condition being trivial when $e_{j_1,j_2}=\emptyset$); that is, according to our notation

$$p_t\in H^{1+\delta}(\Gamma_\tau). \tag{3.17}$$

Moreover, by using (3.15), (3.16) and the norm indicated in (3.13), we obtain

$$\lim_{h\to 0} \|p_{ch}|_{\Gamma_\tau} - p_t\|_{1+\delta,\Gamma_\tau} = 0. \tag{3.18}$$

Step 4. *Some auxiliary scalar fields.* By using Poincaré's inequality and the Lax–Milgram lemma, we define $q_{ch} \in H^1_{0,\Gamma_\tau}(\Omega)$ by

$$\begin{aligned} &\text{find } q_{ch} \in H^1_{0,\Gamma_\tau}(\Omega): \ (\operatorname{grad} q_{ch}, \operatorname{grad} t)_{0,\Omega} \\ &\quad = -(\operatorname{grad} p_{ch}, \operatorname{grad} t)_{0,\Omega} \quad \forall t \in H^1_{0,\Gamma_\tau}(\Omega) \end{aligned} \tag{3.19}$$

and $r_{ch} = p_{ch} + q_{ch} \in H^1(\Omega)$. Note that, $\forall h \in I$, r_{ch} is the unique field in $H^1(\Omega)$ satisfying $r_{ch}|_{\Gamma_\tau} = p_{ch}|_{\Gamma_\tau}$ and $(\operatorname{grad} r_{ch}, \operatorname{grad} t)_{0,\Omega} = 0 \quad \forall t \in H^1_{0,\Gamma_\tau}(\Omega)$.

The same decomposition of p_h is possible. Thus we define $q_h \in P_{1,hV}$ by

$$\begin{aligned} &\text{find } q_h \in P_{1,hV}: \ (\operatorname{grad} q_h, \operatorname{grad} t_h)_{0,\Omega} \\ &\quad = -(\operatorname{grad} p_h, \operatorname{grad} t_h)_{0,\Omega} \quad \forall t_h \in P_{1,hV} \end{aligned} \tag{3.20}$$

and $r_h = p_h + q_h \in P_{1,h}$. For all $h \in I$, r_h is the unique field in $P_{1,h}$ satisfying $r_h|_{\Gamma_\tau} = p_h|_{\Gamma_\tau}$ and $(\operatorname{grad} r_h, \operatorname{grad} t_h)_{0,\Omega} = 0 \quad \forall t_h \in P_{1,hV}$.

Step 5. *A sufficient condition for (DCP) holds true.* The orthogonal decomposition $V = V_0 \oplus V_1$ and hypothesis H1 imply

$$\mathbf{v}_h = \mathbf{v}_{0ch} + \mathbf{v}_{1ch}, \tag{3.21}$$

$\mathbf{v}_{0ch} \in V_0$, $\mathbf{v}_{1ch} \in V_1$, $\|\mathbf{v}_{0ch}\|_{curl,\Omega} \le C_0 \ \forall h \in I$ and $\|\mathbf{v}_{1ch}\|_{curl,\Omega} \le C_0 \ \forall h \in I$. Since V_1 is compactly imbedded in $(L^2(\Omega))^3$ [18, Proposition 7.3], a sufficient condition for the convergence in $(L^2(\Omega))^3$ of $\mathbf{v}_h$ is the convergence in $(L^2(\Omega))^3$ of $\mathbf{v}_{0ch}$.

But note that $\operatorname{grad} q_{ch} = \mathbf{v}_{0ch}$. As a matter of fact, $\operatorname{grad} q_{ch}$ is the V_0 $(= \operatorname{grad}(H^1_{0,\Gamma_\tau}(\Omega)))$ component of $\operatorname{grad} p_{ch} = \mathbf{u}_{0ch} = \mathbf{v}_h - \mathbf{u}_{1ch}$, and $\mathbf{u}_{1ch} \in U_1$ is orthogonal to $U_0 \supset V_0$. Thus inequality (3.5) implies

$$\|\operatorname{grad} r_{ch} - \mathbf{v}_{0ch} - \operatorname{grad} r_h + \operatorname{grad} q_h\|_{0,\Omega} \le C_4 h^s. \tag{3.22}$$

The following sufficient condition for (DCP),

$$\lim_{h\to 0} \|\operatorname{grad} r_{ch} - \operatorname{grad} r_h\|_{0,\Omega} = 0, \tag{3.23}$$

would imply $\lim_{h\to 0} \|\mathbf{v}_{0ch} - \operatorname{grad} q_h\|_{0,\Omega} = 0$, which, in turn, implies $\lim_{h\to 0} \|\mathbf{v}_{0ch}\|_{0,\Omega} = 0$ since $\mathbf{v}_{0ch}$ is orthogonal (in $(L^2(\Omega))^3$) to $V_{0h} \ \forall h \in I$ and by H4 to $\operatorname{grad} q_h \in \operatorname{grad}(P_{1,hV}) \ \forall h \in I$ (as $\mathbf{v}_h \in V_{1h}$ and $\mathbf{v}_{1ch} \in V_1$).

The next two lemmas prove that condition (3.23) is satisfied. □

Remark 3.2. The sequences appearing in condition (3.23) have peculiar properties. As a matter of fact, one could note that on one hand, r_{ch} is the weak solution of a problem for the Laplace operator with mixed boundary conditions given by $r_{ch}|_{\Gamma_\tau} = p_{ch}|_{\Gamma_\tau}$ on Γ_τ and the homogeneous Neumann condition on Γ_ν, with $p_{ch}|_{\Gamma_\tau}$ satisfying (3.18). On the other hand, r_h is its standard finite element approximation (see, for example, [17, p. 147]) obtained by setting $r_h(b_{ih}) = p_{ch}(b_{ih})$, where b_{ih} are the nodes of the triangulation $\mathcal{T}_h$ belonging to $\overline{\Gamma}_\tau$ (the above conditions make sense since $r_{ch}|_{\Gamma_\tau} = p_{ch}|_{\Gamma_\tau} \in C^0(\overline{\Gamma}_\tau)$ (see (3.10)). As a matter of fact, $r_h|_{\Gamma_\tau} = p_h|_{\Gamma_\tau}$ and by definition $p_h = \pi_h p_{ch}$.

Remark 3.3. A result analogous to the one provided by the next lemma can be found, for example, in [24, exercise 5.x.10, p. 132]. However, for the sake of completeness we provide a proof.

LEMMA 3.4. *If* $p_{ch}|_{\Gamma_\tau} = p_t \ \forall h \in I$, *then condition* (3.23) *is satisfied.*

Proof. We use a continuous linear extension operator R_1 from $H^{1+\delta}(\Gamma_\tau)$ into $H^{1+\delta}(\Gamma)$ to define $ep_t = R_1(p_t)$ such that $ep_t|_{\Gamma_\tau} = p_t$ and $\|ep_t\|_{1+\delta,\Gamma} \leq C_1\|p_t\|_{1+\delta,\Gamma_\tau}$. This is possible [21] since Γ_τ is a Lipschitz submanifold of Γ which, in turn, is the Lipschitz continuous boundary of a polyhedron. The same notation as that used in Step 3 of the proof of Proposition 3.1 for spaces of functions with domain Γ_τ or Γ is here adopted.

Now, by using Theorem 2 of [22] there exists a continuous inverse R_2 of the first-order trace operator $\gamma^{(1)}$ [22] such that $r_{02} = R_2(ep_t) \in H^{3/2+\delta}(\Omega)$, $r_{02}|_\Gamma = ep_t$, $r_{02}|_{\Gamma_\tau} = p_t$, and $\|r_{02}\|_{3/2+\delta,\Omega} \leq C_2\|ep_t\|_{1+\delta,\Gamma} \leq C_2C_1\|p_t\|_{1+\delta,\Gamma_\tau}$.

By using Poincaré's inequality and the Lax–Milgram lemma, we define $\phi_2 \in H^1_{0,\Gamma_\tau}(\Omega)$ by

$$\text{find } \phi_2 \in H^1_{0,\Gamma_\tau}(\Omega): \ (\operatorname{grad}\phi_2, \operatorname{grad} t)_{0,\Omega} = -(\operatorname{grad} r_{02}, \operatorname{grad} t)_{0,\Omega} \quad \forall t \in H^1_{0,\Gamma_\tau}(\Omega) \tag{3.24}$$

and $r_2 = \phi_2 + r_{02} \in H^1(\Omega)$. Note that $|\phi_2|_{1,\Omega} \leq |r_{02}|_{1,\Omega} \leq \|r_{02}\|_{3/2+\delta,\Omega} \leq C_2C_1 \|p_t\|_{1+\delta,\Gamma_\tau}$ and that $|r_2|_{1,\Omega} \leq |r_{02}|_{1,\Omega} + |\phi_2|_{1,\Omega} \leq 2C_2C_1\|p_t\|_{1+\delta,\Gamma_\tau}$. Note, moreover, that r_2 is the unique field in $H^1(\Omega)$ satisfying $r_2|_{\Gamma_\tau} = r_{02}|_{\Gamma_\tau} = p_t$ and $(\operatorname{grad} r_2, \operatorname{grad} t)_{0,\Omega} = 0 \quad \forall t \in H^1_{0,\Gamma_\tau}(\Omega)$. The decomposition $r_2 = \phi_2 + r_{02}$ is useful since the irregular component ϕ_2 satisfies homogeneous boundary conditions on Γ_τ.

Since $r_{02} \in H^{3/2+\delta}(\Omega)$ with $\delta > 0$ we can define $r_{02h} = \pi_h r_{02}$. By Theorem 5.48 of [9] we deduce

$$\|r_{02h} - r_{02}\|_{1,\Omega} \leq C_4\|r_{02}\|_{3/2+\delta,\Omega}\, h^{1/2+\delta} \leq C_4C_2C_1\|p_t\|_{1+\delta,\Gamma_\tau}\, h^{1/2+\delta}. \tag{3.25}$$

By definition we have $r_{02h}(b_{ih}) = p_t(b_{ih})$, where b_{ih} are the nodes of the triangulation $\mathcal{T}_h$ belonging to $\overline{\Gamma}_\tau \ \forall h \in I$.

Now we define ϕ_{2h} as follows:

$$\text{find } \phi_{2h} \in P_{1,hV}: \ (\operatorname{grad}\phi_{2h}, \operatorname{grad} t_h)_{0,\Omega} = -(\operatorname{grad} r_{02h}, \operatorname{grad} t_h)_{0,\Omega} \quad \forall t_h \in P_{1,hV}. \tag{3.26}$$

We define also $\phi_{2h,1}$ and $\phi_{2h,2}$ by substituting r_{02h} with $r_{02h} - r_{02}$ and r_{02h} with r_{02} in the above problem, respectively. Thus $\phi_{2h} = \phi_{2h,1} + \phi_{2h,2}$ and we have that $|\phi_{2h,1}|_{1,\Omega} \leq |r_{02h} - r_{02}|_{1,\Omega} \leq C_4C_2C_1\|p_t\|_{1+\delta,\Gamma_\tau}\, h^{1/2+\delta}$. Moreover, it is well known [17, Theorem 18.2] that $\|\operatorname{grad}\phi_2 - \operatorname{grad}\phi_{2h,2}\|_{0,\Omega} = 0$.

Finally, let us define $r_{2h} = \phi_{2h} + r_{02h}$. We have $r_{2h}(b_{ih}) = p_t(b_{ih})$, $\forall b_{ih} \in \overline{\Gamma}_\tau$, $\forall h \in I$ and $(\operatorname{grad} r_{2h}, \operatorname{grad} t_h)_{0,\Omega} = 0 \quad \forall t_h \in P_{1,hV}$. Thus, on one hand, r_{2h} is the unique field in $P_{1,h}$ having these properties and, on the other hand,

$$\begin{aligned} |r_2 - r_{2h}|_{1,\Omega} &= |\phi_2 + r_{02} - \phi_{2h} - r_{02h}|_{1,\Omega} \\ &\leq |\phi_2 - \phi_{2h,1} - \phi_{2h,2}|_{1,\Omega} + |r_{02} - r_{02h}|_{1,\Omega} \\ &\leq |\phi_2 - \phi_{2h,2}|_{1,\Omega} + |\phi_{2h,1}|_{1,\Omega} + |r_{02} - r_{02h}|_{1,\Omega} \\ &\leq |\phi_2 - \phi_{2h,2}|_{1,\Omega} + 2C_4C_2C_1\|p_t\|_{1+\delta,\Gamma_\tau}\, h^{1/2+\delta}. \end{aligned} \tag{3.27}$$

We can conclude the proof by observing that r_2 and r_{2h} can be easily identified with r_{ch} and r_h of Proposition 3.1, respectively, subject to the boundary conditions stated in the present lemma. ∎

LEMMA 3.5. *Condition* (3.23) *is satisfied whenever condition* (3.18) *holds true.*

Proof. This is a trivial consequence of Lemma 3.4. Let us consider $p_{ch}|_{\Gamma_\tau} = p_{ch}|_{\Gamma_\tau} - p_t + p_t = p_{th1} + p_t \ \forall h \in I$.

Consider first just the part of the boundary condition given by p_{th1}. Let us denote by r_{1ch} ($\forall h \in I$) the unique field in $H^1(\Omega)$ corresponding to r_2 in the proof of Lemma 3.4. This means that r_{1ch} is defined by the same procedure we used to define r_2 in the proof of Lemma 3.4, but, in this case, the boundary conditions are given by p_{th1} instead of p_t. Then $|r_{1ch}|_{1,\Omega} \leq 2C_2C_1\|p_{th1}\|_{1+\delta,\Gamma_\tau}$. The components ϕ_{1ch} and r_{01ch} corresponding to ϕ_2 and r_{02} of Lemma 3.4, respectively, satisfy $|\phi_{1ch}|_{1,\Omega} \leq C_2C_1\|p_{th1}\|_{1+\delta,\Gamma_\tau}$ and $\|r_{01ch}\|_{3/2+\delta,\Omega} \leq C_2C_1\|p_{th1}\|_{1+\delta,\Gamma_\tau}$. Finally, the discrete field r_{01h} corresponding to r_{02h} satisfies $\|r_{01h} - r_{01ch}\|_{1,\Omega} \leq C_4C_2C_1\|p_{th1}\|_{1+\delta,\Gamma_\tau}h^{1/2+\delta}$. Condition (3.18) means that $\|p_{th1}\|_{1+\delta,\Gamma_\tau} \to 0$ as $h \to 0$. Thus $|r_{1ch}|_{1,\Omega} \to 0$ and $|r_{01ch}|_{1,\Omega} \to 0$ as $h \to 0$ and therefore $|r_{01h}|_{1,\Omega} \to 0$ as $h \to 0$. Finally, by (3.26) even ϕ_{1h} (corresponding to ϕ_{2h}) is such that $|\phi_{1h}|_{1,\Omega} \to 0$ as $h \to 0$ so that r_{1h} (corresponding to r_{2h}) is such that $|r_{1h}|_{1,\Omega} \to 0$ as $h \to 0$. We deduce that $|r_{1ch} - r_{1h}|_{1,\Omega} \to 0$ as $h \to 0$.

Now, we consider the part of the boundary condition given by $p_{ch}|_{\Gamma_\tau} = p_t$. By Lemma 3.4 we deduce $|r_2 - r_{2h}|_{1,\Omega} \to 0$ as $h \to 0$.

Since r_{ch} and r_h of Proposition 3.1 are such that $r_{ch} = r_{1ch} + r_2$ and $r_h = r_{1h} + r_{2h}$ we deduce condition (3.23). ∎

In the next section we prove that when the sequence of subspaces $\{V_h\}$ is defined by using Nedelec's tetrahedral edge elements (i.e., $V_h = R_{l,hV}\ \forall h \in I$ for any fixed $l \in \mathbb{N},\ l > 0$, or $V_h = Q_{l,hV}\ \forall h \in I$ for any fixed $l \in \mathbb{N},\ l > 0$), conditions H1, H2, H3, and H4 are satisfied.

4. Discrete compactness property for all elements of the two Nedelec families defined on tetrahedra. In this section we simply try to verify that conditions H1, H2, H3, and H4 hold true for the elements of interest.

First, let us point out that H4 is satisfied, provided that $V_h \supset R_{1,hV}$. This is trivially true for $V_h = R_{l,hV}$ and $V_h = Q_{l,hV}$, $l \in \mathbb{N},\ l > 0$, as is well known. However, the same is true by definition (2.11) when $V_h = P_{m,1,hV},\ m \in \mathbb{N}, m > 0$ [7, pp. 338–342 and p. 345].

Second, again by definitions (2.7), (2.8), and (2.11), condition H1 is satisfied for all the above spaces.

Moreover, condition H2 simply requires H^1 regularity on an element-by-element basis. Thus H2 is trivially satisfied by all standard elements and by the elements of interest in particular.

Thus we have to work just on H3. When first-order edge elements of the first Nedelec family are considered, this condition is satisfied by Lemma 5 of [1].

In order to obtain the same result for all other elements we work on the spaces $V_h = P_{m,1,hV},\ m \in \mathbb{N},\ m > 0$. Suppose, for a moment, that we know that the space sequence so generated satisfies (DCP), for any fixed $m > 0$. Since, as already pointed out, $P_{m,1,hV} \supset R_{1,hV}\ \forall m \in \mathbb{N},\ m > 0$ [7, p. 345], a standard result [17] implies that (CDK) is satisfied. Now observe that $\forall h \in I$

- $Q_{1,hV} = R_{1,hV} \oplus F_{1,hV}$ [7, Lemma 19],
- $R_{l,hV} = P_{l-1,1,hV} \oplus F_{1,hV} \oplus F_{2,hV} \oplus \cdots \oplus F_{l-1,hV},\ l > 1$, [7],
- $Q_{l,hV} = P_{l-1,1,hV} \oplus F_{1,hV} \oplus F_{2,hV} \oplus \cdots \oplus F_{l,hV},\ l > 1$, [7],

and $F_{l,hV} \subset V_0\ \forall l > 0$ [7, Lemma 20]. Then a direct application of Lemma 27 of [7] implies that the sequences generated by $V_h = R_{l,hV}\ \forall h \in I$ or by $V_h = Q_{l,hV}\ \forall h \in I$ satisfies (DCP) for any fixed $l \in \mathbb{N},\ l > 0$.

It just remains to prove (DCP) for the space sequence generated by $P_{m,1,hV}$ $\forall m \in \mathbb{N},\ m > 0$. This is done by exploiting the following slight modification of Lemma 30 of [7] ($\hat{\mathbf{z}} = 0$ in the original statement is replaced by $\operatorname{grad} \hat{\mathbf{z}} = 0$ in the following lemma).

LEMMA 4.1. *If Z is a finite dimensional subspace of $(H^1(\hat{K}))^3$ such that $\forall \hat{\mathbf{z}} \in Z$ $\operatorname{curl} \hat{\mathbf{z}} = 0$ implies $\operatorname{grad} \hat{\mathbf{z}} = 0$, then $\exists C > 0$, C independent of $\hat{\mathbf{z}}$, such that $|\hat{\mathbf{z}}|_{1,\hat{K}} \leq C \|\operatorname{curl} \hat{\mathbf{z}}\|_{0,\hat{K}}$.*

Proof. Let $Z_{0c} = \{\mathbf{z} \in Z \mid \operatorname{curl} \mathbf{z} = \mathbf{0}\}$ and $Z_{0g} = \{\mathbf{z} \in Z \mid \operatorname{grad} \mathbf{z} = \mathbf{0}\}$. One of the hypotheses easily implies $Z_{0c} = Z_{0g}$. In order to simplify the notation we denote this space by Z_0. The scalar products $(\ ,\ )_{0,\Omega}$, $(\ ,\ )_{1,\Omega}$, and $(\ ,\ )_{curl,\Omega}$ are all exactly the same when at least one of the two fields involved belongs to Z_0.

Let us now define $Z_0^{\perp} = \{\mathbf{z} \in Z \mid (\mathbf{z}, \mathbf{z}_0)_{0,\Omega} = (\mathbf{z}, \mathbf{z}_0)_{1,\Omega} = (\mathbf{z}, \mathbf{z}_0)_{curl,\Omega} = 0\ \forall \mathbf{z}_0 \in Z_0\}$. We have that in the finite dimensional space $Z_0^{\perp}$ (subspace of Z) the seminorms $|\ |_{1,\hat{K}}$ and $|\ |_{curl,\hat{K}}$ are actually norms, since $Z_0 \cap Z_0^{\perp} = \{0\}$. Thus they are equivalent; i.e., $\exists C > 0$, C independent of $\overline{\mathbf{z}}_1$, such that $|\overline{\mathbf{z}}_1|_{1,\hat{K}} \leq C \|\operatorname{curl} \overline{\mathbf{z}}_1\|_{0,\hat{K}}\ \forall \overline{\mathbf{z}}_1 \in Z_0^{\perp}$.

Since any component $\overline{\mathbf{z}}_0 \in Z_0$ of any $\overline{\mathbf{z}} \in Z$ ($\overline{\mathbf{z}} = \overline{\mathbf{z}}_0 + \overline{\mathbf{z}}_1$) is not able to affect the two seminorms appearing on the left- and right-hand sides of the last inequality, the lemma is proved. ∎

Then Lemma 33 and Corollary 2 of [7] hold true with the same modification. This modified version of Corollary 2 of [7] is the statement we needed since, provided its hypotheses are verified, it implies H3. To help the reader, we report below such a corollary with its original hypotheses.

COROLLARY. *Let us suppose $K \in \mathcal{T}_h$ is affine equivalent to $\hat{K}$. Moreover, let $\sigma > 0$ be such that $\frac{h_K}{\rho_K} \leq \sigma\ \forall K \in \mathcal{T}_h$, $\forall h \in I$. If $Z \subset H^1(K)^3$ is finite dimensional, $\hat{Z} = B_K^T(Z)$ and $\forall \underline{\hat{z}} \in \hat{Z}$, $\operatorname{curl} \underline{\hat{z}} = 0$ implies $\underline{\hat{z}} = 0$, then $\exists C > 0$, C independent of $\underline{z}$, K, and h such that $\|\underline{z}\|_{1,K} \leq C \|\underline{z}\|_{\operatorname{curl},K}\ \forall \underline{z} \in Z$.*

But we always consider a regular family of meshes, so that, by using the notation of the above corollary, there exists $\sigma > 0$ such that $\frac{h_K}{\rho_K} \leq \sigma\ \forall K \in \mathcal{T}_h$, $\forall h \in I$ [17]. Moreover, $\forall \mathbf{w}_h \in P_{m,1,hV}$ we have $\mathbf{w}_h|_K \in P_{m,1} = R_1 \oplus E_2 \oplus \cdots \oplus E_{m+1}\ \forall K \in \mathcal{T}_h, \forall h \in I$, and by Lemma 18 of [7] (see also [23]) the spaces R_l and E_{l+1} are invariant $\forall l > 0$ under the usual affine transformation [23] so that, with the notation of the corollary, $\hat{Z} = B_K^T(Z)\ \forall K \in \mathcal{T}_h$ affine equivalent to $\hat{K}$. Finally, if $\hat{\mathbf{z}} = \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2 + \cdots + \hat{\mathbf{y}}_{m+1}$, $\hat{\mathbf{z}} \in P_{m,1} \subset R_{m+1}$, $\hat{\mathbf{y}}_1 \in R_1 \subset Q_1$, $\hat{\mathbf{y}}_i \in E_i \subset Q_i$, $i = 2, \ldots, m+1$, satisfies $\operatorname{curl} \hat{\mathbf{z}} = 0$, we have by Lemma 9 of [7] that $\hat{\mathbf{z}} \in Q_m$. But $\hat{\mathbf{y}}_i \in E_i \subset Q_i \subset Q_m$, $i = 2, \ldots, m$, and $\hat{\mathbf{y}}_1 \in R_1 \subset Q_1 \subset Q_m$. Then also $\hat{\mathbf{y}}_{m+1} \in Q_m$. Since by Lemma 16 of [7] $E_{m+1} \cap Q_m = \{0\}$, we obtain $\hat{\mathbf{y}}_{m+1} = 0$. In an analogous way we can prove $\hat{\mathbf{y}}_i = 0$, $2 \leq i \leq m$. Thus $\hat{\mathbf{z}} \in R_1$, which implies $\hat{\mathbf{z}} \in P_0^3$, i.e., $\operatorname{grad} \hat{\mathbf{z}} = 0$. Then all hypotheses of the modified version of Corollary 2 of [7] are satisfied and we thus have proved the following proposition.

PROPOSITION 4.2. *Suppose that the hypotheses reported in section 2 concerning the domain, its boundary, the homogeneous material, and the triangulation of the domain are satisfied. Then (DCP) is satisfied for the sequence of subspaces $\{V_h\}$ defined by $V_h = R_{l,hV}\ \forall h \in I$ for any fixed $l \in \mathbb{N},\ l > 0$ or by $V_h = Q_{l,hV}\ \forall h \in I$ for any fixed $l \in \mathbb{N},\ l > 0$.*

Remark 4.3. With the above procedure we have also proved (DCP) for many other sequences of spaces.

Remark 4.4. By using Proposition 2.27 of [6] we could remove the hypothesis concerning the homogeneous material from the statement of Propositions 3.1 and 4.2. As a matter of fact, a generalization of Proposition 2.27 of [6] can be proved when the

dielectric permittivity is a matrix-valued complex function which satisfies conditions H1 and H2 of [25, section 2] (see [9] for a result in this direction). Thus, for the first time, to the best of the author's knowledge, we have proved that (DCP) holds true for all Nedelec edge elements (of any order and of both families defined on tetrahedra) in most cases of interest in engineering applications.

Acknowledgment. The author would like to thank three anonymous reviewers for their very helpful comments.

REFERENCES

[1] F. Kikuchi, *On a discrete compactness property for the Nedelec finite elements*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 36 (1989), pp. 479–490.
[2] D. Boffi, *Fortin operator and discrete compactness for edge elements*, Numer. Math., 87 (2000), pp. 229–246.
[3] D. Boffi, *A note on the de Rham complex and a discrete compactness property*, Appl. Math. Lett., 14 (2001), pp. 33–38.
[4] P. Monk and L. Demkowicz, *Discrete compactness and the approximation of Maxwell's equations in* $\mathbb{R}^3$, Math. Comp., 70 (2001), pp. 507–523.
[5] A. Bermudez and D. G. Pedreira, *Mathematical analysis of a finite element method without spurious solutions for computation of dielectric waveguides*, Numer. Math., 61 (1992), pp. 39–57.
[6] S. Caorsi, P. Fernandes, and M. Raffetto, *On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems*, SIAM J. Numer. Anal., 38 (2000), pp. 580–607.
[7] S. Caorsi, P. Fernandes, and M. Raffetto, *Spurious-free approximations of electromagnetic eigenproblems by means of Nedelec-type elements*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 331–354.
[8] D. Boffi and L. Gastaldi, *Edge finite elements for the approximation of Maxwell resolvent operator*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 293–305.
[9] P. Monk, *Finite Element Methods for Maxwell's Equations*, Numer. Math. Sci. Comput., Oxford University Press, New York, 2003.
[10] J. C. Nedelec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[11] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 627–649.
[12] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Masson, Paris, 1992.
[13] J. C. Nedelec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 50 (1986), pp. 57–81.
[14] R. E. Collin, *Field Theory of Guided Waves*, McGraw–Hill, New York, 1960.
[15] C. A. Balanis, *Antenna Theory: Analysis and Design*, Wiley, New York, 1997.
[16] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[17] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, Handb. Numer. Anal. II, North–Holland, Amsterdam, 1991, pp. 17–351.
[18] P. Fernandes and G. Gilardi, *Magnetostatic and electrostatic problems in inhomogeneous anisotropic media with irregular boundary and mixed boundary conditions*, Math. Models Methods Appl. Sci., 7 (1997), pp. 957–991.
[19] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potential in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.
[20] A. Buffa, M. Costabel, and D. Sheen, *On traces for* $H(curl, \Omega)$ *in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.
[21] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, 1985.
[22] C. Bernardi, M. Dauge, and Y. Moday, *Compatibilité de traces aux arêtes et coins d'un polyhèdre*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 679–684.
[23] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.

[24] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1994.

[25] A. Alonso and M. Raffetto, *Unique solvability for electromagnetic boundary value problems in the presence of partly lossy inhomogeneous anisotropic media and mixed boundary conditions*, Math. Models Methods Appl. Sci., 13 (2003), pp. 597–611.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1830–1845

ASYMPTOTIC MESH INDEPENDENCE OF NEWTON'S METHOD REVISITED*

MARTIN WEISER†, ANTON SCHIELA‡, AND PETER DEUFLHARD§

Abstract. The paper presents a new affine invariant theory on asymptotic mesh independence of Newton's method for discretized nonlinear operator equations. Compared to earlier attempts, the new approach is both much simpler and more intuitive from the algorithmic point of view. The theory is exemplified at finite element methods for elliptic PDE problems.

Key words. asymptotic mesh independence, Newton's method, affine invariance

AMS subject classifications. 65J15, 65L10, 65N30

DOI. 10.1137/S0036142903434047

Introduction. The term "mesh independence" characterizes the observation that finite dimensional Newton methods, when applied to a nonlinear PDE on successively finer discretizations with comparable initial guesses, show roughly the same convergence behavior on all sufficiently fine discretizations. The "mesh independence principle" has been stated and even exploited for mesh design in papers by Allgower and Böhmer [1] and McCormick [19]. Further theoretical investigations of the phenomenon have been given in [2] by Allgower, Böhmer, Potra, and Rheinboldt. Those papers, however, lacked certain important features in the theoretical characterization that made their application to discretized PDEs difficult. This drawback has been avoided in the affine invariant theoretical study by Deuflhard and Potra in [8]; from that analysis, the modified term "asymptotic mesh independence" naturally emerged. The present paper suggests a different approach, which is also affine invariant but much simpler and more natural from the algorithmic point of view.

In a number of papers subsequent to [2], mesh independence principles for different problem settings or different algorithms were established; we mention generalized equations [11, 3], SQP methods [20, 21], shape design [18], constrained Gauss–Newton methods [15], Newton-like methods [16], and gradient projection [17].

The paper is organized as follows. In section 1 we first revisit the theoretical approaches given up to now to treat mesh independence for operator equations. In section 2 we compare discrete versus continuous Newton methods, again in affine invariant terms; in contrast to the earlier treatment in [8], we use only terminology that naturally arises from the algorithmic point of view, such as Newton sequences and approximation errors. The new theory is then exemplified at finite element methods (FEM) for elliptic PDEs (section 3).

1. Preliminary considerations. Let a nonlinear operator equation be denoted by

$$F(x) = 0,$$

*Received by the editors August 29, 2003; accepted for publication (in revised form) April 30, 2004; published electronically January 20, 2005. This research was supported by the DFG Research Center "Mathematics for key technologies" in Berlin. A preprint of this paper appeared as ZIB-Report 03-13, Zuse Institute Berlin, Berlin, Germany, 2003.
http://www.siam.org/journals/sinum/42-5/43404.html
†Corresponding author. Zuse Institute Berlin (ZIB), 14195 Berlin, Germany (weiser@zib.de).
‡Zuse Institute Berlin (ZIB), 14195 Berlin, Germany (schiela@zib.de).
§Free University of Berlin (FU), 14195 Berlin, Germany (deuflhard@zib.de).

where $F : D \to Y$ is defined on a convex domain $D \subset X$ of a Banach space X with values in a Banach space Y. Throughout the paper we assume the existence of a unique solution x^* of this operator equation. The corresponding ordinary Newton method in Banach space may be written as

$$F'(x^k)\Delta x^k = -F(x^k), \quad x^{k+1} = x^k + \Delta x^k, \quad k = 0, 1, \ldots, \tag{1.1}$$

assuming, of course, that the derivatives are invertible. In each Newton step, the linearized operator equation must be solved, which is why this approach is often also called *quasilinearization*. For F, we assume that Theorem 1 from [7] holds, an affine invariant version of the classical Newton–Mysovskikh theorem, whose essence we recall here for the purpose of later reference.

THEOREM 1.1. *Let $F : D \to Y$ be a continuously differentiable mapping with $D \subset X$ convex. Let $\|\cdot\|$ denote the norm in the domain space X. Suppose that $F'(x)$ is invertible for each $x \in D$. Assume that, for collinear x, y, $z \in D$, the following affine invariant Lipschitz condition holds:*

$$\big\|F'(z)^{-1}\big(F'(y) - F'(x)\big)v\big\| \le \omega\|y - x\|\,\|v\|. \tag{1.2}$$

For the initial guess $x^0 \in D$ assume that

$$h_0 = \omega\|\Delta x^0\| < 2$$

and that $\bar{S}(x^0, \rho) \subset D$ for $\rho = \frac{\|\Delta x^0\|}{1 - h_0/2}$.

Then the sequence $\{x^k\}$ of ordinary Newton iterates remains in $S(x^0, \rho)$ and converges to a unique solution $x^* \in \bar{S}(x^0, \rho)$. Its convergence speed can be estimated as

$$\|x^{k+1} - x^k\| \le \tfrac{1}{2}\omega\|x^k - x^{k-1}\|^2.$$

In actual computation, we can solve only discretized nonlinear equations of finite dimension, at best on a sequence of successively finer mesh levels, say,

$$F_j(x_j) = 0, \quad j = 0, 1, \ldots, \tag{1.3}$$

where $F_j : D_j \to Y_j$ denotes a nonlinear mapping defined on a convex domain $D_j \subset X_j$ of a finite dimensional subspace $X_j \subset X$ with values in a finite dimensional space Y_j. We assume F_j results from a Petrov–Galerkin discretization, such that $F_j(x_j) = r_j F(x_j)$ with some linear restriction $r_j : Y \to Y_j$. The corresponding finite dimensional ordinary Newton method reads

$$F_j'(x_j^k)\Delta x_j^k = -F_j(x_j^k), \quad x_j^{k+1} = x_j^k + \Delta x_j^k, \quad k = 0, 1, \ldots.$$

In each Newton step, a system of linear equations must be solved. Since $(F_j)' = r_j F'$, this system can equally well be interpreted either as a discretization of the linearized operator equation (1.1) or as a linearization of the discrete nonlinear system (1.3). Again we assume that Theorem 1.1 holds, this time for the finite dimensional mapping F_j. Let ω_j denote the corresponding affine invariant Lipschitz constant. Then the quadratic convergence of this Newton method is governed by the relation

$$\|x_j^{k+1} - x_j^k\| \le \tfrac{1}{2}\omega_j\|x_j^k - x_j^{k-1}\|^2.$$

Under the assumptions of Theorem 1.1 there exist unique discrete solutions x_j^* on each level j. Of course, we want to choose appropriate discretization schemes such that

$$\lim_{j\to\infty} x_j^* = x^*. \tag{1.4}$$

From the synopsis of the discrete and the continuous Newton method, we immediately see that any comparison of the convergence behavior on different discretization levels j will direct us toward a comparison of the affine covariant Lipschitz constants ω_j. Of particular interest is the connection with the Lipschitz constant ω of the underlying operator equation.

In the earlier papers [1, 2] on mesh independence two assumptions of the kind

$$\|F_j'(x_j)^{-1}\| \le \beta_j, \quad \|F_j'(x_j + v_j) - F_j'(x_j)\| \le \gamma_j \|v_j\|$$

have been made in combination with the *uniformity* requirements

$$\beta_j \le \beta, \quad \gamma_j \le \gamma. \tag{1.5}$$

Obviously, these assumptions lack affine invariance. More important, however, and as a consequence of the noninvariance, these conditions are phrased in terms of *operator norms*, which, in turn, depend on the relation of norms in the domain and the image space of the mappings F_j and F, respectively. For typical PDEs and typical choices of norms we would obtain

$$\lim_{j\to\infty} \beta_j \to \infty,$$

which clearly contradicts the uniformity assumption (1.5). Consequently, an analysis in terms of β_j and γ_j would not be applicable to this important case.

The situation is different with the affine invariant Lipschitz constants ω_j: They depend only on the choice of *norms in the domain space*. It is easy to verify that

$$\omega_j \le \beta_j \gamma_j.$$

In section 2 below we will show that the ω_j remain bounded in the limit $j \to \infty$, as long as ω is bounded—even if either β_j or γ_j blow up. Moreover, even when the product $\beta_j\gamma_j$ remains bounded, the Lipschitz constant ω_j may be considerably lower, i.e.,

$$\omega_j \ll \beta_j \gamma_j.$$

A prerequisite for the asymptotic property (1.4) to hold is that the elements of the infinite dimensional space X can be well approximated by elements of the finite dimensional subspaces X_j. In general, however, the solution x^* has "better smoothness properties" than the generic elements of the space X. For this reason, the earlier papers [2, 8] had restricted their analysis to some smoother subset $W^* \subset X$ and explicitly assumed that

$$x^*, x^k, \Delta x^k, x^k - x^* \in W^*, \quad k = 0, 1, \ldots.$$

However, such an assumption is hard to confirm in the concrete case. That is why we will drop it for our analysis to be presented.

Next, we revisit the paper [8] in some necessary detail. In that paper a family of linear projections

$$\pi_j : X \to X_j, \quad j = 0, 1, \ldots,$$

was introduced, assumed to satisfy the *stability condition*

$$q_j = \sup_{x \in W^*, x \neq 0} \frac{\|\pi_j x\|}{\|x\|} \leq \overline{q} < \infty, \quad j = 0, 1, \ldots. \tag{1.6}$$

The projection property $\pi_j^2 = \pi_j$ immediately gives rise to the lower bound

$$q_j \geq 1. \tag{1.7}$$

As a measure of the *approximation quality* that paper defined

$$\delta_j = \sup_{x \in W^*, x \neq 0} \frac{\|x - \pi_j x\|}{\|x\|}, \quad j = 0, 1, \ldots. \tag{1.8}$$

The rather natural idea that a refinement of the discretization improves the approximation quality was expressed by the asymptotic assumption

$$\lim_{j \to \infty} \delta_j = 0. \tag{1.9}$$

The triangle inequality and (1.6) supplied the upper bound

$$q_j \leq 1 + \delta_j. \tag{1.10}$$

By combination of (1.7), (1.9), and (1.10), asymptotic stability arose as

$$\lim_{j \to \infty} q_j = 1. \tag{1.11}$$

However, as has been pointed out by Braess [6], the above theory has some weak points. In fact, from (1.6) we conclude that $x = 0$ implies $\pi_j x = 0$. The reverse, however, will not be true in general. Hence, one must be aware of pathological elements $x \neq 0$ with corresponding approximations $\pi_j x = 0$. On a uniform one-dimensional grid, such a pathological element might look just like $x(t)$ represented graphically in Figure 1.1. Insertion of such elements into (1.8) would yield

$$\delta_j \geq 1$$

on each level j, on which such pathological elements exist. If one were to accept such an occurrence on *all* levels, then this would be in clear contradiction to the desired asymptotic property (1.9) and its consequence (1.11).

In order to close this gap of that theory, one would have to relate the subset W^* and the projections π_j such that the occurrence of pathological elements would

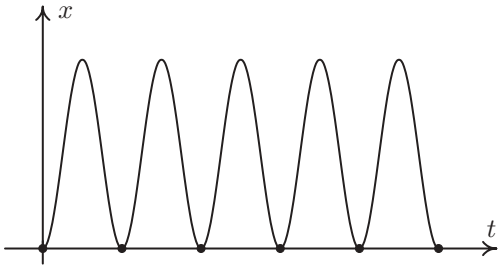


FIG. 1.1. *Pathological element $x \neq 0$ with $\pi_j x = 0$ (•: mesh nodes).*

be asymptotically excluded. As an example, assume we have *nested* subspaces X_j, e.g., constructed by uniform mesh refinement. Suppose we begin with a "sufficiently good" initial projection π_0 on a "sufficiently" fine mesh, which already captures the main qualitative behavior of the solution x^* correctly. Then "pathological" elements would no longer be expected to occur on finer meshes in actual computation. Thus, upon carefully choosing appropriate subsets of W^*, the theory from [8] could, in principle, be repaired. However, the technicalities of such a theory tend to obscure the underlying simple idea.

For this reason, here we abandon that approach and turn to a different one, which seems to us both simpler and more intuitive from the algorithmic point of view: We will avoid the (anyway computationally unavailable) projections π_j and define the approximation quality δ_j differently, just exploiting usual approximation results for discretization schemes.

2. Discrete versus continuous Newton sequences. In this section, we study the comparative behavior of discrete versus continuous Newton sequences. If not explicitly stated otherwise, the notation is taken from the previous section.

We will consider the phenomenon of mesh independence of Newton's method in two steps. First, we will show that the discrete Newton sequence tracks the continuous Newton sequence closely, with a maximal distance bounded in terms of the mesh size; both of the Newton sequences behave nearly identically until, eventually, a small neighborhood of the solution is reached. Second, we prove the existence of affine invariant Lipschitz constants ω_j for the discretized problems, which approach the Lipschitz constant ω of the continuous problem in the limit $j \to \infty$; again, the distance can be bounded in terms of the mesh size. Upon combining these two lines, we finally establish the existence of locally unique discrete solutions x_j^* in a vicinity of the continuous solution x^*.

To begin with, we prove the following nonlinear perturbation lemma.

LEMMA 2.1. *Consider two Newton sequences $\{x^k\}$, $\{y^k\}$ starting at initial guesses x^0, y^0 and continuing as*

$$x^{k+1} = x^k + \Delta x^k, \quad y^{k+1} = y^k + \Delta y^k,$$

where $\Delta x^k, \Delta y^k$ are the corresponding ordinary Newton corrections. Assume the affine invariant Lipschitz condition (1.2) *is satisfied. Then the following contraction result holds:*

$$\|x^{k+1} - y^{k+1}\| \leq \omega \left(\tfrac{1}{2}\|x^k - y^k\| + \|\Delta x^k\|\right) \|x^k - y^k\|. \tag{2.1}$$

Proof. Dropping the iteration index k, we start with

$$\begin{aligned}
x + \Delta x - y - \Delta y
&= x - F'(x)^{-1}F(x) - y + F'(y)^{-1}F(y) \\
&= x - F'(x)^{-1}F(x) + F'(x)^{-1}F(y) - F'(x)^{-1}F(y) - y + F'(y)^{-1}F(y) \\
&= x - y - F'(x)^{-1}\big(F(x) - F(y)\big) + F'(x)^{-1}\big(F'(y) - F'(x)\big)F'(y)^{-1}F(y) \\
&= F'(x)^{-1}\left(F'(x)(x-y) - \int_{t=0}^{1} F'(y + t(x-y))(x-y)\,dt\right) \\
&\quad + F'(x)^{-1}(F'(y) - F'(x))\Delta y.
\end{aligned}$$

Upon using assumption (1.2), we conclude that

$$\begin{aligned}
\|x^{k+1} - y^{k+1}\| &\le \int_{t=0}^{1} \|F'(x^k)^{-1}\big(F'(x^k) - F'(y^k + t(x^k - y^k))\big)(x^k - y^k)\|\,dt \\
&\quad + \|F'(x^k)^{-1}(F'(y^k) - F'(x^k))\Delta y^k\| \\
&\le \frac{\omega}{2}\|x^k - y^k\|^2 + \omega\|x^k - y^k\|\,\|\Delta y^k\|,
\end{aligned}$$

which confirms (2.1). ∎

With the above auxiliary result, we are now ready to study the relative behavior of discrete versus continuous Newton sequences.

THEOREM 2.2. *In addition to the notation as already introduced, let $x^0 = x_j^0 \in X_j$ denote a given starting value such that the assumptions of Theorem* 1.1 *hold for the continuous Newton iteration, including*

$$h_0 = \omega\|\Delta x^0\| < 2.$$

For the discrete mapping F_j and all arguments $x_j \in D_j = D \cap X_j$ define

$$F_j'(x_j)\Delta x_j = -F_j(x_j), \quad F'(x_j)\Delta x = -F(x_j). \tag{2.2}$$

Assume that the discretization is fine enough such that

$$\|\Delta x_j - \Delta x\| \le \delta_j \le \frac{\min\{1, 2 - h_0\}}{2\omega} \tag{2.3}$$

uniformly for $x_j \in D_j$. Assume furthermore $\bar{S}\left(x^0, \rho_j\right) \cap X_j \subset D_j$ for

$$\rho_j := \frac{\|\Delta x_0\|}{1 - h_0/2} + \frac{2\delta_j}{\min\{1, 2 - h_0\}}.$$

Then the sequence of the discrete Newton iterates x_j^k remains in $B(x_0, \rho_j) \cap X_j$ and the following error estimates hold:

$$\|x_j^k - x^k\| \le \frac{2\delta_j}{\min\{1, 2 - h_0\}} \le \frac{1}{\omega} \quad \textit{for all } k \in \mathbb{N}, \tag{2.4}$$

$$\limsup_{k\to\infty} \|x_j^k - x^k\| \le 2\delta_j. \tag{2.5}$$

Proof. In [14, pp. 99, 160], Hairer, Nørsett, and Wanner introduced "Lady Windermere's fan" as a tool to prove discretization error results for evolution problems

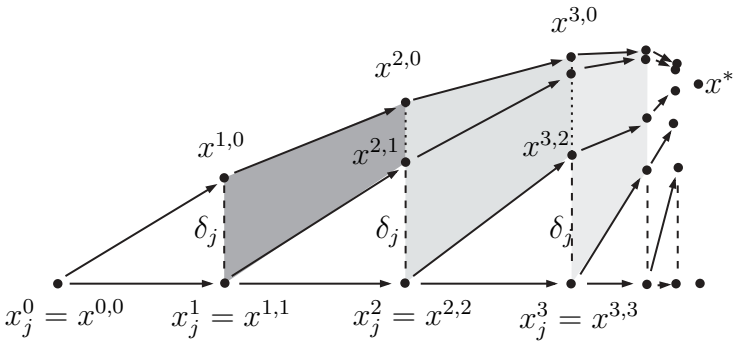


FIG. 2.1. *"Lady Windermere's fan" for the discrete and the continuous Newton method.*

based on some linear perturbation lemma. We may copy this idea and exploit our nonlinear perturbation Lemma 2.1 in the present case. The situation is represented graphically in Figure 2.1.

The discrete Newton sequence starting at the given initial point $x_j^0 = x^{0,0}$ is written as $\{x^{k,k}\}$. The continuous Newton sequence, written as $\{x^{k,0}\}$, starts at the same initial point $x^0 = x^{0,0}$ and runs toward the solution point x^*. In between we define further continuous Newton sequences, written as $\{x^{i,k}\}, k = i, i+1, \ldots$, which start at the discrete Newton iterates $x_j^i = x^{i,i}$ and also run toward x^*. Note that the existence or even uniqueness of a discrete solution point x_j^* is not implied by the assumptions of the theorem.

For the purpose of repeated induction, we assume that

$$\|x_j^{k-1} - x^0\| < \rho_j,$$

which certainly holds for $k = 1$. In order to characterize the deviation between discrete and continuous Newton sequences, we introduce the two majorants

$$\omega\|\Delta x^k\| \le h_k, \quad \|x_j^k - x^k\| \le \epsilon_k.$$

Recall from Theorem 1.1 that

$$h_{k+1} = \frac{1}{2}h_k^2. \tag{2.6}$$

For the derivation of a second majorant recursion, we apply the triangle inequality in the form

$$\|x^{k+1,k+1} - x^{k+1,0}\| \le \|x^{k+1,k+1} - x^{k+1,k}\| + \|x^{k+1,k} - x^{k+1,0}\|.$$

The first term can be treated using assumption (2.3) so that

$$\begin{aligned}\|x^{k+1,k+1} - x^{k+1,k}\| = \|x_j^k + \Delta x_j^k - \left(x^{k,k} + \Delta x^{k,k}\right)\| &= \|\Delta x_j^k - \Delta x^{k,k}\| \\ &\le \delta_j.\end{aligned} \tag{2.7}$$

For the second term, we may apply our nonlinear perturbation Lemma 2.1 (see the shaded regions in Figure 2.1) to obtain

$$\|x^{k+1,k} - x^{k+1,0}\| \le \omega\left(\frac{1}{2}\|x^{k,k} - x^{k,0}\| + \|\Delta x^{k,0}\|\right)\|x^{k,k} - x^{k,0}\|.$$

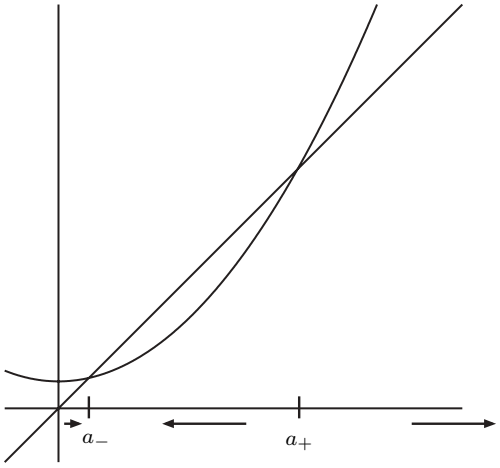


FIG. 2.2. *Fixed point recursion* a_k.

Combining these results then leads to

$$\|x^{k+1,k+1} - x^{k+1,0}\| \le \delta_j + \frac{\omega}{2}\epsilon_k^2 + h_k\epsilon_k.$$

The above right-hand side may be defined to be ϵ_{k+1}. Hence, together with (2.6), we arrive at the following set of majorant equations:

$$\begin{aligned} h_{k+1} &= \frac{1}{2}h_k^2, & h_0 &= \omega\|\Delta x_0\|, \\ \epsilon_{k+1} &= \delta_j + \frac{1}{2}\omega\epsilon_k^2 + h_k\epsilon_k, & \epsilon_0 &= 0. \end{aligned}$$

Now for $\beta \ge 1$ we multiply the second recursion by $\beta\omega$ and add both recursions. This yields the following recursion for $\beta\omega\epsilon_k + h_k$:

$$\beta\omega\epsilon_{k+1} + h_{k+1} = \beta\omega\delta_j + \frac{1}{2}(\beta\omega\epsilon_k + h_k)^2 - \left[\tfrac{1}{2}(\beta-1)\beta\omega^2\epsilon_k^2\right].$$

Since the term in squared brackets is positive, the sequence a_k defined by

$$a_{k+1} = \beta\omega\delta_j + \frac{1}{2}a_k^2, \qquad a_0 = h_0, \tag{2.8}$$

is a majorant to $\beta\omega\epsilon_k + h_k$. Solving (2.8) yields the equilibrium points

$$a_\pm = 1 \pm \sqrt{1 - 2\beta\omega\delta_j} \tag{2.9}$$

if $2\beta\omega\delta_j \le 1$, which is always possible to guarantee by choosing $1 \le \beta \le (2\omega\delta_j)^{-1}$ due to (2.3). The sequence converges monotonically toward the stable fixed point a_- in case $h_0 < a_+$ (see Figure 2.2). We consider the two cases $h_0 \le 1$ and $h_0 > 1$ separately. If $h_0 \le 1$, we choose

$$\beta = \frac{1}{2\omega\delta_j},$$

such that $h_0 \le a_- = 1$. Due to monotonicity the sequence a_k is bounded from above by $a_- = 1$. We then derive the upper bound

$$\epsilon_k \le \frac{a_-}{\beta\omega} \le 2\delta_j.$$

Both (2.4) and (2.5) are covered by this result. For $1 < h_0 < 2$, we choose $\sigma > 0$ sufficiently small and

$$\beta = \frac{h_0(2-h_0)}{(2+\sigma)\omega\delta_j},$$

such that both $\beta \ge 1$ and $h_0 < a_+$ are satisfied. Due to monotonicity, the sequence a_k is bounded from above by $a_0 = h_0$, and we obtain

$$\epsilon_k \le \frac{h_0}{\beta\omega} = \frac{(2+\sigma)\delta_j}{2-h_0}. \tag{2.10}$$

Since (2.10) holds for all sufficiently small $\sigma > 0$, we obtain

$$\epsilon_k \le \frac{2\delta_j}{2-h_0},$$

which proves (2.4). The asymptotic result (2.5) is now an immediate consequence of $a_k \to a_-$.

Finally, with application of the triangle inequality

$$\|x_j^{k+1} - x^0\| \le \|x^{k+1} - x^0\| + \epsilon_{k+1} < \frac{\|\Delta x_0\|}{1-h_0/2} + \frac{2\delta_j}{\min\{1, 2-h_0\}} = \rho_j,$$

the induction and therefore the whole proof are completed. ∎

We are interested in the question of whether a discrete solution point x_j^* exists. The above tracking theorem, however, states only that the discrete Newton sequence stays close to the continuous Newton sequence and therefore has an accumulation point close to the continuous solution.

COROLLARY 2.3. *Under the assumptions of Theorem* 2.2*, there exists at least one accumulation point*

$$\hat{x}_j \in \bar{S}\,(x^*, 2\delta_j) \cap X_j \;\subset\; S\left(x^*, \frac{1}{\omega}\right) \cap X_j,$$

which need not be a solution point of the discrete equations $F_j(x_j) = 0$.

In order to prove more, Theorem 1.1 directs us to study whether a Lipschitz condition of the kind (1.2) additionally holds.

LEMMA 2.4. *Assume Theorem* 1.1 *holds for the mapping* $F : X \to Y$. *For collinear* $x_j, y_j, z_j \in X_j$, *define* $u_j \in X_j$ *and* $u \in X$ *according to*

$$F'(x_j)u = (F'(z_j) - F'(y_j))\, v_j, \tag{2.11}$$

$$F_j'(x_j)u_j = \left(F_j'(z_j) - F_j'(y_j)\right) v_j \tag{2.12}$$

for arbitrary $v_j \in X_j$. *Assume that the discretization method satisfies*

$$\|u - u_j\| \le \sigma_j \|z_j - y_j\|\, \|v_j\|. \tag{2.13}$$

Then there exist constants

$$\omega_j \le \omega + \sigma_j\,, \tag{2.14}$$

such that the affine invariant Lipschitz condition

$$\|u_j\| \le \omega_j \|z_j - y_j\|\,\|v_j\|$$

holds.

Proof. The proof is a simple application of the triangle inequality:

$$\begin{aligned}\|u_j\| \le \|u\| + \|u_j - u\| &\le \omega\|z_j - y_j\|\,\|v_j\| + \sigma_j\|z_j - y_j\|\,\|v_j\| \\ &= (\omega + \sigma_j)\,\|z_j - y_j\|\,\|v_j\|. \qquad \square\end{aligned}$$

Finally, the existence of a unique discrete solution x_j^* close to the continuous solution x^* is a direct consequence.

COROLLARY 2.5. *Under the assumptions of Theorem* 2.2 *and Lemma* 2.4 *the discrete Newton sequence* $\{x_j^k\}, k = 0, 1, \ldots,$ *converges q-quadratically to a unique discrete solution point*

$$x_j^* \in \bar{S}\,(x^*, 2\delta_j) \cap X_j \;\subset\; S\left(x^*, \frac{1}{\omega}\right) \cap X_j.$$

Proof. We just need to apply Theorem 1.1 to the finite dimensional mapping F_j with the starting value $x_j^0 = x^0$, and the affine invariant Lipschitz constant ω_j from (2.14). $\square$

Summarizing, we come to the following conclusion, at least in terms of the analyzed upper bounds: If the asymptotic properties

$$\lim_{j\to\infty} \delta_j = 0, \quad \lim_{j\to\infty} \sigma_j = 0,$$

can be shown to hold, then the convergence speed of the discrete ordinary Newton method is asymptotically just the same as that of the continuous ordinary Newton method. Moreover, if related initial guesses x^0 and x_j^0 and a common termination criterion are chosen, then even the number of iterations will be nearly the same.

3. Application to discretization schemes. In order to apply the abstract mesh independence principles of section 2 to discretization schemes for differential equations, we have to show two features. First,

$$\|\Delta x - \Delta x_j\| \le \delta_j, \qquad \lim_{j\to\infty} \delta_j = 0, \tag{3.1}$$

where Δx is the exact and Δx_j is the approximate solution of the Newton equations (2.2), respectively.

Second,

$$\|u - u_j\| \le \sigma_j \|z_j - y_j\|\,\|v_j\|, \qquad \lim_{j\to\infty} \sigma_j = 0, \tag{3.2}$$

where u and u_j are the solutions of the Lipschitz equations (2.11) and (2.12), respectively.

The structure of the argumentation will be straightforward. The first step is to apply classical error estimates for the numerical method under consideration. These estimates usually depend on the regularity of the exact solution y of the linear correction problems. The second step is then to show appropriate regularity results for y.

We concentrate on FEM for elliptic PDEs. Collocation methods for ODEs are discussed in [22].

FEM for semilinear elliptic PDEs. Assume $f : \mathbb{R} \to \mathbb{R}$ is monotonically increasing and locally Lipschitz continuously differentiable with

$$|f'(x) - f'(y)| \le L(1 + \max(|x|, |y|))|x - y|. \tag{3.3}$$

This implies the growth condition $f = \mathcal{O}(|x|^3)$, which in turn implies that the nonlinear superposition (or Nemyckii) operator $\mathbf{f}$ generated by f maps $H_0^1(\Omega)$ continuously into $L_2(\Omega)$ on some convex polygonal domain $\Omega \subset \mathbb{R}^d$, $d \le 3$, via the embedding $H_0^1(\Omega) \hookrightarrow L_6(\Omega)$ (cf. [4, 12]). We define the continuous problem $F(x) = 0$ as the boundary value problem

$$F(x) = -\mathrm{div}(\kappa \nabla x) + \mathbf{f}(x) = 0, \quad x \in H_0^1(\Omega), \tag{3.4}$$

with $0 < \underline{\kappa} \le \kappa \le \overline{\kappa}$. The discretizations F_j are provided by finite element approximations on shape-regular triangulations $\mathcal{T}_j$ with mesh size $\tau_j = \max_{T \in \mathcal{T}_j} \mathrm{diam}\, T$. We consider piecewise linear finite element spaces $X_j \subset H_0^1(\Omega)$ on the triangulations $\mathcal{T}_j$.

THEOREM 3.1. *Let a bounded set $D \subset H_0^1(\Omega)$ be given. Then there exist constants $M_1, M_2 < \infty$ depending only on D and the problem setting $P = (\Omega, \kappa, f)$, such that the Newton-FEM discretizations F_j satisfy the Newton approximation condition* (3.1) *with $\delta_j = M_1 \tau_j$,*

$$\|\Delta x - \Delta x_j\|_{H^1} \le M_1 \tau_j \quad \textit{uniformly for } x_j \in D \cap X_j, \tag{3.5}$$

and the Lipschitz approximation condition (3.2) *with $\sigma_j = M_2 \tau_j$,*

$$\|u - u_j\|_{H^1} \le M_2 \tau_j \|z_j - y_j\|_{H^1} \|v_j\|_{H^1} \tag{3.6}$$

uniformly for all $y_j, z_j \in D \cap X_j$ and $v_j \in X_j$.

Proof. First we prove (3.5). Let Δx satisfy $F'(x_j)\Delta x = -F(x_j)$ and let Δx_j be its FEM approximation. Returning to (2.7) we notice that $x^{k+1,k}$ is more regular than $\Delta x^{k,k}$. Thus we introduce $w = x_j + \Delta x$, which satisfies

$$-\mathrm{div}(\kappa \nabla w) + \mathbf{f}'(x_j) w = -\mathbf{f}(x_j) + \mathbf{f}'(x_j) x_j. \tag{3.7}$$

The growth condition (3.3) implies $\mathbf{f}(x_j) \in L_2$ and $\mathbf{f}'(x_j) \in L_3$, such that the right-hand side of (3.7) is contained in L_2. We may estimate

$$\begin{aligned}
\|\mathbf{f}(x_j) - \mathbf{f}'(x_j) x_j\|_{L_2} &= \left\| \int_{t=0}^{1} (\mathbf{f}'(t x_j) - \mathbf{f}'(x_j)) x_j \, dt + \mathbf{f}(0) \right\|_{L_2} \\
&\le \int_{t=0}^{1} L(1-t) \|(1 + |x_j|) x_j^2\|_{L_2} \, dt + c \\
&\le \frac{L}{2} (\|x_j^2\|_{L_2} + \|x_j^3\|_{L_2}) + c \\
&= c(\|x_j\|_{L_4}^2 + \|x_j\|_{L_6}^3 + 1) \\
&\le c(\|x_j\|_{H^1}^2 + \|x_j\|_{H^1}^3 + 1) \\
&\le c,
\end{aligned}$$

where c denotes a generic constant independent of the discretization and x_j. Since the Helmholtz term in (3.7) is positive semidefinite due to the monotonicity of f, the inverse of the differential operator can be bounded in terms of the ellipticity constant of its main part only, which is independent of x_j. Thus we obtain

$$\|w\|_{H^1} \le c \|\mathbf{f}(x_j) - \mathbf{f}'(x_j) x_j\|_{L_2} \le c.$$

Using Hölder's inequality and the embedding $H_0^1(\Omega) \hookrightarrow L_6(\Omega)$, we estimate

$$\|\mathbf{f}'(x_j)w\|_{L_2} \le \|w\|_{L_6}\|\mathbf{f}'(x_j)\|_{L_3} \le \|w\|_{H^1} c(1 + \|x_j\|_{L_6}^2) \le c.$$

We now rewrite (3.7) as

$$-\mathrm{div}(\kappa\nabla w) = -\mathbf{f}(x_j) + \mathbf{f}'(x_j)x_j - \mathbf{f}'(x_j)w.$$

Since the right-hand side is contained in L_2, the solution w is H^2-regular (cf. [13]) with

$$\|w\|_{H^2} \le c\|\mathbf{f}(x_j) - \mathbf{f}'(x_j)x_j + \mathbf{f}'(x_j)w\|_{L_2} \le c.$$

We thus obtain an approximation error

$$\|w_j - w\|_{H^1} \le c\tau_j\|w\|_{H^2} \le c\tau_j$$

for its FEM approximation $w_j = x_j + \Delta x_j$ (cf. [5, p. 79]), uniformly for all x_j. For the approximation error $\Delta x_j - \Delta x$ we now obtain

$$\|\Delta x_j - \Delta x\|_{H^1} = \|w_j - w\|_{H^1} \le c\tau_j.$$

Second, we prove (3.6). u is defined by

$$F'(x_j)u = (F'(z_j) - F'(y_j))v_j = (\mathbf{f}'(z_j) - \mathbf{f}'(y_j))v_j.$$

As before, the right-hand side is contained in L_2 and the solution u is H^2-regular, such that we obtain

$$\|u_j - u\|_{H^1} \le c\tau_j\|(\mathbf{f}'(z_j) - \mathbf{f}'(y_j))v_j\|_{L_2}.$$

Upon using Hölder's inequality twice we conclude that

$$\begin{aligned}\|(\mathbf{f}'(z_j) - \mathbf{f}'(y_j))v_j\|_{L_2} &\le \|L^2(1 + \max(|y_j|, |z_j|))^2(z_j - y_j)^2 v_j^2\|_{L_1}^{1/2} \\ &\le \left(\|L^2(1 + \max(|y_j|, |z_j|))^2\|_{L_3}\|(z_j - y_j)^2\|_{L_3}\|v_j^2\|_{L_3}\right)^{1/2} \\ &= L\|1 + \max(|y_j|, |z_j|)\|_{L_6}\|z_j - y_j\|_{L_6}\|v_j\|_{L_6} \le c,\end{aligned}$$

which completes the proof. ∎

Combining Theorem 2.2 and Lemma 2.4 with Theorem 3.1 we obtain asymptotic mesh independence for FEM approximations of semilinear elliptic equations.

COROLLARY 3.2. *Assume that there exists a convex and bounded set $D \subset H^1$, such that on D the assumptions of Theorem* 1.1 *(in particular $\omega < \infty$) and Theorem* 3.1 *are satisfied for the nonlinear equation* (3.4).

Then there exists a constant M_1 and a mesh size $\tau_0 > 0$, such that for all discretizations X_j with corresponding mesh size $\tau_j < \tau_0$ and starting values $x^0 = x_j^0 \in X_j$ with

$$h_0 = \omega\|\Delta x^0\|_{H^1} < 2 \tag{3.8}$$

and

$$\bar{S}\left(x^0, \frac{\|\Delta x_0\| + 2M_1\tau_j}{1 - h_0/2}\right) \subset D, \tag{3.9}$$

the discrete Newton sequence remains in D, and its distance to the continuous Newton sequence is bounded by

$$\|x_j^k - x^k\|_{H^1} \le \frac{2M_1\tau_j}{1-h_0/2}. \tag{3.10}$$

Moreover, both sequences converge q-quadratically to solutions x_j^* and x^*, respectively, with

$$\|x_j^* - x^*\|_{H^1} \le 2M_1\tau_j. \tag{3.11}$$

Proof. Application of Theorem 3.1 on D yields constants $M_1, M_2 < \infty$ such that $\|\Delta x - \Delta x_j\|_{H^1} \le M_1\tau_j$ and $\|u-u_j\|_{H^1} \le M_2\tau_j\|z_j-y_j\|_{H^1}\|v_j\|_{H^1}$ hold for all $x_j \in D$ in terms of (3.1) and (3.2). We will verify Corollary 3.2 for

$$\tau_j < \tau_0 := \min\left\{\frac{1-h_0/2}{2\omega M_1}, \frac{1}{2M_2}\left(\sqrt{\omega^2 + \frac{M_2(1-h_0/2)}{M_1}} - \omega\right)\right\}.$$

Note that the continuous Newton sequence satisfying (3.8) and (3.9) remains in $S(x^0, \frac{\|\Delta x_0\|_{H^1}}{1-h_0/2}) \subset D$ due to Theorem 1.1. Because of

$$\tau_j < \tau_0 \le \frac{1-h_0/2}{2\omega M_1} \le \frac{\min\{1, 2-h_0\}}{2\omega M_1},$$

condition (2.3) is clearly satisfied, such that we can apply Theorem 2.2 and obtain (3.10), (3.11), and $x_j^k \in D$.

Now we turn to q-quadratic convergence of the discrete Newton sequence. A direct consequence of (3.10) is the estimate

$$\begin{aligned}\|\Delta x_j^k\|_{H^1} &\le \|x_j^{k+1} - x^{k+1}\|_{H^1} + \|\Delta x^k\|_{H^1} + \|x^k - x_j^k\|_{H^1} \\ &\le \frac{4M_1\tau_j}{1-h_0/2} + \|\Delta x^k\|_{H^1}.\end{aligned}$$

As $\lim_{k\to\infty}\|\Delta x^k\|_{H^1} = 0$ by Theorem 1.1 we can find an index k_0 such that

$$\|\Delta x_j^k\|_{H^1} \le \frac{8M_1\tau_j}{1-h_0/2} \qquad \text{for all } k \ge k_0.$$

Application of Lemma 2.4 yields $\omega_j \le \omega + \tau_j M_2$ and therefore

$$h_j^k := \omega_j\|\Delta x_j^k\|_{H^1} \le (\omega + \tau_j M_2)\frac{8M_1\tau_j}{1-h_0/2} \qquad \text{for all } k \ge k_0.$$

Now

$$\tau_j < \tau_0 \le \frac{1}{2M_2}\left(\sqrt{\omega^2 + \frac{M_2(1-h_0/2)}{M_1}} - \omega\right)$$

implies $h_j^{k_0} < 2$, such that Theorem 1.1 yields q-quadratic convergence of the discrete Newton iteration starting at $x_j^{k_0}$. ∎

FEM for strongly nonlinear elliptic PDEs. For strongly nonlinear PDEs with a second order differential operator depending on the solution, the analytic treatment of the approximation conditions (3.1) and (3.2) is considerably more difficult. The global regularity of the right-hand side is, in general, only H^{-1}, which results in sharp edges in the Newton correction. These bucklings, however, coincide geometrically with the edges of the triangulation, such that the finite element approximation quality does not deteriorate. This effect is indeed observed in actual computation.

The regularity theory necessary for addressing such problems is beyond the scope of the present paper. As a substitute, we give a numerical example from [10], where the phenomenon of asymptotic mesh independence may be studied.

Example: Parametric minimal surface. Consider the variational problem

$$\min \int_\Omega \sqrt{1+|\nabla u|^2}\, dx$$

subject to the boundary conditions

$$\begin{aligned} u &= \cos(x)\cos(y) && \text{on } \Gamma_D = \partial\Omega \backslash \Gamma_N, \\ \frac{\partial u}{\partial n} &= 0 && \text{on } \Gamma_N \end{aligned}$$

on $\Omega = [-\pi/2, 0]^2$. The functional gives rise to the first and second order expressions

$$\begin{aligned} \langle F(u), v\rangle &= \int_\Omega \left(1+|\nabla u|^2\right)^{-1/2} \nabla u^T \nabla v\, dx, \\ \langle F'(u)v, w\rangle &= \int_\Omega \Big(-\left(1+|\nabla u|^2\right)^{-3/2} \nabla w^T (\nabla u \nabla u^T) \nabla v \\ &\qquad + \left(1+|\nabla u|^2\right)^{-1/2} \nabla w^T \nabla v \Big)\, dx. \end{aligned}$$

We define two different problem settings by choosing

(a) $\Gamma_N = [-\pi/2, 0] \times \{0\}$,

(b) $\Gamma_N = [-\pi/2, 0] \times \{0\} \cup \{0\} \times [-\pi/4, 0]$.

Note that by symmetry, problem (a) represents a Dirichlet problem on a convex domain, whereas the deliberate choice of boundary conditions (b) leads to a Dirichlet problem on a highly nonconvex slit domain, on which no physically meaningful solution exists.

The adaptive Newton-multilevel code Newton-KASKADE [9, 10] has been run on both problems, providing affine invariant computational estimates $[\omega_j] \le \omega_j$ on each mesh refinement level j. On each level, a few Newton steps have been computed using the approximation from the level before, and the maximum estimate encountered in these steps has been selected as $[\omega_j]$. As can be seen from Table 3.1, the Lipschitz constants for the well-defined problem (a) remain bounded and rather independent of the refinement level, apart from some fluctuation due to the finite sampling of ω_j. In contrast to that, the estimates for the Lipschitz constant of problem (b) are dramatically increasing by five orders of magnitude. This indicates that the problem has finite dimensional solutions on each of the successive meshes, each unique within the corresponding finite dimensional Kantorovich ball with radius $\rho_j \sim 1/\omega_j$; however, these balls shrink from radius $\rho_0 \approx 1$ to $\rho_{12} \approx 10^{-5}$. Frank extrapolation of this effect insinuates the conjecture that there exists no continuous unique solution of the underlying minimization problem.

TABLE 3.1
Estimated Lipschitz constants $[\omega_j]$ on different refinement levels j.

	Problem (a)		Problem (b)	
j	♯nodes	$[\omega_j]$	♯nodes	$[\omega_j]$
0	4	1.32	5	7.5
1	7	1.17	10	4.2
2	18	4.55	17	7.3
3	50	6.11	26	9.6
4	123	5.25	51	22.5
5	158	20.19	87	50.3
6	278	19.97	105	1486.2
7	356	9.69	139	2715.6
8	487	8.47	196	5178.6
9	632	11.73	241	6837.2
10	787	44.21	421	12040.2
11	981	49.24	523	167636.0
12	1239	20.10	635	1405910.0
13	1610	32.93		
14	2054	37.22		

Acknowledgment. One of the authors (P.D.) wishes to thank Dietrich Braess, Bochum, for detailed discussions about the earlier paper [8]; these discussions eventually led to our decision to analyze asymptotic mesh independence in a simpler and more intuitive framework. Moreover, the authors gratefully acknowledge his constructive comments on an earlier draft of this paper.

REFERENCES

[1] E. L. ALLGOWER AND K. BÖHMER, *Application of the mesh independence principle to mesh refinement strategies*, SIAM J. Numer. Anal., 24 (1987), pp. 1335–1351.
[2] E. L. ALLGOWER, K. BÖHMER, F. A. POTRA, AND W. C. RHEINBOLDT, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.
[3] W. ALT, *Mesh-independence of the Lagrange-Newton method for nonlinear optimal control problems and their discretizations*, Ann. Oper. Res., 101 (2001), pp. 101–117.
[4] J. APPELL AND P. P. ZABREJKO, *Nonlinear Superposition Operators*, Cambridge University Press, Cambridge, UK, 1990.
[5] D. BRAESS, *Finite Elements*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.
[6] D. BRAESS, *private communication*, 2002.
[7] P. DEUFLHARD AND G. HEINDL, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal., 16 (1979), pp. 1–10.
[8] P. DEUFLHARD AND F. POTRA, *Asymptotic mesh independence of Newton–Galerkin methods via a refined Mysovskii theorem*, SIAM J. Numer. Anal., 29 (1992), pp. 1395–1412.
[9] P. DEUFLHARD AND M. WEISER, *Local inexact Newton multilevel FEM for nonlinear elliptic problems*, in Computational Science for the 21st Century, M.-O. Bristeau, G. Etgen, W. Fitzgibbon, J.-L. Lions, J. Periaux, and M. Wheeler, eds., Wiley-Interscience, New York, 1997, pp. 129–138.
[10] P. DEUFLHARD AND M. WEISER, *Global inexact Newton multilevel FEM for nonlinear elliptic problems*, in Multigrid Methods V, Lect. Notes Comput. Sci. Eng. 3, W. Hackbusch and G. Wittum, eds., Springer-Verlag, Berlin, 1998, pp. 71–89.
[11] A. DONTCHEV, W. HAGER, AND V. VELIOV, *Uniform convergence and mesh independence of Newton's method for discretized variational problems*, SIAM J. Control Optim., 39 (2000), pp. 961–980.
[12] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, New York, 1977.
[13] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman, Boston, 1985.
[14] E. HAIRER, S. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations. I: Nonstiff Problems*, Springer Ser. Comput. Math. 8, 2nd ed., Springer-Verlag, Berlin, 1993.

[15] M. Heinkenschloss, *Mesh independence for nonlinear least squares problems with norm constraints*, SIAM J. Optim., 3 (1993), pp. 81–117.
[16] C. Kelley and E. Sachs, *Mesh independence of Newton-like methods for infinite dimensional problems*, J. Integral Equations Appl., 3 (1991), pp. 549–573.
[17] C. T. Kelley and E. W. Sachs, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477–493.
[18] M. Laumen, *Newton's mesh independence principle for a class of optimal shape design problems*, SIAM J. Control Optim., 37 (1999), pp. 1070–1088.
[19] S. McCormick, *A revised mesh refinement strategy for Newton's method applied to nonlinear two-point boundary value problems*, in Numerical Treatment of Differential Equations Applications, Lecture Notes in Math. 679, Springer-Verlag, Berlin, 1978, pp. 15–23.
[20] S. Volkwein, *Mesh-independence for an augmented Lagrangian-SQP method in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 767–785.
[21] S. Volkwein, *Mesh-independence of Lagrange-SQP methods with Lipschitz-continuous Lagrange multiplier updates*, Optim. Methods Softw., 17 (2002), pp. 77–111.
[22] M. Weiser, A. Schiela, and P. Deuflhard, *Asymptotic Mesh Independence of Newton's Method Revisited*, preprint, ZIB Report 03-13, Zuse Institute Berlin, Berlin, Germany, 2003.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1846–1861

ON COMPARING THE WRITHE OF A SMOOTH CURVE TO THE WRITHE OF AN INSCRIBED POLYGON*

JASON CANTARELLA†

Abstract. We find bounds on the difference between the writhing numbers of a smooth curve and a polygonal curve inscribed within. The proof is based on an extension of Fuller's difference of writhe formula to the case of polygonal curves. The results establish error bounds useful in the numerical computation of writhe in terms of bounds on the edge lengths of the polygon and the derivatives of the curve. The bounds are "adaptive" in the sense that they improve when regions of the smooth curve with larger derivatives are approximated by shorter edges of the polygon.

Key words. writhing number, numerical quadrature

AMS subject classifications. 53A04, 65D18, 65D30

DOI. 10.1137/S0036142902403164

1. Introduction.

The writhing number measures the wrapping and coiling of space curves. Writhe has proved useful in molecular biology, where it is used to study the geometry of tangled strands of DNA [19]; often with the famous Călugăreanu–White formula for a curve C in space with a normal field V [5, 6, 21, 15]:

$$\mathrm{Lk}(C, C + \epsilon V) = \mathrm{Tw}(C, V) + \mathrm{Wr}(C).$$

Writhe has proved important in the study of elastic rods in biology [20] and has even been used as the basis for a system for automatically classifying protein structures [16].

In these applications, and in numerical simulations performed by biologists and mathematicians, it is often required to compute writhing numbers using numerical methods.

For polygonal curves, there is a well-developed set of algorithms for computing writhe. These methods range from Banchoff's original exact sum formula for the writhe of a polygonal curve [3], to Agarwal, Edelsbrunner, and Wang's fast sweepline algorithm [1], which is based on deep results from computational geometry. Many other methods are surveyed by Klenin and Langowski in [13] (see also [9, 19]).

For smooth curves, the only existing methods are the standard tools for numerical integration, such as adaptive quadrature. It would be desirable to apply the sophisticated algorithms developed for the polygonal case in this setting, but there is a missing ingredient: We must be able to bound the error introduced in approximating a smooth curve by an inscribed polygonal curve. The purpose of this paper is to bridge this gap by proving the following theorem.

THEOREM 1. *Suppose $C(t)$ is a simple, closed curve of class $\mathcal{C}^4$. We assume $C(t)$ is parametrized so that $|C'(t)| \geq 1$, and that we have upper bounds $B_1, \ldots, B_4$ on $|C'(t)|, \ldots, |C^{(4)}(t)|$. Let $C_n(t)$ be any n-edge polygonal curve inscribed in C with maximum edge length x and $1/x > 5B_2$.*

*Received by the editors February 25, 2002; accepted for publication (in revised form) May 3, 2004; published electronically January 20, 2005. This work was supported by a National Defense Science and Engineering Graduate Fellowship, an NSF Postdoctoral Research Fellowship (DMS-99-02397), and an NSF Individual Investigator grant (DMS-02-04862).
http://www.siam.org/journals/sinum/42-5/40316.html
†Department of Mathematics, University of Georgia, Athens, GA 30602 (jason@math.uga.edu).

If the ribbon formed by joining $C_n(t)$ to $C(t)$ for every t is embedded,

$$|\operatorname{Wr}(C) - \operatorname{Wr}(C_n)| < \alpha\, n x^3 + n O(x^4), \tag{1}$$

where α is a numerical constant less than $B_2(5B_2^2 + B_3)$.

That is, if the lengths of the edges of C_n are approximately constant, the error is bounded by a multiple of $1/n^2$. We also derive a "local" version of these bounds which is useful when the derivatives of the curve are large in some regions and small in others. In this case, we may split the curve and derive independent bounds on each region's contribution to the overall approximation error (see Corollary 17). This allows us to use more edges of C_n to approximate regions where C has large derivatives.

The proof is based on Fuller's ΔWr formula, which gives the difference in writhing number between two curves as the spherical area of the ribbon bounded by the curves on S^2 swept out by their unit tangent vectors [11]. (Following Solomon [18], we will refer to such curves as *tantrices*, though they are classically referred to as *tangent indicatrices*.)

We begin by defining the writhing number in section 2. Sections 3 and 4 then introduce the original form of Fuller's ΔWr formula. In sections 5 and 6 we extend Fuller's formula to the case where one curve is polygonal and the other is of class $\mathcal{C}^2$ using a natural geometric idea: The tantrix of a polygonal curve should be defined to be the chain of geodesic segments on S^2 joining the (isolated) tangent vectors of the curve (this was pointed out by Chern in [8]). In the process, we discover a surprising fact: The writhe of a polygonal curve is *equal* to the writhe of any smooth curve obtained by carefully rounding off its corners!

Section 8 contains the remainder of our work: estimating the terms in our improved version of the ΔWr formula to obtain Theorem 13. We test our error bounds in section 9 by computing the writhe of a collection of polygonal curves inscribed in a smooth curve of known writhe.

The last section contains a discussion of some open problems inspired by the present work. We state the most important of them now: Like most of the theory of writhing numbers, the proof of our main theorem depends essentially on the fact that C is closed. Can these methods be extended to open curves?

2. Definitions. The writhing number of a space curve is defined by the following definition.

Definition 2. *The* writhe *of a piecewise differentiable curve $C(s)$ is given by*

$$\operatorname{Wr}(C) = \frac{1}{4\pi} \int_{C\times C} \frac{C'(s) \times C'(t) \cdot (C(s) - C(t))}{|C(s) - C(t)|^3}\, \mathrm{d}s\, \mathrm{d}t. \tag{2}$$

Definition 2 is inspired by the Gauss formula for the linking number of two space curves, $A(s)$ and $B(s)$ (see Epple [10] for a fascinating discussion of the history of this formula):

$$\operatorname{Lk}(A, B) = \frac{1}{4\pi} \int_{A\times B} \frac{A'(s) \times B'(t) \cdot (A(s) - B(t))}{|A(s) - B(t)|^3}\, \mathrm{d}s\, \mathrm{d}t. \tag{3}$$

When the two curves A and B become a single curve, their linking number becomes the writhing number. This introduces a potential singularity on the diagonal of $C \times C$, but a careful calculation shows that the integral still converges. In fact, the integrand of (2) approaches 0 on the diagonal of $C \times C$, even when the curve C has a corner.

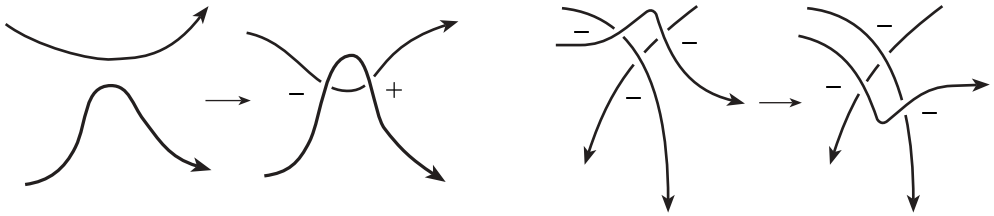


FIG. 1. *Changing the projection direction within a cell can only alter the diagram by one of these two moves. Neither changes the signed crossing number of the diagram, as we can see by counting the + and − markers at the crossings of C.*

From now on, we will assume that C is simple. With this assumption, another way to look at the integral of Definition 2 is to observe that the integrand is the pullback of the area form on S^2 under the Gauss map $C \times C \to S^2$ defined by

$$(C(s), C(t)) \mapsto \frac{C(s) - C(t)}{|C(s) - C(t)|}. \tag{4}$$

From this perspective, we can see that the (signed) multiplicity of the Gauss map at any point p on S^2 is just the number of self-crossings of the projection of C in direction p.

3. Fuller's ΔWr formula. Suppose we have a differentiable curve $C(t)$, with unit tangent vector $T(t)$. As we mentioned in section 1, the curve $T(t)$ on the unit sphere is known as the tantrix of C. This curve divides the unit sphere into a number of cells. Within each cell, the signed crossing number of the projection of C is constant: Changing projection directions within the cell amounts to altering the projection of the knot by a regular isotopy consisting of Reidemeister moves of types II and III (pictured in Figure 1). Neither of these moves changes the signed crossing number of the knot.

This observation motivates the idea that the writhe of a closed space curve is related to the fraction of the sphere's area enclosed by its tantrix. In 1978, Brock Fuller stated the following formula.

THEOREM 3 (Fuller's spherical area formula). *For any closed space curve $C(s)$ of class $\mathcal{C}^3$, let A be the spherical area enclosed by the tantrix of C. Then*

$$1 + \mathrm{Wr}(C) = \frac{A}{2\pi} \mod 2. \tag{5}$$

Fuller used this formula to conclude that the difference in writhe between two curves X_0 and X_1 whose tantrices T_0 and T_1 are sufficiently close is given by a certain formula, which represents the spherical area of the ribbon between T_0 and T_1.

To be more specific, suppose that X_0 and X_1 are simple closed space curves of class $\mathcal{C}^2$, with regular parametrization (that is, parametrized so that X_0' and X_1' never vanish), and unit tangent vectors T_0 and T_1. Let $F : S^1 \times [0,1] \to \mathbf{R}^3$ be a continuous deformation of X_0 into X_1, where $F(t,\lambda) = X_\lambda(t)$ and the X_λ are simple curves of class $\mathcal{C}^1$, with unit tangent vectors $T_\lambda(t)$ continuous in (t,λ).

THEOREM 4 (Fuller's ΔWr formula). *If $T_1(t)$ and $T_\lambda(t)$ are not antipodal for all (t,λ), then*

$$\mathrm{Wr}(X_1) - \mathrm{Wr}(X_0) = \frac{1}{2\pi} \int_C \frac{T_0(t) \times T_1(t)}{1 + T_0(t) \cdot T_1(t)} \cdot [T_0'(t) + T_1'(t)] \, \mathrm{d}t. \tag{6}$$

We observe that this formula does not require an arc-length parametrization of X_0 and X_1.

4. Justifying Fuller's interpretation of the ΔWr formula. While Fuller stated both of these theorems in 1978, he did not provide complete proofs for either. The first rigorous proofs of Theorems 3 and 4 were given by Aldinger, Klapper, and Tabor [2] in 1995. While these authors proved both theorems as stated, they did not show that the formula in Theorem 4 represents the spherical area of the ribbon between T_0 and T_1 (in [2], the right-hand side of (6) describes the difference between the twist of two frames on X_0 and X_1.)

In the spirit of their paper, we now justify Fuller's original intuition about (6).

PROPOSITION 5. *Given two curves $T_0(t), T_1(t) : [0,1] \to S^2$, where $T_0(t)$ and $T_1(t)$ are never antipodal, the area of the spherical region R bounded by T_0, T_1 and the great circle arcs joining their endpoints is given by*

$$\text{Area}(R) = \int \frac{T_0(t) \times T_1(t)}{1 + T_0(t) \cdot T_1(t)} \cdot (T_0' + T_1') \, \mathrm{d}t. \tag{7}$$

Proof. We let

$$u(\theta, t) = \cos\theta \; T_0(t) + \sin\theta \; T_1(t)$$

and parametrize the region R by

$$v(\theta, t) = \frac{u(\theta, t)}{|u(\theta, t)|},$$

where θ ranges from 0 to $\pi/2$. Plugging this parametrization into the area form on S^2 and using the properties of the triple product, we find

$$\mathrm{d}\,\text{Area} = \frac{1}{|u|^3} \left(\frac{\partial u}{\partial \theta} \times \frac{\partial u}{\partial s} \cdot u \right) \mathrm{d}\theta \wedge \mathrm{d}t.$$

Using the definition of $u(\theta, t)$, this simplifies to

$$\mathrm{d}\,\text{Area} = T_0 \times T_1 \cdot \left(\frac{\cos\theta}{(1 + 2\cos\theta \sin\theta \; T_0 \cdot T_1)^{\frac{3}{2}}} T_0' + \frac{\sin\theta}{(1 + 2\cos\theta \sin\theta \; T_0 \cdot T_1)^{\frac{3}{2}}} T_1' \right) \mathrm{d}\theta \wedge \mathrm{d}t.$$

Using the formula $\sin 2\theta = 2\cos\theta \sin\theta$ and the fact that the definite integrals of each of the trigonometric expressions above from 0 to $\pi/2$ are equal, we have

$$\text{Area}(R) = \int_0^1 T_0 \times T_1 \cdot \left[\int_0^{\pi/2} \frac{\cos\theta}{(1 + \sin 2\theta \; T_0 \cdot T_1)^{3/2}} \, \mathrm{d}\theta \right] (T_0' + T_1') \, \mathrm{d}t.$$

This can be solved by the general integration formula

$$\int \frac{\cos\theta}{(1 + a\sin 2\theta)^{3/2}} \, \mathrm{d}\theta = \frac{-a\cos\theta - \sin\theta}{(a^2 - 1)\sqrt{1 + a\sin 2\theta}}, \tag{8}$$

which yields the formula in the statement of the proposition. ∎

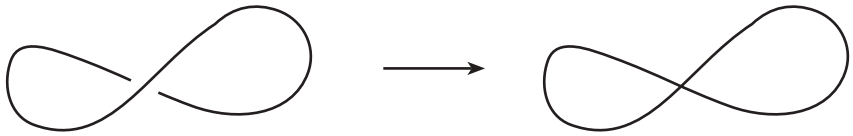


FIG. 2. *The family of almost-planar curves on the left converge in any C^k norm to the planar figure eight curve on the right. However, the writhe of the curves on the left approaches one, while the writhe of the planar figure eight is zero. This shows that writhe is not continuous in any C^k norm on curves.*

5. Extending Fuller's formulas to polygonal curves: I. To measure the difference in writhe between a smooth curve and a polygonal curve inscribed in the smooth curve, we must extend Theorem 4 to polygonal curves. This raises an immediate objection: Theorem 4 deals with the area enclosed by a curve's tantrix, while the set of tangent vectors of a polygon is a collection of isolated points!

To make sense of Fuller's theorem in this context, we recall the following definition from Chern [8].

DEFINITION 6. *The tantrix of a piecewise C^1 curve $C(s)$ with positive corner angles is the image of $T(s)$ on the unit sphere, together with the great circle arcs joining the pairs of tangent vectors at each corner of the curve.*

We will show that Theorem 4 holds for polygons under this definition. To do so, we intend to approximate each polygonal curve with a family of smooth curves so that the writhe of the smooth curves converges to the writhe of the polygonal curve.

Examining Definition 2, it might seem that this would be easy. For instance, one might conjecture that Wr was continuous in the C^1 norm on curves and hope to obtain an approximating family using standard techniques. Unfortunately, the situation is not so simple; as the example in Figure 2 shows, writhe is not continuous in any C^k norm on curves. Thus, our proof depends explicitly on the hypothesis that the limit curve is polygonal; it cannot be easily extended to the case where the limit curve is merely piecewise C^2.

To prepare for the proof, we establish some notation for polygonal curves. Let $C(t)$ be a polygonal curve with corners at cyclically ordered parameter values $t_0 < t_1 < \cdots < t_n = t_0$. We let $T(t)$ denote the unit tangent to C and set up the convention that $T(t_i)$ will be the tangent vector *leaving* $C(t_i)$.

We now construct a family of smooth curves approximating our polygonal curve.

PROPOSITION 7. *Given an embedded polygonal curve C with corners at $t_0, \ldots, t_{n-1}$, $t_n = t_0$, there exists a family of smooth curves C_i converging pointwise to C with*

1. *$C_i = C$ outside a neighborhood of each corner point $C(t_j)$ of radius $1/i$;*
2. *near each corner, the tangent vectors of C_i interpolate between $T(t_{j-1})$ and $T(t_j)$;*
3. *$\mathrm{Wr}(C_i) \to \mathrm{Wr}(C)$.*

Proof. It is easy to construct a family of $C_i \to C$ obeying conditions 1 and 2 by rounding off each corner of C. We claim that this can be done in such a way that the writhe integrand has a uniform upper bound on all the C_i. Since condition 1 implies that the $C_i \to C$ pointwise in the $\mathcal{C}^1$ norm, the bounded convergence theorem [17, p. 81] will then yield condition 3.

Since any pair of adjacent edges is planar, we can choose the C_i so that the region of each C_i approximating a pair of adjacent edges is also planar. This means that for some universal ϵ, the writhe integrand of each C_i vanishes in an ϵ-neighborhood of the diagonal of $C_i \times C_i$.

Since C has no self-intersections and the angle at each corner of C_i is positive, the distance between any pair of nonadjacent edges of C is bounded below by some constant. Since the C_i converge to C pointwise, we may assume the same for the portions of the C_i approximating any pair of disjoint edges. Throwing away finitely many of the C_i if necessary, this means that for any $\delta > 0$, there exists a universal lower bound (depending on δ) on the distance between any pair of points in $C_i \times C_i$ *outside* an δ-neighborhood of the diagonal.

But for any pair of points on C_i, the writhe integrand is bounded above by the inverse square of the distance between them. Thus, our lower bound on self-distances yields a universal upper bound on the writhe integrand for C and all the C_i outside a δ-neighborhood of the diagonal. Choosing $\delta < \epsilon$, this completes the proof of the proposition. □

6. Extending Fuller's formulas to polygonal curves: II. We now state our extension of Fuller's theorem. Our formula will apply to the following situation (cf. section 3): Suppose that X_0 and X_1 are simple closed space curves, with X_0 of class $\mathcal{C}^2$ and X_1 polygonal, with regular parametrization (that is, parametrized so that X_0' and X_1' never vanish where they are defined), and unit tangent vectors T_0 and T_1.

Let $F : S^1 \times [0,1] \to \mathbf{R}^3$ be a C^0 deformation of X_0 into X_1, where $F(t,\lambda) = X_\lambda(t)$, so that the X_λ are simple curves of class C^1 for $\lambda \in [0,1)$, with unit tangent vectors $T_\lambda(t)$ continuous in (t,λ). As above, we take the corners of X_1 to be at parameter values $t_0, t_1, \ldots, t_n = t_0$. We let T_1 denote the unit tangent vector to X_1, and let $T_1(t_i)$ be the tangent vector *leaving* $X_1(t_i)$.

THEOREM 8. *If each corner angle of X_1 is strictly greater than $\pi/2$ and each $T_1(t)$ and $T_\lambda(t)$ are at an angle less than $\pi/2$, then*

$$\mathrm{Wr}(X_1) - \mathrm{Wr}(X_0) = \frac{1}{2\pi} \operatorname{Area} R(T_0, T_1),$$

where $R(T_0, T_1)$ is the spherical region bounded by the tantrices of X_0 and X_1 (using Definition 6*) and great circle segments joining their endpoints, and* Area *represents oriented area on S^2.*

COROLLARY 9. *If each corner angle of X_1 is strictly greater than $\pi/2$ and each $T_1(t)$ and $T_\lambda(t)$ are at an angle less than $\pi/2$, then*

$$\begin{aligned}\mathrm{Wr}(X_1) - \mathrm{Wr}(X_0) = \frac{1}{2\pi} \sum_{i=1}^{n} &\operatorname{Area} R(T_0(t_i), T_0(t_{i+1}), T_1(t_i)) \\ &+ \operatorname{Area} \triangle T_0(t_i) T_1(t_{i-1}) T_1(t_i).\end{aligned}$$

Here $R(T_0(t_i), T_0(t_{i+1}), T_1(t_i))$ is the spherical region bounded by geodesics from $T_1(t_i)$ to $T_0(t_i)$ and $T_0(t_{i+1})$ and the portion of T_0 between t_i and t_{i+1}, $\triangle T_0(t_i) T_1(t_{i-1}) T_1(t_i)$ is the spherical triangle with these three vertices, and Area *represents oriented area on S^2.*

Proof. Construct a sequence of smooth curves $C_j \to X_1$ using Proposition 7. For large enough j, each of these curves can be homotoped to X_1 through a family of simple $\mathcal{C}^1$ curves with a continuous family of tangent vectors, as in the setup for the statement of Theorem 8.

Joining these homotopies to the homotopy from X_1 to X_0 assumed by our hypotheses generates a family of (nonsmooth) homotopies from the X_0 to each of the

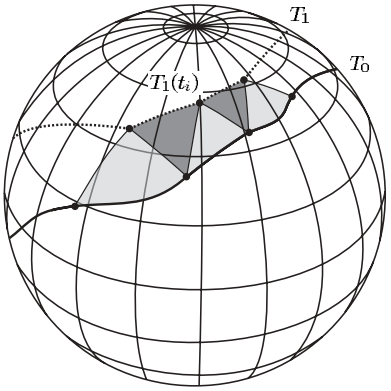


FIG. 3. *This figure shows the two types of regions in the sum in the statement of Theorem* 8. *The top (dotted) curve shows the great circle arcs joining the tangent vectors $T(t_i)$ of the polygonal curve X_1. The bottom curve shows the continuous curve of unit tangents T_0 to the smooth curve X_0. The light gray regions show the first terms in the sum, while the dark gray spherical triangles show the second terms.*

C_j. We wish to smooth each of these to obtain homotopies from X_0 to C_j which obey the conditions of Fuller's ΔWr formula (Theorem 4).

We first prove that the tangent vectors of each of the intermediate curves in each homotopy from X_0 to C_j are never antipodal to the corresponding tangent vectors T_j of C_j. By hypothesis, for each t and λ, $\angle T_\lambda(t), T_1(t) < \pi/2$. On the other hand, since the difference between the tangent vectors to X_1 at any corner is less than $\pi/2$, for large enough j, $\angle T_1(t), T_j(t) < \pi/2$. Putting these equations together, we see that $\angle T_\lambda(t), T_j(t) < \pi$, and so these vectors are never antipodal.

It is easy to smooth the combined homotopy from X_0 to C_j so that each of the intermediate curves is of class $\mathcal{C}^1$ while preserving this condition. Since the smoothed homotopy satisfies the hypotheses of Fuller's ΔWr formula (Theorem 4), Proposition 5 tells us that the difference between $\mathrm{Wr}(X_0)$ and $\mathrm{Wr}(C_j)$ is the spherical area of the ribbon joining T_0 and T_j.

For each i, the contribution to the spherical area from the straight part of C_j between t_i and t_{i+1} comes from the ribbon between $T_1(t_i)$ and the portion of T_0 with $t \in (t_i + 1/j, t_{i+1} - 1/j)$. As $j \to \infty$, this area converges to the area of the ribbon between the portion of T_0 with $t \in (t_i, t_{i+1})$ and $T_1(t_i)$. This is the first term in our sum above.

At each vertex t_i of X_1, the contribution to our spherical area from the curved part of C_j comes from the ribbon between the great circle arc connecting $T_1(t_{i-1})$ and $T_1(t_i)$ and a portion of T_0 of parameter length $2/j$. As $j \to \infty$, the area of this ribbon converges to the area of the spherical triangle with vertices $T_0(t_i)$, $T_1(t_i)$, $T_1(t_{i-1})$. This is the second term in our sum above. Figure 3 shows both of these terms on the unit sphere.

We have shown that the right-hand side of the statement of this theorem is equal to the limit $\lim_{j\to\infty}(\mathrm{Wr}(C_j) - \mathrm{Wr}(X_0))$. However, by Proposition 7, $\lim_{j\to\infty} \mathrm{Wr}(C_j) = \mathrm{Wr}(X_1)$. Thus

$$\lim_{j\to\infty} \mathrm{Wr}(C_j) - \mathrm{Wr}(X_0) = \mathrm{Wr}(X_1) - \mathrm{Wr}(X_0), \tag{9}$$

which is the left-hand side in the statement of this theorem. This completes the proof. □

We now make a surprising observation: Since the tantrices of the C_j differ as curves on S^2 only in parametrization, the area between each of these curves and the tantrix of X_0 is constant. Thus, by Fuller's formula, each C_j has the same writhe! And since (by Proposition 7) these writhing numbers converge to the writhe of X_1, each $\mathrm{Wr}(C_j)$ is equal to $\mathrm{Wr}(X_1)$ as well! So we have the following corollary.

COROLLARY 10. *If C_n is a polygonal curve, and C is a smooth curve obtained by rounding off the corners of C_n under the conditions of Proposition* 7, *then*

$$\mathrm{Wr}(C_n) = \mathrm{Wr}(C). \tag{10}$$

7. Extending Fuller's formulas to polygonal curves: III. Corollary 10 allows us to close this circle of ideas by observing that we have also extended Theorem 3 to polygons (cf. Proposition 3 in Cimasoni [9]).

THEOREM 11 (Fuller's spherical area formula for polygons). *For any closed space polygon P, let A be the spherical area enclosed by the tantrix of P, where we define this tantrix by Definition* 6. *Then*

$$1 + \mathrm{Wr}(P) = \frac{A}{2\pi} \mod 2. \tag{11}$$

Proof. Round off the corners of P under the conditions of Proposition 7 to obtain a C^3 curve C. The tantrices of C and of P have the same spherical image, and so enclose the same area, while C and P have the same writhe by Corollary 10. But the classical form of Fuller's spherical area formula applies to C. □

Our proof of Fuller's difference of writhe formula above also applies to a pair of polygons (one simply approximates each by smooth curves and constructs a three-stage homotopy between pairs of smooth curves), under somewhat stricter hypotheses, as seen in the following theorem.

THEOREM 12. *Suppose X_0 and X_1 are polygons, and let $F : S^1 \times [0,1] \to \mathbf{R}^3$ be a C^0 deformation of X_0 into X_1, where $F(t,\lambda) = X_\lambda(t)$, so that the X_λ are simple curves of class C^1 for $\lambda \in (0,1)$, with unit tangent vectors $T_\lambda(t)$ continuous in (t,λ).*

If each corner angle of X_i is strictly greater than $\pi/4$ and each $T_i(t)$ and $T_\lambda(t)$ are at an angle less than $\pi/2$, then

$$\mathrm{Wr}(X_1) - \mathrm{Wr}(X_0) = \frac{1}{2\pi} \operatorname{Area} R(T_0, T_1),$$

where $R(T_0, T_1)$ is the spherical region bounded by the tantrices of X_0 and X_1 (using Definition 6*) and great circle segments joining their endpoints, and* Area *represents oriented area on S^2.*

8. Bounding the ΔWr formula. We now prove our main theorem by finding asymptotic bounds for Fuller's ΔWr formula. Our theorem deals with the following situation: Assume that $C(t)$ is a simple closed curve of class $\mathcal{C}^4$, parametrized so that $|C'(t)| \geq 1$. (Given any initial parametrization, this can be accomplished by rescaling.) Further, assume we have upper bounds $B_1, \ldots, B_4$ on the norms of the first four derivatives of C. In particular, we do not require that C be parametrized by arclength and state our bounds in terms of curvature and torsion because in practice it is very difficult to obtain an arc-length parametrization of a given curve, while it is comparatively easy to obtain values for these derivative bounds.

Let $C_n(t)$ be any n-edge polygonal curve inscribed in C. We assume that the maximum edge length of C is bounded by x.

THEOREM 13. *If the ribbon formed by joining $C_n(t)$ to $C(t)$ for every t is embedded and $1/x > 5B_2$,*

$$|\operatorname{Wr}(C) - \operatorname{Wr}(C_n)| < \alpha\, n x^3 + nO(x^4), \tag{12}$$

where α is a numerical constant less than $B_2(5B_2^2 + B_3)$.

We make a few comments on this theorem before diving into the proof. First, we observe that if the lengths of the edges of C_n are all of the same order of magnitude, the difference between the writhe of C and the writhe of C_n is of order $1/n^2$.

Next, we discuss the role of the additional hypotheses in the statement above, that is, that the ribbon between C and C_n be embedded and that $1/x$ be greater than $5B_2$. Both are intended to exert enough control over the approximation to guarantee the existence of a homotopy from C to C_n obeying the requirements of Theorem 8.

We can guarantee that C_n satisfies the first hypothesis by proving that C_n lies in an embedded tubular neighborhood of C. Since C is of class $\mathcal{C}^4$, and has no self-intersections, such a neighborhood is guaranteed to exist: For a discussion of how to compute the radius of this tube (which is known as the *thickness* of C), see the literature on *ropelength* of knots (e.g., [12, 7, 14]).

Proof. We begin by reparametrizing our curve by arclength. This forces us to recompute our bounds for the derivatives of $C(t)$ (a standard computation), arriving at

$$|C'(s)| = 1, \qquad |C''(s)| < K := 2B_2, \qquad |C'''(s)| < T := 2B_3 + 10B_2^2, \tag{13}$$

while $C^{(4)}(s)$ is again bounded above. To remind ourselves of the connection between these bounds and the curvature and torsion of our curve, we will refer to the bound for the second derivative as K and the bound for the third derivative as T. Further, we note that the curvature $\kappa(s)$ of our curve is bounded above by K and that our hypotheses imply that $1/x > (5/2)K$.

We also establish the convention that the corners of C_n are at parameter values cyclically ordered as $s_0, \ldots, s_{n-1}, s_n = s_0$.

By smoothing the linear interpolation between C and C_n, we can construct a homotopy between C and C_n according to the conditions of Theorem 8 as long as the following hold:

1. the ribbon joining C to C_n is embedded,
2. the angle at each corner of C_n is at least $\pi/2$,
3. the angle between $T(s)$ and $T_n(s)$ is at most $\pi/2$ for any s.

Borrowing from Lemma 16 (below), we see that our assumption that $1/x > (5/2)K$ is enough to bound the angle in condition 3 by $0.20402 < \pi/4$. At any corner s_i, the same lemma implies that the corner angle is the supplement of at most twice 0.20402, so this is enough to ensure that condition 2 holds as well.

Theorem 8 now tells us that

$$\begin{aligned} |\operatorname{Wr}(C) - \operatorname{Wr}(C_n)| \leq \frac{1}{2\pi} \sum_{i=1}^{n} &|\operatorname{Area} R\left(T\left(s_i\right), T\left(s_{i+1}\right), T_n\left(s_i\right)\right)| \\ &+ \left|\operatorname{Area} \triangle T\left(s_i\right) T_n\left(s_{i-1}\right) T_n\left(s_i\right)\right|, \end{aligned} \tag{14}$$

where the first term is the area of the spherical region bounded by the geodesics from $T_n(s_i)$ to $T(s_i)$ and $T(s_{i+1})$ and the portion of T between s_i and s_{i+1}, and the second term is the area of the spherical triangle. Our job now is to estimate the areas of these regions. To do so, we first recall Taylor's theorem.

THEOREM 14 (Taylor's theorem). *Suppose $C(s)$ is a curve of class $\mathcal{C}^4$, with fourth derivative bounded by B_4'. Then (choosing coordinates so that $C(0)$ is at the origin),*

$$C(s) = sC'(0) + \frac{s^2}{2}C''(0) + \frac{s^3}{6}C'''(0) + R_4(s), \tag{15}$$

where $|R_4(s)| < s^4 B_4'$.

We will use this expression for $C(s)$ frequently in our work below.

LEMMA 15. *For any s, we have*

$$|s - |C(s)|| < \frac{K^2}{24}|s^3| + \frac{1}{120}|s^5| \qquad \text{and} \qquad |C(s)| \le |s|. \tag{16}$$

Further, for any edge of C_n, the difference $|s_{i+1} - s_i|$ is at most $1.01\,x$.

Proof. We assume without loss of generality that s is positive. By Schur's lemma [8], since the curvature of C is bounded above by K, $|C(x)|$ is at least the length of a chord across an arc of length s on a circle of radius $1/K$, or $(2/K)\sin(K/2)s$. This means that we have

$$\frac{2}{K}\sin\frac{K}{2}s = s - \frac{K^2}{24}s^3 + R_5(s),$$

where $R_5(s)$ is the term of order s^5 which comes from the usual Taylor expansion of $\sin s$. In particular,

$$\begin{aligned} |s - |C(s)|| &< s - \frac{2}{K}\sin\frac{K}{2}s \\ &< \frac{K^2}{24}s^3 - R_5(s), \end{aligned}$$

where $R_5(s) < \frac{1}{120}s^5$. The upper bound on $|C(s)|$ comes from the fact that C is unit-speed.

The second statement is another Schur's lemma calculation, this time invoking our hypothesis that $x > (5/2)K$ and observing that $1.01\sin y > y$ for y between 0 and $1/5$. □

We will also need an upper bound on the angle between $T(s)$ and $C_n(s)$.

LEMMA 16. *The angle between the tangent vector $T(s)$ and the corresponding tangent vector $T_n(s)$ to C_n is bounded above by*

$$\angle T(s)T_n(s) < 0.51005Kx. \tag{17}$$

Proof. Assume that s is between s_i and s_{i+1}. Then

$$\sin\angle T(s)T_n(s) = \frac{|[C(s_{i+1}) - C(s_i)] \times T(s)|}{|C(s_{i+1}) - C(s_i)|}. \tag{18}$$

But we have

$$C(s_{i+1}) - C(s_i) = \int_{s_i}^{s_{i+1}} T(t)\,dt,$$

and for any t, we have

$$T(t) = T(s) + \int_s^t T'(u)\,du.$$

This means that

$$[C(s_{i+1}) - C(s_i)] \times T(s) = \int_{s_i}^{s_{i+1}} T(t) \times T(s)\, dt \tag{19}$$

$$= \int_{s_i}^{s_{i+1}} \int_{s}^{t} T'(u) \times T(s)\, du\, dt. \tag{20}$$

Since $|T'(u) \times T(s)| \leq |T'(u)||T(s)| \leq \kappa(u) < K$, and s is between s_i and s_{i+1}, a small computation reveals that this integral is bounded by $\frac{K}{2}(s_{i+1} - s_i)^2$.

Since the length $|C(s_{i+1}) - C(s_i)|$ is bounded below by $(1/1.01)(s_{i+1} - s_i)$ by Lemma 15, we get

$$\sin \angle T(t)T_n(t) < \frac{1.01}{2} Kx. \tag{21}$$

Since $1/x > (5/2)K$, this is always bounded above by $1.01/5$, and so

$$\angle T(t)T_n(t) < \frac{1.01^2}{2} Kx. \qquad \square \tag{22}$$

We are now ready to embark on the main work of the proof: estimating the areas in (14). We begin with the first term: the area bounded by the portion of $T(s)$ between s_i and s_{i+1}, together with the great circle arcs joining $T(s_i)$ and $T(s_{i+1})$ to $T_n(s_i)$. Without loss of generality, we may assume that $i = 0$, that $s_0 = 0$, and that $C(0) = \mathbf{0}$, and apply the Taylor expansion of (15) to C at 0. Our strategy is to prove that this region is contained in a neighborhood of the great circle arc joining $T(0)$ and $T(s_1)$. Suppose s is between 0 and s_1. We want to bound the height of $T(s)$ above the $T(0)$, $T(s_1)$ plane, or

$$h(s) := \frac{C'(s) \cdot C'(0) \times C'(s_1)}{|C'(0) \times C'(s_1)|}. \tag{23}$$

First, we have

$$C'(s_1) = C'(0) + s_1 C''(0) + \frac{s_1^2}{2} C'''(0) + R_3(s_1),$$
$$C'(s) = C'(0) + sC''(0) + \frac{s^2}{2} C'''(0) + R_3(s).$$

Using the triple product identities, we can rewrite $h(s)$ in terms of the inner product of $C'(0)$ and the cross product of these vectors. For the triple product, we get

$$\left[\frac{s_1^2 s}{2} - \frac{s^2 s_1}{2}\right] C'(0) \cdot C''(0) \times C'''(0) + C'(0) \cdot [R_3(s_1) \times C'(s) + C'(s_1) \times R_3(s)]\,. \tag{24}$$

Expanding the last term, we see that it is the sum of a term of order $s_1 s^3$ and a term of order $s s_1^3$. Thus, to leading order, the norm of the entire triple product is bounded above by

$$|h(s)| < H := \frac{KT}{2|C'(0) \times C'(s_1)|} s_1^3 + O(s_1^4), \tag{25}$$

since $s \in [0, s_1]$. We now consider the height of $T_n(0)$ above the $T(0)$, $T(s_1)$ plane. Since $T_n(0)$ is the normalization of $C(s_1) - C(0) = C(s_1)$, this height is given by

$$\frac{C(s_1)}{|C(s_1)|} \cdot \frac{C'(0) \times C'(s_1)}{|C'(0) \times C'(s_1)|}. \tag{26}$$

As before, we get

$$C'(0) \times C'(s_1) = s_1 C'(0) \times C''(0) + \frac{s_1^2}{2} C'(0) \times C'''(0) + O(s_1^3). \tag{27}$$

Taking the dot product with the Taylor expansion of $C(s_1)$, we get only terms of order $O(s_1^4)$ and higher. Thus, to leading order, this region is contained in a rectangle based on the great circle arc joining $C'(0)$ and $C'(s_1)$ of height H. We now estimate the area of this rectangle.

First, we note that the length of the great circle joining $C'(0)$ and $C'(s_1)$ is given by the angle θ between $C'(0)$ and $C'(s_1)$. Since $s_1 < 1.01\,x$ by Lemma 15, this length is bounded above by $1.01\,Kx$, which is less than 0.404 by our hypotheses on x. Since H is small compared to s, we may assume that the entire rectangle is contained within a spherical disk of radius 0.5.

We project the rectangle to the plane by central projection: This map is increasing on lengths and areas, and increases length by at most a factor of 1.01. The area of the rectangle in the plane is overestimated by the product $1.01\,\theta H$. On the other hand, we have $|C'(0) \times C'(s_1)| = \sin\theta$. And for $\theta < 0.404$, $1.02 \sin\theta > \theta$. Keeping track of the various constants involved, and using the fact that $s_1 < 1.01\,x$ again, the area of this spherical region is overestimated by

$$\text{Area}\, R(T(s_i), T(s_{i+1}), T_n(s_i)) < KT\,x^3 + O(x^4). \tag{28}$$

We now turn to the second term in (14): the area of the spherical triangle bounded by $T(s_i)$, $T_n(s_{i-1})$, and $T_n(s_i)$. Without loss of generality we assume that $i = 1$, that $s_1 = 0$, and that $C(0) = \mathbf{0}$, and we expand C around 0 using (15). We wish to compute

$$\text{Area}\, \triangle \left(\frac{C(s_0)}{|C(s_0)|}, \frac{C(s_2)}{|C(s_2)|}, C'(0) \right) = \left| \left(\frac{C(s_0)}{|C(s_0)|} - C'(0) \right) \times \left(\frac{C(s_2)}{|C(s_2)|} - C'(0) \right) \right|. \tag{29}$$

If we factor out $1/|C(s_0)||C(s_2)|$, we are left with the norm of the cross product of two terms:

$$C(s_0) - |C(s_0)|C'(0) = (s_0 - |C(s_0)|)\, C'(0) + \frac{s_0^2}{2} C''(0) + \frac{s_0^3}{6} C'''(0) + R_4(s_0),$$
$$C(s_2) - |C(s_2)|C'(0) = (s_2 - |C(s_2)|)\, C'(0) + \frac{s_2^2}{2} C''(0) + \frac{s_2^3}{6} C'''(0) + R_4(s_2).$$

Using Lemma 15, we see that $|s - |C(s)|| < (K^2/24)s^3 + O(s^5)$, and we see that the leading term of this expression contains fifth powers of s_0 and s_2, and is bounded by

$$s_0^2 s_2^2 \left(\frac{K^3}{48} + \frac{KT}{12} \right) (s_0 + s_2). \tag{30}$$

However, we must still divide by $|C(s_0)||C(s_2)|$. By Lemma 15, we see that the ratios $s_0/|C(s_0)|$ and $s_2/|C(s_2)|$ are bounded above by 1.01. Thus, using the same lemma to

conclude that s_2 and s_0 are less than $1.01\,x$, and making a central projection argument as before, we are left with

$$\text{Area}\,\triangle(T_n(s_i), T_n(s_{i-1}), T(s_i)) < \frac{K^3 + KT}{3}x^3 + O(x^4). \tag{31}$$

Summing over i and dividing by 2π, then writing K and T in terms of B_2 and B_3, we obtain the statement of Theorem 13. Note that we have overestimated the numerical constants to simplify the resulting formula. □

If a curve has a small region of high curvature and larger regions of low curvature, it may be desirable to approximate the curve more carefully in the regions of high curvature in order to save time in the computation of writhe. Since our error bound is additive along the curve, these methods are well suited to this case. We have the following corollary.

COROLLARY 17. *Suppose C is a $\mathcal{C}^4$ curve and C_n is a curve inscribed in C so that C and C_n obey the hypotheses of Theorem* 13.

If C and C_n are divided into regions R_i, each containing n_i edges which are bounded above in length by x_i, and so that the derivatives of C are bounded by $B_{1i}, \ldots, B_{4i}$ and $1/x_i > 5B_{2i}$, then

$$|\,\mathrm{Wr}(C) - \mathrm{Wr}(C_p)| < \sum_i \alpha_i\, n_i x_i^3 + n_i O(x_i^4),$$

where each α_i is a numerical constant less than $B_{2i}(5B_{2i}^2 + B_{3i})$.

We make one more observation.

PROPOSITION 18. *Let C be a simple, closed space curve of class $\mathcal{C}^2$ and let C_p be a polygonal approximating curve as in Theorem* 13 *or Corollary* 17.

If the arc joining the endpoints of a sequence of n edges of C_p is planar, then the $n-2$ edges interior to this region contribute nothing to the error bound in the theorem.

In particular, this means that the derivative bounds in both statements can be taken to be bounds on the derivatives of the nonplanar *regions of the curve C.*

Proof. On these edges, the tantrix of the smooth curve and the polygonal curve parametrize the same great circle arc on S^2. Thus, the ribbon between these curves has zero area. □

9. Example computations. We are now prepared to test Theorem 1 by computing the writhing numbers of various polygonal approximations of a smooth curve and comparing the results to the exact writhe of the smooth curve. To control the numerical error introduced in these calculations, all of these computations were performed using an arbitrary-precision implementation of Banchoff's formula [3] for the writhing number of a polygonal curve. The initial runs were performed with 45 decimal digits of precision. They were checked against runs performed with 54 digits of precision. Since the results agreed, we feel confident that roundoff error does not affect the computations reported on below.

The curve whose writhe we computed is an example of Fuller [11] (see Figure 4).

Using Theorem 3 and the Călugăreanu–White formula, it is easy to see that the writhe of this curve is $3(1-\sin 0.33) \simeq 2.0278709$. After all, the area enclosed by the tantrix of this curve C is that of a hemisphere, plus three enclosures of a spherical cap of radius $\pi/2 - 0.33$. Thus the writhe of the curve is equal to $1 - \sin 0.33 \mod 2$. To complete the computation, one sets up a frame on the curve and computes its twist and linking number. (Details for this computation can be found in [11].)

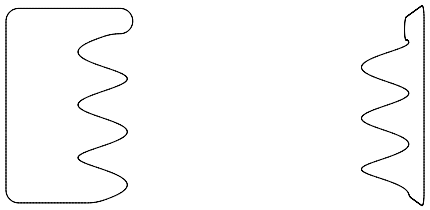


FIG. 4. *This example of Fuller's "closed helix" is composed of* 3 *turns of a helix of radius* 1 *with pitch angle* 0.33*, with ends joined by a planar curve.*

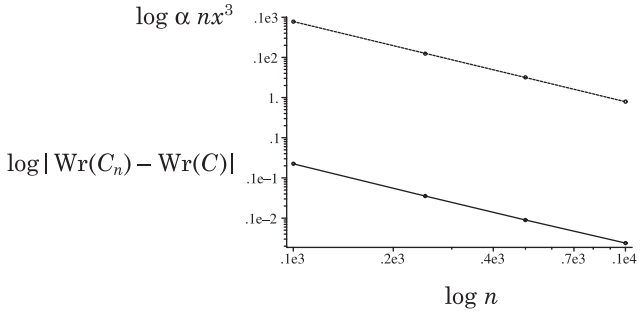


FIG. 5. *This graph shows a log-log plot of the actual error in computing the writhing number for one of Fuller's "closed helices" with various numbers of edges (lower solid line), together with our error bounds (upper dotted line). The fact that the lines are parallel shows that the convergence is of order* n^2*, as predicted by Theorem* 13*.*

We now take a series of polygonal approximations to C and compare the difference between their writhing numbers and the writhe of C to the bounds of Theorem 13.

We begin by finding bounds on the derivatives of C and the edge length of our approximations. By Proposition 18, it suffices to find derivative bounds for the helical region of C. Since the helix has unit radius, both B_2 and B_3 can be taken to be one. The curve is parametrized so that $|C'(s)| \geq 1$, and we can take $\alpha = 6$.

Here are the results of computing writhe with various numbers of edges:

n	$\mathrm{Wr}(C_n)$	$\lvert\mathrm{Wr}(C_n) - \mathrm{Wr}(C)\rvert$	x	$\alpha\, nx^3$
100	2.00541	0.02246	0.506	77.73
250	2.02434	0.00353	0.203	12.55
500	2.02697	0.0009	0.101	3.09
1000	2.02763	0.00024	0.051	0.786

It is worth examing a graph of these results (see Figure 5).

10. Discussion of results and further directions. In this paper, we have given a set of asymptotic error bounds that allow us to compute the writhe of a closed space curve with defined accuracy by computing the writhe of a polygonal approximation to this curve. The example we computed in section 9 shows that our bounds are of the right order of magnitude: Roughly speaking, the writhe converges quadratically in the number of edges of the approximation.

We have now built a new family of algorithms for computing the writhe integral with bounded error: Estimate the number of edges required using Theorem 1 and then compute the writhe of that inscribed polygon using any of the methods mentioned in the introduction.

To test one of these algorithms in practice, we implemented the most basic of the polygonal writhe computation algorithms (Banchoff's exact sum formula [3]) and tested it against the well-developed adaptive quadrature routine `CUHRE` [4]. Our tests reveal that our simple method does not yet outperform `CUHRE`, but the routines can be comparable. For instance, on a sample computation of Fuller's closed helix, `CUHRE` made 11895 evaluations of the writhe integrand for an actual error of 0.0014732694, while our algorithm computed 44850 terms in an exact sum for an actual error of 0.00150432 (the error estimates for each algorithm were 0.2052 and 0.5121, respectively). Since the more advanced algorithms for computing the writhe of polygonal curves are often orders of magnitude faster than Banchoff's algorithm, this holds out the hope that a better method based on our theorem may improve significantly on adaptive quadrature. Clearly, more work is required in this area.

We have proved versions of Fuller's spherical area formula (Theorem 11) and Fuller's difference of writhe formula (Theorem 12). But it is puzzling that our approximation theorem for curves with corners (Proposition 7) should depend on the hypothesis that the limit curve is polygonal.

We suspect that the following conjecture holds.

CONJECTURE 19. *Fuller's spherical area formula (Theorem* 3*) and Fuller's* $\Delta\mathrm{Wr}$ *formula (Theorem* 4*) hold for piecewise* C^2 *curves with the extended definition of tantrix given by Definition* 6.

The proofs of both of these theorems depend on the Călugăreanu–White formula, which applies only to closed curves. Thus all of our results are restricted to closed curves. This leaves open the following much more important problem.

PROBLEM 20. *Extend all these theorems (the Călugăreanu–White formula, Fuller's spherical area formula, and Fuller's* $\Delta\mathrm{Wr}$ *formula) to open curves.*

In particular, extending the results of this paper to open curves would be useful for applications in biology, where the curves of interest are not necessarily closed. We note that while Fuller's ΔWr formula makes sense for open curves, computational examples show that it does not give the correct answer: Boundary terms must be added to account for the ends of the curves.

Acknowledgments. I am grateful to Herbert Edelsbrunner, Herman Gluck, Issac Klapper, John Maddocks, Kathleen Rogers, and David Swigon, among many others, for fruitful conversations. I would also like to acknowledge Mark Peletier and Robert Planque, whose insightful questions and perceptive observations have helped me rethink the writhing number.

REFERENCES

[1] P. K. AGARWAL, H. EDELSBRUNNER, AND Y. WANG, *Computing the writhing number of a polygonal knot*, Discrete Comput. Geom., to appear.

[2] J. ALDINGER, I. KLAPPER, AND M. TABOR. *Formulae for the calculation and estimation of writhe*, J. Knot Theory Ramifications, 4 (1995), pp. 343–372.

[3] T. BANCHOFF, *Self linking numbers of space polygons*, Indiana Univ. Math. J., 25 (1976), pp. 1171–1188.

[4] J. BERNTSEN, T. O. ESPELID, AND A. GENZ, *Algorithm* 698*: DCUHRE: An adaptive multidimensional integration routine for a vector of integrals*, ACM Trans. Math. Software, 17 (1991), pp. 452–456.

[5] G. Călugăreanu, *L'intégrale de Gauss et l'analyse des nœuds tridimensionnels*, Rev. Math. Pures Appl., 4 (1959), pp. 5–20.
[6] G. Călugăreanu, *Sur les classes d'isotopie des nœuds tridimensionnels et leurs invariants*, Czechoslovak Math. J., 11 (1961), pp. 588–625.
[7] J. Cantarella, R. B. Kusner, and J. M. Sullivan, *On the minimum ropelength of knots and links*, Invent. Math., 150 (2002), pp. 257–286.
[8] S. S. Chern, *Curves and surfaces in Euclidean space*, in Studies in Global Geometry and Analysis, S. S. Chern, ed., Math. Assoc. Amer., 1967, pp. 16–56.
[9] D. Cimasoni, *Computing the writhe of a knot*, J. Knot Theory Ramifications, 10 (2001), pp. 387–395.
[10] M. Epple, *Orbits of asteroids, a braid, and the first link invariant*, Math. Intelligencer, 20 (1998), pp. 45–52.
[11] F. Brock Fuller, *Decomposition of the linking number of a closed ribbon: A problem from molecular biology*, Proc. Natl. Acad. Sci. USA, 75 (1978), pp. 3557–3561.
[12] O. Gonzalez and J. H. Maddocks, *Global curvature, thickness, and the ideal shapes of knots*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 4769–4773.
[13] K. Klenin and J. Langowski, *Computation of writhe in modeling of supercoiled DNA*, Biopolymers, 54 (2000), pp. 307–317.
[14] R. A. Litherland, J. Simon, O. Durumeric, and E. Rawdon, *Thickness of knots*, Topology Appl., 91 (1999), pp. 233–244.
[15] W. F. Pohl, *The self-linking number of a closed space curve*, J. Math. Mech., 17 (1967/1968), pp. 975–985.
[16] P. Rogen and B. Fain, *Automatic classification of protein structure by using Gauss integrals*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 119–124.
[17] H. L. Royden, *Real Analysis*, Macmillan, New York, 1963.
[18] B. Solomon, *Tantrices of spherical curves*, Amer. Math. Monthly, 103 (1996), pp. 30–39.
[19] D. W. Sumners, *Lifting the curtain: Using topology to probe the hidden action of enzymes*, Match, 34 (1996), pp. 51–76.
[20] J. M. T. Thompson, G. H. M. van der Heijden, and S. Neukirch, *Supercoiling of DNA plasmids: Mechanics of the generalized ply*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 959–985.
[21] J. White, *Self-linking and the Gauss integral in higher dimensions*, Amer. J. Math., 91 (1969), pp. 693–728.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1862–1874

AN EXPLICIT FINITE DIFFERENCE METHOD AND A NEW VON NEUMANN-TYPE STABILITY ANALYSIS FOR FRACTIONAL DIFFUSION EQUATIONS*

S. B. YUSTE† AND L. ACEDO†

Abstract. A numerical method for solving the fractional diffusion equation, which could also be easily extended to other fractional partial differential equations, is considered. In this paper we combine the forward time centered space (FTCS) method, well known for the numerical integration of ordinary diffusion equations, with the Grünwald–Letnikov discretization of the Riemann–Liouville derivative to obtain an explicit FTCS scheme for solving the fractional diffusion equation. The stability analysis of this scheme is carried out by means of a powerful and simple new procedure close to the well-known von Neumann method for nonfractional partial differential equations. The analytical stability bounds are in excellent agreement with numerical test. A comparison between exact analytical solutions and numerical predictions is made.

Key words. fractional diffusion equation, von Neumann stability analysis, parabolic integro-differential equations, finite difference methods

AMS subject classifications. 26A33, 74S20, 45K05, 60J60

DOI. 10.1137/030602666

1. Introduction.

The study of fractional differential equations has been a highly specialized and isolated field of mathematics for many years [1]. However, in the last decade there has been increasing interest in the description of physical and chemical processes by means of equations involving fractional derivatives and integrals. This mathematical technique has a broad potential range of application [2]: relaxation in polymer systems, dynamics of protein molecules, and the diffusion of contaminants in complex geological formations are some of the most recently suggested applications [3].

Fractional kinetic equations have proved particularly useful in the context of anomalous slow diffusion (subdiffusion) [4]. Anomalous diffusion is characterized by an asymptotic long-time behavior of the mean square displacement of the form

$$\left\langle x^2(t)\right\rangle \sim \frac{2K_\gamma}{\Gamma(1+\gamma)}t^\gamma, \quad t\to\infty, \tag{1.1}$$

where γ is the anomalous diffusion exponent. The process is usually referred to as subdiffusive when $0<\gamma<1$. Ordinary (or Brownian) diffusion corresponds to $\gamma=1$ with $K_1=D$ (the diffusion coefficient). From a continuous point of view, the diffusion process is described by the diffusion equation $u_t(x,t)=D\,u_{xx}(x,t)$, where $u(x,t)$ represents the probability density of finding a "particle" at x at time t, and where $u_{\eta\zeta\ldots}$ is the partial derivative with respect to the variables η,$\zeta\ldots$. It turns out that the probability density function $u(x,t)$ that describes anomalous (sub)diffusive particles follows the fractional diffusion equation [4, 5, 6, 7]:

*Received by the editors April 1, 2004; accepted for publication (in revised form) May 19, 2004; published electronically January 20, 2005. This research was partially supported by the Ministerio de Ciencia y Tecnología (Spain) through grant FIS2004-01399 and by the European Community's Human Potential Programme under contract HPRN-CT-2002-00307, DYGLAGEMEM.

http://www.siam.org/journals/sinum/42-5/60266.html

†Departamento de Física, Universidad de Extremadura, E-06071 Badajoz, Spain (santos@unex.es, acedo@unex.es).

$$\frac{\partial}{\partial t}u(x,t) = K_\gamma\, {}_0D_t^{1-\gamma}\frac{\partial^2}{\partial x^2}u(x,t), \quad t \geq 0, \tag{1.2}$$

where ${}_0D_t^{1-\gamma}$ is the fractional derivative defined through the Riemann–Liouville operator (see section 2). Fractional subdiffusion-advection equations, and fractional Fokker–Planck equations, have also been proposed [8, 9, 10, 11], and even subdiffusion-limited reactions have been discussed within this framework [12, 13]. These equations are also referred to as parabolic integrodifferential equations with weakly singular kernels [14].

These current applications of fractional differential equations, and many others that may well be devised in the near future, make it imperative to search for methods of solution. Some exact analytical solutions for a few cases, although important, are already known in terms of special functions such as the Wright function and Fox's H-function [6, 7, 15, 16]. Some of these results have been obtained by means of the Mellin transform [6, 7] and the method of images [16]. The powerful method of separation of variables can also be applied to fractional equations in the same way as for the usual diffusion equations (an example is given in section 4). Another route to solving fractional equations is through the integration of the product of the solution of the corresponding nonfractional equation (the Brownian counterpart obtained by setting $\gamma \to 1$) and a one-sided Lévy stable density [4, 17]. However, as also for the Brownian case, the availability of numerical methods for solving (1.2) would be most desirable, especially for those cases where no analytical solution is available. One possibility was discussed recently by Gorenflo and Mainardi [18], Gorenflo, De Fabritiis, and Mainardi [19], and Gorenflo et al. [20], who presented a scheme for building discrete models of random walks suitable for the Monte Carlo simulation of random variables with a probability density governed by fractional diffusion equations. Another, more standard, approach is to build difference schemes of the type used for solving Volterra-type integrodifferential equations [14]. Along this line, some implicit (*backward* Euler and Crank–Nicholson) methods have been proposed [14, 21, 22, 23, 24, 25].

In this paper we shall use the *forward* Euler difference formula for the time derivative $\partial u/\partial t$ in (1.2) to build an *explicit* method that we will call the fractional forward time centered space (FTCS) method. For Brownian ($\gamma = 1$) diffusion equations, this explicit procedure is the simplest numerical methods workhorse [26, 27]. However, for fractional diffusion equations, this explicit method has been overlooked, perhaps because of the difficulty in finding the conditions under which the procedure is stable. This problem is solved here by means of a new stability analysis procedure close to the usual Fourier–von Neumann method for nonfractional partial differential equations.

The plan of the paper is as follows. In section 2 we give a short introduction to some results and definitions in fractional calculus. The numerical procedure for solving the fractional diffusion equation (1.2) by means of the explicit FTCS method is given in section 3. In this section we also discuss the stability and the truncating errors of the FTCS scheme. In section 4 we compare exact analytical solutions with numerical ones and check the reliability of the analytical stability condition. Some concluding remarks are given in section 5.

2. Basic concepts of fractional calculus. The notion of fractional calculus was anticipated by Leibniz, one of the founders of standard calculus, in a letter written in 1695 [1, 4]. But it was in the next two centuries that this subject fully developed into a field of mathematics with the work of Laplace, Cayley, Riemann, Liouville, and many others.

There are two alternative definitions for the fractional derivative ${}_0D_t^{1-\gamma}$ of a function $f(t)$. On the one hand, there is the Riemann–Liouville operator definition

$$ {}_0D_t^{1-\gamma} f(t) = \frac{1}{\Gamma(\gamma)} \frac{\partial}{\partial t} \int_0^t d\tau \frac{f(\tau)}{(t-\tau)^{1-\gamma}} , \tag{2.1} $$

with $0 < \gamma < 1$. For $\gamma = 1$ one recovers the identity operator and for $\gamma = 0$ the ordinary first-order derivative. On the other hand, the fractional derivative of order $1-\gamma$ of a function $f(t)$ in the Grünwald–Letnikov form is

$$ {}_0D_t^{1-\gamma} f(t) = \lim_{h\to 0} \frac{1}{h^{(1-\gamma)}} \sum_{k=0}^{[t/h]} \omega_k^{(1-\gamma)} f(t-kh), \quad t \geq 0, \tag{2.2} $$

where $[t/h]$ means the integer part of t/h and $\omega_k^{(1-\gamma)} = (-1)^k \binom{1-\gamma}{k}$. The Grünwald–Letnikov definition is simply a generalization of the ordinary discretization formulas for integer order derivatives [1]. The Riemann–Liouville and the Grünwald–Letnikov approaches coincide under relatively weak conditions: if $f(t)$ is continuous and $f'(t)$ is integrable in the interval $[0,t]$, then for every order $0 < 1-\gamma < 1$ both the Riemann–Liouville and the Grünwald–Letnikov derivatives exist and coincide for any time inside the interval $[0,t]$ [1, sect. 2.3.7]. This theorem of fractional calculus ensures the consistency of both definitions for most physical applications, where the functions are expected to be sufficiently smooth.

The Grünwald–Letnikov definition is important for our purposes because it allows us to estimate ${}_0D_t^{1-\gamma} f(t)$ numerically in a simple and efficient way:

$$ {}_0D_t^{1-\gamma} f(t) = \frac{1}{h^{(1-\gamma)}} \sum_{k=0}^{[t/h]} \omega_k^{(1-\gamma)} f(t-kh) + O(h^p) . \tag{2.3} $$

This formula is not unique because there are many different valid choices for $\omega_k^{(\alpha)}$ that lead to approximations of different order p [28]. Let $\omega(z,\alpha)$ be the generating function of the coefficients $\omega_k^{(\alpha)}$, i.e.,

$$ \omega(z,\alpha) = \sum_{k=0}^{\infty} \omega_k^{(\alpha)} z^k. \tag{2.4} $$

If the generating function is

$$ \omega(z,\alpha) = (1-z)^\alpha, \tag{2.5} $$

then we get the backward difference formula of order $p = 1$ (BDF1) [1, 28]. This is also called the backward Euler formula of order 1 or, simply, the Grünwald–Letnikov formula. These coefficients are $\omega_k^{(\alpha)} = (-1)^k \binom{\alpha}{k}$ and can be evaluated recursively:

$$ \omega_0^{(\alpha)} = 1, \qquad \omega_k^{(\alpha)} = \left(1 - \frac{\alpha+1}{k}\right) \omega_{k-1}^{(\alpha)}. \tag{2.6} $$

The generating function for the backward difference formula of order $p = 2$ (BDF2) is [1, 28]

$$ \omega(z,\alpha) = \left(\frac{3}{2} - 2z + \frac{1}{2}z^2\right)^\alpha . \tag{2.7} $$

These coefficients can be easily calculated using fast Fourier transforms [1]. However, for the fractional FTCS method discussed in this paper, we will show in the next section that nothing is gained by using second-order approximations for the fractional derivative. Additionally, the stability bound is smaller if one uses the BDF2 formula. Finally, it is important to note that the error estimates given in (2.3) are valid only if either $t/h \gg 1$ [1] or $u(x,t)$ is sufficiently smooth at the time origin $t = 0$ [29].

3. Fractional FTCS method. We will use the customary notation $x_j = j\Delta x$, $t_m = m\Delta t$, and $u(x_j, t_m) \equiv u_j^{(m)} \simeq U_j^{(m)}$, where $U_j^{(m)}$ stands for the numerical estimate of the exact value of $u(x,t)$ at the point (x_j, t_m). In the usual FTCS method, the diffusion equation is replaced with a difference recurrence system for the quantities $u_j^{(m)}$:

$$\frac{u_j^{(m+1)} - u_j^{(m)}}{\Delta t} = D\,\frac{u_{j-1}^{(m)} - 2u_j^{(m)} + u_{j+1}^{(m)}}{(\Delta x)^2} + T(x,t), \tag{3.1}$$

with $T(x,t)$ being the truncation term [26]. In the same way, the fractional equation is replaced with

$$\frac{u_j^{(m+1)} - u_j^{(m)}}{\Delta t} = K_\gamma\ {}_0D_t^{1-\gamma}\frac{u_{j-1}^{(m)} - 2u_j^{(m)} + u_{j+1}^{(m)}}{(\Delta x)^2} + T(x,t). \tag{3.2}$$

The estimate of the truncation term will be given in section 3.2. Inserting the Grünwald–Letnikov definition of the fractional derivative given in (2.3) into (3.2), neglecting the truncation term, and rearranging the terms, we finally get the explicit FTCS difference scheme

$$U_j^{(m+1)} = U_j^{(m)} + S_\gamma \sum_{k=0}^{m} \omega_k^{(1-\gamma)} \left[U_{j-1}^{(m-k)} - 2U_j^{(m-k)} + U_{j+1}^{(m-k)}\right], \tag{3.3}$$

where $S_\gamma = K_\gamma \Delta t/[h^{1-\gamma}(\Delta x)^2]$. In this scheme, $U_j^{(m+1)}$, for every position j, is given *explicitly* in terms of all the previous states $U_j^{(n)}$, $n = 0, 1, \ldots, m$. Because the estimates $U_j^{(m)}$ of $u(x_j, t_m)$ are made at the times $m\Delta t$, $m = 1, 2, \ldots$, and because the evaluation of ${}_0D_t^{1-\gamma}u(x_j,t)$ by means of (2.3) requires knowing $u(x_j,t)$ at the times nh, $n = 0, 1, 2, \ldots$, it is natural to choose $h = \Delta t$. In this case,

$$S_\gamma = K_\gamma \frac{\Delta t^\gamma}{(\Delta x)^2}. \tag{3.4}$$

We assume that the system is prepared in an initial state $u_j^{(0)} = U_j^{(0)}$ with $u_j^{(n)} = 0$ if $n \le -1$. The iteration process described by (3.3) is easily implementable as a computer algorithm, but the resulting program is far more memory hungry than the elementary Markov diffusive analogue because, in evaluating $U_j^{(m+1)}$, one has to save all the previous estimates $U_{j-1}^{(n)}$, $U_j^{(n)}$, and $U_{j+1}^{(n)}$ for $n = 0, 1, \ldots, m$. However, the use of the short-memory principle [1] or the nested mesh procedure [30] could alleviate this burden. Regardless, before tackling (3.3) seriously we must first discuss two fundamental questions concerning any integration algorithm: its stability and the magnitude of the errors committed by the replacement of the continuous equation with the discrete algorithm.

3.1. Stability of the fractional FTCS method. Here we will show that the stability of the fractional numerical schemes can be analyzed very easily and efficiently with a method close to the well-known Von Neumann (or Fourier) method of nonfractional partial differential equations. In this section, we will apply it to the fractional FTCS difference scheme (3.3).

We start by assuming a solution (a subdiffusion mode or eigenfunction) with the form $u_j^{(m)} = \zeta_m e^{iqj\Delta x}$, where q is a real spatial wave number. Inserting this expression into (3.3) one gets

$$\zeta_{m+1} = \zeta_m - 4S \sin^2\left(\frac{q\Delta x}{2}\right) \sum_{k=0}^{m} \omega_k^{(1-\gamma)} \zeta_{m-k}. \tag{3.5}$$

It is interesting to note that this equation is the discretized version of

$$\frac{d\psi(t)}{dt} = -4C \sin^2\left(\frac{q\Delta x}{2}\right) \, {}_0D_t^{1-\gamma}\psi(t) \tag{3.6}$$

(with $C = S(\Delta t)^\gamma$) whose solution can be expressed in terms of the Mittag–Leffler function $E_\gamma(-\lambda t^\gamma)$ [2, 4]. This result is not unexpected because the subdiffusion modes of (1.2) decay as Mittag–Leffler functions [4] (e.g., see (4.4)).

The stability of the solution is determined by the behavior of ζ_m. Unfortunately, solving (3.5) is much more difficult than solving the corresponding equation for the diffusive case. However, let us write

$$\zeta_{m+1} = \xi \zeta_m, \tag{3.7}$$

and let us *assume* for the moment that $\xi \equiv \xi(q)$ is independent of time. Then (3.5) implies a closed equation for the amplification factor ξ of the subdiffusion mode:

$$\xi = 1 - 4S_\gamma \sin^2\left(\frac{q\Delta x}{2}\right) \sum_{k=0}^{m} \omega_k^{(1-\gamma)} \xi^{-k} \, . \tag{3.8}$$

If $|\xi| > 1$ for some q, the temporal factor of the solution grows to infinity according to (3.7) and the mode is unstable. Considering the extreme value $\xi = -1$, we obtain from (3.8) the following stability bound on S_γ:

$$S_\gamma \sin^2\left(\frac{q\Delta x}{2}\right) \le \frac{1/2}{\sum_{k=0}^{m} (-1)^k \omega_k^{(1-\gamma)}} \equiv S_{\gamma,m}^{\times} \, . \tag{3.9}$$

The bound expressed in (3.9) depends on the number of iterations m. Nevertheless, this dependence is weak: $S_{\gamma,m}^{\times}$ approaches $S_\gamma^{\times} \equiv \lim_{m\to\infty} S_{\gamma,m}^{\times}$ in the form of oscillations with small decaying amplitudes (see Figure 3.1). The value of $S_\gamma^{\times}$ can be deduced from (3.9) by taking into account that $\sum_{k=0}^{\infty} (-1)^k \omega_k^{(1-\gamma)} \equiv \omega(-1, 1-\gamma)$ (see (2.4)). Therefore, we find that the FTCS method is stable as long as

$$S_\gamma \sin^2\left(\frac{q\Delta x}{2}\right) \le S_\gamma^{\times} \tag{3.10}$$

with

$$S_\gamma^{\times} = \frac{1}{2\omega(-1, 1-\gamma)} \, . \tag{3.11}$$

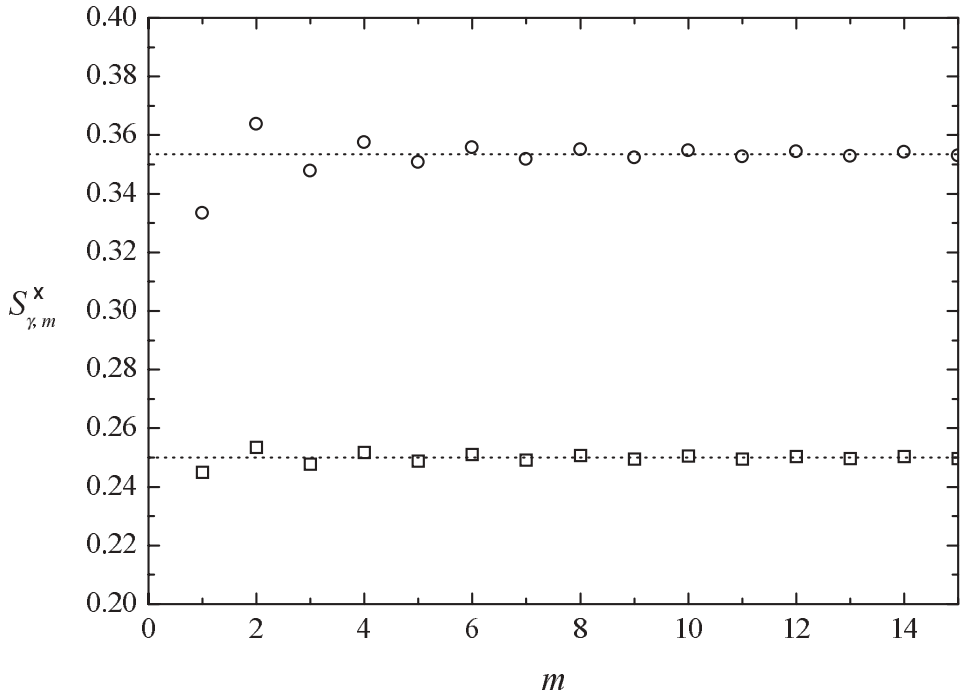


FIG. 3.1. *First values of $S_{\gamma,m}$ versus m for $\gamma = 1/2$ when the first-order coefficients (circles) and second-order coefficients (squares) are used. The lines mark the corresponding limit values $S^{\times}_{\gamma}$ given by* (3.12) *and* (3.13).

In particular, when the BDF1 coefficients given by (2.5) are used, one gets

$$S^{\times}_{\gamma} = \frac{1}{2(1-\xi)^{1-\gamma}\big|_{\xi\to -1}} = \frac{1}{2^{2-\gamma}}\ . \tag{3.12}$$

Similarly, when the BDF2 coefficients given by (2.7) are used, one gets

$$S^{\times}_{\gamma} = \frac{1}{2\left(\frac{3}{2} - 2\xi + \frac{1}{2}\xi^2\right)^{1-\gamma}\Big|_{\xi\to -1}} = \frac{1}{4^{3/2-\gamma}}. \tag{3.13}$$

We will verify numerically in section 4 that the explicit integration method as given by (3.3) is stable when

$$S_{\gamma} \le \frac{S^{\times}_{\gamma}}{\sin^2\left(\frac{q\Delta x}{2}\right)} \tag{3.14}$$

and unstable otherwise. As the maximum value of the square of the sine function is bounded by 1, we can give a more conservative but simpler bound: the fractional FTCS method will be stable when

$$S_{\gamma} = K_{\gamma}\frac{\Delta t^{\gamma}}{(\Delta x)^2} \le S^{\times}_{\gamma}. \tag{3.15}$$

The physical interpretation of this restriction is the same as for the diffusive case, namely, (3.15) means that the maximum allowed time step Δt is, up to a numerical factor, the (sub)diffusion time across a distance of length Δx (cf. (1.1)).

Notice that the value of $S_\gamma^\times = 1/4^{3/2-\gamma}$ given by (3.13) is smaller than $1/2^{2-\gamma}$ for any $\gamma < 1$ (if $\gamma = 1$, we recover the bound $S^\times = 1/2$ of the usual explicit FTCS method for the ordinary diffusion equation [26, 27]). Consequently, the fractional FTCS method that uses a second-order approximation in the fractional derivative is "less robust" than the fractional FTCS method that uses the first-order coefficients $\omega_k^{(1-\gamma)}$. Taking into account that the two methods (BDF1 and BDF2) have the same precision (see section 3.2) we note that nothing is gained by using the fractional derivative with higher precision. Therefore, in practical applications, here we will use only the first-order coefficients (2.6).

3.2. Truncating error of the fractional FTCS method. The truncating error $T(x,t)$ of the fractional FTCS difference scheme is (see (3.2))

$$T(x,t) = \frac{u_j^{(m+1)} - u_j^{(m)}}{\Delta t} - K_\gamma \, D_t^{1-\gamma} \left[\frac{u_{j-1}^{(m)} - 2u_j^{(m)} + u_{j+1}^{(m)}}{(\Delta x)^2} \right] . \tag{3.16}$$

But

$$\frac{u_j^{(m+1)} - u_j^{(m)}}{\Delta t} = u_t + \frac{1}{2} u_{tt} \Delta t + O(\Delta t)^2 \tag{3.17}$$

and

$$ {}_0D_t^{1-\gamma} \left[u_{j-1}^{(m)} - 2u_j^{(m)} + u_{j+1}^{(m)} \right] = \frac{1}{h^{1-\gamma}} \sum_{k=0}^{m} w_k^{1-\gamma} \left[u_{xx} + \frac{1}{12} u_{xxxx} \, (\Delta x)^2 + \cdots \right] + O\left(h^p\right) \tag{3.18}$$

so that, taking into account that $u(x,t)$ is the exact solution of (1.2), we finally get from (3.16), (3.17), and (3.18) the following result:

$$T(x,t) = O(h^p) + \frac{1}{2} u_{tt} \Delta t - \frac{K_\gamma (\Delta x)^2}{12} \, {}_0D_t^{1-\gamma} u_{xxxx} + \cdots \tag{3.19}$$

$$= O(h^p) + O(\Delta t) + O(\Delta x)^2 \, . \tag{3.20}$$

Therefore, (i) assuming that the initial boundary data for u are consistent (as assumed for the usual FTCS method [26]) and (ii) assuming that u is sufficiently smooth at the origin $t = 0$ (see the remark below (2.7)), we conclude that the method discussed in this paper is unconditionally consistent for any order p because $T(x,t) \to 0$ as h, Δt, $\Delta x \to 0$. As remarked above, in practical calculations it is convenient to use $h = \Delta t$ so that, due to the term $O(\Delta t)$ in (3.20), no improvements are achieved by considering higher orders than $p = 1$ in the fractional derivative. It is interesting to note that for the diffusion equation ($\gamma = 1$) it is possible to cancel out the last two terms in (3.19) with the choice $\Delta t = (\Delta x)^2/(6K_\gamma)$, thereby obtaining a scheme that is "second-order accurate" [26]. This is not possible for the fractional case because of the fractional operator.

4. Numerical solutions and the stability bound on S_γ. The objective of this section is twofold: first, we want to test the reliability of the numerical algorithm defined in (3.3) by applying it to two fractional problems with known exact solutions and, second, we want to check the stability bounds obtained in section 3.1.

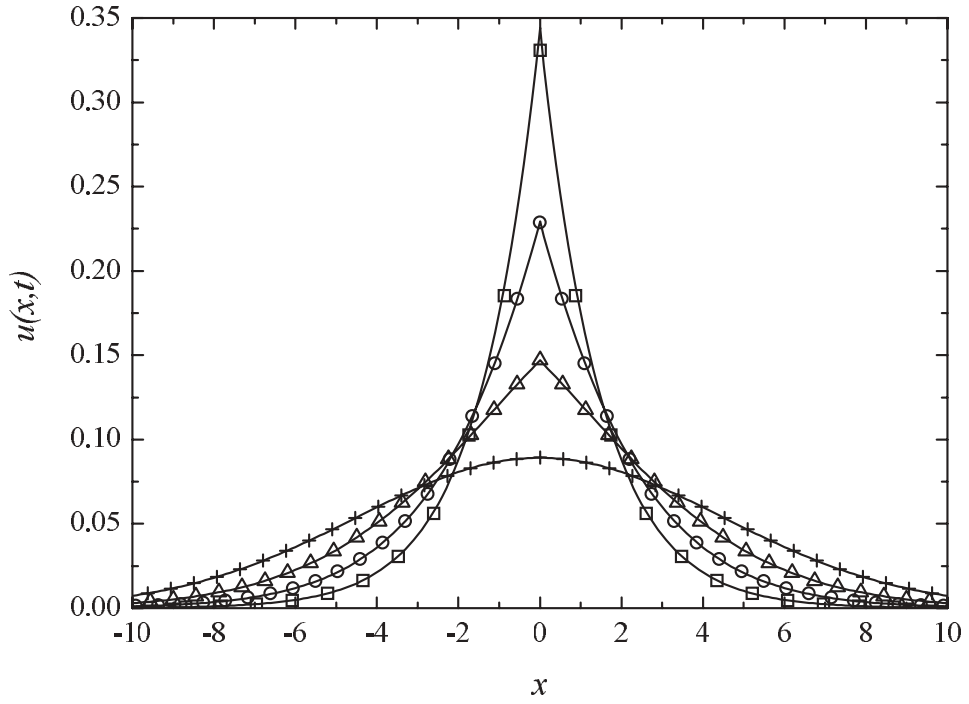


FIG. 4.1. *Numerical solution of the subdiffusion equation for the problem defined in the unbounded space,* $-\infty < x < \infty$ *with initial condition* $u(x, t = 0) = \delta(x)$ *for* $\gamma = 1/4$ *(squares),* $\gamma = 1/2$ *(circles),* $\gamma = 3/4$ *(triangles), and* $\gamma = 1$ *(crosses), and* $t = 10$. *The lines correspond to the exact analytical solution.*

4.1. Numerical solution versus exact solution: Two examples. The fundamental solution of the subdiffusion equation in (1.2) corresponds to the problem defined in the unbounded space, $-\infty < x < \infty$, where the initial condition is $u(x, t = 0) = \delta(x)$. This solution is called the propagator, or Green's function, and can be expressed in terms of Fox's H-function [4]:

$$u(x,t) = \frac{1}{\sqrt{4\pi K_\gamma t^\gamma}}\, H_{11}^{10}\left[\frac{|x|}{\sqrt{K_\gamma t^\gamma}}\,\middle|\,\begin{matrix}(1-\gamma/2,\gamma/2)\\(0,1)\end{matrix}\right]. \tag{4.1}$$

In our numerical solution we used the boundary conditions $u(-L, t) = u(L, t) = 0$ with a sufficiently large L in order to avoid finite size effects. In Figure 4.1 we compare the numerical integration results with the exact solution (4.1) for $\gamma = 1/4$, $1/2$, $3/4$, 1 at $t = 10$. The time step used was $\Delta t = 0.01$ and $\Delta x = \sqrt{K_\gamma (\Delta t)^\gamma / S_\gamma}$ with $K_\gamma = 1$ and $S_\gamma = 0.28$, 0.33, 0.4, and 0.5. All these values of S_γ are just below the stability bound $S_\gamma^\times$ (see (3.12)). The agreement is excellent except for $\gamma = 1/4$ and $x = 0$, but this minor discrepancy is surely due to the large spatial cell $\Delta x \simeq 1.06$ used in this case.

We have also considered a problem with absorbing boundaries, $u(0, t) = u(1, t) = 0$, and initial condition $u(x, t = 0) = x(1 - x)$, with $0 \le x \le 1$. The exact analytical solution of (1.2) is easily found by the method of separation of variables: $u(x, t) =$

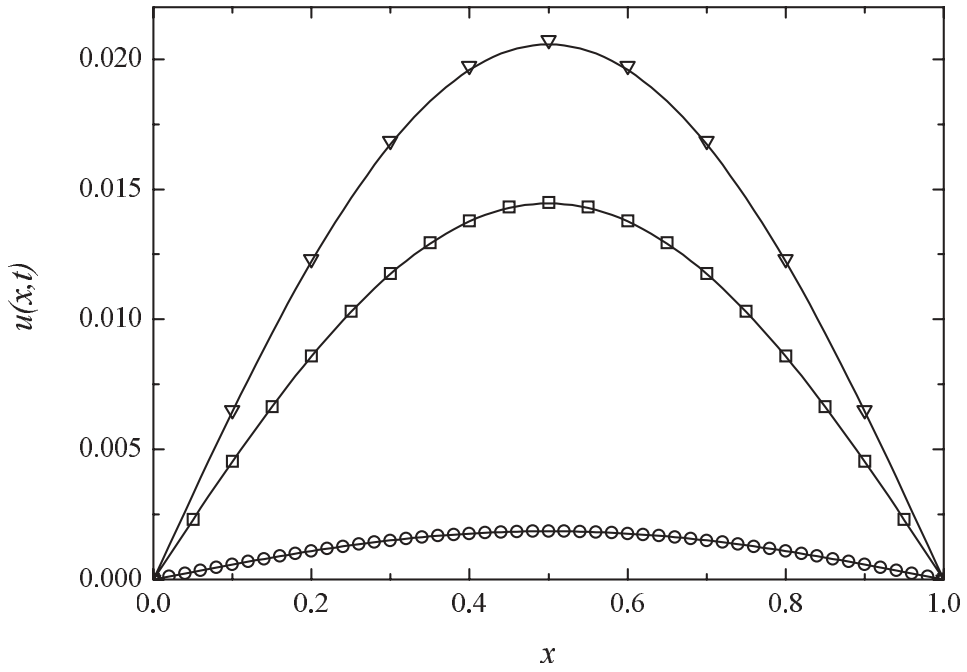


FIG. 4.2. *Numerical solution of the subdiffusion equation for the problem with absorbing boundary conditions* $u(0,t) = u(1,t) = 0$ *and initial condition* $u(x,0) = x(1-x)$, $0 \le x \le 1$, *for* $t = 0.5$. *The solution* $u(x,t)$ *is shown for* $\gamma = 0.5$ *(triangles),* $\gamma = 0.75$ *(squares), and* $\gamma = 1$ *(circles). The lines correspond to the exact analytical solution.*

$X(x)T(t)$. We thus find $X_n(x) = \sin(n\pi x)$ and

$$\frac{dT}{dt} = -K_\gamma \lambda_n^2 \, {}_0D_t^{1-\gamma} T \, , \tag{4.2}$$

where $\lambda_n = n\pi$, $n = 1, 2, \ldots$. The solution of (4.2) is found in terms of the Mittag–Leffler function [4]:

$$T_n(t) = E_\gamma(-K_\gamma n^2 \pi^2 t^\gamma) \, . \tag{4.3}$$

Imposing the initial condition we obtain

$$u(x,t) = \frac{8}{\pi^3} \sum_{n=0}^{\infty} \frac{1}{(2n+1)^3} \sin[(2n+1)\pi x] E_\gamma[-K(2n+1)^2 \pi^2 t^\gamma] \, . \tag{4.4}$$

In Figure 4.2 we compare this exact solution with the results of the numerical integration scheme for $\gamma = 0.5$, $\gamma = 0.75$, and $\gamma = 1$ for $t = 0.5$ and $K_\gamma = 1$. The values of S_γ used were $S_\gamma = 0.33$, 0.4, and 0.5 with $\Delta x = 1/10$, $1/20$, and $1/50$, respectively. The values of Δt for fixed S_γ and Δx stem from the definition of S_γ:

$$\Delta t = \left[\frac{S_\gamma (\Delta x)^2}{K_\gamma} \right]^{1/\gamma} \, . \tag{4.5}$$

Excellent agreement is observed for the three values of γ, it being slightly poorer for the smallest value, which is not surprising because in this case the mesh size $\Delta x = 1/10$ used is the largest.

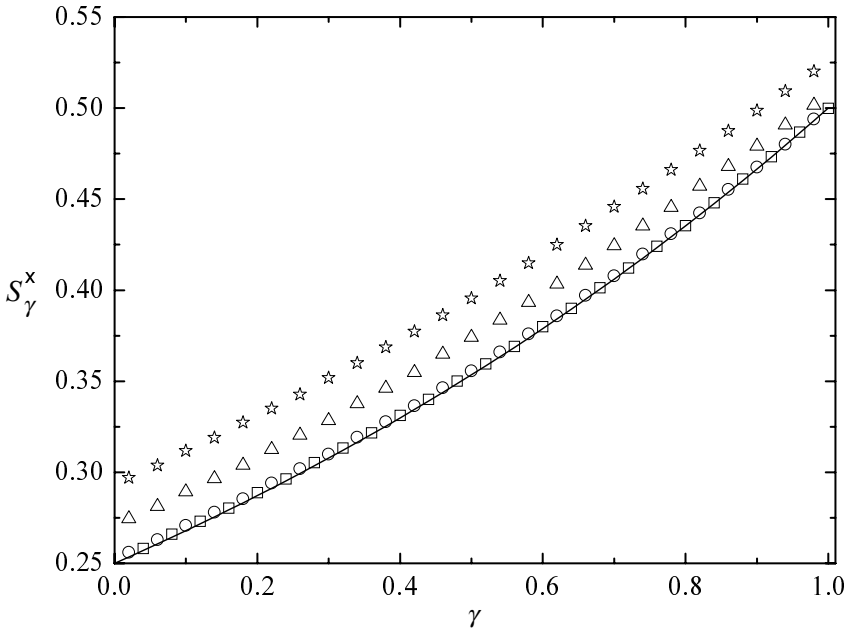


FIG. 4.3. *Values of $S_\gamma^\times$ corresponding to the onset of instability versus the subdiffusion exponent γ. The solid line is the prediction of the Fourier–von Neumann analysis and the symbols denote the results of the numerical tests with the criterion in* (4.6)*: stars, triangles, and squares for the absorbing boundary problem with $u(x,0) = x(1-x)$ with $M = 50$, 100 and 1000, respectively, and circles for the propagator with $M = 1000$.*

4.2. Numerical check of the stability analysis. We checked the stability bound on the value of the S_γ given in (3.12) in the following way. For a set of values of γ in the interval $[0,1]$, and for values of S_γ starting at $S_\gamma = 0.98 S_\gamma^\times$ (in particular, for $S_\gamma = 0.98/2^{2-\gamma} + 0.001\,n$, $n = 0, 1, 2, \dots$) we applied the fractional FTCS integration until step M. We say that the resulting integration for given values of γ and S_γ is unstable when the following condition is satisfied at any position j:

$$\left| \frac{u_j^{m-1}}{u_j^m} - \Xi \right| > \Xi \quad \text{for any} \quad m = M - \Delta M, M - \Delta M + 1, \dots, M, \tag{4.6}$$

where $\Xi = 5$ and $\Delta M = 10$. This means that the numerical solution is considered unstable if the quotient u_j^{m-1}/u_j^m becomes negative or larger than 2Ξ at any of the last ΔM steps. (Of course, this criterion is arbitrary; however, the results do not change substantially for any other reasonable choice of Ξ and ΔM.) Let $S_\gamma^{\min}$ be the smallest value of $S_\gamma = 0.98/2^{2-\gamma} + 0.001\,n$ that verifies the criterion (4.6). For the absorbing boundary problem we calculate these values using $\Delta x = 1/2N$ with $N = 5$, $M = 50$, $M = 100$, and $M = 1000$. For the propagator, we calculate $S_\gamma^{\min}$ using $M = 1000$ and $\Delta t = 5 \times 10^{-4}$ in a lattice with absorbing frontiers placed at $x = -N\Delta x$ and $x = N\Delta x$ with $N = 50$. It is well known that for a lattice with $2N + 1$ points (including the absorbing boundaries) the maximum value of $\sin(q\Delta x/2)$ in (3.10) occurs for $q\Delta x = (2N-1)\pi/(2N)$, so that in Figure 4.3 we plot $S_\gamma^{\min} \sin^2[(2N-1)\pi/(4N)]$. We observe that for large M the stability bound predicted by (3.12) agrees with the result of the numerical test. The larger values obtained for smaller M mean that the method must

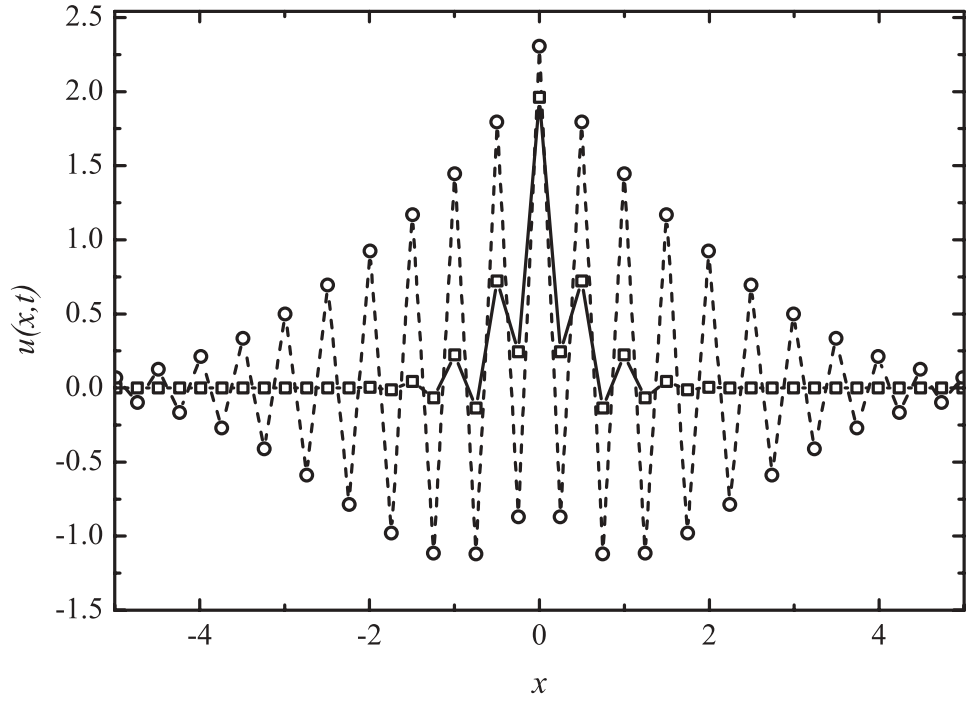


FIG. 4.4. *Unstable numerical solution of the subdiffusion equation for the problem defined in the unbounded space with initial condition* $u(x, t = 0) = \delta(x)$ *for* $\gamma = 1/2$, $K_\gamma = 1$, $S = 0.36$ *and* $t = 0.005$ *(squares), and* $t = 0.05$ *(circles). The time step is* $\Delta t = 0.0005$ *and the spatial mesh* Δx *is obtained according to* (4.5). *The lines are plotted as a visual guide.*

be "very unstable" to fulfill our instability criterion in so few steps. The success of the numerical test is truly remarkable and supports the application of our Fourier–von Neumann-type stability analysis to the fractional FTCS scheme made in section 3.1.

In Figure 4.4 we plot the numerical solution when $S_\gamma = 0.36 > S_\gamma^\times$ in the case of the propagator with $\gamma = 1/2$. This kind of awkward oscillatory behavior in the unstable domain is also typical for ordinary partial differential equations.

5. Concluding remarks. The availability of efficient numerical algorithms for the integration of fractional equations is important, as these equations are becoming essential tools for the description of a wide range of systems [3]. In this paper we have discussed a numerical algorithm for the solution of the fractional (sub)diffusion equation (1.2). Although we have dealt with this particular equation, our procedure could be extended to any fractional integrodifferential equation (for example, to fractional diffusion-wave equations) by means of an obvious combination of the Grünwald–Letnikov definition of the fractional derivative [1, 2, 4] with standard discretization algorithms used in the context of ordinary partial differential equations [26]. Furthermore, the method (given its explicit nature) can be trivially extended to d-dimensional problems, which is not such an easy task when implicit methods are considered.

In our numerical method the state of the system at a given time $t = m\Delta t$ is given explicitly in terms of the previous states at $t = (m-1)\Delta t, \ldots, \Delta t, 0$ by means of the FTCS scheme (3.3). We verified that for some standard initial conditions with exact analytical solution, namely, (a) the propagator in an unlimited

system with $u(x, t = 0) = \delta(x)$, and (b) a system with absorbing boundaries and $u(x, t = 0) = x(1 - x)$, the present algorithm leads to numerical solutions which are in excellent agreement with the exact solutions. Using a Fourier–von Neumann technique we have provided the conditions for which the fractional FTCS method is stable (cf. (3.10) and (3.11)). For $\gamma = 1$ the well-known bound $S = D\Delta t/(\Delta x)^2 \leq 1/2$ of the ordinary explicit method for the diffusion equation is recovered.

Concerning the implementation of the method, we must remark that the evaluation of the state of the system at a given time step $m\Delta t$ requires information about all previous states at $t = (m-1)\Delta t, (m-2)\Delta t, \ldots, \Delta t, 0$ and not merely the immediately preceding state as in ordinary diffusion. This is a consequence of the non-Markovian nature of subdiffusion and implies the need for massive computer memory in order to store the evolution of the system, which is especially cumbersome in computations of long-time asymptotic behaviors. This could be palliated by using the "short-memory" principle [1] or the nested mesh procedure [30]. Another feature of the explicit numerical scheme is the interdependence of the temporal and spatial discrete steps for a fixed S_γ. If, as usual, one intends to integrate an equation with a given mesh Δx, then the corresponding step size Δt for a given $S_\gamma < S_\gamma^\times$ is of the order $(\Delta x)^{2/\gamma}$. As a consequence, Δt could become extremely small even for no too small values of Δx, especially when the problem is far from the diffusion limit, i.e., for small values of γ, so that the number of steps needed to reach even moderate times would become prohibitively large. In this case, resorting to other methods (e.g., implicit methods [14, 21, 22, 23, 24, 25]) that are stable for larger values of Δt is compulsory.

REFERENCES

[1] I. Podlubny, *Fractional Differential Equations*, Academic Press, San Diego, 1999.

[2] R. Hilfer, ed., *Applications of Fractional Calculus in Physics*, World Scientific, Singapore, 2000.

[3] I. M. Sokolov, J. Klafter, and A. Blumen, *Fractional kinetics*, Phys. Today, 55 (2002), pp. 48–54.

[4] R. Metzler and J. Klafter, *The random walk's guide to anomalous diffusion: A fractional dynamics approach*, Phys. Rep., 339 (2000), pp. 1–77.

[5] V. Balakrishnan, *Anomalous diffusion in one dimension*, Phys. A, 132 (1985), pp. 569–580.

[6] W. Wyss, *Fractional diffusion equation*, J. Math. Phys., 27 (1986), pp. 2782–2785.

[7] W. R. Schneider and W. Wyss, *Fractional diffusion and wave equations*, J. Math. Phys., 30 (1989), pp. 134–144.

[8] R. Metzler, E. Barkai, and J. Klafter, *Anomalous diffusion and relaxation close to thermal equilibrium: A fractional Fokker-Planck equation approach*, Phys. Rev. Lett., 82 (1999), pp. 3563–3567.

[9] A. Compte, *Continuous time random walks on moving fluids*, Phys. Rev. E, 55 (1997), pp. 6821–6831.

[10] A. Compte and M. O. Cáceres, *Fractional dynamics in random velocity fields*, Phys. Rev. Lett., 81 (1998), pp. 3140–3143.

[11] R. Metzler, J. Klafter, and I. M. Sokolov, *Anomalous transport in external fields: Continuous time random walks and fractional diffusion equations extended*, Phys. Rev. E, 58 (1998), pp. 1621–1633.

[12] S. B. Yuste and K. Lindenberg, *Subdiffusion limited A+A reactions*, Phys. Rev. Lett., 87 (2001), article 118301.

[13] S. B. Yuste and K. Lindenberg, *Subdiffusion-limited reactions*, Chem. Phys., 284 (2002), pp. 169–180.

[14] C. Chuanmiao and S. Tsimin, *Finite Element Methods for Integrodifferential Equations*, World Scientific, Singapore, 1998.

[15] R. Gorenflo, Y. Luchko, and F. Mainardi, *Analytical properties and applications of the Wright function*, Fract. Calc. Appl. Anal., 2 (1999), pp. 383–414.

[16] R. Metzler and J. Klafter, *Boundary value problems for fractional diffusion equations*, Phys. A, 278 (2000), pp. 107–125.

[17] E. Barkai and R. J. Silbey, *Fractional Kramers equation*, J. Phys. Chem. B, 104 (2000), pp. 3866–3874.
[18] R. Gorenflo and F. Mainardi, *Random walk models for space-fractional diffusion processes*, Fract. Calc. Appl. Anal., 1 (1998), pp. 167–191.
[19] R. Gorenflo, G. De Fabritiis, and F. Mainardi, *Discrete random walk models for symmetric Lévy-Feller diffusion processes*, Phys. A, 269 (1999), pp. 79–89.
[20] R. Gorenflo, F. Mainardi, D. Moretti, G. Pagnini, and P. Paradisi, *Discrete random walk models for space-time fractional diffusion*, Chem. Phys., 284 (2002), pp. 521–541.
[21] J. M. Sanz-Serna, *A numerical method for a partial integro-differential equation*, SIAM J. Numer. Anal., 25 (1988), pp. 319–327.
[22] J. C. López-Marcos, *A difference scheme for a nonlinear partial integrodifferential equation*, SIAM J. Numer. Anal., 27 (1990), pp. 20–31.
[23] Ch. Lubich, I. H. Sloan, and V. Thomée, *Nonsmooth data error estimates for approximations of an evolution equation with a positive-type memory term*, Math. Comp., 65 (1996), pp. 1–17.
[24] W. McLean and V. Thomée, *Numerical solution of an evolution equation with a positive-type memory term*, J. Austral. Math. Soc. Ser. B, 35 (1993), pp. 23–70.
[25] W. McLean, V. Thomée, and L. B. Wahlbin, *Discretization with variable time steps of an evolution equation with a positive-type memory term*, J. Comput. Appl. Math., 69 (1996), pp. 49–69.
[26] K. W. Morton and D. F. Mayers, *Numerical Solution of Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1994.
[27] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran* 77: *The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.
[28] Ch. Lubich, *Discretized fractional calculus*, SIAM J. Math. Anal., 17 (1986), pp. 704–719.
[29] R. Gorenflo, *Fractional calculus: Some numerical methods*, in Fractals and Fractional Calculus in Continuum Mechanics, A. Carpinteri and F. Mainardi, eds., Springer-Verlag, New York, 1997, pp. 277–290.
[30] N. J. Ford and C. Simpson, *The numerical solution of fractional differential equations: Speed versus accuracy*, Numer. Algorithms, 26 (2001), pp. 333-346.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1875–1898

A POSTERIORI ERROR ANALYSIS FOR THE MEAN CURVATURE FLOW OF GRAPHS*

OMAR LAKKIS† AND RICARDO H. NOCHETTO‡

Abstract. We study the equation describing the motion of a nonparametric surface according to its mean curvature flow. This is a nonlinear nonuniformly parabolic PDE that can be discretized in space via a finite element method. We conduct an a posteriori error analysis of the spatial discretization and derive upper bounds on the error in terms of computable estimators based on local residual indicators. The reliability of the estimators is illustrated with two numerical simulations, one of which treats the case of a singular solution.

Key words. finite element, mean curvature, error analysis, a posteriori, nonlinear PDE, parabolic equation, geometric motion, convergence, reliability, efficiency, sharp estimates, effectivity index

AMS subject classifications. Primary, 65N30, 65G20, 35K60; Secondary, 57R99, 40A30

DOI. 10.1137/S0036142903430207

1. Introduction. The objective of this article is the derivation of reliable a posteriori error estimates for the *mean curvature flow (MCF)* of a d-dimensional time-dependent submanifold $\Gamma(t)$ of the Euclidean space $\mathbb{R}^{d+1}$. We pay special attention to the physically relevant cases ($d = 1, 2, 3$), and we refer to $\Gamma(t)$ simply as a *moving surface*. A geometric definition of the MCF, whose details can be found in Huisken [15] and the references therein, is given by

$$\boldsymbol{V}(\boldsymbol{x}, t) = -\boldsymbol{\kappa}(\boldsymbol{x}, t) \quad \text{for } \boldsymbol{x} \in \Gamma(t),\ t \in \mathbb{R}, \tag{1.1}$$

where $\boldsymbol{V}$ and $\boldsymbol{\kappa}$ are respectively the velocity and the vector mean curvature of Γ. More general definitions of MCF are found in the literature [5, 11, 4], but will not be used.

In this paper we are interested in the *graph* (also called *nonparametric*) *description* in which the moving surface is described as the graph of a function u defined on a space-time domain $\Omega \times [0, T] \subset \mathbb{R}^d \times \mathbb{R}$. This description leads to the following PDE, referred to as the *mean curvature flow of graphs (MCFG)*:

$$\frac{\partial_t u(\boldsymbol{x}, t)}{Qu(\boldsymbol{x}, t)} - \frac{1}{d} \operatorname{div} \frac{\nabla u(\boldsymbol{x}, t)}{Qu(\boldsymbol{x}, t)} = 0 \quad \text{for } \boldsymbol{x} \in \Omega,\ t \in [0, T], \tag{1.2}$$

where ∇ denotes the derivative with respect to $\boldsymbol{x}$ and Q the *elementary area* operator defined by

$$Qw := (1 + |\nabla w|^2)^{1/2}. \tag{1.3}$$

*Received by the editors June 19, 2003; accepted for publication (in revised form) April 28, 2004; published electronically January 20, 2005. This work was partially supported by NSF grants DMS-9971450, DMS-0204670, and INT-9910086.

http://www.siam.org/journals/sinum/42-5/43020.html

†Institute for Applied and Computational Mathematics, FORTH PO Box 1527, GR-71110 Heraklion, Greece (omar@iacm.forth.gr, http://www.iacm.forth.gr/∼omar). This author was also supported by *Progetto MURST Cofin* 2000 at *Università di Milano*.

‡Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu).

We drop the factor $1/d$, through a time rescaling by d, and we study the following initial-boundary value problem associated with (1.2).

PROBLEM 1.1 (Cauchy–Dirichlet problem for the MCFG). *Given functions* $f : \Omega \times (0,T] \to \mathbb{R}$ *and* $g : \partial_{\mathrm{p}}(\Omega \times (0,T)) \to \mathbb{R}$, *find* $u : \overline{\Omega} \times [0,T] \to \mathbb{R}$ *such that*

$$\frac{\partial_t u(\boldsymbol{x},t)}{Qu(\boldsymbol{x},t)} - \operatorname{div} \frac{\nabla u(\boldsymbol{x},t)}{Qu(\boldsymbol{x},t)} = f(\boldsymbol{x},t) \quad \text{for } (\boldsymbol{x},t) \in \Omega \times (0,T], \tag{1.4}$$

$$u(\boldsymbol{x},t) = g(\boldsymbol{x},t) \quad \text{for } (\boldsymbol{x},t) \in \partial_{\mathrm{p}}(\Omega \times (0,T)), \tag{1.5}$$

where $\partial_{\mathrm{p}}(\Omega \times (0,T))$ *is the* parabolic boundary *defined as* $\Omega \times \{0\} \cup \partial\Omega \times [0,T]$.

Arguably the MCF plays the role of model geometric motion, in the same way as the heat equation plays the role of model diffusion equation. For more than two decades the MCF has been the object of mathematical analysis [1, 4, 5, 11, 15, 16] as well as computer simulations [5, 8, 23, 20] and numerical analysis [6, 7, 8, 28]. It has also attracted the interest of practitioners, especially in the fields of materials science and phase transition where the MCF, or some closely related geometric motion, often models the motion of a free boundary [3, 13, 24].

A straightforward way to approximate numerically the solution of Problem 1.1 is first to discretize the spatial variable through a finite element method—which comes naturally, as (1.4) is written in "divergence form"—and secondly to discretize the time variable with a finite difference scheme known as *semi-implicit*, in which the nonlinearity is treated explicitly and the linear part implicitly [8]. The first stage of this process, discussed in sections 2.1–2.5, is referred to as the *spatial (semi-) discretization*. Deckelnick and Dziuk [7] and Dziuk [8] have derived a priori error estimates for both the spatially discrete and the semi-implicit fully discrete scheme.

The study of *a posteriori error estimates* for evolution equations, which has developed in the last 15 years, is mainly motivated by their successful use in deriving adaptive mesh refinement algorithms. The lack of such estimates in the case of the MCFG and the interest in adaptive methods for this problem are the driving motives behind this article. Our main results, discussed in section 3, are a posteriori upper bounds on the error for the spatially discrete approximation. A posteriori error estimates have been established for linear parabolic problems [9, 19] and used to derive adaptive mesh refinement algorithms. Analogous results have also been derived for certain nonlinear elliptic [12, 26] and parabolic [10, 17, 18] equations, but these cannot be applied to the MCFG.

As observed since the early days of adaptive finite element methods (FEM) [2], an adaptive mesh refinement algorithm must satisfy two fundamental properties: *reliability* and *efficiency*. These two algorithmic concepts are closely related to the nature of the error bounds. Indeed, an algorithm is called *reliable* if the error between its output and the exact solution is bounded from above by a given tolerance; in terms of estimators, reliability is achieved if the error/estimator ratio—known as the *effectivity index* in the literature—is bounded from above by a positive constant. On the other hand, an algorithm is called *efficient* if it produces a result with a prescribed error in the least amount of computational time; the efficiency of an algorithm translates, in the language of estimators, into the effectivity index being bounded from below. For an estimator to be both reliable and efficient, it is necessary for it to be *sharp*, meaning that the order of convergence of the error and that of the estimator must be equal, as the meshsize goes to zero. In particular, *sharpness* allows the estimators to be used in stopping criteria for adaptive algorithms. In this paper, besides proving reliable error estimates (upper error bounds), we will also conduct numerical experiments to

understand whether these estimates are also sharp. For this, the numerical examples we shall present in section 7 are mainly designed toward comparing the numerical asymptotic convergence rates of the error and of the estimators.

The MCFG is an example of an evolution equation that is not covered by any of the general techniques developed so far for the derivation of a posteriori error estimates for nonlinear equations [10, 17, 27]. This is mainly due to the nonuniformly parabolic nature of the equation and, more philosophically, to the fact that general nonlinear theories end up being less reliable and harder to apply. In this paper we employ an ad hoc energy technique to derive the estimates. To the best of our knowledge, the energy technique is the only practical way to achieve our aim. A distinctive feature of this paper is the use of special quantities to quantify the error. Like in most nonuniformly parabolic equations, the Sobolev norms are extremely hard to handle in the MCFG context, and we are naturally led to use the *geometric errors*, which are introduced next. These are not Sobolev norms of the error $u - u_h$, where u and u_h are respectively the exact and approximate solutions, but more specialized measures of the error (see section 2.5). The geometric errors are not even symmetric in u and u_h, yet they satisfactorily quantify the error and are easy to use.

DEFINITION 1.2 (geometric error). *Let u be the solution of Problem* 1.1 *and u_h be the finite element solution given by Problem* 2.5. *For each $t \in [0,T]$, define*

$$A(t) := \int_\Omega |\boldsymbol{N}u_h(\boldsymbol{x},t) - \boldsymbol{N}u(\boldsymbol{x},t)|^2 \, Qu(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}, \tag{1.6}$$

$$B(t) := \int_0^t \int_\Omega \left(Vu_h(\boldsymbol{x},s) - Vu(\boldsymbol{x},s)\right)^2 Qu(\boldsymbol{x},s)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}s, \tag{1.7}$$

where

$$\begin{aligned} &\mathrm{W}^1(\Omega) \ni w \mapsto \boldsymbol{N}w := \frac{(\nabla w; -1)}{Qw} \in \mathrm{L}_\infty(\Omega)^{d+1} \\ \text{and} \quad &\mathrm{W}^1_1(\Omega\times(0,T)) \ni w \mapsto Vw := \frac{\partial_t w}{Qw} \in \mathrm{L}_1^{\mathrm{loc}}(\Omega\times(0,T)) \end{aligned} \tag{1.8}$$

are, respectively, the normal vector *and the* normal velocity *operators. We will denote by $\mathrm{C}^k(\Omega)$ (resp., $\mathrm{W}^k(\Omega)$) the space of k times continuously (resp., weakly) differentiable functions, by $\mathrm{W}^k_p(\Omega)$ the usual Sobolev space of functions in $\mathrm{W}^k(\Omega)$ with derivatives in $\mathrm{L}_p(\Omega)$, and by $\overset{\circ}{\mathrm{W}}{}^k_p(\Omega)$ the subspace of functions with vanishing trace. The functions of time A and B are the building blocks of the* total geometric error E *defined by*

$$E(t)^2 := B(t) + \sup_{[0,t]} A(s) = \int_0^t \int_\Omega (Vu_h - Vu)^2\, Qu + \sup_{(0,t)} \int_\Omega |\boldsymbol{N}u_h - \boldsymbol{N}u|^2\, Qu. \tag{1.9}$$

We refer to $\sup_{[0,t]} A^{1/2}$ *and* $B^{1/2}$ *as the* geometric energy error *and* normal velocity error*, respectively.*

The integrals of the form $\int_\Omega \cdot\, Qu(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}$ in (1.9) can be interpreted as integrals over the moving surface $\Gamma(t)$, which give us the $\mathrm{L}_2(\Gamma)$ norm of the *difference of normals* and the *difference of normal velocities*. A comparison with the integrals appearing on the left-hand side of (2.6) explains in part why they "fit" the problem.

We point out that, despite the natural relation between our notion of error and the MCFG, no related concept of error has yet been used in the context of a posteriori error control for parabolic equations. In fact, the geometric nature contrasts sharply with the pure analytic setting found, for instance, in Verfürth's monograph [26]. A related, symmetric, geometric error is employed by Fierro and Veeser for the stationary case [12].

It is important to observe that the sharpest estimate in this article, given by Theorem 3.6, is a *conditional estimate*. By conditional we mean that the estimate is valid only if a certain condition on how close the approximate solution is to the exact solution is satisfied. A relevant feature of our result in this respect is that the condition can be machine-checked since it entails computable quantities. This is of paramount importance for a result to be fully "a posteriori" (see Remark 3.7). In this sense, to the best of our knowledge, our result is the first conditional a posteriori estimate for nonlinear parabolic equations. Conditional results have also been derived for the prescribed mean curvature (elliptic) equation by Fierro and Veeser [12]. We notice that Verfürth has also established conditional results, but the conditions are not fully a posteriori and cannot be machine-checked [26]. In order to appreciate the sharpness of the conditional result of Theorem 3.6, an *unconditional estimate* is given in Theorem 3.4 for the sake of comparison. Our numerical results provide a practical comparison between the two theoretical bounds and show that the conditional estimate is sharp while the unconditional estimate is not.

Dziuk has shown an a priori error bound of rate $\mathrm{O}(h)$ on the geometric error in the spatially discrete case [8]. The geometric error introduced in Definition 1.2 is similar to the one used by Dziuk, but in his case the integrals are evaluated on the discrete surface, while we compute them on the exact surface. In this respect our a posteriori viewpoint can be seen, roughly speaking, as dual to the a priori approach. We notice, however, that our results are valid under weaker regularity assumptions on the exact solution u (see Example 7.5). Our analysis also includes *time-dependent boundary value* g and nonhomogeneous right-hand side f, while Dziuk's analysis is limited to the homogeneous and time-independent boundary value case.

The rest of this paper is organized as follows. In section 2 we discuss some properties of Problem 1.1 and introduce the associated spatial finite element method. In section 3 we state the main results and make some observations. Next, in sections 4–6 we prove these results. Finally, numerical tests are discussed in section 7.

2. The Cauchy–Dirichlet problem and its spatial discretization.

ASSUMPTION 2.1 (solvability and regularity). *Unless otherwise stated, the following conditions will be assumed to hold:*

(a) Classical solvability: *Problem* 1.1 *admits a unique classical solution* u *in* $\mathrm{C}^{2,1}(\Omega \times (0,T]) \cap \mathrm{C}^0(\overline{\Omega} \times [0,T])$ *for some* $T > 0$.

(b) Boundary regularity of contact angle:[1]

$$\frac{\nabla u(t)}{Qu(t)} \in \mathrm{W}^1_d(\Omega) \qquad \forall t \in [0,T]. \tag{2.1}$$

(c) Regularity of normal velocity:

$$Vu(t) = \frac{\partial_t u(t)}{Qu(t)} \in \mathrm{L}_d(\Omega) \qquad \forall t \in [0,T]. \tag{2.2}$$

[1] We use the following convention throughout this article: whenever a space-time function $w : \Omega \times [0,T] \to \mathbb{R}^N$ $(N = 1, d)$ is written with only one argument, it means that the argument is a time variable and that its value—e.g., $w(t)$ or $w(1/2)$—is a function with domain Ω.

(d) Regularity of vertical velocity:

$$\partial_t u(t) \in \mathrm{W}_1^1(\Omega) \cap \mathrm{L}_2(\Omega) \qquad \forall t \in [0,T]. \tag{2.3}$$

Remark 2.2 (about the regularity assumptions). Assumption 2.1(a) is backed up by the fact that Problem 1.1 admits classical solutions under certain sufficient conditions relating the mean-convexity of $\partial\Omega$ and the function $|f|$ [16, section 12.8]. Solutions, which are classical up to blow-up, can also exist in more general situations where the domain is non–mean-convex or compatibility conditions are violated [25]. There are two implicit assumptions that are immediate consequences of Assumption 2.1(a): we necessarily have $f \in \mathrm{C}^0(\Omega \times (0,T])$ and $g \in \mathrm{C}^0(\partial_{\mathrm{p}}(\Omega \times (0,T]))$. Although a "weak form" of Problem 1.1 will be derived in section 2.3, we do not know of any satisfactory concept of a "weak solution" for it.

The reason that we assume (2.3) is technical: this assumption will be needed to test (1.4) by $\partial_t u$ (see sections 2.3, 2.4, and 4.3). Notice that for $d \le 2$, in view of the Sobolev embedding, this assumption can be simplified to $\partial_t u \in \mathrm{W}_1^1(\Omega)$ and implies (2.2). Notice also that, for $d \ge 1$, the Sobolev embedding and (2.3) imply that $\partial_t u \in \mathrm{L}_{d'}(\Omega)$ for $d' = d/(d-1)$.

Proposition 2.3 (weak form). *Let $u \in \mathrm{C}^{2,1}(\Omega \times (0,T]) \cap \mathrm{C}^0(\overline{\Omega} \times [0,T])$ be a given function that satisfies* (2.1) *and* (2.2). *The function u is a classical solution of Problem* 1.1 *if and only if*

$$\left\langle \frac{\partial_t u(t)}{Qu(t)}, \phi \right\rangle + \left\langle \frac{\nabla u(t)}{Qu(t)}, \nabla\phi \right\rangle = \langle f(t), \phi\rangle \qquad \forall \phi \in \mathring{\mathrm{W}}_1^1(\Omega),\ t \in (0,T], \tag{2.4}$$

$$u(t) - \tilde{g}(t) \in \mathring{\mathrm{W}}_1^1(\Omega) \quad \forall t \in (0,T], \quad \textit{and} \quad u(0) = g(0), \tag{2.5}$$

where $\tilde{g}(t)$ is an extension of $g(t)$ to all of Ω.

We use the notation $\langle v, w\rangle_D := \int_D v(\boldsymbol{x}) w(\boldsymbol{x})\, \mathrm{d}\mu(\boldsymbol{x})$ for functions v and w such that $vw \in \mathrm{L}_1(D,\mu)$, $D \subset \mathbb{R}^d$, and $\mathrm{d}\mu$ is the Lebesgue measure "$\mathrm{d}\cdot$" or the $(d-1)$-dimensional Hausdorff measure, depending on the Hausdorff dimension of D. If $D = \Omega$, we omit the subscript in the brackets.

The proof of Proposition 2.3 follows basic PDE techniques and is omitted. We observe that the existence of $\tilde{g}$, for $g(t) \in \mathrm{L}_1(\partial\Omega)$, is guaranteed in view of [22, eq. (5.5)].

Lemma 2.4 (stability estimate). *If we have $f \in \mathrm{L}_2(0,T;\mathrm{L}_\infty(\Omega))$ and $g \in \mathrm{W}_1^1(\partial_{\mathrm{p}}(\Omega \times (0,T)))$, then*

$$\begin{aligned} &\frac{1}{2}\int_0^t \int_\Omega |Vu|^2\, Qu + \int_\Omega Qu(t) \\ &\qquad \le \exp\left(\frac{1}{2}\int_0^t \|f\|^2_{\mathrm{L}_\infty(\Omega)}\right)\left(\|Qg(0)\|_{\mathrm{L}_1(\Omega)} + \|\partial_t g\|_{\mathrm{L}_1(\partial\Omega\times(0,t))}\right). \end{aligned} \tag{2.6}$$

Proof. Test (1.4) by $\partial_t u \in \mathrm{L}_{d'}(\Omega)$ and, owing to (2.1) and (2.3), apply the integration by parts formula on Ω:

$$0 = \int_\Omega |Vu|^2\, Qu + \int_\Omega \frac{\nabla u}{Qu} \cdot \nabla \partial_t u - \int_{\partial\Omega} \frac{\nabla u \cdot \boldsymbol{\nu}}{Qu} \partial_t u - \int_\Omega f \partial_t u. \tag{2.7}$$

The first term, which is equal to $\int_\Omega Vu\, \partial_t u$, is well defined thanks to (2.3) and (2.2). The third and fourth terms are bounded as follows:

$$\int_{\partial\Omega} \frac{\nabla u \cdot \boldsymbol{\nu}}{Qu} \partial_t u = \int_{\partial\Omega} \left(\frac{\nabla u}{Qu} \cdot \boldsymbol{\nu}\right) \partial_t g \le \|\partial_t g\|_{\mathrm{L}_1(\partial\Omega)}, \tag{2.8}$$

$$\int_\Omega f\partial_t u = \int_\Omega f\sqrt{Qu}\frac{\partial_t u}{\sqrt{Qu}} \le \frac{1}{2}\int_\Omega |Vu|^2\, Qu + \frac{1}{2}\|f\|^2_{\mathrm{L}_\infty(\Omega)}\int_\Omega Qu. \tag{2.9}$$

Next we observe that the basic identity

$$\partial_t Qu(\boldsymbol{x},t) = \partial_t\sqrt{1+|\nabla u|^2} = \frac{\nabla u\cdot\partial_t\nabla u}{Qu} \tag{2.10}$$

implies

$$\frac{1}{2}\int_\Omega |Vu|^2\, Qu + \mathrm{d}_t\int_\Omega Qu \le \|\partial_t g\|_{\mathrm{L}_1(\partial\Omega)} + \frac{1}{2}\|f\|^2_{\mathrm{L}_\infty(\Omega)}\int_\Omega Qu. \tag{2.11}$$

The result is obtained by integrating on $[0,t]$ and applying the Gronwall lemma. ∎

Inequality (2.6) acquires a geometric meaning upon observing that $\int_\Omega Qu$ is the area of graph(u). This gives us a control on the growth of the area in time in terms of the data. In particular, if the forcing term $f=0$ and the boundary conditions are time-independent, then (2.6) quantifies the decrease in area of the graph that tends toward a nonparametric minimal surface as time grows. The MCF, with $f=0$, is thus interpreted as the gradient descent method for the area functional with respect to the $\mathrm{L}_2(\Gamma(t))$ norm.

2.1. Finite element discretization. We start by introducing $\{\mathscr{T}_h\}_h$, a shape-regular family of triangulations (simplicial partitions) of the domain Ω. This means that there exists a constant $\sigma_0\in\mathbb{R}^+$, independent of the particular triangulation $\mathscr{T}_h$, such that

$$\frac{\sup\{\rho\in\mathbb{R}^+:\ \mathrm{B}_\rho(\boldsymbol{x})\subset K\}}{\operatorname{diam}(K)} \ge \sigma_0 \qquad \forall K\in\mathscr{T}_h. \tag{2.12}$$

We will refer to σ_0 as the *shape-regularity* of the family $\{\mathscr{T}_h\}_h$. We assume that the *approximate domain* $\Omega_h = \operatorname{int}\left(\bigcup_{K\in\mathscr{T}_h}\overline{K}\right)$ coincides with Ω; this is a simplifying assumption that could be removed at the cost of seriously complicating the analysis, without adding much content to the results we intend to present. The symbol h stands for both the *local meshsize function* and the global meshsize of $\mathscr{T}_h$; this abuse of notation should not cause confusion.

Given a simplex $K\in\mathscr{T}_h$ and $\psi:\Omega\to\mathbb{R}$, we denote by ψ_K the restriction $\psi|_K$—e.g., if $\psi=h$, we have $h_K=\operatorname{diam}(K)$—and by $\mathscr{U}_K^h$, the $\mathscr{T}_h$*-neighborhood* of K,

$$\mathscr{U}_K^h := \operatorname{int}\Big(\bigcup\big\{\overline{K'}\in\mathscr{T}_h:\ \overline{K'}\cap\overline{K}\neq\emptyset\big\}\Big). \tag{2.13}$$

We also associate with $\mathscr{T}_h$ its *internal mesh* $\Sigma_h := \bigcup_{S\in\mathscr{S}_h^\circ} S$, where $\mathscr{S}_h^\circ$ is the set of internal edges (or faces) of the simplexes in $\mathscr{T}_h$. The *finite element spaces*, constructed on $\mathscr{T}_h$, that will be employed are

$$\mathbb{V}_h := \big\{\phi\in\mathrm{W}_1^1(\Omega):\ \phi_K\in\mathbb{P}^\ell\ \forall K\in\mathscr{T}_h\big\} \quad\text{and}\quad \mathring{\mathbb{V}}_h := \mathbb{V}_h\cap\mathring{\mathrm{W}}_1^1(\Omega), \tag{2.14}$$

where $\ell\in\mathbb{Z}^+$ and $\mathbb{P}^\ell$ is the space of polynomials of degree at most ℓ. A spatial finite element discretization of Problem 1.1 can be now derived from (2.4).

Problem 2.5 (spatially discrete scheme for the MCFG). *Let $\tilde g_h(t)\in\mathbb{V}_h$ be an interpolant of $\tilde g(t)$. Find $u_h\in\mathrm{C}^1([0,T];\mathbb{V}_h)$ such that, for each $t\in[0,T]$,*

$$u_h(t) - \tilde g_h(t) \in \mathring{\mathbb{V}}_h, \tag{2.15}$$

$$\left\langle \frac{\partial_t u_h(t)}{Qu_h(t)}, \phi_h \right\rangle + \left\langle \frac{\nabla u_h(t)}{Qu_h(t)}, \nabla \phi_h \right\rangle = \langle f(t), \phi_h \rangle \qquad \forall \phi_h \in \mathring{\mathbb{V}}_h. \tag{2.16}$$

Solvability of Problem 2.5 and a priori error estimates are studied by Dziuk [8, Thm. 1]. Throughout the paper, u_h will denote the solution of Problem 2.5.

3. A posteriori error estimates. In this section we state our main results. We start by introducing some definitions.

DEFINITION 3.1 (residual functions). *For each $t \in [0,T]$, let $r(t)$ be the* internal residual *and let $j(t)$ be the* jump residual *associated with u_h. These two functions are defined on $\Omega \setminus \Sigma_h$ and Σ_h, respectively, and are given by*

$$r(\boldsymbol{x},t) := \frac{\partial_t u_h(\boldsymbol{x},t)}{Qu_h(\boldsymbol{x},t)} - f(\boldsymbol{x},t) - \operatorname{div}\left(\frac{\nabla u_h(\boldsymbol{x},t)}{Qu_h(\boldsymbol{x},t)}\right) \qquad \text{for } \boldsymbol{x} \in \Omega \setminus \Sigma_h, \tag{3.1}$$

$$j(\boldsymbol{x},t) := \left[\frac{\nabla u_h(\boldsymbol{x},t)}{Qu_h(\boldsymbol{x},t)}\right]_S \qquad \text{for } \boldsymbol{x} \in S \in \mathscr{S}_h^\circ, \tag{3.2}$$

where the jump of a vector field $\boldsymbol{\psi}$ *across an edge S is defined as*

$$[\boldsymbol{\psi}]_S(\boldsymbol{x}) := \lim_{\varepsilon \to 0} \left(\boldsymbol{\psi}(\boldsymbol{x} + \varepsilon \boldsymbol{\nu}_S) - \boldsymbol{\psi}(\boldsymbol{x} - \varepsilon \boldsymbol{\nu}_S)\right) \cdot \boldsymbol{\nu}_S \tag{3.3}$$

with $\boldsymbol{x} \in S$ and $\boldsymbol{\nu}_S$ denoting one of the two normals to S (the choice is arbitrary and does not affect the definition).

DEFINITION 3.2 (local indicators and weights). *Denote by C_1 and C_2 the Scott–Zhang interpolation inequality constants, which depend only on the shape-regularity σ_0 of $\mathscr{T}_h$ and which we introduce later in inequalities* (6.2) *and* (6.3)*, respectively. With each $K \in \mathscr{T}_h$ we associate the* local

$$\text{elliptic indicator} \qquad \eta_0^K(t) := h_K^{d/2}\left(C_1 \|r(t)\|_{\mathrm{L}_d(K)} + C_2 \|j(t)\|_{\mathrm{L}_\infty(\partial K)}\right), \tag{3.4}$$

$$\text{parabolic indicator} \qquad \eta_1^K(t) := h_K^{d/2}\left(C_1 \|\partial_t r(t)\|_{\mathrm{L}_d(K)} + C_2 \|\partial_t j(t)\|_{\mathrm{L}_\infty(\partial K)}\right), \tag{3.5}$$

and the local weights

$$\omega^K(t) := \sup_{\boldsymbol{x} \in \mathscr{U}_K^h} Qu_h(\boldsymbol{x},t)^2, \qquad \alpha^K(t) := \omega^K(t)^2 \sup_{\boldsymbol{x} \in \mathscr{U}_K^h} \frac{1}{Qu(\boldsymbol{x},t)}. \tag{3.6}$$

DEFINITION 3.3 (a posteriori error estimators). *Denote by M and γ two positive constants, depending only on the shape-regularity σ_0, which we will introduce in detail in the proof of Lemma* 6.4*. We define the* elliptic part of the proper estimator

$$\mathscr{E}_{2,0}(t) := \sup_{s \in [0,t]} \hat{\mathscr{E}}_{2,0}(s), \qquad \text{where} \quad \hat{\mathscr{E}}_{2,0}(t)^2 := \gamma^2 \sum_{K \in \mathscr{T}_h} \alpha^K(t) \eta_0^K(t)^2, \tag{3.7}$$

the parabolic part of the proper estimator

$$\mathscr{E}_{2,1}(t) := \int_0^t \dot{\mathscr{E}}_{2,1}(s)\,\mathrm{d}s, \qquad \text{where} \quad \dot{\mathscr{E}}_{2,1}(t)^2 := \gamma^2 \sum_{K \in \mathscr{T}_h} \alpha^K(t) \eta_1^K(t)^2, \tag{3.8}$$

the elliptic part of the vicinity estimator

$$\mathscr{E}_{\infty,0}(t) := \sup_{s \in [0,t]} \hat{\mathscr{E}}_{\infty,0}(s), \quad \text{where} \quad \hat{\mathscr{E}}_{\infty,0}(t) := M \max_{K \in \mathscr{T}_h} \left(h_K^{-d/2} \omega^K(t) \eta_0^K(t)\right), \tag{3.9}$$

and the parabolic part of the vicinity estimator

$$\mathscr{E}_{\infty,1}(t) := \int_0^t \dot{\mathscr{E}}_{\infty,1}(s)\,\mathrm{d}s, \quad \textit{where} \quad \dot{\mathscr{E}}_{\infty,1}(t) := M \max_{K\in\mathscr{T}_h} \left(h_K^{-d/2} \omega^K(t) \eta_1^K(t) \right). \tag{3.10}$$

These definitions allow us to introduce the proper estimator *and the* vicinity estimator*, respectively, as*

$$\mathscr{E}_2(t) := \left(\mathscr{E}_{2,0}(t)^2 + \mathscr{E}_{2,1}(t)^2 \right)^{1/2} \quad \textit{and} \quad \mathscr{E}_\infty(t) := \mathscr{E}_{\infty,0}(t) + \mathscr{E}_{\infty,1}(t). \tag{3.11}$$

We finally introduce the initial estimator *and* total estimator*, respectively, as*

$$\mathscr{E}_0 := \left((1 + 2\mathscr{E}_{\infty,0}(0))\, A(0) + 2\mathscr{E}_{2,0}(0) \sqrt{A(0)} \right)^{1/2} \tag{3.12}$$

and

$$\mathscr{E}(t) := \left(\mathscr{E}_0^2 + \mathscr{E}_2(t)^2 + \mathscr{E}_\infty(t) \right)^{1/2}. \tag{3.13}$$

The motivation for our terminology will become clear in Theorem 3.6 *below: there the vicinity estimator $\mathscr{E}_\infty$ does not enter directly into the conditional estimate, but dictates a "closeness condition" that must be satisfied for the estimate to hold. This conditional estimate then involves the initial and proper estimators $\mathscr{E}_0$ and $\mathscr{E}_2$.*

We are now ready to state the main results, whose proofs are spread through sections 4–6.

THEOREM 3.4 (unconditional a posteriori estimate). *Let u be the solution of Problem* 1.1*, and u_h the finite element solution of Problem* 2.5*. For all $t \in [0,T]$ there exist $C = C[u_h, f, t]$ and $C' = C'[f, g, t]$ such that*

$$C \le \exp \int_0^t \left(2 \left\| \partial_t u_h(s) \right\|^2_{\mathrm{L}_\infty(\Omega)} + 4 \left\| \nabla \partial_t u_h(s) \right\|_{\mathrm{L}_\infty(\Omega)} \right) \mathrm{d}s, \tag{3.14}$$

$$C' \le \exp \left(\frac{1}{2} \left\| f(s) \right\|^2_{\mathrm{L}_\infty(\Omega)} \right) \left(\left\| Qg(0) \right\|_{\mathrm{L}_1(\Omega)} + \left\| \partial_t g \right\|_{\mathrm{L}_1(\partial\Omega\times(0,t))} \right), \tag{3.15}$$

$$\int_0^t \int_\Omega (Vu_h - Vu)^2 Qu + \frac{1}{2} \sup_{[0,t]} \int_\Omega \left| \boldsymbol{N} u_h - \boldsymbol{N} u \right|^2 Qu \le C \left({\mathscr{E}_0}^2 + 4\mathscr{E}_2(t)^2 + 8C'\mathscr{E}_\infty(t) \right). \tag{3.16}$$

Remark 3.5 (the sharpness of the unconditional estimate). The estimate (3.16) holds, regardless of whether the approximate solution u_h is close to or far from the exact solution u. The presence of the vicinity estimator $\mathscr{E}_\infty$ on the right-hand side is undesirable because, even under the most optimistic assumptions of regularity on u, there is no indication that this estimator will have the same order of convergence, as h goes to zero, as the square of the geometric error on the left-hand side. In fact, the numerical tests described in section 7 bear strong evidence that $\mathscr{E}_\infty$ does not decay with a sufficiently high power of h. This means that the above estimate is not sharp and that it cannot be relied upon as a stopping criterion in an adaptive scheme. A crucial point of this paper is that *this estimate can be improved, provided that u_h is sufficiently close to u,* as stated in the next theorem.

THEOREM 3.6 (conditional a posteriori estimate). *Let u be the solution of Problem* 1.1*, and u_h the finite element solution of Problem* 2.5*. For each $t \in [0,T]$, if*

$$\mathscr{E}_\infty(t) \leq \frac{1}{8}, \tag{3.17}$$

then there exists a constant $C = C[u_h, t]$ such that

$$C \leq \exp \int_0^t \left(2 \left\| \partial_t u_h(s) \right\|^2_{\mathrm{L}_\infty(\Omega)} + 4 \left\| \nabla \partial_t u_h(s) \right\|_{\mathrm{L}_\infty(\Omega)} \right) \mathrm{d}s, \tag{3.18}$$

$$\int_0^t \int_\Omega (Vu_h - Vu)^2 Qu + \frac{1}{2} \sup_{[0,t]} \int_\Omega \left| \boldsymbol{N} u_h - \boldsymbol{N} u \right|^2 Qu \leq C \left(\mathscr{E}_0{}^2 + 8 \mathscr{E}_2(t)^2 \right). \tag{3.19}$$

Remark 3.7 (a posteriori nature of condition (3.17)). Theorem 3.6 is a conditional result, typical in nonlinear analysis. The condition (3.17) can be interpreted as follows: the approximate solution u_h needs to be sufficiently close to the exact solution u for the estimate to hold. The technique we use can be thought of as a linearization of the equation about u_h, instead of a linearization about u, which would be natural in an a priori setting. This leads to the important fact that *condition* (3.17) *can be effectively verified* since it involves exclusively a posteriori, and therefore computable, quantities. Thus, in a practical adaptive method where a stopping criterion is needed, Theorem 3.4 would be used in the early preasymptotic stages in order to get close enough to the exact solution; the estimate of Theorem 3.6 would then provide a sharper criterion once the algorithm enters a second stage in which the condition (3.17) is satisfied.

4. The error equation. We divide the proof of Theorems 3.4 and 3.6 into several steps that will spread over the next two sections. Here we introduce the residual-based energy technique and we formulate the error equation.

4.1. The residual. The *residual* is defined as the difference between the exact operator acting on the approximate solution and the exact operator acting on the exact solution. In our setting, the result has to be understood in the following weak sense:

$$\langle \mathscr{R} \,|\, \phi \rangle := \left\langle \frac{\partial_t u_h}{Qu_h} - \frac{\partial_t u}{Qu}, \phi \right\rangle + \left\langle \frac{\nabla u_h}{Qu_h} - \frac{\nabla u}{Qu}, \nabla \phi \right\rangle \qquad \forall \phi \in \mathring{\mathrm{W}}{}^1_1(\Omega). \tag{4.1}$$

Here $\langle \cdot \,|\, \cdot \rangle$ stands for the duality pairing. The distribution $\mathscr{R}$ is time-dependent and, owing to Assumption 2.1, $\mathscr{R}(t)$ is a bounded linear functional on $\mathring{\mathrm{W}}{}^1_1(\Omega)$ for all $t \in [0,T]$. We will refer to $\mathscr{R}$ as the *residual functional*. The use of (2.4) and an integration by parts in the space variable lead to the residual functional *representation*

$$\begin{aligned} \langle \mathscr{R} \,|\, \phi \rangle &= \left\langle \frac{\partial_t u_h}{Qu_h} - f - \operatorname{div} \left(\frac{\nabla u_h}{Qu_h} \right), \phi \right\rangle + \left\langle \left[\frac{\nabla u_h}{Qu_h} \right], \phi \right\rangle_{\Sigma_h} \\ &= \langle r, \phi \rangle + \langle j, \phi \rangle_{\Sigma_h} \qquad \forall \phi \in \mathring{\mathrm{W}}{}^1_1(\Omega), \end{aligned} \tag{4.2}$$

where the residual functions r and j are those introduced in section 3.1.

4.2. Galerkin orthogonality and the error equation. The starting point of our residual-based a posteriori estimation is exploiting the property that $\mathscr{R}$ vanishes

on $\mathring{\mathbb{V}}_h$. This is the so-called *Galerkin orthogonality* property, which yields the following *error equation*:

$$\left\langle \frac{\partial_t u_h}{Qu_h} - \frac{\partial_t u}{Qu}, \phi \right\rangle + \left\langle \frac{\nabla u_h}{Qu_h} - \frac{\nabla u}{Qu}, \nabla\phi \right\rangle = \langle \mathscr{R} \,|\, \phi - \phi_h \rangle \tag{4.3}$$

for all $\phi \in \mathring{\mathrm{W}}^1_1(\Omega)$, $\phi_h \in \mathring{\mathbb{V}}_h$.

4.3. Choice of the test function. The energy technique relies on an appropriate choice of test functions ϕ and ϕ_h in (4.3). Let us denote by e the error

$$e(\boldsymbol{x}, t) := u_h(\boldsymbol{x}, t) - u(\boldsymbol{x}, t) \tag{4.4}$$

and make the following choices for the test functions:

$$\phi(\boldsymbol{x}, t) := \partial_t e(\boldsymbol{x}, t), \tag{4.5}$$

$$\phi_h(\boldsymbol{x}, t) := I_h \phi(\boldsymbol{x}, t), \tag{4.6}$$

where I_h is the Scott–Zhang interpolation operator which will be briefly discussed in section 6.2 and section 6.3. For $\partial_t e$ to be admissible as a test function ϕ in (4.3), it must vanish on $\partial\Omega$, which is not necessarily true. This motivates the following temporary assumption, which will be removed in section 6.3 where we deal with general boundary data.

ASSUMPTION 4.1 (exact boundary data resolution). *Until* section 6.3*, let either*
(a) *the boundary value g be approximated exactly by g_h, or*
(b) *g be time independent.*

5. Coercivity. Our objective in this section is to derive a lower bound on the left-hand side of (4.3) with the choice made in (4.5). To achieve this objective we exhibit as much coercivity as the nonlinearity allows; we will make a liberal use of the word "coercivity" in this sense. The geometric error functions of time A and B, introduced in section 1.2, will be used extensively in this section and in the next one. We begin by stating some simple yet fundamental geometric relations observed by Dziuk.

LEMMA 5.1 (basic geometry [8]). *Given $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \mathbb{R}^d$, if $q_i := (1 + |\boldsymbol{p}_i|^2)^{1/2}$ and $\boldsymbol{n}_i := (\boldsymbol{p}_i; -1)/q_i \in \mathbb{R}^d$ for $i = 1, 2$, then the following geometric relations hold:*

$$1 - \frac{1 + \boldsymbol{p}_1 \cdot \boldsymbol{p}_2}{q_1 q_2} = \frac{1}{2} |\boldsymbol{n}_1 - \boldsymbol{n}_2|^2, \tag{5.1}$$

$$\left| \left(\frac{1}{q_1} - \frac{1}{q_2} \right) \left(\frac{\boldsymbol{p}_1}{q_1} - \frac{\boldsymbol{p}_2}{q_2} \right) \right| \le \frac{1}{2} |\boldsymbol{n}_1 - \boldsymbol{n}_2|^2, \tag{5.2}$$

$$\frac{|\boldsymbol{p}_1 - \boldsymbol{p}_2|}{q_1} \le (1 + |\boldsymbol{p}_2|) |\boldsymbol{n}_1 - \boldsymbol{n}_2|. \tag{5.3}$$

LEMMA 5.2 (Dziuk identity [8]). *If v and w are sufficiently differentiable functions on $\Omega \times [0, T]$, then*

$$\begin{aligned} \frac{1}{2}\partial_t \left(|\boldsymbol{N}v - \boldsymbol{N}w|^2 Qw \right) = {} & \left(\frac{\nabla v}{Qv} - \frac{\nabla w}{Qw} \right) \cdot \nabla \left(\partial_t v - \partial_t w \right) \\ & - \nabla \partial_t v \cdot \left(\frac{\nabla w}{Qv} - \frac{\nabla w}{Qw} + \frac{\nabla v}{Qv} - \frac{1 + \nabla w \cdot \nabla v}{(Qv)^2} \frac{\nabla v}{Qv} \right). \end{aligned} \tag{5.4}$$

The first term in the left-hand side of (4.3) is handled through the following inequality.

LEMMA 5.3 (coercivity of the velocity term). *With the notation*

$$\varrho_1(t) := \frac{1}{2}\left\|\partial_t u_h(t)\right\|^2_{\mathrm{L}_\infty(\Omega)}, \tag{5.5}$$

we have that, for all $t \in [0,T]$,

$$\left\langle \frac{\partial_t u_h}{Qu_h} - \frac{\partial_t u}{Qu}, \partial_t u_h - \partial_t u \right\rangle \geq \frac{1}{2}\,\mathrm{d}_t B(t) - \varrho_1 A(t). \tag{5.6}$$

Proof. Basic manipulations imply

$$\begin{aligned}
&\left\langle \frac{\partial_t u_h}{Qu_h} - \frac{\partial_t u}{Qu}, \partial_t u_h - \partial_t u \right\rangle \\
&\quad = \int_\Omega (Vu_h - Vu)^2 Qu + \int_\Omega \partial_t u_h \left(\frac{1}{Qu} - \frac{1}{Qu_h}\right)(Vu_h - Vu)Qu \\
&\quad \geq \mathrm{d}_t B(t) - \left\|\partial_t u_h\right\|_{\mathrm{L}_\infty(\Omega)} \int_\Omega \left|\frac{1}{Qu} - \frac{1}{Qu_h}\right| \sqrt{Qu}\,|Vu_h - Vu|\,\sqrt{Qu} \\
&\quad \geq \mathrm{d}_t B(t) - \left\|\partial_t u_h\right\|_{\mathrm{L}_\infty(\Omega)} \left(\int_\Omega \left|\boldsymbol{N}u_h - \boldsymbol{N}u\right|^2 Qu\right)^{1/2} \left(\int_\Omega (Vu_h - Vu)^2 Qu\right)^{1/2}.
\end{aligned}$$

Consequently

$$\left\langle \frac{\partial_t u_h}{Qu_h} - \frac{\partial_t u}{Qu}, \partial_t u_h - \partial_t u \right\rangle \geq \mathrm{d}_t B(t) - \frac{1}{2}\,\mathrm{d}_t B(t) - \varrho_1(t)A(t) = \frac{1}{2}\,\mathrm{d}_t B(t) - \varrho_1 A(t),$$

as asserted. ∎

LEMMA 5.4 (coercivity for normals and gradients). *With the notation*

$$\varrho_2(t) := \left\|\nabla \partial_t u_h(t)\right\|_{\mathrm{L}_\infty(\Omega)}, \tag{5.7}$$

we have that, for all $t \in [0,T]$,

$$\left\langle \frac{\nabla u_h}{Qu_h} - \frac{\nabla u}{Qu}, \nabla(\partial_t u_h - \partial_t u) \right\rangle \geq \frac{1}{2}\,\mathrm{d}_t A(t) - \varrho_2(t)A(t). \tag{5.8}$$

Proof. Integrating in space both sides of (5.4) and rearranging terms yields

$$\begin{aligned}
\left\langle \frac{\nabla u_h}{Qu_h} - \frac{\nabla u}{Qu}, \nabla(\partial_t u_h - \partial_t u) \right\rangle &= \frac{1}{2}\,\mathrm{d}_t A(t) \\
&\quad + \int_\Omega \nabla \partial_t u_h \cdot \left(\frac{\nabla u}{Qu_h} - \frac{\nabla u}{Qu} + \frac{\nabla u_h}{Qu_h} - \frac{1 + \nabla u \cdot \nabla u_h}{(Qu_h)^2}\frac{\nabla u_h}{Qu_h}\right).
\end{aligned}$$

To show the result it is sufficient to show that the last integral above is bounded from below by $-\varrho_2(t)A(t)$. To do this we add and subtract $-(Qu\nabla u_h)/(Qu_h)^2$ and rewrite this term as the sum of two integrals:

$$\begin{aligned}
I_1 + I_2 := &\int_\Omega \nabla \partial_t u_h \cdot \left(\frac{\nabla u}{Qu_h} - \frac{\nabla u}{Qu} + \frac{\nabla u_h}{Qu_h} - \frac{Qu\nabla u_h}{(Qu_h)^2}\right) \\
&+ \int_\Omega \nabla \partial_t u_h \cdot \left(\frac{Qu\nabla u_h}{(Qu_h)^2} - \frac{1 + \nabla u \cdot \nabla u_h}{(Qu_h)^2}\frac{\nabla u_h}{Qu_h}\right).
\end{aligned} \tag{5.9}$$

The integrals in (5.9) are bounded, by using (5.2) for the first one,

$$I_1 \geq -\left\|\nabla \partial_t u_h\right\|_{\mathrm{L}_\infty(\Omega)} \int_\Omega \left| \left(\frac{1}{Qu} - \frac{1}{Qu_h} \right) \left(\frac{\nabla u_h}{Qu_h} - \frac{\nabla u}{Qu} \right) Qu \right| \geq -\frac{\varrho_2(t)}{2} A(t),$$

and with the help of (5.1) for the second one,

$$I_2 \geq -\left\|\nabla \partial_t u_h\right\|_{\mathrm{L}_\infty(\Omega)} \int_\Omega \left| \left(1 - \frac{1 + \nabla u_h \cdot \nabla u}{Qu_h Qu} \right) Qu \frac{\nabla u_h}{(Qu_h)^2} \right| \geq -\frac{\varrho_2(t)}{2} A(t).$$

This proves the assertion. ∎

LEMMA 5.5 (estimate of the geometric terms). *With the notation*

$$\varrho(t) := \varrho_1(t) + \varrho_2(t), \tag{5.10}$$

we have that, for all $t \in [0,T]$,

$$A(t) + B(t) \leq A(0) + 2\int_0^t \varrho(s) A(s)\, \mathrm{d}s + 2\int_0^t \langle \mathscr{R}(s) \,|\, \partial_t \left(e(s) - I_h e(s) \right) \rangle \, \mathrm{d}s. \tag{5.11}$$

Proof. Using (4.5), (4.6), (5.6), and (5.8) in (4.3), we obtain

$$\frac{1}{2} \left(\mathrm{d}_t A(t) + \mathrm{d}_t B(t) \right) \leq \langle \mathscr{R} \,|\, \partial_t e(t) - I_h \partial_t e(t) \rangle + \varrho(t) A(t) \tag{5.12}$$

for all $t \in [0,T]$. An integration in time over the interval $[0,t]$ yields the result. ∎

6. Bounding the residual by the estimators. We prove in this section Theorems 3.4 and 3.6 by estimating $\int_0^t \langle \mathscr{R} \,|\, \partial_t \left(e - I_h e \right) \rangle$ appearing in (5.11). We will denote by $d' = d/(d-1)$ the conjugate exponent of d, the latter being the surface's dimension. We start by stating two lemmas bearing a fundamental geometric relationship and an interpolation theory result, respectively.

LEMMA 6.1 (Fierro–Veeser inequality [12]). *Adopting the same notation as in Lemma* 5.1*, the following inequality holds:*

$$\left| \boldsymbol{p}_1 - \boldsymbol{p}_2 \right| \frac{1}{q_1^2} \leq 2 \left| \boldsymbol{n}_1 - \boldsymbol{n}_2 \right| + \left| \boldsymbol{n}_1 - \boldsymbol{n}_2 \right|^2 q_2. \tag{6.1}$$

LEMMA 6.2 (Scott–Zhang interpolation [22]). *If* I_h *denotes the averaging interpolation operator that was introduced by Scott and Zhang—called the* Scott–Zhang interpolator *in what follows—then the following interpolation inequalities hold:*

$$\left\| \psi - I_h \psi \right\|_{\mathrm{L}_{d'}(K)} \leq C_1 \left| \psi \right|_{\mathrm{W}_1^1(\mathscr{U}_K^h)}, \tag{6.2}$$

$$\left\| \psi - I_h \psi \right\|_{\mathrm{L}_1(\partial K)} \leq 2 C_2 \left| \psi \right|_{\mathrm{W}_1^1(\mathscr{U}_K^h)}, \tag{6.3}$$

where $\mathscr{U}_K^h$ *is the* $\mathscr{T}_h$*-neighborhood of* K *defined in* (2.13).

Remark 6.3. The particular choice of the norms in Lemma 6.2 is motivated by our wish for $\sqrt{A(t)}$ to appear in an upper bound on the right-hand side of (5.11).

Indeed, estimating the residual $\mathscr{R}$ in energy norms would typically lead to dealing with $|\nabla u_h - \nabla u|$. In light of the geometric errors A and B in the left-hand side of (5.11), a straightforward idea would be to bound its L_2 norm, that is, $|\nabla u_h - \nabla u|^2$, from above by $C \left| \boldsymbol{N} u_h - \boldsymbol{N} u \right|^2 Qu$, with the constant $C = C[u_h]$ independent of u

(think of u_h being unrelated to u in this paragraph). The only practical way to derive such a bound would be a pointwise geometric relation like

$$\frac{|\boldsymbol{p}_1 - \boldsymbol{p}_2|^2}{\kappa(\boldsymbol{p}_1)\,|\boldsymbol{n}_1 - \boldsymbol{n}_2|^2\, q_2} \le 1, \tag{6.4}$$

where $\boldsymbol{p}_1 = \nabla u_h$, $\boldsymbol{n}_1 = N u_h$, $q_1 = Q u_h$, the quantities with subscript 2 refer to u, and κ is some function of $\boldsymbol{p}_1$ only. Unfortunately this is not possible because (6.4) is false. To see this, fix $\boldsymbol{p}_1$ and observe that $\boldsymbol{n}_1 - \boldsymbol{n}_2$ is bounded; by letting $|\boldsymbol{p}_2| \to \infty$, we obtain, in contrast with (6.4),

$$\frac{|\boldsymbol{p}_1 - \boldsymbol{p}_2|^2}{\kappa(\boldsymbol{p}_1)\,|\boldsymbol{n}_1 - \boldsymbol{n}_2|^2\, q_2} \ge C\frac{|\boldsymbol{p}_1 - \boldsymbol{p}_2|^2}{q_2} = \mathrm{O}(|\boldsymbol{p}_2|) \to \infty. \tag{6.5}$$

This difficulty can be circumvented by using the L_1 norm of $|\nabla u - \nabla u_h|$, instead of the L_2 norm, and the Fierro–Veeser inequality (6.1), which reads

$$\begin{aligned} |\nabla u_h - \nabla u| &= (Q u_h)^2 \frac{|\nabla u_h - \nabla u|}{(Q u_h)^2} \\ &\le (Q u_h)^2 (2\,|\boldsymbol{N} u_h - \boldsymbol{N} u| + |\boldsymbol{N} u_h - \boldsymbol{N} u|^2\, Q u). \end{aligned} \tag{6.6}$$

Notice that the last term is cumbersome because its power is too high—it is the "price to pay." This term will yield a term of the form $\varrho(t)A(t)$ on the right-hand side which has to be handled carefully in order to close the estimate.

Recalling first the notation in section 1.2 and section 3.3, we now state and prove the central result of this paper.

LEMMA 6.4 (residual estimate). *The following inequality holds for all* $t \in [0,T]$:

$$\begin{aligned} A(t) + B(t) &\le \mathscr{E}_0{}^2 + 2\hat{\mathscr{E}}_{2,0}(t) A(t)^{1/2} + 2\hat{\mathscr{E}}_{\infty,0}(t) A(t) \\ &\quad + 2\int_0^t \dot{\mathscr{E}}_{2,1}(s) A(s)^{1/2}\,\mathrm{d}s + 2\int_0^t \dot{\mathscr{E}}_{\infty,1}(s) A(s)\,\mathrm{d}s + 2\int_0^t \varrho(s) A(s)\,\mathrm{d}s. \end{aligned} \tag{6.7}$$

Proof. Apply the representation formula (4.2) with $\phi = \partial_t \delta_h e$, where $\delta_h e(t) := e(t) - I_h e(t)$, integrate by parts in time, and use the commutativity property $\partial_t I_h = I_h \partial_t$, to obtain

$$\begin{aligned} \int_0^t \langle \mathscr{R}(s) \,|\, \partial_t \delta_h e(s) \rangle\,\mathrm{d}s &= \int_0^t \langle r(s), \partial_t \delta_h e(s) \rangle + \langle j(s), \partial_t \delta_h e(s) \rangle_{\Sigma_h}\,\mathrm{d}s \\ &= \big[\langle r, \delta_h e \rangle + \langle j, \delta_h e \rangle_{\Sigma_h}\big]_0^t - \int_0^t \langle \partial_t r(s), \delta_h e(s) \rangle + \langle \partial_t j(s), \delta_h e(s) \rangle_{\Sigma_h}\,\mathrm{d}s. \end{aligned}$$

Hence

$$\begin{aligned} &\int_0^t \langle \mathscr{R}(s) \,|\, \partial_t \delta_h e(s) \rangle\,\mathrm{d}s \\ &\quad \le \sum_{K \in \mathscr{T}_h} \left(\sum_{s \in \{0,t\}} \left(\|r(s)\|_{\mathrm{L}_d(K)}\, \|\delta_h e(s)\|_{\mathrm{L}_{d'}(K)} + \frac{1}{2}\, \|j(s)\|_{\mathrm{L}_\infty(\partial K)}\, \|\delta_h e(s)\|_{\mathrm{L}_1(\partial K)} \right) \right. \\ &\qquad \left. + \int_0^t \|\partial_t r(s)\|_{\mathrm{L}_d(K)}\, \|\delta_h e(s)\|_{\mathrm{L}_{d'}(K)} + \frac{1}{2}\, \|\partial_t j(s)\|_{\mathrm{L}_\infty(\partial K)}\, \|\delta_h e(s)\|_{\mathrm{L}_1(\partial K)}\,\mathrm{d}s \right). \end{aligned}$$

Owing to the approximation properties of the Scott–Zhang interpolator in Lemma 6.2, and using the local indicators η_i^K introduced in Definition 3.2, we may write

$$\int_0^t \langle \mathscr{R}(s) \,|\, \partial_t \delta_h e(s) \rangle \,\mathrm{d}s \le \sum_{K\in\mathscr{T}_h} h_K^{-d/2} \Big(\eta_0^K(0) \,\|\nabla e(0)\|_{\mathrm{L}_1(\mathscr{U}_K^h)} + \eta_0^K(t) \,\|\nabla e(t)\|_{\mathrm{L}_1(\mathscr{U}_K^h)} + \int_0^t \eta_1^K(s) \,\|\nabla e(s)\|_{\mathrm{L}_1(\mathscr{U}_K^h)} \,\mathrm{d}s \Big). \tag{6.8}$$

We proceed by observing that inequality (6.6) implies

$$\|\nabla e(t)\|_{\mathrm{L}_1(\mathscr{U}_K^h)} \le \sup_{\mathscr{U}_K^h} (Qu_h)^2 \int_{\mathscr{U}_K^h} \left(\frac{2\mathscr{N}(t)}{\sqrt{Qu(t)}} + \mathscr{N}(t)^2 \right), \tag{6.9}$$

where, in order to simplify notation, we introduce the shorthand

$$\mathscr{N} := |\boldsymbol{N} u_h - \boldsymbol{N} u| \,\sqrt{Qu}. \tag{6.10}$$

We continue the bound in (6.8) by using (6.9) as follows:

$$\begin{aligned} &\int_0^t \langle \mathscr{R}(s) \,|\, \partial_t e(s) - I_h \partial_t e(s) \rangle \,\mathrm{d}s \\ &\quad \le \sum_{K\in\mathscr{T}_h} \eta_0^K(0) h_K^{-d/2} \omega^K(0) \int_{\mathscr{U}_K^h} \left(\frac{2\mathscr{N}(0)}{\sqrt{Qu(0)}} + \mathscr{N}(0)^2 \right) \\ &\quad + \sum_{K\in\mathscr{T}_h} \eta_0^K(t) h_K^{-d/2} \omega^K(t) \int_{\mathscr{U}_K^h} \left(\frac{2\mathscr{N}(t)}{\sqrt{Qu(t)}} + \mathscr{N}(t)^2 \right) \\ &\quad + \sum_{K\in\mathscr{T}_h} \int_0^t \eta_1^K(s) h_K^{-d/2} \omega^K(s) \int_{\mathscr{U}_K^h} \left(\frac{2\mathscr{N}(s)}{\sqrt{Qu(s)}} + \mathscr{N}(s)^2 \right) \mathrm{d}s. \end{aligned} \tag{6.11}$$

The first two terms in (6.11) can be bounded at once through the following inequality (where we simply take $t = 0$ for the first term):

$$\begin{aligned} &\sum_{K\in\mathscr{T}_h} \eta_0^K(t) h_K^{-d/2} \omega^K(t) \int_{\mathscr{U}_K^h} \left(\frac{2\mathscr{N}(t)}{\sqrt{Qu(t)}} + \mathscr{N}(t)^2 \right) \\ &\quad \le 2 \left(\sum_{K\in\mathscr{T}_h} \eta_0^K(t)^2 h_K^{-d} \omega^K(t)^2 \,|\mathscr{U}_K^h| \sup_{\mathscr{U}_K^h} \frac{1}{Qu(t)} \right)^{1/2} \left(\sum_{K\in\mathscr{T}_h} \int_{\mathscr{U}_K^h} \mathscr{N}(t)^2 \right)^{1/2} \\ &\qquad + \max_{K\in\mathscr{T}_h} \left(\eta_0^K(t) h_K^{-d/2} \omega^K(t) \right) \left(\sum_{K\in\mathscr{T}_h} \int_{\mathscr{U}_K^h} \mathscr{N}(t)^2 \right). \end{aligned}$$

Likewise, the last term in (6.11) is bounded by

$$\begin{aligned} &\int_0^t \left(2 \left(\sum_{K\in\mathscr{T}_h} \eta_1^K(s)^2 h_K^{-d} \omega^K(s)^2 \,|\mathscr{U}_K^h| \sup_{\mathscr{U}_K^h} \frac{1}{Qu(s)} \right)^{1/2} \left(\sum_{K\in\mathscr{T}_h} \int_{\mathscr{U}_K^h} \mathscr{N}(s)^2 \right)^{1/2} \right. \\ &\qquad \left. + \max_{K\in\mathscr{T}_h} \left(\eta_1^K(s) h_K^{-d/2} \omega^K(s) \right) \left(\sum_{K\in\mathscr{T}_h} \int_{\mathscr{U}_K^h} \mathscr{N}(s)^2 \right) \right) \mathrm{d}s. \end{aligned}$$

To conclude the proof, we observe that the shape-regularity of $\mathscr{T}_h$ (2.12) implies the existence of two constants $\gamma_0 \in \mathbb{R}^+$ and $M \in \mathbb{Z}^+$, depending only on σ_0 and the space dimension d, such that the number of simplexes of $\mathscr{T}_h$ contained in $\mathscr{U}_K^h$ does not exceed M and $\left|\mathscr{U}_K^h\right| \leq M\gamma_0^2 h_K^d$. Defining $\gamma := 2M\gamma_0$, it follows that

$$\begin{aligned}
&\int_0^t \langle \mathscr{R}(s) \,|\, \partial_t \delta_h e(s) \rangle \, \mathrm{d}s \\
&\quad \leq \gamma \left(\sum_{K \in \mathscr{T}_h} \alpha^K(0) \eta_0^K(0)^2 \right)^{1/2} A(0)^{1/2} + M \max_{K \in \mathscr{T}_h} \left(h_K^{-d/2} \omega^K(0) \eta_0^K(0) \right) A(0) \\
&\quad + \gamma \left(\sum_{K \in \mathscr{T}_h} \alpha^K(t) \eta_0^K(t)^2 \right)^{1/2} A(t)^{1/2} + M \max_{K \in \mathscr{T}_h} \left(h_K^{-d/2} \omega^K(t) \eta_0^K(t) \right) A(t) \\
&\quad + \gamma \int_0^t \left(\sum_{K \in \mathscr{T}_h} \alpha^K(s) \eta_1^K(s)^2 \right)^{1/2} A(s)^{1/2} \, \mathrm{d}s + M \int_0^t \max_{K \in \mathscr{T}_h} \left(h_K^{-d/2} \omega^K(s) \eta_1^K(s) \right) A(s) \, \mathrm{d}s.
\end{aligned}$$

Recalling Definition 3.3, we combine the last inequality with (5.11) and obtain (6.7), as asserted. ∎

Next we prove the theorems stated in section 3 with the aid of Lemma 6.4. For (6.7) to be useful we must control the terms containing $A(t)$ on the right-hand side by those on the left-hand side. We distinguish two main ways of doing this. The first way, which is direct and somewhat naive, uses the stability Lemma 2.4 and leads to the unconditional a posteriori estimate in Theorem 3.4. The second, more careful, way results in the conditional but sharper estimate in Theorem 3.6. To shorten the discussion, we first show the latter and then the former, which is simpler.

6.1. Proof of Theorem 3.6. Our starting point is inequality (6.7). Introduce the notation $A^*(t) := \sup_{[0,t]} A$, apply the Hölder inequality, and use the Young inequality with a parameter μ at our disposal, to obtain

$$\begin{aligned}
A(t) + B(t) \leq {\mathscr{E}_0}^2 + \mu A(t) + \frac{1}{\mu} \hat{\mathscr{E}}_{2,0}(t)^2 + \mu A^*(t) + \frac{1}{\mu} \mathscr{E}_{2,1}(t)^2 \\
+ 2\hat{\mathscr{E}}_{\infty,0}(t) A(t) + 2A^*(t) \mathscr{E}_{\infty,1}(t) + 2 \int_0^t \varrho(s) A(s) \, \mathrm{d}s.
\end{aligned} \tag{6.12}$$

Choosing $\mu = 1/8$; taking the supremum over $[0,t]$ on both sides; recalling that B, $\mathscr{E}_{2,1}$, and $\mathscr{E}_{\infty,1}$ are nondecreasing; and using Definition 3.3, we can write

$$A^*(t) + B(t) \leq \mathscr{E}_0^2 + \frac{1}{4} A^*(t) + 8\mathscr{E}_2(t)^2 + 2\mathscr{E}_\infty A^*(t) + 2 \int_0^t \varrho(s) A^*(s) \, \mathrm{d}s. \tag{6.13}$$

The condition (3.17), i.e., $\mathscr{E}_\infty \leq 1/8$, and the last inequality imply

$$\frac{1}{2} A^*(t) + B(t) \leq {\mathscr{E}_0}^2 + 8\mathscr{E}_2(t)^2 + 2 \int_0^t \varrho(s) A^*(s) \, \mathrm{d}s. \tag{6.14}$$

To conclude the proof, it suffices now to apply the Gronwall lemma in the above inequality, and to recall (5.10), (5.5), and (5.7), in order to derive (3.18) and (3.19). ∎

6.2. Proof of Theorem 3.4. The proof is a direct combination of Lemma 6.4 and the elementary fact that

$$A(t) = \int_\Omega |\boldsymbol{N}u_h(t) - \boldsymbol{N}u(t)|^2 \, Qu(t) \le 4 \int_\Omega Qu(t). \tag{6.15}$$

The stability Lemma 2.4 provides us with an upper bound on the last integral in terms of the data f and g. To conclude, it is enough to proceed along the lines of section 6.1 with $\mu = 1/4$ and apply the Gronwall lemma. ∎

Remark 6.5 (slowly varying solutions). Notice that if $\int_0^t \varrho$ is small enough (for which it is necessary for $\|\partial_t u_h\|_{\mathrm{L}_1(\mathrm{W}^1_\infty)}$ to be small), the Gronwall lemma argument is not needed and the exponential bound on C can be dropped. This is particularly true for solutions that are close to stationary points, i.e., if $\partial_t f$ and $\partial_t g$ are very small. We will not pursue this issue further in this paper, but we remark that this condition is also a posteriori and could be checked automatically if needed.

6.3. Time-dependent Dirichlet boundary data. As promised earlier, we now remove Assumption 4.1; that is, we allow

$$\partial_t (u_h - u)|_{\partial\Omega} = \partial_t (g_h - g) \neq 0. \tag{6.16}$$

We study the case where the boundary value g is discretized as follows:

$$\tilde{g}_h := I_h \tilde{g} \quad \text{and} \quad g_h := \tilde{g}_h|_{\partial\Omega}\,, \tag{6.17}$$

where I_h it the Scott–Zhang interpolator of Lemma 6.2 and $\tilde{g}$ denotes the extension of g to the whole domain Ω [22, eq. (5.5)]. The error $e = u_h - u$ can thus be decomposed as follows:

$$e = e_0 + \epsilon := (u_h - \tilde{g}_h - u + \tilde{g}) + (\tilde{g}_h - \tilde{g}). \tag{6.18}$$

The residual $\mathscr{R}$, as defined in (4.1), can be naturally extended to be a functional on $\mathrm{W}^1_1(\Omega)$. It follows that if we take $\phi = \partial_t e$ in (4.3), we have

$$\langle \mathscr{R} \,|\, \partial_t e \rangle = \langle \mathscr{R} \,|\, \partial_t e_0 \rangle + \langle \mathscr{R} \,|\, \partial_t \epsilon \rangle = \langle \mathscr{R} \,|\, \partial_t e_0 - I_h \partial_t e_0 \rangle + \langle \mathscr{R} \,|\, \partial_t \epsilon \rangle\,. \tag{6.19}$$

Notice that a Galerkin orthogonality argument can be applied directly to the part with the admissible error $e_0 \in \mathring{\mathrm{W}}^1_1$. As for the last term in (6.19), we use the $\mathbb{V}_h$-invariance property of the Scott–Zhang interpolator I_h, namely $I_h I_h \psi = I_h \psi$ for all $\psi \in \mathrm{W}^1_1(\Omega)$, and (6.17) to conclude that

$$I_h \epsilon = I_h \tilde{g}_h - I_h \tilde{g} = I_h I_h \tilde{g} - I_h \tilde{g} = 0.$$

This implies that $\partial_t I_h \epsilon = 0$, and thus $\langle \mathscr{R} \,|\, \partial_t \epsilon \rangle = \langle \mathscr{R} \,|\, \partial_t (\epsilon - I_h \epsilon) \rangle$, whence the following representation formula follows from (6.19) and elementwise integration by parts:

$$\langle \mathscr{R} \,|\, \partial_t e \rangle = \langle r, \partial_t e - I_h \partial_t e \rangle + \langle j, \partial_t e - I_h \partial_t e \rangle + \langle \beta - \beta_h, \partial_t \epsilon \rangle_{\partial\Omega}\,, \tag{6.20}$$

where $\beta := (\nabla u \cdot \boldsymbol{\nu})/Qu$ and $\beta_h = (\nabla u_h \cdot \boldsymbol{\nu})/Qu_h$.

In order to obtain a lower bound on the left-hand side of (6.20), which is equal to the left-hand side of (4.3) with $\phi = \partial_t e$, we proceed in the same fashion as in section 5 and thereby we again derive (5.11). The first two terms on the right-hand

side of (6.20) can be dealt with exactly as in section 6, while the fact that $\beta, \beta_h \leq 1$ implies the following bound for the last term:

$$\langle \beta - \beta_h, \partial_t \epsilon \rangle_{\partial\Omega} \leq 2 \left\| \partial_t \epsilon \right\|_{\mathrm{L}_1(\partial\Omega)} = 2 \left\| \partial_t g - \partial_t g_h \right\|_{\mathrm{L}_1(\partial\Omega)}. \tag{6.21}$$

This proves the following generalization of Lemma 6.4.

LEMMA 6.6 (residual estimate with boundary values). *With the notation* $\mathscr{E}_\partial(t) := \int_0^t \left\| \partial_t (g - g_h) \right\|_{\mathrm{L}_1(\partial\Omega)}$, *we have that, for all* $t \in [0,T]$,

$$\begin{aligned} A(t) + B(t) &\leq {\mathscr{E}_0}^2 + 2\mathscr{E}_\partial(t) + 2\hat{\mathscr{E}}_{2,0}(t) A(t)^{1/2} + 2\hat{\mathscr{E}}_{\infty,0} A(t) \\ &\quad + 2\int_0^t \dot{\mathscr{E}}_{2,1}(s) A(s)^{1/2} \, \mathrm{d}s + 2\int_0^t \dot{\mathscr{E}}_{\infty,1}(s) A(s) \, \mathrm{d}s + 2\int_0^t \varrho(s) A(s) \, \mathrm{d}s. \end{aligned} \tag{6.22}$$

This lemma enables us to obtain extended versions of Theorems 3.6 and 3.4 by just adding $\mathscr{E}_\partial$ to the estimators therein. We omit the statement of these results as they can be written in a straightforward manner.

7. Numerical experiments. We now present some numerical computations that we have performed in order to confirm the reliability and test the sharpness of the error estimates derived in Theorems 3.6 and 3.4. Many of the comments in this section are given as figure captions in order to make the reading easier.

DEFINITION 7.1 (fully discrete semi-implicit scheme [7, 8]). *Let* $N \in \mathbb{Z}^+$ *and* $0 = t_0 < t_1 < \cdots < t_N = T$ *be a partition of the time interval* $[0,T]$. *For each* $n \in [1:N]$, *denote by* $\tau_n := t_n - t_{n-1}$ *the* n*th step size. Given* U_h^0 *(an approximation of* $g(0)$*) and* $\tilde{g}_h^n$ *(the extension to* Ω *of an interpolant of* $g(t_n)$*), find a sequence of functions* $U_h^n \in \mathbb{V}_h$ *such that, for each* $n \in [1:N]$,

$$\left\langle \frac{\nabla U_h^n}{Q U_h^{n-1}}, \nabla \phi_h \right\rangle + \left\langle \frac{U_h^n}{\tau_n Q U_h^{n-1}}, \phi_h \right\rangle = \left\langle \frac{U_h^{n-1}}{\tau_n Q U_h^{n-1}} + f^n, \phi_h \right\rangle \quad \forall \phi_h \in \mathring{\mathbb{V}}_h, \tag{7.1}$$

$$U_h^n - \tilde{g}_h^n \in \mathring{\mathbb{V}}_h. \tag{7.2}$$

We implemented this scheme, which is due to Dziuk [8], with the help of the C finite element toolbox ALBERT of Schmidt and Siebert [21]. All the computations are based on piecewise linear ($\mathbb{P}^1$) finite elements.

7.1. Main goal of the numerical results. With reference to Definition 3.3, we introduce the *full proper estimator* defined as $\widetilde{\mathscr{E}} := ({\mathscr{E}_0}^2 + {\mathscr{E}_2}^2)^{1/2}$, and we recall that we denote by $\mathscr{E}_\infty$ the *vicinity estimator*, by $\mathscr{E}$ the *total estimator*, and by E the *geometric error*, introduced in (1.9). With this notation the *unconditional estimate* of Theorem 3.4 can be written as

$$E \leq C\mathscr{E} = C\left(\widetilde{\mathscr{E}}^2 + C' \mathscr{E}_\infty \right)^{1/2}, \tag{7.3}$$

while the *conditional estimate* provided by Theorem 3.6 can be summarized as follows:

$$\mathscr{E}_\infty \leq c \quad \Rightarrow \quad E \leq \widetilde{C} \widetilde{\mathscr{E}}. \tag{7.4}$$

The main goal of our numerical experiments is to see that the error bound (7.4) is sharp whereas (7.3) is not. This will be illustrated by comparing the *experimental order of convergence (EOC)* of E, $\widetilde{\mathscr{E}}$, and $\mathscr{E}_\infty^{1/2}$. The EOC is defined as follows: for a given finite sequence of uniform triangulations $\{\mathscr{T}_{h_i}\}_{i=1,\ldots,I}$ of meshsize h_i, the EOC

of a corresponding sequence of some triangulation-dependent quantity $e(i)$ (like an error or an estimator) is itself a sequence defined as

$$\operatorname{EOC} e(i) = \frac{\log(e(i+1)/e(i))}{\log(h_{i+1}/h_i)}. \tag{7.5}$$

Notice that for (7.4) to be sharp it is sufficient to have $\operatorname{EOC} E \approx \operatorname{EOC} \widetilde{\mathscr{E}}$ and $\mathscr{E}_\infty = \mathrm{o}(1)$, as i increases—this will be satisfied in our numerical tests—whereas for (7.3) to be sharp it is necessary to have the stronger requirement that $\operatorname{EOC} E \approx \operatorname{EOC} \widetilde{\mathscr{E}}$ and $\operatorname{EOC} E \approx \operatorname{EOC} \mathscr{E}_\infty^{1/2}$—this will fail in our numerical tests. We will focus also on understanding when $\mathscr{E}_\infty = \mathrm{o}(1)$ might fail and on computing the *effectivity index*, which is a practical bound on the constant $\widetilde{C}$, and is defined as $E/\widetilde{\mathscr{E}}$, at the finest level I. Since we view the errors and the estimators as functions of time, the EOC and the effectivity index are also presented as functions of time.

Remark 7.2 (practical version of the error estimators). To test the reliability and the sharpness of the upper bound given by the estimators, we compute a fully discrete version of the spatially discrete global estimators introduced in Definition 3.3. These estimators are sums of the local indicators

$$\eta_i^K(t) := h_K^{d/2}(C_1 \left\|(\partial_t)^i r(t)\right\|_{\mathrm{L}_d(K)} + C_2 \left\|(\partial_t)^i j(t)\right\|_{\mathrm{L}_\infty(\partial K)}), \quad i = 0, 1, \tag{7.6}$$

which involve the L_∞ norm that is not so practical. Since we use piecewise linear elements, the jump residuals are constant functions on each edge, and thus the L_∞ norm can be replaced by the L_2 norm using the inverse estimate

$$\|v\|_{\mathrm{L}_\infty(\partial K)} \le C h_K^{(1-d)/2} \|v\|_{\mathrm{L}_2(\partial K)} \tag{7.7}$$

for all v that are constants on each edge of ∂K. It is hence legitimate to use, instead of η_K^i, the handier local indicators

$$\bar{\eta}_i^K(t) := h_K^{d/2} C_1 \left\|(\partial_t)^i r(t)\right\|_{\mathrm{L}_d(K)} + h_K^{1/2} C_2 \left\|(\partial_t)^i j(t)\right\|_{\mathrm{L}_2(\partial K)}, \quad i = 0, 1. \tag{7.8}$$

All the integrals are in fact quadratures: while ALBERT's built-in Gaussian quadrature is used to approximate the space integrals, a simple midpoint rule is used for the time integrals. Time derivatives are replaced by backward finite differences.

Remark 7.3 (the discrete initial condition). In our computations, we take the *minimal surface projection* for the discrete initial values, i.e., $U_h^0 := u_h(0) = M_h g(0)$, where $M_h v$ is defined, for each $v \in \mathrm{W}_1^1(\Omega)$, as the unique function in $\mathbb{V}_h$ such that

$$\left\langle \frac{\nabla M_h v}{Q M_h v}, \nabla \phi_h \right\rangle = \left\langle \frac{\nabla v}{Qv}, \nabla \phi_h \right\rangle \qquad \forall \phi_h \in \mathring{\mathbb{V}}_h, \tag{7.9}$$

and that interpolates v on the boundary.

This choice of the discrete initial value reduces the initial transients that can occur with other choices for the discrete initial values such as Lagrange interpolation.

Example 7.4 (smooth exact solution on a square). Our first series of tests use the following exact solution as a benchmark:

$$u(x, y; t) = t(\sin(t) - \sin(t - x(1-x)y(1-y))), \quad (x, y, t) \in [0, 1]^2 \times [0, 8]. \tag{7.10}$$

The function u is smooth, it has zero initial and boundary values, which allows us to focus on the effect of the estimators only, and it is the solution of Problem 1.1

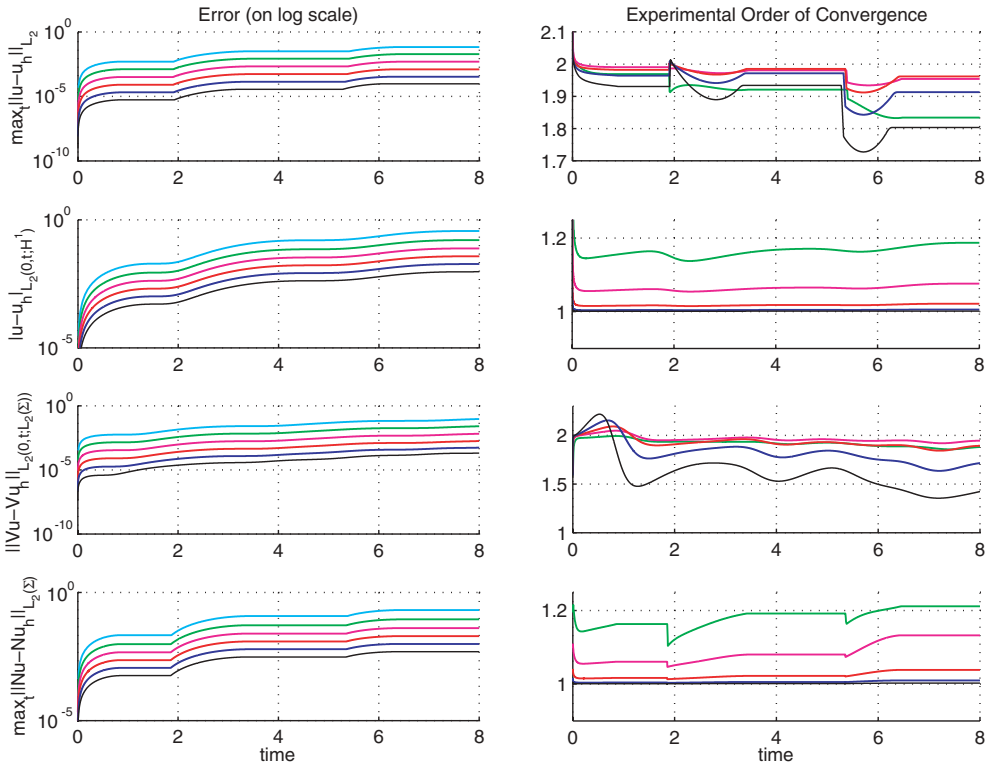


FIG. 7.1. *Errors and EOC vs. time for Example* 7.4. *In the left column, we plot the errors in the customary Sobolev norms, related to the heat equation, and the geometric energy error and normal velocity error introduced in Definition* 1.2. *In the right column, we plot the corresponding EOC. The different gray tones, from light to dark, correspond to the decreasing meshsizes* h. *Notice that the behavior of the Sobolev energy norm error* $|e|_{\mathrm{L}_2(\mathrm{H}^1_0)}$ *and the geometric energy error is similar and that both have EOC close to* 1.

where the right-hand side f is obtained by applying the differential operator of (1.2) on u. We performed a series of computations, on uniform meshes, with the meshsizes $h_i = (0.5)^i$ for $1 \leq i \leq 6$. We report the results in the form of graphs, where the abscissa always denotes the time variable; this allows us to track the behavior of the errors and estimators in time. Figure 7.1 shows the behavior of the exact spatial errors, namely, the geometric errors and those in the customary Sobolev norms for evolution equations. Figure 7.2 shows the behavior of the proper and vicinity estimators with respect to time.

As shown by the right-hand subfigure in Figure 7.1, the $\mathrm{EOC}\, E \approx 1$—this is to be expected from the a priori results, derived in the case of smooth solutions [8]. Although the normal velocity error tends to decrease faster, the geometric error decreases like the geometric energy error, which has order 1.

The sharpness of estimate (7.4) can be seen from the fact that $\mathrm{EOC}\, \widetilde{\mathscr{E}} \approx 1$ and that $\mathscr{E}_\infty \to 0$. On the other hand, we notice that $\mathrm{EOC}\, \mathscr{E}_\infty \leq 1$, which implies that $\mathrm{EOC}(\mathscr{E}_\infty^{1/2}) \prec \mathrm{EOC}\, E$, and thus indicates that the unconditional estimate (7.3) is not sharp.

The effectivity index $\widetilde{C}$, relative to the estimate (7.4), is plotted in Figure 7.3(a) as a function of time. In this example, the effectivity index is bounded in time, and

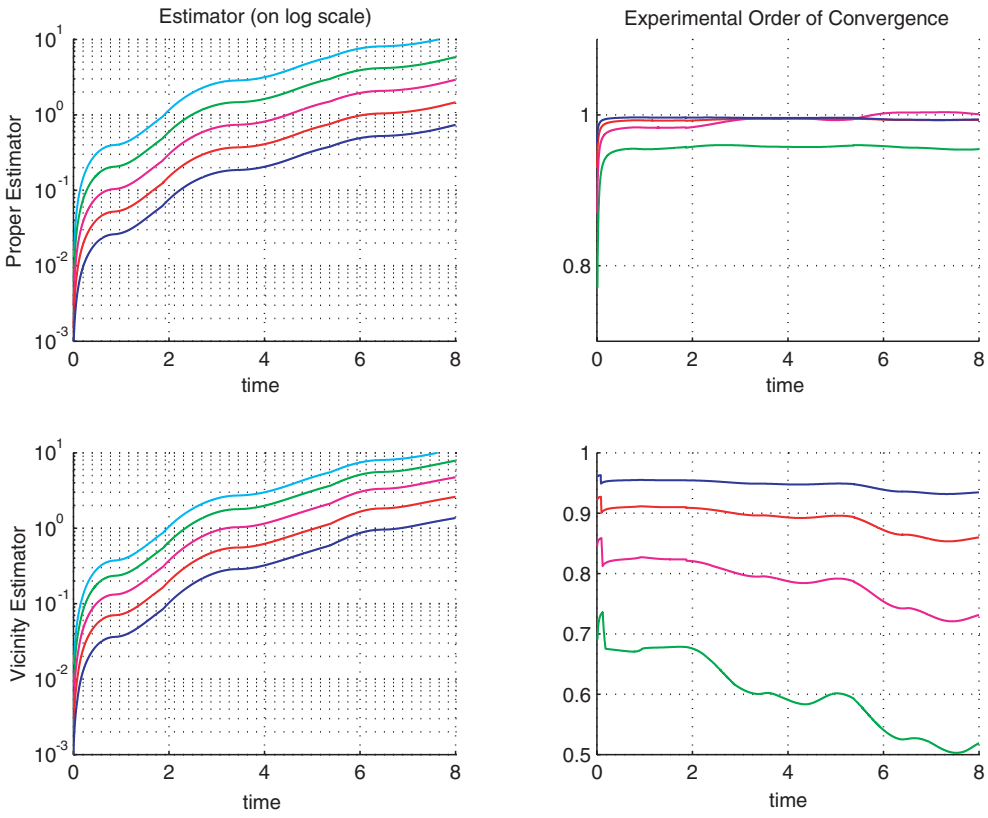


FIG. 7.2. *Estimators and EOC vs. time for Example* 7.4 *for* $i \geq 2$. *The upper panels show the proper estimator* $\widetilde{\mathscr{E}}$ *and the corresponding EOC (notice that in this particular figure we do not plot the proper estimator for the first meshsize* $h = 0.5$*, for reasons of clarity). The lower panels exhibit the vicinity estimator behavior which is seen to converge to zero. According to Figure* 7.1*, we have* $\mathrm{EOC}\, E \approx 1$ *and, according to the current figure, we have* $\mathrm{EOC}\, \widetilde{\mathscr{E}} \approx 1$ *which means that the proper estimator is sharp as expected. Notice also that, for the vicinity estimator, we have* $\mathrm{EOC}\, \mathscr{E}_\infty \approx 0.95 \leq 1$*, which implies that* $\mathrm{EOC}\, \mathscr{E} \leq 1/2$*: this is strong numerical evidence that the unconditional estimate* (7.3) *cannot be sharp, in that the estimators decay with a much lower order than the errors, and justifies the need for the sharper conditional result of Theorem* 3.6.

we do not detect the exponential behavior predicted by the worst-case-scenario bound in (3.19).

Example 7.5 (shrinking spherical segment). This second numerical example is inspired by a simple geometric situation. A sphere that moves by mean curvature flow shrinks to a point in finite time [14]. If we assume that the initial radius of the sphere equals 2 and that the center is fixed at $(0, 0, 0)$, then the segment of the surface that lies above the square $[0, 1] \times [0, 1] \times \{0\} \in \mathbb{R}^3$ is the graph of the function

$$u(\boldsymbol{x}, t) = \sqrt{4 - 4t - |\boldsymbol{x}|^2}, \quad (\boldsymbol{x}, t) \in [0, 1]^2 \times [0, 0.5]. \tag{7.11}$$

The function u thus constitutes a solution of Problem 1.1 with zero right-hand side f and time-dependent Dirichlet boundary value g. This is an interesting example because of a blow-up of the gradient which occurs at the space-time boundary point $(1, 1; 1/2)$.

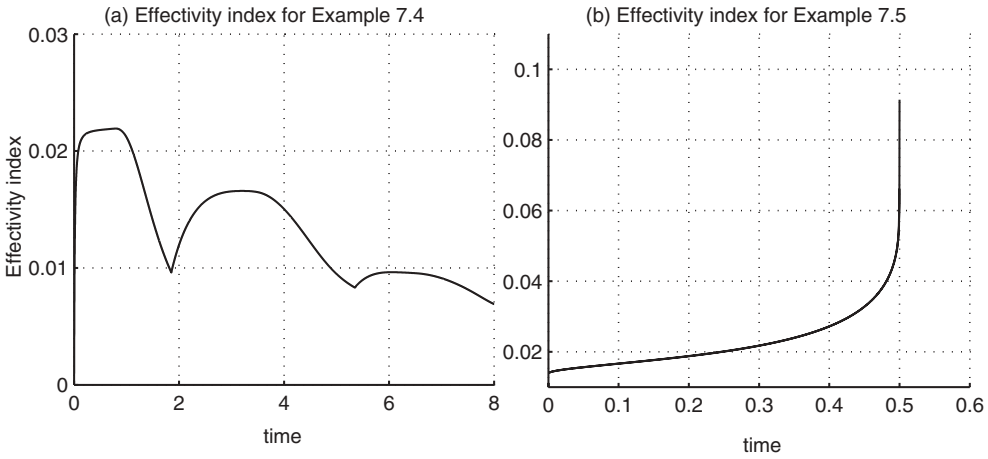


FIG. 7.3. *Effectivity indexes for Examples* 7.4 *and* 7.5. *These indexes, which are defined for the finest mesh, are numerical realizations of the constants on the right-hand side of* (3.16) *or* (3.18). *Panel* (a) *refers to the smooth exact solution of Example* 7.4*; the effectivity index behaves well in time. Panel* (b) *shows the effectivity index for the proper estimator* 7.5*, which has a blow-up at time* $t = 0.5$. *Consequently, the exponential behavior predicted for the factor* C *in Theorem* 3.6 *might be sharp. The behavior of the graph in* (b) *close to* $t = 0.5$ *is to be taken with care, though, as the vicinity estimator blows up there according to Figure* 7.5*, and thus the conditional estimate is not guaranteed to hold anymore.*

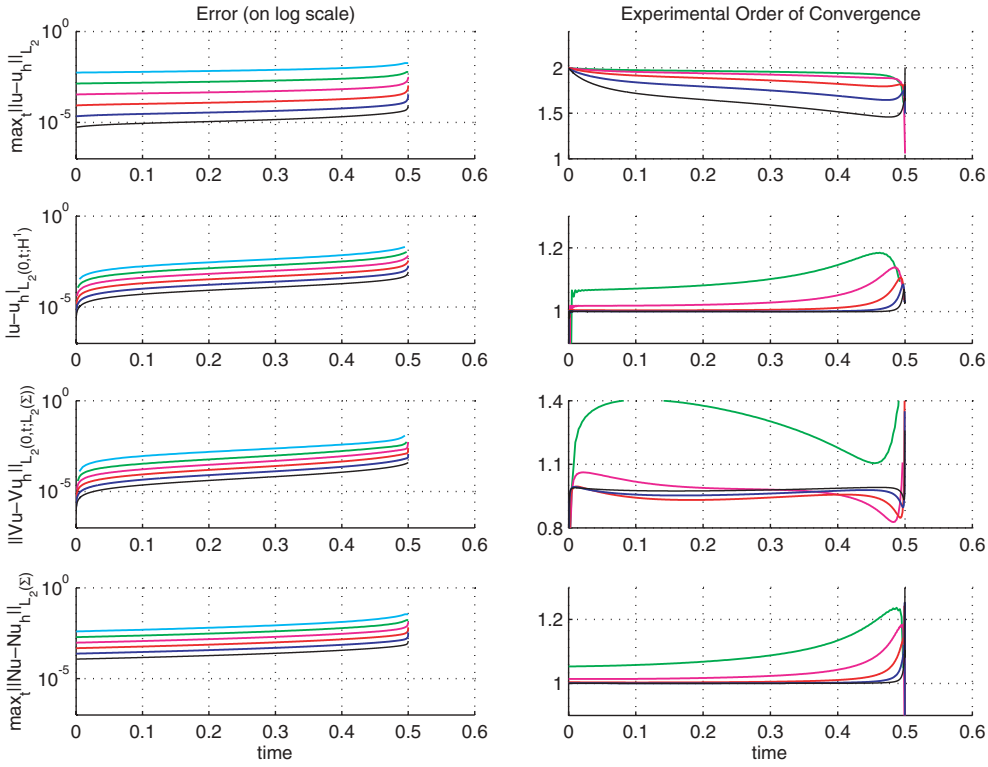


FIG. 7.4. *Sobolev norm errors and geometric errors for the shrinking sphere of Example* 7.5. *The different gray tones correspond to decreasing meshsize* h. *In this example a blow-up in the gradient occurs at the boundary at time* $t = 0.5$.

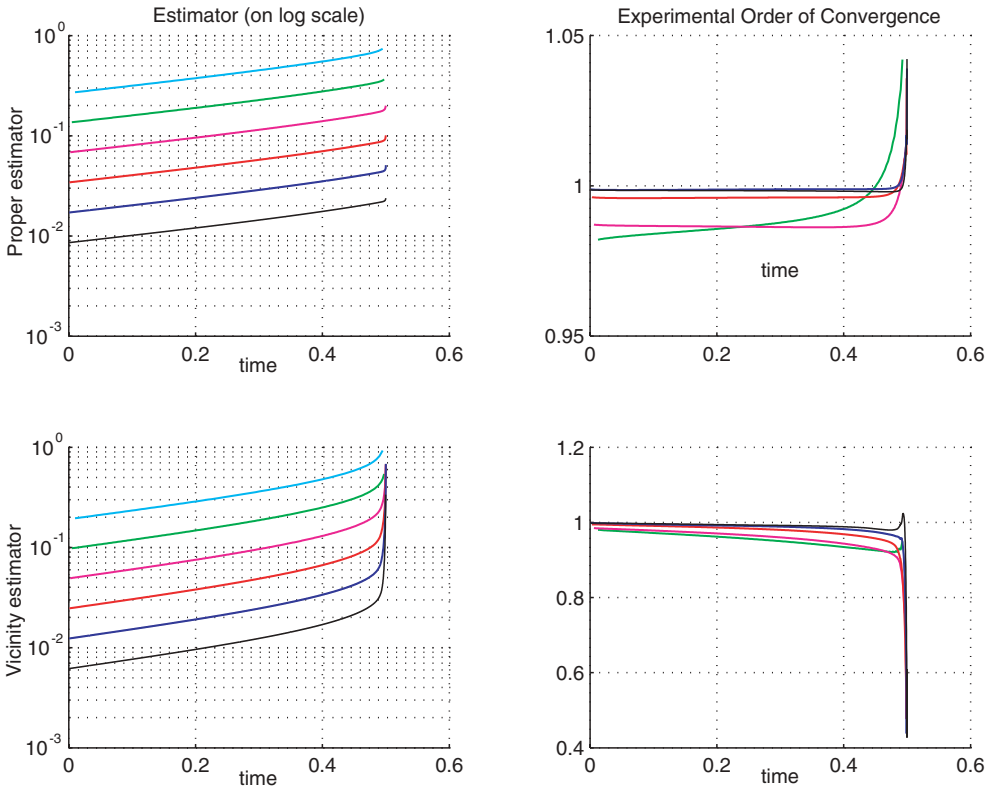


FIG. 7.5. *Estimators and EOC vs. time for the shrinking sphere segment of Example* 7.5. *We exhibit the behavior of the proper estimator in the upper row and that of the vicinity estimator in the lower row. Darker gray tones correspond to decreasing meshsizes* h. *We can observe two stages as time approaches the blow-up* $t = 0.5$. *In the first stage, the same observations made for Example* 7.4 *are valid in that* $\mathrm{EOC}\,\widetilde{\mathscr{E}} \approx \mathrm{EOC}\,E$ *and* $\mathscr{E}_\infty \to 0$ *(justifying once more the need for Theorem* 3.6*). In the second stage, the vicinity estimator* $\mathscr{E}_\infty$ *exhibits a blow-up, which means that the condition* (3.17) *of Theorem* 3.6 *is violated and that we can no longer rely on the proper estimator* $\widetilde{\mathscr{E}}$. *The vicinity estimator blow-up can be interpreted as numerical evidence of the boundary gradient blow-up occurring at* $t = 0.5$.

Despite this singular behavior, the function u still satisfies Assumption 2.1, and our a posteriori error analysis applies. Notice that the a priori error analysis of Deckelnick and Dziuk [7, Prop. 3] does not apply in this case because of the overly stringent regularity assumptions. This example allows us to appreciate the exponential worst-case-scenario bound on $\widetilde{C}$ (factor C in Theorem 3.6), as that bound is expected to behave like $\exp(1/\sqrt{0.5-t})$ as $t \to 0.5$. Numerical solutions have been computed on uniform triangulations with meshsizes $h_i = (0.5)^i$, $i = 2, \ldots, 7$. The type of data we report is similar to that in section 7.4: the errors and their asymptotic behavior are reported in Figure 7.4, while Figure 7.5 shows the behavior of the estimators. We refer to the caption for a comment on the blow-up at $t = 0.5$ and its effect on the estimators and estimate validity. In Figure 7.3(b) we report the effectivity index of the proper estimator, which justifies in part the exponential behavior predicted by the theory. Notice that because of the blow-up behavior, the effectivity index is not so meaningful in the last part of the graph, close to $t = 0.5$, where the vicinity estimator is too big.

Acknowledgments. We express our thanks to Gerhard Dziuk for fruitful discussions, Kunibert Siebert and Alfred Schmidt for introducing us to ALBERT [21], and Claudio Verdi for his advice and support during the stay of O. Lakkis at the Dipartimento di Matematica, Università di Milan, Italy. We are also grateful to the Isaac Newton Institute, Cambridge, UK, where this work was finished. The program *Computational Challenges in PDEs* 2003 at the Isaac Newton Institute provided excellent support as well as a stimulating and friendly scientific environment. Finally, we thank the referees for their careful reading and constructive criticism.

REFERENCES

[1] L. Ambrosio, *Lecture Notes on Geometric Evolution Problems, Distance Function, and Viscosity Solutions*, Pubblicazioni 1029, Istituto di Analisi Numerica del CNR, Pavia, Italy, 1997.

[2] I. Babuška and W. C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[3] E. Bänsch, *Finite element discretization of the Navier–Stokes equations with a free capillary surface*, Numer. Math., 88 (2001), pp. 203–235.

[4] G. Barles and P. E. Souganidis, *A new approach to front propagation problems: Theory and applications*, Arch. Ration. Mech. Anal., 141 (1998), pp. 237–296.

[5] K. A. Brakke, *The Motion of a Surface by Its Mean Curvature*, Princeton University Press, Princeton, NJ, 1978.

[6] K. Deckelnick, *Error bounds for a difference scheme approximating viscosity solutions of mean curvature flow*, Interfaces Free Bound., 2 (2000), pp. 117–142.

[7] K. Deckelnick and G. Dziuk, *Error estimates for a semi-implicit fully discrete finite element scheme for the mean curvature flow of graphs*, Interfaces Free Bound., 2 (2000), pp. 341–359.

[8] G. Dziuk, *Numerical schemes for the mean curvature flow of graphs*, in Variations of Domain and Free-Boundary Problems in Solid Mechanics (Paris, 1997), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 63–70.

[9] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* I. *A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

[10] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* IV. *Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.

[11] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature.* I, J. Differential Geom., 33 (1991), pp. 635–681.

[12] F. Fierro and A. Veeser, *On the a posteriori error analysis for equations of prescribed mean curvature*, Math. Comp., 72, (2003), pp. 1611–1634.

[13] M. Fried and A. Veeser, *Simulation and numerical analysis of dendritic growth*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, Springer, Berlin, 2001, pp. 225–252 (with color plates on pp. 812–813).

[14] G. Huisken, *Asymptotic behavior for singularities of the mean curvature flow*, J. Differential Geom., 31 (1990), pp. 285–299.

[15] G. Huisken, *Local and global behaviour of hypersurfaces moving by mean curvature*, in Differential Geometry: Partial Differential Equations on Manifolds (Los Angeles, CA, 1990), AMS, Providence, RI, 1993, pp. 175–191.

[16] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific Publishing, River Edge, NJ, 1996.

[17] R. H. Nochetto, G. Savaré, and C. Verdi, *A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations*, Comm. Pure Appl. Math., 53 (2000), pp. 525–589.

[18] R. H. Nochetto, A. Schmidt, and C. Verdi, *A posteriori error estimation and adaptivity for degenerate parabolic problems*, Math. Comp., 69 (2000), pp. 1–24.

[19] M. Picasso, *Adaptive finite elements for a linear parabolic problem*, Comput. Methods Appl. Mech. Engrg., 167 (1998), pp. 223–237.

[20] A. Schmidt, *Computation of three dimensional dendrites with finite elements*, J. Comput. Phys., 125 (1996), pp. 293–312.

[21] A. Schmidt and K. G. Siebert, *ALBERT—Software for scientific computations and applications*, Acta Math. Univ. Comenian. (N.S.), 70 (2000), pp. 105–122.

[22] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[23] J. Sethian and S. J. Osher, *The design of algorithms for hypersurfaces moving with curvature-dependent speed*, in Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications (Aachen, 1988), Vieweg, Braunschweig, 1989, pp. 544–551.
[24] J. E. Taylor, J. W. Cahn, and C. A. Handwerker, *Geometric models of crystal growth*, Acta Metall. Mater., 40 (1992), pp. 1443–1474.
[25] N. N. Uraltseva, *Boundary regularity for flows of nonparametric surfaces driven by mean curvature*, in Motion by Mean Curvature and Related Topics (Trento, 1992), G. Buttazzo and A. Visintin, eds., De Gruyter, Berlin, 1994, pp. 198–209.
[26] R. Verfürth, *A Review of* A Posteriori *Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester-Stuttgart, 1996.
[27] R. Verfürth, *A posteriori error estimates for nonlinear problems: $L^r(0,T;W^{1,\rho}(\omega))$-error estimates for finite element discretizations of parabolic equations*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 487–518.
[28] N. J. Walkington, *Algorithms for computing motion by mean curvature*, SIAM J. Numer. Anal., 33 (1996), pp. 2215–2238.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1899–1913

CONVERGENCE OF THE BINOMIAL TREE METHOD FOR AMERICAN OPTIONS IN A JUMP-DIFFUSION MODEL*

XIAO-SONG QIAN†, CHENG-LONG XU‡, LI-SHANG JIANG‡, AND BAO-JUN BIAN‡

Abstract. The paper studies the binomial tree method for American options in a jump-diffusion model. We employ the theory of viscosity solution to show uniform convergence of the binomial tree method for American options. We also prove existence and convergence of the optimal exercise boundary in the binomial tree approximation. In addition, the terminal value of the optimal exercise boundary is given for American options in jump-diffusion models.

Key words. binomial tree method, explicit difference method, American option, jump-diffusion model, integrodifferential equation, viscosity solution

AMS subject classifications. 65M12, 35R35, 45K05, 49L25, 91B24

DOI. 10.1137/S0036142902409744

1. Introduction.

Consider a financial market with two assets (B_t, S_t). The first is a risk-free asset whose price B_t is governed by the equation $dB_t = rB_t dt$, where r is the constant positive interest rate; the other is a risky stock. In a given probability space $(\Omega, \mathcal{F}, \mathrm{P})$, the stock price evolves according to the stochastic differential equation

$$\frac{dS_t}{S_t} = (\mu - q)dt + \sigma dW_t + d\left(\sum_{j=1}^{N_t} U_j\right), \tag{1.1}$$

where the coefficients μ, q, σ are positive constants, q is the dividend yield, $(W_t)_{t\geq 0}$ is a standard Brownian motion, $(N_t)_{t\geq 0}$ is a Poisson process with constant intensity λ, and the sequence $(U_j)_{j\geq 1}$ are square integrable independently and identically distributed (i.i.d.) random variables taking values in $(-1, +\infty)$ (since the price of a financial asset should be positive).

Let K be the striking price of option, let S be the price of stock and $z^+ = \max\{z, 0\}$. Then the payoff of an American option is $\varphi(S)$, where $\varphi(S) = (S-K)^+$ (call options) or $\varphi(S) = (K-S)^+$(put options). Because the market here is not complete, according to Harrison and Kreps [7], each equivalent martingale measure Q defines an admissible price of contingent claim. To simplify analysis and focus attention on early exercise considerations, we assume $Q = P$; then we must have $\mu = r - \lambda k$, where $k = E(U_1)$ and E is the expectation operator over the random variable U_1 (see [6]). Using an argument similar to that of Pham [6], it can be shown that an American option's price $V_t = V(S, t)$ solves the following parabolic variational

*Received by the editors June 18, 2002; accepted for publication (in revised form) January 19, 2004; published electronically January 20, 2005. This work was supported by National Science Foundation of China grant 10171078.

http://www.siam.org/journals/sinum/42-5/40974.html

†Department of Mathematics, Yangzhou University, Yangzhou 225002, People's Republic of China and Department of Applied Mathematics, Tongji University, Shanghai 200092, People's Republic of China (qian@yzu.edu.cn).

‡Department of Applied Mathematics, Tongji University, Shanghai 200092, People's Republic of China and Division of Computational Science, E-Institute of Shanghai University, Shanghai 200436, People's Republic of China (clxu601@online.sh.cn, jianglsk@online.sh.cn, bianbj@263.net). The work of the second author was supported partly by the E-Institute of Shanghai Municipal Education Commission (N.E03004).

inequality:

$$\text{(1.2)}\qquad \begin{cases} \min\{-\hat{L}V(S,t), V(S,t)-\varphi(S)\}=0, & S>0,\ 0\le t<T,\\ V(S,T)=\varphi(S), & S>0, \end{cases}$$

where $\hat{L}$ is the parabolic integrodifferential operator

$$\hat{L}V=\frac{\partial V}{\partial t}+\frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2}+(r-q-\lambda k)S\frac{\partial V}{\partial S}-(r+\lambda)V+\lambda\int_{-1}^{\infty}V(S(1+y),t)dN(y)$$

and $N(y)$ is the distribution function of random variable U_1.

There is usually no explicit formula for the price of American options, so numerical methods for computing the price of American options are necessary and important. The binomial tree method, as a discrete time model first proposed by Cox, Ross, and Rubinstein [4] is the most popular numerical approach to pricing options in diffusion models. Amin [2] generalized their algorithm to jump-diffusion models. Xu, Qian, and Jiang [10] gave an optimal error estimation of European options in Amin's model. In essence, the binomial tree method belongs to the probabilistic one; however, it can be proved that the binomial tree method is consistent with a certain explicit difference scheme. By virtue of the notion of viscosity solution (see [5]), Barles, Daher, and Romano [3] and Jiang and Dai [8] presented a framework to prove the convergence of difference schemes for parabolic equations in diffusion models. However, in jump-diffusion models, the equations associated to option pricing are so-called integrodifferential equations. Fortunately, Alvarez and Tourin [1] developed the notion of viscosity solution for second order integrodifferential equations. The main purpose of this paper is to use numerical analysis and the theory of viscosity solution to prove the convergence of Amin's binomial tree method for American options in jump-diffusion models. We also show existence and convergence of optimal exercise boundary in the binomial tree approximation.

The rest of this paper is organized as follows. In the next section, the equivalence of the binomial tree method and an explicit difference scheme are discussed. Section 3 is devoted to the convergence proof of the binomial tree method. In section 4, we prove the existence and convergence of approximate optimal exercise boundary. Finally, we give a proof of comparison principle, which is used in section 3.

2. Binomial tree method. First, we recall the binomial tree method developed by Amin [2] when the underlying asset follows a jump-diffusion process. Let $\mathbf{Z}=\{l: l=0,\pm1,\pm2,\ldots\}$, let N be the number of discrete time points, $\Delta t=\frac{T}{N}$, $\rho=e^{r\Delta t}=1+r\Delta t+O(\Delta t^2)$, and let V_j^n be the option price at time point $n\Delta t$ with stock price $e^{j\sigma\sqrt{\Delta t}}$. Then we have (see Amin [2])

$$\text{(2.1)}\qquad \begin{cases} V_j^n=\max\Big\{\frac{1}{\rho}\big[(1-\lambda\Delta t)\left(pV_{j+1}^{n+1}+(1-p)V_{j-1}^{n+1}\right)+\lambda\Delta t\sum_{l\in\mathbf{Z}}V_{j+l}^{n+1}\hat{p}_l\big],\\ \qquad\qquad (e^{j\sigma\sqrt{\Delta t}}-K)^+\Big\}, \qquad j\in\mathbf{Z},\ 0\le n\le N-1,\\ V_j^N=(e^{j\sigma\sqrt{\Delta t}}-K)^+, \qquad j\in\mathbf{Z}, \end{cases}$$

where

$$p=\frac{\frac{e^{(r-q)\Delta t}-\lambda\Delta t\sum_{l\in\mathbf{Z}}e^{l\sigma\sqrt{\Delta t}}\hat{p}_l}{1-\lambda\Delta t}-e^{j\sigma\sqrt{\Delta t}}}{e^{\sigma\sqrt{\Delta t}}-e^{-\sigma\sqrt{\Delta t}}}$$

and

$$\begin{aligned}\hat{p}_l &= \text{Prob}\left(\ln(1+U_1) \in [(l-\tfrac{1}{2})\sigma\sqrt{\Delta t}, (l+\tfrac{1}{2})\sigma\sqrt{\Delta t})\right)\\ &= N(\mathrm{e}^{(l+\frac{1}{2})\sigma\sqrt{\Delta t}} - 1) - N(\mathrm{e}^{(l-\frac{1}{2})\sigma\sqrt{\Delta t}} - 1).\end{aligned}$$

In view of [2], we have

$$\begin{aligned}&\textstyle\sum_{l\in\mathbf{Z}} e^{l\sigma\sqrt{\Delta t}}\hat{p}_l = 1 + k + O(\Delta t),\\ &p = \frac{1}{2} + \left(r - q - \lambda k - \frac{\sigma^2}{2}\right)\frac{\sqrt{\Delta t}}{2\sigma} + O(\Delta t^{\frac{3}{2}}).\end{aligned} \tag{2.2}$$

Then we deduce from (2.1) and (2.2) that

$$\begin{aligned}V_j^n = \max\Bigg\{ &\frac{1-\lambda\Delta t}{1+r\Delta t + O(\Delta t^2)}\left[\left(\frac{1}{2} + \left(r - q - \lambda k - \frac{\sigma^2}{2}\right)\frac{\sqrt{\Delta t}}{2\sigma} + O(\Delta t^{\frac{3}{2}})\right)V_{j+1}^{n+1}\right.\\ &\left. + \left(\frac{1}{2} - \left(r - q - \lambda k - \frac{\sigma^2}{2}\right)\frac{\sqrt{\Delta t}}{2\sigma} + O(\Delta t^{\frac{3}{2}})\right)V_{j-1}^{n+1}\right]\\ &+ \frac{\lambda\Delta t}{1+r\Delta t + O(\Delta t^2)}\sum_{l\in\mathbf{Z}} V_{j+l}^{n+1}\hat{p}_l, \quad (e^{j\sigma\sqrt{\Delta t}} - K)^+\Bigg\}.\end{aligned} \tag{2.3}$$

Next, we discuss the equivalence of the binomial tree method and explicit difference scheme. Let $\mathbf{R}$ be the set of all real numbers, $Q_T = \{(x,t) : 0 \le t < T, x \in \mathbf{R}\}$. Using a simple transformation $S = \mathrm{e}^x$, $v(x,t) = V(S,t)$, (1.2) changes to the constant-coefficients problem

$$\begin{cases} \min\Big\{ - Lv(x,t), v(x,t) - \varphi(e^x)\Big\} = 0, & (x,t) \in Q_T,\\ v(x,T) = \varphi(e^x), & x \in \mathbf{R},\end{cases} \tag{2.4}$$

where L is the operator

$$Lv = \frac{\partial v}{\partial t} + \frac{\sigma^2}{2}\frac{\partial^2 v}{\partial x^2} + \left(r - q - \lambda k - \frac{\sigma^2}{2}\right)\frac{\partial v}{\partial x} - (r+\lambda)v + \lambda\int_{\mathbf{R}} v(x+y,t)d\tilde{N}(y),$$

and $\tilde{N}(y) = N(\mathrm{e}^y - 1)$.

We now present an explicit finite difference scheme for (2.4). Given mesh size $\Delta x, \Delta t, N\Delta t = T$, let $Q_h = \{(n\Delta t, j\Delta x) : 0 \le n \le N, j \in \mathbf{Z}\}$ stand for the set of lattice, v_j^n represent the value of numerical approximation of $v(x,t)$ at $(n\Delta t, j\Delta x) \in Q_h$, and $\varphi_i = \varphi(e^{j\Delta x}) = (e^{j\Delta x} - K)^+$. Then we have

$$\begin{cases} \min\left\{ -\dfrac{v_j^{n+1} - v_j^n}{\Delta t} - \dfrac{\sigma^2}{2}\dfrac{v_{j+1}^{n+1} - 2v_j^{n+1} + v_{j-1}^{n+1}}{\Delta x^2} - \left(r - q - \lambda k - \dfrac{\sigma^2}{2}\right)\dfrac{v_{j+1}^{n+1} - v_{j-1}^{n+1}}{2\Delta x}\right.\\ \qquad \left. + (r+\lambda)v_j^n - \lambda\displaystyle\sum_{l\in\mathbf{Z}} v_{j+l}^{n+1}p_l,\ v_j^n - \varphi_j\right\} = 0, \qquad j \in \mathbf{Z},\ 0 \le n \le N-1,\\ v_j^N = \varphi_j, \qquad j \in \mathbf{Z},\end{cases} \tag{2.5}$$

or

$$v_j^n = \max\left\{ \frac{1}{1+r\Delta t}\frac{1}{1+\frac{\lambda\Delta t}{1+r\Delta t}}\left((1-\frac{\sigma^2\Delta t}{\Delta x^2})v_j^{n+1} + (\frac{\sigma^2\Delta t}{2\Delta x^2} + \left(r-q-\lambda k-\frac{\sigma^2}{2}\right)\frac{\Delta t}{2\Delta x})v_{j+1}^{n+1} \right.\right.$$
$$\left.\left. + (\frac{\sigma^2\Delta t}{2\Delta x^2} - (r-q-\lambda k-\frac{\sigma^2}{2})\frac{\Delta t}{2\Delta x})v_{j-1}^{n+1}\right) + \frac{1}{1+r\Delta t}\cdot\frac{\lambda\Delta t}{1+\frac{\lambda\Delta t}{1+r\Delta t}}\sum_{l\in\mathbf{Z}} v_{j+l}^{n+1}p_l,\ \varphi_i \right\},$$

where

$$p_l = \int_{(l-\frac{1}{2})\Delta x}^{(l+\frac{1}{2})\Delta x} d\tilde{N}(y) = N(e^{(l+\frac{1}{2})\Delta x}-1) - N(e^{(l-\frac{1}{2})\Delta x}-1).$$

If $\sigma^2\Delta t/\Delta x^2 = 1$, then we have

$$p_l = \hat{p}_l \tag{2.6}$$

and

$$\begin{cases} v_j^n = \max\left\{ \frac{1-\lambda\Delta t+O(\Delta t^2)}{1+r\Delta t}\left[(\frac{1}{2}+(r-q-\lambda k-\frac{\sigma^2}{2})\frac{\sqrt{\Delta t}}{2\sigma})v_{j+1}^{n+1}\right.\right. \\ \qquad\left.\left. + (\frac{1}{2}-(r-q-\lambda k-\frac{\sigma^2}{2})\frac{\sqrt{\Delta t}}{2\sigma})v_{j-1}^{n+1}\right] + \frac{\lambda\Delta t+O(\Delta t^2)}{1+r\Delta t}\sum_{l\in\mathbf{Z}} v_{j+l}^{n+1}p_l,\ \varphi_j \right\}, \\ \qquad\qquad j\in\mathbf{Z},\ 0\le n\le \mathbf{N}-1, \\ v_j^N = \varphi_j = (e^{j\sigma\sqrt{\Delta t}}-K)^+, \qquad j\in\mathbf{Z}. \end{cases} \tag{2.7}$$

Comparing (2.3) and (2.7), we deduce the following result.

THEOREM 2.1. *The binomial tree method* (2.1) *is equivalent to the explicit difference scheme* (2.5) *with* $\frac{\sigma^2\Delta t}{\Delta x^2} = 1$ *in the sense of neglecting a higher order* $\Delta t^{\frac{3}{2}}$.

3. Convergence. This section is devoted to the convergence of the binomial tree method. From now on, we will concentrate on American call option. It is easy to generalize all main results to American put option.

Without loss of generality, we assume $\sigma^2\Delta t/\Delta x^2 = 1$ and

$$0 < p < 1, \tag{3.1}$$

which always holds for Δt small enough. In addition, we should assume $q > 0$, because when $q = 0$, an American call option is reduced to the European option.

Now we investigate the properties of the binomial tree method (2.1).

LEMMA 3.1. *The binomial tree method* (2.1) *has the following properties:*

(1) $V_j^n \le V_{j+1}^n$ *for all* j, n.

(2) $V_j^{n+1} \le V_j^n$ *for all* $n < N,\ j$.

(3) $V_j^n \le e^{j\Delta x}$ *for all* j, n.

(4) $V_j^n - V_i^n \le e^{j\Delta x} - e^{i\Delta x}$ *for all* $i \le j,\ n$.

Proof. We use the induction to prove the properties.

(1) Clearly, $V_j^N = \varphi_j \le \varphi_{j+1} = V_{j+1}^N$. If $V_j^{m+1} \le V_{j+1}^{m+1}$ for all j, then

$$\begin{aligned} V_j^m &= \max\left\{\frac{1}{\rho}\left[(1-\lambda\Delta t)(pV_{j+1}^{m+1}+(1-p)V_{j-1}^{m+1})+\lambda\Delta t\sum_{l\in\mathbf{Z}}V_{j+l}^{m+1}p_l\right],\ \varphi_j\right\} \\ &\le \max\left\{\frac{1}{\rho}\left[(1-\lambda\Delta t)(pV_{j+2}^{m+1}+(1-p)V_{j}^{m+1})+\lambda\Delta t\sum_{l\in\mathbf{Z}}V_{j+1+l}^{m+1}p_l\right],\ \varphi_{j+1}\right\} \\ &= V_{j+1}^m, \end{aligned}$$

which is the desired result.

(2) By (2.1), $V_j^{N-1} \ge \varphi_j = V_j^N$. If $V_j^{m+1} \ge V_j^{m+2}$ for all j, then

$$\begin{aligned} V_j^m &= \max\left\{\frac{1}{\rho}\left[(1-\lambda\Delta t)(pV_{j+1}^{m+1}+(1-p)V_{j-1}^{m+1})+\lambda\Delta t\sum_{l\in\mathbf{Z}}V_{j+l}^{m+1}p_l\right],\ \varphi_j\right\} \\ &\ge \max\left\{\frac{1}{\rho}\left[(1-\lambda\Delta t)(pV_{j+1}^{m+2}+(1-p)V_{j-1}^{m+2})+\lambda\Delta t\sum_{l\in\mathbf{Z}}V_{j+l}^{m+2}p_l\right],\ \varphi_j\right\} \\ &= V_j^{m+1}. \end{aligned}$$

Property (2) is proved.

(3) Let V^m denote $\{V_j^m\}_{j\in\mathbf{Z}}$. We introduce a weighted norm $\|\cdot\|$ as follows:

$$\|V^m\| = \sup_{j\in\mathbf{Z}} |e^{-j\Delta x}V_j^m|.$$

It suffices to show that $\|V^m\| \le 1$ for all m. Clearly, $\|V^N\| = \sup_{j\in\mathbf{Z}}|e^{-j\Delta x}\varphi_j| = 1$. If $\|V^{m+1}\| \le 1$, then

$$\begin{aligned} e^{-j\Delta x}V_j^m &= \max\left\{\tfrac{1}{\rho}\left[(1-\lambda\Delta t)(pe^{\Delta x}e^{-(j+1)\Delta x}V_{j+1}^{m+1}+e^{-\Delta x}e^{-(j-1)\Delta x}(1-p)V_{j-1}^{m+1})\right.\right. \\ &\qquad \left.\left. +\ \lambda\Delta t\sum_{l\in\mathbf{Z}}e^{l\Delta x}e^{-(j+l)\Delta x}V_{j+l}^{m+1}p_l\right],\ e^{-j\Delta x}\varphi_j\right\} \\ &\le \max\left\{\frac{\|V^{m+1}\|}{\rho}\left[(1-\lambda\Delta t)(pe^{\Delta x}+(1-p)e^{-\Delta x})+\lambda\Delta t\sum_{l\in\mathbf{Z}}e^{l\Delta x}p_l\right],\ 1\right\}. \end{aligned}$$

Due to

$$pe^{\Delta x}+(1-p)e^{-\Delta x} = \frac{e^{(r-q)\Delta t}-\lambda\Delta t\sum_{l\in\mathbf{Z}}e^{l\Delta x}p_l}{1-\lambda\Delta t}, \tag{3.2}$$

we have

$$\begin{aligned} \|V^m\| &\le \max\left\{\frac{\|V^{m+1}\|}{\rho}e^{(r-q)\Delta t},\ 1\right\} \\ &\le \max\{e^{-q\Delta t},\ 1\} \le 1, \end{aligned}$$

which yields the desired result.

(4) For $i \le j$,

$$V_j^N - V_i^N = (e^{j\Delta x} - K)^+ - (e^{i\Delta x} - K)^+ = \begin{cases} e^{j\Delta x} - e^{i\Delta x} & \text{if } e^{j\Delta x} \ge e^{i\Delta x} \ge K, \\ e^{j\Delta x} - K \le e^{j\Delta x} - e^{i\Delta x} & \text{if } e^{j\Delta x} \ge K \ge e^{i\Delta x}, \\ 0 \le e^{j\Delta x} - e^{i\Delta x} & \text{if } K \ge e^{j\Delta x} \ge e^{i\Delta x}. \end{cases}$$

So we deduce

$$V_j^N - V_i^N = \varphi_j - \varphi_i \le e^{j\Delta x} - e^{i\Delta x} \quad \text{for all } i \le j. \tag{3.3}$$

Suppose $V_j^{m+1} - V_i^{m+1} \le e^{j\Delta x} - e^{i\Delta x}$ for all $i \le j$. To simplify writing, we set

$$I_j^m = (1 - \lambda\Delta t)(pV_{j+1}^m + (1-p)V_{j-1}^m) + \lambda\Delta t \sum_{l\in\mathbf{Z}} V_{j+l}^m p_l.$$

Then

$$\begin{aligned} &\tfrac{1}{\rho} I_j^{m+1} - \tfrac{1}{\rho} I_i^{m+1} \\ &= \tfrac{1}{\rho}\left[[(1-\lambda\Delta t)\big(p(V_{j+1}^{m+1} - V_{i+1}^{m+1}) + (1-p)(V_{j-1}^{m+1} - V_{i-1}^{m+1})\big) \right. \\ &\qquad \left. + \lambda\Delta t \sum_{l\in\mathbf{Z}} (e^{(j+l)\Delta x} - e^{(i+l)\Delta x}) p_l \right] \\ &\le \tfrac{1}{\rho}(e^{j\Delta x} - e^{i\Delta x}) \left[(1-\lambda\Delta t)\big(pe^{j\Delta x} + (1-p)e^{-j\Delta x}\big) + \lambda\Delta t \sum_{l\in\mathbf{Z}} e^{l\Delta x} p_l \right] \\ &= e^{-q\Delta t}\big[e^{j\Delta x} - e^{i\Delta x}\big] \\ &\le e^{j\Delta x} - e^{i\Delta x} \qquad \text{for all } \ i \le j, \end{aligned} \tag{3.4}$$

where the third equality follows from (3.2). So we have

$$V_j^m - V_i^m = \begin{cases} \frac{1}{\rho} I_j^{m+1} - \frac{1}{\rho} I_i^{m+1} & \text{if } \frac{1}{\rho} I_j^{m+1} \ge \varphi_j \text{ and } \frac{1}{\rho} I_i^{m+1} \ge \varphi_i, \\ \varphi_j - \varphi_i & \text{if } \frac{1}{\rho} I_j^{m+1} \le \varphi_j \text{ and } \frac{1}{\rho} I_i^{m+1} \le \varphi_i, \\ \frac{1}{\rho} I_j^{m+1} - \varphi_i \le \frac{1}{\rho} I_j^{m+1} - \frac{1}{\rho} I_i^{m+1} & \text{if } \frac{1}{\rho} I_j^{m+1} \ge \varphi_j \text{ and } \frac{1}{\rho} I_i^{m+1} \le \varphi_i, \\ \varphi_j - \frac{1}{\rho} I_i^{m+1} \le \varphi_j - \varphi_i & \text{if } \frac{1}{\rho} I_j^{m+1} \le \varphi_j \text{ and } \frac{1}{\rho} I_i^{m+1} \ge \varphi_i. \end{cases}$$

In all cases we can deduce from (3.3) and (3.4) that

$$V_j^m - V_i^m \le e^{j\Delta x} - e^{i\Delta x} \qquad \text{for all } \ i \le j,$$

which is the desired result. The proof is completed. □

To simplify notation, (2.1) can be written as

$$V^n = F(\Delta t)(V^{n+1}),$$

where $V^m = \{V_j^m\}_{j\in\mathbf{Z}}$. It is easy to check that $F(\Delta t)$ has the following properties:

LEMMA 3.2. (1) $F(\Delta t)$ *is monotone, i.e.,*

$$F(\Delta t)U \le F(\Delta t)V \qquad \text{if } U \le V.$$

(2) $F(\Delta t)$ *can commute with nonnegative constants, i.e.,*

$$F(\Delta t)(V+K) \le F(\Delta t)V + K, \quad K \ge 0.$$

Let $u_{\Delta t}(x,t)$ be defined as the extension function of V_j^n as follows:

$$u_{\Delta t}(x,t) = \begin{cases} V_j^n \ \text{ for } \ x \in \big[(j-\frac{1}{2})\Delta x, (j+\frac{1}{2})\Delta x\big), t \in \big[(n-\frac{1}{2})\Delta t, (n+\frac{1}{2})\Delta t\big), \\ \qquad j \in \mathbf{Z}, 1 \le n \le N-1, \\ V_j^N \text{ for } \ x \in \big[(j-\frac{1}{2})\Delta x, (j+\frac{1}{2})\Delta x\big), t \in \big[(N-\frac{1}{2})\Delta t, T\big), \ j \in \mathbf{Z}, \\ V_j^0 \ \text{ for } \ x \in \big[(j-\frac{1}{2})\Delta x, (j+\frac{1}{2})\Delta x\big), t \in \big[0, \frac{1}{2}\Delta t\big), \ j \in \mathbf{Z}. \end{cases}$$

By definition and Lemma 3.1(3), we have

$$u_{\Delta t}(x,t) = \big(F(\Delta t)u_{\Delta t}(\,\cdot\,, t+\Delta t)\big)(x) \quad \text{for all } (x,t) \in \mathbf{R} \times [0, T-\Delta t]$$

and

$$0 \le u_{\Delta t}(x,t) \le e^{x+\frac{\Delta x}{2}} \quad \text{for small } \Delta t. \tag{3.5}$$

We will show the convergence of the binomial tree method for American options.

THEOREM 3.3. *As $\Delta t \to 0$, we have that $u_{\Delta t}(x,t)$ converges locally uniformly to $u(x,t)$ in $\mathbf{R} \times [0,T]$, where $u(x,t)$ is the unique solution to the problem* (2.4).

The proof of Theorem 3.3 relies on the notion of viscosity solution. It will be useful to have the notations

$$\mathrm{USC}\big(\mathbf{R} \times [0,T]\big) = \big\{\text{upper semicontinuous functions } u : \ \mathbf{R} \times [0,T] \to \mathbf{R}\big\},$$
$$\mathrm{LSC}\big(\mathbf{R} \times [0,T]\big) = \big\{\text{lower semicontinuous functions } u : \ \mathbf{R} \times [0,T] \to \mathbf{R}\big\}.$$

DEFINITION 3.4. (1) *Any $u \in \mathrm{USC}(\mathbf{R} \times [0,T])(\mathrm{LSC}(\mathbf{R} \times [0,T]))$ is a viscosity subsolution (supersolution) of* (2.4) *if*

$$\begin{cases} \min\Big\{-\dfrac{\partial \phi}{\partial t} - \dfrac{\sigma^2}{2}\dfrac{\partial^2 \phi}{\partial x^2} - \left(r - q - \lambda k - \dfrac{\sigma^2}{2}\right)\dfrac{\partial \phi}{\partial x} + (r+\lambda)u \\ \qquad\qquad - \lambda \displaystyle\int_{\mathbf{R}} \phi(x+y,t)d\tilde{N}(y), u - (e^x - K)^+\Big\} \le 0 \ (\ge 0), \quad (x,t) \in \mathbf{R} \times [0,T), \\ u(x,T) \le \ (\ge) \ (e^x - K)^+, \qquad x \in \mathbf{R}, \end{cases}$$

whenever $\phi \in C^2(\mathbf{R} \times [0,T])$ and $u - \phi$ has a global maximum (minimum) at $(x,t) \in \mathbf{R} \times [0,T)$.

(2) *u is a viscosity solution of* (2.4) *if it is simultaneously a subsolution and a supersolution.*

Now we establish the comparison principle for problem (2.4).

THEOREM 3.5. *Suppose u and v are viscosity subsolution and supersolution of* (2.4) *and*

$$|u(x,t)|, |v(x,t)| \le e^x.$$

Then

$$u \le v \qquad \text{on } \mathbf{R} \times (0,T].$$

The proof of this theorem will be given in the appendix of this paper.

Remark 3.6. By Theorem 3.5, we obtain the uniqueness of problem (2.4) immediately.

Proof of Theorem 3.3. Denote

$$u^*(x,t) = \limsup_{\Delta t \to 0,(y,s)\to(x,t)} u_{\Delta t}(y,s),$$

$$u_*(x,t) = \liminf_{\Delta t \to 0,(y,s)\to(x,t)} u_{\Delta t}(y,s).$$

Owing to (3.5), u^* and u_* are well defined and

$$0 \le u_*(x,t) \le u^*(x,t) \le e^x. \tag{3.6}$$

Obviously, $u^* \in \mathrm{USC}(\mathbf{R}\times[0,T])$ and $u_* \in \mathrm{LSC}(\mathbf{R}\times[0,T])$. If we can show u^* and u_* are subsolution and supersolution of (2.4), respectively, then in terms of Theorem 3.5, we deduce $u^* \le u_*$ and thus $u^* = u_* = u(x,t)$, which guarantees that the whole sequence converges to the unique viscosity solution $u(x,t)$.

We only need to show that u^* is a subsolution of (2.4). It can be shown that $u^*(x,T) = (e^x - K)^+$. Suppose that for $\phi \in C^2(\mathbf{R}\times[0,T))$, $u^* - \phi$ attains a global maximum at $(x_0,t_0) \in \mathbf{R}\times[0,T)$ and $(u^*-\phi)(x_0,t_0) = 0$. We then should assume that (x_0,t_0) is a strict global maximum. Set $\Phi = \phi - \varepsilon, \varepsilon > 0$; then $u^* - \Phi$ attains a strict global maximum at (x_0,t_0) and

$$(u^* - \Phi)(x_0,t_0) > 0. \tag{3.7}$$

By the definition of u^*, there exists a sequence $u_{\Delta t_m}(y_m,s_m)$ such that

$$\Delta t_m \to 0,\ (y_m,s_m)\to(x_0,t_0),\ u_{\Delta t_m}(y_m,s_m)\to u^*(x_0,t_0) \qquad \text{as } m\to+\infty.$$

Assuming that $(\hat y_m,\hat s_m)$ is a global maximum point of $u_{\Delta t_m} - \Phi$ on $\mathbf{R}\times[0,T)$, we can deduce that there is a subsequence $u_{\Delta t_{m_i}}(\hat y_{m_i},\hat s_{m_i})$ such that

$$\Delta t_{m_i} \to 0,\ (\hat y_{m_i},\hat s_{m_i}) \to (x_0,t_0),\ (u_{\Delta t_{m_i}} - \Phi)(\hat y_{m_i},\hat s_{m_i}) \to (u^*-\Phi)(x_0,t_0) \tag{3.8}$$

as $m_i \to +\infty$. Indeed, suppose $(\hat y_{m_i},\hat s_{m_i}) \to (\hat y,\hat s)$; then

$$\begin{aligned}(u^*-\Phi)(x_0,t_0) &= \lim_{m_i\to\infty}(u_{\Delta t_{m_i}} - \Phi)(y_{m_i},s_{m_i})\\ &\le \lim_{m_i\to\infty}(u_{\Delta t_{m_i}} - \Phi)(\hat y_{m_i},\hat s_{m_i}) \le (u^*-\Phi)(\hat y,\hat s),\end{aligned}$$

which forces $(\hat y,\hat s) = (x_0,t_0)$ since (x_0,t_0) is a strict global maximum point of $u^* - \Phi$. Therefore,

$$(u_{\Delta t_{m_i}} - \Phi)(\,\cdot\,,\hat s_{m_i} + \Delta t_{m_i}) \le (u_{\Delta t_{m_i}} - \Phi)(\hat y_{m_i},\hat s_{m_i}) \quad \text{in } \mathbf{R},$$

that is,

$$u_{\Delta t_{m_i}}(\,\cdot\,,\hat s_{m_i} + \Delta t_{m_i}) \le \Phi(\,\cdot\,,\hat s_{m_i} + \Delta t_{m_i}) + (u_{\Delta t_{m_i}} - \Phi)(\hat y_{m_i},\hat s_{m_i}) \quad \text{in } \mathbf{R}. \tag{3.9}$$

From (3.7) and (3.8), we also deduce the important fact

$$(u_{\Delta t_{m_i}} - \Phi)(\hat y_{m_i},\hat s_{m_i}) > 0 \tag{3.10}$$

when m_i is large enough.

Then by (3.7)–(3.10) and Lemma 3.2, we have

$$\begin{aligned} u_{\Delta t_{m_i}}(\hat{y}_{m_i},\hat{s}_{m_i}) &= \Big(F(\Delta t_{m_i})u_{\Delta t_{m_i}}(\,\cdot\,,\hat{s}_{m_i}+\Delta t_{m_i})\Big)(\hat{y}_{m_i}) \\ &\le \Big(F(\Delta t_{m_i})\big(\Phi(\,\cdot\,,\hat{s}_{m_i}+\Delta t_{m_i})+(u_{\Delta t_{m_i}}-\Phi)(\hat{y}_{m_i},\hat{s}_{m_i})\big)\Big)(\hat{y}_{m_i}) \\ &\le \Big(F(\Delta t_{m_i})\Phi(\,\cdot\,,\hat{s}_{m_i}+\Delta t_{m_i})\Big)(\hat{y}_{m_i})+(u_{\Delta t_{m_i}}-\Phi)(\hat{y}_{m_i},\hat{s}_{m_i}). \end{aligned}$$

Thus

$$\Phi(\hat{y}_{m_i},\hat{s}_{m_i})-(F(\Delta t_{m_i})\Phi(\,\cdot\,,\hat{s}_{m_i}+\Delta t_{m_i}))(\hat{y}_{m_i})\le 0. \tag{3.11}$$

Let m_i tend to infinity, and by Theorem 2.1, we get from (3.11) that

$$\begin{aligned} \min\Bigg\{ &-\frac{\partial\Phi}{\partial t}-\frac{\sigma^2}{2}\frac{\partial^2\Phi}{\partial x^2}-\left(r-q-\lambda k-\frac{\sigma^2}{2}\right)\frac{\partial\Phi}{\partial x}+(r+\lambda)\Phi \\ &-\lambda\int_{\mathbf{R}}\Phi(x+y,t)d\tilde{N}(y),\ \Phi-(e^x-K)^+\Bigg\}(x_0,t_0)\le 0. \end{aligned}$$

Letting ε tend to zero, we have

$$\begin{aligned} \min\Bigg\{ &-\frac{\partial\phi}{\partial t}-\frac{\sigma^2}{2}\frac{\partial^2\phi}{\partial x^2}-\left(r-q-\lambda k-\frac{\sigma^2}{2}\right)\frac{\partial\phi}{\partial x}+(r+\lambda)\phi \\ &-\lambda\int_{\mathbf{R}}\phi(x+y,t)d\tilde{N}(y),\ \phi-(e^x-K)^+\Bigg\}(x_0,t_0)\le 0. \end{aligned} \tag{3.12}$$

Recalling that $u^*(x_0,t_0)=\phi(x_0,t_0)$, we conclude from (3.12) that u^* is a subsolution of (2.4). Similarly, we can show that u_* is a supersolution of (2.4). Thus, we have proved $u_{\Delta t}(x,t)$ converges to $u(x,t)$ as $\Delta t\to 0$. Because $u(x,t)$ is continuous and monotone with respect to x and t (see [6]), and $u_{\Delta t}(x,t)$ is also a monotone function of x and t from Lemma 3.1, we know [8, Lemma 4.5] that this convergence is uniform on any semibounded domain $(-\infty,M)\times[0,T]$, which is the desired result. □

Remark 3.7. From Theorem 3.5, we can only deduce that $u^*\le u_*$ on $\mathbf{R}\times(0,T]$. However, we can make some extension of $u_{\Delta t}(x,t)$ by (2.1) for $n=-1,-2,\ldots,-N'$ to let u^* and u_* have definition in $\mathbf{R}\times[-\delta,T]$ for some $\delta=N'\Delta t>0$. It can still be shown that u^* and u_*, satisfying (3.6), are viscosity subsolution and supersolution of problem (2.4) on $\mathbf{R}\times[-\delta,T]$. Then we conclude by Theorem 3.5 that $u^*\le u_*$ on $\mathbf{R}\times(\delta,T]$, which implies $u^*=u_*$ on $\mathbf{R}\times[0,T]$.

4. Optimal exercise boundary. This section investigates properties of optimal exercise boundary (i.e., free boundary). We will show existence and convergence of optimal exercise boundary in the binomial tree method. We also achieve the terminal value $s(T)$ of the optimal exercise boundary $s(t)$ of problem (2.4).

To prove the existence of approximate optimal exercise boundary in the binomial tree method, it suffices to show the following.

LEMMA 4.1. *Let Δt be sufficiently small. For each $n<N$, there exists an integer j_n such that*

$$\begin{cases} V_j^n=\varphi_j>\dfrac{1}{\rho}\left[(1-\lambda\Delta t)\big(pV_{j+1}^{n+1}+(1-p)V_{j-1}^{n+1}\big)+\lambda\Delta t\displaystyle\sum_{l\in\mathbf{Z}}V_{j+l}^{n+1}p_l\right], & j\ge j_n, \\ V_j^n=\dfrac{1}{\rho}\left[(1-\lambda\Delta t)\big(pV_{j+1}^{n+1}+(1-p)V_{j-1}^{n+1}\big)+\lambda\Delta t\displaystyle\sum_{l\in\mathbf{Z}}V_{j+l}^{n+1}p_l\right]\ge\varphi_j, & j<j_n. \end{cases} \tag{4.1}$$

Furthermore, we have

$$j_n \le j_{n-1}.$$

Proof. We use induction to prove this lemma. Let $m_1 = \inf\{j :\ e^{j\Delta x} - K \ge 0\}$. If $j \ge m_1 + 1$, in terms of (3.2), we can get

$$\begin{aligned} I_j^N &= (1-\lambda\Delta t)\big[pV_{j+1}^N + (1-p)V_{j-1}^N\big] + \lambda\Delta t\sum_{l\in\mathbf{Z}} V_{j+l}^N p_l \\ &= (1-\lambda\Delta t)\big[p(e^{(j+1)\Delta x} - K) + (1-p)(e^{(j-1)\Delta x} - K)\big] + \lambda\Delta t\sum_{l\in\mathbf{Z}}(e^{(j+l)\Delta x} - K)^+ p_l \\ &= (1-\lambda\Delta t)e^{j\Delta x}\big[pe^{\Delta x} + (1-p)e^{-\Delta x}\big] - (1-\lambda\Delta t)K + \lambda\Delta t\sum_{l\ge m_1-j}(e^{(j+l)\Delta x} - K)p_l \\ &= e^{j\Delta x+(r-q)\Delta t} - K + \lambda\Delta t\sum_{l<m_1-j}(K - e^{(j+l)\Delta x})p_l, \qquad j \ge m_1+1. \end{aligned}$$

Then

$$\begin{aligned} I_j^N - \rho\varphi_j &= e^{j\Delta x}(e^{(r-q)\Delta t} - e^{r\Delta t}) + K(e^{r\Delta t} - 1) + \lambda\Delta t\sum_{l<m_1-j}(K - e^{(j+l)\Delta x})p_l \\ &= \left[rK - qe^{j\Delta x} + \lambda\sum_{l<m_1-j}(K - e^{(j+l)\Delta x})p_l\right]\frac{\Delta x^2}{\sigma^2} + O(\Delta x^4), \qquad j \ge m_1+1. \end{aligned}$$

Noting that

$$\begin{aligned} \sum_{l<m_1-j}(K - e^{(j+l)\Delta x})p_l &= \sum_{l<m_1-(j+1)}(K - e^{(j+l)\Delta x})p_l + (K - e^{(m_1-1)\Delta x})p_{k_1-j-1} \\ &\ge \sum_{l<m_1-(j+1)}(K - e^{(j+1+l)\Delta x})p_l \end{aligned}$$

and

$$0 \le \sum_{l<m_1-j}(K - e^{(j+l)\Delta x})p_l \le K,$$

it is easy to see that, for Δx small enough and $j \ge m_1 + 1$, $I_j^N - \rho\varphi_j$ is strictly monotonically decreasing with respect to $j\Delta x$ and when j is sufficiently large, we must have $I_j^N - \rho\varphi_j < 0$. So take

$$m_2 = \inf\left\{j \ge m_1 + 1 :\ rK - qe^{j\Delta x} + \lambda\sum_{l<m_1-j}(K - e^{(j+l)\Delta x})p_l \le 0\right\}. \tag{4.2}$$

When $j \ge m_2 + 1$, due to the strict monotonicity of $I_j^N - \rho\varphi_j$, we get $V_j^{N-1} = \varphi_j > \frac{1}{\rho}I_j^N$; when $j < m_1$, $V_j^{N-1} = \frac{1}{\rho}I_j^N \ge 0 = \varphi_j$. At the same time, we have

$$\begin{aligned} I_{m_1}^N - \rho\varphi_{m_1} &= (1-\lambda\Delta t)\big[p\varphi_{m_1+1} + (1-p)\varphi_{m_1-1}\big] + \lambda\Delta t\sum_{l\in\mathbf{Z}}\varphi_{m_1+l}p_l \\ &> \left[rK - qe^{m_1\Delta x} + \lambda\sum_{l<0}\big(K - e^{(m_1+l)\Delta x}\big)p_l\right]\frac{\Delta x^2}{\sigma^2} + O(\Delta x^4) \\ &> I_{m_1+1}^N - \rho\varphi_{m_1+1}, \end{aligned}$$

where the second inequality follows from $\varphi_{m_1-1} = 0 > e^{(m_1-1)\Delta x} - K$. Thus we know that $I_j^N - \rho\varphi_j$ is also strictly monotonically decreasing for $j \geq m_1$. So, when $m_2 = m_1 + 1$, if $\varphi_{m_1} > \frac{1}{\rho} I_{m_1}^N$ we choose $j_{N-1} = m_1 (= m_2 - 1)$; otherwise we choose $j_{N-1} = m_2$ or $m_2 + 1$. When $m_2 > m_1 + 1$, which implies $I_{m_1}^N - \rho\varphi_{m_1} \geq I_{m_1+j}^N - \rho\varphi_{m_1+j} \geq 0$ for $1 \leq j \leq m_2 - m_1 - 1$, we choose $j_{N-1} = m_2$ or $m_2 + 1$. Thus we have shown that there exists $j_{N-1} \in [m_2 - 1, m_2 + 1]$ such that (4.1) holds.

Suppose (4.1) is true when $n = m + 1$. When $j < j_{m+1}$, due to Lemma 3.1(2), we have $\frac{1}{\rho} I_j^{m+1} \geq \frac{1}{\rho} I_j^{m+2} \geq \varphi_j$ (which implies that $j_m \geq j_{m+1}$ if j_m exists); when $j \geq j_{m+1} + 1$,

$$\begin{aligned} I_j^{m+1} - \rho\varphi_j &= (1 - \lambda\Delta t)\big[pV_{j+1}^{m+1} + (1-p)V_{j-1}^{m+1}\big] + \lambda\Delta t \sum_{l\in\mathbf{Z}} V_{j+l}^{m+1} p_l - \rho\varphi_j \\ &= (1 - \lambda\Delta t)\Big[p\varphi_{j+1} + (1-p)\varphi_{j-1}\Big] + \lambda\Delta t \sum_{l\in\mathbf{Z}} V_{j+l}^{m+1} p_l - \rho\varphi_j \\ &= \left[rK - qe^{j\Delta x} + \lambda \sum_{l\in\mathbf{Z}} \big(V_{j+l}^{m+1} - (e^{(j+l)\Delta x} - K)\big) p_l\right] \frac{\Delta x^2}{\sigma^2} + O(\Delta x^4) \\ &= \left[rK - qe^{j\Delta x} + \lambda \sum_{l<j_{m+1}-j} \big(V_{j+l}^{m+1} - e^{(j+l)\Delta x} + K\big) p_l\right] \frac{\Delta x^2}{\sigma^2} + O(\Delta x^4). \end{aligned}$$

From Lemma 3.1(4), we can deduce that $I_j^{m+1} - \rho\varphi_j$ is strictly monotonically decreasing with respect to $j\Delta x$ for Δx small enough and $j \geq j_{m+1} + 1$. From Lemma 3.1(3) we know that $I_j^{m+1} - \rho\varphi_j < 0$ when j is sufficiently large. Take

$$m_3 = \inf\left\{ j \geq j_{m+1} + 1 :\ rK - qe^{j\Delta x} + \lambda \sum_{l<j_{m+1}-j} (V_{j+l}^{m+1} - e^{(j+l)\Delta x} + K) p_l \leq 0 \right\}.$$

Similar to the case of $n = N - 1$, we can show that there exists $j_m \in [m_3 - 1, m_3 + 1]$ such that (4.1) holds. Thus the proof is completed. ∎

By Lemma 4.1, we can define the approximate optimal exercise boundary.

DEFINITION 4.2. *For fixed Δt, define the approximate optimal exercise boundary $x = s_{\Delta t}(t)$ as follows: for $t \in [(n-1)t, n\Delta t]$, $1 \leq n < N$,*

$$s_{\Delta t}(t) = \frac{t - (n-1)\Delta t}{\Delta t} j_n \Delta x + \frac{n\Delta t - t}{\Delta t} j_{n-1} \Delta x.$$

By definition, $s_{\Delta t}(t)$ is monotonically decreasing.

Similar to the proof of Jiang [8, Theorem 4.1(2)] and Lamberton [9, Theorem 3.1], we have the following.

THEOREM 4.3. *If the genuine optimal exercise boundary $s(t)$ of problem (2.4) is continuous, then $s_{\Delta t}(t)$ converges uniformly to $s(t)$ as $\Delta t \to 0$.*

Now we can give the terminal value $s(T)$ of the genuine optimal exercise boundary $s(t)$.

THEOREM 4.4. *If $s(t)$ is continuous, we have*

$$s(T) = \max\{\ln K, x_0\}, \tag{4.3}$$

where x_0 is the unique solution of the equation

$$rK - qe^x + \lambda \int_{-\infty}^{\ln K - x} (K - e^{x+y}) d\tilde{N}(y) = 0. \tag{4.4}$$

Proof. From the proof of Lemma 4.1, we know that $s_{\Delta t}(T - \Delta t) = j_{N-1}\Delta x \in [(m_2 - 1)\Delta x, (m_2 + 1)\Delta x]$, where m_2 is defined by (4.2). Let $\Delta t \to 0$; then we can easily deduce (4.3) from Theorem 4.3 and (4.2) . We only have to verify that equation (4.4) has an unique solution. Let

$$f(x) = rK - qe^x + \lambda \int_{-\infty}^{\ln K - x} (K - e^{x+y}) d\tilde{N}(y);$$

then

$$f'(x) = -qe^x - \lambda \int_{-\infty}^{\ln K - x} e^{x+y} d\tilde{N}(y) < 0$$

and

$$\lim_{x \to +\infty} f(x) = -\infty, \quad \lim_{x \to -\infty} f(x) \geq rK > 0.$$

We conclude that the equation $f(x) = 0$ has a unique solution x_0. The proof is completed. □

Remark 4.5. From Theorem 4.4, we know that for American call options in jump-diffusion models, the terminal value of optimal exercise boundary is $\max\{K, e^{x_0}\}$. When there is no jump, i.e., $\lambda = 0$, the value is $\max\{K, \frac{r}{q}K\}$, which is a well-known result in diffusion models.

Remark 4.6. It's easy to see that the solution of (4.4) tends to infinity as q tends to zero. Then the optimal exercise boundary also tends to infinity. Indeed, when $q = 0$, American call options are equal to European ones. However, when dividends are considered, the valuation of American call options, just as puts, is a free-boundary problem.

Remark 4.7. In this paper American call options are considered because the approximate sequence to call options is not uniformly bounded in the L^∞-norm as opposed to puts. However, it is easy to generalize all main results to American put options. Similarly, for American puts, we can prove that the terminal value of optimal exercise boundary is $\min\{K, e^{y_0}\}$, where y_0 is the unique solution of the following equation:

$$-rK + qe^y + \lambda \int_{\ln K - y}^{+\infty} (e^{x+y} - K) d\tilde{N}(x) = 0.$$

Appendix. Proof of Theorem 3.5. For all $u \in \mathrm{USC}(\mathbf{R} \times [0, T])$ and $(x, t) \in \mathbf{R} \times [0, T)$, we define the parabolic superjet:

$$\begin{aligned} \mathcal{P}^{2,+}u(x,t) = \Big\{ & (a, p, X) \in \mathbf{R} \times \mathbf{R} \times \mathbf{R} / u(y, s) \leq u(x, t) + a(s - t) + p(y - x) \\ & + \frac{X}{2}(y - x)^2 + o(|s - t| + |y - x|^2) \quad \text{as} \quad (y, s) \to (x, t) \Big\} \end{aligned}$$

and its closure

$$\begin{aligned} \bar{\mathcal{P}}^{2,+}u(x,t) = \{ & (a, p, X) = \lim_{n \to \infty} (a_n, p_n, X_n) \text{ with } (a_n, p_n, X_n) \in \mathcal{P}^{2,+}u(x_n, t_n) \\ & \text{and } \lim_{n \to \infty} (x_n, t_n, u(x_n, t_n)) = (x, t, u(x, t)) \}. \end{aligned}$$

Similarly, if $v \in \mathrm{LSC}(\mathbf{R} \times [0,T])$, we consider the parabolic subject $\mathcal{P}^{2,-}v(x,t) = -\mathcal{P}^{2,-}(-v)(x,t)$ and its closure $\overline{\mathcal{P}}^{2,-}v(x,t) = -\overline{\mathcal{P}}^{2,-}(-v)(x,t)$.

Repeating the argument of Alvarez and Tourin, we can obtain another formulation for viscosity solution, as follows.

LEMMA A.1. *Let $u \in \mathrm{USC}(\mathbf{R} \times [0,T])(\mathrm{LSC}(\mathbf{R} \times [0,T])$ and $|u(x,t)| \le e^x$; then u is a viscosity subsolution (supersolution) of* (2.4) *if*

$$\begin{cases} \min\left\{-a - \dfrac{\sigma^2}{2}X - \left(r - q - \lambda k - \dfrac{\sigma^2}{2}\right)p + (r+\lambda)u - \lambda\displaystyle\int_{\mathbf{R}} u(x+y,t)d\tilde{N}(y),\right. \\ \qquad \left. u - (e^x - K)^+\right\} \le 0\ (\ge 0), \quad (x,t) \in \mathbf{R} \times (0,T), \\ \qquad\qquad (a,p,X) \in \ \mathcal{P}^{2,+}u(x,t)(\mathcal{P}^{2,-}u(x,t)), \\ u(x,T) \le\ (\ge)\ (e^x - K)^+, \qquad x \in \mathbf{R}. \end{cases}$$

Using the transformation $v(x,t) = e^x u(x,t)$, (2.4) is transformed into the following equation:

$$\text{(A.1)} \quad \begin{cases} \min\left\{ -\dfrac{\partial u}{\partial t} - \dfrac{\sigma^2}{2}\dfrac{\partial^2 u}{\partial x^2} - \left(r - q - \lambda k + \dfrac{\sigma^2}{2}\right)\dfrac{\partial v}{\partial x} + (q + \lambda + \lambda k)u \right. \\ \qquad \left. - \lambda\displaystyle\int_{\mathbf{R}} u(x+y,t)e^y d\tilde{N}(y), u - g(x)\right\} = 0, \qquad (x,t) \in \mathbf{R} \times [0,T), \\ u(x,T) = g(x), \qquad x \in \mathbf{R}, \end{cases}$$

where $g(x) = (1 - e^{-x}K)^+$.

So we only have to prove comparison for (A.1) with bounded solutions.

Let u and v be subsolution and supersolution of (A.1), respectively, and $|u|, |v| \le 1$. We argue by contradiction and thus assume that

$$\text{(A.2)} \qquad (z,s) \in \mathbf{R} \times (0,T) \quad \text{and} \quad u(z,s) - v(z,s) = \delta > 0.$$

For all $\varepsilon, \alpha, \beta > 0$, we define a function $\Phi(x,y,t)$ in $\mathbf{R} \times \mathbf{R} \times (0,T]$:

$$\Phi(x,y,t) = u(x,t) - v(y,t) - \frac{\varepsilon}{2}(x^2 + y^2) - \frac{\alpha}{2}|x-y|^2 - \frac{\beta}{t}.$$

Since $|u|, |v| \le 1$ and $\lim_{t \to 0} \Phi(x,y,t) = -\infty$, Φ admits a maximum at $(\hat{x}, \hat{y}, \hat{t}) \in \mathbf{R} \times \mathbf{R} \times (0,T]$ (we omit the dependence on ε and α to alleviate notations) and $\hat{x}^2 + \hat{y}^2 < +\infty$ (otherwise $\Phi(\hat{x}, \hat{y}, \hat{t}) \to -\infty$). By (A.2),

$$\text{(A.3)} \qquad \Phi(\hat{x}, \hat{y}, \hat{t}) \ge \delta - \varepsilon z^2 - \frac{\beta}{s} \ge \frac{\delta}{2} > 0 \quad \text{when } \varepsilon \text{ and } \beta \text{ is small enough.}$$

It is easy [5, Lemma 3.1] to check

$$\text{(A.4)} \qquad \lim_{\alpha \to \infty} \alpha|\hat{x} - \hat{y}|^2 = 0.$$

If $\hat{t} = T$, we have

$$0 < \frac{\delta}{2} \le \Phi(\hat{x}, \hat{y}, T) \le g(\hat{x}) - g(\hat{y}) - \frac{\alpha}{2}|\hat{x} - \hat{y}|^2;$$

however, the right-hand side above tends to zero as $\alpha \to \infty$ by (A.4). So $\hat{t} < T$ if α is large.

Applying Theorem 8.3 of [5] to the function $\Phi(x,y,t)$ at $(\hat{x},\hat{y},\hat{t}) \in \mathbf{R}\times\mathbf{R}\times(0,T)$, we can find $a,b,X,Y \in \mathbf{R}$ such that

$$(a, \alpha(\hat{x}-\hat{y})+\varepsilon\hat{x}, X) \in \bar{\mathcal{P}}^{2,+}u(\hat{x},\hat{t}),\ (b, \alpha(\hat{x}-\hat{y})-\varepsilon\hat{y}, Y) \in \bar{\mathcal{P}}^{2,-}v(\hat{y},\hat{t})$$

and

$$a-b=-\frac{\beta}{\hat{t}^2},\quad X-Y \le 2\varepsilon+\frac{2\varepsilon^2}{\alpha}. \tag{A.5}$$

The fact that u and v are, respectively, subsolution and supersolution of (A.1) and Lemma A.1 yields

$$\begin{aligned}\min\Big\{ &-a-\frac{\sigma^2}{2}X-\left(r-q-\lambda k+\frac{\sigma^2}{2}\right)[\alpha(\hat{x}-\hat{y})+\varepsilon\hat{x}]+(q+\lambda+\lambda k)u(\hat{x},\hat{t})\\ &-\lambda\int_{\mathbf{R}} u(\hat{x}+y,\hat{t})e^y d\tilde{N}(y),\ u(\hat{x},\hat{t})-g(\hat{x})\Big\}\le 0\end{aligned}$$

and

$$\begin{aligned}\min\Big\{ &-b-\frac{\sigma^2}{2}Y-\left(r-q-\lambda k+\frac{\sigma^2}{2}\right)[\alpha(\hat{x}-\hat{y})-\varepsilon\hat{y}]+(q+\lambda+\lambda k)v(\hat{y},\hat{t})\\ &-\lambda\int_{\mathbf{R}} v(\hat{y}+y,\hat{t})e^y d\tilde{N}(y),\ v(\hat{y},\hat{t})-g(\hat{y})\Big\}\ge 0.\end{aligned}$$

Subtracting these two inequalities and remarking that $\min(c,d)-\min(e,f)\le 0$ implies either $c-e\le 0$ or $d-f\le 0$; we divide our consideration into two cases:

(1) The first case is

$$\begin{aligned}&(q+\lambda+\lambda k)[u(\hat{x},\hat{t})-v(\hat{y},\hat{t})]-\lambda\int_{\mathbf{R}}[u(\hat{x}+y,\hat{t})-v(\hat{y}+y,\hat{t})]e^y d\tilde{N}(y)\\ &\le -\frac{\beta}{\hat{t}^2}+\frac{\sigma^2}{2}(X-Y)+(r-q-\lambda k+\tfrac{\sigma^2}{2})\varepsilon(\hat{x}+\hat{y}).\end{aligned} \tag{A.6}$$

Because

$$\Phi(\hat{x},\hat{y},\hat{t}) \ge \Phi(\hat{x}+y,\hat{y}+y,\hat{t}),$$

we have

$$u(\hat{x}+y,\hat{t})-v(\hat{y}+y,\hat{t}) \le u(\hat{x},\hat{t})-v(\hat{y},\hat{t})+\varepsilon[y(\hat{x}+\hat{y})+y^2]. \tag{A.7}$$

We conclude from (A.5), (A.6), and (A.7) that

$$\begin{aligned}&\left(q+\lambda+\lambda k-\lambda\int_{\mathbf{R}} e^y d\tilde{N}(y)\right)[u(\hat{x},\hat{t})-v(\hat{y},\hat{t})]\\ &\le -\frac{\beta}{\hat{t}^2}+\frac{\sigma^2}{2}(2\varepsilon+\frac{2\varepsilon^2}{\alpha})+\varepsilon(r-q-\lambda k+\frac{\sigma^2}{2})(\hat{x}+\hat{y})+\varepsilon\lambda\int_{\mathbf{R}} e^y[y(\hat{x}+\hat{y})+y^2]d\tilde{N}(y).\end{aligned} \tag{A.8}$$

We recall that

$$\begin{aligned}&\int_{\mathbf{R}} e^y d\tilde{N}(y)=\int_{-1}^{+\infty}(1+y)dN(y)=1+k,\\ &\int_{\mathbf{R}} e^y y^2 d\tilde{N}(y)=\int_{-1}^{+\infty}(1+y)\ln^2(1+y)dN(y)\le\int_{-1}^{+\infty} y^2 dN(y)<+\infty,\\ &\hat{x}^2+\hat{y}^2<+\infty.\end{aligned} \tag{A.9}$$

At the same time, by (A.3),

$$u(\hat{x},\hat{t}) - v(\hat{y},\hat{t}) \geq \Phi(\hat{x},\hat{y},\hat{t}) \geq \frac{\delta}{2} > 0 \quad \text{if } \varepsilon \text{ is small enough.} \tag{A.10}$$

So by choosing α sufficiently large and sending $\varepsilon \to 0, \beta \to 0$ in (A.8), we deduce from (A.8), (A.9), and (A.10) that

$$0 < \frac{q\delta}{2} \leq 0,$$

which is a contradiction.

(2) The second case occurs if

$$u(\hat{x},\hat{t}) - v(\hat{y},\hat{t}) \leq g(\hat{x}) - g(\hat{y}). \tag{A.11}$$

In view of (A.4), (A.10), and continuity of $g(x)$, we can easily deduce a contradiction from (A.11) when $\alpha \to \infty$. □

REFERENCES

[1] O. Alvarez and A. Tourin, *Viscosity solution of nonlinear integro-differential equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 293–317.

[2] K. I. Amin, *Jump diffusion option valuation in discrete time*, J. Finance, 48 (1993), pp. 1833–1863.

[3] G. Barles, Ch. Daher, and M. Romano, *Convergence of numerical schemes for parabolic equations arising in finance theory*, Math. Models Methods Appl. Sci., 5 (1995), pp. 125–143.

[4] J. Cox, S. Ross, and M. Rubinstein, *Option pricing: A simplified approach*, J. Finan. Econom., 7 (1979), pp. 229–264.

[5] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[6] H. Pham, *Optimal stopping, free boundary and American option in a jump-diffusion model*, Appl. Math. Optim., 35 (1997), pp. 145–164.

[7] J. M. Harrison and D. M. Kreps, *Martingales and arbitrage in multiperiods securities markets*, J. Econom. Theory, 20 (1979), pp. 381–408.

[8] L. Jiang and M. Dai, *Convergence of binomial tree method for American options*, in Proceeding of the Conference on Partial Differential Equations and their Applications, Singapore, 1999, pp. 106–118.

[9] D. Lamberton, *Convergence of the critical price in the approximation of American options*, Math. Finance, 3 (1993), pp. 179–190.

[10] C. Xu, X. Qian, and L. Jiang, *Numerical analysis on binomial tree methods for a jump-diffusion model*, J. Comput. Math., 156 (2003), pp. 23–45.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1914–1931

A FAMILY OF RECTANGULAR MIXED ELEMENTS WITH A CONTINUOUS FLUX FOR SECOND ORDER ELLIPTIC PROBLEMS*

TODD ARBOGAST† AND MARY F. WHEELER‡

Abstract. We present a family of mixed finite element spaces for second order elliptic equations in two and three space dimensions. Our spaces approximate the vector flux by a continuous function. Our spaces generalize certain spaces used for approximation of Stokes problems. The finite element method incorporates projections of the Dirichlet data and certain low order terms. The method is locally conservative on the average. Suboptimal convergence is proven and demonstrated numerically. The key result is to construct a flux π-projection operator that is bounded in the Sobolev space H^1, preserves a projection of the divergence, and approximates optimally. Moreover, the corresponding Raviart–Thomas flux preserving π-projection operator is an L^2-projection when restricted to this family of spaces.

Key words. mixed finite element method, continuous flux, elliptic equation, error estimates

AMS subject classifications. 65N15, 65N30, 35J20

DOI. 10.1137/S0036142903435247

1. Introduction.

Mixed finite element methods have been used effectively to solve many problems, including second order elliptic problems [10, 15, 31, 34]. Both the scalar variable and its vector flux are approximated directly. While it is necessary to approximate the flux in $H(\mathrm{div})$, the space of L^2 vectors whose divergence is also in L^2, it is not necessary that the flux be fully continuous. In the usual mixed spaces (see, e.g., [12, 13, 14, 17, 29, 31]), only the normal component of the approximate flux is continuous. The tangential components are discontinuous across element boundaries.

In some applications, it is desirable that the flux be continuous. The applications we have in mind come from simulating fluid flow in a porous medium [7, 30, 32]. Miscible displacement in a petroleum reservoir or groundwater transport problems require the solution to a system of equations in model form:

$$ap + \nabla \cdot \mathbf{u} = b, \quad \Omega, \tag{1.1}$$

$$d\mathbf{u} = -\nabla p + \mathbf{c}, \quad \Omega, \tag{1.2}$$

$$\phi \frac{\partial c}{\partial t} - \nabla \cdot \big(D(\mathbf{u})\nabla c - \mathbf{u}c\big) = \hat{c}f^+ - cf^-, \quad \Omega,\ t > 0, \tag{1.3}$$

where p is the pressure, $\mathbf{u}$ is the Darcy velocity (i.e., the flux), c is the concentration of a dissolved chemical that is transported by the flow, and a, b, $\mathbf{c}$, d, ϕ, D, and $\hat{c}$ are

*Received by the editors September 19, 2003; accepted for publication (in revised form) May 7, 2004; published electronically January 20, 2005. This work was supported in part by the U.S. Department of Energy and the State of Texas Governor's Energy Office.

http://www.siam.org/journals/sinum/42-5/43524.html

† Institute for Computational Engineering and Sciences, University of Texas, 1 University Station C0200, Austin, TX 78712, and Mathematics Department, University of Texas, 1 University Station C1200, Austin, TX 78712–0257 (arbogast@ices.utexas.edu). This author was supported in part by U.S. National Science Foundation grant DMS-0074310.

‡Institute for Computational Engineering and Sciences, University of Texas, 1 University Station C0200, Austin, TX 78712, and Department of Aerospace Engineering & Engineering Mechanics, University of Texas, 1 University Station C0600, Austin, TX 78712, and Department of Petroleum & Geosystems Engineering, University of Texas, 1 University Station C0300, Austin, TX 78712 (mfw@ices.utexas.edu).

various parameters. To these equations, we must add boundary and initial conditions. For the subsystem (1.1)–(1.2), let $\partial\Omega$ be decomposed into Γ_N and Γ_D, and set

$$\mathbf{u}\cdot\nu = g_N, \quad \Gamma_N, \tag{1.4}$$

$$p = p_D, \quad \Gamma_D, \tag{1.5}$$

where ν is the outer unit normal vector, g_N is the given boundary flux, and p_D is the given boundary pressure.

Since $D(\mathbf{u}) \approx 0$, equation (1.3) is nearly hyperbolic. Characteristic methods have been successful in treating this equation (see, e.g., [4, 20, 22, 23]); however, they rely heavily on the velocity $\mathbf{u}$. To obtain good characteristic trace-backs, $\mathbf{u}_h \approx \mathbf{u}$ should satisfy

(i) $\nabla\cdot\mathbf{u}_h = \mathcal{P}_W f$, where $\mathcal{P}_W$ is an appropriate projection;

(ii) $\mathbf{u}_h$ is continuous.

Property (i) gives a proper divergence to the flow so that mass is conserved, while (ii) is required for consistency in tracing regions through the flow field.

A second application involves the coupling of Stokes flow with Darcy flow in a region with open channel flow adjacent to a porous medium [8]. Usual approaches require that the Stokes equations be approximated by a continuous velocity, since it must remain in H^1. However, this is not properly matched on the open/porous interface to a discontinuous Darcy velocity. A continuous Darcy velocity would therefore be desirable. These spaces have been exploited in [1, 2, 3].

Current mixed methods achieve (i) at the expense of (ii). Our goal is to relax (i), so that it holds only "on the average", but maintain (ii). Brezzi, Fortin, and Marini [16] presented a stabilization technique that allows the use of continuous finite element spaces. Their technique involves a modification of the usual mixed equations. Herein, we provide a family of mixed methods that is stable for the original set of mixed equations. These methods are defined on rectangular grids, and they generalize the Stokes elements of Fortin [24] (cf. Bernardi and Raugel [9]). We present the full development for two dimensions, and discuss the three dimensional case in the last section.

2. Some general notation. Throughout the paper, for domain ω, we denote by $L^p(\omega)$ the usual Lebesgue space of index p, $1 \le p \le \infty$, and by $W^{k,p}(\omega)$ the usual Sobolev space of k weak derivatives in $L^p(\omega)$. We denote by $(\cdot,\cdot)_\omega$ the $L^2(\omega)$ inner product (i.e., Lebesgue integration over ω). Moreover, $\|\cdot\|_{k,\omega}$ is the norm of $H^k(\omega) \equiv W^{k,2}(\omega)$, and $|\cdot|_{r,\omega}$ denotes the $H^r(\omega)$ seminorm. In the notation we may suppress ω when it is Ω. On a domain boundary e, we use the notation $\langle\cdot,\cdot\rangle_e$ for the $L^2(e)$ inner product. For d-dimensional set S, $|S|$ is its d-dimensional Lebesgue measure.

Let $P_k(\omega)$ denote the space of polynomials of degree at most k over the set ω. Moreover, in $\mathbb{R}^2$, let $Q_{i,j}(\omega)$ be the set of polynomials of degree at most i in x and j in y over ω, and similarly define $Q_{i,j,k}(\omega)$ in $\mathbb{R}^3$. We will make use of scaling arguments, so let us define $\hat{R} = [-1,1]^2$ as our reference rectangle. Moreover, let $\hat{\lambda}_j$ denote the Legendre polynomial of degree j on $[-1,1]$. Recall that they are $L^2([-1,1])$-orthogonal polynomials, and that by convention, they are normalized so that $\hat{\lambda}_j(1) = 1$; then also $\hat{\lambda}_j(-1) = (-1)^j$.

Finally, if X is a closed subspace of $L^2(\omega)$, we denote by $\mathcal{P}_X : L^2(\omega) \to X$ the $L^2(\omega)$-projection operator onto X defined for $\psi \in L^2(\omega)$ as the unique $\mathcal{P}_X\psi \in X$ such that

$$(\psi - \mathcal{P}_X\psi, \varphi)_\omega = 0 \quad \forall\varphi \in X.$$

3. An illustrative example. To illustrate our finite element spaces, consider (1.1)–(1.2) in mixed form for $\Omega \subset \mathbb{R}^2$ a bounded domain, with the natural boundary conditions $\Gamma_N = \partial\Omega$ and $g_N = 0$, and a and $\mathbf{c}$ set to 0:

$$(\nabla \cdot \mathbf{u}, w) = (b, w) \qquad \forall w \in L^2(\Omega)/\mathbb{R}, \tag{3.1}$$

$$(d\mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) = 0 \quad \forall \mathbf{v} \in H_0(\mathrm{div}; \Omega), \tag{3.2}$$

where $H_0(\mathrm{div}; \Omega)$ is the subset of $H(\mathrm{div}; \Omega)$ with vanishing normal trace on $\partial\Omega$. Let $\mathbf{V}_h \times W_h$ denote some mixed finite element space, and solve (3.1)–(3.2) for $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times W_h$ with the restrictions that $w \in W_h$ and $\mathbf{v} \in \mathbf{V}_h$.

If $\mathbf{V}_h \times W_h$ is the lowest order Raviart–Thomas space [29, 31, 34], the solution $\mathbf{u}_h$ is discontinuous. On a rectangular element R, the x-coordinate of $\mathbf{u}_h$, $u_{h,1}$, is in

$$V_{h,1}^{RT_0}(R) = Q_{1,0}(R).$$

This space has one degree of freedom for each edge normal to the x-direction. If we add four degrees of freedom by defining

$$V_{h,1}(R) = Q_{1,2}(R),$$

the extra corner degrees of freedom allow us to enforce continuity, while the two edge degrees of freedom allow us to maintain the *average* flux across each normal edge, i.e., the proper average divergence of the flow. A similar modification in y, $V_{h,2}(R) = Q_{2,1}(R)$, gives a new element with the required properties on all edges of R. This is a Stokes element due to Fortin [24].

4. The spaces in two dimensions on rectangles. Assume $\Omega \subset \mathbb{R}^2$. Let $\mathcal{T}_h$ denote a quasi-regular, conforming finite element partition of Ω into rectangles of diameter bounded by h. By quasi-regular, we mean that the aspect ratio of the rectangles is bounded by a fixed constant. We now define our family of mixed spaces; it contains the element of the previous section as its lowest order member.

DEFINITION 4.1. *For each integer $k \geq 1$, the linear space $\mathbf{V}_h^k \times W_h^{k-1}$ is defined so that, for any rectangle $R \in \mathcal{T}_h$,*

$$\begin{aligned} \mathbf{V}_h^k(R) &= Q_{k,k+1}(R) \times Q_{k+1,k}(R), \\ W_h^{k-1}(R) &= Q_{k-1,k-1}(R), \end{aligned}$$

and

$$\begin{aligned} \mathbf{V}_h^k &= \{\mathbf{v} \in (C^0(\Omega))^2 : \mathbf{v}|_R \in \mathbf{V}_h^k(R) \quad \forall R \in \mathcal{T}_h\}, \\ W_h^{k-1} &= \{w \in L^2(\Omega) : w|_R \in W_h^{k-1}(R) \quad \forall R \in \mathcal{T}_h\}. \end{aligned}$$

A local basis for $W_h^{k-1}(R)$ is trivial to construct; moreover, since W_h^{k-1} is discontinuous across element boundaries, a global basis can be constructed immediately.

A local nodal basis can be defined for $\mathbf{V}_h^k(R)$ by the degrees of freedom given in the next lemma. Moreover, because these degrees of freedom uniquely determine the function on each edge, they can be pieced together across edges and vertices to form a global basis for the continuous function space $\mathbf{V}_h^k$.

LEMMA 4.2. *For R a rectangle, $\mathbf{u}_h \in \mathbf{V}_h^k(R)$ is uniquely defined by the following degrees of freedom:*

(i) *for each corner point $P \in \partial R$ and Cartesian direction $j = 1, 2$,*

$$DOF^{(\mathrm{i})}_{P,j}(\mathbf{u}_h) = u_{h,j}(P) = \mathbf{u}_h(P) \cdot \mathbf{e}_j;$$

(ii) *on each edge $e \subset \partial R$,*

$$DOF^{(\mathrm{ii})}_{e,\lambda}(\mathbf{u}_h) = \langle \mathbf{u}_h \cdot \tau_e, \lambda \rangle_e \quad \forall \lambda \in P_{k-2}(e),$$

where τ_e is a unit tangential direction;

(iii) *on each edge $e \subset \partial R$,*

$$DOF^{(\mathrm{iii})}_{e,\lambda}(\mathbf{u}_h) = \langle \mathbf{u}_h \cdot \nu_e, \lambda \rangle_e \quad \forall \lambda \in P_{k-1}(e),$$

where ν_e is the outer unit normal direction;

(iv) *over R,*

$$DOF^{(\mathrm{iv})}_{\mathbf{v}}(\mathbf{u}_h) = (\mathbf{u}_h, \mathbf{v})_R \quad \forall \mathbf{v} \in Q_{k-2,k-1}(R) \times Q_{k-1,k-2}(R).$$

Moreover, on each edge $e \subset \partial R$, $\mathbf{u}_h|_e$ is uniquely defined by the degrees of freedom (i)–(iii) *restricted to e.*

Proof. We restrict our analysis to the case where $R = \hat{R} = [-1, 1]^2$; an affine map can be used to show the result for a general rectangle R. As usual, since our function spaces are finite-dimensional vector spaces, the degrees of freedom uniquely determine the function if and only if both the dimension of the function space and the number of independent degrees of freedom agree; whenever the degrees of freedom vanish, the function also vanishes.

We begin by considering $\mathbf{u}_h$ on an edge e. Now $\mathbf{u}_h \cdot \tau_e \in P_k(e)$ and $\mathbf{u}_h \cdot \nu_e \in P_{k+1}(e)$, so the total dimension of this space is $2k + 3$. The number of degrees of freedom that act on e are 4 for (i), $k - 1$ for (ii), k for (iii), and 0 for (iv), leading to the same number $2k + 3$ degrees of freedom on e. Suppose that the degrees of freedom (i)–(iii) restricted to e vanish. We conclude from (i) that

$$\mathbf{u}_h \cdot \tau_e(\xi) = (1 - \xi^2)\, q(\xi),$$

where $q \in P_{k-2}$ (and $\mathbf{u}_h \cdot \tau_e \equiv 0$ if $k = 1$). Then (ii) implies that $q = 0$, and we conclude that $\mathbf{u}_h \cdot \tau_e \equiv 0$ on e. Similarly, we conclude from (i) and (iii) that $\mathbf{u}_h \cdot \nu_e \equiv 0$ on e. We have thereby demonstrated the last statement of the lemma.

Note that the total number of degrees of freedom is 8 for (i), $4(k - 1)$ for (ii), $4k$ for (iii), and $2(k - 1)k$ for (iv), so that the total is $2k^2 + 6k + 4$. This is the same as

$$\dim \mathbf{V}^k_h(R) = 2(k + 1)(k + 2) = 2k^2 + 6k + 4.$$

So suppose that all degrees of freedom of $\mathbf{u}_h$ vanish. We have already shown that then $\mathbf{u}_h|_{\partial R}$ vanishes, so

$$\mathbf{u}_h(x, y) = (1 - x^2)(1 - y^2)\, \mathbf{v}(x, y)$$

for some $\mathbf{v}(x, y) \in Q_{k-2,k-1}(R) \times Q_{k-1,k-2}(R)$. By degree of freedom (iv), we conclude that $\mathbf{v} \equiv 0$, and so also $\mathbf{u}_h \equiv 0$, completing the proof. ∎

For completeness and future reference, we show how to construct an explicit nodal basis on the reference rectangle $\hat{R} = [-1, 1]^2$. A nodal basis has the property that each member has one degree of freedom evaluate to 1 and the rest to 0. Recall that $\hat{\lambda}_j$

is the Legendre polynomial of degree j on $\hat{R}$, that $\langle \hat{\lambda}_i, \hat{\lambda}_j \rangle_{[-1,1]} = 0$ if $i \neq j$, and that $\hat{\lambda}_j(1) = 1$ and $\hat{\lambda}_j(-1) = (-1)^j$. For degrees of freedom (ii)–(iv), we need to select a basis for the test spaces. Since $\{\hat{\lambda}_j\}_{j \leq k}$ forms a basis for $P_k(-1,1)$, we can restrict the polynomials in the degrees of freedom (ii) and (iii) to Legendre polynomials, and to tensor products of such in (iv).

For degree of freedom (i), we define for corner point $P = (-1,-1)$ and direction 1 the basis function

$$\hat{\mathbf{v}}^{(\mathrm{i})}_{P,1} = -\frac{1}{4}(\hat{\lambda}_k(\hat{x}) - \hat{\lambda}_{k-1}(\hat{x}))(\hat{\lambda}_{k+1}(\hat{y}) - \hat{\lambda}_k(\hat{y}))\mathbf{e}_1.$$

This function has the property that it is $\mathbf{e}_1$ at P and vanishes at the three other corner points, and the degrees of freedom (ii)–(iv) vanish. We similarly define a basis function for the other three corner points, and for direction 2. For degree of freedom (ii), with edge $e = (-1,1) \times \{-1\}$ and $\lambda = \hat{\lambda}_j(\hat{x})$, $0 \leq j \leq k-2$, we define

$$\hat{\mathbf{v}}^{(\mathrm{ii})}_{e,j} = \alpha^{(\mathrm{ii})}_{e,j}(\hat{\lambda}_j(\hat{x}) - \hat{\lambda}_{k_j}(\hat{x}))(\hat{\lambda}_{k+1}(\hat{y}) - \hat{\lambda}_k(\hat{y}))\mathbf{e}_1,$$

where k_j is either $k-1$ or k so that k_j and j have the same even/odd parity (i.e., so that $\hat{\lambda}_{k_j}(-1) = \hat{\lambda}_j(-1)$), and $\alpha^{(\mathrm{ii})}_{e,j}$ is chosen to fix the normalization

$$\mathrm{DOF}^{(\mathrm{ii})}_{e,\lambda_j}(\hat{\mathbf{v}}^{(\mathrm{ii})}_{e,j}) = \langle \hat{\mathbf{v}}^{(\mathrm{ii})}_{e,j} \cdot \tau_e, \hat{\lambda}_j \rangle_e = 1.$$

The other basis functions of type (ii) are defined similarly. For degree of freedom (iii), with edge $e = \{-1\} \times (-1,1)$ and $\lambda = \hat{\lambda}_j(\hat{y})$, $0 \leq j \leq k-1$, we define

$$\hat{\mathbf{v}}^{(\mathrm{iii})}_{e,j} = \alpha^{(\mathrm{iii})}_{e,j}(\hat{\lambda}_k(\hat{x}) - \hat{\lambda}_{k-1}(\hat{x}))(\hat{\lambda}_j(\hat{y}) - \hat{\lambda}_{k_j}(\hat{y}))\mathbf{e}_1,$$

where k_j is either k or $k+1$ so that k_j and j have the same parity and $\alpha^{(\mathrm{iii})}_{e,j}$ is chosen to make

$$\mathrm{DOF}^{(\mathrm{iii})}_{e,\lambda_j}(\hat{\mathbf{v}}^{(\mathrm{iii})}_{e,j}) = \langle \hat{\mathbf{v}}^{(\mathrm{iii})}_{e,j} \cdot \nu_e, \hat{\lambda}_j \rangle_e = 1.$$

The other basis functions of type (iii) are defined similarly. Finally, for degree of freedom (iv), with $\mathbf{v} = \hat{\lambda}_j(\hat{x})\hat{\lambda}_\ell(\hat{y})\mathbf{e}_1$, $0 \leq j \leq k-2$, $0 \leq \ell \leq k-1$, we define

$$\hat{\mathbf{v}}^{(\mathrm{iv})}_{j,\ell} = \alpha^{(\mathrm{iv})}_{j,\ell}(\hat{\lambda}_j(\hat{x}) - \hat{\lambda}_{k_j}(\hat{x}))(\hat{\lambda}_\ell(\hat{y}) - \hat{\lambda}_{k_\ell}(\hat{y}))\mathbf{e}_1,$$

where k_j is either $k-1$ or k so that $\hat{\lambda}_{k_j}(-1) = \hat{\lambda}_j(-1)$, and k_ℓ is either k or $k+1$ so that $\hat{\lambda}_{k_\ell}(-1) = \hat{\lambda}_\ell(-1)$, and $\alpha^{(\mathrm{iv})}_{j,\ell}$ is defined so that

$$\mathrm{DOF}^{(\mathrm{iv})}_{\hat{\lambda}_j\hat{\lambda}_\ell}(\hat{\mathbf{v}}^{(\mathrm{iv})}_{j,\ell}) = (\hat{\mathbf{v}}^{(\mathrm{ii})}_{j,\ell}, \hat{\lambda}_j\hat{\lambda}_\ell)_{\hat{R}} = 1.$$

The other type (iv) nodal basis functions are defined similarly.

The nodal basis on $\mathcal{T}_h$ is constructed from these local basis functions via local affine mappings. On $R \in \mathcal{T}_h$, we would map R to $\hat{R}$ by $x \mapsto \hat{x} \equiv (x - x_0)/h_1$ and $y \mapsto \hat{y} \equiv (y - y_0)/h_2$, where R has side lengths h_1 and h_2 and lower left corner (x_0, y_0). The important part of the construction is that the reference basis functions $\hat{\mathbf{v}}^{(\mathrm{i})}, \ldots, \hat{\mathbf{v}}^{(\mathrm{iv})}$ are independent of h.

We close this section with a remark about the finite element basis. We chose degrees of freedom and a local basis that are useful for the numerical analysis that

follows. This is primarily due to degree of freedom (iii), which says that moments of the normal flux are controlled. However, an equivalent set of degrees of freedom would be to replace (ii) and (iii) by evaluation at an appropriate number of points along the boundary. This would be a better basis for implementation, since it is simpler to construct.

5. A π operator. For simplicity, let $\mathcal{P}_W^{k-1} : L^2(\Omega) \to W_h^{k-1}$ denote the $L^2(\Omega)$-projection $\mathcal{P}_{W_h^{k-1}}$. As is usual for mixed spaces, we will define a π operator [15, 21, 31] for our spaces. Our π operator should map $(H^1(\Omega))^2$ onto $\mathbf{V}_h^k$. When $\mathbf{u} \in (H^1(\Omega))^2$, the normal or tangential trace of $\mathbf{u}$ on an edge e is in $H^{1/2}(e)$, and so degrees of freedom (ii) and (iii), and also (iv), are defined. However, degree of freedom (i) causes some problems.

We can resolve the difficulty by using the Clément [19] (or the Scott–Zhang [33]) interpolant $\mathcal{I}^k : H^1(\Omega) \to Q_h^k$, where Q_h^k is the space of continuous functions with $Q_h^k|_R = Q_{k,k}(R)$ for any $R \in \mathcal{T}_h$. For completeness, we define $\mathcal{I}^k$ as in [19]. The interpolant of $\psi \in L^2(\Omega)$ is defined at the nodal points of Q_h^k by setting the value to a local L^2-projection of ψ. Let N denote the set of nodal points of Q_h^k; these are, for example, the union over $R \in \mathcal{T}_h$ of the $(k+1)^2$ grid points of a $k \times k$ uniform grid over R. For each $P \in N$, let

$$\Delta_P = \bigcup_{R \in \mathcal{T}_h \text{ with } P \in R} R$$

be the union of the rectangles containing P. Then define $\psi_P \in Q_{k,k}(\Delta_P)$ by

$$(\psi - \psi_P, \varphi)_{\Delta_P} = 0 \quad \forall \varphi \in Q_{k,k}(\Delta_P),$$

and set $\mathcal{I}^k\psi(P) = \psi_P(P)$. This uniquely defines $\mathcal{I}^k\psi \in Q_h^k$; moreover, we have the estimate

$$\|\psi - \mathcal{I}^k\psi\|_j \le C\|\psi\|_r h^{r-j}, \quad j \le r \le k+1,\ j = 0, 1. \tag{5.1}$$

DEFINITION 5.1. *For $\mathbf{u} \in (H^1(\Omega))^2$, let $\pi^k\mathbf{u} \in \mathbf{V}_h^k$ be defined as the interpolant of the degrees of freedom from Lemma 4.2, modified by the Clément operator $\mathcal{I}^k$. That is, for each $R \in \mathcal{T}_h$, we require the following:*

(1) *For each corner point $P \in \partial R$ and direction $j = 1, 2$,*

$$\pi^k\mathbf{u}(P) \cdot \mathbf{e}_j = \mathcal{I}^k\mathbf{u}(P) \cdot \mathbf{e}_j.$$

(2) *On each edge $e \subset \partial R$,*

$$\langle \pi^k\mathbf{u} \cdot \tau_e, \lambda \rangle_e = \langle \mathbf{u} \cdot \tau_e, \lambda \rangle_e \quad \forall \lambda \in P_{k-2}(e).$$

(3) *On each edge $e \subset \partial R$,*

$$\langle \pi^k\mathbf{u} \cdot \nu_e, \lambda \rangle_e = \langle \mathbf{u} \cdot \nu_e, \lambda \rangle_e \quad \forall \lambda \in P_{k-1}(e).$$

(4) *Over R,*

$$(\pi^k\mathbf{u}, \mathbf{v})_R = (\mathbf{u}, \mathbf{v})_R \quad \forall \mathbf{v} \in Q_{k-2,k-1}(R) \times Q_{k-1,k-2}(R).$$

The operator π^k is linear, and it is well defined by Lemma 4.2. Before deriving properties of this operator, we consider its explicit construction using the global nodal basis given in section 4. Let us denote this basis as $\{\mathbf{v}_{j_1}^{(\mathrm{i})}, \mathbf{v}_{j_2}^{(\mathrm{ii})}, \mathbf{v}_{j_3}^{(\mathrm{iii})}, \mathbf{v}_{j_4}^{(\mathrm{iv})}\}_{j_1,j_2,j_3,j_4}$, where the superscript designates the degree of freedom type, and for consistency of notation below, the index ranges do not overlap. Basis function $\mathbf{v}_{j_2}^{(\mathrm{ii})}$, of degree of freedom type (ii), is defined with respect to some grid edge e_{j_2} and Legendre polynomial λ^{j_2} on e_{j_2}, and similarly for $\mathbf{v}_{j_3}^{(\mathrm{iii})}$. For $\mathbf{v}_{j_4}^{(\mathrm{iv})}$, it is defined with respect to rectangle R_{j_4} and a tensor product of Legendre polynomials $\mathbf{v}^{j_4}$.

We claim that we can represent, for $\mathbf{u} \in (H^1(\Omega))^2$,

$$\begin{aligned}\pi^k\mathbf{u} = \mathcal{I}^k\mathbf{u} &+ \sum_{j_2} \frac{1}{|e_{j_2}|}\langle(\mathbf{u}-\mathcal{I}^k\mathbf{u})\cdot\tau_{e_{j_2}}, \lambda^{j_2}\rangle_{e_{j_2}}\,\mathbf{v}_{j_2}^{(\mathrm{ii})} \\ &+ \sum_{j_3} \frac{1}{|e_{j_3}|}\langle(\mathbf{u}-\mathcal{I}^k\mathbf{u})\cdot\nu_{e_{j_3}}, \lambda^{j_3}\rangle_{e_{j_3}}\,\mathbf{v}_{j_3}^{(\mathrm{iii})} \\ &+ \sum_{j_4} \frac{1}{|R_{j_4}|}(\mathbf{u}-\mathcal{I}^k\mathbf{u}, \mathbf{v}^{j_4})\,\mathbf{v}_{j_4}^{(\mathrm{iv})},\end{aligned} \tag{5.2}$$

which is indeed in $\mathbf{V}_h^k$. We note that, after a local change of variables,

$$\langle\mathbf{v}\cdot\xi,\lambda\rangle_e = |e|\langle\hat{\mathbf{v}}\cdot\hat{\xi},\hat{\lambda}\rangle_{\hat{e}} \quad\text{and}\quad (\mathbf{v},\psi)_R = |R|(\hat{\mathbf{v}},\hat{\psi})_{\hat{R}},$$

and thus by our normalization of the reference basis functions, the degrees of freedom (ii)–(iv) of $\pi^k\mathbf{u}$ match $\mathbf{u}$, and $\pi^k\mathbf{u}$ has the correct values at the grid points.

LEMMA 5.2. *Assume that $u \in (H^1(\Omega))^2$.*

(a) *The linear operator π^k is bounded on $(H^1(\Omega))^2$ independently of h.*

(b) *There exists some constant C independent of h such that for $R \in \mathcal{T}_h$ and $\mathbf{u} \in (H^r(\Delta_R))^2$,*

$$|\pi^k\mathbf{u}-\mathbf{u}|_{j,R} \le C|\mathbf{u}|_{r,\Delta_R}h_R^{r-j}, \quad 1 \le r \le k+1,\ j = 0,1,$$

where $h_R = \operatorname{diam}(R)$ and Δ_R is the union of R and its nearest neighboring elements in $\mathcal{T}_h$.

(c) *For $\mathbf{u} \in (H^r(\Omega))^2$,*

$$|\pi^k\mathbf{u}-\mathbf{u}|_j \le C|\mathbf{u}|_r h^{r-j}, \quad 1 \le r \le k+1,\ j = 0,1.$$

(d) $\mathcal{P}_W^{k-1}\nabla\cdot\mathbf{u} = \mathcal{P}_W^{k-1}\nabla\cdot\pi^k\mathbf{u}$.

Proof. To show (a), we use the representation (5.2) derived above, which when squared implies

$$\begin{aligned}|\pi^k\mathbf{u}|^2 \le |\mathcal{I}^k\mathbf{u}|^2 &+ \sum_{j_2} \frac{1}{|e_{j_2}|^2}|\langle(\mathbf{u}-\mathcal{I}^k\mathbf{u})\cdot\tau_{e_{j_2}}, \lambda^{j_2}\rangle_{e_{j_2}}\,\mathbf{v}_{j_2}^{(\mathrm{ii})}|^2 \\ &+ \sum_{j_3} \frac{1}{|e_{j_3}|^2}|\langle(\mathbf{u}-\mathcal{I}^k\mathbf{u})\cdot\nu_{e_{j_3}}, \lambda^{j_3}\rangle_{e_{j_3}}\mathbf{v}_{j_3}^{(\mathrm{iii})}|^2 \\ &+ \sum_{j_4} \frac{1}{|R_{j_4}|^2}|(\mathbf{u}-\mathcal{I}^k\mathbf{u}, \mathbf{v}^{j_4})\,\mathbf{v}_{j_4}^{(\mathrm{iv})}|^2,\end{aligned}$$

since at each point $x \in \Omega$ the sums are finite with the number depending on k but not on h. A similar expression holds for the gradient, and so, after integrating,

$$\begin{aligned}\|\pi^k \mathbf{u}\|_1^2 \le \|\mathcal{I}^k \mathbf{u}\|_1^2 &+ \sum_{j_2} \frac{1}{|e_{j_2}|^2} \|\langle (\mathbf{u} - \mathcal{I}^k \mathbf{u}) \cdot \tau_{e_{j_2}}, \lambda^{j_2} \rangle_{e_{j_2}}^2 \, \mathbf{v}_{j_2}^{(\mathrm{ii})}\|_1^2 \\ &+ \sum_{j_3} \frac{1}{|e_{j_3}|^2} \|\langle (\mathbf{u} - \mathcal{I}^k \mathbf{u}) \cdot \nu_{e_{j_3}}, \lambda^{j_3} \rangle_{e_{j_3}}^2 \, \mathbf{v}_{j_3}^{(\mathrm{iii})}\|_1^2 \\ &+ \sum_{j_4} \frac{1}{|R_{j_4}|^2} \|(\mathbf{u} - \mathcal{I}^k \mathbf{u}, \mathbf{v}^{j_4})^2 \, \mathbf{v}_{j_4}^{(\mathrm{iv})}\|_1^2. \end{aligned} \tag{5.3}$$

We begin with the second term on the right-hand side above. By the standard affine change of variables $R \mapsto \hat{R}$ and the quasi regularity of the grid, we deduce that for any ψ

$$\|\psi\|_{0,R} = |R|^{1/2} \, \|\hat{\psi}\|_{0,\hat{R}}, \tag{5.4}$$

$$|\psi|_{1,R} \le C|R|^{1/2} h^{-1} \, |\hat{\psi}|_{1,\hat{R}} \le C \, |\hat{\psi}|_{1,\hat{R}} \le C \, |\psi|_{1,R}, \tag{5.5}$$

since in two dimensions $|R|^{1/2} = O(h)$. Now

$$\|\mathbf{v}_{j_2}^{(\mathrm{ii})}\|_1 = \left\{ \sum_R \|\mathbf{v}_{j_2}^{(\mathrm{ii})}\|_{1,R}^2 \right\}^{1/2} \le C \left\{ \sum_R \|\hat{\mathbf{v}}_{j_2}^{(\mathrm{ii})}\|_{1,\hat{R}}^2 \right\}^{1/2} \le C,$$

since $\hat{\mathbf{v}}_{j_2}^{(\mathrm{ii})}$ is supported in at most four rectangles and is independent of h. Moreover, if $e_{j_2} \subset R' \in \mathcal{T}_h$, then by the trace theorem (see, e.g., [26]),

$$\begin{aligned}\frac{1}{|e_{j_2}|} |\langle \mathbf{v} \cdot \tau_{e_{j_2}}, \lambda^{j_2} \rangle_{e_{j_2}}| &= |\langle \hat{\mathbf{v}} \cdot \hat{\tau}_{\hat{e}_{j_2}}, \hat{\lambda}^{j_2} \rangle_{\hat{e}_{j_2}}| \\ &\le C \|\hat{\mathbf{v}}\|_{1,\hat{R}'} \|\hat{\lambda}^{j_2}\|_{0,\hat{e}_{j_2}} \le C\big\{ h^{-1} \|\mathbf{v}\|_{0,R'} + |\mathbf{v}|_{1,R'} \big\}, \end{aligned}$$

and so

$$\begin{aligned}\sum_{j_2} \frac{1}{|e_{j_2}|^2} \|\langle \mathbf{v} \cdot \tau_{e_{j_2}}, \lambda^{j_2} \rangle_{e_{j_2}} \mathbf{v}_{j_2}^{(\mathrm{ii})}\|_1^2 &\le C \sum_{j_2} \big\{ h^{-1} \|\mathbf{v}\|_{0,R'} + |\mathbf{v}|_{1,R'} \big\}^2 \\ &\le C\big\{ h^{-2} \|\mathbf{v}\|_0^2 + |\mathbf{v}|_1^2 \big\}, \end{aligned} \tag{5.6}$$

since a given rectangle $R' \supset e_{j_2}$ appears at most four times in the sum. Similarly,

$$\sum_{j_3} \frac{1}{|e_{j_3}|^2} \|\langle \mathbf{v} \cdot \nu_{e_{j_3}}, \lambda^{j_3} \rangle_{e_{j_3}} \mathbf{v}_{j_3}^{(\mathrm{iii})}\|_1^2 \le C\big\{ h^{-2} \|\mathbf{v}\|_0^2 + |\mathbf{v}|_1^2 \big\} \tag{5.7}$$

and

$$\sum_{j_4} \frac{1}{|R_{j_4}|^2} \|(\mathbf{v}, \mathbf{v}^{j_4}) \, \mathbf{v}_{j_4}^{(\mathrm{iv})}\|_1^2 \le C h^{-2} \|\mathbf{v}\|_0^2 \, . \tag{5.8}$$

Thus from (5.3) we deduce that

$$\|\pi^k \mathbf{u}\|_1 \le C\big\{ \|\mathcal{I}^k \mathbf{u}\|_1 + h^{-1} \|\mathbf{u} - \mathcal{I}^k \mathbf{u}\|_0 + |\mathbf{u} - \mathcal{I}^k \mathbf{u}|_1 \big\} \le C \|\mathbf{u}\|_1. \tag{5.9}$$

Now (b), with $j = 1$, follows from the Bramble–Hilbert lemma [10], since locally π^k is a kth degree polynomial preserving operator. To prove the $j = 0$ case when $k > 1$, we need only compute

$$\|\mathbf{u} - \pi^k \mathbf{u}\|_{0,R}^2 = \|\mathbf{u} - \bar{\mathbf{u}} - (\pi^k \mathbf{u} - \overline{\pi^k \mathbf{u}})\|_{0,R}^2 \le C \|\mathbf{u} - \pi^k \mathbf{u}\|_{1,R}^2 h^2,$$

where the over-line denotes the local average and we use the fact that $\bar{\mathbf{u}} = \overline{\pi^k \mathbf{u}}$ by degree of freedom (iv). The case of $k = 1$ follows directly from a careful scaling analysis of (5.2) as above and (5.1). The argument appears in [26]: noting (5.4) and modifying (5.6)–(5.8), we see as in (5.9) that

$$\begin{aligned}
\|\mathbf{u} - \pi^k \mathbf{u}\|_0 &\le \|\mathbf{u} - \mathcal{I}^k \mathbf{u}\|_0 + \sum_{j_2} \frac{1}{|e_{j_2}|} \|\langle (\mathbf{u} - \mathcal{I}^k \mathbf{u}) \cdot \tau_{e_{j_2}}, \lambda^{j_2} \rangle_{e_{j_2}} \mathbf{v}_{j_2}^{(\mathrm{ii})}\|_0 \\
&\quad + \sum_{j_3} \frac{1}{|e_{j_3}|} \|\langle (\mathbf{u} - \mathcal{I}^k \mathbf{u}) \cdot \nu_{e_{j_3}}, \lambda^{j_3} \rangle_{e_{j_3}} \mathbf{v}_{j_3}^{(\mathrm{iii})}\|_0 \\
&\quad + \sum_{j_4} \frac{1}{|R_{j_4}|} \|(\mathbf{u} - \mathcal{I}^k \mathbf{u}, \mathbf{v}^{j_4}) \mathbf{v}_{j_4}^{(\mathrm{iv})}\|_0 \\
&\le C \big\{ \|\mathbf{u} - \mathcal{I}^k \mathbf{u}\|_0 + h |\mathbf{u} - \mathcal{I}^k \mathbf{u}|_1 \big\} \le C h^r \|\mathbf{u}\|_r,
\end{aligned}$$

where $r = 1$ or 2.

Result (c) follows from (b). For (d), let $w \in W_h^{k-1}$ and compute

$$\begin{aligned}
(\nabla \cdot \mathbf{u}, w)_R &= \langle \mathbf{u} \cdot \nu, w \rangle_{\partial R} - (\mathbf{u}, \nabla w)_R \\
&= \langle \pi^k \mathbf{u} \cdot \nu, w \rangle_{\partial R} - (\pi^k \mathbf{u}, \nabla w)_R = (\nabla \cdot \pi^k \mathbf{u}, w)_R,
\end{aligned}$$

since $w|_e \in P_{k-1}(e)$ and $\nabla w|_R \in Q_{k-2,k-1}(R) \times Q_{k-1,k-2}(R)$. ∎

Remark 5.1. The usual mixed spaces satisfy

$$\mathcal{P}_{W_{h,usual}} \nabla \cdot \mathbf{u} = \nabla \cdot \pi_{usual} u, \tag{5.10}$$

since $\nabla \cdot \mathbf{V}_{h,usual} \subset W_{h,usual}$. This is *not* true for the spaces of Definition 4.1.

Our spaces can be viewed as generalizations of the Raviart–Thomas spaces RT_{k-1} [29, 31]. Our analysis will make strong use of the RT_{k-1}-projection operator $\tilde{\pi}^{k-1}$. We briefly review RT_{k-1}. By definition,

$$RT_{k-1} = \tilde{\mathbf{V}}_h^{k-1} \times W_h^{k-1},$$

where

$$\begin{aligned}
&\tilde{\mathbf{V}}_h^{k-1}(R) = Q_{k,k-1}(R) \times Q_{k-1,k}(R), \\
&\tilde{\mathbf{V}}_h^{k-1} = \{ \mathbf{v} \in H(\mathrm{div}; \Omega) : \mathbf{v}|_R \in \tilde{\mathbf{V}}_h^{k-1}(R) \quad \forall R \in \mathcal{T}_h \}.
\end{aligned}$$

The operator

$$\tilde{\pi}^{k-1} : (H^1(\Omega))^2 \to \tilde{\mathbf{V}}_h^{k-1}$$

is defined on a rectangle R by the degrees of freedom

(i) $\langle \tilde{\pi}^{k-1} \mathbf{u} \cdot \nu, \lambda \rangle_e = \langle \mathbf{u} \cdot \nu, \lambda \rangle_e$ for all $\lambda \in P_{k-1}(e)$ and edges $e \subset \partial R$,

(ii) $(\tilde{\pi}^{k-1} \mathbf{u}, \mathbf{v})_R = (\mathbf{u}, \mathbf{v})_R$ for all $\mathbf{v} \in Q_{k-2,k-1}(R) \times Q_{k-1,k-2}(R)$.

We recall that

$$\nabla \cdot \tilde{\pi}^{k-1}\mathbf{u} = \mathcal{P}_W^{k-1}\nabla \cdot \mathbf{u}, \tag{5.11}$$

$$\|\mathbf{u} - \tilde{\pi}^{k-1}\mathbf{u}\|_0 \le C\|\mathbf{u}\|_r h^r, \quad 1 \le r \le k. \tag{5.12}$$

Let $\mathcal{P}_{\tilde{V}}^{k-1} : (L^2(\Omega))^2 \to \tilde{\mathbf{V}}_h^{k-1}$ denote projection $\mathcal{P}_{\tilde{V}_h^{k-1}}$. Relations between π^k and $\tilde{\pi}^{k-1}$ are given as follows.

LEMMA 5.3. *Assume that* $\mathbf{u} \in (H^1(\Omega))^2$.

(a) $\tilde{\pi}^{k-1}\pi^k\mathbf{u} = \tilde{\pi}^{k-1}\mathbf{u}$.

(b) $\tilde{\pi}^{k-1}\mathbf{u} = \mathcal{P}_{\tilde{V}}^{k-1}\pi^k\mathbf{u}$.

(c) $\tilde{\pi}^{k-1}\mathbf{u} = \mathcal{P}_{\tilde{V}}^{k-1}\mathbf{u}$, *provided* $\mathbf{u} \in \mathbf{V}_h^k$.

Proof. For (a), π^k preserves the degrees of freedom of $\tilde{\pi}^{k-1}$. Result (b) is a corollary of (a) and (c).

For (c), since $\tilde{\pi}^{k-1}$ is constructed locally on each element, we restrict our analysis to a rectangle R. Assume without loss of generality that $R = \hat{R}$. Since x and y components are independent, we consider only x-components. Then

$$\tilde{V}_{h,1}^{k-1} = \text{span}\{\lambda_i(x)\lambda_j(y) : i \le k\ ,\ j \le k-1\},$$
$$V_{h,1}^{k} = \text{span}\{\lambda_i(x)\lambda_j(y) : i \le k\ ,\ j \le k+1\}.$$

We wish to show that for $\mathbf{u} \in \mathbf{V}_h^k(R)$,

$$((\tilde{\pi}^{k-1}\mathbf{u})_1 - u_1, v_1)_R = 0 \quad \forall\ v_1 \in \tilde{V}_{h,1}^{k-1}(R).$$

If $\mathbf{u} \in \tilde{\mathbf{V}}_h^{k-1}(R)$, the result is trivial, so assume

$$u_1 \in \text{span}\{\lambda_i(x)\lambda_j(y) : i \le k, \quad j = k, k+1\}.$$

Then $(u_1, v_1)_R = 0$ for all $v_1 \in \tilde{V}_{h,1}^{k-1}(R)$. Moreover, if e is an edge of ∂R with x constant (i.e., with $\nu = \mathbf{e}_1$), then by the orthogonality of the Legendre polynomials,

(i) $\langle u_1(x,\cdot), q\rangle_e = 0$ for all $q(y) \in P_{k-1}(e)$,

(ii) $(u_1, v_1)_R = 0$ for all $v_1 \in Q_{k-2,k-1}(R)$.

These are the degrees of freedom that define the x-component of $\tilde{\pi}^{k-1}$, and so $(\tilde{\pi}^{k-1}\mathbf{u})_1 = 0 = (\mathcal{P}_{\tilde{V}}^{k-1}\mathbf{u})_1$. ∎

6. A mixed finite element method. To approximate (1.1)–(1.2), (1.4)–(1.5), we first rewrite the equations in mixed variational form. Let

$$\mathbf{V}_0 = \{\mathbf{v} \in H(\text{div};\Omega) : \mathbf{v} \cdot \nu = 0 \text{ on } \Gamma_N\},$$

and $W = L^2(\Omega)$, unless $a \equiv 0$ and $\Gamma_N = \partial\Omega$ (the pure Neumann problem), in which case $W = \{w \in L^2(\Omega) : \int_\Omega w = 0\}$. In the latter case, we also assume the usual compatibility condition between b and g_N.

We find $(\mathbf{u}, p) \in H(\text{div};\Omega) \times W$ such that $\mathbf{u} \cdot \nu = g_N$ on Γ_N and

$$(ap, w) + (\nabla \cdot \mathbf{u}, w) = (b, w) \qquad \forall w \in W, \tag{6.1}$$

$$(d\mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) = -\langle p_D, \mathbf{v} \cdot \nu\rangle_{\Gamma_D} + (\mathbf{c}, \mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{V}_0. \tag{6.2}$$

Normally one merely restricts to the finite element spaces to define the mixed method. For our spaces, however, this is not the proper definition of the method

when nonhomogeneous boundary terms appear. We assume that the grid is such that for any edge $e \subset R \in \mathcal{T}_h$, either $e \subset \Gamma_N$ or $e \subset \Gamma_D$.

We begin with Dirichlet conditions. Let us define for edge $e \subset \Gamma_N \cap \partial R$, $R \in \mathcal{T}_h$,

$$\Lambda_h^k(e) = \mathbf{V}_h^k(R) \cdot \nu = P_{k+1}(e),$$
$$\tilde{\Lambda}_h^{k-1}(e) = \tilde{\mathbf{V}}_h^{k-1}(R) \cdot \nu = P_{k-1}(e),$$

and the full spaces Λ_h^k of continuous functions and $\tilde{\Lambda}_h^{k-1}$ of discontinuous functions over Γ_N. With the definition $\mathcal{P}_{\tilde{\Lambda}}^{k-1} = \mathcal{P}_{\tilde{\Lambda}_h^{k-1}}$, we replace p_D by $\mathcal{P}_{\tilde{\Lambda}}^{k-1} p_D$ in the equations.

For Neumann conditions, we need to set $\mathbf{u}_h \cdot \nu$ on Γ_N. This can be done in any reasonable way, but for the error analysis to follow it is convenient to set $\mathbf{u}_h \cdot \nu$ to $\pi^k \mathbf{u}$. Since $\mathbf{u}$ is unknown, we need to define it using only g_N. Near the boundary only, we use the Scott–Zhang [33] modification of the Clément operator $\mathcal{I}^k$ [19] considered in section 5, which has similar properties, except that in (5.1) we must have $r \geq 1$. We set the corner values to a local L^2-projection defined entirely on the boundary of the domain, instead of over rectangles in the domain, so that $\mathcal{I}^k \mathbf{u} \cdot \nu|_{\Gamma_N}$ is defined entirely by $\mathbf{u} \cdot \nu|_{\Gamma_N} = g_N$. Further, we can define an operator $\mathcal{I}_\Lambda^k : L^2(\Gamma_N) \to \Lambda_h^k$ by

$$\mathcal{I}_\Lambda^k g(P) = \mathcal{I}^k g(P)$$

for each corner point P of the grid restricted to Γ_N, and, on each grid edge $e \subset \Gamma_N$,

$$\langle \mathcal{I}_\Lambda^k g, \lambda \rangle_e = \langle g, \lambda \rangle_e \quad \forall \lambda \in P_k(e).$$

Then it is easy to check that

$$\pi^k \mathbf{u} \cdot \nu|_{\Gamma_N} = \mathcal{I}_\Lambda^k g_N.$$

Finally, to set the Neumann boundary condition, we set $\mathbf{u}_h \cdot \nu = \mathcal{I}_\Lambda^k g_N$ on Γ_N.

Let

$$\mathbf{V}_{h,0}^k = \{\mathbf{v} \in \mathbf{V}_h^k : \mathbf{v} \cdot \nu = 0 \text{ on } \Gamma_N\}.$$

We now define our mixed finite element method for (6.1)–(6.2). Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h^k \times W_h^{k-1}$ such that $\mathbf{u}_h \cdot \nu = \mathcal{I}_\Lambda^k g_N$ on Γ_N and

$$(ap_h, w_h) + (\nabla \cdot \mathbf{u}_h, w_h) = (b, w_h) \qquad \forall w_h \in W_h, \tag{6.3}$$
$$(d\mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) = -\langle \mathcal{P}_{\tilde{\Lambda}}^{k-1} p_D, \mathbf{v}_h \cdot \nu \rangle_{\Gamma_D} + (\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} \mathbf{c}, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in \mathbf{V}_{h,0}^k. \tag{6.4}$$

7. Mixed method error analysis. We now analyze the error in approximating (6.1)–(6.2) by (6.3)–(6.4). In (6.2) let $\mathbf{v}$ be replaced by $\tilde{\pi}^{k-1} \mathbf{v}_h$ for $\mathbf{v}_h \in \mathbf{V}_{h,0}^k$. By Lemma 5.3, $\tilde{\pi}^{k-1} \mathbf{v}_h = \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} \mathbf{v}_h$, and with (5.11) we obtain

$$\begin{aligned}(\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u}, \mathbf{v}_h) - (\mathcal{P}_W^{k-1} p, \nabla \cdot \mathbf{v}_h) &= -\langle \mathcal{P}_{\tilde{\Lambda}}^{k-1} p_D, \mathbf{v}_h \cdot \nu \rangle_{\Gamma_D} \\ &\quad + (\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} \mathbf{c}, \mathbf{v}_h), \quad \mathbf{v}_h \in \mathbf{V}_{h,0}^k. \end{aligned} \tag{7.1}$$

Note that in (6.1), if $w \in W_h^{k-1}$, we can replace $\mathbf{u}$ by $\pi^k \mathbf{u}$. Thus the difference of (6.1), (7.1) and (6.3)–(6.4) is

$$(a(p - p_h), w_h) + (\nabla \cdot (\pi^k \mathbf{u} - \mathbf{u}_h), w_h) = 0, \qquad w_h \in W_h^{k-1}, \tag{7.2}$$
$$(\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u} - d\mathbf{u}_h, \mathbf{v}_h) - (\mathcal{P}_W^{k-1} p - p_h, \nabla \cdot \mathbf{v}_h) = 0, \quad \mathbf{v}_h \in \mathbf{V}_{h,0}^k. \tag{7.3}$$

Select the standard test functions

$$\mathbf{v}_h = \pi^k \mathbf{u} - \mathbf{u}_h \in \mathbf{V}_{h,0}^k \quad \text{and} \quad w_h = \mathcal{P}_W^{k-1} p - p_h \in W_h^{k-1} \ .$$

Then the sum of (7.2)–(7.3) is

$$\begin{aligned}(a(p-p_h), p-p_h) &+ (d(\mathbf{u}-\mathbf{u}_h), \mathbf{u}-\mathbf{u}_h) \\ &= (a(p-p_h), p-\mathcal{P}_W^{k-1}p) + (d(\mathbf{u}-\mathbf{u}_h), \mathbf{u}-\pi^k\mathbf{u}) + (d\mathbf{u} - \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u}, \pi^k\mathbf{u} - \mathbf{u}_h),\end{aligned}$$

from which we conclude, using standard approximation theory [10, 18] and Lemma 5.2, that

$$\begin{aligned}&\|\sqrt{a}(p-p_h)\|_0 + \|d^{1/2}(\mathbf{u}-\mathbf{u}_h)\|_0 \\ &\quad \le C\big\{\|p - \mathcal{P}_W^{k-1}p\|_0 + \|\mathbf{u}-\pi^k\mathbf{u}\|_0 + \|d\mathbf{u} - \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u}\|_0\big\} \\ &\quad \le C\big\{\|p\|_r + \|\mathbf{u}\|_r\big\}h^r \qquad (7.4)\end{aligned}$$

for any $1 \le r \le k$.

THEOREM 7.1. *If $\mathbf{u} \in (H^1(\Omega))^2$, the coefficients $a \ge 0$ and d are sufficiently smooth, and d is uniformly elliptic, then there is some constant C, independent of h, such that*

$$\begin{aligned}&\text{(a) } \|\sqrt{a}(p-p_h)\|_0 + \|\mathbf{u}-\mathbf{u}_h\|_0 \le C\big\{\|p\|_r + \|\mathbf{u}\|_r\big\}h^r, && 1 \le r \le k, \\ &\text{(b) } \|\mathcal{P}_W^{k-1}p - p_h\|_0 \le C\big\{\|p\|_r + \|\mathbf{u}\|_{r+1}\big\}h^{r+1}, && 0 \le r \le k, \\ &\text{(c) } \|p-p_h\|_0 \le C\big\{\|p\|_r + \|\mathbf{u}\|_r\big\}h^r, && 1 \le r \le k, \\ &\text{(d) } \|\mathcal{P}_W^{k-1}\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_0 \le C\big\{\|p\|_r + \|\mathbf{u}\|_r\big\}h^r, && 1 \le r \le k, \\ &\text{(e) } \|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_0 \le C\big\{\|p\|_r + \|\mathbf{u}\|_r\big\}h^{r-1}, && 1 \le r \le k,\end{aligned}$$

where for (b) *and* (c) *we have assumed that problem* (7.5)–(7.7) *below is* 2*-regular, i.e., its solution satisfies* (7.8)*, and for* (e) *we assume that the finite element partition is quasi-uniform.*

While results (c) and (d) are optimal in the rate of convergence, results (a) and (e) are only suboptimal. Result (b) exhibits a superconvergence phenomenon typical of mixed methods. If $k = 1$, we do not control the full divergence error; however, we at least have stability. See, e.g., [25, 27] for conditions that imply 2-regularity.

Proof. We have shown (a) above. For (b) and (c), let $\psi \in L^2(\Omega)$ be such that $\|\psi\|_0 \le 1$ and consider the solution $\varphi \in H^2(\Omega)$ to

$$a\varphi - \nabla\cdot d^{-1}\nabla\varphi = \psi, \quad \Omega, \qquad (7.5)$$

$$d^{-1}\nabla\varphi\cdot\nu = 0, \quad \Gamma_N, \qquad (7.6)$$

$$\varphi = 0\ , \quad \Gamma_D. \qquad (7.7)$$

By hypothesis the problem is 2-regular, so by definition there is some constant $C > 0$ such that

$$\|\varphi\|_2 \le C\|\psi\|_0 \le C. \qquad (7.8)$$

Then, using Lemma 5.2(d),

$$\begin{aligned}(\mathcal{P}_W^{k-1}p - p_h, \psi) &= (a(\mathcal{P}_W^{k-1}p - p_h), \varphi) - (\mathcal{P}_W^{k-1}p - p_h, \nabla\cdot d^{-1}\nabla\varphi) \\ &= (a(\mathcal{P}_W^{k-1}p - p_h), \varphi) - (\mathcal{P}_W^{k-1}p - p_h, \nabla\cdot \pi^k d^{-1}\nabla\varphi) \\ &= (a(\mathcal{P}_W^{k-1}p - p_h), \varphi) - (\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u} - d\mathbf{u}_h, \pi^k d^{-1}\nabla\varphi),\end{aligned}$$

by (7.3), since $\pi^k d^{-1}\nabla\varphi \in \mathbf{V}_{h,0}^k$. Our immediate goal is to cancel the coefficients d and d^{-1} in the last term, so we compute

$$\begin{aligned}(\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u} - d\mathbf{u}_h, \pi^k d^{-1}\nabla\varphi) &= (\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u} - d\mathbf{u}_h, \pi^k d^{-1}\nabla\varphi - \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d^{-1}\nabla\varphi)\\ &\quad + (\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u} - d\mathbf{u}_h, \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d^{-1}\nabla\varphi)\\ &= (\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u} - d\mathbf{u}_h, \pi^k d^{-1}\nabla\varphi - \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d^{-1}\nabla\varphi)\\ &\quad + (d(\mathbf{u} - \mathbf{u}_h), \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d^{-1}\nabla\varphi - d^{-1}\nabla\varphi) + (\mathbf{u} - \mathbf{u}_h, \nabla\varphi).\end{aligned}$$

We now wish to integrate the last term above by parts and use the other error equation (7.2). However, we must do this carefully, so as to obtain the superconvergence claimed in the theorem. We compute

$$\begin{aligned}(\mathbf{u} - \mathbf{u}_h, \nabla\varphi) &= (\mathbf{u} - \pi^k\mathbf{u}, \nabla\varphi) + (\pi^k\mathbf{u} - \mathbf{u}_h, \nabla\varphi - \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1}\nabla\varphi)\\ &\quad + (\pi^k\mathbf{u} - \mathbf{u}_h, \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1}\nabla\varphi).\end{aligned}$$

Now, using Lemma 5.3, (5.11), and (7.2),

$$\begin{aligned}(\pi^k\mathbf{u} - \mathbf{u}_h, \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1}\nabla\varphi) &= (\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1}(\pi^k\mathbf{u} - \mathbf{u}_h), \nabla\varphi)\\ &= (\tilde{\pi}^{k-1}(\pi^k\mathbf{u} - \mathbf{u}_h), \nabla\varphi)\\ &= -(\nabla\cdot\tilde{\pi}^{k-1}(\pi^k\mathbf{u} - \mathbf{u}_h), \varphi)\\ &= -(\nabla\cdot(\pi^k\mathbf{u} - \mathbf{u}_h), \mathcal{P}_W^{k-1}\varphi)\\ &= (a(p - p_h), \mathcal{P}_W^{k-1}\varphi)\\ &= (a(\mathcal{P}_W^{k-1}p - p_h), \mathcal{P}_W^{k-1}\varphi) + ((a - \mathcal{P}_W^0 a)(p - \mathcal{P}_W^{k-1}p), \mathcal{P}_W^{k-1}\varphi),\end{aligned}$$

since $\mathcal{P}_W^{k-1}(\mathcal{P}_W^0 a(p - \mathcal{P}_W^{k-1}p)) = 0$. Combining and using the approximation properties of the various projections and (7.4), we obtain that

$$\begin{aligned}(\mathcal{P}_W^{k-1}p - p_h, \psi) &= (a(\mathcal{P}_W^{k-1}p - p_h), \varphi - \mathcal{P}_W^{k-1}\varphi)\\ &\quad - (\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u} - d\mathbf{u}_h, \pi^k d^{-1}\nabla\varphi - \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d^{-1}\nabla\varphi)\\ &\quad - (d(\mathbf{u} - \mathbf{u}_h), \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d^{-1}\nabla\varphi - d^{-1}\nabla\varphi) - (\mathbf{u} - \pi^k\mathbf{u}, \nabla\varphi)\\ &\quad - (\pi^k\mathbf{u} - \mathbf{u}_h, \nabla\varphi - \mathcal{P}_{\tilde{\mathbf{V}}}^{k-1}\nabla\varphi) - ((a - \mathcal{P}_W^0 a)(p - \mathcal{P}_W^{k-1}p), \mathcal{P}_W^{k-1}\varphi)\\ &\le C\big\{\big(\|\sqrt{a}(\mathcal{P}_W^{k-1}p - p_h)\|_0 + \|\mathbf{u} - \mathbf{u}_h\|_0 + \|\mathcal{P}_{\tilde{\mathbf{V}}}^{k-1} d\mathbf{u} - d\mathbf{u}\|_0\\ &\qquad + \|\pi^k\mathbf{u} - \mathbf{u}\|_0 + \|a\|_{W^{1,\infty}}\|p - \mathcal{P}_W^{k-1}p\|_0\big)\, h + \|\mathbf{u} - \pi^k\mathbf{u}\|_0\big\}\|\varphi\|_2\\ &\le C\big\{\|p\|_r + \|\mathbf{u}\|_{r+1}\big\}h^{r+1},\end{aligned}$$

where $0 \le r \le k$. Thus (b) and (c) follow.

Substituting $w_h = \mathcal{P}_W^{k-1}\nabla\cdot(\mathbf{u} - \mathbf{u}_h) \in W_h^{k-1}$ into (7.2) and using (7.4) leads to (d). For (e), we have a standard inverse inequality argument [10]. On a quasi-uniform partition (i.e., one that is quasi-regular and has the size of the largest to smallest rectangle bounded independently of h), we compute

$$\begin{aligned}\|\nabla\cdot(\mathbf{u} - \mathbf{u}_h)\|_0 &\le \|\nabla\cdot(\mathbf{u} - \pi^k\mathbf{u})\|_0 + \|\nabla\cdot(\pi^k\mathbf{u} - \mathbf{u}_h)\|_0\\ &\le \|\mathbf{u} - \pi^k\mathbf{u}\|_1 + Ch^{-1}\|\pi^k\mathbf{u} - \mathbf{u}_h\|_0,\end{aligned}$$

and the result follows easily. □

8. The inf-sup condition. Our spaces satisfy the celebrated LBB or inf-sup condition of Ladyzhenskaya [28], Babuška [6], Brezzi [11], and Brezzi and Fortin [15].

THEOREM 8.1. *If Ω is a polygonal domain, then there exists a constant $\gamma > 0$ such that*

$$\inf_{w \in W_h^{k-1}} \sup_{\mathbf{v} \in \mathbf{V}_{h,0}^k} \frac{(\nabla \cdot \mathbf{v}, w)}{\|\mathbf{v}\| \|w\|_0} \geq \gamma > 0,$$

wherein functions of W_h^{k-1} have mean zero if $\Gamma_N = \partial\Omega$ and we take the norm on $\mathbf{V}_h^k$ to be any one of

$$\|\mathbf{v}\| = \left\{ \|\mathbf{v}\|_0^2 + \|\nabla \cdot \mathbf{v}\|_0^2 \right\}^{1/2} \equiv \|\mathbf{v}\|_{H(\mathrm{div})},$$
$$\|\mathbf{v}\| = \left\{ \|\mathbf{v}\|_0^2 + \|\mathcal{P}_W^{k-1} \nabla \cdot \mathbf{v}\|_0^2 \right\}^{1/2} \equiv |||v|||,$$

or $\|\mathbf{v}\| = \|\mathbf{v}\|_1$.

Proof. It suffices to show the result for the $(H^1(\Omega))^2$-norm on V_h^k, since

$$|||\mathbf{v}||| \leq \|\mathbf{v}\|_{H(\mathrm{div})} \leq \|\mathbf{v}\|_1.$$

There exists a $\beta > 0$ such that, given $w \in L^2(\Omega)$ (or $w \in L^2(\Omega)/\mathbb{R}$ if $\Gamma_N = \partial\Omega$), there is some $\psi \in (H^1(\Omega))^2$ such that both $\nabla \cdot \psi = w$ and $\|\psi\|_1 \leq \beta \|w\|_0$ (see [5]). Then for $w \in W_h^{k-1}$,

$$\sup_{\mathbf{v} \in \mathbf{V}_{h,0}^k} \frac{(\nabla \cdot \mathbf{v}, w)}{\|\mathbf{v}\|_1 \|w\|_0} \geq \frac{(\nabla \cdot \pi^k \psi, w)}{\|\pi^k \psi\|_1 \|w\|_0} \geq \frac{\|w\|_0}{C \|\psi\|_1} \geq \frac{1}{\beta C} > 0,$$

since π^k is bounded on $(H^1(\Omega))^2$. ∎

We analyzed our finite element method in a direct way in section 7 above, rather than use the inf-sup theory of saddle point problems [6, 10, 11, 15]. This is because it is not straightforward to apply the theory. To do so would require both that the form $(\nabla \cdot \mathbf{v}, w)$ be continuous and that $(d\mathbf{v}, \mathbf{v})$ be coercive on the set $Z = \{\mathbf{v} \in \mathbf{V}_{h,0}^k : (\nabla \cdot \mathbf{v}, w) = 0 \ \forall w \in W_h^{k-1}\}$. For the former, we would need to take the norm on $\mathbf{V}_h^k$ to be at least $\|\cdot\|_{H(\mathrm{div})}$, while for the latter we would need the norm to be no stronger than $|||\cdot|||$. The problem is that testing $(\nabla \cdot \mathbf{v}, w)$ by $w \in W_h^{k-1}$ does not control the full divergence of $\mathbf{v}$, but only its projection $\mathcal{P}_W^{k-1} \nabla \cdot \mathbf{v}$.

9. Numerical examples. In this section we present some numerical results on the problem (1.1)–(1.2), (1.5) with $\Omega = (0,1)^2$, $\Gamma_D = \partial\Omega$, $a = 0$, and $\mathbf{c} = 0$. We fix the true solution and then define p_D and b so that the equations are satisfied. The test cases are summarized in Table 9.1, wherein d and p are defined, as well as a statement as to whether the grid is uniform or not. The observed convergence errors are shown in Table 9.2, and in Table 9.3 we show the observed convergence rates. The norms of the errors were computed using a tensor product 3-point Gauss rule, and the convergence rates were obtained by fitting the norms of the errors to $\exp(m \log h + b)$, with m being the convergence rate.

As can be seen from Table 9.3, p converges to p_h as $O(h)$, and $\mathcal{P}_W^0 p$ as $O(h^2)$, both as expected from Theorem 7.1. We also see that $\mathbf{u}$ converges to $\mathbf{u}_h$ somewhat better than expected. It appears that on the uniform grid we attain $O(h^{3/2})$ superconvergence in the L^2-norm, and $O(h^{1/2})$ in the H^1-seminorm. On the random grid, the superconvergence appears to be lost, at least mostly so. For these test cases,

TABLE 9.1
Numerical test cases.

Case	Coefficient d	True solution	Grid
1	1	$y^4 e^x$	uniform
2	1	$\dfrac{\cos(x^2 y)}{x^2+xy+1}$	uniform
3	$\begin{pmatrix} e^{2xy^2} & 0 \\ 0 & 1/(1.1+x^2-y) \end{pmatrix}$	$\dfrac{\cos(x^2 y)}{x^2+xy+1}$	uniform
4	$\begin{pmatrix} e^{2xy^2} & 0 \\ 0 & 1/(1.1+x^2-y) \end{pmatrix}$	$\dfrac{\cos(x^2 y)}{x^2+xy+1}$	random

TABLE 9.2
Observed errors.

Case	h	$\|p-p_h\|_0$	$\|\mathcal{P}_W^0 p-p_h\|_0$	$\|\mathbf{u}-\mathbf{u}_h\|_0$	$\|\nabla(\mathbf{u}-\mathbf{u}_h)\|_0$	$\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_0$
1	1/8	9.90e−2	1.86e−3	8.11e−2	3.43e+0	9.40e−1
	1/16	4.98e−2	4.31e−4	3.01e−2	2.60e+0	4.60e−1
	1/32	2.49e−2	1.03e−4	1.09e−2	1.92e+0	2.27e−1
	1/64	1.25e−2	2.64e−5	3.92e−3	1.38e+0	1.13e−1
2	1/8	2.55e−2	2.57e−4	1.23e−2	5.16e−1	9.82e−2
	1/16	1.27e−2	5.95e−5	4.51e−3	3.91e−1	4.80e−2
	1/32	6.37e−3	1.43e−5	1.62e−3	2.86e−1	2.38e−2
	1/64	3.18e−3	3.22e−6	5.78e−4	2.05e−1	1.18e−2
3	1/8	2.55e−2	4.57e−4	3.06e−2	1.26e+0	2.88e−1
	1/16	1.27e−2	1.17e−4	1.17e−2	9.93e−1	1.57e−1
	1/32	6.37e−3	2.96e−5	4.29e−3	7.45e−1	8.00e−2
	1/64	3.18e−3	6.30e−6	1.54e−3	5.43e−1	4.00e−2
4	1/8	2.68e−2	4.81e−4	3.17e−2	1.23e+0	3.00e−1
	1/16	1.34e−2	1.18e−4	1.20e−2	9.95e−1	1.50e−1
	1/32	6.68e−3	3.38e−5	5.37e−3	8.73e−1	8.03e−2
	1/64	3.38e−3	5.80e−6	2.44e−3	6.85e−1	4.04e−2

TABLE 9.3
Observed convergence rates. Best fit to $y = \exp(m \log h + b)$, with m reported below (where y is the norm of the error).

Case	$\|p-p_h\|_0$	$\|\mathcal{P}_W^0 p-p_h\|_0$	$\|\mathbf{u}-\mathbf{u}_h\|_0$	$\|\nabla(\mathbf{u}-\mathbf{u}_h)\|_0$	$\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_0$
1	0.996	2.049	1.457	0.437	1.020
2	1.000	2.100	1.471	0.445	1.018
3	1.000	2.052	1.438	0.405	0.950
4	0.997	2.092	1.227	0.271	0.958

$\mathcal{P}_W^0 \nabla\cdot\mathbf{u}_h = \mathcal{P}_W^0 \nabla\cdot\mathbf{u}$. However, it appears that $\nabla\cdot\mathbf{u}$ approaches $\nabla\cdot\mathbf{u}_h$ with rate $O(h)$. We have not been able to demonstrate that this is true in general.

If the projection of the Dirichlet data is *not* used in (6.4), then the results degrade significantly. Thus, it is important to include this projection in the method. In our tests, it amounts to simplifying the computation by replacing the Dirichlet value on each boundary element edge by its average value.

10. The spaces in three dimensions. There are several ways to define our mixed finite element spaces in three dimensions. We present the version with as few degrees of freedom as seems possible.

Let

$$\tilde{Q}_{j,j} = \{p \in Q_{j,j} : p \text{ has exact degree } j\}.$$

Then $\dim Q_{j,j} = 2j+1$. Moreover, let

$$\tilde{Q}^x_{i,j,j} = \{p \in Q_{i,j,j} : p \text{ has exact degree } j \text{ in } y \text{ and } z\},$$

and define similarly $\tilde{Q}^y_{j,i,j}$ and $\tilde{Q}^z_{j,j,i}$, each with dimension $(i+1)(2j+1)$.

We next define some spaces of "bubble" functions. For each integer $k \geq 1$, let $R = (a_0, a_1) \times (b_0, b_1) \times (c_0, c_1)$ and

$$\begin{aligned} B^k_x(R) = \big\{ & p : p(x,y,z) \\ & = (y-b_0)(b_1-y)(z-c_0)(c_1-z)\big[(a_1-x)\, q_0(y,z) + (x-a_0)\, q_1(y,z)\big] \\ & \text{for some } q_0, q_1 \in \tilde{Q}_{k-1,k-1}((b_0,b_1)\times(c_0,c_1))\big\} \\ \subset\ & Q_{1,k+1,k+1}(R), \end{aligned}$$

with a similar definition for $B^k_y(R)$ and $B^k_z(R)$. Also let

$$\begin{aligned} \mathbf{B}^k(R) = \big\{ & \mathbf{v} : \mathbf{v}(x,y,z) = (x-a_0)(a_1-x)(y-b_0)(b_1-y)(z-c_0)(c_1-z)\,\psi(x,y,z) \\ & \text{for some } \psi \in \tilde{Q}^x_{k-2,k-1,k-1}(R) \times \tilde{Q}^y_{k-1,k-2,k-1}(R) \times \tilde{Q}^z_{k-1,k-1,k-2}(R)\big\} \\ \subset\ & Q_{k,k+1,k+1}(R) \times Q_{k+1,k,k+1}(R) \times Q_{k+1,k+1,k}(R). \end{aligned}$$

DEFINITION 10.1. *For each integer $k \geq 1$, for any $R \in \mathcal{T}_h$, let*

$$\begin{aligned} \mathbf{V}^k_h(R) &= \big(Q_{k,k,k}(R)\big)^3 + B^k_x(R) \times B^k_y(R) \times B^k_z(R) + \mathbf{B}^k(R) \\ &\subset Q_{k,k+1,k+1}(R) \times Q_{k+1,k,k+1}(R) \times Q_{k+1,k+1,k}(R), \\ W^{k-1}_h(R) &= Q_{k-1,k-1,k-1}(R), \end{aligned}$$

and define $\mathbf{V}^k_h \times W^{k-1}_h$ as in Definition 4.1.

LEMMA 10.2. *Let R be a rectangular parallelepiped in three dimensions. For each edge $e \in \partial R$, fix a set $\mathcal{S}_e$ of $k+1$ distinct points, including the endpoints. Then $\mathbf{u}_h \in \mathbf{V}^k_h(R)$ is uniquely defined by the following degrees of freedom:*

(i) *For each edge $e \in \partial R$, point $P \in \mathcal{S}_e$, and Cartesian direction $j = 1, 2, 3$,*

$$DOF^{(\mathrm{i})}_{P,j}(\mathbf{u}_h) = u_{h,j}(P) = \mathbf{u}_h(P) \cdot \mathbf{e}_j.$$

(ii) *On each face $f \subset \partial R$,*

$$DOF^{(\mathrm{ii})}_{f,\tau_f,\lambda}(\mathbf{u}_h) = \langle \mathbf{u}_h \cdot \tau_f, \lambda \rangle_f \quad \forall \lambda \in Q_{k-2,k-2}(f),$$

where τ_f is one of the two linearly independent Cartesian unit tangential directions.

(iii) *On each face $f \subset \partial R$,*

$$DOF^{(\mathrm{iii})}_{f,\lambda}(\mathbf{u}_h) = \langle \mathbf{u}_h \cdot \nu_f, \lambda \rangle_f \quad \forall \lambda \in Q_{k-1,k-1}(f),$$

where ν_f is the outer unit normal direction.

(iv) *Over R,*

$$DOF^{(\mathrm{iv})}_{\mathbf{v}}(\mathbf{u}_h) = (\mathbf{u}_h, \mathbf{v})_R \quad \forall \mathbf{v} \in Q_{k-2,k-1,k-1}(R) \times Q_{k-1,k-2,k-1}(R) \times Q_{k-1,k-1,k-2}(R).$$

Moreover, on each edge $e \subset \partial R$, $\mathbf{u}_h|_e$ is uniquely defined by the degrees of freedom (i) *restricted to e, and on each face $f \subset \partial R$, $\mathbf{u}_h|_f$ is uniquely defined by the degrees of freedom* (i)–(iii) *restricted to f.*

The dimension of $\mathbf{V}_h^k(R)$ is

$$\begin{aligned}\dim \mathbf{V}_h^k(R) &= 3(k+1)^3 + 6(2k-1) + 3(k-1)(2k-1) = 3(k^3+5k^2+4k)\\ &= 3k(k+1)(k+4).\end{aligned}$$

The number of independent degrees of freedom for (i) is $3(12(k-1)+8) = 36k-12$, for (ii) is $12(k-1)^2 = 12k^2-24k+12$, for (iii) is $6k^2$, and for (iv) is $3k^2(k-1)$, for a total matching the dimension of the space. Therefore the proof of the lemma is similar to that for Lemma 4.2, and we omit it.

The theoretical results of the previous sections can be developed for these spaces in a relatively straightforward way. In particular, we have analogues of the definition of π^k, Lemmas 5.2 and 5.3, the development of the mixed finite element method, and Theorems 7.1 and 8.1.

REFERENCES

[1] T. Arbogast and D. S. Brunson, *A computational method for approximating a Darcy–Stokes system governing a vuggy porous medium*, submitted.

[2] T. Arbogast, D. S. Brunson, S. L. Bryant, and J. J. W. Jennings, *A Computational Investigation of a Macro-Model for Vuggy Porous Media*, manuscript.

[3] T. Arbogast, D. S. Brunson, S. L. Bryant, and J. W. Jennings, *A preliminary computational investigation of a macro-model for vuggy porous media*, in Computational Methods in Water Resources XV, C. T. Miller, M. W. Farthing, W. G. Gray, and G. F. Pindar, eds., Elsevier, New York, 2004.

[4] T. Arbogast and M. F. Wheeler, *A characteristics-mixed finite element method for advection-dominated transport problems*, SIAM J. Numer. Anal., 32 (1995), pp. 404–424.

[5] D. N. Arnold, L. R. Scott, and M. Vogelius, *Regular inversion of the divergence operator with Dirichlet boundary conditions on a polygon*, Ann. Scuola Norm. Sup. Pisa Cl. Sci.-Serie IV, 15 (1988), pp. 169–192.

[6] I. Babuška, *The finite element method with Lagrangian multipliers*, Numer. Math., 20 (1973), pp. 179–192.

[7] J. Bear, *Dynamics of Fluids in Porous Media*, Dover, New York, 1972.

[8] G. S. Beavers and D. D. Joseph, *Boundary conditions at a naturally permeable wall*, J. Fluid Mech., 30 (1967), pp. 197–207.

[9] C. Bernardi and G. Raugel, *Analysis of some finite elements for the Stokes problem*, Math. Comp., 44 (1985), pp. 71–79.

[10] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[11] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, RAIRO Oper. Res., 8 (1974), pp. 129–151.

[12] F. Brezzi, J. Douglas, Jr., R. Durán, and M. Fortin, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.

[13] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

[14] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

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

[16] F. Brezzi, M. Fortin, and L. D. Marini, *Mixed finite element methods with continuous stresses*, Math. Models Methods Appl. Sci., 3 (1993), pp. 275–287.

[17] Z. Chen and J. Douglas, Jr., *Prismatic mixed finite elements for second order elliptic problems*, Calcolo, 26 (1989), pp. 135–148.

[18] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[19] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numèr., 9 (1975), pp. 77–84.
[20] C. N. Dawson, T. F. Russell, and M. F. Wheeler, *Some improved error estimates for the modified method of characteristics*, SIAM J. Numer. Anal., 26 (1989), pp. 1487–1512.
[21] J. Douglas, Jr., and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.
[22] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.
[23] R. E. Ewing, T. F. Russell, and M. F. Wheeler, *Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics*, Comput. Methods Appl. Mech. Engrg., 47 (1984), pp. 73–92.
[24] M. Fortin, *Old and new finite elements for incompressible flows*, Internat. J. Numer. Methods Fluids, 1 (1981), pp. 347–364.
[25] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[26] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[27] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[28] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, 2nd ed., Gordon and Breach, New York, 1969.
[29] J. C. Nédélec, *Mixed finite elements in* $\mathbf{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[30] D. W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, Amsterdam, 1977.
[31] R. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, I. Galligani and E. Magenes, eds., Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.
[32] T. F. Russell and M. F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media*, in The Mathematics of Reservoir Simulation, R. E. Ewing, ed., Frontiers Appl. Math. 1, SIAM, Philadelphia, 1983, pp. 35–106.
[33] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[34] J. M. Thomas, *Sur l'analyse numerique des methodes d'elements finis hybrides et mixtes*, Ph.D. thesis, Sciences Mathematiques, à l'Universite Pierre et Marie Curie, Paris, 1977.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1932–1958

ERROR ESTIMATES FOR A FINITE VOLUME ELEMENT METHOD FOR ELLIPTIC PDEs IN NONCONVEX POLYGONAL DOMAINS*

P. CHATZIPANTELIDIS† AND R. D. LAZAROV†

Abstract. We consider standard finite volume piecewise linear approximations for second order elliptic boundary value problems on a nonconvex polygonal domain. Based on sharp shift estimates, we derive error estimations in H^1-, L_2- and L_∞-norms, taking into consideration the regularity of the data. Numerical experiments and counterexamples illustrate the theoretical results.

Key words. finite volume element method, nonconvex polygons, error estimations

AMS subject classifications. 65N15, 65N30

DOI. 10.1137/S0036142903427639

1. Introduction.

We analyze the standard finite volume element method for the discretization of second order linear elliptic PDEs on a nonconvex polygonal domain $\Omega \subset \mathbb{R}^2$. Namely, for a given function f, we seek u such that

$$Lu \equiv -\operatorname{div}(A\nabla u) = f \quad \text{in } \Omega, \quad \text{and} \quad u = 0 \quad \text{on } \partial\Omega \tag{1.1}$$

with $A = (a_{ij})_{i,j=1}^2$ a given symmetric matrix function with real-value entries $a_{ij} \in W_\infty^1$, $1 \le i, j \le 2$. We assume that the matrix A is uniformly positive definite in Ω, i.e., there exists a positive constant α_0 such that

$$\xi^T A(x)\xi \ge \alpha_0 \xi^T \xi \quad \forall \xi \in \mathbb{R}^2, \forall x \in \bar{\Omega}. \tag{1.2}$$

The class of finite volume methods is based on some approximation of the balance relation

$$-\int_{\partial b} A\nabla u \cdot n \, ds = \int_b f \, dx, \tag{1.3}$$

which is valid for any subdomain $b \subset \Omega$. Here n denotes the outer unit normal vector to the boundary of b.

There are various approaches to the finite volume method. One, the finite volume element method, uses a finite element partition of Ω, where the solution space consists of continuous piecewise linear functions, a collection of vertex-centered control volumes, and a test space of piecewise constant functions over the control volumes (cf., e.g., [6, 10, 25, 28]). A second approach, usually called the finite volume difference method, uses cell-centered grids and approximates the derivatives in the balance equation by finite differences (cf., e.g., [22, 29, 33]). Another approach uses mixed reformulation of the problem [12, 16]. The first approach is quite close to the finite element method but nevertheless has some new properties that make it attractive for the applications [1, 20]. The second approach is closer to the classical finite difference method and extends it to more general than rectangular meshes. It is used mostly on perpendicular bisection or Voronoi type meshes. Approximations on such rectangular

*Received by the editors May 5, 2003; accepted for publication (in revised form) February 6, 2004; published electronically January 20, 2005.
http://www.siam.org/journals/sinum/42-5/42763.html
†Department of Mathematics, Texas A&M University, College Station, TX, 77843 (chatzipa@math.tamu.edu, lazarov@math.tamu.edu).

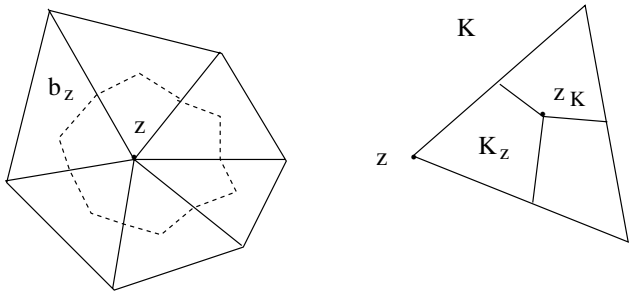


FIG. 1.1. *Left-hand side: A sample region with dotted lines indicating the corresponding box b_z. Right-hand side: A triangle K partitioned into three subregions K_z.*

and triangular meshes were studied, for example, in [34] and [26], respectively. The third approach is close to mixed and hybrid finite element methods and can deal, for example, with irregular quadrilateral and hexahedral cells [12, 30]. Finite volume discretizations for more general convection-diffusion-reaction problems were studied by many authors. For a comprehensive presentation and more references of existing results we refer to the monographs on the finite volume difference method [22] and on the finite volume element method [28], and for various applications on the special issue [21].

We shall consider a finite volume element discretization of (1.1), in the standard conforming space of piecewise linear functions,

$$X_h = \{\chi \in C(\Omega) : \ \chi|_K \text{ is linear } \forall K \in T_h \text{ and } \chi|_{\partial\Omega} = 0\}$$

with $\{T_h\}_{0<h<1}$ a given family of triangulations of Ω with h denoting the maximum diameter of the triangles of T_h. For simplicity we shall assume that T_h is a quasi-uniform triangulation. However, this assumption is only required to show L_∞-norm error estimates. For L_2- and H^1-norm error estimations, *nondegenerate* triangulations [9, equation (4.4.16)] are sufficient.

The finite volume problem will satisfy a relation similar to (1.3) for b in a finite collection of subregions of Ω called control volumes, the number of which will be equal to the dimension of the finite element space X_h. These control volumes are constructed in the following way: Let z_K be the barycenter of $K \in T_h$. We connect z_K with line segments to the midpoints of the edges of K, thus partitioning K into three quadrilaterals K_z, $z \in Z_h(K)$, where $Z_h(K)$ are the vertices of K. Then with each vertex $z \in Z_h = \cup_{K\in T_h} Z_h(K)$ we associate a control volume (also called a box) b_z, which consists of the union of the subregions K_z, sharing the vertex z (see Figure 1.1). We denote the set of interior vertices of Z_h by Z_h^0.

The finite volume element method is then to find $u_h \in X_h$ such that

$$-\int_{\partial b_z} (A\nabla u_h)\cdot n\, ds = \int_{b_z} f\, dx, \quad \forall z \in Z_h^0. \tag{1.4}$$

Before we start our description of this work we introduce some notation. We will use the standard notation for the Sobolev spaces W_p^s and $H^s = W_2^s$ (cf. [2]). Namely, $L_p(V)$, $1 \le p < \infty$, denotes the space of p-integrable real functions over $V \subset \mathbb{R}^2$, $(\cdot,\cdot)_V$ the inner product in $L_2(V)$, $|\cdot|_{H^s(V)}$ and $\|\cdot\|_{H^s(V)}$ the seminorm and norm, respectively, in $H^s(V)$, $|\cdot|_{W_p^s(V)}$ and $\|\cdot\|_{W_p^s(V)}$ the seminorm and norm, respectively,

in $W_p^s(V)$, $p \geq 1$, and $s \in \mathbb{R}$. In addition, if $V = \Omega$ we suppress the index V, and if $p = 2$ and $s = 0$ we also suppress these indexes and denote $\|\cdot\|_{W_2^0} = \|\cdot\|$. Further, we shall denote with p' the adjoint of p, i.e., $\frac{1}{p} + \frac{1}{p'} = 1$, $p > 1$.

We begin with some comments. It is well known that in the case of a polygonal Ω, if $f \in L_p$, $1 < p < \infty$, then the solution u of (1.1) is not always in W_p^2 (cf., e.g., [24] and section 2). However, it is always in $W_{\bar{p}}^2$ or in a fractional order space H^{1+s} for some $0 < s < 1$, where s and $\bar{p}$, given in section 2, depend on both the maximal interior angle of Ω and p. In short, for p large, s and $\bar{p}$ depend on the maximal interior angle, while for p close to 1, they depend on p.

In this paper we study the influence of the corner singularities imposed by the nonconvex polygonal domain Ω and the possible insufficient regularity of the right-hand side f, say, $f \in L_p(\Omega)$, $p < 2$, or $f \in H^{-\ell}(\Omega)$, $0 \leq \ell < 1/2$, on the convergence rate of the finite volume element method. For domains with smooth boundary and convex polygonal domains, H^1- and L_2-norm error estimates were derived in [15] and [23], respectively, taking into account the regularity of f.

Note that we use the conservative version of the method, namely the right-hand side of the scheme is computed by the L_2–inner product of f with the characteristic functions of the finite volumes (or equivalently by the duality between H^ℓ and $H^{-\ell}$ for $0 \leq \ell < 1/2$). The reason for $\ell < 1/2$ is that (1.4) makes sense for at least $f \in L_1$. For results concerning finite volume schemes for problems with more singular f, i.e., $f \in H^{-1}$, we refer to [19], where an approximation of $\int_b f$ is considered.

As a model for our analysis we shall consider the corresponding Galerkin finite element method, which is to find $\underline{u}_h \in X_h$ such that

$$a(\underline{u}_h, \chi) = (f, \chi), \quad \forall \chi \in X_h, \tag{1.5}$$

with $a(\cdot,\cdot)$ the bilinear form defined by

$$a(v, w) = \int_\Omega A\nabla v \cdot \nabla w \, dx.$$

It is known that $\underline{u}_h$ satisfies (cf., e.g., [3, 8] and [9, Chapter 12])

$$\|u - \underline{u}_h\| + h^\delta \|u - \underline{u}_h\|_{H^1} \leq Ch^{s+\delta} \begin{cases} \|u\|_{H^{1+s}}, \\ \|u\|_{W_{\bar{p}}^2}, \end{cases} \quad \text{any } \delta < \pi/\omega, \tag{1.6}$$

where s is given by (2.4) or (2.6), $\bar{p}$ by (2.3), and ω denotes the biggest interior angle of Ω (cf. section 2). Note that the convergence rate of the finite element method (1.5) is optimal in the H^1-norm and suboptimal in the L_2-norm, since X_h has the following approximation properties (cf., e.g., [9, p. 285]):

$$\inf_{\chi \in X_h} (\|v - \chi\| + h\|v - \chi\|_{H^1}) \leq \begin{cases} Ch^{1+s}\|v\|_{H^{1+s}}, & \forall v \in H^{1+s} \cap H_0^1, \ 0 < s < 1, \\ Ch^{3-2/p}\|v\|_{W_p^2}, & \forall v \in W_p^2 \cap H_0^1, \ 1 \leq p \leq 2. \end{cases} \tag{1.7}$$

In the literature there are various techniques for improving the convergence rate of a finite element method in nonconvex domains, e.g., mesh refinement, augmenting the basis functions with appropriate singular functions (cf., e.g., [8, 11]). Also, recently in [18] such a method was analyzed for some finite volume element methods. Here, we are interested in the analysis of (1.4) in a mesh T_h, which does not have any prior knowledge of the singularity imposed by the domain.

TABLE 1.1

Theoretical convergence rate of the finite volume element method versus the finite element method in a nonconvex polygonal domain, when the exact solution u of problem (1.1) is in H^{1+s}, where s is defined by (2.4) or (2.6), and any $\delta < \pi/\omega$.

$p_\omega = 2/(2-\pi/\omega)$, $s_0 = 1-\pi/\omega$		H^1-norm		L_2-norm		L_∞-norm	
$\tilde{p}_\omega = 2p_\omega/(3p_\omega - 2)$		FVE	FE	FVE	FE	FVE	FE
$p_\omega < 2$	$1 < p < \tilde{p}_\omega$	s		$s+\delta$	$s+\delta$	$\approx s$	
$\tilde{p}_\omega < p_\omega$	$\tilde{p}_\omega < p < p_\omega$			$\min(1, s+\delta)$			
$f \in L_p$	$p_\omega \le p$			1			
$f \in W^t_\alpha$	$1 < \alpha < \tilde{p}_\omega$			$s+\delta$			
	$\tilde{p}_\omega \le \alpha \le 2$			$\min(s+\delta, 1+t)$			
$s_0 < 1/2$	$\ell < s_0$			$1-\ell$			
$f \in H^{-\ell}$	$s_0 < \ell < 1/2$			$1-\ell$			

In Theorems 4.3 and 5.2, we show optimal order H^1-norm error estimates for the finite volume element method (1.4), if $f \in L_p$, $p > 1$, and $f \in H^{-\ell}$, $\ell \in (0, 1/2)$. Thus, the finite element (1.5) and finite volume element method (1.4) converge with the same rate in the H^1-norm.

However, as in the convex case (cf., e.g., [13, 27]), the situation in the L_2-norm error estimate is quite different. The convergence rate in the L_2-norm of the finite volume element method (1.4) is suboptimal and lower than the corresponding finite element method. In Theorem 4.3, for $f \in L_p$, $p > 1$, we show L_2-norm error estimations where the order cannot be higher than 1. However, assuming additional regularity for f, namely, $f \in W^t_\alpha$, $t \in (0, 1]$, $\alpha \in (1, 2]$, we are able to show, in Theorem 4.6, L_2-norm error estimations that, depending on α and t, could be of the same order as the finite element method. For example, this is true for α or t sufficiently close to 1. Also, in Theorem 4.8 we derive almost optimal order L_∞-norm error estimates.

In section 5, we consider the case where $f \in H^{-\ell}$, $\ell \in (0, 1/2)$ with $A = I$ and show optimal order H^1-norm, suboptimal L_2-norm, and almost optimal L_∞-norm error estimates. In Theorem 5.2, we show again that the convergence rate of the finite volume element method (1.4) in the L_2-norm is suboptimal and lower than the corresponding suboptimal rate of the finite element method.

In Table 1.1, we summarize the theoretical results concerning the convergence rate of the finite volume element method in the H^1-, L_2-, and L_∞-norms obtained in sections 4 and 5 and compare them with the corresponding known results for the finite element method. According to (1.6) the rate of the finite element method in the H^1-norm and L_2-norm is s and $s+\delta$, respectively, for any $\delta < \pi/\omega$ and s given by either (2.4) or (2.6), depending on whether $f \in L_p$ or $f \in H^{-\ell}$. Note that if we assume that $f \in W^t_\alpha$, with $t \in (0, 1]$ and $\alpha \in (1, 2]$, both methods give the same convergence rate, if $\alpha < \tilde{p}_\omega = 2p_\omega/(3p_\omega - 2)$ with $p_\omega = 2/(2-\pi/\omega)$. Otherwise, this is determined by $\min(s+\pi/\omega, 1+t)$.

Also, in section 7 we present some numerical results for Poisson's equation on a Γ-shaped domain. The particular examples we consider justify the theoretical results of Theorems 4.3, 4.6, and 4.8. However, these do not show the lower convergence rate in the L_2-norm of Theorem 4.3, which occurs if $f \in L_p$, $p > 1$, and $f \notin W^t_\alpha$, for any $\alpha \in (1, 2]$ and $t \in (0, 1)$. To show that the L_2-norm estimates of Theorems 4.3 and 5.2 are sharp, following [27], we consider two counterexamples.

A short presentation of parts of this work can be found in [14]. For simplicity we choose not to include convection terms in the differential equation (1.1). But they

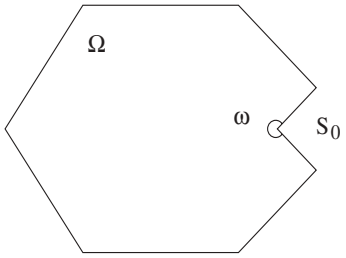


FIG. 2.1. *A nonconvex domain* Ω.

can be included provided they are bounded and the diffusion term is dominating. A brief description of this paper is the following: In section 2 we give, in short, known sharp regularity estimates for the exact solution of problems (1.1) and (2.1), based on [24, 4]. In section 3 we present the finite volume element method. In sections 4 and 5, we analyze the finite volume element method (1.4) and derive error estimates in the H^1-, L_2- and L_∞-norms. The approach follows the one developed in [13] and uses known sharp regularity results for the solutions of elliptic boundary value problems (cf. [24]). In section 6, we derive some auxiliary results, needed in proving Theorems 4.3, 4.6, 4.8, 5.2, 5.4, and 5.5. Finally in section 7, we present numerical examples that illustrate the theoretical results of section 4.

2. Preliminaries. Let us first consider the Dirichlet problem for Poisson's equation: Given $f \in L_p$, $p > 1$, find a function $u : \Omega \to \mathbb{R}^2$ such that

$$-\Delta u = f, \quad \text{in } \Omega, \quad \text{and} \quad u = 0 \quad \text{on } \partial\Omega \tag{2.1}$$

with Ω a bounded, nonconvex, polygonal domain in $\mathbb{R}^2$. For simplicity we assume that Ω has only one interior angle greater than π, namely $\omega \in (\pi, 2\pi)$ (cf. Figure 2.1). It is well known that for such domains there exists a unique solution $u \in H_0^1$ of (2.1).

The solution u could be represented in the form $u = u_S + u_R$, where $u_R \in W_p^2 \cap H_0^1$, and $u_S = cr^{\lambda_m} \frac{1}{\sqrt{\omega}\lambda_m} \sin(\lambda_m\theta)\eta(re^{i\theta})$, expressed in polar coordinates (r, θ) with respect to the vertex S_0 with angle ω (cf. [24]). Here c is a constant, $\lambda_m = \frac{m\pi}{\omega}$, $m \in \mathbb{N}$, and η is a cutoff function which is one near S_0 and zero away from S_0. A crucial role in determining the regularity of the solution u of (1.1) is played by the constant $p_\omega \equiv \frac{2}{2-\pi/\omega}$. According to [24, p. 233],

$$\text{if } f \in L_p, \ p > 1, \quad \text{then } u \in W_{\bar{p}}^2, \tag{2.2}$$

where

$$\bar{p} = \begin{cases} p, & p < p_\omega, \\ \gamma, \quad \text{any } \gamma < p_\omega, & p \geq p_\omega, \end{cases} \qquad p_\omega \equiv \frac{2}{2-\pi/\omega}. \tag{2.3}$$

Using also the imbedding $W_{\bar{p}}^2 \subset H^{1+s}$, for $s = 2 - 2/\bar{p}$ (cf., e.g., [24, p. 34]), we obtain the following:

$$\text{if } f \in L_p, \ p > 1, \quad \text{then } u \in H^{1+s}, \text{ for } s = 2 - \frac{2}{\bar{p}}. \tag{2.4}$$

Also, for problem (2.1) we have (cf., e.g., [4]),

$$\text{if } f \in H^{-\ell}, \ 0 \leq \ell \leq 1, \quad \text{then } u \in H^{1+s}, \tag{2.5}$$

where

$$s = \begin{cases} 1-\ell, & s_0 < \ell \le 1, \\ \delta, \quad \text{any } \delta < \pi/\omega, & 0 \le \ell \le s_0, \end{cases} \qquad s_0 = \frac{2}{p_0} - 1 = 1 - \frac{\pi}{\omega}. \tag{2.6}$$

For the more general problem (1.1), similar results hold. Let S be a vertex of Ω, and denote the corresponding interior angle of Ω by $\omega(S)$. Let $\mathcal{A}$ and $\mathcal{T}$ be matrices such that $\mathcal{A} = (a_{ij}(S))_{i,j=1}^2$ and $-\mathcal{T}\mathcal{A}\mathcal{T}^T = I$, and let $\omega_A(S)$ be the angle at the vertex $\mathcal{T}\mathcal{S}$ of the transformed domain $\mathcal{T}\Omega = \{\mathcal{T}x : x \in \Omega\}$. Define

$$\omega = \max_S \omega_A(S) \quad \text{and} \quad p_\omega = \frac{2}{2 - \pi/\omega}.$$

3. The finite volume element method. In order to analyze the finite volume element method (1.4) we shall need to rewrite it in a variational form resembling the one for the finite element problem (1.5) (cf., e.g., [13]). For this purpose we introduce the space

$$Y_h = \{\eta \in L_2(\Omega) : \ \eta|_{b_z} \text{ is constant}, \quad z \in Z_h^0, \ \eta|_{b_z} = 0 \text{ if } z \in \partial\Omega\}.$$

For an arbitrary $\eta \in Y_h$ we multiply the integral relation (1.4) by $\eta(z)$ and sum over all $z \in Z_h^0$. Thus we obtain the following Petrov–Galerkin formulation of the finite volume element method: Find $u_h \in X_h$ such that

$$a_h(u_h, \eta) = (f, \eta), \quad \forall \eta \in Y_h, \tag{3.1}$$

where the bilinear form $a_h(\cdot,\cdot) : X_h \times Y_h \to \mathbb{R}$ is defined by

$$a_h(v, \eta) = -\sum_{z \in Z_h^0} \eta(z) \int_{\partial b_z} (A\nabla v) \cdot n \, ds, \quad v \in X_h, \ \eta \in Y_h. \tag{3.2}$$

Further, we consider the interpolation operator $I_h : C(\Omega) \to Y_h$, defined by

$$I_h v = \sum_{z \in Z_h^0} v(z)\varphi_z, \tag{3.3}$$

where φ_z is the characteristic function of b_z. Then, we can rewrite (1.4) as

$$a_h(u_h, I_h\chi) = \sum_{z \in Z_h^0} \chi(z) \int_{b_z} f \, dx, \quad \forall \chi \in X_h. \tag{3.4}$$

Note that for every $f \in L_p$ and $\chi \in X_h$,

$$(f, I_h\chi) = \sum_{z \in Z_h} \chi(z) \int_\Omega f\varphi_z \, dx = \sum_{z \in Z_h^0} \chi(z) \int_{b_z} f \, dx. \tag{3.5}$$

Thus (3.4) can be written equivalently in the form

$$a_h(u_h, I_h\chi) = (f, I_h\chi), \quad \forall \chi \in X_h. \tag{3.6}$$

Existence of u_h follows from the fact that a_h is coercive, for h sufficiently small (cf., e.g., [13] or [28, Theorem 3.2.1]),

$$\exists\, c_0 > 0 : \quad c_0|\chi|_{H^1}^2 \le a_h(\chi, I_h\chi), \quad \forall \chi \in X_h.$$

Then this, the local stability of I_h,

$$\|I_h\chi\|_{L_p(K)} \le C\|\chi\|_{L_p(K)}, \quad \forall\chi \in X_h,\ K \in T_h,\ p > 1,$$

and the Sobolev imbedding

$$\|\chi\|_{L_p} \le C\|\chi\|_{H^1}, \quad \forall\chi \in X_h,\ p > 1,$$

give the stability of the finite volume scheme (3.6),

$$\|u_h\|_{H^1} \le C\|f\|_{L_p}, \quad p > 1. \tag{3.7}$$

Also, note that if $A(x)$ is a constant matrix over each finite element $K \in T_h$, then $a_h(\chi, I_h\psi) = a(\chi,\psi),\ \forall\chi,\psi \in X_h$ (cf., e.g., [27]). In particular, if $A = I$, we have

$$a_h(\chi, I_h\psi) = a(\chi,\psi) = \int_\Omega \nabla\chi\cdot\nabla\psi\,dx, \quad \forall\chi,\psi \in X_h \tag{3.8}$$

(cf., e.g., [6]). Thus, (3.6) takes the form

$$a(u_h,\chi) = (f, I_h\chi), \quad \forall\chi \in X_h. \tag{3.9}$$

In the case of general matrix $A(x)$, the identity (3.8) is not valid. However, following [13], we are able to rewrite a_h in a form similar to a. Indeed, we transform the left-hand side of (1.4) using integration by parts to get, for $z \in Z_h^0$ and $K \in T_h$,

$$\int_{K_z} L\chi\,dx + \int_{\partial K_z\cap\partial K} A\nabla\chi\cdot n\,ds = -\int_{\partial K_z\cap\partial b_z} A\nabla\chi\cdot n\,ds, \quad \forall\chi \in X_h.$$

Thus, multiplying by $\psi(z)$, $\psi \in X_h$, and summing over the triangles having z as a vertex and then over the vertices $z \in Z_h^0$, we obtain

$$a_h(\chi, I_h\psi) = \sum_K\{(L\chi, I_h\psi)_K + (A\nabla\chi\cdot n, I_h\psi)_{\partial K}\}, \quad \forall\chi,\psi \in X_h. \tag{3.10}$$

This is similar to

$$a(\chi,\psi) \equiv (A\nabla\chi,\nabla\psi) = \sum_K\{(L\chi,\psi)_K + (A\nabla\chi\cdot n,\psi)_{\partial K}\}, \quad \forall\chi,\psi \in X_h.$$

Due to this similarity and for convenience, in what follows we shall use (3.10) as a definition of the bilinear form a_h.

4. Nonsmooth data: L_p case. In this section we shall derive H^1-, L_2-, and L_∞-norm estimates of the error $u - u_h$ for $f \in L_p$, $p > 1$. First, we shall demonstrate that the finite element method (1.5) and finite volume element method (3.6) have the same convergence rate in the H^1-norm. The L_2-norm error estimate is quite different, and we derive two separate results. First, we will show suboptimal order L_2-norm error estimates for $f \in L_p$, $p > 1$, where the order is less than in the corresponding order for the finite element scheme (1.5). Next, assuming higher regularity for f, namely $f \in W_\alpha^t$, $t \in (0, 1]$, $\alpha \in (1, 2]$, we will show again suboptimal order L_2-norm error estimates, but now depending on α and t, these could be of the same order as the corresponding estimates of the finite element scheme. Finally, we show almost optimal L_∞-norm estimates of the error $u - u_h$.

For the analysis of the finite volume element method (3.6) we shall need to estimate the errors ε_h and ε_a defined by

$$\varepsilon_h(f,\chi) = (f,\chi) - (f, I_h\chi), \qquad \forall f \in L_p,\ \chi \in X_h,$$
$$\varepsilon_a(\chi,\psi) = a(\chi,\psi) - a_h(\chi, I_h\psi), \quad \forall \chi,\psi \in X_h.$$

In section 6 we will give the proof of the following two lemmas.

LEMMA 4.1. *There exists a constant C such that for every $\chi \in X_h$,*

$$|\varepsilon_h(f,\chi)| \le Ch\|f\|_{L_p}\,|\chi|_{W^1_{p'}}, \qquad \forall f \in L_p,\ \frac{1}{p}+\frac{1}{p'}=1, \tag{4.1}$$

$$|\varepsilon_h(f,\chi)| \le Ch^{1+t}\|f\|_{W^t_p}\,|\chi|_{W^1_{p'}}, \quad \forall f \in W^t_p,\ 0<t\le 1. \tag{4.2}$$

LEMMA 4.2. *Assume that $A \in W^2_\infty$. Then there exists a positive constant $C = C(A)$ such that*

$$|\varepsilon_a(\psi,\chi)| \le Ch|\psi|_{W^1_{\bar p}}\,|\chi|_{W^1_{\bar p'}}, \qquad \forall \chi,\psi \in X_h, \tag{4.3}$$

$$|\varepsilon_a(u_h,\chi)| \le Ch\big(\|\nabla(u-u_h)\|_{L_2} + h\|u\|_{W^2_{\bar p}}\big)|\chi|_{W^1_{\bar p'}}, \quad \forall \chi \in X_h. \tag{4.4}$$

Next, we derive H^1- and L_2-norm error estimates for the finite volume element method (1.4).

THEOREM 4.3. *Let u and u_h be the solutions of* (1.1) *and* (1.4)*, respectively, with $f \in L_p$, $p>1$. Then, there exists a constant C, independent of h, such that*

$$\|u-u_h\|_{H^1} \le C\big(h^s\|u\|_{W^2_{\bar p}} + h^{\min(1,2-2/p)}\|f\|_{L_p}\big) \le Ch^s\|f\|_{L_p}, \tag{4.5}$$

$$\|u-u_h\| \le C\big(h^{s+\delta}\|u\|_{W^2_{\bar p}} + h^{\min(1,s+\delta)}\|f\|_{L_p}\big), \quad \textit{for any } \delta < \pi/\omega, \tag{4.6}$$

with $\bar p$ and s given by (2.3) *and* (2.4)*, respectively.*

Remark 4.4. The H^1-norm error estimation (4.5) is of optimal order (cf. (1.7)). However, the L_2-norm error estimation is not of the same order as the finite element approximation (cf. (1.6)) for every p. For example, for p sufficiently close to 1, $s+\delta<1$, thus, $\|u-u_h\| = O(h^{s+\delta})$. However, for $p \ge 2$, $s = 2-2/\bar p \approx \pi/\omega$. Therefore, since $s+\delta \approx 2\pi/\omega > 1$, $\|u-u_h\| = O(h)$. The most interesting outcome of this theorem is that the convergence rate for the L_2-norm is suboptimal and lower than the rate of the finite element method (1.5). This estimate is sharp, as first demonstrated by a counterexample in [27], for convex domains. Later in section 7 we give a similar example to the one in [27], which shows the sharpness of the L_2-error estimate (4.6).

Proof. In view of (2.2), $u \in W^2_{\bar p}$, with $\bar p$ defined by (2.3). Using the triangle inequality,

$$\|u-u_h\|_{H^1} \le \|u-\chi\|_{H^1} + \|u_h-\chi\|_{H^1}, \quad \forall \chi \in X_h, \tag{4.7}$$

and the approximation properties (1.7) of X_h, it suffices to consider the last term of (4.7). The positive definiteness of A, (1.2), gives

$$\alpha_0\|u_h-\chi\|^2_{H^1} \le a(u_h-\chi, u_h-\chi), \quad \forall \chi \in X_h. \tag{4.8}$$

Thus, in view of

$$a(u_h-\chi,u_h-\chi)=a(u-u_h,u_h-\chi)+a(u-\chi,u_h-\chi)\\ \le a(u-u_h,u_h-\chi)+C\|u-\chi\|_{H^1}\|u_h-\chi\|_{H^1},\quad \forall\chi\in X_h,$$

and (4.8), we get for every $\chi\in X_h$,

$$\|u_h-\chi\|_{H^1}^2\le C|a(u-u_h,u_h-\chi)|+C\|u-\chi\|_{H^1}^2. \tag{4.9}$$

In addition, using the definitions of ε_h and ε_a, we have

$$a(u-u_h,u_h-\chi)=a(u,u_h-\chi)-a_h(u_h,I_h(u_h-\chi))-\varepsilon_a(u_h,u_h-\chi)\\ =\varepsilon_h(f,u_h-\chi)-\varepsilon_a(u_h,u_h-\chi),\quad \forall\chi\in X_h. \tag{4.10}$$

Applying then, to this relation, (4.1), (4.3), and the inverse inequality

$$|\chi|_{W^1_{p'}}\le Ch^{2/p'-1}|\chi|_{H^1},\quad p'>2,\ \forall\chi\in X_h,$$

we obtain

$$|a(u-u_h,u_h-\chi)|\le C(h^{\min(1,2-2/p)}\|f\|_{L_p}+h\|u_h\|_{H^1})\|u_h-\chi\|_{H^1}. \tag{4.11}$$

Thus, for h sufficiently small, this estimate, (3.7) and (4.9) yield

$$\|u_h-\chi\|_{H^1}\le C\|u-\chi\|_{H^1}+Ch^{\min(1,2-2/p)}\|f\|_{L_p},\quad \forall\chi\in X_h, \tag{4.12}$$

which combined with (1.7) and (4.7) gives

$$\|u-u_h\|_{H^1}\le C\big(h^s\|u\|_{W^2_{\bar p}}+h^{\min(1,2-2/p)}\|f\|_{L_p}\big).$$

Using now the fact that for $p<p_\omega$, $s=2-2/p$ and for $p\ge p_\omega$, $s<\min(1,2-2/p)$, we get

$$\|u-u_h\|_{H^1}\le Ch^s\big(\|u\|_{W^2_{\bar p}}+\|f\|_{L_p}\big). \tag{4.13}$$

Finally, employing the a priori regularity estimation of u, (2.2), we obtain the desired estimate (4.5).

We now prove (4.6) by using a duality argument. We consider the following auxiliary problem: Seek $\varphi\in H^1_0$ such that

$$L\varphi=u-u_h\quad\text{in }\Omega\quad\text{and}\quad\varphi=0\quad\text{on }\partial\Omega. \tag{4.14}$$

In view of (2.2) and the fact that $u-u_h\in L_2$, we have $\varphi\in W^2_\gamma$, where $\gamma<p_\omega$, i.e. $2/\gamma=2-\pi/\omega+\varepsilon$, with arbitrary small $\varepsilon>0$, and satisfies the a priori estimate

$$\|\varphi\|_{W^2_\gamma}\le C\|u-u_h\|,\quad \gamma<p_\omega. \tag{4.15}$$

Now let $\Pi_h: W^2_\gamma\cap H^1_0\to X_h$ denote the standard nodal interpolation operator. It is well known that Π_h has the following approximation property (cf., e.g., [17, Theorem 3.1.6] and [2, Theorem 5.4]),

$$\|\Pi_h v-v\|_{H^1}\le Ch^{\pi/\omega-\varepsilon}\|v\|_{W^2_\gamma},\quad \forall v\in W^2_\gamma\cap H^1_0, \tag{4.16}$$

and Π_h is bounded in $\|\cdot\|_{W_q^1}$ (cf., e.g., [17, Theorem 3.1.6] and [2, Theorem 5.4]),

$$\|\Pi_h v\|_{W_q^1} \le C\|v\|_{W_\gamma^2}, \quad \forall v \in W_\gamma^2 \cap H_0^1, \quad q \le p'_\gamma = 2\gamma/(2-\gamma), \tag{4.17}$$

where $p_\gamma = 2\gamma/(3\gamma-2)$.

Using (4.14) and Green's formula, we easily obtain

$$\begin{aligned} \|u-u_h\|^2 &= -(u-u_h, L\varphi) = a(u-u_h,\varphi) \\ &= a(u-u_h, \varphi-\Pi_h\varphi) + a(u-u_h,\Pi_h\varphi) := I + II. \end{aligned} \tag{4.18}$$

The first term, I, can obviously be bounded in the following way by using (4.13) and (4.16):

$$|I| \le C\|u-u_h\|_{H^1}\,\|\varphi-\Pi_h\varphi\|_{H^1} \le Ch^{s+\pi/\omega-\varepsilon}\big(\|u\|_{W_{\bar p}^2} + \|f\|_{L_p}\big)\,\|\varphi\|_{W_\gamma^2}. \tag{4.19}$$

Also, in view of (4.10), the second term, II, can be written in the form

$$II = a(u-u_h,\Pi_h\varphi) = \varepsilon_h(f,\Pi_h\varphi) - \varepsilon_a(u_h,\Pi_h\varphi). \tag{4.20}$$

Then using (4.1) and (4.4), II can be estimated by

$$|II| \le Ch\|f\|_{L_p}|\Pi_h\varphi|_{W_{p'}^1} + h\big(\|\nabla(u-u_h)\|_{L_2} + h\|u\|_{W_{\bar p}^2}\big)\,|\Pi_h\varphi|_{W_{\bar p'}^1}. \tag{4.21}$$

In order to bound $|\Pi_h\varphi|_{W_{p'}^1}$ and $|\Pi_h\varphi|_{W_{\bar p'}^1}$ in (4.21) we consider two different cases for p: (1) $p \ge p_\gamma = 2\gamma/(3\gamma-2)$, and (2) $1 < p < p_\gamma$. We can easily see that $p_\gamma < p_\omega$. Thus, in view of the definition of $\bar p$ (cf. (2.3)), for $p \ge p_\gamma$ we also have $\bar p \ge p_\gamma$ and for $1 < p < p_\gamma$, $\bar p < p_\gamma$.

Let us first consider the case $p \ge p_\gamma$. Then we have $p' \le p'_\gamma$ and $\bar p' \le p'_\gamma$ so for the respective norms of $\Pi_h\varphi$ in (4.21) we can apply the estimate (4.17). Using also (4.13) we get

$$\begin{aligned} |II| &\le C(h\|f\|_{L_p} + h^{1+s}(\|u\|_{W_{\bar p}^2} + \|f\|_{L_p}))\|\varphi\|_{W_\gamma^2} \\ &\le C(h\|f\|_{L_p} + h^{1+s}\|u\|_{W_{\bar p}^2})\|\varphi\|_{W_\gamma^2}. \end{aligned} \tag{4.22}$$

Combining now this estimation with (4.19), (4.15), and (4.18), we obtain the desired estimate (4.6), for $p \ge p_\gamma$.

In the remaining case $1 < p < p_\gamma$ we cannot directly employ (4.17) for the estimation of II. However, the inverse inequality

$$|\chi|_{W_q^1} \le Ch^{2/q-2/p'_\gamma}|\chi|_{W_{p'_\gamma}^1}, \quad q > p'_\gamma, \quad \forall \chi \in X_h,$$

and (4.17), give

$$|\Pi_h v|_{W_q^1} \le Ch^{2/q-1+\pi/\omega-\varepsilon}\|v\|_{W_\gamma^2}, \quad \forall v \in W_\gamma^2 \cap H_0^1,\ q > p'_\gamma. \tag{4.23}$$

Using now this estimation in (4.21) and the fact that for $1 < p < p_\gamma$, $2/p' = 2-2/p = s$, we get

$$\begin{aligned} |II| &\le Ch^{2/p'+\pi/\omega-\varepsilon}\big(\|f\|_{L_p} + h\|u\|_{W_{\bar p}^2}\big)\,\|\varphi\|_{W_\gamma^2} \\ &\le Ch^{s+\pi/\omega-\varepsilon}\big(\|f\|_{L_p} + h^s\|u\|_{W_{\bar p}^2}\big)\,\|\varphi\|_{W_\gamma^2}. \end{aligned} \tag{4.24}$$

Then, combining this estimation with (4.19), (4.15), and (4.18), we obtain

$$\|u-u_h\| \le Ch^{s+\delta}(\|u\|_{W^2_{\bar p}} + \|f\|_{L_p}).$$

Finally, (4.6) follows from the fact that for $1 < p < p_\gamma$, $s = 2 - \frac{2}{p} < 2 - \frac{2}{p_\gamma} = \frac{2}{\gamma} - 1 = 1 - \frac{\pi}{\omega} + \varepsilon$. ∎

Remark 4.5. For the proof of Theorems 4.3 and 4.6 it is not necessary to assume a quasi-uniform mesh T_h. This is done in order to simplify the proof, and it is only required for the validity of the inverse inequalities that are used. This assumption can be avoided by applying local inverse inequalities which hold in more general triangulations.

Next, we shall demonstrate that under some additional assumptions on the smoothness of the data the convergence rate in the L_2-norm can be improved and be equal to the rate of the corresponding finite element method.

THEOREM 4.6. *Let u and u_h be the solutions of* (1.1) *and* (1.4)*, respectively. Assume that $f \in W^t_\alpha$, $1 < \alpha \le 2$, $0 < t \le 1$, and $A \in W^2_\infty$. Then there exists a constant C, independent of h, such that*

$$\|u-u_h\| \le C(h^{s+\delta}\|f\|_{L_p} + h^{1+t+\min(0,1-2/\alpha+\delta)}\|f\|_{W^t_a}), \text{ for any } \delta < \pi/\omega, \tag{4.25}$$

with $p = 2\alpha/(2-t\alpha)$ and $\bar p$ and s given by (2.3) *and* (2.4)*, respectively.*

Remark 4.7. For $\alpha < \tilde p_\omega = \frac{2p_\omega}{3p_\omega - 2}$, we have $2/\alpha > 1 + \pi/\omega$. Thus $1+t+\min(0, 1-2/\alpha+\delta) = 2+t-2/\alpha+\delta = 2-2/p+\delta \ge s+\delta$. Therefore, $\|u-u_h\| = O(h^{s+\delta})$, i.e., in this case the L_2-norm error estimate of the finite volume element method has the same convergence rate as the corresponding finite element method. If $\alpha \ge \tilde p_\omega$, i.e., $2/\alpha \le 1+\delta$, then the order of $\|u-u_h\|$ is $\min(s+\delta, 1+t)$.

Proof. The proof will be similar to the one for (4.6). First, let us note that since $f \in W^t_\alpha$, we have by imbedding (cf. [2, Theorem 7.57]) that $f \in L_p$, with $p = 2\alpha/(2-t\alpha)$. Thus, in view of (2.2), $u \in W^2_{\bar p}$ with $\bar p$ given by (2.3).

Let again $\gamma < p_\omega$, such that $2/\gamma = 2 - \pi/\omega + \varepsilon$, with arbitrary small $\varepsilon > 0$, and let $\varphi \in W^2_\gamma \cap H^1_0$ be the solution of the auxiliary problem (4.14). Obviously, in order to show a higher order L_2-norm error estimation of $u - u_h$, we need to derive "better" bounds for I and II of (4.18). It is obvious that the estimation of I, (4.19), derived in Theorem 4.3 is of the desired order. Thus, it suffices to show a better estimate for II than the ones derived in Theorem 4.3 (cf. (4.22) and (4.24)).

Using (4.2) and (4.4) in (4.20), we get

$$|II| \le C(h^{1+t}\|f\|_{W^t_\alpha}|\Pi_h\varphi|_{W^1_{\alpha'}} + h^{1+s}\|f\|_{L_p}|\Pi_h\varphi|_{W^1_{\bar p'}}). \tag{4.26}$$

Similarly, as in Theorem 4.3 we need to derive bounds for

$$|\Pi_h\varphi|_{W^1_{\alpha'}} \quad \text{and} \quad |\Pi_h\varphi|_{W^1_{\bar p'}},$$

and we will need to consider various cases for α and p with respect to $p_\gamma = 2\gamma/(3\gamma-2)$.

Since $p = 2\alpha/(2-t\alpha)$, we can easily see that $p > \alpha$; thus we have the following three cases: (1) $p > \alpha \ge p_\gamma$, (2) $p \ge p_\gamma > \alpha$, and (3) $p_\gamma > p > \alpha$.

First, we consider the case $p > \alpha \ge p_\gamma$. For such p, according to (2.3), we have $\bar p > p_\gamma$. Thus, using (4.17) in (4.26), and the fact that $1/2 < \pi/\omega < 1$, we get

$$|II| \le C(h^{1+t}\|f\|_{W^t_\alpha} + h^{s+\pi/\omega-\varepsilon}\|f\|_{L_p})\|\varphi\|_{W^2_\gamma}. \tag{4.27}$$

Therefore, combining this estimation, (4.19), (4.15), (4.18), and the fact that if $\alpha > p_\gamma$, then $2/\alpha < 2/p_\gamma = 1 + \pi/\omega - \varepsilon$, we obtain the desired result, (4.25), in the case $p > \alpha \geq p_\gamma$.

Now let $p \geq p_\gamma > \alpha$. Again we can easily see that $\bar{p} \geq p_\gamma$. Therefore, applying (4.23) and (4.17) in (4.26) and using the fact that $1/2 < \pi/\omega < 1$, we obtain

$$\begin{aligned} |II| &\leq C(h^{t+2/\alpha'+\pi/\omega-\varepsilon}\|f\|_{W^t_\alpha} + h^{1+s}\|f\|_{L_p})\|\varphi\|_{W^2_\gamma} \\ &\leq C(h^{1+t+1-2/\alpha+\pi/\omega-\varepsilon}\|f\|_{W^t_\alpha} + h^{s+\pi/\omega-\varepsilon}\|f\|_{L_p})\|\varphi\|_{W^2_\gamma}. \end{aligned} \tag{4.28}$$

Therefore, combining this estimation, (4.19), (4.15), (4.18), and the fact that if $\alpha \leq p_\gamma$, then $2/\alpha > 1 + \pi/\omega - \varepsilon$, we obtain the desired result, (4.25), if $p > p_\gamma \geq \alpha$.

In the remaining case $p_\gamma > p > \alpha$, we have $\bar{p} = p < p_\gamma$. Thus, applying (4.23) in (4.26) and using the fact that $1/2 < \pi/\omega < 1$, we have

$$\begin{aligned} |II| &\leq C(h^{t+2/\alpha'+\pi/\omega-\varepsilon}\|f\|_{W^t_\alpha} + h^{2/p'+\pi/\omega-\varepsilon+s}\|f\|_{L_p})\|\varphi\|_{W^2_\gamma} \\ &\leq C(h^{1+t+1-2/\alpha+\pi/\omega-\varepsilon}\|f\|_{W^t_\alpha} + h^{s+\pi/\omega-\varepsilon}\|f\|_{L_p})\|\varphi\|_{W^2_\gamma}. \end{aligned} \tag{4.29}$$

Therefore, combining this estimation, (4.19), (4.15), and (4.18) we obtain the desired result, (4.25), for the remaining case $p_\gamma \geq p > \alpha$. ∎

Finally, we will show an almost optimal L_∞-norm error estimate.

THEOREM 4.8. *Let u and u_h be the solutions of* (1.1) *and* (1.4)*, respectively, with $f \in L_p$, $p > 1$, and $A \in W^2_\infty$. Then there exists a constant C, independent of h, such that*

$$\|u - u_h\|_{L_\infty} \leq Ch^s \log \frac{1}{h}\|f\|_{L_p}. \tag{4.30}$$

Proof. We split the error $u - u_h$ by adding and subtracting the Galerkin finite element approximation $\underline{u}_h$ (cf. (1.5)); thus $u - u_h = (u - \underline{u}_h) + (\underline{u}_h - u_h)$. The estimation of $\|u - \underline{u}_h\|_{L_\infty}$ is well known (cf., e.g., [32]). However, we shall briefly demonstrate it.

In view of [32, equation (0.8)] and the standard imbedding $W^2_{\bar{p}} \subset C^{0,2-2/\bar{p}}$ (cf., e.g., [24, Theorem 1.4.5.2]), we have

$$\|u - \underline{u}_h\|_{L_\infty} \leq Ch^s \log \frac{1}{h}\|u\|_{C^{0,s}} \leq Ch^s \log \frac{1}{h}\|u\|_{W^2_{\bar{p}}},$$

where $s = 2 - 2/\bar{p}$ and $C^{m,\ell}$ is the space of m times continuously differentiable functions whose mth order derivative fulfills a uniform Hölder condition of order ℓ.

Then, combining this with the elliptic regularity estimate,

$$\|u\|_{W^2_{\bar{p}}} \leq C_{\bar{p}}\|f\|_{L_p} \tag{4.31}$$

(cf. [24, Theorem 5.2.7]), we obtain

$$\|u - \underline{u}_h\|_{L_\infty} \leq C_{\bar{p}}h^s \log \frac{1}{h}\|f\|_{L_p}, \quad p > 1. \tag{4.32}$$

We turn now to the estimation of $\|\underline{u}_h - u_h\|_{L_\infty}$. Let $x_0 \in K_0 \in \mathcal{T}_h$ such that $\|\underline{u}_h - u_h\|_{L_\infty} = |(\underline{u}_h - u_h)(x_0)|$ and $\delta_{x_0} = \delta \in C_0^\infty(\bar{\Omega})$ a regularized Dirac δ-function satisfying

$$(\delta, \chi) = \chi(x_0), \quad \forall \chi \in X_h.$$

For such a function δ (cf., e.g., [9]) we have

$$\operatorname{supp}\delta \subset B = \{x \in \Omega : |x - x_0| \le h/2\}, \quad \int_\Omega \delta = 1, \quad 0 \le \delta \le Ch^{-2},$$
$$\|\delta\|_{L_p} \le Ch^{2(1-p)/p}, \quad 1 < p < \infty.$$

Also let us consider the corresponding regularized Green's function $G \in H_0^1$, defined by

$$a(G, v) = (\delta, v), \quad \forall v \in H_0^1. \tag{4.33}$$

Then, we have

$$\begin{aligned} \|\underline{u}_h - u_h\|_{L_\infty} &= (\delta, \underline{u}_h - u_h) = a(G, \underline{u}_h - u_h) = a(G_h, \underline{u}_h - u_h) \\ &= a(u - u_h, G_h) = \varepsilon_h(f, G_h) - \varepsilon_a(u_h, G_h), \end{aligned} \tag{4.34}$$

where $G_h \in X_h$ is the finite element approximation of G, i.e.,

$$a(G, \chi) = a(G_h, \chi), \quad \forall \chi \in X_h.$$

Further, using (4.1), (4.3), and the inverse inequality

$$|\chi|_{W_q^1} \le Ch^{2/q-1}|\chi|_{H^1}, \quad \forall \chi \in X_h,\ q > 2,$$

in (4.34) we obtain

$$\begin{aligned} \|\underline{u}_h - u_h\|_{L_\infty} &\le C\big\{h\big(\|\nabla(u - u_h)\|_{L_2} + h\|u\|_{W_{\bar p}^2}\big)|G_h|_{W_{\bar p'}^1} + h\|f\|_{L_p}|G_h|_{W_{p'}^1}\big\} \\ &\le C\big\{h^{2-2/\bar p}\big(\|\nabla(u - u_h)\|_{L_2} + h\|u\|_{W_{\bar p}^2}\big) \\ &\qquad + h^{\min(1,2-2/p)}\|f\|_{L_p}\big\}\|G_h\|_{H^1}, \end{aligned} \tag{4.35}$$

with $p > 1$. In addition, in view of [31, Lemma 3.1] we get

$$\|G_h\|_{H^1} \le C\|\nabla G\|_{L_2} \le C\frac{1}{(q-1)^{1/2}}\,\|\delta\|_{L_q} \tag{4.36}$$

with $q \downarrow 1$. Choosing now $q = 1 + (\log \frac{1}{h})^{-1}$ we have

$$\|G_h\|_{H^1} \le C\left(\log\frac{1}{h}\right)^{1/2}. \tag{4.37}$$

Combining now (4.34)–(4.37) and Theorem 4.3, we obtain

$$\|\underline{u}_h - u_h\|_{L_\infty} \le Ch^{2s}\left(\log\frac{1}{h}\right)^{1/2}\|u\|_{W_{\bar p}^2} + Ch^{\min(1,2-2/p)}\left(\log\frac{1}{h}\right)^{1/2}\|f\|_{L_p}. \tag{4.38}$$

From this, (4.31), and (4.32) we get the desired estimation (4.30). ∎

Remark 4.9. Assuming $f \in L_\infty$ will not improve the convergence rate in (4.30), we can easily see that in this case (4.38) does not contribute terms of order higher than 1. However, (4.32) gives terms of order almost $2-2/\bar p$, which is less than 1. Also, if we assume $f \in W_\alpha^t$, then similarly as in Theorem 4.6 we can show $\|\underline{u}_h - u_h\|_{L_\infty} = O(h^{1+t})$, but again the error $\|u - \underline{u}_h\|_{L_\infty}$ will be at most of order $2 - 2/\bar p$.

5. Nonsmooth data: $H^{-\ell}$ case. In this section we will consider problem (2.1), i.e., $A = I$, and we shall derive H^1-, L_2- and L_∞-norm estimates of the error $u - u_h$ for $f \in H^{-\ell}$, $\ell \in (0, 1/2)$. We will show optimal H^1-, suboptimal L_2-, and almost optimal L_∞-norm error estimates. The H^1- and L_∞-norm estimations are of the same order with the corresponding estimations for the finite element scheme, whereas the L_2-norm estimates are smaller.

This time for the analysis of the finite volume element method (3.6) we shall need in addition the following lemma, which we prove in section 6.

LEMMA 5.1. *There exists a constant C such that for every $\chi \in X_h$,*

$$|\varepsilon_h(f,\chi)| \le Ch^{1-\ell}\|f\|_{H^{-\ell}}\,|\chi|_{H^1}, \quad \forall f \in H^{-\ell},\ 0 < \ell < 1/2. \tag{5.1}$$

THEOREM 5.2. *Let u and u_h be the solutions of (2.1) and (1.4), respectively, with $f \in H^{-\ell}$, $0 \le \ell < 1/2$. Then there exists a constant C, independent of h, such that*

$$\|u - u_h\|_{H^1} \le C\big(h^s\|u\|_{H^{1+s}} + h^{1-\ell}\|f\|_{H^{-\ell}}\big) \le Ch^s\|f\|_{H^{-\ell}}, \tag{5.2}$$

$$\|u - u_h\| \le C\big(h^{s+\delta}\|u\|_{H^{1+s}} + h^{1-\ell}\|f\|_{H^{-\ell}}\big), \quad \textit{any } \delta < \pi/\omega. \tag{5.3}$$

Remark 5.3. The convergence rate of the H^1-norm is of optimal order (cf. (1.7)). However, since $s + \delta > 1 \ge 1 - \ell$, for δ arbitrarily close to π/ω and $\delta < \pi/\omega$, the convergence rate in the L_2-norm is suboptimal and lower than the rate of the corresponding finite element method (cf. (1.5)). Later in section 7 we give an example similar to the one in [27], which shows the sharpness of the L_2-error estimate (5.3).

Proof. The proof is similar as in Theorem 4.3, thus it suffices to estimate the first term of the right-hand side of (4.9). If $f \in H^{-\ell}$, with $0 < \ell < 1/2$, then in view of (2.5), $u \in H^{1+s}$, with s defined by (2.6). Since $A = I$, $\varepsilon_a \equiv 0$. Therefore, using (5.1) in (4.10) we obtain

$$|a(u - u_h, u_h - \chi)| \le Ch^{1-\ell}\|f\|_{H^{-\ell}}|u_h - \chi|_{H^1}, \quad \forall \chi \in X_h. \tag{5.4}$$

Then, in view of the approximation property (1.7) of X_h we get

$$\|u - u_h\|_{H^1} \le C\big(h^s\|u\|_{H^{1+s}} + h^{1-\ell}\|f\|_{H^{-\ell}}\big).$$

Using now the fact that for $s_0 < \ell < 1/2$, $s = 1 - \ell$ (cf. (2.6)), and for $0 \le \ell \le s_0$, $s < 1 - \ell$, and the a priori regularity estimate (2.5), we obtain the desired estimate (5.2).

We now turn to (5.3). Using again the same arguments as in Theorem 4.3 it suffices to estimate term II of (4.18). Let again $\gamma < p_\omega$, such that $2/\gamma = 2 - \pi/\omega + \varepsilon$, with arbitrarily small $\varepsilon > 0$, and let $\varphi \in W^2_\gamma \cap H^1_0$ be the solution of the auxiliary problem (4.14). Combining (4.10) and (5.1), we have

$$|II| \le Ch^{1-\ell}\|f\|_{H^{-\ell}}|\Pi_h\varphi|_{H^1}. \tag{5.5}$$

Finally, since, $p_\gamma < 2$, we can employ (4.17) in the estimation above, and then combining (4.18), (4.16), (5.2), and (4.15), we obtain the desired estimate (5.3). ∎

In Theorem 5.2 we demonstrated that $\|u - u_h\|_{H^1} \approx Ch^s$, for $u \in H^{1+s}$, $s < \pi/\omega$. In general, we know that $u \notin H^{1+\pi/\omega}$, even if f is smooth. In Theorem 5.4, we will show that for $f \in H^{-\ell}$, with $\ell \in (0, s_0)$, $\|u - u_h\|_{H^1} \approx C_\ell h^{\pi/\omega}$, where the constant C_ℓ blows up when $\ell \to s_0$. This is a slight improvement of the result of Theorem 5.2, which in this case gives $\|u - u_h\|_{H^1} \approx Ch^{\pi/\omega-\varepsilon}$ with $\varepsilon > 0$ arbitrarily small. Here we use the technique developed in [5].

THEOREM 5.4. *Let u and u_h be the solutions of (2.1) and (1.4), respectively, with $f \in H^{-\ell}$, $0 \le \ell < s_0$. Then there exists a constant C, independent of h, such that*

$$\|u - u_h\|_{H^1} \le C \frac{1}{s_0 - \ell} h^{\pi/\omega} \|f\|_{H^{-\ell}}. \tag{5.6}$$

Proof. Obviously, if $f \in H^{-\ell}$, with $0 \le \ell < s_0$, then $f \in H^{-\tilde{\ell}}$, $\tilde{\ell} \in (s_0, 1/2)$. Then according to Theorem 5.2, we have that

$$\|u - u_h\|_{H^1} \le C h^{1-\tilde{\ell}} \big(\|u\|_{H^{2-\tilde{\ell}}} + \|f\|_{H^{-\tilde{\ell}}} \big). \tag{5.7}$$

Also, since $u \in H^{2-\tilde{\ell}}$, we have

$$(f, v) = a(u, v) \le \|u\|_{H^{2-\tilde{\ell}}} \|v\|_{H^{\tilde{\ell}}}, \quad \forall v \in H^1_0; \tag{5.8}$$

thus, $\|f\|_{H^{-\tilde{\ell}}} \le \|u\|_{H^{2-\tilde{\ell}}}$, which in view of (5.7) gives

$$\|u - u_h\|_{H^1} \le C h^{1-\tilde{\ell}} \|u\|_{H^{2-\tilde{\ell}}}. \tag{5.9}$$

In addition, we can easily see that if $u \in H^2 \cap H^1_0$,

$$\|u - u_h\|_{H^1} \le C h \|u\|_{H^2}. \tag{5.10}$$

Then, by interpolation between (5.9) and (5.10), we get

$$\|u - u_h\|_{H^1} \le C h^{1-s_0} \|u\|_X, \tag{5.11}$$

where $X = [H^2 \cap H^1_0, H^{2-\tilde{\ell}} \cap H^1_0]_{s_0/\tilde{\ell}, \infty}$. Here $[V, W]_{\theta, q}$, $0 \le \theta \le 1$, $1 \le q \le \infty$, denote the Banach spaces intermediate between V and W defined by the K-functional, which are used in interpolation theory (cf., e.g., [7, Chapter 5]). Denote now with $L_{2,\psi}$ the orthogonal space with respect to the L_2-inner–product to the space spanned by the function $\psi = \varphi + u_R$, where $\varphi = r^{-\pi/\omega} \sin(\vartheta\pi/\omega)\eta$, and $u_R \in H^1_0$ the variational solution of $-\Delta u_R = \Delta\varphi$. Then, in view of [5, Theorem 4.1], we have

$$\|u\|_X \le C \|f\|_Y \tag{5.12}$$

with $Y = [L_{2,\psi}, H^{-\tilde{\ell}}]_{s_0/\tilde{\ell}, \infty}$; thus

$$\|u - u_h\|_{H^1} \le C h^{1-s_0} \|f\|_Y. \tag{5.13}$$

Further, since $\tilde{\ell} > s_0$, $[L_{2,\psi}, H^{-1}]_{\tilde{\ell}, 2} = [L_2, H^{-1}]_{\tilde{\ell}, 2} = H^{-\tilde{\ell}}$ (cf., e.g., [5, equation (3.16)]). Therefore, in view of the reiteration theorem for the interpolation of spaces (cf., e.g., [7, Chapter 5]), we get

$$Y = [L_{2,\psi}, H^{-1}]_{s_0, \infty}. \tag{5.14}$$

In addition, in view of [5, Theorem 3.1 and Remark 3.1], we have

$$\|f\|_{[L_{2,\psi}, H^{-1}]_{s_0,\infty}} \le C \frac{1}{s_0 - \ell} \|f\|_{H^{-\ell}}, \quad \forall f \in H^{-\ell}. \tag{5.15}$$

Thus, combining (5.13)–(5.15) we get the desired estimate. ∎

Finally, we will show an almost optimal L_∞-norm error estimate.

THEOREM 5.5. *Let u and u_h be the solutions of (2.1) and (1.4), respectively, with $f \in H^{-\ell}$, $0 < \ell < 1/2$. Then there exists a constant C, independent of h, such that*

$$\|u - u_h\|_{L_\infty} \le Ch^s \log \frac{1}{h} \|f\|_{H^{-\ell}}. \tag{5.16}$$

Proof. The proof is similar to the one for Theorem 4.8. Hence, we will derive bounds for $\|u - \underline{u}_h\|_{L_\infty}$ and $\|\underline{u}_h - u_h\|_{L_\infty}$.

This time using [32, equation (0.8)] and the standard imbedding $H^{1+s} \subset C^{0,s}$ (cf., e.g., [24, Theorem 1.4.5.2]), we have

$$\|u - \underline{u}_h\|_{L_\infty} \le Ch^s \log \frac{1}{h} \|u\|_{C^{0,s}} \le Ch^s \log \frac{1}{h} \|u\|_{H^{1+s}}.$$

Then, combining this with the elliptic regularity estimate,

$$\|u\|_{H^{1+s}} \le C_\ell \|f\|_{H^{-\ell}}$$

(cf. [4]), we obtain

$$\|u - \underline{u}_h\|_{L_\infty} \le C_\ell h^s \log \frac{1}{h} \|f\|_{H^{-\ell}}, \quad 0 \le \ell < 1/2. \tag{5.17}$$

We turn now to the estimation of $\|\underline{u}_h - u_h\|_{L_\infty}$. Since $A = I$, (4.34) gives

$$\|\underline{u}_h - u_h\|_{L_\infty} = \varepsilon_h(f, G_h), \tag{5.18}$$

where $G_h \in X_h$ is the finite element approximation of the regularized Green function G (cf. (4.33)). Then, using Lemma 5.1 and (4.37), we obtain

$$\|\underline{u}_h - u_h\|_{L_\infty} \le Ch^{1-\ell} \left(\log \frac{1}{h} \right)^{1/2} \|f\|_{H^{-\ell}}.$$

From this and (5.17) we get the desired estimation (5.16). ∎

6. Auxiliary results. In this section we shall prove Lemmas 4.1, 4.2, and 5.1 of the previous sections.

Proof of Lemma 4.1. We can easily see that the interpolation operator I_h satisfies the property

$$\begin{aligned} \|\chi - I_h\chi\|^q_{L_q(K)} &= \sum_{z \in Z_h(K} \int_{K_z} (\chi - \chi(z))^q \, dx \\ &\le h_K^q |\chi|^q_{W^1_q(K)}, \quad \forall \chi \in X_h, \ q > 1, \end{aligned} \tag{6.1}$$

with $Z_h(K)$ the set of the vertices of K. Also, since in the construction of the control volumes we choose z_K to be the barycenter of K, we have

$$\int_K \chi \, dx = \int_K I_h\chi \, dx, \quad \forall K \in T_h, \ \forall \chi \in X_h. \tag{6.2}$$

In view of (6.1), (4.1) follows easily. Let now $\bar{f}_K$ be the mean value of f in K. Thus,

$$\|f - \bar{f}_K\|_{L_p(K)} \le Ch_K |f|_{W^1_p(K)}, \quad \forall f \in W^1_p(K), \ p > 1. \tag{6.3}$$

Then, by interpolation of this estimate and $\|f-\bar f_K\|_{L_p(K)} \le C\|f\|_{L_p(K)}$, we get, for $f \in W_p^t(K)$, $p>1$, and $0<t\le 1$

$$\|f-\bar f_K\|_{L_p(K)} \le Ch_K^t |f|_{W_p^t(K)}. \tag{6.4}$$

Since $\bar f_K$ is constant over K, due to (6.2), we have

$$(f, \chi - I_h\chi)_K = (f-\bar f_K, \chi - I_h\chi)_K, \quad \forall \chi \in X_h.$$

Thus, due to this, (6.4), and (6.1), we get for every $\chi \in X_h$,

$$|(f, \chi - I_h\chi)_K| = |(f-\bar f_K, \chi - I_h\chi)_K| \le Ch_K^{1+t}|f|_{W_p^t(K)}\,|\chi|_{W_{p'}^1(K)},$$

which concludes the proof of (4.2). ∎

We now turn to the proof of Lemma 4.2. For this we shall need the following auxiliary result.

LEMMA 6.1. *Let K be a triangle and e a side of K. Then for $\varphi \in W_p^1(K)$, $p>1$, there exists a constant C independent of K such that*

$$\left|\int_e \varphi(\chi - I_h\chi)\,ds\right| \le Ch|\varphi|_{W_p^1(K)}|\chi|_{W_{p'}^1(K)}, \quad \forall \chi \in \mathbb{P}_1(K).$$

Proof of Lemma 6.1. It is obvious that, for c constant, $I_h c = c$ and

$$\int_e I_h\chi\,ds = \int_e \chi\,ds, \quad \forall \chi \in X_h,\ \forall e \in E_h. \tag{6.5}$$

Thus, we have for every $\chi \in \mathbb{P}_1(K)$ and $\varphi \in L_2(e)$,

$$\int_e \varphi(\chi - I_h\chi)\,ds = \int_e (\varphi - c_1)(\chi - c_2 - I_h(\chi - c_2))\,ds,$$

for all constants $c_1, c_2 \in \mathbb{R}$, $K \in T_h$, and $e \in E_h(K)$. Using now in the relation above the fact that $\|I_h\chi\|_{L_\infty(e)} \le \|\chi\|_{L_\infty(e)}$ and a local inverse inequality, we get for all constants $c_1, c_2 \in \mathbb{R}$, $\chi \in \mathbb{P}_1(K)$, and $\varphi \in W_p^1(K)$,

$$\begin{aligned}\left|\int_e \varphi(\chi - I_h\chi)\,ds\right| &\le \|\varphi - c_1\|_{L_p(e)}\|\chi - c_2 - I_h(\chi - c_2)\|_{L_{p'}(e)}\\ &\le h_e^{1/p'}\|\varphi - c_1\|_{L_p(e)}\,\|\chi - c_2 - I_h(\chi - c_2)\|_{L_\infty(e)}\\ &\le Ch_e^{1/p'}\|\varphi - c_1\|_{L_p(e)}\|\chi - c_2\|_{L_\infty(e)}\\ &\le C\|\varphi - c_1\|_{L_p(e)}\,\|\chi - c_2\|_{L_{p'}(e)}\end{aligned} \tag{6.6}$$

with $h_e = |e|$. In view of the Bramble–Hilbert lemma and a standard homogeneity argument, we can easily show

$$\inf_{c\in\mathbb{R}} \|\varphi - c\|_{L_p(e)} \le Ch_e^{1-1/p}|\varphi|_{W_p^1(K)}, \quad \forall \varphi \in W_p^1(K),\ p>1.$$

Finally, combining this with (6.6) we obtain the desired estimate. ∎

We now turn to the proof of Lemma 4.2.

Proof of Lemma 4.2. First we will show (4.3). In view of Green's formula, we have

$$\varepsilon_a(\psi,\chi)=\sum_K (L\psi,\chi-I_h\chi)_K+\sum_K (A\nabla\psi\cdot n,\chi-I_h\chi)_{\partial K}=I+II. \tag{6.7}$$

For the first term we have from (6.1),

$$|I|\le C\sum_K \|L\psi\|_{L_p(K)}\,\|\chi-I_h\chi\|_{L_{p'}(K)}\le C\sum_K h_K|\psi|_{W^1_{\bar p}(K)}\,|\chi|_{W^1_{\bar p'}(K)}.$$

The bound for II follows at once from Lemma 6.1 since $|A\nabla\psi\cdot n|_{W^1_{\bar p}(K)}\le C|\psi|_{W^1_{\bar p}(K)}$.

We now turn to (4.4). Let $\psi=u_h$ in (6.7) and $(\nabla A)_K$ be the average over K. Then in view of (6.1)–(6.3) we have for every $\chi\in X_h$,

$$(Lu_h,\chi-I_h\chi)_K=([\nabla A-(\nabla A)_K]\nabla u_h,\chi-I_h\chi))_K\le Ch_K^2|u_h|_{W^1_{\bar p}(K)}\,|\chi|_{W^1_{\bar p'}(K)}$$

with $\bar p$ given by (2.3). From the estimation above we easily obtain the desired bound for I. Let now $E_h(K)$ be the set of edges of $K\in T_h$ and $\bar A_e=A(m_e)$, where m_e is the midpoint of the edge e. We will show that for every $\chi\in X_h$,

$$II=\sum_K\sum_{e\in E_h(K)}((A-\bar A_e)\nabla(u_h-u)\cdot n,\chi-I_h\chi)_e. \tag{6.8}$$

Provided that this holds, we may apply Lemma 6.1 and the estimate

$$|(A-\bar A_e)\nabla(u_h-u)|_{W^1_{\bar p}(K)}\le C\big(\|\nabla(u-u_h)\|_{L_2(K)}+h\|u\|_{W^2_{\bar p}(K)}\big) \tag{6.9}$$

to obtain

$$|II|\le Ch\big(\|\nabla(u-u_h)\|_{L_2}+h\|u\|_{W^2_{\bar p}}\big)\,|\chi|_{W^1_{\bar p'}},\quad \forall\chi\in X_h,$$

which gives the desired estimate for II. Therefore, it remains to prove (6.8). We will show, for every $\psi\in X_h$,

$$\sum_K(A\nabla u\cdot n,\psi-I_h\psi)_{\partial K}=\sum_K\sum_{e\in E_h(K)}(\bar A_e\nabla u\cdot n,\psi-I_h\psi)_e=0. \tag{6.10}$$

In the first sum we have by Green's formula for every $\psi\in X_h$,

$$\sum_K(A\nabla u\cdot n,\psi)_{\partial K}=\sum_K(A\nabla u,\nabla\psi)_K-(Lu,\psi)_K=(A\nabla u,\nabla\psi)-(Lu,\psi)=0.$$

In addition, $\sum_K(A\nabla u\cdot n,I_h\psi)_{\partial K}=0$ because $I_h\psi$ is piecewise constant on each interior edge e and $A\nabla u\cdot n$ is continuous across e (in the trace sense), and $I_h\psi=0$ on $\partial\Omega$. Since the first sum in (6.10) vanishes for each smooth A and is continuous in A on $L_1(\cup\partial K)$, the second sum is the limit of sums with a smooth A and, therefore, also vanishes. Finally, since $\bar A_e\nabla u_h\cdot n$ is constant on each e, in view of (6.5) we have

$$\sum_K\sum_{e\in E_h(K)}(\bar A_e\nabla u_h\cdot n,\chi-I_h\chi)_e=0,\quad\forall\chi\in X_h.\qquad\square$$

It remains now to prove Lemma 5.1.

Proof of Lemma 5.1. In view of the definition of ε_h, it suffices to show

$$|\chi - I_h\chi|_{H^\ell} \le Ch^{1-\ell}\|\nabla\chi\|_{L_2}, \quad 0 < \ell < 1/2.$$

The fractional order seminorm $|\cdot|_{H^\ell}$ is given by

$$|w|^2_{H^\ell} = \int_\Omega\int_\Omega \frac{|w(x)-w(y)|^2}{|x-y|^{2(1+\ell)}}\,dy\,dx; \tag{6.11}$$

therefore,

$$\begin{aligned}|\chi - I_h\chi|^2_{H^\ell} &= \sum_{z,w\in Z_h}\int_{b_z}\int_{b_w}\frac{|(\chi-I_h\chi)(x)-(\chi-I_h\chi)(y)|^2}{|x-y|^{2(1+\ell)}}\,dy\,dx\\ &\le 4\sum_{\substack{z,w\in Z_h\\ z\neq w}}\int_{b_z}\int_{b_w}|\nabla\chi(z)|^2\frac{|x-z|^2}{|x-y|^{2(1+\ell)}}\,dy\,dx\\ &\quad+\sum_{z\in Z_h}\int_{b_z}\int_{b_z}|\nabla\chi(z)|^2|x-y|^{-2+2(1-\ell)}\,dy\,dx = 4I + II.\end{aligned}$$

For the estimation of II we rewrite the integral with respect to the y variable in polar coordinates (r,θ) having as center x; thus $|x-y| = r$ and

$$\int_{b_z}|x-y|^{-2+2(1-\ell)}\,dy \le C\int_0^h r^{(1-\ell)p-2+1}\,dr = Ch^{2(1-\ell)}.$$

Therefore,

$$\int_{b_z}\int_{b_z}|\nabla\chi|^2|x-y|^{-2+2(1-\ell)}\,dy\,dx \le Ch^{2(1-\ell)}\|\nabla\chi\|^2_{L_2(b_z)}, \tag{6.12}$$

which gives the desired estimate for II. Let us consider now $z \neq w$ and fix temporarily an $x \in K_z$. Using again polar coordinates with center x we estimate the integral with respect to y,

$$\int_{b_w}|x-y|^{-2(1+\ell)}\,dy \le C\int_{r_0(x)}^\infty r^{1-2(1+\ell)}\,dr \le Cr_0^{-2\ell}(x) = III,$$

where $r_0(x) = \operatorname{dist}(x, b_w)$. Let us assume that vertices z and w are in a different triangle and $|r_0(x)| > kh$; therefore, $|III| \le C|r_0(x)|^{-2\ell} \le Ch^{-2\ell}$. Thus,

$$\int_{b_z}\int_{b_w}|\nabla\chi(z)|^2|x-z|^2|x-y|^{-2(1+\ell)}\,dy\,dx \le Ch^{2(1-\ell)}\|\nabla\chi\|^2_{L_2(b_z)}. \tag{6.13}$$

Finally, let us consider the case that $z \neq w$ and are vertices of the same triangle K. Then $r_0(x)$ could be arbitrarily small and in order for

$$\int_{b_z} r_0(x)^{-2\ell}(x)\,dx < +\infty,$$

we need to assume that $\ell < 1/2$. In a such case, we have

$$\int_{b_z}\int_{b_w}|\nabla\chi(z)|^2|x-z|^2|x-y|^{-2(1+\ell)}\,dy\,dx \le Ch^2\int_{b_z}|\nabla\chi(z)|^2r_0^{2\ell}(x)\,dx. \tag{6.14}$$

Next we will estimate the right-hand side of the relation above. For this, it suffices to bound $\int_{K_z} |\nabla\chi(z)|^2 r_0^{2\ell}$. Let us denote with x_1 and x_2 the two coefficients of a point x in K_z and introduce a rotation and translation of the (x_1, x_2)-coordinate system to $(\tilde{x}_1, \tilde{x}_2)$, where $\tilde{x}_1$-axis is the common edge $K_z \cap K_w$. We can easily see that for any point in $x \in K_z$,

$$r_0(x) = \operatorname{dist}(x, b_w) \geq \operatorname{dist}((\tilde{x}_1, 0), b_w) = \tilde{x}_1.$$

Therefore, $r_0^{-2\ell}(x) \leq \tilde{x}_1^{-2\ell}$, $\forall x \in K_z$. Then

$$\begin{aligned} h^2 \int_{K_z} |\nabla\chi(z)|^2 r_0^{-2\ell}(x)\, dx &\leq Ch^2 |\nabla\chi(z)|^2 \int_0^h \int_0^h \tilde{x}_1^{-2\ell}\, d\tilde{x}_1\, d\tilde{x}_2 \\ &\leq Ch^{2(1-\ell)} \|\nabla\chi\|^2_{L_2(K_z)}, \end{aligned}$$

assuming $\ell < 1/2$. Hence, the relation above and (6.14) give

$$\int_{b_z} \int_{b_w} |\nabla\chi(z)|^2 |x - z|^2 |x - y|^{-2(1+\ell)}\, dy\, dx \leq Ch^{2(1-\ell)} \|\nabla\chi\|^2_{L_2(b_z)}.$$

Combining this with (6.12) and (6.13), we obtain the desired estimate. ∎

7. Numerical results. In this section we will illustrate on several numerical examples the theoretical results of section 4. Our examples are similar to the ones considered in [8, 27].

First, we will show that the theoretical L_2-norm convergence rate of Theorem 4.6 is satisfied for the model Dirichlet boundary value problem for the Poisson equations in a Γ-shaped domain (cf. Figure 7.1), with vertices $(0,0)$, $(1,0)$, $(1,1)$, $(-1,1)$, $(-1,-1)$, and $(0,-1)$. As in [8], we consider the following two singular functions for this Γ-shaped domain:

$$S_1(r,\theta) = \phi(r) r^{2/3} \sin\left(\frac{2}{3}\theta\right), \qquad S_2(r,\theta) = \phi(r) r^{\beta} \sin\left(\frac{2}{3}\theta\right),$$

where $\beta \in (0,1)$ and ϕ is a cutoff function defined by

$$\phi(r) = \begin{cases} 1 & 0 \leq r \leq 1/4, \\ -192r^5 + 480r^4 - 440r^3 + 180r^2 - \frac{135}{4}r + \frac{27}{8}, & 1/4 \leq r \leq 3/4, \\ 0 & 3/4 \leq r. \end{cases}$$

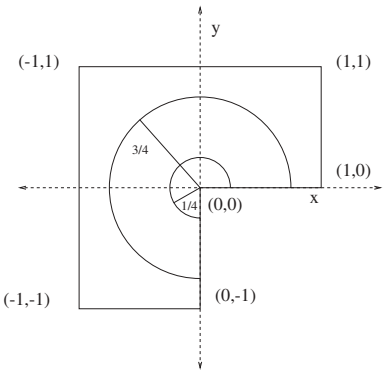


FIG. 7.1. *A Γ-shaped domain.*

TABLE 7.1
Approximate theoretical convergence rate for exact solution $u = S_1 + S_2 + (x - x^3)(y^2 - y^4)$.

$p_\omega = 3/2, \quad \tilde{p}_\omega = 6/5$		β			
		1/3	1/2	2/3	3/4
f $(-\Delta u)$ almost in W^t_α		$W^{10/66}_{11/10}$	$W^{1/6}_{6/5}$	$W^{4/3}_{6/5}$	$W^{5/12}_{6/5}$
rate in H^1-norm $= s$		1/3	1/2	2/3	2/3
rate in L_2–norm $\approx$	$s + 2/3, \ (\alpha < \tilde{p}_\omega)$	$s + \frac{2}{3} = 1$			
	$\min(s + 2/3, 1 + t), (\alpha \geq \tilde{p}_\omega)$		$s + \frac{2}{3} = \frac{7}{6}$	$s + \frac{2}{3} = \frac{4}{3}$	$s + \frac{2}{3} = \frac{4}{3}$
rate in L_∞-norm $\approx s$		1/3	1/2	2/3	2/3

For $f = -\Delta(S_1 + S_2) + 6x(y^2 - y^4) + (x - x^3)(12y^2 - 2)$ the exact solution is $u = S_1 + S_2 + (x - x^3)(y^2 - y^4)$. We can easily see that

$$\Delta(S_1 + S_2) = \phi(r)(\beta^2 - (2/3)^2)r^{\beta-2}\sin\left(\frac{2}{3}\theta\right) + (2\beta + 1)\phi'(r)r^{\beta-1}\sin\left(\frac{2}{3}\theta\right) + \phi''(r)r^{\beta}\sin\left(\frac{2}{3}\theta\right) + \frac{7}{3}\phi'(r)r^{-1/3}\sin\left(\frac{2}{3}\theta\right) + \phi''(r)r^{2/3}\sin\left(\frac{2}{3}\theta\right).$$

Since ϕ is a smooth cutoff function and $6x(y^2-y^4)+(x-x^3)(12y^2-2)$ is a polynomial, the nonsmoothness of f results from $-\Delta(S_1 + S_2)$ and for $\beta \in (0, 1)$ this is dictated from the term $r^{\beta-2}$, except in the case $\beta = 2/3$, where the leading term is $r^{-1/3}$.

According to [24, Theorem 1.4.5.3], if a function g can be written as $g = r^\gamma \varphi(\vartheta)$, in polar coordinates, where φ is smooth function, then $g \in W^t_\alpha$, with $t > 0$ and $\alpha > 1$, for $\gamma > t - 2/\alpha$. Thus, applying this to f, we have that f is almost in $W^{\beta-2+2/\alpha}_\alpha$, with $\alpha \in (1, 2/(2-\beta))$, for $\beta \neq 2/3$, and $f \in W^{-1/3+2/\alpha}_\alpha$, with $\alpha \in (1, 6)$, for $\beta = 2/3$. In addition, in view of the imbedding $L_p \subset W^t_a$, with $p = 2\alpha/(2 - t\alpha)$, for $\beta \neq 2/3$, then $f \in L_p$, with $p = 2/(2 - \beta)$, and for $\beta = 2/3$, $f \in L_6$.

Since we have considered a Γ-shaped domain, the largest interior angle is $3\pi/2$; therefore, $p_\omega = 2/(2 - (\pi/\frac{3\pi}{2})) = 3/2$. Thus, in view of (2.2), the solution u of the Poisson problem is almost in $W^2_{\bar{p}}$, with $\bar{p} = \min(2/(2 - \beta), 3/2)$, or else u is almost in H^{1+s} with $s = \min(\beta, 2/3)$.

For example, we consider $\beta = 1/3$, $1/2$, $2/3$, and $3/4$. Then f is almost in the Sobolev spaces $W^{10/66}_{11/10}$, $W^{1/6}_{6/5}$, $W^{4/3}_{6/5}$, and $W^{5/12}_{6/5}$ for $\beta = 1/3$, $1/2$, $2/3$, and $3/4$, respectively. In Table 7.1 we present the theoretical and in Tables 7.2 and 7.3 the computed rates of convergence of the finite volume element method which illustrate the results of Theorem 4.6. The computation is done in the following way: For a given triangulation with number of nodes N and stepsize $2h$, we compute the finite volume solution and the norms of the errors $\|u - u_{2h}\|_T$, where $T = H^1, L_2, L_\infty$. Then we split each triangle into four similar triangles and compute the solution u_h and the corresponding norms of the errors, $\|u - u_h\|_T$. Then the computed rates are given by $\log_2 \frac{\|u-u_{2h}\|_T}{\|u-u_h\|_T}$. This procedure is repeated up to seven levels of refinement. The integrals in the finite volume formulation were approximated with a 13-point Gaussian quadrature. For the solution of the corresponding linear system we used a multigrid preconditioner.

One may argue that the suboptimal order of the L_2-norm error estimates in Theorems 4.3 and 5.4 of the finite volume element method might be an artifact of the proof and expect the same rate as in the finite element method. However, this is not correct. In what follows, we consider a counterexample which is based on a

TABLE 7.2
Experimental convergence rate for $\beta = \frac{1}{3}$ and $\beta = \frac{1}{2}$.

# of nodes	$\beta = 1/3$			$\beta = 1/2$		
	H^1	L_2	L_∞	H^1	L_2	L_∞
225	0.75	1.26	0.28	0.86	1.54	0.48
833	0.68	1.14	0.38	0.83	1.39	0.54
3201	0.59	1.09	0.38	0.81	1.32	0.54
12545	0.49	1.06	0.37	0.75	1.26	0.54
49665	0.42	1.04	0.36	0.69	1.23	0.54
197633	0.38	1.03	0.36	0.63	1.22	0.53
788481	0.37	1.02	0.35	0.60	1.21	0.53
Theoretical $\approx$	0.33	1	0.33	0.5	1.17	0.5

TABLE 7.3
Experimental convergence rate for $\beta = \frac{2}{3}$ and $\beta = \frac{3}{4}$.

# of nodes	$\beta = 2/3$			$\beta = 3/4$		
	H^1	L_2	L_∞	H^1	L_2	L_∞
225	0.89	1.76	0.92	0.90	1.84	1.15
833	0.89	1.60	0.66	0.91	1.66	0.69
3201	0.91	1.55	0.67	0.93	1.63	0.70
12545	0.90	1.46	0.67	0.93	1.54	0.69
49665	0.87	1.41	0.67	0.91	1.47	0.69
197633	0.83	1.37	0.67	0.88	1.42	0.69
788481	0.79	1.36	0.67	0.85	1.40	0.69
Theoretical $\approx$	0.66	1.33	0.66	0.66	1.33	0.66

similar argument given in [27]. The following arguments can easily be modified and apply to a model problem in a convex domain. This can then be used to illustrate the theoretical convergence rates derived in [14, 23].

First we will show that the L_2-norm estimate in Theorem 4.3 is sharp. We consider the model problem

$$-\Delta u = f \quad \text{in } \Omega, \quad \text{and} \quad u = 0 \quad \text{on } \partial\Omega, \tag{7.1}$$

where $f \in L_2$ and Ω is the Γ-shaped domain with vertices $(0,0)$, $(2,0)$, $(2,2)$, $(-2,2)$, $(-2,-2)$, and $(0,-2)$. Since $\pi/\omega = 2/3$, according to Theorem 4.3 we know that

$$\|u - u_h\|_{H^1} \le Ch^s \|f\|_{L_2}$$

with $s = 2/3 - \varepsilon$, $\varepsilon > 0$ arbitrarily small. Let us assume then that (4.6) is not true and the finite volume and finite element methods converge in the L_2-norm with the same rate, i.e.,

$$\|u - u_h\|_{L_2} \le Ch^{2s} \|f\|_{L_2}.$$

Obviously,

$$\|u - u_h\|_{L_2} = \sup_{\phi \in L_2 \setminus \{0\}} \frac{(u - u_h, \phi)}{\|\phi\|_{L_2}}.$$

Hence, our assumption leads to

$$|(u - u_h, \phi)| \le Ch^{2s} \|\phi\|_{L_2} \|f\|_{L_2}. \tag{7.2}$$

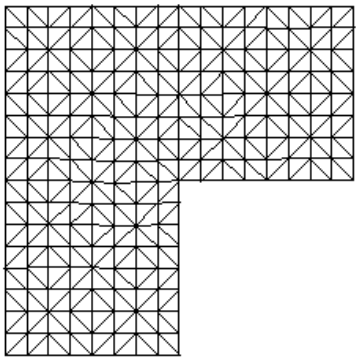


FIG. 7.2. *An example of a triangulation. The successive uniform refinement occurs by splitting the triangles into four.*

Next, let us denote $\psi \in H^{1+s} \cap H_0^1$ the solution of the auxiliary problem

$$-\Delta\psi = \phi \quad \text{in } \Omega, \quad \text{and} \quad \psi = 0 \quad \text{on } \partial\Omega. \tag{7.3}$$

Thus,

$$(u - u_h, \phi) = a(u - u_h, \psi) = a(u - u_h, \psi - \Pi_h\psi) + a(u - u_h, \Pi_h\psi), \tag{7.4}$$

where $\Pi_h\psi$ is the interpolant of ψ in X_h. Obviously, then

$$a(u - u_h, \Pi_h\psi) = (f, \Pi_h\psi - I_h\Pi_h\psi). \tag{7.5}$$

We can easily see that

$$a(u - u_h, \psi - \Pi_h\psi) \le Ch^{2s}\|\phi\|_{L_2}\|f\|_{L_2}.$$

Thus combining (7.2)–(7.5), we get

$$(f, \Pi_h\psi - I_h\Pi_h\psi) \le Ch^{2s}\|\phi\|_{L_2}\|f\|_{L_2}.$$

Since f is an arbitrary function of L_2, this leads to

$$\|\Pi_h\psi - I_h\Pi_h\psi\|_{L_2} \le Ch^{2s}\|\phi\|_{L_2}.$$

Hence,

$$\|\psi - I_h\Pi_h\psi\|_{L_2} \le Ch^{2s}\|\phi\|_{L_2}.$$

Then, since ϕ is also an arbitrary function, this should be true for any function $\psi \in H^{1+s} \cap H_0^1$. Therefore, let us consider a function $\psi \in H^{1+s} \cap H_0^1$ such that

$$\psi(x_1, x_2) = x_1(1 - x_1), \quad (x_1, x_2) \in \Omega_1 = [1/2, 3/2] \times [1/2, 3/2]. \tag{7.6}$$

For this ψ we should get

$$\|\psi - I_h\Pi_h\psi\|_{L_2(\Omega_1)} \le Ch^{2s}. \tag{7.7}$$

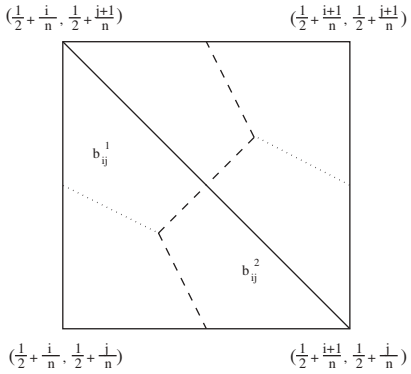


FIG. 7.3. *A sample square K_{ij}. The two regions b_{ij}^1 and b_{ij}^2 are separated with the dashed line.*

We discretize Ω_1 into n^2 equal size squares with length $h = 1/n$, and each square is divided further into two right triangles in the same direction. Next, we construct the relative control volumes by connecting the barycenter of its triangle with the middle of the edges. Let us denote z_{ij} the vertices $(1/2 + i/n, 1/2 + j/n)$, $i, j = 0, \ldots, n-1$. Also, let K_{ij} be the square $[1/2 + i/n, 1/2 + (i+1)/n] \times [1/2 + j/n, 1/2 + (j+1)/n]$, $i, j = 1, \ldots, n$, and $b_{ij}^1 = K_{ij} \cap (b_{ij} \cup b_{i(j+1)})$ and $b_{ij}^2 = K_{ij} \cap (b_{(i+1)j} \cup b_{(i+1)(j+1)})$ (cf. Figure 7.3). Then, since ψ depends only on x, $I_h\psi_I$ has the same value on the control volumes b_{ij}, $j = 0, \ldots, n-1$, for every $i = 0, \ldots, n-1$. For this reason, on the square K_{ij}, $I_h\psi = \psi(z_{ij})$ on b_{ij}^1 and $I_h\psi = \psi(z_{(i+1)j})$ on b_{ij}^2. Then we have

$$\begin{aligned}
\|\psi - I_h\Pi_h\psi\|_{L_2(\Omega_1)}^2 &= \int_{\Omega_1} \psi^2 \, dx_1 \, dx_2 + \int_{\Omega_1} (I_h\Pi_h\psi)^2 \, dx_1 \, dx_2 \\
&\quad - 2\int_{\Omega_1} \psi I_h\Pi_h\psi \, dx_1 \, dx_2 \\
&= \int_{1/2}^{3/2} x_1^2(1-x_1)^2 \, dx_1 + \sum_{i,j=0}^{n-1} (\psi^2(z_{ij})|b_{ij}^1| + \psi^2(z_{(i+1)j})|b_{ij}^2|) \\
&\quad - 2\sum_{i,j=0}^{n-1} \left(\psi(z_{ij}) \int_{b_{ij}^1} x_1(1-x_1) \, dx_1 \, dx_2 \right. \\
&\qquad \left. + \psi(z_{(i+1)j}) \int_{b_{ij}^2} x_1(1-x_1) \, dx_1 \, dx_2 \right) \\
&= \frac{10}{81n^2} + \frac{1}{405n^4}.
\end{aligned}$$

Finally, we have

$$\|\psi - I_h\Pi_h\psi\|_{L_2(\Omega_1)} = \frac{\sqrt{10}}{9}\frac{1}{n} + o\left(\frac{1}{n}\right) = O(h).$$

Combining this with (7.7) we get a contradiction, since $2s \approx 4/3$.

Similar arguments can be used in order to show now the sharpness of the L_2-norm error estimate in Theorem 5.2. Thus, let us consider this time the model problem (7.1), with $f \in H^{-1/3}$.

According to Theorem 5.2 we have that

$$\|u-u_h\|_{H^1} \le Ch^s \|f\|_{H^{-1/3}},$$

with $s = 2/3-\varepsilon$, $\varepsilon > 0$, arbitrarily small. Let us assume that the L_2-norm error estimate in Theorem 5.4 does not hold and the finite volume and finite element methods converge in L_2-norm with the same rate, i.e.,

$$\|u-u_h\|_{L_2} \le Ch^{2s} \|f\|_{H^{-1/3}}.$$

Repeating similar arguments as in the previous counterexample, the function $\psi \in H^{1+s} \cap H_0^1$ that satisfies (7.6) and (7.3) should also satisfy

$$\|\Pi_h\psi - I_h\Pi_h\psi\|_{H^{1/3}} \le Ch^{2s}\|\phi\|_{L_2}. \tag{7.8}$$

We discretize again Ω_1 in the same way as before, into n^2 equal size squares with length $h = 1/n$, and each square is divided further into two right triangles in the same direction. We construct the control volumes b_{ij} in the same manner as before and denote z_{ij} the vertices $(1/2+i/n, 1/2+j/n)$, $i,j = 0,\dots,n-1$. Then, using the definition of $|\cdot|_{H^\ell}$ (6.11), we can estimate $\|\Pi_h\psi - I_h\Pi_h\psi\|_{H^{1/3}(\Omega_1)}$ from below by

$$\|\Pi_h\psi - I_h\Pi_h\psi\|^2_{H^{1/3}(\Omega_1)} \ge \sum_{i,j=1}^{n-1} \int_{b_{z_{ij}}}\int_{b_{z_{ij}}} \frac{|\nabla\Pi_h\psi(z_{ij})\cdot(x-y)|^2}{|x-y|^{2(1+1/3)}}\,dy\,dx. \tag{7.9}$$

Also, let $x=(x_1,x_2)$, $y=(y_1,y_2)$, and $\tilde{K}_{ij} = [1/2+i/n, 1/2+i/n+1/3n]\times[1/2+j/n, 1/2+j/n+1/3n]$, and since ψ is invariant in the x_2-direction and $|x-y| \le \sqrt{2}/3n$, (7.9) gives

$$\begin{aligned}
&\|\Pi_h\psi - I_h\Pi_h\psi\|^2_{H^{1/3}(\Omega_1)} \\
&\quad\ge \sum_{i,j=1}^{n-1} \int_{\tilde{K}_{ij}}\int_{\tilde{K}_{ij}} \frac{|\nabla\Pi_h\psi(z_{ij})\cdot(x-y)|^2}{|x-y|^{2(1+1/3)}}\,dy\,dx \\
&\quad\ge \left(\frac{3n}{\sqrt{2}}\right)^{2(1+1/3)} (n-1)\sum_{i=1}^{n-1} \int_{\tilde{K}_{i1}}\int_{\tilde{K}_{i1}} \left|\frac{\partial\Pi_h\psi(z_{i1})}{\partial x_1}\right|^2 (x_1-y_1)^2\,dy\,dx.
\end{aligned} \tag{7.10}$$

Next, we can easily see that $|\frac{\partial\Pi_h\psi(z_{i1})}{\partial x_1}| = (2i+1)/n$ and

$$\int_{\tilde{K}_{i1}}\int_{\tilde{K}_{i1}} (x_1-y_1)^2\,dy\,dx = \frac{1}{2\cdot 3^7 n^6}.$$

Thus,

$$\sum_{i=1}^{n-1} \int_{\tilde{K}_{i1}}\int_{\tilde{K}_{i1}} \left|\frac{\partial\Pi_h\psi(z_{i1})}{\partial x_1}\right|^2 (x_1-y_1)^2\,dy\,dx \ge \frac{1}{2\cdot 3^7 n^8}\sum_{i=1}^{n-1} i^2 = \frac{n(n-1)(2n-1)}{4\cdot 3^8 n^8}.$$

Finally, employing this in (7.10) we get

$$\|\Pi_h\psi - I_h\Pi_h\psi\|_{H^{-1/3}(\Omega_1)} \ge \frac{C}{n^{2/3}} + o\left(\frac{1}{n^{2/3}}\right) = O(h^{2/3}).$$

Combining this with (7.8) we get a contradiction, since $2s \approx 4/3$.

Acknowledgment. The authors thank Dr. S. Tomov for performing and discussing some of the numerical experiments presented in this paper.

REFERENCES

[1] I. Aavatsmark, T. Barkve, Ø. Bøe, and T. Mannseth, *Discretization on unstructured grids for inhomogeneous, anisotropic media. Part* I: *Derivation of the methods*, SIAM J. Sci. Comput., 19 (1998), pp. 1700–1716.
[2] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[3] C. Bacuta, J. H. Bramble, and J. E. Pasciak, *New interpolation results and applications to finite element methods for elliptic boundary value problems*, East-West J. Numer. Math., 9 (2001), pp. 179–198.
[4] C. Bacuta, J. H. Bramble, and J. E. Pasciak, *Using finite element tools in proving shift theorems for elliptic boundary value problems*, Numer. Linear Algebra Appl., 10 (2003), pp. 33–64.
[5] C. Bacuta, J. H. Bramble, and J. Xu, *Regularity estimates for elliptic boundary value problems in Besov spaces*, Math. Comp., 72 (2003), pp. 1577–1595.
[6] R. E. Bank and D. J. Rose, *Some error estimates for the box method*, SIAM J. Numer. Anal., 24 (1987), pp. 777–787.
[7] C. Bennett and R. Sharpley, *Interpolation of Operators*, Academic Press, New York, 1988.
[8] S. C. Brenner, *Multigrid methods for the computation of singular solutions and stress intensity factors* I: *Corner singularities*, Math. Comp., 68 (1999), pp. 559–583.
[9] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[10] Z. Cai, *On the finite volume element method*, Numer. Math., 58 (1991), pp. 713–735.
[11] Z. Cai and S. Kim, *A finite element method using singular functions for the Poisson equation: Corner singularities*, SIAM J. Numer. Anal., 39 (2001), pp. 286–299.
[12] Z. Cai, J. E. Jones, S. F. McCormick, and T. F. Russell, *Control-volume mixed finite element methods*, Comput. Geosci., 1 (1997), pp. 289–315.
[13] P. Chatzipantelidis, *Finite volume methods for elliptic pde's: A new approach*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 307–324.
[14] P. Chatzipantelidis and R. D. Lazarov, *The finite volume element method in nonconvex polygonal domains*, in Finite Volumes for Complex Applications III, R. Herbin and D. Kröner, eds., Hermes Penton Science, London, 2002, pp. 171–178.
[15] S.-H. Chou and Q. Li, *Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach*, Math. Comp., 69 (1999), pp. 103–120.
[16] S.-H. Chou and P. S. Vassilevski, *A general mixed covolume framework for constructing conservative schemes for elliptic problems*, Math. Comp., 68 (1999), pp. 991–1011.
[17] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.
[18] K. Djadel, S. Nicaise, and J. Tabka, *Some refined finite volume methods for elliptic problems with corner singularities*, Internat. J. Finite Volume (electronic journal), 2003.
[19] J. Droniou and T. Gallouët, *Finite volume methods for convection–diffusion equations with right–hand side in H^{-1}*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 705–724.
[20] M. G. Edwards and C. F. Roger, *Finite volume discretization with imposed flux continuity for the general tensor pressure equation*, Comput. Geosci., 2 (1998), pp. 259–290.
[21] M. G. Edwards, R. D. Lazarov, and I. Yotov, eds., Special issue on *Locally conservative numerical methods for flow in porous media*, Comput. Geosci., 6 (2002), pp. 225–564.
[22] R. Eymard, T. Gallouët, and R. Herbin, *Finite Volume Methods*, in Handbook of Numerical Analysis, Vol. VII, North–Holland, Amsterdam, 2001, pp. 713–1020.
[23] R. E. Ewing, T. Lin, and Y. Lin, *On the accuracy of the finite volume element method based on piecewise linear polynomials*, SIAM J. Numer. Anal., 39 (2002), pp. 1865–1888.
[24] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[25] W. Hackbusch, *On first and second order box schemes*, Computing, 41 (1989), pp. 277–296.
[26] R. Herbin, *An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 165–173.
[27] H. Jianguo and X. Shitong, *On the finite volume element method for general self-adjoint elliptic problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1762–1774.
[28] R. Li, Z. Chen, and W. Wu, *Generalized Difference Methods for Differential Equations*, Pure Appl. Math. 226, Marcel Dekker, New York, 2000.
[29] I. D. Mishev, *Finite volume methods on Voronoi mesh*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 193–212.

[30] R. L. Naff, T. F. Russell, and J. D. Wilson, *Shape functions for velocity interpolation in general hexahedral cells*, Comput. Geosci., 6 (2002), pp. 285–314.
[31] R. Nochetto, *Pointwise a posteriori estimates for elliptic problems on highly graded meshes*, Math. Comp., 64 (1995), pp. 1–22.
[32] A. H. Schatz, *A weak discrete maximum principle and stability of the finite element method in L_∞ on plane polygonal domains*. I, Math. Comp., 34 (1980), pp. 77–91.
[33] M. Shashkov, *Conservative Finite-Difference Methods on General Grids*, Symbolic and Numeric Computation Series, CRC Press, Boca Raton, 1996.
[34] E. Süli, *Convergence of finite volume schemes for Poisson's equation on nonuniform meshes*, SIAM J. Numer. Anal., 28 (1991), pp. 1419–1430.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1959–1977

LOCALLY CONSERVATIVE COUPLING OF STOKES AND DARCY FLOWS*

BÉATRICE RIVIÈRE† AND IVAN YOTOV†

Abstract. A locally conservative numerical method for solving the coupled Stokes and Darcy flows problem is formulated and analyzed. The approach employs the mixed finite element method for the Darcy region and the discontinuous Galerkin method for the Stokes region. A discrete inf-sup condition and optimal error estimates are derived.

Key words. multiphysics, porous media flow, incompressible fluid flow, discontinuous Galerkin, mixed finite element, error estimates, inf-sup condition

AMS subject classifications. 35Q35, 65N30, 65N15, 76D07, 76S05

DOI. 10.1137/S0036142903427640

1. Introduction.

The numerical modeling of reactive transport necessitates the use of numerical schemes that do not create artificial mass [14]. Mixed finite element (MFE) and discontinuous Galerkin (DG) methods are two examples of locally mass conservative methods that are used in the geosciences. MFE methods are quite popular for porous media problems [16, 34, 17, 4] and DG methods are attractive for modeling flow on unstructured meshes [33, 31, 30, 32].

Many applications involve different physical processes in different parts of the simulation domain. In this paper we propose a numerical method for approximating the solution to the coupled Darcy–Stokes problem. Such systems arise, for example, in modeling the interaction between surface water (river) and groundwater (aquifer). There are few works in the literature that address the numerical analysis of the coupled Darcy–Stokes problem. In [25], Layton, Schieweck, and Yotov consider a formulation based on the Beavers–Joseph–Saffman interface conditions [5, 35, 24], prove the existence and uniqueness of a weak solution, and analyze a continuous finite element scheme coupled with MFE. A similar formulation is studied by Discacciati, Miglio, and Quarteroni [15], where continuous finite elements are used in both regions. An application of this formulation to vugular porous media is studied in [3]. A singularly perturbed Stokes problem, which models Darcy flow as a limiting case, is considered by Mardal, Tai, and Winther [27]. There, a new finite element is proposed which behaves uniformly in the perturbation parameter. Ewing, Iliev, and Lazarov [18] employ finite difference methods for a similar model involving the Navier–Stokes equations with an added Darcy term.

The model we consider, which is similar to the one in [25], is based on imposing the correct local equations in each region, coupled with appropriate interface conditions. In particular, the fluid region is modeled by the Stokes equations and the porous media region is modeled by the Darcy's law. Continuity of flux, balance of forces, and the Beavers–Joseph–Saffman slip with friction condition (see (2.10) below) are imposed on the interface. In this work we emphasize locally mass conservative discretizations. Conserving mass locally is especially important when the flow equations are coupled

*Received by the editors May 2, 2003; accepted for publication (in revised form) February 27, 2004; published electronically January 20, 2005. This research was partially supported by NSF grants DMS 0107389 and DMS 0112239.

http://www.siam.org/journals/sinum/42-5/42764.html

†Department of Mathematics, 301 Thackeray, University of Pittsburgh, Pittsburgh, PA 15260 (riviere@math.pitt.edu, yotov@math.pitt.edu).

with the reactive transport of chemical species. In the porous media region, the fluid velocity and pressure are obtained by MFE, and in the incompressible flow region, the fluid velocity and pressure are approximated by DG. An advantage of our approach is the possibility of coupling existing highly optimized MFE-based porous media simulators with the flexibility and easy implementation of DG methods for incompressible flows. The meshes at the interface between the two regions may be nonmatching. The estimates are derived for two-dimensional problems. The results are also valid in higher dimension, and depend on the existence of approximation operators (see Remark 4.4 below).

The outline of the paper is as follows. In section 2, the model problem, notation, and scheme are presented. Section 3 contains the derivation of the discrete inf-sup condition. In section 4, approximation results and optimal a priori error estimates are proved. Some concluding remarks follow.

2. Model problem, notation, and scheme. Let Ω be a domain in $\mathbb{R}^d$, $d = 2$, subdivided into two subdomains Ω_1, Ω_2. Let Γ_{12} be the interface $\partial\Omega_1 \cap \partial\Omega_2$. Define $\Gamma_i = \partial\Omega_i \backslash \Gamma_{12}$, $i = 1, 2$. Denote by $\boldsymbol{n}$ the outward normal vector to $\partial\Omega$. Let $\boldsymbol{n}_{12}$ (resp., $\boldsymbol{\tau}_{12}$) be the unit normal (resp., tangential) vector to Γ_{12} outward of Ω_1. Denote by $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2)$ the fluid velocity and by $p = (p_1, p_2)$ the fluid pressure, where $\boldsymbol{u}_i = \boldsymbol{u}|_{\Omega_i}$ and $p_i = p|_{\Omega_i}$. The flow in the domain Ω_1 is assumed to be of Stokes type, and therefore the following equations are satisfied:

$$-\nabla \cdot \boldsymbol{T}(\boldsymbol{u}_1, p_1) = \boldsymbol{f}_1 \quad \text{in } \Omega_1, \tag{2.1}$$

$$\nabla \cdot \boldsymbol{u}_1 = 0 \quad \text{in } \Omega_1, \tag{2.2}$$

$$\boldsymbol{u}_1 = 0 \quad \text{on } \Gamma_1. \tag{2.3}$$

Here $\boldsymbol{T}$ is the stress tensor

$$\boldsymbol{T}(\boldsymbol{u}_1, p_1) = -p_1 \boldsymbol{I} + 2\mu \boldsymbol{D}(\boldsymbol{u}_1)$$

which depends on the viscosity $\mu > 0$ and the strain tensor

$$\boldsymbol{D}(\boldsymbol{u}_1) = \frac{1}{2}(\nabla \boldsymbol{u}_1 + \nabla \boldsymbol{u}_1^T).$$

In the region Ω_2, the fluid pressure and velocity satisfy the single phase Darcy flow equations

$$\nabla \cdot \boldsymbol{u}_2 = f_2 \quad \text{in } \Omega_2, \tag{2.4}$$

$$\boldsymbol{u}_2 = -\boldsymbol{K} \nabla p_2 \quad \text{in } \Omega_2, \tag{2.5}$$

$$\boldsymbol{u}_2 \cdot \boldsymbol{n} = 0 \quad \text{on } \Gamma_2, \tag{2.6}$$

where $\boldsymbol{K}$ is a symmetric and positive definite tensor representing the permeability divided by the viscosity and satisfying, for some $0 < \kappa_0 \le \kappa_1 < \infty$,

$$\kappa_0 \xi^T \xi \le \xi^T \boldsymbol{K}(x) \xi \le \kappa_1 \xi^T \xi \quad \forall x \in \Omega_2,\ \forall \xi \in \mathbb{R}^d. \tag{2.7}$$

The physical quantities are coupled through appropriate interface conditions

$$\boldsymbol{u}_1 \cdot \boldsymbol{n}_{12} = \boldsymbol{u}_2 \cdot \boldsymbol{n}_{12}, \tag{2.8}$$

$$p_1 - 2\mu((\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12}) \cdot \boldsymbol{n}_{12}) = p_2, \tag{2.9}$$

$$\boldsymbol{u}_1 \cdot \boldsymbol{\tau}_{12} = -2G(\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12}) \cdot \boldsymbol{\tau}_{12}. \tag{2.10}$$

Note that condition (2.8) represents the mass conservation across the interface, condition (2.9) imposes balance of forces across the interface, and condition (2.10) is the Beavers–Joseph–Saffman law, where $G > 0$ is a friction constant that can be determined experimentally. The reader should refer to [5, 35, 24, 25] for a detailed description and motivation for the choice of these interface conditions.

For $i = 1, 2$, let $\mathcal{E}_h^i$ be a nondegenerate quasi-uniform subdivision of Ω_i [11] such that the partition $\mathcal{E}_h^1$ consists of triangles and $\mathcal{E}_h^2$ consists of either triangles or rectangles. Let Γ_h^i be the set of interior edges and let h_i denote the maximum diameter of elements in $\mathcal{E}_h^i$. The meshes at the interface between the two domains Ω_i may not match. For $s \geq 0$, $p > 1$, and a domain $E \subset \mathbb{R}^d$, let $W^{s,p}(E)$ be the usual Sobolev spaces [1], let $H^s(E) = W^{s,2}(E)$ be equipped with the usual norm $\|\cdot\|_{s,E}$, and let $L_0^2(E)$ denote the space of L^2 functions with zero average. In the formulation for the Stokes region, we need that both the gradient of $\boldsymbol{u}_1$ and the pressure p_1 have a trace on line segments. For this, it suffices to define the following velocity-pressure spaces for the Stokes region:

$$\boldsymbol{X}^1 = \{\boldsymbol{v}_1 \in (L^2(\Omega_1))^d : \ \forall E \in \mathcal{E}_h^1, \ \boldsymbol{v}_1|_E \in (W^{2,4/3}(E))^d\},$$
$$M^1 = \{q_1 \in L^2(\Omega_1) : \ \forall E \in \mathcal{E}_h^1, \ q_1|_E \in W^{1,4/3}(E)\},$$

with norms

$$|||\boldsymbol{v}_1|||_{s,\Omega_1}^2 = \sum_{E \in \mathcal{E}_h^1} \|\boldsymbol{v}_1\|_{s,E}^2,$$
$$\|\boldsymbol{v}_1\|_{X^1}^2 = |||\nabla \boldsymbol{v}_1|||_{0,\Omega_1}^2 + \sum_{e \in \Gamma_h^1 \cup \Gamma_1} \frac{\sigma_e}{|e|} \|[\boldsymbol{v}_1]\|_{0,e}^2 + \frac{\mu}{G} \sum_{e \in \Gamma_{12}} \|\boldsymbol{v}_1 \cdot \boldsymbol{\tau}_{12}\|_{0,e}^2,$$
$$\|q_1\|_{M^1} = \|q_1\|_{0,\Omega_1}.$$

Here, the parameter $\sigma_e \geq 0$ takes a constant value over each edge e, and $|e|$ denotes the measure (or length) of e. Given a fixed normal vector $\boldsymbol{n}_e$ on each edge $e = \partial E_e^1 \cap \partial E_e^1$, directed from E_e^1 to E_e^2, the average and jump of functions in $\boldsymbol{X}^1$ and M^1 can be defined as

$$\{w\} = \frac{1}{2}(w|_{E_e^1}) + \frac{1}{2}(w|_{E_e^2}), \quad [w] = (w|_{E_e^1}) - (w|_{E_e^2}) \quad \forall e = \partial E_e^1 \cap \partial E_e^2,$$
$$\{w\} = w|_{E_e^1}, \quad [w] = w|_{E_e^1} \quad \forall e = \partial E_e^1 \cap \partial \Omega_1.$$

The velocity-pressure spaces for the Darcy region are

$$\boldsymbol{X}^2 = \left\{\boldsymbol{v} \in H(\mathrm{div}; \Omega_2) : \int_{\partial \Omega_2} \boldsymbol{v} \cdot \boldsymbol{n} w = 0 \ \forall w \in H_{0,\Gamma_{12}}^1(\Omega_2)\right\},$$
$$M^2 = L^2(\Omega_2),$$

where $H(\mathrm{div}; \Omega_2)$ is the space of vectors in $(L^2(\Omega_2))^d$ whose divergence lies in $L^2(\Omega_2)$ and

$$H_{0,\Gamma_{12}}^1(\Omega_2) = \{w \in H^1(\Omega_2) : w = 0 \text{ on } \Gamma_{12}\}.$$

The norms associated with $(\boldsymbol{X}^2, M^2)$ are

$$\|\boldsymbol{v}_2\|_{X^2}^2 = \|\boldsymbol{v}_2\|_{0,\Omega_2}^2 + \|\nabla \cdot \boldsymbol{v}_2\|_{0,\Omega_2}^2, \qquad \|q_2\|_{M^2} = \|q_2\|_{0,\Omega_2}. \tag{2.11}$$

We can now define $\boldsymbol{X} = \boldsymbol{X}^1 \times \boldsymbol{X}^2$ and $M = (M^1 \times M^2) \cap L_0^2(\Omega)$, the spaces for the coupled formulation with the usual norms

$$\|\boldsymbol{v}\|_X^2 = \|\boldsymbol{v}_1\|_{X^1}^2 + \|\boldsymbol{v}_2\|_{X^2}^2, \qquad \|q\|_M^2 = \|q_1\|_{M^1}^2 + \|q_2\|_{M^2}^2. \tag{2.12}$$

In [25], it was shown that there exists a unique weak solution $(\boldsymbol{u}, p)$ of the coupled problem (2.1)–(2.10), with $\boldsymbol{u}_1 \in (H^1(\Omega_1))^d$, $\boldsymbol{u}_2 \in \boldsymbol{X}^2$, and $p \in M$. We will assume that the solution $(\boldsymbol{u}, p)$ is regular enough, so that it is a strong solution of (2.1)–(2.10). Next, we introduce the bilinear forms $a_1 : \boldsymbol{X}^1 \times \boldsymbol{X}^1 \to \mathbb{R}$ and $b_1 : \boldsymbol{X}^1 \times M^1 \to \mathbb{R}$,

$$\begin{aligned} a_1(\boldsymbol{u}_1, \boldsymbol{v}_1) = {} & 2\mu \sum_{E \in \mathcal{E}_h^1} \int_E \boldsymbol{D}(\boldsymbol{u}_1) : \boldsymbol{D}(\boldsymbol{v}_1) + \sum_{e \in \Gamma_h^1 \cup \Gamma_1} \frac{\sigma_e}{|e|} \int_e [\boldsymbol{u}_1] \cdot [\boldsymbol{v}_1] \\ & - 2\mu \sum_{e \in \Gamma_h^1 \cup \Gamma_1} \int_e \{\boldsymbol{D}(\boldsymbol{u}_1)\} \boldsymbol{n}_e \cdot [\boldsymbol{v}_1] + 2\mu\epsilon \sum_{e \in \Gamma_h^1 \cup \Gamma_1} \int_e \{\boldsymbol{D}(\boldsymbol{v}_1)\} \boldsymbol{n}_e \cdot [\boldsymbol{u}_1] \\ & + \frac{\mu}{G} \sum_{e \in \Gamma_{12}} \int_e \boldsymbol{u}_1 \cdot \boldsymbol{\tau}_{12} \boldsymbol{v}_1 \cdot \boldsymbol{\tau}_{12}, \end{aligned} \tag{2.13}$$

$$b_1(\boldsymbol{v}_1, p_1) = - \sum_{E \in \mathcal{E}_h} \int_E p_1 \nabla \cdot \boldsymbol{v}_1 + \sum_{e \in \Gamma_h^1 \cup \Gamma_1} \int_e \{p_1\} [\boldsymbol{v}_1] \cdot \boldsymbol{n}_e. \tag{2.14}$$

Here, ϵ is a constant that takes the value -1 or $+1$, which makes the bilinear form a_1 symmetric or nonsymmetric. The bilinear forms corresponding to the Darcy region are $a_2 : \boldsymbol{X}^2 \times \boldsymbol{X}^2 \to \mathbb{R}$ and $b_2 : \boldsymbol{X}^2 \times M^2 \to \mathbb{R}$:

$$a_2(\boldsymbol{u}_2, \boldsymbol{v}_2) = \int_{\Omega_2} \boldsymbol{K}^{-1} \boldsymbol{u}_2 \cdot \boldsymbol{v}_2, \tag{2.15}$$

$$b_2(\boldsymbol{v}_2, q_2) = - \int_{\Omega_2} q_2 \nabla \cdot \boldsymbol{v}_2. \tag{2.16}$$

Let k_1, k_2, and l_2 be positive integers. Let $\boldsymbol{X}_h$ and M_h be finite-dimensional subspaces of $\boldsymbol{X}$ and M, respectively, such that

$$\boldsymbol{X}_h = \boldsymbol{X}_h^1 \times \boldsymbol{X}_h^2, \qquad M_h = M_h^1 \times M_h^2,$$

where $(\boldsymbol{X}_h^1, M_h^1)$ is the pair of discontinuous finite element spaces

$$\begin{aligned} \boldsymbol{X}_h^1 &= \{\boldsymbol{v}_1 \in \boldsymbol{X}^1 : \ \forall E \in \mathcal{E}_h^1, \ \boldsymbol{v}_1 \in (\mathbb{P}_{k_1}(E))^d\}, \\ M_h^1 &= \{q_1 \in M^1 : \ \forall E \in \mathcal{E}_h^1, \ q_1 \in \mathbb{P}_{k_1 - 1}(E)\}. \end{aligned}$$

The discrete spaces corresponding to the Darcy region consist of the standard mixed finite element spaces (such as RT spaces [29], BDM spaces [9], BDFM spaces [8], and BDDF spaces [7]). The mixed spaces $\boldsymbol{X}_h^2$ and M_h^2 contain all polynomials of degree at least k_2 and l_2, respectively. Note that for the Raviart–Thomas (RT) spaces, the condition $l_2 = k_2$ holds. We also assume that

$$\forall \boldsymbol{v}_2 \in \boldsymbol{X}_h^2, \quad \boldsymbol{v}_2 \cdot \boldsymbol{n} = 0 \quad \text{on } \Gamma_2.$$

Let E be a mesh element with diameter h_E. Given $p \in L_0^2(\Omega)$, we denote by $\tilde{p}$ the L^2 projection of p in M_h satisfying

$$\forall q \in \mathbb{P}_{k_1 - 1}(E), \quad \int_E q(\tilde{p} - p) = 0 \quad \forall E \in \mathcal{E}_h^1, \tag{2.17}$$

$$\forall q \in \mathbb{P}_{l_2}(E), \quad \int_E q(\tilde{p} - p) = 0 \quad \forall E \in \mathcal{E}_h^2, \tag{2.18}$$

and, if $p|_{\Omega_1} \in H^{k_1}(\Omega_1)$ and $p|_{\Omega_2} \in H^{l_2+1}(\Omega_2)$, then

$$\|p-\tilde{p}\|_{m,E} \le Ch_E^{k_1-m}|p|_{k_1,E}, \quad E \subset \Omega_1,\ m=0,1, \tag{2.19}$$

$$\|p-\tilde{p}\|_{m,E} \le Ch_E^{l_2+1-m}|p|_{l_2+1,E}, \quad E \subset \Omega_2,\ m=0,1. \tag{2.20}$$

Remark 2.1. One advantage of the DG method is that one can vary the polynomial degrees from element to element. Here we assume that k_1 is the minimum of the polynomial degrees used in the Stokes region.

Here and throughout the paper, C denotes a varying constant that is independent of the diameter of the mesh elements. We also make use of the quasi-local interpolant $\boldsymbol{\Pi}_h^1 : (H^1(\Omega_1))^d \to \boldsymbol{X}_h^1$ [13, 19, 12, 22] satisfying, for all $\boldsymbol{v}_1 \in (H^1(\Omega_1))^d$,

$$b_1(\boldsymbol{\Pi}_h^1\boldsymbol{v}_1 - \boldsymbol{v}_1, q_1) = 0 \quad \forall q_1 \in M_h^1, \tag{2.21}$$

$$\forall e \in \Gamma_h^1 \cup \Gamma_1,\ \int_e [\boldsymbol{\Pi}_h^1\boldsymbol{v}_1]\cdot q_1 = 0 \quad \forall \boldsymbol{v}_1 \in (H^1(\Omega_1))^d : \boldsymbol{v}_1 = 0 \text{ on } \Gamma_1,\ \forall q_1 \in M_h^1, \tag{2.22}$$

$$|||\boldsymbol{\Pi}_h^1\boldsymbol{v}_1|||_{1,\Omega_1} \le C\|\boldsymbol{v}_1\|_{1,\Omega_1}. \tag{2.23}$$

The operator $\boldsymbol{\Pi}_h^1$ has the optimal approximation properties

$$|\boldsymbol{\Pi}_h^1\boldsymbol{v}_1 - \boldsymbol{v}_1|_{m,E} \le Ch_E^{s-m}|\boldsymbol{v}_1|_{s,\delta(E)} \quad \forall 1 \le s \le k_1+1,\ \forall \boldsymbol{v}_1 \in H^s(\Omega_1),\ m=0,1, \tag{2.24}$$

where $\delta(E)$ is a suitable macro-element containing E. Moreover, it holds that for at least one edge e of every element $E \in \mathcal{E}_h^1$,

$$\int_e (\boldsymbol{\Pi}_h^1\boldsymbol{v}_1 - \boldsymbol{v}_1) = 0 \quad \forall \boldsymbol{v}_1 \in (H^1(\Omega_1))^d. \tag{2.25}$$

We note that (2.25) holds true for all edges in the cases $k=1$ and $k=2$. For $k=3$, we can assume, without loss of generality, that (2.25) is satisfied for all edges in Γ_{12}. We will make use of the following bounds on $\boldsymbol{\Pi}_h^1$.

LEMMA 2.2. *Let* $1 \le s \le k_1+1$. *For all* $\boldsymbol{v}_1 \in (H^s(\Omega_1))^d$,

$$\|\boldsymbol{\Pi}_h^1\boldsymbol{v}_1 - \boldsymbol{v}_1\|_{X^1} \le Ch_1^{s-1}|\boldsymbol{v}_1|_{s,\Omega_1}, \tag{2.26}$$

$$\|\boldsymbol{\Pi}_h^1\boldsymbol{v}_1\|_{X^1} \le C\|\boldsymbol{v}_1\|_{1,\Omega_1}. \tag{2.27}$$

Proof. From Lemma 3.10 of [22] and from (2.24), we have

$$\|\boldsymbol{v}_1 - \boldsymbol{\Pi}_h^1\boldsymbol{v}_1\|_{X^1} \le C|||\nabla(\boldsymbol{v}_1 - \boldsymbol{\Pi}_h^1\boldsymbol{v}_1)|||_{0,\Omega_1} \le Ch_1^{s-1}|\boldsymbol{v}_1|_{s,\Omega_1}. \tag{2.28}$$

The bound (2.27) follows easily from the triangle inequality and (2.26) with $s=1$, using that $\|\boldsymbol{v}_1\|_{X^1} \le C\|\boldsymbol{v}_1\|_{1,\Omega_1}$ for $\boldsymbol{v}_1 \in (H^1(\Omega_1))^d$. ∎

We also recall the MFE interpolant $\boldsymbol{\Pi}_h^2 : \boldsymbol{X}^2 \cap (H^\theta(\Omega_2))^d \to \boldsymbol{X}_h^2$ for any $\theta > 0$, satisfying [10], for any $\boldsymbol{v}_2 \in \boldsymbol{X}^2 \cap (H^\theta(\Omega_2))^d$,

$$b_2(\boldsymbol{\Pi}_h^2\boldsymbol{v}_2 - \boldsymbol{v}_2, q_2) = 0 \quad \forall q_2 \in M_h^2, \tag{2.29}$$

$$\int_e ((\boldsymbol{\Pi}_h^2\boldsymbol{v}_2 - \boldsymbol{v}_2)\cdot\boldsymbol{n}_e)\boldsymbol{w}_2\cdot\boldsymbol{n}_e = 0 \quad \forall e \in \Gamma_h^2,\ \forall \boldsymbol{w}_2 \in \boldsymbol{X}_h^2. \tag{2.30}$$

Moreover, $\boldsymbol{\Pi}_h^2$ satisfies the approximation properties

$$\|\boldsymbol{v}_2 - \boldsymbol{\Pi}_h^2\boldsymbol{v}_2\|_{0,E} \le Ch_E^s|\boldsymbol{v}_2|_{s,E}, \quad 1 \le s \le k_2+1, \tag{2.31}$$

$$\|\nabla\cdot(\boldsymbol{v}_2 - \boldsymbol{\Pi}_h^2\boldsymbol{v}_2)\|_{0,E} \le Ch_E^s|\nabla\cdot\boldsymbol{v}_2|_{s,E}, \quad 0 \le s \le l_2+1. \tag{2.32}$$

It has been shown by Mathew in [28] for the Raviart–Thomas elements [29] that

$$\|\boldsymbol{\Pi}_h^2 \boldsymbol{v}_2\|_{H(\mathrm{div};\Omega_2)} \leq C(\|\boldsymbol{v}_2\|_{\theta,\Omega_2} + \|\nabla \cdot \boldsymbol{v}_2\|_{0,\Omega_2}), \tag{2.33}$$

a result that can be trivially extended to the other families of MFE spaces. Recall the basic trace inequalities on any mesh element E with diameter h_E

$$\forall \phi \in H^1(E),\ \forall e \subset \partial E, \quad \|\phi\|_{0,e}^2 \leq C(h_E^{-1}\|\phi\|_{0,E}^2 + h_E|\phi|_{1,E}^2), \tag{2.34}$$

$$\forall \phi \in H^2(E),\ \forall e \subset \partial E, \quad \|\nabla\phi \cdot \boldsymbol{n}\|_{0,e}^2 \leq C(h_E^{-1}\|\phi\|_{1,E}^2 + h_E|\phi|_{2,E}^2), \tag{2.35}$$

$$\forall \phi \in \mathbb{P}_k(E),\ \forall e \subset \partial E, \quad \|\nabla\phi \cdot \boldsymbol{n}\|_{0,e} \leq C h_E^{-1/2}|\phi|_{1,E}, \tag{2.36}$$

Recall also the Korn's inequality proved in [6]

$$\forall \boldsymbol{v} \in \boldsymbol{X}_h^1, \quad C|||\nabla \boldsymbol{v}|||_{0,\Omega_1}^2 \leq |||\boldsymbol{D}(\boldsymbol{v})|||_{0,\Omega_1}^2 + \sum_{e \in \Gamma_h^1 \cup \Gamma_1} \frac{1}{|e|}\|[\boldsymbol{v}]\|_{0,e}^2. \tag{2.37}$$

Define the finite-dimensional space of functions on the interface $\Lambda_h = X_h^2 \cdot \boldsymbol{n}_{12}$ and let

$$\boldsymbol{V}_h = \left\{ \boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2) \in \boldsymbol{X}_h : \sum_{e \in \Gamma_{12}} \int_e \eta(\boldsymbol{v}_1 - \boldsymbol{v}_2) \cdot \boldsymbol{n}_{12} = 0\ \forall \eta \in \Lambda_h \right\}.$$

Defining $a = a_1 + a_2$ and $b = b_1 + b_2$, the numerical scheme is, Find $(\boldsymbol{U}, P) \in \boldsymbol{V}_h \times M_h$ such that

$$a(\boldsymbol{U}, \boldsymbol{v}) + b(\boldsymbol{v}, P) = \int_{\Omega_1} \boldsymbol{f}_1 \cdot \boldsymbol{v} \quad \forall \boldsymbol{v} \in \boldsymbol{V}_h, \tag{2.38}$$

$$b(\boldsymbol{U}, q) = \int_{\Omega_2} f_2 q \quad \forall q \in M_h. \tag{2.39}$$

Remark 2.3. This scheme is locally mass conservative. Indeed, if one chooses the test function in (2.39) such that $q = 1$ on E and $q = 0$ on the rest of the domain, we have

$$\int_{\partial E} \{\boldsymbol{U}\} \cdot \boldsymbol{n}_E = 0 \quad \forall E \subset \mathcal{E}_h^1,$$
$$\int_{\partial E} \boldsymbol{U} \cdot \boldsymbol{n}_E = \int_E f_2 \quad \forall E \subset \mathcal{E}_h^2.$$

Remark 2.4. The space of weakly-continuous-normal velocities $\boldsymbol{V}_h$ is introduced to facilitate the analysis of the numerical method. A direct construction of this space may, however, be difficult. An equivalent formulation to (2.38)–(2.39) is given in section 5. It is only based on the space $\boldsymbol{X}_h$ and is more suitable for implementation. The space Λ_h plays the role of a Lagrange multiplier or mortar space for imposing continuity of the normal velocities on Γ_{12}. The choice $\Lambda_h = X_h^2 \cdot \boldsymbol{n}_{12}$ is critical for the stability and accuracy of the numerical scheme, even in the case of nonmatching grids across Γ_{12}. This choice differs from the mortar space used in [2] in the case of MFE discretizations on nonmatching grids.

In the rest of the section, we show that the solution of the coupled problem satisfies the scheme up to an interface consistency error. We also prove uniqueness and existence of the discrete solution.

LEMMA 2.5. *If $(\boldsymbol{u},p)\in \boldsymbol{X}\times M$ solves the coupled Stokes–Darcy flow problem (2.1)–(2.10), such that $\boldsymbol{u}_i=\boldsymbol{u}|_{\Omega_i}$ and $p_i=p|_{\Omega_i}$, then $(\boldsymbol{u},p)$ satisfies the variational problem*

$$a(\boldsymbol{u},\boldsymbol{v})+b(\boldsymbol{v},p)=\int_{\Omega_1}\boldsymbol{f}_1\cdot\boldsymbol{v}_1-\sum_{e\in\Gamma_{12}}\int_e p_2(\boldsymbol{v}_1-\boldsymbol{v}_2)\cdot\boldsymbol{n}_{12}\quad\forall\boldsymbol{v}\in\boldsymbol{V}_h, \tag{2.40}$$

$$b(\boldsymbol{u},q)=\int_{\Omega_2}f_2q\quad\forall q\in M_h. \tag{2.41}$$

Proof. Multiplying the Stokes equation (2.1) by $\boldsymbol{v}_1\in\boldsymbol{X}_h^1$ and integrating by parts over one element E,

$$\int_E T(\boldsymbol{u}_1,p_1):\nabla\boldsymbol{v}_1-\int_{\partial E}T(\boldsymbol{u}_1,p_1)\boldsymbol{n}_E\cdot\boldsymbol{v}_1=\int_E\boldsymbol{f}_1\cdot\boldsymbol{v}_1.$$

Summing over all elements E,

$$\begin{aligned}&\sum_E\int_E(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1)):\nabla\boldsymbol{v}_1-\sum_{e\in\Gamma_h^1}\int_e[(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))]\boldsymbol{n}_e\cdot\boldsymbol{v}_1\\&\quad-\int_{\Gamma_{12}}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}_{12}\cdot\boldsymbol{v}_1-\int_{\Gamma_1}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}\cdot\boldsymbol{v}_1=\int_{\Omega_1}\boldsymbol{f}_1\cdot\boldsymbol{v}_1.\end{aligned}$$

It is easy to show that $\boldsymbol{D}(\boldsymbol{u}_1):\nabla\boldsymbol{v}_1=\boldsymbol{D}(\boldsymbol{u}_1):\boldsymbol{D}(\boldsymbol{v}_1)$ and that $\boldsymbol{I}:\nabla\boldsymbol{v}_1=\nabla\cdot\boldsymbol{v}_1$. Thus, the equation becomes

$$\begin{aligned}&\sum_E\int_E(2\mu\boldsymbol{D}(\boldsymbol{u}_1):\boldsymbol{D}(\boldsymbol{v}_1)-p_1\nabla\cdot\boldsymbol{v}_1)\\&\quad-\sum_{e\in\Gamma_h^1}\int_e\{-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1)\}\boldsymbol{n}_e\cdot[\boldsymbol{v}_1]-\sum_{e\in\Gamma_h^1}\int_e[-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1)]\boldsymbol{n}_e\cdot\{\boldsymbol{v}_1\}\\&\quad-\int_{\Gamma12}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}_{12}\cdot\boldsymbol{v}_1-\int_{\Gamma_1}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}\cdot\boldsymbol{v}_1=\int_{\Omega_1}\boldsymbol{f}_1\cdot\boldsymbol{v}_1.\end{aligned}$$

By regularity of the true solution, we have

$$\begin{aligned}&\sum_E\int_E(2\mu\boldsymbol{D}(\boldsymbol{u}_1):\boldsymbol{D}(\boldsymbol{v}_1)-p_1\nabla\cdot\boldsymbol{v}_1)-\int_{\Gamma_{12}}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}_{12}\cdot\boldsymbol{v}_1\\&\quad-\sum_{e\in\Gamma_h^1}\int_e\{-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1)\}\boldsymbol{n}_e\cdot[\boldsymbol{v}_1]+\epsilon\sum_{e\in\Gamma_h^1}\int_e\{2\mu\boldsymbol{D}(\boldsymbol{v}_1)\}\boldsymbol{n}_e\cdot[\boldsymbol{u}_1]\\&\quad-\int_{\Gamma_1}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}\cdot\boldsymbol{v}_1+\epsilon\int_{\Gamma_1}2\mu\boldsymbol{D}(\boldsymbol{v}_1)\boldsymbol{n}\cdot\boldsymbol{u}_1=\int_{\Omega_1}\boldsymbol{f}_1\cdot\boldsymbol{v}_1.\end{aligned}$$

Let us now consider the interface term

$$(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}_{12}=-p_1\boldsymbol{n}_{12}+(2\mu(\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12})\cdot\boldsymbol{\tau}_{12})\boldsymbol{\tau}_{12}+(2\mu(\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12})\cdot\boldsymbol{n}_{12})\boldsymbol{n}_{12},$$

which, combined with $\boldsymbol{v}_1=(\boldsymbol{v}_1\cdot\boldsymbol{\tau}_{12})\boldsymbol{\tau}_{12}+(\boldsymbol{v}_1\cdot\boldsymbol{n}_{12})\boldsymbol{n}_{12}$, gives

$$\begin{aligned}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}_{12}\cdot\boldsymbol{v}_1=&-p_1(\boldsymbol{v}_1\cdot\boldsymbol{n}_{12})+2\mu(\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12})\cdot\boldsymbol{\tau}_{12}(\boldsymbol{v}_1\cdot\boldsymbol{\tau}_{12})\\&+2\mu(\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12})\cdot\boldsymbol{n}_{12}(\boldsymbol{v}_1\cdot\boldsymbol{n}_{12}).\end{aligned}$$

Thus,

$$-\int_{\Gamma_{12}}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}_{12}\cdot\boldsymbol{v}_1=-\int_{\Gamma_{12}}(-p_1+2\mu(\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12})\cdot\boldsymbol{n}_{12})(\boldsymbol{v}_1\cdot\boldsymbol{n}_{12})$$
$$-\int_{\Gamma_{12}}2\mu(\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12})\cdot\boldsymbol{\tau}_{12}(\boldsymbol{v}_1\cdot\boldsymbol{\tau}_{12}).$$

With the interface conditions (2.9) and (2.10), we obtain

$$-\int_{\Gamma_{12}}(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}_{12}\cdot\boldsymbol{v}_1=\int_{\Gamma_{12}}p_2(\boldsymbol{v}_1\cdot\boldsymbol{n}_{12})+\frac{\mu}{G}\int_{\Gamma_{12}}(\boldsymbol{u}_1\cdot\boldsymbol{\tau}_{12})(\boldsymbol{v}_1\cdot\boldsymbol{\tau}_{12}).$$

Thus,

$$\sum_E\int_E(2\mu\boldsymbol{D}(\boldsymbol{u}_1):\boldsymbol{D}(\boldsymbol{v}_1)-p_1\nabla\cdot\boldsymbol{v}_1)$$
$$-\sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e\{(-p_1\boldsymbol{I}+2\mu\boldsymbol{D}(\boldsymbol{u}_1))\boldsymbol{n}_e\}\cdot[\boldsymbol{v}_1]+\epsilon\sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e\{2\mu\boldsymbol{D}(\boldsymbol{v}_1)\boldsymbol{n}_e\}\cdot[\boldsymbol{u}_1]$$
$$+\int_{\Gamma_{12}}p_2\boldsymbol{v}_1\cdot\boldsymbol{n}_{12}+\frac{\mu}{G}\int_{\Gamma_{12}}\boldsymbol{u}_1\cdot\boldsymbol{\tau}_{12}\boldsymbol{v}_1\cdot\boldsymbol{\tau}_{12}=\int_{\Omega_1}\boldsymbol{f}_1\cdot\boldsymbol{v}_1$$

which is equivalent to

$$a_1(\boldsymbol{u}_1,\boldsymbol{v}_1)+b_1(\boldsymbol{v}_1,p_1)+\int_{\Gamma_{12}}p_2\boldsymbol{v}_1\cdot\boldsymbol{n}_{12}=\int_{\Omega_1}\boldsymbol{f}_1\cdot\boldsymbol{v}_1\quad\forall\boldsymbol{v}_1\in\boldsymbol{X}_h^1. \tag{2.42}$$

The Darcy's law (2.5) can be rewritten as $\boldsymbol{K}^{-1}\boldsymbol{u}_2=-\nabla p_2$. As usual, multiplication by $\boldsymbol{v}_2\in\boldsymbol{X}_h^2$ and integration by parts on the Darcy region yields

$$\int_{\Omega_2}\boldsymbol{K}^{-1}\boldsymbol{u}_2\cdot\boldsymbol{v}_2=-\int_{\Omega_2}\nabla p_2\cdot\boldsymbol{v}_2=\int_{\Omega_2}p_2\nabla\cdot\boldsymbol{v}_2-\int_{\partial\Omega_2}p_2\boldsymbol{v}_2\cdot\boldsymbol{n}$$
$$=\int_{\Omega_2}p_2\nabla\cdot\boldsymbol{v}_2-\int_{\Gamma_2}p_2\boldsymbol{v}_2\cdot\boldsymbol{n}+\int_{\Gamma_{12}}p_2\boldsymbol{v}_2\cdot\boldsymbol{n}_{12},$$

or equivalently,

$$a_2(\boldsymbol{u}_2,\boldsymbol{v}_2)+b_2(\boldsymbol{v}_2,p_2)-\int_{\Gamma_{12}}p_2\boldsymbol{v}_2\cdot\boldsymbol{n}_{12}=0\quad\forall\boldsymbol{v}_2\in\boldsymbol{X}_h^2. \tag{2.43}$$

Adding (2.42) and (2.43) yields (2.40). Clearly, (2.2) and the regularity of the solution gives

$$b_1(\boldsymbol{u}_1,q)=0\quad\forall q\in M_h^1.$$

Finally, a simple integration in (2.4) yields

$$b_2(\boldsymbol{u}_2,q)=\int_{\Omega_2}f_2q\quad\forall q\in M_h^2,$$

and adding to the previous equation gives the result. ∎

Next, we prove a coercivity lemma that holds true under the following condition.

Hypothesis A. In the definition of the bilinear form $a_1(\cdot,\cdot)$, let us assume that either the condition (a) or (b) holds true.

(a) $\epsilon = 1$ and $\sigma_e > 1$ for all edges in $\Gamma_h^1 \cup \Gamma_1$. For instance, one may choose $\sigma_e = 2$.
(b) $\epsilon = -1$ and $\sigma_e \geq \sigma_0 > 0$ for σ_0 large enough.

LEMMA 2.6. *Assuming Hypothesis A, there exists a positive constant C_0 such that*

$$C_0\|\boldsymbol{v}\|_X^2 \leq a(\boldsymbol{v},\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \boldsymbol{X}_h : \nabla\cdot\boldsymbol{v} = 0 \text{ a.e. in } \Omega_2.$$

Proof. Let $\boldsymbol{v} \in \boldsymbol{X}_h$. Then $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$ with $\boldsymbol{v}_i \in \boldsymbol{X}_h^i$, $i = 1, 2$. Using (2.13) and (2.15),

$$\begin{aligned} a(\boldsymbol{v},\boldsymbol{v}) = {} & 2\mu \sum_{E\in\mathcal{E}_h^1} \int_E \boldsymbol{D}(\boldsymbol{v}_1) : \boldsymbol{D}(\boldsymbol{v}_1) + \sum_{e\in\Gamma_h^1\cup\Gamma_1} \frac{\sigma_e}{|e|}\int_e [\boldsymbol{v}_1]^2 \\ & - 2(1-\epsilon)\mu \sum_{e\in\Gamma_h^1\cup\Gamma_1} \int_e \{\boldsymbol{D}(\boldsymbol{v}_1)\}\boldsymbol{n}_e\cdot[\boldsymbol{v}_1] + \frac{\mu}{G}\sum_{e\in\Gamma_{12}}\int_e (\boldsymbol{v}_1\cdot\boldsymbol{\tau}_{12})^2 + \int_{\Omega_2} \boldsymbol{K}^{-1}\boldsymbol{v}_2\cdot\boldsymbol{v}_2. \end{aligned}$$

Using Korn's inequality (2.37) and the bound on $\boldsymbol{K}$ (2.7) gives

$$\begin{aligned} a(\boldsymbol{v},\boldsymbol{v}) \geq {} & C\mu|||\nabla\boldsymbol{v}|||_{0,\Omega_1}^2 + C\sum_{e\in\Gamma_h^1\cup\Gamma_1} \frac{\sigma_e - 1}{|e|}\int_e [\boldsymbol{v}_1]^2 \\ & - 2(1-\epsilon)\mu \sum_{e\in\Gamma_h^1\cup\Gamma_1} \int_e \{\boldsymbol{D}(\boldsymbol{v}_1)\}\boldsymbol{n}_e\cdot[\boldsymbol{v}_1] + \frac{\mu}{G}\sum_{e\in\Gamma_{12}}\int_e (\boldsymbol{v}_1\cdot\boldsymbol{\tau}_{12})^2 + \frac{1}{\kappa_1}\|\boldsymbol{v}_2\|_{0,\Omega_2}^2. \end{aligned}$$

If $\epsilon = 1$, then the result is straightforward. If $\epsilon = -1$, we have from trace inequality (2.36)

$$\begin{aligned} 2(1-\epsilon)\mu \sum_{e\in\Gamma_h^1\cup\Gamma_1} \int_e \{\boldsymbol{D}(\boldsymbol{v}_1)\}\boldsymbol{n}_e\cdot[\boldsymbol{v}_1] &\leq 4\mu \sum_{e\in\Gamma_h^1\cup\Gamma_1} h_1^{-1/2}\|\nabla\boldsymbol{v}_1\|_{0,E_e}\left(\frac{|e|}{|e|}\right)^{1/2}\|[\boldsymbol{v}_1]\|_{0,e} \\ &\leq \frac{C}{2}\mu|||\nabla\boldsymbol{v}_1|||_{0,\Omega_1}^2 + \tilde{C}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{1}{|e|}\int_e [\boldsymbol{v}_1]^2. \end{aligned}$$

Thus, we obtain if $\epsilon = -1$,

$$\begin{aligned} a(\boldsymbol{v}_1,\boldsymbol{v}_1) \geq {} & \frac{3}{4}\mu|||\nabla\boldsymbol{v}_1|||_{0,\Omega_1}^2 + \sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{C(\sigma_e-1)-\tilde{C}}{|e|}\int_e [\boldsymbol{v}_1]^2 \\ & + \frac{\mu}{G}\sum_{e\in\Gamma_{12}}\int_e (\boldsymbol{v}_1\cdot\boldsymbol{\tau}_{12})^2 + \frac{1}{\kappa_1}\|\boldsymbol{v}_2\|_{0,\Omega_2}^2 \geq C_0(\|\boldsymbol{v}_1\|_{X^1}^2 + \|\boldsymbol{v}_2\|_{0,\Omega_2}^2) \end{aligned}$$

with C_0 positive constant, assuming that σ_e is large enough:

$$(C(\sigma_e - 1) - \tilde{C} \geq C_0 > 0). \qquad \square$$

We are now ready to prove that the discrete scheme (2.38)–(2.39) is solvable.

LEMMA 2.7. *If Hypothesis A holds, then there exists a unique solution to the problem* (2.38)–(2.39).

Proof. Since the problem (2.38)–(2.39) is finite dimensional, it suffices to show that the solution is unique. Set $f_i = 0$ and choose $\boldsymbol{v} = \boldsymbol{U}$ and $q = P$. Then

$$a(\boldsymbol{U},\boldsymbol{U}) = 0.$$

In addition,

$$b(\boldsymbol{U}, q) = 0 \quad \forall q \in M_h,$$

which implies that $\nabla \cdot \boldsymbol{U} = 0$ in Ω_2, since $\nabla \cdot \boldsymbol{X}_h^2 = M_h^2$. Therefore Lemma 2.6 directly implies that $\boldsymbol{U} = 0$. Thus, the pressure satisfies

$$b(\boldsymbol{v}, P) = 0 \quad \forall \boldsymbol{v} \in \boldsymbol{V}_h.$$

The inf-sup condition (3.1) proved below implies that $P = 0$. ∎

3. A discrete inf-sup condition. In this section, a discrete inf-sup condition is proved.

THEOREM 3.1. *There exists a positive constant β such that*

$$\inf_{q_h \in M_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_h} \frac{b(\boldsymbol{v}_h, q_h)}{\|\boldsymbol{v}_h\|_X \|q_h\|_M} \geq \beta. \tag{3.1}$$

Proof. Let $q_h \in M_h$ be given. Then there exists [20, 21] $\boldsymbol{v} \in (H^1(\Omega))^d$ such that

$$\nabla \cdot \boldsymbol{v} = -q_h \quad \text{in } \Omega, \qquad \boldsymbol{v} = 0 \quad \text{on } \partial\Omega,$$

satisfying

$$\|\boldsymbol{v}\|_{1,\Omega} \leq C \|q_h\|_{0,\Omega}.$$

Note that

$$b(v, q_h) = -\int_\Omega (\nabla \cdot \boldsymbol{v}) q_h = \|q_h\|_M^2,$$

which, together with the above a priori bound, implies

$$b(\boldsymbol{v}, q_h) \geq \frac{1}{C} \|\boldsymbol{v}\|_{1,\Omega} \|q_h\|_M.$$

Next, we need to construct an operator $\pi_h : \boldsymbol{X}^1 \times (\boldsymbol{X}^2 \cap (H^1(\Omega_2))^d) \to \boldsymbol{V}_h$ satisfying

$$b(\pi_h \boldsymbol{v} - \boldsymbol{v}, q_h) = 0 \quad \forall q_h \in M_h, \qquad \text{and} \qquad \|\pi_h \boldsymbol{v}\|_X \leq C \|\boldsymbol{v}\|_{1,\Omega}. \tag{3.2}$$

Let $\pi_h \boldsymbol{v} = (\pi_h^1 \boldsymbol{v}, \pi_h^2 \boldsymbol{v}) \in \boldsymbol{X}_h^1 \times \boldsymbol{X}_h^2$. We take $\pi_h^1 \boldsymbol{v} = \boldsymbol{\Pi}_h^1 \boldsymbol{v}_1$ where $\boldsymbol{\Pi}_h^1 : \boldsymbol{X}^1 \to \boldsymbol{X}_h^1$ is the quasi-local interpolant defined in (2.21). Clearly, due to (2.27),

$$\|\pi_h^1 \boldsymbol{v}\|_{X^1} \leq C \|\boldsymbol{v}\|_{1,\Omega_1}. \tag{3.3}$$

To define $\pi_h^2 \boldsymbol{v}$, consider the auxiliary problem

$$\nabla \cdot \nabla \varphi = 0 \quad \text{in } \Omega_2, \tag{3.4}$$

$$\nabla \varphi \cdot \boldsymbol{n} = 0 \quad \text{on } \Gamma_2, \tag{3.5}$$

$$\nabla \varphi \cdot \boldsymbol{n}_{12} = (\pi_h^1 \boldsymbol{v} - \boldsymbol{v}) \cdot \boldsymbol{n}_{12} \quad \text{on } \Gamma_{12}. \tag{3.6}$$

The problem is well posed, since

$$\int_{\Gamma_{12}} (\pi_h^1 \boldsymbol{v} - \boldsymbol{v}) \cdot \boldsymbol{n}_{12} = 0,$$

due to (2.25). Let $\boldsymbol{z} = \nabla\varphi$. We note that the piecewise smooth function $\pi_h^1\boldsymbol{v}\cdot\boldsymbol{n}_{12} \in H^\theta(\Gamma_{12})$ for any $0 < \theta < 1/2$. By elliptic regularity [26],

$$\|\boldsymbol{z}\|_{\theta,\Omega_2} \le C\|(\pi_h^1\boldsymbol{v}-\boldsymbol{v})\cdot\boldsymbol{n}_{12}\|_{\theta-1/2,\Gamma_{12}}, \quad 0\le\theta\le 1/2. \tag{3.7}$$

Let $\boldsymbol{w} = \boldsymbol{v}+\boldsymbol{z}$. Clearly $\nabla\cdot\boldsymbol{w} = \nabla\cdot\boldsymbol{v}$ in Ω_2 and $\boldsymbol{w}\cdot\boldsymbol{n}_{12} = \pi_h^1\boldsymbol{v}\cdot\boldsymbol{n}_{12}$ on Γ_{12}. We now define $\pi_h^2\boldsymbol{v} := \boldsymbol{\Pi}_h^2\boldsymbol{w}$, where $\boldsymbol{\Pi}_h^2 : \boldsymbol{X}^2 \cap (H^\theta(\Omega_2))^d \to \boldsymbol{X}_h^2$ is the MFE interpolant defined in (2.29). Note that, using (2.29),

$$\begin{aligned} b_2(\pi_h^2\boldsymbol{v}, q_h) &= b_2(\boldsymbol{\Pi}_h^2\boldsymbol{w}, q_h) = b_2(\boldsymbol{w}, q_h) \\ &= -\int_{\Omega_2}(\nabla\cdot\boldsymbol{w})q_h = -\int_{\Omega_2}(\nabla\cdot\boldsymbol{v})q_h = b_2(\boldsymbol{v}, q_h) \quad \forall q_h\in M_h^2, \end{aligned}$$

thus the so-constructed $\pi_h\boldsymbol{v} = (\pi_h^1\boldsymbol{v}, \pi_h^2\boldsymbol{v})$ satisfies

$$b(\pi_h\boldsymbol{v}-\boldsymbol{v}, q_h) = 0 \quad \forall q_h\in M_h.$$

It is easy to see that $\pi_h\boldsymbol{v}\in\boldsymbol{V}_h$. Indeed, for every $e\in\Gamma_h^{12}$ and $\eta\in\Lambda_h$, using (2.30) and the fact that $\Lambda_h = X_h^2\cdot\boldsymbol{n}_{12}$,

$$\int_e \pi_h^2\boldsymbol{v}\cdot\boldsymbol{n}_{12}\eta = \int_e \boldsymbol{\Pi}_h^2\boldsymbol{w}\cdot\boldsymbol{n}_{12}\eta = \int_e \boldsymbol{w}\cdot\boldsymbol{n}_{12}\eta = \int_e \pi_h^1\boldsymbol{v}\cdot\boldsymbol{n}_{12}\eta.$$

It remains to show the bound in (3.2). Using (2.31), (2.32), and (3.7),

$$\begin{aligned} \|\pi_h^2\boldsymbol{v}\|_{X^2} &= \|\boldsymbol{\Pi}_h^2\boldsymbol{w}\|_{X^2} \\ &\le \|\boldsymbol{\Pi}_h^2\boldsymbol{v}\|_{X^2} + \|\boldsymbol{\Pi}_h^2\boldsymbol{z}\|_{X^2} \\ &\le C(\|\boldsymbol{v}\|_{1,\Omega_2} + \|\boldsymbol{z}\|_{\theta,\Omega_2}) \\ &\le C(\|\boldsymbol{v}\|_{1,\Omega_1} + \|(\pi_h^1\boldsymbol{v}-\boldsymbol{v})\cdot\boldsymbol{n}\|_{\Gamma_{12}}) \end{aligned}$$

The last term can be bounded as follows. For every $e\in\Gamma_{12}$, and edge (face) of $E\in\mathcal{E}_h^1$, using (2.34) and (2.24),

$$\|(\pi_h^1\boldsymbol{v}-\boldsymbol{v})\cdot\boldsymbol{n}_{12}\|_e \le C(h_E^{-1/2}\|\pi_h^1\boldsymbol{v}-\boldsymbol{v}\|_{0,E} + h_E^{1/2}|\pi_h^1\boldsymbol{v}-\boldsymbol{v}|_{1,E}) \le Ch_E^{1/2}|\boldsymbol{v}|_{1,\delta(E)}. \tag{3.8}$$

Therefore

$$\|\pi_h^2\boldsymbol{v}\|_{X^2} \le C\|\boldsymbol{v}\|_{1,\Omega},$$

which, combined with (3.3), implies the bound in (3.2). Now using (3.2),

$$\frac{1}{C}\|q_h\|_M \le \frac{b(\boldsymbol{v}, q_h)}{\|\boldsymbol{v}\|_{1,\Omega}} = \frac{b(\pi_h\boldsymbol{v}, q_h)}{\|\boldsymbol{v}\|_{1,\Omega}} \le \frac{b(\pi_h\boldsymbol{v}, q_h)}{\frac{1}{C}\|\pi_h\boldsymbol{v}\|_X} \quad \forall q_h\in M_h,$$

which proves (3.1). ∎

4. A priori error estimates. In this section, optimal error estimates in the energy norm are obtained for the velocity field. Also, optimal error estimates in the L^2 norm of the error for the pressure are obtained. We start with an approximation result for the weakly normal-continuous velocity space $\boldsymbol{V}_h$.

LEMMA 4.1. *For $\boldsymbol{v} \in (H^1(\Omega))^d$ such that $\boldsymbol{v}|_{\Omega_1} \in (H^{k_1+1}(\Omega_1))^d$, $\boldsymbol{v}|_{\Omega_2} \in (H^{k_2+1}(\Omega_2))^d$, and $\nabla \cdot \boldsymbol{v}|_{\Omega_2} \in (H^{l_2+1}(\Omega_2))^d$, there exists $\tilde{\boldsymbol{v}} \in \boldsymbol{V}_h$ such that*

$$b(\boldsymbol{v} - \tilde{\boldsymbol{v}}, q) = 0 \quad \forall q \in M_h, \tag{4.1}$$

$$\forall e \in \Gamma_h^1 \cup \Gamma_1, \ \int_e [\tilde{\boldsymbol{v}}] \cdot \boldsymbol{q} = 0 \quad \forall \boldsymbol{q} \in (\mathbb{P}_{k_1-1}(e))^d, \tag{4.2}$$

$$\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|_X \le C\{h_1^{k_1}|\boldsymbol{v}|_{k_1+1,\Omega_1} + h_2^{k_2+1}|\boldsymbol{v}|_{k_2+1,\Omega_2} + h_2^{l_2+1}|\nabla \cdot \boldsymbol{v}|_{l_2+1,\Omega_2}\}. \tag{4.3}$$

Proof. We will show that the interpolant $\pi_h \boldsymbol{v}$ constructed in Theorem 3.1 satisfies the above conditions. Indeed, (4.1) and (4.2) follow directly from the construction of $\pi_h \boldsymbol{v}$. To show (4.3), we first note that (2.26) implies that

$$\|\boldsymbol{v} - \pi_h \boldsymbol{v}\|_{X^1} \le C h_1^{k_1} |\boldsymbol{v}|_{k_1+1,\Omega_1}. \tag{4.4}$$

Next,

$$\|\boldsymbol{v} - \pi_h \boldsymbol{v}\|_{X^2} = \|\boldsymbol{v} - \boldsymbol{\Pi}_h^2 \boldsymbol{w}\|_{X^2} \le \|\boldsymbol{v} - \boldsymbol{\Pi}_h^2 \boldsymbol{v}\|_{X^2} + \|\boldsymbol{\Pi}_h^2(\boldsymbol{w} - \boldsymbol{v})\|_{X^2}. \tag{4.5}$$

For the first term on the right in (4.5), using (2.31) and (2.32),

$$\|\boldsymbol{v} - \boldsymbol{\Pi}_h^2 \boldsymbol{v}\|_{X^2} \le C h_2^{k_2+1} |\boldsymbol{v}|_{k_2+1,\Omega_2} + h_2^{l_2+1} |\nabla \cdot \boldsymbol{v}|_{l_2+1,\Omega_2}. \tag{4.6}$$

The last term in (4.5) can be bounded as follows, using (2.33), (3.7), (3.8), and (2.24):

$$\begin{aligned} \|\boldsymbol{\Pi}_h^2(\boldsymbol{w} - \boldsymbol{v})\|_{X^2} &= \|\boldsymbol{\Pi}_h^2 \boldsymbol{z}\|_{X^2} \le \|\boldsymbol{z}\|_{\theta,\Omega_2} \\ &\le C\|(\pi_h^1 \boldsymbol{v} - \boldsymbol{v}) \cdot \boldsymbol{n}_{12}\|_{0,\Gamma_{12}} \le C h_1^{k_1+1/2} |\boldsymbol{v}|_{k_1+1,\Omega_1}. \end{aligned} \tag{4.7}$$

A combination of (4.4)–(4.7) completes the proof. ∎

THEOREM 4.2. *Let $(\boldsymbol{u}, p) \in \boldsymbol{X} \times M$ be the solution of the coupled problem (2.1)–(2.10). Assume that $\boldsymbol{u}|_{\Omega_i} \in H^{k_i+1}(\Omega_i)$ for $i = 1, 2$. Assume that $p|_{\Omega_1} \in H^{k_1}(\Omega_1)$ and that $p|_{\Omega_2} \in H^{l_2+1}(\Omega_2)$. Assume that Hypothesis A holds. Let $(\boldsymbol{U}, P)$ be the discrete solution of (2.38)–(2.39) Then, the following estimate holds:*

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{U}\|_X \le{}& C h_1^{k_1}(|\boldsymbol{u}|_{k_1+1,\Omega_1} + |p|_{k_1,\Omega_1}) + C h_2^{k_2+1} |\boldsymbol{u}|_{k_2+1,\Omega_2} \\ &+ C(h_2^{l_2+1} + h_2^{l_2+1/2} h_1^{1/2}) |p|_{l_2+1,\Omega_2}. \end{aligned}$$

Proof. Let $\tilde{\boldsymbol{u}}$ be the interpolant of $\boldsymbol{u}$ defined in Lemma 4.1 and let $\tilde{p}$ be the interpolant of p, satisfying (2.17)–(2.20). From (2.40), (2.41), and (2.38)–(2.39), the error equation is

$$\begin{aligned} a(\boldsymbol{U} - \tilde{\boldsymbol{u}}, \boldsymbol{v}) + b(\boldsymbol{v}, P - \tilde{p}) &= a(\boldsymbol{u} - \tilde{\boldsymbol{u}}, \boldsymbol{v}) + b(\boldsymbol{v}, p - \tilde{p}) \\ &\quad - \sum_{e \in \Gamma_{12}} \int_e p_2(\boldsymbol{v}_1 - \boldsymbol{v}_2) \cdot \boldsymbol{n}_{12} \quad \forall \boldsymbol{v} \in \boldsymbol{V}_h, \end{aligned} \tag{4.8}$$

$$b(\boldsymbol{U} - \tilde{\boldsymbol{u}}, q) = b(\boldsymbol{u} - \tilde{\boldsymbol{u}}, q) \quad \forall q \in M_h. \tag{4.9}$$

Note that (4.1) implies that $b(\boldsymbol{U}-\tilde{\boldsymbol{u}},q)=0$ for all $q\in M_h$, which implies that

$$\nabla\cdot(\boldsymbol{U}-\tilde{\boldsymbol{u}})=0 \quad \text{in } \Omega_2,$$

since $\nabla\cdot\boldsymbol{X}_h^2=M_h^2$. Define $\boldsymbol{\chi}=\boldsymbol{U}-\tilde{\boldsymbol{u}}$ and $\xi=P-\tilde{p}$. Choose $\boldsymbol{v}=\boldsymbol{\chi}$ and $q=\xi$. Then,

$$a(\boldsymbol{\chi},\boldsymbol{\chi})+b(\boldsymbol{\chi},\xi)=a(\boldsymbol{u}-\tilde{\boldsymbol{u}},\boldsymbol{\chi})+b(\boldsymbol{\chi},p-\tilde{p})-\sum_{e\in\Gamma_{12}}\int_e p_2(\boldsymbol{\chi}_1-\boldsymbol{\chi}_2)\cdot\boldsymbol{n}_{12},$$
$$b(\boldsymbol{\chi},\xi)=0.$$

Equivalently,

$$a(\boldsymbol{\chi},\boldsymbol{\chi})=a(\boldsymbol{u}-\tilde{\boldsymbol{u}},\boldsymbol{\chi})+b(\boldsymbol{\chi},p-\tilde{p})-\sum_{e\in\Gamma_{12}}\int_e p_2(\boldsymbol{\chi}_1-\boldsymbol{\chi}_2)\cdot\boldsymbol{n}_{12}. \tag{4.10}$$

The first term on the right can be estimated as follows:

$$\begin{aligned}a_1(\boldsymbol{u}-\tilde{\boldsymbol{u}},\boldsymbol{\chi})&=2\mu\sum_{E\in\mathcal{E}_h^1}\int_E\boldsymbol{D}(\boldsymbol{u}-\tilde{\boldsymbol{u}}):\boldsymbol{D}(\boldsymbol{\chi})\\&\quad-2\mu\sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e\{\boldsymbol{D}(\boldsymbol{u}-\tilde{\boldsymbol{u}})\}\boldsymbol{n}_e\cdot[\boldsymbol{\chi}]+2\mu\epsilon\sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e\{\boldsymbol{D}(\boldsymbol{\chi})\}\boldsymbol{n}_e\cdot[\boldsymbol{u}-\tilde{\boldsymbol{u}}]\\&\quad+\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\int_e[\boldsymbol{u}-\tilde{\boldsymbol{u}}]\cdot[\boldsymbol{\chi}]+\frac{\mu}{G}\sum_{e\in\Gamma_{12}}\int_e(\boldsymbol{u}-\tilde{\boldsymbol{u}})\cdot\boldsymbol{\tau}_{12}\boldsymbol{\chi}\cdot\boldsymbol{\tau}_{12}\\&=T_1+\cdots+T_5.\end{aligned}$$

Using Cauchy–Schwarz inequality, and the approximation result (4.3), we have

$$\begin{aligned}T_1&\le 2\mu\sum_{E\in\mathcal{E}_h^1}\|\nabla(\boldsymbol{u}-\tilde{\boldsymbol{u}})\|_{0,E}\|\nabla\boldsymbol{\chi}\|_{0,E}\le\frac{1}{8}|||\nabla\boldsymbol{\chi}|||_{0,\Omega_1}^2+C|||\nabla(\boldsymbol{u}-\tilde{\boldsymbol{u}})|||_{0,\Omega_1}^2\\&\le\frac{1}{8}|||\nabla\boldsymbol{\chi}|||_{0,\Omega_1}^2+Ch_1^{2k_1}|\boldsymbol{u}|_{k_1+1,\Omega_1}^2.\end{aligned}$$

Let $L_h(\boldsymbol{u})$ denote the standard Lagrange interpolant of degree k_1 defined in Ω_1 and let us insert it in the second integral term. Note that $L_h(\boldsymbol{u})$ satisfies the optimal error estimates

$$|L_h(\boldsymbol{u})-\boldsymbol{u}|_{m,E}\le Ch_E^{s-m}|\boldsymbol{u}|_{s,E}\quad\forall 2\le s\le k_1+1,\ m=0,1,2. \tag{4.11}$$

For e a segment of $\Gamma_h^1\cup\Gamma_1$, we have

$$\int_e\{\boldsymbol{D}(\boldsymbol{u}-\tilde{\boldsymbol{u}})\}\boldsymbol{n}_e\cdot[\boldsymbol{\chi}]=\int_e\{\boldsymbol{D}(\boldsymbol{u}-L_h(\boldsymbol{u}))\}\boldsymbol{n}_e\cdot[\boldsymbol{\chi}]+\int_e\{\boldsymbol{D}(L_h(\boldsymbol{u})-\tilde{\boldsymbol{u}})\}\boldsymbol{n}_e\cdot[\boldsymbol{\chi}].$$

Expanding the first integral, we obtain from the trace inequality (2.35) and from the

fact that the Lagrange interpolant satisfies (4.11)

$$\begin{aligned}
&\sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e\{\boldsymbol{D}(\boldsymbol{u}-L_h(\boldsymbol{u}))\}\boldsymbol{n}_e\cdot[\boldsymbol{\chi}]\\
&\quad\le\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e^{1/2}}{|e|^{1/2}}\|[\boldsymbol{\chi}]\|_{0,e}\frac{|e|^{1/2}}{\sigma_e^{1/2}}\|\{\boldsymbol{D}(\boldsymbol{u}-L_h(\boldsymbol{u}))\}\boldsymbol{n}_e\|_{0.e}\\
&\quad\le\frac{1}{8}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{\chi}]\|_{0,e}^2+C\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{|e|}{\sigma_e}(h_e^{-1}|\boldsymbol{u}-L_h(\boldsymbol{u})|_{1,E_e^{12}}^2+h_e|\boldsymbol{u}-L_h(\boldsymbol{u})|_{2,E_e^{12}}^2)\\
&\quad\le\frac{1}{8}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{\chi}]\|_{0,e}^2+Ch_1^{2k_1}|\boldsymbol{u}|_{k_1+1,\Omega_1}^2.
\end{aligned}$$

Similarly, using the trace inequality (2.36), triangle inequality, and (4.3)

$$\begin{aligned}
\sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e\{\boldsymbol{D}(L_h(\boldsymbol{u})-\tilde{\boldsymbol{u}})\}\boldsymbol{n}_e\cdot[\boldsymbol{\chi}]&\le\frac{1}{8}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{\chi}]\|_{0,e}^2\\
&\quad+C\sum_{e\in\Gamma_h^1\cup\Gamma_1}|\tilde{\boldsymbol{u}}-L_h(\boldsymbol{u})|_{1,E_e^{12}}^2\\
&\le\frac{1}{8}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{\chi}]\|_{0,e}^2+Ch_1^{2k_1}|\boldsymbol{u}|_{k_1+1,\Omega_1}^2.
\end{aligned}$$

Therefore,

$$T_2\le\frac{1}{4}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{\chi}]\|_{0,e}^2+Ch_1^{2k_1}|\boldsymbol{u}|_{k_1+1,\Omega_1}^2.$$

The third term vanishes because of the continuity of $\boldsymbol{u}$ and property (4.2) of $\tilde{\boldsymbol{u}}$:

$$T_3=0. \tag{4.12}$$

Using Cauchy–Schwarz inequality, the jump term is bounded by virtue of (2.24) and (2.34):

$$\begin{aligned}
T_4&\le\frac{1}{8}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{\chi}]\|_{0,e}^2+C\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{u}-\tilde{\boldsymbol{u}}]\|_{0,e}^2\\
&\le\frac{1}{8}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{\chi}]\|_{0,e}^2+Ch_1^{2k_1}|\boldsymbol{u}|_{k+1,\Omega_1}^2.
\end{aligned}$$

The last term is bounded as follows, from the trace inequality (2.34):

$$\begin{aligned}
T_5&\le\frac{\mu}{G}\sum_{e\in\Gamma_{12}}\|\boldsymbol{u}-\tilde{\boldsymbol{u}}\|_{0,e}\|\boldsymbol{\chi}\cdot\boldsymbol{\tau}_{12}\|_{0,e}\\
&\le\frac{\mu}{2G}\sum_{e\in\Gamma_{12}}\|\boldsymbol{\chi}\cdot\boldsymbol{\tau}_{12}\|_{0,e}^2+C\sum_{e\in\Gamma_1}(h_e^{-1}\|\boldsymbol{u}-\tilde{\boldsymbol{u}}\|_{0,E}^2+h_e|\boldsymbol{u}-\tilde{\boldsymbol{u}}|_{1,E}^2)\\
&\le\frac{\mu}{2G}\sum_{e\in\Gamma_{12}}\|\boldsymbol{\chi}\cdot\boldsymbol{\tau}_{12}\|_{0,e}^2+Ch_1^{2k_1}|\boldsymbol{u}|_{k_1+1,\Omega_1}^2.
\end{aligned}$$

Let us now estimate $a_2(\boldsymbol{u}-\tilde{\boldsymbol{u}},\boldsymbol{\chi})$, using the result (4.3),

$$a_2(\boldsymbol{u}-\tilde{\boldsymbol{u}},\boldsymbol{\chi}) = \int_{\Omega_2} \boldsymbol{K}^{-1}(\boldsymbol{u}-\tilde{\boldsymbol{u}})\cdot\boldsymbol{\chi} \le \frac{1}{8}\|\boldsymbol{K}^{-1/2}\boldsymbol{\chi}\|^2_{0,\Omega_2} + h_2^{2k_2+2}|\boldsymbol{u}|^2_{k_2+1,\Omega_2}.$$

Let us now estimate $b_1(\boldsymbol{\chi}, p-\tilde{p})$. By property (2.17), (2.19), and the trace estimate (2.34),

$$\begin{aligned} b_1(\boldsymbol{\chi}, p-\tilde{p}) &= -\sum_{E\in\mathcal{E}_h}\int_E (p-\tilde{p})\nabla\cdot\boldsymbol{\chi} + \sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e \{p-\tilde{p}\}[\boldsymbol{\chi}]\cdot\boldsymbol{n}_e \\ &= \sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e \{p-\tilde{p}\}[\boldsymbol{\chi}]\cdot\boldsymbol{n}_e \\ &\le \frac{1}{8}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\int_e [\boldsymbol{\chi}]^2 + Ch_1^{2k_1}|p|^2_{k_1,\Omega_1}. \end{aligned}$$

Now estimate $b_2(\boldsymbol{\chi}, p-\tilde{p})$ using Cauchy–Schwarz inequality and approximation result (2.20)

$$b_2(\boldsymbol{\chi}, p-\tilde{p}) = -\int_{\Omega_2}(p-\tilde{p})\nabla\cdot\boldsymbol{\chi} \le \frac{1}{8}|||\nabla\boldsymbol{\chi}|||^2_{0,\Omega_2} + Ch_2^{2l_2+2}|p|^2_{l_2+1,\Omega_2}$$

It remains to bound the last term in (4.10). Since $\boldsymbol{\chi}$ belongs to $\boldsymbol{V}_h$, we have

$$\sum_{e\in\Gamma_{12}}\int_e p_2(\boldsymbol{\chi}_1-\boldsymbol{\chi}_2)\cdot\boldsymbol{n}_{12} = \sum_{e\in\Gamma_{12}}\int_e (p_2-\tilde{p}_2^e)(\boldsymbol{\chi}_1-\boldsymbol{\chi}_2)\cdot\boldsymbol{n}_{12},$$

where $\tilde{p}_2^e\in\Lambda_h$ is the L^2 projection of p_2 with respect to the L^2 inner product on the edge e. Therefore, by definition of the projection and since $\Lambda_h = \boldsymbol{X}_h^2\cdot\boldsymbol{n}_{12}$, we have

$$\sum_{e\in\Gamma_{12}}\int_e (p_2-\tilde{p}_2^e)\boldsymbol{\chi}_2\cdot\boldsymbol{n}_{12} = 0.$$

We also note that for any edge e and any constant vector $\boldsymbol{c}_e$, we have

$$\begin{aligned} \sum_{e\in\Gamma_{12}}\int_e (p_2-\tilde{p}_2^e)\boldsymbol{\chi}_1\cdot\boldsymbol{n}_{12} &= \sum_{e\in\Gamma_{12}}\int_e (p_2-\tilde{p}_2^e)(\boldsymbol{\chi}_1-\boldsymbol{c}_e)\cdot\boldsymbol{n}_{12} \\ &\le \sum_{e\in\Gamma_{12}}\|p_2-\tilde{p}_2^e\|_{0,e}\|\boldsymbol{\chi}_1-\boldsymbol{c}_e\|_{0,e}. \end{aligned}$$

Assume that each edge e of Γ_{12} is shared by the element $E_e^2\in\mathcal{E}_h^2$ and parts of the elements $E_{e,i}^1\in\mathcal{E}_h^1$, $i=1,n_e$. Then, from the approximation properties and the trace inequality (2.34), we obtain

$$\int_e (p_2-\tilde{p}_2^e)\boldsymbol{\chi}_1\cdot\boldsymbol{n}_{12} \le Ch_2^{l_2+1/2}\|p_2\|_{l_2+1,E_e^2}\sum_{i=1}^{n_e}(h_1^{-1/2}\|\boldsymbol{\chi}_1-\boldsymbol{c}_e\|_{0,E_{e,i}^1} + h_1^{1/2}\|\nabla\boldsymbol{\chi}_1\|_{0,E_{e,i}^1}),$$

thus

$$\begin{aligned} \sum_{e\in\Gamma_{12}}\int_e (p_2-\tilde{p}_2^e)\boldsymbol{\chi}_1\cdot\boldsymbol{n}_{12} &\le C\sum_{e\in\Gamma_{12}} h_2^{l_2+1/2}|p_2|_{l_2+1,E_e^2}\sum_{i=1}^{n_e} h_1^{1/2}\|\nabla\boldsymbol{\chi}_1\|_{0,E_{e,i}^1} \\ &\le \frac{1}{8}|||\nabla\boldsymbol{\chi}|||^2_{0,\Omega_1} + Ch_2^{2l_2+1}h_1|p_2|^2_{l_2+1,\Omega_2}. \end{aligned}$$

Combining all bounds above yields

$$a(\boldsymbol{\chi},\boldsymbol{\chi}) \le \frac{1}{4}|||\nabla\boldsymbol{\chi}|||^2_{0,\Omega_1} + \frac{3}{4}\sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\|[\boldsymbol{\chi}]\|^2_{0,e} + \frac{\mu}{2G}\sum_{e\in\Gamma_{12}}\|\boldsymbol{\chi}\cdot\boldsymbol{\tau}_{12}\|^2_{0,e}$$
$$+\frac{1}{4}\|\boldsymbol{K}^{-1/2}\boldsymbol{\chi}\|^2_{0,\Omega_2} + Ch_2^{2k_2+2}|\boldsymbol{u}|_{k_2+1,\Omega_2} + C(h_2^{2l_2+2}+h_2^{2l_2+1}h_1)|p|^2_{l_2+1,\Omega_2}$$
$$+Ch_2^{2k_1}|\boldsymbol{u}|^2_{k_1+1,\Omega_1} + Ch_1^{2k_1}|p|^2_{k_1,\Omega_1}.$$

Equivalently,

$$a(\boldsymbol{\chi},\boldsymbol{\chi}) \le Ch_2^{2k_2+2}|\boldsymbol{u}|^2_{k_2+1,\Omega_2} + C(h_2^{2l_2+2}+h_2^{2l_2+1}h_1)|p|^2_{l_2+1,\Omega_2}$$
$$+Ch_1^{2k_1}(|\boldsymbol{u}|^2_{k_1+1,\Omega_1} + |p|^2_{k_1,\Omega_1}).$$

Now, since $\nabla\cdot\boldsymbol{\chi}=0$ in Ω_2, the coercivity Lemma 2.6 implies

$$\|\boldsymbol{u}-\boldsymbol{U}\|_X \le \|\boldsymbol{u}-\tilde{\boldsymbol{u}}\|_X + \|\boldsymbol{U}-\tilde{\boldsymbol{u}}\|_X$$
$$\le \|\boldsymbol{u}-\tilde{\boldsymbol{u}}\|_X + \frac{1}{\sqrt{C_0}}a(\boldsymbol{\chi},\boldsymbol{\chi})^{1/2}$$

which concludes the proof, using (4.3). ∎

THEOREM 4.3. *Under the assumptions and notation of Theorem* 4.2*, we have*

$$\|p-P\|_{0,\Omega} \le Ch_1^{k_1}(|\boldsymbol{u}|_{k_1+1,\Omega_1} + |p|_{k_1,\Omega_1}) + Ch_2^{k_2+1}|\boldsymbol{u}|_{k_2+1,\Omega_2}$$
$$+C(h_2^{l_2+1} + h_2^{l_2+1/2}h_1^{1/2})|p|_{l_2+1,\Omega_2},$$

where C is a constant independent of h_1, h_2.

Proof. The error equation (4.8) can be written as

$$\forall\boldsymbol{v}\in\boldsymbol{V}_h,\ \ a(\boldsymbol{U}-\boldsymbol{u},\boldsymbol{v}) + b(\boldsymbol{v},P-\tilde{p}) = b(\boldsymbol{v},p-\tilde{p}) - \sum_{e\in\Gamma_{12}}\int_e p_2(\boldsymbol{v}_1-\boldsymbol{v}_2)\cdot\boldsymbol{n}_{12}. \tag{4.13}$$

From the discrete inf-sup condition (3.1),

$$\|P-\tilde{p}\|_{0,\Omega} \le \frac{1}{\beta}\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\frac{b(\boldsymbol{v}_h,P-\tilde{p})}{\|\boldsymbol{v}_h\|_X}. \tag{4.14}$$

Using (4.13), for any $\boldsymbol{v}_h\in\boldsymbol{V}_h$,

$$b(\boldsymbol{v}_h,P-\tilde{p}) = -a(\boldsymbol{U}-\boldsymbol{u},\boldsymbol{v}_h) + b(\boldsymbol{v}_h,p-\tilde{p}) - \sum_{e\in\Gamma_{12}}\int_e p_2(\boldsymbol{v}_{h1}-\boldsymbol{v}_{h2})\cdot\boldsymbol{n}_{12}.$$

For the first term on the right,

$$a(\boldsymbol{U}-\boldsymbol{u},\boldsymbol{v}_h) = 2\mu\sum_{E\in\mathcal{E}_h^1}\int_E \boldsymbol{D}(\boldsymbol{U}-\boldsymbol{u}):\boldsymbol{D}(\boldsymbol{v}_h) + \sum_{e\in\Gamma_h^1\cup\Gamma_1}\frac{\sigma_e}{|e|}\int_e[\boldsymbol{U}-\boldsymbol{u}]\cdot[\boldsymbol{v}_h]$$
$$-2\mu\sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e\{\boldsymbol{D}(\boldsymbol{U}-\boldsymbol{u})\boldsymbol{n}_e\}\cdot[\boldsymbol{v}_h] + 2\mu\epsilon\sum_{e\in\Gamma_h^1\cup\Gamma_1}\int_e\{\boldsymbol{D}(\boldsymbol{v}_h)\boldsymbol{n}_e\}\cdot[\boldsymbol{U}-\boldsymbol{u}]$$
$$+\frac{\mu}{G}\sum_{e\in\Gamma_{12}}\int_e(\boldsymbol{U}-\boldsymbol{u})\cdot\boldsymbol{\tau}_{12}\boldsymbol{v}_h\cdot\boldsymbol{\tau}_{12} + \int_{\Omega_2}\boldsymbol{K}^{-1}(\boldsymbol{U}-\boldsymbol{u})\cdot\boldsymbol{v}_h$$
$$= Q_1+\cdots+Q_6.$$

We now bound each Q_i term. From Cauchy–Schwarz inequality, the terms Q_1, Q_2, Q_5, and Q_6 are easily bounded

$$Q_1 + Q_2 + Q_5 + Q_6 \leq C\|\boldsymbol{v}_h\|_X \|\boldsymbol{U} - \boldsymbol{u}\|_X.$$

We now bound Q_3,

$$\begin{aligned} Q_3 &\leq C \sum_{e\in\Gamma_h^1\cup\Gamma_1} \left(\frac{|e|}{\sigma_e}\right)^{1/2} \|\nabla(\boldsymbol{U}-\boldsymbol{u})\|_{0,e} \left(\frac{\sigma_e}{|e|}\right)^{1/2} \|[\boldsymbol{v}_h]\|_{0,e} \\ &\leq C\|\boldsymbol{v}_h\|_X \left(\sum_{e\in\Gamma_h^1\cup\Gamma_1} (h_1\|\nabla(\boldsymbol{U}-\tilde{\boldsymbol{u}})\|_{0,e}^2 + h_1\|\nabla(\boldsymbol{u}-\tilde{\boldsymbol{u}})\|_{0,e}^2) \right)^{1/2} \\ &\leq C\|\boldsymbol{v}_h\|_X (\|U-\tilde{\boldsymbol{u}}\|_X^2 + Ch_1^{2k_1}|\boldsymbol{u}|_{k_1+1,\Omega_1}^2)^{1/2}. \end{aligned}$$

Now, Q_4 is bounded similarly, from trace inequality (2.36),

$$\begin{aligned} Q_4 &\leq C \sum_{e\in\Gamma_h^1\cup\Gamma_1} \|\{\boldsymbol{D}(\boldsymbol{v}_h)\boldsymbol{n}_e\}\|_{0,e} \|[\boldsymbol{U}-\boldsymbol{u}]\|_{0,e} \\ &\leq C \sum_{e\in\Gamma_h^1\cup\Gamma_1} h^{-1/2}\|\nabla\boldsymbol{v}_h\|_{0,E_e^{12}} \left(\frac{\sigma_e}{|e|}\right)^{1/2-1/2} \|[\boldsymbol{U}-\boldsymbol{u}]\|_{0,e} \\ &\leq C\|\boldsymbol{v}_h\|_X \|\boldsymbol{U}-\boldsymbol{u}\|_X. \end{aligned}$$

Let us now estimate $b(\boldsymbol{v}_h, p-\tilde{p})$. From the property (2.17), it is reduced to

$$\begin{aligned} b(\boldsymbol{v}_h, p-\tilde{p}) &= \sum_{e\in\Gamma_h^1\cup\Gamma_1} \int_e \{p-\tilde{p}\}[\boldsymbol{v}_h]\cdot\boldsymbol{n}_e \\ &\leq \sum_{e\in\Gamma_h^1\cup\Gamma_1} \left(\frac{\sigma_e}{|e|}\right)^{1/2} \|[\boldsymbol{v}_h]\|_{0,e} \left(\frac{|e|}{\sigma_e}\right)^{1/2} \|\{p-\tilde{p}\}\|_{0,e} \\ &\leq \|\boldsymbol{v}_h\|_X Ch_1^{k_1}|p|_{k_1,\Omega_1}. \end{aligned}$$

Finally, following the same approach as in the proof of Theorem 4.2, we bound the interface integral

$$\begin{aligned} \sum_{e\in\Gamma_{12}} \int_e p_2(\boldsymbol{v}_{h1}-\boldsymbol{v}_{h2})\cdot\boldsymbol{n}_{12} &= \sum_{e\in\Gamma_{12}} \int_e (p_2-\tilde{p}_2^e)\boldsymbol{v}_{h1}\cdot\boldsymbol{n}_{12} \\ &\leq C\|\boldsymbol{v}_h\|_X h_2^{l_2+1/2} h_1^{1/2} |p_2|_{l_2+1,\Omega_2}. \end{aligned}$$

Combining all the bounds with (4.14) yields

$$\|P-\tilde{p}\|_{0,\Omega} \leq C\left(\|\boldsymbol{U}-\boldsymbol{u}\|_X + h_1^{k_1}(|\boldsymbol{u}|_{k_1+1,\Omega_1} + |p|_{k_1,\Omega_1}) + h_2^{l_2+1/2}h_1^{1/2}\|p\|_{l_2+1,\Omega_2}\right).$$

Using Theorem 4.2 concludes the proof. ∎

Remark 4.4. The results proven in this section are valid and unchanged in three-dimensional domains, assuming there exist interpolants $\boldsymbol{\Pi}_h^1$ and $\boldsymbol{\Pi}_h^2$ defined in (2.21) and (2.29). The existence of $\boldsymbol{\Pi}_h^1$ for $k=1$ in three dimensions is given in [13]. The existence of $\boldsymbol{\Pi}_h^2$ in any dimension is a well-known fact [10].

5. Implementation issues and conclusions. In this paper, the convergence of a numerical scheme for solving the coupled Darcy–Stokes problem is proved. In order to parallelize the implementation of the scheme, a Lagrange multiplier $\lambda \in \Lambda_h$ approximating p_2 on Γ_{12} can be introduced. We recall the definition of $\Lambda_h = \boldsymbol{X}_h^2 \cdot \boldsymbol{n}_{12}$ given in section 2. Defining the bilinear form on the interface,

$$\Lambda(\eta, \boldsymbol{v}) = \sum_{e \in \Gamma_{12}} \int_e \eta(\boldsymbol{v}_1 - \boldsymbol{v}_2) \cdot \boldsymbol{n}_{12} \quad \forall \eta \in \Lambda_h, \ \forall \boldsymbol{v} \in \boldsymbol{X}_h,$$

the scheme can be rewritten as: Find $(\boldsymbol{U}, P, \lambda) \in \boldsymbol{X}_h \times M_h \times \Lambda_h$ such that $\boldsymbol{U}_i = \boldsymbol{U}|_{\Omega_i}$ and $P_i = P|_{\Omega_i}$ satisfy

$$a_1(\boldsymbol{U}_1, \boldsymbol{v}_1) + b_1(\boldsymbol{v}_1, P_1) + \Lambda(\lambda, \boldsymbol{v}_1) = \int_{\Omega_1} \boldsymbol{f}_1 \cdot \boldsymbol{v}_1 \quad \forall \boldsymbol{v}_1 \in \boldsymbol{X}_h^1, \tag{5.1}$$

$$b_1(\boldsymbol{U}_1, q_1) = 0 \quad \forall q_1 \in M_h^1, \tag{5.2}$$

$$a_2(\boldsymbol{U}_2, \boldsymbol{v}_2) + b_2(\boldsymbol{v}_2, P_2) - \Lambda(\lambda, \boldsymbol{v}_2) = 0 \quad \forall \boldsymbol{v}_2 \in \boldsymbol{X}_h^2, \tag{5.3}$$

$$b_2(\boldsymbol{U}_2, q_2) = \int_{\Omega_2} f_2 q_2 \quad \forall q_2 \in M_h^2, \tag{5.4}$$

$$\Lambda(\eta, \boldsymbol{U}) = 0 \quad \forall \eta \in \Lambda_h. \tag{5.5}$$

It can easily be shown that the two discrete formulations are equivalent. Formulation (5.1)–(5.5) is suitable for a parallel implementation. In particular, using an approach from [23], a nonoverlapping domain decomposition algorithm can be formulated that reduces the coupled system to a symmetric and positive definite interface problem for λ. In addition to its parallel efficiency, this approach allows for existing codes solving the Stokes or the Darcy equations to be utilized.

REFERENCES

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

[2] T. Arbogast, L. C. Cowsar, M. F. Wheeler, and I. Yotov, *Mixed finite element methods on nonmatching multiblock grids*, SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.

[3] T. Arbogast and H. Lehr, *Homogenization of a Darcy-Stokes System Modeling Vuggy Porous Media*, Technical report 02-44, University of Texas at Austin, 2002.

[4] T. Arbogast, M. F. Wheeler, and I. Yotov, *Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences*, SIAM J. Numer. Anal., 34 (1997), pp. 828–852.

[5] G. Beavers and D. Joseph, *Boundary conditions at a naturally impermeable wall*, J. Fluid Mech., 30 (1967), pp. 197–207.

[6] S. Brenner, *Korn's inequalities for piecewise H^1 vector fields*, Math. Comp., 73 (2004), pp. 1067–1087.

[7] F. Brezzi, J. Douglas, Jr., R. Duràn, and M. Fortin, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.

[8] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modèl. Math. Anal. Numèr., 21 (1987), pp. 581–604.

[9] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 88 (1985), pp. 217–235.

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

[11] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[12] M. Crouzeix and R. Falk, *Non conforming finite elements for the Stokes problem*, Math. Comp., 52 (1989), pp. 437–456.

[13] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations.* I, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–75.

[14] C. Dawson, *Conservative, shock-capturing transport methods with nonconservative velocity approximations*, Comput. Geosci., 3 (1999), pp. 205–227.

[15] M. Discacciati, E. Miglio, and A. Quarteroni, *Mathematical and numerical models for coupling surface and groundwater flows*, Appl. Numer. Math., 43 (2002), pp. 57–74.

[16] J. Douglas, Jr., R. E. Ewing, and M. F. Wheeler, *The approximation of the pressure by a mixed method in the simulation of miscible displacement*, RAIRO Anal. Numér., 17 (1983), pp. 17–34.

[17] L. J. Durlofsky, *Accuracy of mixed and control volume finite element approximations to Darcy velocity and related quantities*, Water Resources Research, 30 (1994), pp. 965–973.

[18] R. E. Ewing, O. P. Iliev, and R. D. Lazarov, *Numerical Simulation of Contamination Transport Due to Flow in Liquid and Porous Media*, Technical report 1992-10, Enhanced Oil Recovery Institute, University of Wyoming, 1992.

[19] M. Fortin and M. Soulie, *A non-conforming piecewise quadratic finite element on triangles*, Internat. J. Numer. Methods Engrg., 19 (1983), pp. 505–520.

[20] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Vol.* I, Springer-Verlag, New York, 1994.

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

[22] V. Girault, B. Riviere, and M. Wheeler, *A discontinuous Galerkin method with non-overlapping domain decomposition for the Stokes and Navier-Stokes problems*, Math. Comp., 74 (2004), pp. 53–84.

[23] R. Glowinski and M. F. Wheeler, *Domain decomposition and mixed finite element methods for elliptic problems*, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Periaux, eds., SIAM, Philadelphia, 1988, pp. 144–172.

[24] W. Jäger and A. Mikelić, *On the interface boundary condition of Beavers, Joseph, and Saffman*, SIAM J. Appl. Math., 60 (2000), pp. 1111–1127.

[25] W. J. Layton, F. Schieweck, and I. Yotov, *Coupling fluid flow with porous media flow*, SIAM J. Numer. Anal., 40 (2003), pp. 2195–2218.

[26] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1972.

[27] K. A. Mardal, X.-C. Tai, and R. Winther, *A robust finite element method for Darcy–Stokes flow*, SIAM J. Numer. Anal., 40 (2002), pp. 1605–1631.

[28] T. P. Mathew, *Domain Decomposition and Iterative Refinement Methods for Mixed Finite Element Discretizations of Elliptic Problems*, Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University, 1989.

[29] R. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.

[30] B. Riviere and M. Wheeler, *Discontinuous Galerkin methods for flow and transport problems in porous media*, Comm. Numer. Methods Engrg., 18 (2002), pp. 63–68.

[31] B. Rivière and M. Wheeler, *Non conforming methods for transport with nonlinear reaction*, in Fluid Flow and Transport in Porous Media: Mathematical and Numerical Treatment, Z. Chen and R. Ewing, eds., Contemp. Math. 295, AMS, Providence, RI, 2002, pp. 421–430.

[32] B. Riviere and M. Wheeler, *A discontinuous Galerkin method for modeling two-phase flow*, Advances in Water Resources, submitted.

[33] B. Rivière, M. Wheeler, and K. Banas, *Part* II. *Discontinuous Galerkin method applied to a single phase flow in porous media*, Comput. Geosci., 4 (2000), pp. 337–349.

[34] T. F. Russell and M. F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media*, in The Mathematics of Reservoir Simulation, Frontiers Appl. Math. 1, R. E. Ewing, ed., SIAM, Philadelphia, 1984, pp. 35–106.

[35] P. Saffman, *On the boundary condition at the surface of a porous media*, Stud. Appl. Math., 50 (1971), pp. 292–315.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1978–1997

ON CONVERGENCE OF MINMOD-TYPE SCHEMES*

SERGEI KONYAGIN†, BOJAN POPOV‡, AND OGNIAN TRIFONOV§

Abstract. A class of nonoscillatory numerical methods for solving nonlinear scalar conservation laws in one space dimension is considered. This class of methods contains the classical Lax–Friedrichs and the second-order Nessyahu–Tadmor schemes. In the case of linear flux, new l_2 stability results and error estimates for the methods are proved. Numerical experiments confirm that these methods are one-sided l_2 stable for convex flux instead of the usual Lip+ stability.

Key words. conservation laws, minmod-type schemes, convergence

AMS subject classifications. 65M15, 65M12

DOI. 10.1137/S0036142903423861

1. Introduction.

We are interested in the scalar hyperbolic conservation law

$$\begin{cases} u_t + f(u)_x = 0, & (x,t) \in \mathbb{R} \times (0,\infty), \\ u(x,0) = u^0(x), & x \in \mathbb{R}, \end{cases} \tag{1.1}$$

where f is a given flux function. In recent years, there has been enormous activity in the development of the mathematical theory and in the construction of numerical methods for (1.1). Even though the existence-uniqueness theory of weak solutions is complete [12], there are many numerically efficient methods for which the questions of convergence and error estimates are still open. For example, there are many nonoscillatory schemes based on the minmod limiter which are numerically robust, at least in many numerical tests, but theoretical results about convergence and error estimates are still missing [3, 6, 7, 18].

In this paper, we consider a class of the so-called Godunov-type schemes for solving (1.1). There are two main steps in such schemes: evolution and projection. In the original Godunov scheme, the projection is onto piecewise constant functions—the cell averages. In the general Godunov-type method, the projection is onto piecewise polynomials. To determine the properties of these schemes it is necessary to study the properties of the projection operator. We limit our attention to the case of piecewise linear projection based on cell averages using minmod limiters for the slope reconstruction, and we call such a scheme *minmod-type*. For example, the Nessyahu–Tadmor scheme [15] is of minmod-type and is based on staggered evolution; other examples include the second-order nonoscillatory central schemes with nonstaggered grids given in [8, 9], and the UNO and TVD2 schemes in [6]. Theoretical results about convergence of such schemes to the entropy solution, or error estimates, are still missing. In most cases the authors give a variation bound for such a scheme, which is enough to conclude that the method converges to a weak solution; see [10]. The only paper which has a convergence result is the paper of Nessyahu and Tadmor [15] in

*Received by the editors February 26, 2003; accepted for publication (in revised form) April 18, 2004; published electronically January 20, 2005.
http://www.siam.org/journals/sinum/42-5/42386.html
†Department of Mathematics, Moscow State University, Moscow, Russia (konyagin@ok.ru). The research of this author was supported by grant NSh-304.2003.1.
‡Department of Mathematics, Texas A&M University, College Station, TX 77845 (popov@math.tamu.edu). The research of this author was supported by ONR grant N00014-91-J-1076.
§Department of Mathematics, University of South Carolina, Columbia, SC 29208 (trifonov@math.sc.edu). The research of this author was supported by NSF DMS grant 9970455.

which the authors prove a single cell entropy inequality for a minor modification of the original minmod scheme. A single entropy inequality is enough to conclude that the scheme is convergent to the unique entropy solution but does not give any rate of convergence. In order to get a rate, one has to have a family of entropy inequalities (see [1, 2, 11, 14]). Alternatively, for a convex flux, one can impose Lip+ stability on the projection and then prove convergence via Tadmor's Lip$'$ theory [16, 19]. Unfortunately, it is well known that minmod-type schemes are incompatible with the Lip+ condition—the Lip+ seminorm is not preserved by a minmod-type projection. It is easy to think about minmod-type schemes in terms of new/old cell averages. That is, we start with a sequence of cell averages $\{w_j\}$, and after one time step (projection and evolution) we get a new sequence $\{w'_j\}$. A scheme is *total variation diminishing* (TVD) if the variation of the new sequence $\sum_j |w'_j - w'_{j-1}|$ is not bigger than the variation of the old one $\sum_j |w_j - w_{j-1}|$, i.e., the l_1 norm of the jumps does not increase in time. In the Lip+ case (for convex flux) the condition on the jumps is that the biggest nonnegative jump does not increase in time:

$$\sup_j (w'_j - w'_{j-1})_+ \le \sup_j (w_j - w_{j-1})_+.$$

In section 3 of this paper, we prove that for linear flux the l_2 norm of the jumps for some minmod-type schemes does not increase in time. This class of schemes include the NT scheme and the TVD2 scheme considered in [6]. Based on that, we use the dual approach (see [16, 19]) to derive a new error estimate in L^2 in section 4. The rate of convergence that we prove is $1/2$ in L^2, which improves the known result of $1/4$ (see [19]). In section 5, we present numerical examples in the case of linear and convex flux and discuss the nonconvex case. Our numerical tests show that for convex flux the minmod schemes preserve the one-sided analogue

$$\sum_j (w'_j - w'_{j-1})_+^2 \le \sum_j (w_j - w_{j-1})_+^2,$$

which suggests a different approach to proving convergence and error estimates for such schemes in the convex case. The l_2 norm of the jumps is a natural candidate norm for the analysis of high-order schemes, such as central or ENO [7] type, due to its numerical viscosity. We view the results of this paper as a step toward obtaining convergence results and estimates for the rate of convergence of minmod-type schemes for solving (1.1) in the case of convex nonlinear flux.

2. Nonoscillatory central schemes. In this section, we are concerned with nonoscillatory central differencing approximations to the scalar conservation law

$$u_t + f(u)_x = 0. \tag{2.1}$$

The prototypes of all central schemes are the staggered form of the Lax–Friedrichs (LxF) scheme and its second-order extension, the Nessyahu–Tadmor (NT) scheme [15]. For an introduction to central schemes, see [8, 9, 13, 15]. For simplicity, we limit our attention to the staggered NT scheme described below. Let $v(x,t)$ be an approximate solution to (2.1), and assume that the space mesh Δx and the time mesh Δt are uniform. Let $x_j := j\Delta x$, $j \in \mathbb{Z}$, $\lambda := \frac{\Delta t}{\Delta x}$, and

$$v_j(t) := \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} v(x,t)\, dx \tag{2.2}$$

be the average of v at time t over $(x_{j-1/2}, x_{j+1/2})$. Let us assume that $v(\cdot,t)$ is a piecewise linear function, and that it is linear on the intervals $(x_{j-1/2}, x_{j+1/2})$, $j \in \mathbb{Z}$, of the form

$$v(x,t) = L_j(x,t) := v_j(t) + (x - x_j)\frac{1}{\Delta x}v'_j, \quad x_{j-1/2} < x < x_{j+1/2}, \tag{2.3}$$

where $\frac{1}{\Delta x}v'_j$ is the numerical derivative of v, which is yet to be determined. Integration of (2.1) over the staggered space-time cell $(x_j, x_{j+1}) \times (t, t+\Delta t)$ yields

$$\begin{aligned} v_{j+1/2}(t+\Delta t) = {} & \frac{1}{\Delta x}\left(\int_{x_j}^{x_{j+1/2}} L_j(x,t)\,dx + \int_{x_{j+1/2}}^{x_{j+1}} L_{j+1}(x,t)\,dx\right) \\ & - \frac{1}{\Delta x}\left(\int_t^{t+\Delta t} f(v(x_{j+1},\tau))\,d\tau - \int_t^{t+\Delta t} f(v(x_j,\tau))\,d\tau\right). \end{aligned} \tag{2.4}$$

The first two integrals on the right-hand side of (2.4) can be evaluated exactly. Moreover, if the CFL condition

$$\lambda \max_{x_j \le x \le x_{j+1}} |f'(v(x,t))| \le \frac{1}{2}, \quad j \in \mathbb{Z}, \tag{2.5}$$

is met, then the last two integrants on the right of (2.4) are smooth functions of τ. Hence, they can be integrated approximately by the midpoint rule with third-order local truncation error. Note that, in the case of zero slopes $\frac{1}{\Delta x}v'_j$ and $\frac{1}{\Delta x}v'_{j+1}$, the time integration is exact for any flux f, and even for nonzero slopes the time integration can be exact for a low degree polynomial flux. Thus, following [15], we arrive at

$$\begin{aligned} v_{j+1/2}(t+\Delta t) = {} & \frac{1}{2}(v_j(t) + v_{j+1}(t)) + \frac{1}{8}(v'_j - v'_{j+1}) \\ & - \lambda\left(f\left(v\left(x_{j+1}, t+\frac{\Delta t}{2}\right)\right) - f\left(v\left(x_j, t+\frac{\Delta t}{2}\right)\right)\right). \end{aligned} \tag{2.6}$$

By Taylor expansion and the conservation law (2.1), we obtain

$$v\left(x_j, t+\frac{\Delta t}{2}\right) = v_j(t) - \frac{1}{2}\lambda f'_j, \tag{2.7}$$

where $\frac{1}{\Delta x}f'_j$ stand for an approximate numerical derivative of the flux $f(v(x=x_j,t))$. The following choices are widely used as approximations of the numerical derivatives (we drop t to simplify the notation):

$$\begin{aligned} v'_j &= \mathrm{m}(v_{j+1} - v_j, v_j - v_{j-1}), \\ f'_j &= \mathrm{m}(f(v_{j+1}) - f(v_j), f(v_j) - f(v_{j-1})), \end{aligned} \tag{2.8}$$

where $\mathrm{m}(a,b)$ stands for the minmod limiter

$$\mathrm{m}(a,b) \equiv \mathrm{MinMod}(a,b) := \frac{1}{2}(\mathrm{sgn}(a) + \mathrm{sgn}(b)) \cdot \min(|a|, |b|) \tag{2.9}$$

with the usual generalization

$$\mathrm{m}(E) := \begin{cases} \inf(E) & \text{if} \quad E \subset \mathbb{R}_+, \\ \sup(E) & \text{if} \quad E \subset \mathbb{R}_-, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

A generalization of this numerical approximation is based the so-called minmod-θ limiters

$$v'_j = \mathrm{m}\left(\theta(v_{j+1}-v_j), \frac{1}{2}(v_{j+1}-v_{j-1}), \theta(v_j - v_{j-1})\right), \tag{2.11}$$
$$f'_j = \mathrm{m}\left(\theta(f(v_{j+1})-f(v_j)), \frac{1}{2}(f(v_{j+1})-f(v_{j-1})), \theta(f(v_j)-f(v_{j-1}))\right).$$

Given the approximate slopes and flux derivatives (2.11), we have a family of central schemes in the predictor-corrector form

$$v\left(x_j, t+\frac{\Delta t}{2}\right) = v_j(t) - \frac{1}{2}\lambda f'_j,$$
$$v_{j+1/2}(t+\Delta t) = \frac{1}{2}(v_j(t)+v_{j+1}(t)) + \frac{1}{8}(v'_j - v'_{j+1}) \tag{2.12}$$
$$- \lambda\left(f\left(v\left(x_{j+1}, t+\frac{\Delta t}{2}\right)\right) - f\left(v\left(x_j, t+\frac{\Delta t}{2}\right)\right)\right),$$

where we start with $v_j(0) := \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u_0(x)\,dx$. Note that we alternate between two uniform partitions of the real line: all intervals with integer end points for $t = 2k\Delta t$, $k \in \mathbb{Z}$, and half integers for $t = (2k+1)\Delta t$, $k \in \mathbb{Z}$. As a special case, we recover the staggered LxF scheme for $\theta = 0$ and the basic minmod scheme for $\theta = 1$ (the middle slope in the minmod limiter (2.11) drops if $\theta \le 1$).

3. l_2 stability for linear flux. In this section we will prove that the central scheme given in (2.12) is l_2 stable for any θ in the interval $[0,1]$. Based on this stability we will also derive a new error estimate in L^2 instead of the usual L^1 estimates in the conservation laws. Note that even for linear flux f, the minmod-type schemes are not linear and the only global property known is that the total variation does not increase in time under an appropriate CFL condition; see [15]. The class of minmod-type schemes is also not Lip+ stable except for the obvious choice $\theta = 0$. Let us consider a linear flux $f(u) = au$, uniform time steps $t_n = n\Delta t$, and restrict the minmod limiter to $\theta \le 1$. We denote $v_j^n := v_j(t_n)$, $\delta_j^n := v_j^n - v_{j-1}^n$. The minmod scheme (2.12) reduces to

$$v_{j+1/2}^{n+1} = \frac{1}{2}(v_j^n + v_{j+1}^n) + \frac{\theta}{8}\left(\mathrm{m}(\delta_j^n, \delta_{j+1}^n) - \mathrm{m}(\delta_{j+1}^n, \delta_{j+2}^n)\right) \tag{3.1}$$
$$- \frac{a\Delta t}{\Delta x}\left(v_{j+1}^n - \frac{a\Delta t}{2\Delta x}\theta\mathrm{m}(\delta_{j+1}^n, \delta_{j+2}^n) - v_j^n + \frac{a\Delta t}{2\Delta x}\theta\mathrm{m}(\delta_j^n, \delta_{j+1}^n)\right).$$

Hence, we have an explicit formula for the new cell averages (at time t_{n+1}) on a staggered grid in terms of the old ones (at time t_n) on a regular grid. In order to simplify the notation, we drop the time dependence and denote $w_j := v_j^n$, $w'_{j+1} := v_{j+1/2}^{n+1}$, $\delta'_j := w'_j - w'_{j-1}$, $\alpha := \frac{1}{2} + \frac{a\Delta t}{\Delta x}$, and $\beta := \frac{1}{2}\alpha(1-\alpha)$. With this notation, we have the following relation between the sequence of the new averages $\{w'_j\}$ and the old ones $\{w_j\}$:

$$w'_j = \alpha w_{j-1} + (1-\alpha)w_j + \theta\beta\left(\mathrm{m}(\delta_{j-1}, \delta_j) - \mathrm{m}(\delta_j, \delta_{j+1})\right). \tag{3.2}$$

Using that $\delta_j = w_j - w_{j-1}$, we derive the formula for the sequence of new jumps in terms of the old ones:

$$\delta'_j = \alpha\delta_{j-1} + (1-\alpha)\delta_j - \theta\beta\mathrm{m}(\delta_{j-2}, \delta_{j-1}) + 2\theta\beta\mathrm{m}(\delta_{j-1}, \delta_j) - \theta\beta\mathrm{m}(\delta_j, \delta_{j+1}). \tag{3.3}$$

The CFL condition (2.5) reduces to $0 \le \alpha \le 1$ because $\alpha = \frac{1}{2} + \frac{a\Delta t}{\Delta x}$ and $|\frac{a\Delta t}{\Delta x}| \le \frac{1}{2}$. The main result in this section is the following stability result.

THEOREM 3.1. *If the initial condition $u_0 \in L^1_{loc}(\mathbb{R})$, then the l_2 norm of the jumps of the approximate solution $v(\cdot, t)$ is nonincreasing in time. That is,*

$$\|\{\delta'_j\}\|_{l_2} \equiv \|\{v^{n+1}_j - v^{n+1}_{j-1}\}\|_{l_2} \le \|\{\delta_j\}\|_{l_2} \equiv \|\{v^n_j - v^n_{j-1}\}\|_{l_2} \tag{3.4}$$

for all $n \ge 1$.

Proof. It is clear that we have to prove the result for one step, assuming that $\|\{\delta_j\}\|_{l_2} < \infty$. We split the proof into two parts. First, we prove the stability for a monotone sequence $\{w_j\}$. By symmetry, it is sufficient to consider the case $\delta_j \ge 0$ for all $j \in \mathbb{Z}$. Then we apply that result locally to derive the l_2 stability for a general sequence.

THEOREM 3.2. *Let us assume that $\delta_j \ge 0$, $j \in \mathbb{Z}$, and δ'_j are given by* (3.3). *Then*

$$\|\{\delta'_j\}\|_{l_2} \le \|\{\delta_j\}\|_{l_2}. \tag{3.5}$$

Proof. Let us recall that $\{\delta_j\}_{-\infty}^{\infty} \in l_2$ and $\delta_j \ge 0$ for all j. It is enough to prove Theorem 3.2 only for $0 < \alpha < 1$. Let $\beta_1 := \theta\beta$; then $0 < \beta_1 \le \beta$. We construct the new sequence $\{\delta'_j\}$ by using the rule

$$\delta'_j = (1-\alpha)\delta_j + \alpha\delta_{j-1} - \beta_1 \min(\delta_{j-2}, \delta_{j-1}) + 2\beta_1 \min(\delta_{j-1}, \delta_j) - \beta_1 \min(\delta_j, \delta_{j+1}) \tag{3.6}$$

for each j. First we assume that $\{\delta_j\}$ has finite support. It is easy to see how to modify the proof in case the support is not finite. Therefore we assume $\delta_j = 0$ for $j \le 3$ and for $j \ge N-3$ for some integer N. Then $\delta'_j = 0$ for $j \le 3$ and $j \ge N-2$. Thus it suffices to prove

$$\sum_{j=1}^{N} \delta_j^2 \ge \sum_{j=1}^{N} (\delta'_j)^2. \tag{3.7}$$

Let us introduce some notation. Let $y_j = \min(\delta_j, \delta_{j+1})$, $\Delta\delta_j = \delta_j - \delta_{j-1}$, $\Delta y_j = y_j - y_{j-1}$, $\Delta^2\delta_j = \delta_j - 2\delta_{j-1} + \delta_{j-2}$, and $\Delta^2 y_j = y_j - 2y_{j-1} + y_{j-2}$. Then (3.6) becomes

$$\delta'_j = ((1-\alpha)\delta_j + \alpha\delta_{j-1}) - \beta_1 \Delta^2 y_j \quad \text{and}$$

$$\sum_{j=1}^{N} (\delta'_j)^2 = \sum_{j=1}^{N} \left(((1-\alpha)\delta_j + \alpha\delta_{j-1})^2 - 2\beta_1((1-\alpha)\delta_j + \alpha\delta_{j-1})\Delta^2 y_j + \beta_1^2 (\Delta^2 y_j)^2 \right).$$

Note that since $\delta_0 = \delta_1 = 0$ and $\delta_{N-1} = \delta_N = 0$, we have

$$\begin{aligned} \sum_{j=1}^{N} \delta_j^2 - ((1-\alpha)\delta_j + \alpha\delta_{j-1})^2 &= \sum_{j=1}^{N} (1 - (1-\alpha)^2 - \alpha^2)\delta_j^2 - 2\alpha(1-\alpha)\delta_j\delta_{j-1} \\ &= 2\beta \sum_{j=1}^{N} (2\delta_j^2 - 2\delta_j\delta_{j-1}) = 2\beta \sum_{j=1}^{N} (\delta_j - \delta_{j-1})^2 = 2\beta \sum_{j=1}^{N} (\Delta\delta_j)^2. \end{aligned}$$

Thus we get

(3.8)
$$\sum_{j=1}^{N}\delta_j^2-\sum_{j=1}^{N}(\delta_j')^2=\sum_{j=1}^{N}\left(2\beta(\Delta\delta_j)^2+2\beta_1((1-\alpha)\delta_j+\alpha\delta_{j-1})\Delta^2y_j-\beta_1^2(\Delta^2y_j)^2\right).$$

To prove Theorem 3.2, we need to prove

$$\sum_{j=1}^{N}\left(2\beta(\Delta\delta_j)^2+2\beta_1((1-\alpha)\delta_j+\alpha\delta_{j-1})\Delta^2y_j-\beta_1^2(\Delta^2y_j)^2\right)\ge 0.$$

Note that

$$\begin{aligned}
&\sum_{j=1}^{N}\left(2\beta(\Delta\delta_j)^2+2\beta_1((1-\alpha)\delta_j+\alpha\delta_{j-1})\Delta^2y_j-\beta_1^2(\Delta^2y_j)^2\right)\\
&\quad=\beta_1\left(\sum_{j=1}^{N}\left(2\left(\frac{\beta}{\beta_1}\right)(\Delta\delta_j)^2+2((1-\alpha)\delta_j+\alpha\delta_{j-1})\Delta^2y_j-\beta_1(\Delta^2y_j)^2\right)\right)\\
&\quad\ge\beta_1\left(\sum_{j=1}^{N}\left(2(\Delta\delta_j)^2+2((1-\alpha)\delta_j+\alpha\delta_{j-1})\Delta^2y_j-\beta(\Delta^2y_j)^2\right)\right)\\
&\quad=\left(\frac{\beta_1}{\beta}\right)\sum_{j=1}^{N}\left(2\beta(\Delta\delta_j)^2+2\beta((1-\alpha)\delta_j+\alpha\delta_{j-1})\Delta^2y_j-\beta^2(\Delta^2y_j)^2\right).
\end{aligned}$$

Therefore it is sufficient to prove the theorem in the case $\beta_1=\beta$, which is the worst case in a certain sense. Now we use $\Delta^2y_j=\Delta y_j-\Delta y_{j-1}$, $\Delta y_j=0$, $\delta_j=0$ for $j\le 1$, $j\ge N-1$, and Abel's transform to obtain

$$\sum_{j=1}^{N}\delta_j\Delta^2y_j=\sum_{j=1}^{N}(\delta_j-\delta_{j+1})\Delta y_j\quad\text{and}\quad\sum_{j=1}^{N}\delta_{j-1}\Delta^2y_j=\sum_{j=1}^{N}(\delta_{j-1}-\delta_j)\Delta y_j.$$

Thus, (3.8) becomes

$$\begin{aligned}
\sum_{j=1}^{N}\delta_j^2-\sum_{j=1}^{N}(\delta_j')^2=2\beta\Bigg(&\sum_{j=1}^{N}(\Delta\delta_j)^2-(1-\alpha)\sum_{j=1}^{N}\Delta\delta_{j+1}\Delta y_j\\
&-\alpha\sum_{j=1}^{N}\Delta\delta_j\Delta y_j-\frac{\beta}{2}\sum_{j=1}^{N}(\Delta^2y_j)^2\Bigg).
\end{aligned}$$

Recall that $y_j=\min(\delta_j,\delta_{j+1})$, $\Delta\delta_j=\delta_j-\delta_{j-1}$, $\Delta y_j=y_j-y_{j-1}$, $\Delta^2\delta_j=\delta_j-2\delta_{j-1}+\delta_{j-2}$, and $\Delta^2y_j=y_j-2y_{j-1}+y_{j-2}$. To finish the proof of Theorem 3.2, it is sufficient to prove the following two lemmas.

LEMMA 3.3.

$$\sum_{j=1}^{N}(\Delta\delta_j)^2-(1-\alpha)\sum_{j=1}^{N}\Delta\delta_{j+1}\Delta y_j-\alpha\sum_{j=1}^{N}\Delta\delta_j\Delta y_j-\beta\sum_{j=1}^{N}(\Delta^2\delta_j)^2\ge 0.$$

LEMMA 3.4.

$$2\sum_{j=1}^{N}(\Delta^2\delta_j)^2 \geq \sum_{j=1}^{N}(\Delta^2 y_j)^2.$$

Proof of Lemma 3.3. We consider that $\sum_j$ denotes $\sum_{j=1}^{N}$. Define

$$A = \sum_j \Delta\delta_{j+1}\Delta y_j \quad \text{and} \quad B = \sum_j \Delta\delta_j \Delta y_j.$$

Our aim is to prove that

$$\sum_j (\Delta\delta_j)^2 - (1-\alpha)A - \alpha B - \beta\sum_j (\Delta^2\delta_j)^2 \geq 0. \tag{3.9}$$

Let $u_+ = \max(u,0)$, $u_- = \min(u,0)$. It is easy to check that

$$\Delta y_j = (\Delta\delta_j)_+ + (\Delta\delta_{j+1})_-. \tag{3.10}$$

We can transform A as follows:

$$\begin{aligned} A &= \sum_j \Delta\delta_{j+1}((\Delta\delta_j)_+ + (\Delta\delta_{j+1})_-) \\ &= \sum_j \Delta\delta_{j+1}(\Delta\delta_j)_+ + \sum_j \Delta\delta_j(\Delta\delta_j)_- = \sum_{\Delta\delta_j\leq 0}(\Delta\delta_j)^2 + \sum_{\Delta\delta_j\geq 0}\Delta\delta_j\Delta\delta_{j+1} \\ &= \sum_{\Delta\delta_j\leq 0}(\Delta\delta_j)^2 + \sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j+1}\leq 0}\Delta\delta_j\Delta\delta_{j+1} + D, \end{aligned} \tag{3.11}$$

where $D = \sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j+1}\geq 0}\Delta\delta_j\Delta\delta_{j+1}$. Further,

$$\begin{aligned} D &= \frac{1}{2}\sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j+1}\geq 0}\left((\Delta\delta_j)^2 + (\Delta\delta_{j+1})^2 - (\Delta^2\delta_{j+1})^2\right) \\ &= \frac{1}{2}\sum_{\Delta\delta_{j-1}\geq 0,\ \Delta\delta_j\geq 0}\left((\Delta\delta_{j-1})^2 + (\Delta\delta_j)^2 - (\Delta^2\delta_j)^2\right) \\ &= \frac{1}{2}\sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j+1}\geq 0}(\Delta\delta_j)^2 + \frac{1}{2}\sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j-1}\geq 0}(\Delta\delta_j)^2 - \frac{1}{2}\sum_{\Delta\delta_{j-1}\geq 0,\ \Delta\delta_j\geq 0}(\Delta^2\delta_j)^2. \end{aligned}$$

By (3.11),

$$\begin{aligned} A = \sum_j(\Delta\delta_j)^2 &- \frac{1}{2}\sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j+1}<0}(\Delta\delta_j)^2 - \frac{1}{2}\sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j-1}<0}(\Delta\delta_j)^2 \\ &- \frac{1}{2}\sum_{\Delta\delta_{j-1}\geq 0,\ \Delta\delta_j\geq 0}(\Delta^2\delta_j)^2 + \sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j+1}\leq 0}\Delta\delta_j\Delta\delta_{j+1}. \end{aligned} \tag{3.12}$$

Transform B in the same way as A:

$$B = \sum_{\Delta\delta_j\geq 0}(\Delta\delta_j)^2 + \sum_{\Delta\delta_j\geq 0,\ \Delta\delta_{j+1}\leq 0}\Delta\delta_j\Delta\delta_{j+1} + E, \tag{3.13}$$

where $E = \sum_{\Delta\delta_j \le 0,\, \Delta\delta_{j+1} \le 0} \Delta\delta_j \Delta\delta_{j+1}$. The quantity E can also be rewritten in the same way as D:

$$E = \frac{1}{2} \sum_{\Delta\delta_j \le 0,\, \Delta\delta_{j+1} \le 0} (\Delta\delta_j)^2 + \frac{1}{2} \sum_{\Delta\delta_{j-1} \le 0,\, \Delta\delta_j \le 0} (\Delta\delta_j)^2 - \frac{1}{2} \sum_{\Delta\delta_{j-1} \le 0,\, \Delta\delta_j \le 0} (\Delta^2\delta_j)^2.$$

Combining this equality with (3.13), we get

$$\begin{aligned} B = {} & \sum_j (\Delta\delta_j)^2 - \frac{1}{2} \sum_{\Delta\delta_j \le 0,\, \Delta\delta_{j+1} > 0} (\Delta\delta_j)^2 - \frac{1}{2} \sum_{\Delta\delta_j \le 0,\, \Delta\delta_{j-1} > 0} (\Delta\delta_j)^2 \\ & - \frac{1}{2} \sum_{\Delta\delta_{j-1} \le 0,\, \Delta\delta_j \le 0} (\Delta^2\delta_j)^2 + \sum_{\Delta\delta_{j+1} \le 0,\, \Delta\delta_j \ge 0} \Delta\delta_j \Delta\delta_{j+1}. \end{aligned} \tag{3.14}$$

By (3.12) and (3.14),

$$\sum_j (\Delta\delta_j)^2 - (1-\alpha)A - \alpha B - \beta \sum_j (\Delta^2\delta_j)^2 = F + G + H + I + J + K + L, \tag{3.15}$$

where

$$F = \frac{1-\alpha}{2} \sum_{\Delta\delta_j \ge 0,\, \Delta\delta_{j+1} < 0} (\Delta\delta_j)^2, \quad G = \frac{1-\alpha}{2} \sum_{\Delta\delta_j \ge 0,\, \Delta\delta_{j-1} < 0} (\Delta\delta_j)^2,$$

$$H = \frac{\alpha}{2} \sum_{\Delta\delta_j \le 0,\, \Delta\delta_{j+1} > 0} (\Delta\delta_j)^2, \quad I = \frac{\alpha}{2} \sum_{\Delta\delta_j \le 0,\, \Delta\delta_{j-1} > 0} (\Delta\delta_j)^2,$$

$$\begin{aligned} J = {} & - \sum_{\Delta\delta_{j+1} \le 0,\, \Delta\delta_j \ge 0} \Delta\delta_j \Delta\delta_{j+1} + \left(\frac{1-\alpha}{2} - \beta \right) \sum_{\Delta\delta_{j-1} \ge 0,\, \Delta\delta_j \ge 0} (\Delta^2\delta_j)^2 \\ & + \left(\frac{\alpha}{2} - \beta \right) \sum_{\Delta\delta_{j-1} \le 0,\, \Delta\delta_j \le 0} (\Delta^2\delta_j)^2, \end{aligned}$$

$$K = -\beta \sum_{\Delta\delta_{j-1} > 0,\, \Delta\delta_j < 0} (\Delta^2\delta_j)^2, \qquad \text{and} \quad L = -\beta \sum_{\Delta\delta_{j-1} < 0,\, \Delta\delta_j > 0} (\Delta^2\delta_j)^2.$$

We have to prove that $F + G + H + I + J + K + L \ge 0$. Among the sums F, G, H, I, J, K, L, only the two last sums might be negative; we will show that and

$$F + I + K \ge 0, \tag{3.16}$$

$$G + H + L \ge 0. \tag{3.17}$$

Indeed,

$$\begin{aligned} & \frac{1-\alpha}{2} (\Delta\delta_{j-1})^2 + \frac{\alpha}{2} (\Delta\delta_j)^2 - \beta (\Delta^2\delta_j)^2 = \frac{1-\alpha}{2} (\Delta\delta_{j-1})^2 + \frac{\alpha}{2} (\Delta\delta_j)^2 \\ & \quad - \frac{(1-\alpha)\alpha}{2} (\Delta\delta_j - \Delta\delta_{j-1})^2 = \frac{1}{2} ((1-\alpha)\Delta\delta_{j-1} + \alpha\Delta\delta_j))^2 \ge 0. \end{aligned}$$

Summing the last inequality over all j with $\Delta\delta_{j-1} > 0$, $\Delta\delta_j < 0$, we get (3.16). The inequality (3.17) can be proved similarly.

Additionally, we have

$$J \geq 0. \tag{3.18}$$

Finally, plugging (3.16), (3.17), and (3.18) into (3.15), we obtain the required (3.9). This completes the proof of Lemma 3.3.

Proof of Lemma 3.4. First, recall that $\Delta^2 y_j = 0$ for $j \leq 1$ and $j \geq N$. Also, from the proof of Lemma 3.3 we have

$$\Delta y_j = \begin{cases} \Delta\delta_{j+1} & \text{if} \quad \Delta\delta_{j+1} \leq 0, \quad \Delta\delta_j \leq 0, \\ \Delta\delta_{j+1} + \Delta\delta_j & \text{if} \quad \Delta\delta_{j+1} \leq 0, \quad \Delta\delta_j \geq 0, \\ \Delta\delta_j & \text{if} \quad \Delta\delta_{j+1} \geq 0, \quad \Delta\delta_j \geq 0, \\ 0 & \text{if} \quad \Delta\delta_{j+1} \geq 0, \quad \Delta\delta_j \leq 0. \end{cases}$$

Similarly,

$$\Delta y_{j-1} = \begin{cases} \Delta\delta_j & \text{if} \quad \Delta\delta_j \leq 0, \quad \Delta\delta_{j-1} \leq 0, \\ \Delta\delta_j + \Delta\delta_{j-1} & \text{if} \quad \Delta\delta_j \leq 0, \quad \Delta\delta_{j-1} \geq 0, \\ \Delta\delta_{j-1} & \text{if} \quad \Delta\delta_j \geq 0, \quad \Delta\delta_{j-1} \geq 0, \\ 0 & \text{if} \quad \Delta\delta_j \geq 0, \quad \Delta\delta_{j-1} \leq 0. \end{cases}$$

Therefore $\Delta\delta_{j-1}$, $\Delta\delta_j$, $\Delta\delta_{j+1}$ and their signs uniquely determine $\Delta^2 y_j$. We have eight cases depending on what the signs of $\Delta\delta_{j-1}$, $\Delta\delta_j$, $\Delta\delta_{j+1}$ are.

- *Case* I. $(+,+,+)$, that is, $\Delta\delta_{j-1} \geq 0$, $\Delta\delta_j \geq 0$, $\Delta\delta_{j+1} \geq 0$. Then, $\Delta y_j = \Delta\delta_j$, $\Delta y_{j-1} = \Delta\delta_{j-1}$, and so $\Delta^2 y_j = \Delta^2\delta_j$; thus, $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_j)^2$ in this case.
- *Case* II. $(+,+,-)$, that is, $\Delta\delta_{j-1} \geq 0$, $\Delta\delta_j \geq 0$, $\Delta\delta_{j+1} < 0$. Then, $\Delta y_j = \Delta\delta_{j+1} + \Delta\delta_j$, $\Delta y_{j-1} = \Delta\delta_{j-1}$, and so $\Delta^2 y_j = \Delta\delta_{j+1} + \Delta\delta_j - \Delta\delta_{j-1}$.
- *Case* III. $(+,-,+)$, that is, $\Delta\delta_{j-1} \geq 0$, $\Delta\delta_j < 0$, $\Delta\delta_{j+1} \geq 0$. Then, $\Delta y_j = 0$, $\Delta y_{j-1} = \Delta\delta_j + \Delta\delta_{j-1}$, and so $\Delta^2 y_j = -\Delta\delta_j - \Delta\delta_{j-1}$. In this case $(\Delta^2 y_j)^2 - (\Delta^2\delta_j)^2 = 4\Delta\delta_j\Delta\delta_{j-1} \leq 0$ and $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_j)^2$.
- *Case* IV. $(+,-,-)$, that is, $\Delta\delta_{j-1} \geq 0$, $\Delta\delta_j < 0$, $\Delta\delta_{j+1} < 0$. Then, $\Delta y_j = \Delta\delta_{j+1}$, $\Delta y_{j-1} = \Delta\delta_j + \Delta\delta_{j-1}$, and so $\Delta^2 y_j = \Delta\delta_{j+1} - \Delta\delta_j - \Delta\delta_{j-1}$.
- *Case* V. $(-,+,+)$, that is, $\Delta\delta_{j-1} < 0$, $\Delta\delta_j \geq 0$, $\Delta\delta_{j+1} \geq 0$. Then, $\Delta y_j = \Delta\delta_j$, $\Delta y_{j-1} = 0$, and so $\Delta^2 y_j = \Delta\delta_j$. In this case $0 \leq \Delta\delta_j < \Delta\delta_j - \Delta\delta_{j-1}$ and $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_j)^2$.
- *Case* VI. $(-,+,-)$, that is, $\Delta\delta_{j-1} < 0$, $\Delta\delta_j \geq 0$, $\Delta\delta_{j+1} < 0$. Then, $\Delta y_j = \Delta\delta_{j+1} + \Delta\delta_j$, $\Delta y_{j-1} = 0$, and so $\Delta^2 y_j = \Delta\delta_{j+1} + \Delta\delta_j$. In this case $(\Delta^2 y_j)^2 - (\Delta^2\delta_{j+1})^2 = 4\Delta\delta_{j+1}\Delta\delta_j \leq 0$ and $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_{j+1})^2$.
- *Case* VII. $(-,-,+)$, that is, $\Delta\delta_{j-1} < 0$, $\Delta\delta_j < 0$, $\Delta\delta_{j+1} \geq 0$. Then, $\Delta y_j = 0$, $\Delta y_{j-1} = \Delta\delta_j$, and so $\Delta^2 y_j = -\Delta\delta_j$. In this case $0 < -\Delta\delta_j \leq \Delta\delta_{j+1} - \Delta\delta_j$ and $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_{j+1})^2$.
- *Case* VIII. $(-,-,-)$, that is, $\Delta\delta_{j-1} < 0$, $\Delta\delta_j < 0$, $\Delta\delta_{j+1} < 0$. Then, $\Delta y_j = \Delta\delta_{j+1}$, $\Delta y_{j-1} = \Delta\delta_j$, and so $\Delta^2 y_j = \Delta^2\delta_{j+1}$. In this case $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_{j+1})^2$.

Therefore in Cases I $(+,+,+)$, III $(+,-,+)$, and V $(-,+,+)$, $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_j)^2$, and in Cases VI $(-,+,-)$, VII $(-,-,+)$, and VIII $(-,-,-)$, $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_{j+1})^2$. There are only two "bad" cases: II $(+,+,-)$ and IV $(+,-,-)$, which need a special treatment.

Next, we define a sequence of $+$ and $-$ signs $\{s_j\}$, where $s_j = +$ if $\Delta\delta_j \geq 0$ and $s_j = -$ if $\Delta\delta_j < 0$. Note that $s_j = +$ for $j \leq 3$ and $j \geq N-2$. There are three types of "bad" quadruples:

- Type A quadruple: $(+,+,-,-)$, that is, $\Delta\delta_{j-1} \geq 0$, $\Delta\delta_j \geq 0$, $\Delta\delta_{j+1} < 0$, $\Delta\delta_{j+2} < 0$ for some j. We claim that in this case the following inequality holds:

$$(\Delta^2 y_j)^2 + (\Delta^2 y_{j+1})^2 \leq 2(\Delta^2\delta_j)^2 + 2(\Delta^2\delta_{j+1})^2 + 2(\Delta^2\delta_{j+2})^2. \tag{3.19}$$

  In this case $\Delta^2 y_j = \Delta\delta_{j+1} + \Delta\delta_j - \Delta\delta_{j-1}$ and $\Delta^2 y_{j+1} = \Delta\delta_{j+2} - \Delta\delta_{j+1} - \Delta\delta_j$. If we denote $\Delta\delta_{j-1}$ by a, $\Delta\delta_j$ by b, $\Delta\delta_{j+1}$ by c, and $\Delta\delta_{j+2}$ by d, the above inequality becomes

$$\begin{aligned}&(c+b-a)^2 + (d-c-b)^2 \leq 2(b-a)^2 + 2(c-b)^2 + 2(d-c)^2\\ &\text{for } a \geq 0,\ b \geq 0,\ c < 0,\ d < 0,\end{aligned}$$

  which is equivalent to $a^2 + 2b^2 + 2c^2 + d^2 - 2ab + 2ac - 8bc + 2bd - 2cd \geq 0$, or

$$(a - b + c)^2 + (b - c + d)^2 - 4bc \geq 0,$$

  which holds since $b \geq 0$, $c < 0$.
- Type B quadruple: $(+,+,-,+)$, that is, $\Delta\delta_{j-1} \geq 0$, $\Delta\delta_j \geq 0$, $\Delta\delta_{j+1} < 0$, $\Delta\delta_{j+2} \geq 0$ for some j. We claim that in this case the following inequality holds:

$$(\Delta^2 y_j)^2 + (\Delta^2 y_{j+1})^2 \leq 2(\Delta^2\delta_j)^2 + 2(\Delta^2\delta_{j+1})^2 + (\Delta^2\delta_{j+2})^2. \tag{3.20}$$

  In this case $\Delta^2 y_j = \Delta\delta_{j+1} + \Delta\delta_j - \Delta\delta_{j-1}$ and $\Delta^2 y_{j+1} = -\Delta\delta_{j+1} - \Delta\delta_j$. Using the notation we just introduced, the inequality becomes

$$\begin{aligned}&(c+b-a)^2 + (c+b)^2 \leq 2(b-a)^2 + 2(c-b)^2 + (d-c)^2\\ &\text{for } a \geq 0,\ b \geq 0,\ c < 0,\ d \geq 0.\end{aligned}$$

  Since $(d-c)^2 \geq c^2$, it is sufficient to prove

$$(c+b-a)^2 + (c+b)^2 \leq 2(b-a)^2 + 2(c-b)^2 + c^2, \text{ or}$$

$$a^2 + 2b^2 + c^2 - 2ab + 2ac - 8bc \geq 0, \text{ or } (a - b + c)^2 + b^2 - 6bc \geq 0,$$

  which holds for $b \geq 0$, $c < 0$.
- Type C quadruple: $(-,+,-,-)$, that is, $\Delta\delta_{j-1} < 0$, $\Delta\delta_j \geq 0$, $\Delta\delta_{j+1} < 0$, $\Delta\delta_{j+2} < 0$ for some j. We claim that in this case the following inequality holds:

$$(\Delta^2 y_j)^2 + (\Delta^2 y_{j+1})^2 \leq (\Delta^2\delta_j)^2 + 2(\Delta^2\delta_{j+1})^2 + 2(\Delta^2\delta_{j+2})^2. \tag{3.21}$$

  In this case $\Delta^2 y_j = \Delta\delta_{j+1} + \Delta\delta_j$ and $\Delta^2 y_{j+1} = \Delta\delta_{j+2} - \Delta\delta_{j+1} - \Delta\delta_j$. Using the notation we just introduced, the inequality becomes

$$\begin{aligned}&(c+b)^2 + (d-c-b)^2 \leq (b-a)^2 + 2(c-b)^2 + 2(d-c)^2\\ &\text{for } a < 0,\ b \geq 0,\ c < 0,\ d < 0.\end{aligned}$$

  Since $(b-a)^2 \geq b^2$, it is sufficient to prove

$$(c+b)^2 + (d+c-b)^2 \leq b^2 + 2(c-b)^2 + 2(d-c)^2, \text{ or}$$

$$b^2 + 2c^2 + d^2 - 8bc + 2bd - 2cd \geq 0, \text{ or } (b - c + d)^2 + c^2 - 6bc \geq 0,$$

  which holds for $b \geq 0$, $c < 0$.

We will call the Type A (3.19), Type B (3.20), and Type C (3.21) inequalities "long" inequalities; we call the inequalities of type $(\Delta^2 y_j)^2 \le (\Delta^2 \delta_j)^2$ and $(\Delta^2 y_j)^2 \le (\Delta^2 \delta_{j+1})^2$ "short" inequalities.

Now, we construct a set of inequalities. We identify all "bad" quadruples and include the corresponding inequality (Type A, B, or C) in the set. Next, for all $j \in [1, N]$ such that s_j is not a middle element of a "bad" quadruple, and such that j does not belong to the "bad" Cases II and IV, we include the corresponding "short" inequality in the set. Finally, we add all inequalities in the set. Taking into account that $\Delta^2 \delta_j = 0$ and $\Delta^2 y_j = 0$ for $j > N$, the resulting inequality is

$$\sum_{j=1}^{N} a_j (\Delta^2 y_j)^2 \le \sum_{j=1}^{N} b_j (\Delta^2 \delta_j)^2, \tag{3.22}$$

where the a_j's and b_j's are nonnegative integers. To finish the proof of the lemma we need to show $a_j \ge 1$ and $b_j \le 2$ for all $j \in [1, N]$.

Note that all "long" inequalities have the form $(\Delta^2 y_j)^2 + (\Delta^2 y_{j+1})^2 \le \cdots$, where s_j and s_{j+1} are the middle elements of a "bad" quadruple. Then $a_j \ge 1$ if s_j is a middle element of a "bad" quadruple. (By middle element of a quadruple we mean second or third element of the quadruple.)

Now, suppose that s_j is not a middle element of a "bad" quadruple. Then j does not belong to the "bad" Cases II and IV. Indeed if j is in Case II: $(s_{j-1}, s_j, s_{j+1}) = (+, +, -)$, then s_j is a middle element of Type B quadruple if $s_{j+2} = +$ and a middle element of Type A quadruple if $s_{j+2} = -$. Similarly, if j is in Case IV: $(s_{j-1}, s_j, s_{j+1}) = (+, -, -)$, then s_j is a middle element of Type A quadruple if $s_{j-1} = +$ and a middle element of Type C quadruple if $s_{j-1} = -$. Therefore a "short" inequality for $(\Delta^2 y_j)^2$ has been included in the set of inequalities. Thus $a_j \ge 1$ in this case as well. We have proved $a_j \ge 1$ for all $j \in [1, N]$.

Now, we prove $b_j \le 2$ for all $j \in [1, N]$. Note that $(\Delta^2 \delta_j)^2$ can appear in only two "short" inequalities: $(\Delta^2 y_j)^2 \le (\Delta^2 \delta_j)^2$ and $(\Delta^2 y_{j-1})^2 \le (\Delta^2 \delta_j)^2$. Therefore $b_j \le 2$ if $(\Delta^2 \delta_j)^2$ does not appear in any "long" inequalities, that is, if s_j is not a second, third, or fourth element of a "bad" quadruple.

The case when s_j is a second, third, or fourth element of a "bad" quadruple requires more work. First, note that two distinct "bad" quadruples have at most two common elements. Indeed all "bad" quadruples are of the form $(*, +, -, *)$, where $*$ denotes $+$ or $-$, and no "bad" quadruple has $(+, -)$ as its first two or last two elements. Next, the only case when two "bad" quadruples have two common elements is the following configuration:

$$(s_{j-1}, s_j, s_{j+1}, s_{j+2}, s_{j+3}, s_{j+4}) = (+, +, -, +, -, -). \tag{3.23}$$

Indeed, a Type A quadruple cannot share exactly two elements with another "bad" quadruple because no "bad" quadruple has $(-, -)$ as its first two elements, or $(+, +)$ as its last two elements. Similar analysis shows that the only way a Type B or Type C quadruple can share exactly two elements with another Type B or Type C quadruple is when the configuration (3.23) occurs.

Let us analyze the configuration (3.23). The "long" inequalities which correspond to the two "bad" quadruples in this configuration are

$$(\Delta^2 y_j)^2 + (\Delta^2 y_{j+1})^2 \le 2(\Delta^2 \delta_j)^2 + 2(\Delta^2 \delta_{j+1})^2 + (\Delta^2 \delta_{j+2})^2, \tag{3.24}$$

$$(\Delta^2 y_{j+2})^2 + (\Delta^2 y_{j+3})^2 \le (\Delta^2 \delta_{j+2})^2 + 2(\Delta^2 \delta_{j+3})^2 + 2(\Delta^2 \delta_{j+4})^2. \tag{3.25}$$

Their sum is

$$(\Delta^2 y_j)^2 + (\Delta^2 y_{j+1})^2 + (\Delta^2 y_{j+2})^2 + (\Delta^2 y_{j+3})^2 \\ \leq 2(\Delta^2 \delta_j)^2 + 2(\Delta^2 \delta_{j+1})^2 + 2(\Delta^2 \delta_{j+2})^2 + 2(\Delta^2 \delta_{j+3})^2 + 2(\Delta^2 \delta_{j+4})^2.$$

In this case s_j, s_{j+1}, s_{j+2}, s_{j+3}, and s_{j+4} appear as second, third, or fourth elements of a "bad" quadruple. Since the configuration (3.23) starts with $(+,+)$ and ends with $(-,-)$, it cannot share two elements with a "bad" quadruple outside the configuration. This means that none of s_j, s_{j+1}, s_{j+2}, s_{j+3}, and s_{j+4} can be a second, third, or fourth element of a "bad" quadruple outside the configuration. In other words, none of $(\Delta^2\delta_j)^2$, $(\Delta^2\delta_{j+1})^2$, $(\Delta^2\delta_{j+2})^2$, $(\Delta^2\delta_{j+3})^2$, and $(\Delta^2\delta_{j+4})^2$ can appear in a "long" inequality other than (3.24) and (3.25). Since, s_j, s_{j+1}, s_{j+2}, and s_{j+3} are middle elements of "bad" quadruples, $(\Delta^2\delta_{j+1})^2$, $(\Delta^2\delta_{j+2})^2$, and $(\Delta^2\delta_{j+3})^2$ cannot appear in "short" inequalities either. Thus $b_{j+1} = b_{j+2} = b_{j+3} = 2$. Also, $(\Delta^2\delta_j)^2$ cannot appear in a "short" inequality. The only way this could happen is $(\Delta^2 y_{j-1})^2 \leq (\Delta^2\delta_j)^2$, which is impossible since $j-1$ is either in Case I $(+,+,+)$ or Case V $(-,+,+)$, depending on what s_{j-2} is, and in both cases the short inequality is $(\Delta^2 y_{j-1})^2 \leq (\Delta^2\delta_{j-1})^2$. Thus $b_j = 2$. Similarly, $(\Delta^2\delta_{j+4})^2$ cannot appear in a "short" inequality. The only way this could happen is $(\Delta^2 y_{j+4})^2 \leq (\Delta^2\delta_{j+4})^2$, which is impossible since $j+4$ is either in Case VII $(-,-,+)$ or Case VIII $(-,-,-)$, depending on what s_{j+5} is, and in both cases the short inequality is $(\Delta^2 y_{j+4})^2 \leq (\Delta^2\delta_{j+5})^2$. Thus $b_{j+4} = 2$. This concludes the analysis of the configuration (3.23).

Now, let s_j be a second, third, or fourth element of a "bad" quadruple but not an element of a configuration (3.23). This means that $(\Delta^2\delta_j)^2$ appears in exactly one "long" inequality (it cannot be a second, third, or fourth element of two distinct "bad" quadruples). If s_j is a third element of a "bad" quadruple, then s_{j-1} and s_j are the middle elements of the quadruple and $(\Delta^2\delta_j)^2$ does not appear in a short inequality. Thus, $b_j \leq 2$ in this case. The cases when s_j is a second or fourth element of a "bad" quadruple need separate consideration.

1. s_j is a second element of a Type A quadruple $(+,+,-,-)$. The only way $(\Delta^2\delta_j)^2$ could appear in a "short" inequality is $(\Delta^2 y_{j-1})^2 \leq (\Delta^2\delta_j)^2$, which is impossible since $j-1$ is either in Case I $(+,+,+)$ or Case V $(-,+,+)$, depending on what s_{j-2} is, and in both cases the short inequality is $(\Delta^2 y_{j-1})^2 \leq (\Delta^2\delta_{j-1})^2$. Thus $b_j = 2$.
2. s_j is a fourth element of a Type A quadruple $(+,+,-,-)$. The only way $(\Delta^2\delta_j)^2$ could appear in a "short" inequality is $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_j)^2$, which is impossible since j is either in Case VII $(-,-,+)$ or Case VIII $(-,-,-)$, depending on what s_{j+1} is, and in both cases the short inequality is $(\Delta^2 y_j)^2 \leq (\Delta^2\delta_{j+1})^2$. Thus $b_j = 2$.
3. s_j is a second element of a Type B quadruple $(+,+,-,+)$. Here the argument is word-by-word like in part 1. The only way $(\Delta^2\delta_j)^2$ could appear in a "short" inequality is $(\Delta^2 y_{j-1})^2 \leq (\Delta^2\delta_j)^2$, which is impossible since $j-1$ is either in Case I $(+,+,+)$ or Case V $(-,+,+)$, depending on what s_{j-2} is, and in both cases the short inequality is $(\Delta^2 y_{j-1})^2 \leq (\Delta^2\delta_{j-1})^2$. Thus $b_j = 2$.
4. s_j is a fourth element of a Type B quadruple $(+,+,-,+)$. Since the coefficient of $(\Delta^2\delta_j)^2$ in the corresponding "long" inequality (3.20) is 1 and $(\Delta^2\delta_j)^2$ could appear in at most one "short" inequality, we conclude $b_j \leq 2$.
5. s_j is a second element of a Type C quadruple $(-,+,-,-)$. Since the coefficient of $(\Delta^2\delta_j)^2$ in the corresponding "long" inequality (3.21) is 1 and $(\Delta^2\delta_j)^2$ could appear in at most one "short" inequality, we conclude $b_j \leq 2$.

6. s_j is a fourth element of a Type C quadruple $(-,+,-,-)$. Here the argument is word-by-word like in Case 2. The only way $(\Delta^2\delta_j)^2$ could appear in a "short" inequality is $(\Delta^2 y_j)^2 \le (\Delta^2\delta_j)^2$, which is impossible since j is either in Case VII $(-,-,+)$ or Case VIII $(-,-,-)$, depending on what s_{j+1} is, and in both cases the short inequality is $(\Delta^2 y_j)^2 \le (\Delta^2\delta_{j+1})^2$. Thus $b_j = 2$.

We have shown that in all six cases $b_j \le 2$ for $j \in [1, N]$, which completes the proof of Lemma 3.4 and Theorem 3.2. □

Now, we continue with the general case, the proof of Theorem 3.1. That is, we want to show that the l_2 norms inequality

$$\|\{\delta'_j\}\|_{l_2} \le \|\{\delta_j\}\|_{l_2} \tag{3.26}$$

holds for *any* initial sequence $\{\delta_j\}$ with finite l_2 norm. We consider the sequence $\{w_j\}$ and restrict the index j to a maximal subset Λ_m on which the piecewise constant function w is monotone, recalling that $\delta_j = w_j - w_{j-1}$. Given a sequence $\{w_j\}$, we can decompose it into monotone subsequences. This decomposition also gives a decomposition of the sequence $\{\delta_j\}$ into subsequences such that in each subsequence all jumps have the same sign (nonnegative or nonpositive). Without any limitations, we assume that the jumps $\{\delta_j\}$ are nonnegative for all $l \le j \le r$, $\delta_{l-1} < 0$, and $\delta_{r+1} < 0$. That is, w_{l-1} is a local minimum and w_r is a local maximum of the piecewise constant function w. Let w^m be the following piecewise constant correction of w:

$$w^m_j := \begin{cases} w_j & \text{if} \quad l \le j \le r, \\ w_{l-1} & \text{if} \quad j < l, \\ w_r & \text{if} \quad j > r. \end{cases} \tag{3.27}$$

Note that $\Lambda_m = \{j : l \le j \le r+1\}$, and the jump sequence $\delta^m := \{\delta^m_j\}$ of w^m is given by

$$\delta^m_j := \begin{cases} w_j - w_{j-1} & \text{if} \quad l \le j \le r, \\ 0 & \text{otherwise.} \end{cases} \tag{3.28}$$

Hence, we have a sequence of monotone functions $\{w^m\}$ and the corresponding jump sequences $\{\delta^m_j\}_j$ such that

$$\sum_m \sum_{j\in\Lambda_m} \|\{\delta^m_j\}\|^2_{l_2} = \sum_{m,j\in\mathbb{Z}} \|\{\delta^m_j\}\|^2_{l_2} = \|\{\delta_j\}\|^2_{l_2},$$

because the sequence of the jumps of $\{\delta_j\}$ is decomposed into disjoint jump subsequences $\{\delta^m_j\}$. There are two types of jumps δ'_j. A jump δ'_j is of *type* 1 if it is equal to the jump $\delta'_j(\delta^m)$, that is, the jump generates with the starting sequence $\{\delta^m_j\}$, where the index m such that $j \in \Lambda_m$. A jump is of *type* 2 if it is not of *type* 1. Note that a *type* 2 jump δ'_{j^*} occurs only inside an interval which contains a strict local extremum. Near a local extremum we have two nonzero jumps, say $\delta^l_{j^*}$ and $\delta^r_{j^*}$, generated by the two monotone w^m's with index sets finishing/starting with j^*. It is easy to verify that

$$|\delta'_{j^*}| = \big||\delta^l_{j^*}| - |\delta^r_{j^*}|\big|.$$

Hence, we have that

$$(\delta'_{j^*})^2 < (\delta^l_{j^*})^2 + (\delta^r_{j^*})^2,$$

and we conclude that

$$\sum_j (\delta'_j)^2 \le \sum_m \sum_{j\in\Lambda_m} (\delta'_j(\delta^m))^2 \le \sum_m \sum_{j\in\Lambda_m} (\delta^m_j)^2 = \sum_n (\delta_j)^2,$$

where we use the notation $\delta'_j(\delta^m)$ for the new jumps generated by δ^m. It is also easy to prove a local inequality but with index set for δ'_j starting from an interval right after an extremum and finishing right before one. □

4. Error estimates for linear flux. Recall that u is the entropy solution to the conservation law $u_t + f(u)_x = 0$ with initial condition u^0, and v is the numerical solution described in (2.12). In the case of linear flux and $0 \le \theta \le 1$, the formula for the new averages of the minmod scheme is given in (3.1), and the conservation law (2.1) reduces to

$$u_t + a u_x = 0. \tag{4.1}$$

Let S_τ be the shift operator defined by $S_\tau g(\cdot) := g(\cdot - \tau)$. Then the exact solution of (4.1) at time t for any initial data u^0 is $u(\cdot, t) = S_{at}u^0$. Let A_h be the averaging operator defined on a uniform partition by $A_h g|_I := \frac{1}{h}\int_I g(s)ds$, where $|I| = h$. It will be useful to define a global approximate solution v. We first define the approximate solution at discrete times by $v^n := v(\cdot, n\Delta t)$, $n = 0, 1, \ldots, N$, in the following way: (i) $v^0 := u^0$; (ii) $v^n := S_{a\Delta t}P_h v^{n-1}$, where for odd n, $1 \le n \le N$, $P_h v$ is the linear function on $I_j := (x_{j-1/2}, x_{j+1/2})$ defined in (2.3) with the minmod slopes (2.8), and for even n we have the analogous definition of P_h on the shifted partition $\{I_{j+1/2} | \, j \in \mathbb{Z}\}$. Note that $P_h v^n = P_h A_h v^n$ because the piecewise linear projection P_h, defined in (2.3) and (2.8), is based only on the averages of v^n on the corresponding partition. The formula (3.1) for the new cell averages can be written as

$$v^n_{j+1/2} = A_h(v^n)|_{I_{j+1/2}} = A_h(S_{a\Delta t}P_h v^{n-1})|_{I_{j+1/2}}$$

for odd n, with A_h based on the staggered partition $\{I_{j+1/2} | \, j \in \mathbb{Z}\}$ and P_h based on regular partition $\{I_j | \, j \in \mathbb{Z}\}$. For even n, we have the same sequence of operators but on the reversed partitions. The global approximate solution v is defined by $v(\cdot, n\Delta t) = v^n$ and $v(\cdot, t) = S_{a(t-n\Delta t)}(P_h v^n)$ for $n\Delta t < t \le (n+1)\Delta t$ and $n = 0, 1, \ldots, N-1$. That is, v solves (4.1) exactly for $n\Delta t < t \le (n+1)\Delta t$ with initial data $P_h v^n$, $n = 0, 1, \ldots, N-1$.

In order to describe the next result, we need to introduce some notation. A function g is of bounded variation, i.e., $g \in \mathrm{BV}(\mathbb{R})$, if

$$|g|_{\mathrm{BV}(\mathbb{R})} := \sup \sum_{i=1}^{n} |g(x_{i+1}) - g(x_i)| < \infty,$$

where the supremum is taken over all finite sequences $x_1 < \cdots < x_n$ in $\mathbb{R}$. Functions of bounded variation have at most countably many discontinuities, and their left and right limits $g(x^-)$ and $g(x^+)$ exist at each point $x \in \mathbb{R}$. Since the values of the initial condition u^0 on a set of measure zero have no influence on the numerical solution v and the entropy solution u, it is desirable to replace the seminorm $|\cdot|_{\mathrm{BV}(\mathbb{R})}$ by a similar quantity independent of the function values on sets of measure zero. The standard approach in conservation laws is to consider the space $\mathrm{Lip}(1, \mathrm{L}^1(\mathbb{R}))$ of all functions $g \in L^1(\mathbb{R})$ such that the seminorm

$$|g|_{\mathrm{Lip}(1,\mathrm{L}^1(\mathbb{R}))} := \limsup_{s>0} \frac{1}{s}\int_{\mathbb{R}} |g(x+s) - g(x)| \, dx \tag{4.2}$$

is finite. It is clear that $|g|_{\mathrm{Lip}(1,\mathrm{L}^1(\mathbb{R}))}$ will not change if g is modified on a set of measure zero. At the same time the above two seminorms are equal for functions $g \in \mathrm{BV}(\mathbb{R})$ such that the value of g at a point of discontinuity lies between $g(x^-)$ and $g(x^+)$ (see Theorem 9.3 in [5]). Similarly, we define the space $\mathrm{Lip}(1, \mathrm{L}^p(\mathbb{R}))$, $1 \le p \le \infty$, which is the set of all functions $g \in L^p(\mathbb{R})$ for which

$$\|g(\cdot - s) - g(\cdot)\|_{L^p(\mathbb{R})} \le Ms, \quad s > 0. \tag{4.3}$$

The smallest $M \ge 0$ for which (4.3) holds is $|g|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}$. It is easy to see that in the case $p = 1$ the seminorm given in (4.3) is the same as the one in (4.2). In the case $p > 1$, the space $\mathrm{Lip}(1, \mathrm{L}^p(\mathbb{R}))$ is essentially the same as $W^1(L^p(\mathbb{R}))$; see [5] for details. With this notation, we have the following result.

THEOREM 4.1. *Let $u(x,t) = u(x - at, 0)$ be the solution to (2.1) with linear flux $f(z) = az$, and let v be the numerical solution described in (3.1) with $0 \le \theta \le 1$. If the CFL condition (2.5) is satisfied, $t_n = n\Delta t$, $0 \le n \le N$, and $T = N\Delta t$, we have*

$$\|u(\cdot, T) - v(\cdot, T)\|_{L^p(\mathbb{R})} \le C(Nh)^{1/2} h^{1/2} |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))} \tag{4.4}$$

for $p = 1, 2$, where C is an absolute constant.

Proof. The L_1 estimate is based on the TVD property of the numerical solution v, and the L_2 estimate is based on the l_2 stability of the jumps proved in Theorem 3.1. Both estimates use a dual argument similar to the one in [19], and in the proof we use an index p, where $p \in \{1, 2\}$. Note that we consider the case of linear flux, and the usual Lip+ stability requirement is not needed in the dual approach because the negative norm stability (4.6) holds for any initial data (not just Lip+). In the proof, C will be an absolute constant that can be different at different places.

Let $e(x,t) := u(x,t) - v(x,t)$ and $E(x,t) := \int_{-\infty}^{x} e(s,t)ds$, where we assume that $u^0 \in L^1(\mathbb{R})$ to guarantee that E is well defined for all $(x,t) \in \mathbb{R} \times (0,T)$. We have that E also satisfies (4.1) for $n\Delta t < t \le (n+1)\Delta t$ with initial data $\int_{-\infty}^{x} u(s, n\Delta t) - P_h v^n(s)ds$, $n = 0, 1, \ldots, N-1$. For a function $g \in L^1(\mathbb{R})$ and $1 \le p \le \infty$, we define a *minus one* norm in the following way:

$$\|g\|_{-1,p} := \left\| \int_{-\infty}^{\cdot} g(s)\, ds \right\|_{L^p(\mathbb{R})}. \tag{4.5}$$

It is easy to verify that for any $\tau \in \mathbb{R}$

$$\|S_\tau g\|_{-1,p} = \|g\|_{-1,p}. \tag{4.6}$$

Recall that $T = N\Delta t$. Then we have the representations $u(\cdot, T) = (S_{a\Delta t})^N u^0$ and $v(\cdot, T) = (S_{a\Delta t} P_h)^N u^0$. Using (4.6), we have

$$\begin{aligned} \|e(\cdot, T)\|_{-1,p} &= \|(S_{a\Delta t})^N u^0 - (S_{a\Delta t} P_h)^N u^0\|_{-1,p} \\ &= \|(S_{a\Delta t})^{N-1} u^0 - P_h (S_{a\Delta t} P_h)^{N-1} u^0\|_{-1,p}, \end{aligned}$$

and by the triangle inequality we obtain

$$\begin{aligned} \|e(\cdot, T)\|_{-1,p} \le{}& \|(S_{a\Delta t})^{N-1} u^0 - (S_{a\Delta t} P_h)^{N-1} u^0\|_{-1,p} \\ &+ \|P_h (S_{a\Delta t} P_h)^{N-1} u^0 - (S_{a\Delta t} P_h)^{N-1} u^0\|_{-1,p}. \end{aligned} \tag{4.7}$$

Let $e^n = \left((S_{a\Delta t})^n - (S_{a\Delta t} P_h)^n\right) u^0$, $n = 0, 1, \ldots, N$. Then (4.7) is equivalent to

$$\|e^N\|_{-1,p} \le \|e^{N-1}\|_{-1,p} + \|P_h v^{N-1} - v^{N-1}\|_{-1,p}, \tag{4.8}$$

and applying (4.8) for $n = N, N-1, \dots, 1$, we get

$$\|e^N\|_{-1,p} \le \sum_{n=1}^{N-1} \|P_h v^n - v^n\|_{-1,p} \tag{4.9}$$

because $e^0 \equiv 0$. To prove the error estimates, we need the following technical lemma.

LEMMA 4.2. *For any $p \in \{1,2\}$ and any $n = 0, 1, \dots, N$, we have*

$$\text{(i)} \quad \|\{\delta_j^n\}\|_{l_p} \le h^{1-\frac{1}{p}} |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))},$$

$$\text{(ii)} \quad \|P_h v^n - A_h v^n\|_{-1,p} \le \left(\frac{2}{p+1}\right)^{\frac{1}{p}} h^{1+\frac{1}{p}} \|\{\delta_j\}\|_{l_p},$$

$$\text{(iii)} \quad \|A_h v^n - v^n\|_{-1,p} \le \left(\frac{4}{p+1}\right)^{\frac{1}{p}} h^2 |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}.$$

Proof. The inequalities (i) and (ii) follow by standard arguments; therefore, we only prove (i) in the case $p = 2$ and omit the rest because their proofs are similar. Recall that $\delta_j^n = v_j^n - v_{j-1}^n$, and by Theorem 3.1 we have

$$\left(\sum_j (\delta_j^n)^2\right)^{1/2} \le \left(\sum_j (\delta_j^0)^2\right)^{1/2},$$

where $\delta_j^0 = u_j^0 - u_{j-1}^0$, $u_j^0 := \frac{1}{h}\int_{I_j} u^0(s)ds$. Hence, to prove (i) for $p = 2$, we need to prove

$$\sum_j (\delta_j^0)^2 \le h |u^0|^2_{\mathrm{Lip}(1,\mathrm{L}^2(\mathbb{R}))}.$$

Since

$$\sum_j (\delta_j^0)^2 = \sum_j \left(\frac{1}{h}\int_{I_j} (u^0(s+h) - u^0(s))ds\right)^2 \le h^{-2} \sum_j \left(\int_{I_j} |u^0(s+h) - u^0(s)|ds\right)^2,$$

and since by the Cauchy–Schwarz inequality $\left(\int_{I_j} |u^0(s+h) - u^0(s)|ds\right)^2 \le h \int_{I_j} |u^0(s+h) - u^0(s)|^2 ds$, we obtain

$$\sum_j (\delta_j^0)^2 \le h^{-1} \int_{\mathbb{R}} |u^0(s+h) - u^0(s)|^2 ds. \tag{4.10}$$

From (4.3), we have $\int_{\mathbb{R}} |u^0(s+h) - u^0(s)|^2 ds \le h^2 |u^0|^2_{\mathrm{Lip}(1,\mathrm{L}^2(\mathbb{R}))}$, and using that in (4.10), we conclude

$$\sum_j (\delta_j^0)^2 \le h |u^0|^2_{\mathrm{Lip}(1,\mathrm{L}^2(\mathbb{R}))},$$

which proves (i) for $p = 2$. To prove (iii), we note that

$$|A_h v^n - v^n|_{I_j} \le \max_{x \in I_j} v^n(x) - \min_{x \in I_j} v^n(x), \tag{4.11}$$

and because $v^n = S_{a\Delta t} v^{n-1}$, we have that

$$\max_{x \in I_j} v^n(x) - \min_{x \in I_j} v^n(x) \le 2 \max(|\delta_{j-1}^{n-1}|, |\delta_j^{n-1}|).$$

The rest of the proof of (iii) is analogous to the proof of (i). ∎

Combining (i)–(iii), we have $\|P_h v^n - v^n\|_{-1,p} \le Ch^2 |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}$, and after applying the above inequality in (4.9), we derive the following estimate:

$$\|e^N\|_{-1,p} \le CNh^2 |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}. \tag{4.12}$$

Because $v^N \notin \mathrm{Lip}(1, \mathrm{L}^2(\mathbb{R}))$, we approximate v^N by

$$\tilde{v} := \frac{1}{h} \int_{x-h/2}^{x+h/2} A_h v^N(s)\, ds.$$

Similar to Lemma 4.2, it is easy to verify that for $p \in \{1, 2\}$ we have

$$\|\tilde{v} - v^N\|_{-1,p} \le Ch^2 |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}, \tag{4.13}$$

$$\|\tilde{v} - v^N\|_{L^p(\mathbb{R})} \le Ch |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}, \tag{4.14}$$

and

$$|\tilde{v}|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))} \le |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}. \tag{4.15}$$

Let $\tilde{e} := u(\cdot, T) - \tilde{v}$. Then $\|\tilde{e}\|_{-1,p} \le \|e^N\|_{-1,p} + \|\tilde{v} - v^N\|_{-1,p}$, and combining the estimates (4.12) and (4.13), we have

$$\|\tilde{e}\|_{-1,p} \le CNh^2 |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}. \tag{4.16}$$

Kolmogorov–Landau inequalities in $L^p(\mathbb{R})$ (p. 156 in [5]) for the functions $\tilde{E}(x) := \int_\infty^x \tilde{e}(s)\, ds$, $\tilde{E}'$, and $\tilde{E}''$ give

$$\|\tilde{e}\|_{L^p} = \|\tilde{E}'\|_{L^p} \le \sqrt{2} \|\tilde{E}\|_{L^p}^{1/2} \|\tilde{E}''\|_{L^p}^{1/2} = \sqrt{2} \|\tilde{e}\|_{-1,p}^{1/2} |\tilde{e}|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}^{1/2}.$$

Using (4.16) and (4.15), we arrive at

$$\|\tilde{e}\|_{L^p} \le C(Nh)^{1/2} h^{1/2} |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}. \tag{4.17}$$

Finally, by the triangle inequality,

$$\|e\|_{L^p} \le \|\tilde{e}\|_{L^p} + \|\tilde{v} - v^N\|_{L^p(\mathbb{R})} \le C(Nh)^{1/2} h^{1/2} |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))},$$

and we combine (4.14) and (4.17) to conclude

$$\|u(\cdot, T) - v(\cdot, T)\|_{L^p(\mathbb{R})} = \|e\|_{L^p} \le C(Nh)^{1/2} h^{1/2} |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}.$$

Note that C can be computed explicitly and is not very big ($C < 20$). In the case $p = 2$ and $u^0 \notin L^1(\mathbb{R})$, we get the same error estimate via an approximation procedure because the estimate is independent of the L^1 norm. ∎

COROLLARY 4.3. *In the case of $Nh \le C$, we get the convergence rate*

$$\|u(\cdot, T) - v(\cdot, T)\|_{L^p(\mathbb{R})} \le Ch^{1/2} |u^0|_{\mathrm{Lip}(1,\mathrm{L}^p(\mathbb{R}))}$$

for $p = 1$ and $p = 2$.

The L^1 estimate is not new—it follows from the arguments in [19]—but the $1/2$ rate in L^2 is new. Note that, using the L^1 estimate, by interpolation arguments we get only a $1/4$ rate in L^2. The rate $1/2$ is optimal for the case $\theta = 0$ because the numerical method in that case reduces to the LxF scheme, a special case of a monotone scheme. In the case $p = 1$, the sharpness of the $1/2$ bound is given in [20], with an extension to the nonlinear case in [17]. The sharpness in the case $p = 2$ follows from the more general result for formal first-order linear schemes; see [4]. The case $\theta > 0$ is more complicated because the schemes are nonlinear, and it will be addressed elsewhere.

5. Numerical examples. In this section, we present numerical evidence for the new l_2 stability result we proved in section 3. Our numerical tests suggest that in the case of linear flux the NT schemes do not increase the l_2 norm of the jumps for either $\theta \leq 1$ (as proved in Theorem 3.1) or for $1 < \theta \leq 2$. In the case of convex flux, we numerically observe the one-sided analogue of this property. We now give generic examples for this l_2 stability in the linear and convex case.

Example 1. Linear equation. We take a piecewise linear initial condition u^0 (see top left of Figure 5.1) and compare three different approximate solutions. It is easy to see that for a bigger value of θ we get smaller numerical diffusion (see top right panel of Figure 5.1). The other two plots on Figure 5.1 give the behaviors of the l_2 and the l_∞ norms of the jumps in time where the time is rescaled from $[0, 0.15]$ to $[0, 1]$ and the l_2 norm is also rescaled. Note the oscillatory behavior of the l_∞ norm

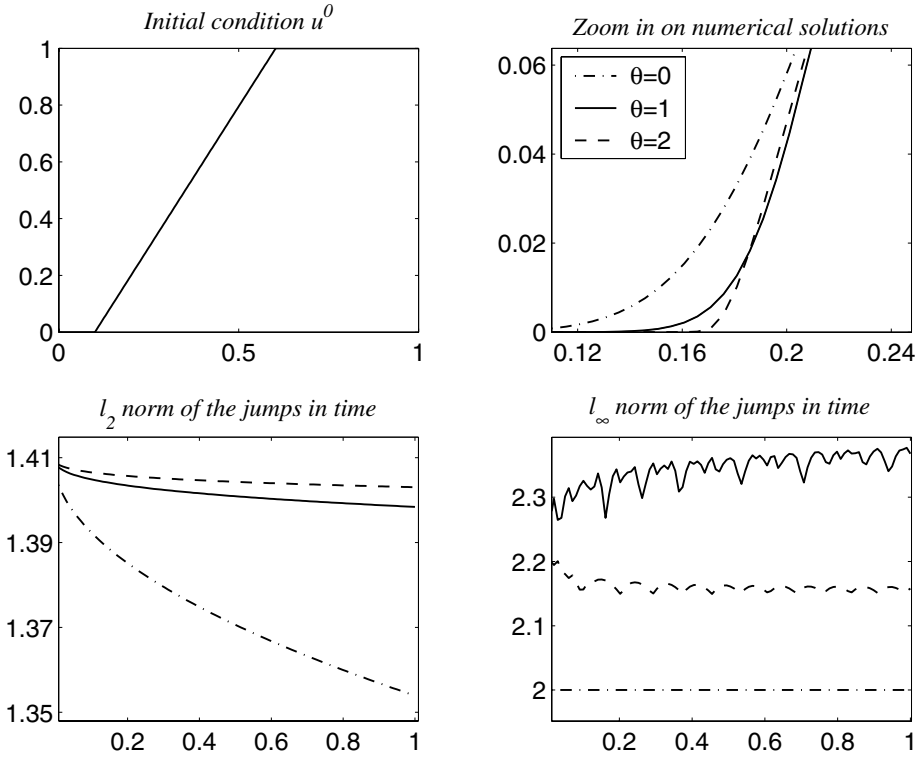


FIG. 5.1. $u_t + 0.5\, u_x = 0$. *The solid line represents* $\theta = 1$, *the dashed line represents* $\theta = 2$, *and the dash-dotted line stands for* $\theta = 0$, *the staggered LxF scheme. The values we used are* $\Delta x = 0.005$, $\lambda = 0.15$, *final time* $T = 0.15$, *and the flux is* $f(u) = 0.5u$.

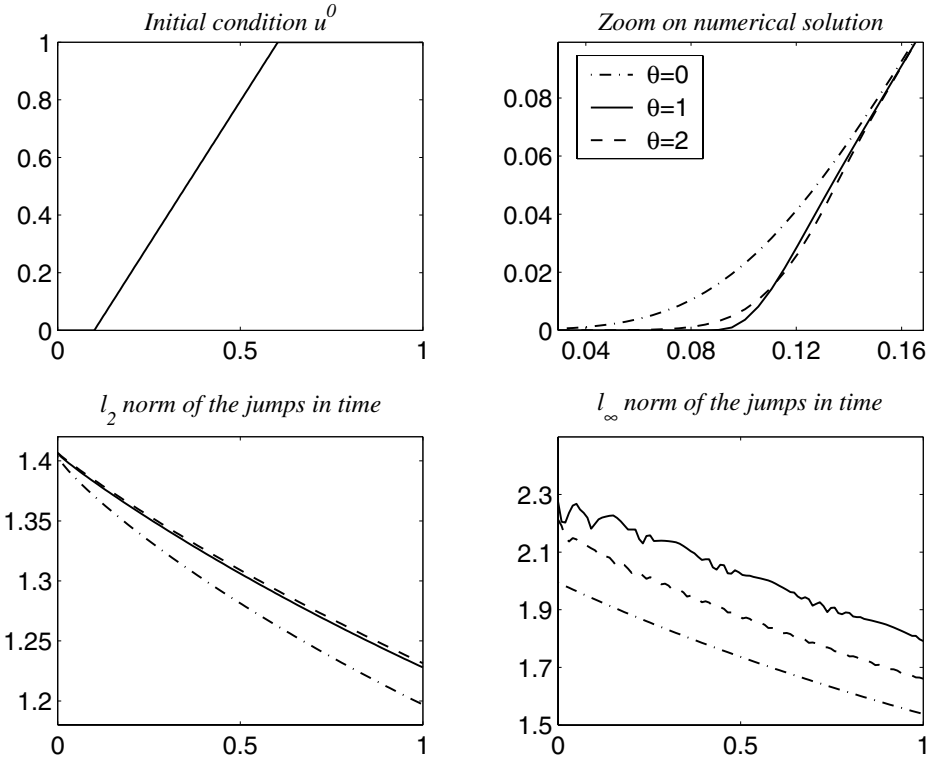


FIG. 5.2. $u_t + (0.5\,u^2)_x = 0$.

and the monotonicity of the l_2 norm for $\theta = 1, 2$.

The presence of shocks or local extrema in the initial data will only make the decrease of the l_2 norm of the jumps faster in the beginning, and then for large time the l_2 norm will decrease very slowly again. In some sense, the total amount of numerical diffusion is given in the decrease of the l_2 norm. In the so-called second-order methods (like $\theta = 1, 2$), the amount of diffusion is much smaller than in first-order methods represented here by the LxF scheme ($\theta = 0$). We will address this issue in a different paper and use it to improve the error estimate for $\theta = 1$.

Example 2. Burgers' equation (see Figure 5.2). We consider the same initial data, numerical schemes, Δx, λ, and T, as in the first example. Note again the oscillatory behavior of the l_∞ norm and the monotonicity of the l_2 norm for $\theta = 1, 2$. This is a generic case of nondecreasing initial data which corresponds to a region of spreading of the characteristics.

The nonlinearity of the flux in such regions helps to decrease overall any norm of the jumps. In the case of a general initial condition, the l_2 norm of the jumps decreases in every region of rarefaction. That is, for convex flux numerical schemes decrease the *one-sided* l_2 norm of the jumps,

$$\sum_j (v_j^{n+1} - v_{j-1}^{n+1})_+^2 \le \sum_j (v_j^n - v_{j-1}^n)_+^2.$$

It is important to note that in the case of convex/concave flux extreme values separate the regions of rarefactions from the regions of shocks, and we observe numerically that

the l_2 norm of the jumps decreases in every interval where the numerical solution is nondecreasing/nonincreasing.

In the nonconvex case (at least one inflection point), the situation is quite different. In one interval of monotonicity we can have both shocks and rarefaction waves. In that case, the NT scheme with $\theta = 2$ converges to a wrong weak solution even for the Buckley–Leverett problem; see Example 3 in [8]. Our numerical tests show that the NT scheme gives a wrong solution to that problem for any value of $\theta \geq 1.2$ and in general it looks like the biggest reliable value of θ for a nonconvex flux is $\theta = 1$.

Acknowledgments. The authors are grateful to Ronald DeVore for his inspiring discussions and constant support. We also thank the anonymous referees. Their comments and suggestions helped improve the paper.

REFERENCES

[1] F. Bouchut, Ch. Bourdarias, and B. Perthame, *A MUSCL method satisfying all entropy inequalities*, Math. Comp., 65 (1996), pp. 1439–1461.

[2] F. Bouchut and B. Perthame, *Kružkov's estimates for scalar conservation laws revisited*, Trans. Amer. Math. Soc., 350 (1998), pp. 2847–2870.

[3] Y. Brenier and S. Osher, *The discrete one-sided Lipschitz condition for convex scalar conservation laws*, SIAM J. Numer. Anal., 25 (1988), pp. 8–23.

[4] P. Brenner, V. Thomée, and L. B. Wahlbin, *Besov Spaces and Applications to Difference Methods for Initial Value Problems*, Lecture Notes in Math. 434, A. Dold and B. Eckmann, eds., Springer-Verlag, Berlin, New York, 1975.

[5] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.

[6] A. Harten and S. Osher, *Uniformly high order accurate non-oscillatory schemes,* I, J. Appl. Numer. Math., 71 (1987), pp. 279–309.

[7] A. Harten, B. Enquist, S. Osher, and S. R. Chakravarthy, *Uniformly high order accurate essentially non-oscillatory schemes,* III, J. Comput. Phys., 71 (1987), pp. 231–303.

[8] G.-S. Jiang, D. Levy, C.-T. Lin, S. Osher, and E. Tadmor, *High-resolution nonoscillatory central schemes with nonstaggered grids for hyperbolic conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2147–2168.

[9] G.-S. Jiang and E. Tadmor, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.

[10] P. Lax and B. Wendroff, *Systems of conservation laws*, Comm. Pure Appl. Math., 13 (1960), pp. 217–237.

[11] K. Kopotun, M. Neamtu, and B. Popov, *Weakly non-oscillatory schemes for scalar conservation laws*, Math. Comp., 72 (2003), pp. 1747–1767.

[12] S. N. Kruzhkov, *First order quasi-linear equations in several independent variables*, Math. USSR Sbornik, 10 (1970), pp. 217–243.

[13] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.

[14] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first order quasi-linear equation*, USSR Comput. Math. Math. Phys., 16 (1976), pp. 105–119.

[15] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.

[16] H. Nessyahu and E. Tadmor, *The convergence rate of approximate solutions for nonlinear scalar conservation laws*, SIAM J. Numer. Anal., 29 (1992), pp. 1505–1519.

[17] F. Şabac, *The optimal convergence rate of monotone finite difference methods for hyperbolic conservation laws*, SIAM J. Numer. Anal., 34 (1997), pp. 2306–2318.

[18] C.-W. Shu, *Numerical experiments on the accuracy of ENO and modified ENO schemes*, J. Comput. Phys., 5 (1990), pp. 127–149.

[19] E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 891–906.

[20] T. Tang and Z.-H. Teng, *The sharpness of Kuznetsov's $O(\sqrt{\Delta x})$ L^1-error estimate for monotone difference scheme*, Math. Comp., 64 (1995), pp. 581–589.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 1998–2019

A MORTAR FINITE ELEMENT METHOD FOR FOURTH ORDER PROBLEMS IN TWO DIMENSIONS WITH LAGRANGE MULTIPLIERS*

LESZEK MARCINKOWSKI†

Abstract. A mortar finite element method with a new Lagrange multipliers space for clamped plate problems is discussed. In the subdomains a conforming Hsieh–Clough–Tocher (HCT) macro element defined on nonmatching triangulations is utilized. The main result of the paper is the proof of an inf-sup condition with new test spaces. Finally, an error bound for Lagrange multipliers is proved.

Key words. plate problem, mortar finite element method, Lagrange multipliers, inf-sup condition, mesh-dependent norm, domain decomposition

AMS subject classifications. 65N15, 74S05, 74K20

DOI. 10.1137/S0036142902387574

1. Introduction. The numerical approximation of partial differential equations is often a very difficult and challenging task. This problem can be solved by using supercomputers; however, their efficiency strongly depends on the utilization of specially adapted numerical algorithms. Domain decomposition methods form a group of such tools, which have been analyzed and successfully used in solving practical problems. One domain decomposition method is the mortar method which allows us to use discretizations of different types with independent discretization parameters in nonoverlapping subdomains. A general presentation of the mortar method in two and three dimensions for elliptic boundary value problems of second order can be found, e.g., in [8], [4], or [5]. In [4], [9], and [29], the mortar methods for second order elliptic problems with saddle point approach were discussed. The mortar approach for discretizations of fourth order elliptic problems was studied in [3], where locally spectral discretizations were utilized, in [21], [22] for discrete Kirchhoff triangle local discretizations, and in [28], [27] for local Hsieh–Clough–Tocher (HCT), reduced HCT, and Adini and Morley finite element discretizations, and some parallel algorithms for solving a discrete problem were considered in [26]. There are also many algorithms for solving problems obtained by the mortar method; cf. [18], [1], [2], [14], [17], and many others.

In this paper, we consider a certain mortar method for local HCT discretizations of plate problems. We introduce a new type of test mortar spaces that contain functions of lower regularity (discontinuous) and prove that they are as good as the ones considered in [28]. The new choice of the test mortar spaces is also due to the more feasible implementation; the mass matrices in the two mortar conditions have a simpler structure. Next we reformulate the mortar discrete problem into a saddle point problem and prove an inf-sup condition in two different norms. Similar types of spaces in the mortar condition have been used for second order elliptic problems in two dimensions in [29] and in three dimensions in [20].

*Received by the editors May 6, 2002; accepted for publication (in revised form) April 27, 2004; published electronically January 20, 2005. This work was partially supported by Polish Scientific grant 2/P03A/005/24.

http://www.siam.org/journals/sinum/42-5/38757.html

†Department of Mathematics, Informatics, and Mechanics, Warsaw University, ul. Banacha 2, 02-097 Warszawa, Poland (lmarcin@mimuw.edu.pl).

In [4] an inf-sup condition in some trace norms defined on interface was proved for second order differential equations, and a few years later in [29] and [9] an inf-sup condition in mesh-dependent norms for second order elliptic problems was stated. We generalized their results to plate problems.

We restrict ourselves to the geometrically conforming version of the mortar method; i.e., the polygonal domain Ω is divided into polygonal subdomains Ω_i which form a coarse triangulation. Locally, in the subdomains the conforming finite element method, i.e., the HCT macro element (cf. [16]), is utilized.

We first introduce independent local discretizations in each subdomain. The two-dimensional triangulations of two neighboring subregions do not necessarily match on their common interface. Then a mortar technique for plate problems which is presented here requires the continuity of the solution at the vertices of subdomains and that the solution on the two neighboring subdomains satisfies two mortar conditions of the L^2 type. We propose two new Lagrange multiplier spaces. We give error bounds analogous to the results of [28]; see also [27].

We also reformulate the mortar discrete problem into a saddle point problem. Then a proper representation of multipliers is given, and error bounds for multipliers are proved in two types of norm, mesh-dependent and dual to proper trace spaces defined on interfaces. We generalize the results of [29], [9], [4], which were obtained for the mortar method for second order elliptic problems, to fourth order problems.

The outline of the paper is as follows. In section 2 we present a differential problem and discuss local discretization of HCT type and the mortar method. Section 3 is devoted to presenting new spaces and proving an error estimate of the mortar method with these new test spaces. Finally, in section 4 inf-sup conditions are proved in both types of norms, and an estimate of an error of Lagrange multipliers is shown.

In the paper the following notation is used: $u \asymp v$, $x \succeq y$, and $w \preceq z$ mean that there exist positive constants c and C independent of the parameter of the fine triangulation of any substructure, and the number of subdomains is such that

$$c\,u \le v \le C\,u, \quad x \ge c\,y, \quad \text{and} \quad w \le C\,z, \ \text{respectively.}$$

2. Discrete problem. Let Ω be a polygonal domain in $\mathbf{R}^2$. The differential problem is to find $u^* \in H_0^2(\Omega)$ such that

$$a(u^*, v) = \int_\Omega f v \, dx \quad \forall v \in H_0^2(\Omega), \tag{2.1}$$

where u^* is the displacement, $f \in L^2(\Omega)$ is the body force, and

$$a(u,v) = \int_\Omega \left[\triangle u \triangle v + (1-\nu)\left(2u_{x_1x_2}v_{x_1x_2} - u_{x_1x_1}v_{x_2x_2} - u_{x_2x_2}v_{x_1x_1}\right)\right] dx.$$

Here

$$H_0^2(\Omega) = \{v \in H^2(\Omega) : \ v = \partial_n v = 0 \ \text{ on } \ \partial\Omega\},$$

∂_n is the normal unit derivative outward to $\partial\Omega$, and $u_{x_ix_j} := \frac{\partial^2 u}{\partial x_i \partial x_j}$ for $i,j = 1,2$. The Poisson ratio ν satisfies $0 < \nu < 1/2$. The Lax–Milgram theorem, utilizing the continuity and ellipticity of the bilinear form $a(\cdot,\cdot)$, yields the existence and the uniqueness of the solution; see, e.g., [11] or [16].

Let Ω be a union of nonoverlapping polygonal subdomains that are arbitrary, i.e.,

$$\overline{\Omega} = \bigcup_{k=1}^{N} \overline{\Omega}_k \ \text{ with } \ \Omega_k \cap \Omega_l = \emptyset, \ \ k \neq l.$$

We assume that the intersection of boundaries of two different subdomains $\partial\Omega_k \cap \partial\Omega_l, k \neq l$, is either the empty set, a vertex, or a common edge. Thus $\{\Omega_k\}$ forms a decomposition of Ω that we call the coarse triangulation with a parameter $H = \max_k H_k$, where $H_k = \operatorname{diam} \Omega_k$. We assume the shape regularity of that decomposition; e.g., cf. section 2, p. 5 in [10]. In the mortar method the interface $\overline{\Gamma} = \bigcup_{k=1}^{N} \overline{\partial\Omega_k \setminus \partial\Omega}$ plays an important role.

We triangulate each subdomain Ω_k into nonoverlapping triangles. We denote an element of this triangulation by τ. We assume that the arising fine triangulation $T_h(\Omega_k)$ is quasi-uniform with parameter $h_k = \max(\operatorname{diam} \tau)$ for $\tau \in T_h(\Omega_k)$; cf. [11].

We now introduce the mortar method that locally uses the HCT macro element; cf. Chapter 7, section 46, p. 279 in [16]. The local finite element space $X_h(\Omega_k)$ is defined by

$$\begin{aligned} X_h(\Omega_k) = \{ v \in C^1(\Omega_k) : \ \ & v_{|\tau} \in P_3(\tau_i) \ \text{ for triangles } \ \tau_i, \ \ i = 1,2,3, \\ & \text{formed by connecting the vertices of } \ \tau \in T_h(\Omega_k) \\ & \text{to its centroid, and } \ v = \partial_n v = 0 \ \text{ on } \ \partial\Omega_k \cap \partial\Omega \}; \end{aligned}$$

cf. [16] and Figure 2.1.

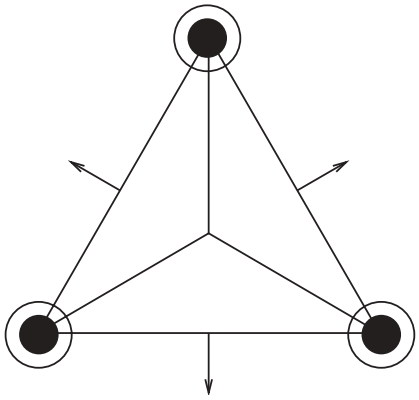


FIG. 2.1. *HCT macro element.*

The degrees of freedom of the HCT element are the following ones: $\{v(p), v_{x_1}(p), v_{x_2}(p), \partial_n v(m)\}$, where p is a vertex of an element and m is a midpoint of an edge of an element (cf. Figure 2.1).

The space $X_h(\Omega_k)$ is equipped with a local bilinear form defined by

$$a_k(u,v) = \int_{\Omega_k} \left[\triangle u \triangle v + (1-\nu) \left(2 u_{x_1 x_2} v_{x_1 x_2} - u_{x_1 x_1} v_{x_2 x_2} - u_{x_2 x_2} v_{x_1 x_1} \right) \right] dx.$$

We next define two global spaces $\tilde{X}_h(\Omega) = \prod_{k=1}^{N} X_h(\Omega_k)$ and its subspace $X_h(\Omega)$ of functions continuous at crosspoints, i.e., at common vertices of subdomains. We

also introduce a bilinear form $a_H(u,v) = \sum_{k=1}^N a_k(u,v)$, a so-called broken norm $\|u\|^2_{H^2_H(\Omega_k)} = \sum_{k=1}^N \|u\|^2_{H^2(\Omega_k)}$, and a broken seminorm $|u|^2_{H^2_H(\Omega_k)} = \sum_{k=1}^N |u|^2_{H^2(\Omega_k)}$.

For each interface $\overline{\Gamma}_{kl} = \partial\Omega_k \cap \partial\Omega_l$, we choose one side as a master (mortar) denoted by $\gamma_{kl} \subset \partial\Omega_k$ and the second one as a slave $\delta_{lk} \subset \partial\Omega_l$. Here the choice of the master side is arbitrary. Thus each interface $\overline{\Gamma}_{kl} = \partial\Omega_k \cap \partial\Omega_l$ has two independent one-dimensional triangulations, the first h_l one: $T^l_h(\delta_{lk})$ inherited from the two-dimensional triangulation of Ω_l and the second h_k one: $T^k_h(\gamma_{kl})$ inherited from the two-dimensional one of Ω_k.

We also introduce an additional notation: Let the set of all vertices of elements of the triangulation of Ω_k, $\overline{\Omega}_k$ $\partial\Omega_k$ and γ_{kl}, δ_{lk} be denoted by $\Omega_{k,h}$, $\overline{\Omega}_{k,h}$, $\partial\Omega_{k,h}$, $\gamma_{kl,h}$, and $\delta_{lk,h}$, respectively.

To properly define a mortar method we have to introduce two test spaces defined on a slave $\delta_{lk} \subset \Gamma$: $M^{1,h_l}(\delta_{lk})$ associated with tangential traces of functions from $X_h(\Omega_l)$ and $M^{2,h_l}(\delta_{lk})$ corresponding to normal traces. We will consider a new type of spaces; cf. [29] for the case of mortar method for second order elliptic equations. The test spaces will be defined below; see section 3.

We say that $u_k \in X_h(\Omega_k)$ and $u_l \in X_h(\Omega_l)$ for $\partial\Omega_l \cap \partial\Omega_k = \overline{\Gamma}_{kl}$ satisfy the mortar conditions if

$$\int_{\delta_{lk}} (u_k - u_l)\psi \, ds = 0 \quad \forall \psi \in M^{1,h_l}(\delta_{lk}) \tag{2.2}$$

and

$$\int_{\delta_{lk}} (\partial_n u_k - \partial_n u_l)\psi \, ds = 0 \quad \forall \psi \in M^{2,h_l}(\delta_{lk}). \tag{2.3}$$

We now define a space V^h as the subspace of $X_h(\Omega)$ formed by functions which satisfy the mortar conditions (2.2) and (2.3) on each interface $\Gamma_{kl} = \delta_{lk} = \gamma_{kl}$. Note that V^h contains functions which are continuous at the crosspoints. This is a technical requirement used in the proof of Lemma 3.14; cf. (3.6).

The discretization of (2.1) in V^h is of the form:
Find $u^*_h \in V^h$ such that

$$a_H(u^*_h, v) = \int_\Omega fv \, dx \quad \forall v \in V^h. \tag{2.4}$$

We now formulate our problem as a saddle point problem. We introduce a new discrete Hilbert space $M_h(\Gamma) = M_1(\Gamma) \times M_2(\Gamma)$ and a bilinear form $b(\cdot,\cdot) : X_h(\Omega) \times M_h(\Gamma) \to \mathbf{R}$, where

$$M_1(\Gamma) = \prod_{\delta_{ji} \subset \Gamma} M^{1,h_j}(\Gamma_{ij}) \ \text{ and } \ M_2(\Gamma) = \prod_{\delta_{ji} \subset \Gamma} M^{2,h_j}(\Gamma_{ij})$$

and

$$b(\psi, u) = \sum_{\delta_{ji} \subset \Gamma} \left(\int_{\Gamma_{ij}} [u]\psi_1 \, ds + \int_{\Gamma_{ij}} [\partial_n u]\psi_2 \, ds \right) \quad \forall u \in X_h(\Omega), \ \ \forall \psi \in M_h(\Gamma).$$

Here $[\cdot]$ is a jump across $\Gamma_{ij} = \delta_{ji} = \gamma_{ij}$, $u = (u_1, \ldots, u_N)$ and $\psi = (\psi_1, \psi_2)$. Then the new problem is to find a pair $(u^*_h, \lambda^*_h) \in X_h(\Omega) \times M_h(\Gamma)$ such that

$$\begin{aligned} a_H(u^*_h, v) + b(\lambda^*_h, v) &= (f, v) \quad \forall v \in X_h(\Omega), \\ b(\psi, u^*_h) &= 0 \qquad \forall \psi \in M_h(\Gamma). \end{aligned} \tag{2.5}$$

We prove below that this problem has a unique solution.

The second equation in (2.5) is equivalent to the mortar conditions (2.2) and (2.3). Thus we have

$$V^h = \{u \in X_h(\Omega) : b(\psi, u) = 0 \ \ \forall \psi \in M_h(\Gamma)\},$$

and the first term of a solution of (2.5) is also a solution of (2.4) and vice versa.

3. New Lagrange multiplier spaces. In this section we introduce a new type of mortar test space for the mortar methods for fourth order problem.

3.1. A new tangential test space. Let $W^{1,h_j}(\Gamma_{ij})$ be the space of traces (tangential) of $X_h(\Omega_j)$ onto δ_{ji} and $W_0^{1,h_j}(\Gamma_{ij}) = H_0^2(\Gamma_{ij}) \cap W^{1,h_j}(\Gamma_{ij})$. We introduce natural nodal bases of them. Basis functions of $W^{1,h_j}(\Gamma_{ij})$ and $W_0^{1,h_j}(\Gamma_{ij})$ are defined by

$$\phi^0_{1,p}(q) = \begin{cases} 1 & p = q, \\ 0 & p \neq q, \end{cases} \quad \partial_s \phi^0_{1,p}(q) = 0,$$

$$\partial_s \phi^1_{1,p}(q) = \begin{cases} 1 & p = q, \\ 0 & p \neq q, \end{cases} \quad \phi^1_{1,p}(q) = 0$$

for $p, q \in \overline{\delta}_{ij,h}$, and $\delta_{ji,h}$, respectively. Here ∂_s is the tangential derivative. We now introduce a mortar test space $M^{1,h}(\delta_{ji}) = \mathrm{Span}\{\theta^0_{1,p}, \theta^1_{1,p} : p \in \delta_{ji,h}\}$ for each interface $\Gamma_{ij} = \gamma_{ij} = \delta_{ji}$. It should be contained in the dual space of $H^{3/2}(\Gamma_{ij})$. The definition of our tangential space is quite simple, namely, let $M^{1,h}(\delta_{ji}) = \mathrm{Span}\{\theta^0_{1,p}, \theta^1_{1,p} : p \in \delta_{ji,h}\}$, where

$$\theta^0_{1,p}(s) = \begin{cases} 1 & \text{for} \quad s \in [m_l, m_r] \\ 0 & \text{for} \quad s \in \overline{\delta}_{ij} \setminus [m_l, m_r] \end{cases},$$

and (cf. Figures 3.1 and 3.2)

$$\theta^1_{1,p}(s) = \begin{cases} (s-p) & \text{for} \quad s \in [m_l, m_r] \\ 0 & \text{for} \quad s \in \overline{\delta}_{ij} \setminus [m_l, m_r] \end{cases}.$$

Here m_l and m_r are the right and left neighboring midpoints of a nodal point p. In the case of a nodal point p next to an end of this slave, we take the respective end instead of m_l in the case of the left end and m_r in the case of the right one. We have $\dim M^{1,h}(\delta_{ji}) = \dim W_0^{1,h_j}(\delta_{ji})$. This property is a very important one.

Note that the space $M^{1,h}(\delta_{ji})$ is much simpler than the one introduced in [28].

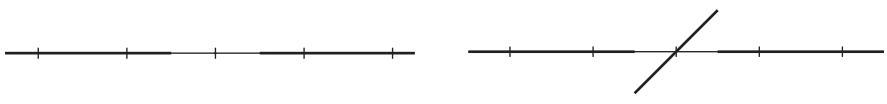


FIG. 3.1. *Tangential basis—interior nodal functions.*

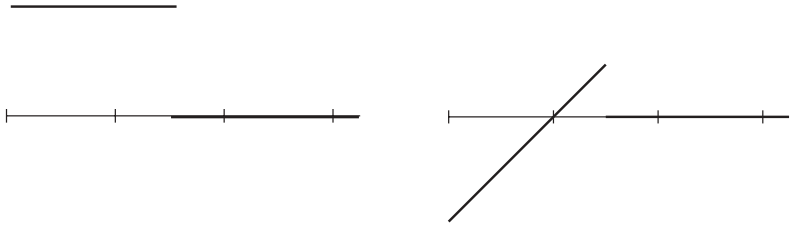


FIG. 3.2. *Tangential basis—near left endpoint of* δ_{ji}.

Let us consider a segment $[0,2]$, the standard cubic Hermitian interpolation basis on this segment, i.e., four functions $\phi_0(x) = 0.25 * (1+x) * (x-2)^2$, $\phi_1(x) = 0.25 * x * (x-2)^2$, $\phi_2(x) = 0.25 * x^2 * (3-x)$, and $\phi_3(x) = 0.25 * x^2 * (x-2)$, and four functions

$$\theta_0(s) = \begin{cases} 1 & \text{for} \quad s \in [0,1] \\ 0 & \text{for} \quad s \in (1,2] \end{cases}, \quad \theta_1(s) = \begin{cases} s & \text{for} \quad s \in [0,1] \\ 0 & \text{for} \quad s \in (1,2] \end{cases},$$

$$\theta_2(s) = \begin{cases} 0 & \text{for} \quad s \in [0,1] \\ 1 & \text{for} \quad s \in (1,2] \end{cases}, \quad \theta_3(s) = \begin{cases} 0 & \text{for} \quad s \in [0,1] \\ s-2 & \text{for} \quad s \in (1,2] \end{cases}.$$

Now we consider a 4×4 matrix $\mathbf{A} = \{\int_0^2 \phi_i \theta_j \, ds\}_{ij}$. We obtain

$$\mathbf{A} = 0.25 * \begin{pmatrix} 13/4 & 29/20 & 3/4 & -11/20 \\ 11/12 & 8/15 & 5/12 & -3/10 \\ 3/4 & 11/20 & 13/4 & -29/20 \\ -5/12 & -3/10 & -11/12 & 8/15 \end{pmatrix}.$$

Simple computation shows that the matrix $(\mathbf{A^T} + \mathbf{A})/2$ is symmetric and positive definite, which is equivalent to

$$(\mathbf{A}\mathbf{u}, \mathbf{u})_{\mathbf{R}^4} > 0 \ \ \forall \mathbf{u} \in \mathbf{R}^4, \ \mathbf{u} \neq 0. \tag{3.1}$$

From this follows a proposition.

PROPOSITION 3.1. *For any* $\tilde{w} = \sum_{p \in \delta_{ji,h}} (w(p)\theta^0_{1,p} + w'(p)\theta^1_{1,p}) \in M^{1,h_j}(\delta_{ji})$ *corresponding to* $w = \sum_{p \in \delta_{ji,h}} (w(p)\phi^0_{1,p} + w'(p)\phi^1_{1,p}) \in W_0^{1,h_j}(\delta_{ji})$ *it holds that*

$$\|w\|^2_{L^2(\delta_{ji})} \asymp \|\tilde{w}\|^2_{L^2(\delta_{ji})} \asymp \int_{\delta_{ji}} w\tilde{w} \, ds.$$

Proof. We show a spectral equivalence of the mass matrices that also proves our proposition: For any vector $\mathbf{w} = (\mathbf{w_0}, \mathbf{w_1})^\mathbf{T}$ corresponding to a function $w = \sum_{p \in \delta_{ji,h}} (w(p)\phi^0_{1,p} + w'(p)\phi^1_{1,p}) \in W_0^{1,h_j}(\delta_{ji})$ it holds that

$$(\mathbf{D}_1\mathbf{w}, \mathbf{w}) \preceq (\mathbf{W}_{1,0}\mathbf{w}, \mathbf{w}) \preceq (\mathbf{M}_1\mathbf{w}, \mathbf{w}) \preceq (\mathbf{D}_1\mathbf{w}, \mathbf{w}),$$

where

$$\mathbf{D}_1 = \begin{pmatrix} \mathbf{D}_{00} & 0 \\ 0 & \mathbf{D}_{11} \end{pmatrix}, \qquad \mathbf{W}_{1,0} = \begin{pmatrix} \mathbf{W}_{00} & 0 \\ 0 & \mathbf{W}_{11} \end{pmatrix}, \qquad \mathbf{M}_1 = \begin{pmatrix} \mathbf{M}_{00} & \mathbf{M}_{01} \\ \mathbf{M}_{10} & \mathbf{M}_{11} \end{pmatrix}$$

for the block diagonal matrices $\mathbf{D}_{kk} = \mathrm{diag}\{\int_{\Gamma_{ij}} (\theta^k_{1,p})^2\, ds\}_{p\in\delta_{ji,h}}$, $\theta^k_{1,p} \in M^{1,h_j}(\delta_{ji})$, and $\mathbf{W}_{kk} = \mathrm{diag}\{\int_{\Gamma_{ij}} (\phi^k_{1,p})^2\, ds\}_{p\in\delta_{ji,h}}$, $\phi^k_{1,p} \in W_0^{1,h_j}(\delta_{ji})$, and for matrices $\mathbf{M}_{kl} = \{\int_{\Gamma_{ij}} \phi^l_{1,p}\theta^k_{1,q}\, ds\}_{p,q\in\delta_{ji,h}}$ for $k, l = 0, 1$.

We first introduce local matrices associated with an element $e \in T^j_h(\delta_{ji})$ and obtained by integrating over this element. Let us define $\mathbf{W}_{kk|e} = \mathrm{diag}\{\int_e (\phi^k_{1,p})^2 ds\}_{p\in\delta_{ji,h}}$, $\mathbf{D}_{kk|e} = \mathrm{diag}\{\int_e (\theta^k_{1,p})^2\, ds\}_{p\in\delta_{ji,h}}$, and $\mathbf{M}_{kl|e} = \{\int_e \phi^l_{1,p}\theta^k_{t,q}\, ds\}_{p,q\in\delta_{ji,h}}$ for $k, l = 0, 1$. We have

$$\mathbf{D}_1 = \sum_{e\in T^j_h(\delta_{ji})} \mathbf{D}_e, \quad \mathbf{W}_{1,0} = \sum_{e\in T^j_h(\delta_{ji})} \mathbf{W}_e, \quad \mathbf{M}_1 = \sum_{e\in T^j_h(\delta_{ji})} \mathbf{M}_e.$$

Then note that

$$(\mathbf{D}_1\mathbf{u}, \mathbf{u}) = \sum_{e\in T^j_h(\delta_{ji})} (\mathbf{D}_e\mathbf{u}, \mathbf{u}), \quad (\mathbf{W}_{1,0}\mathbf{u}, \mathbf{u}) = \sum_{e\in T^j_h(\delta_{ji})} (\mathbf{W}_e\mathbf{u}, \mathbf{u}),$$

$$(\mathbf{M}_1\mathbf{u}, \mathbf{u}) = \sum_{e\in T^j_h(\delta_{ji})} (\mathbf{M}_e\mathbf{u}, \mathbf{u}).$$

The equivalence of $\mathbf{W_e}$ and $\mathbf{D}_e$ is obvious. We have to prove a spectral equivalence of $\mathbf{M}_e$ and $\mathbf{D}_e$. Moving to reference element $[0, 2]$ and utilizing (3.1) (in the case of element e not touching endpoints of δ_{ji}) and Lemma 5, section 4.1, p. 375 in [28] (in the case of end-elements), together with a scaling argument, yields the following bound $(\mathbf{D}_e\mathbf{u}, \mathbf{u}) \preceq (\mathbf{M}_e\mathbf{u}, \mathbf{u})$. The upper bound is obvious. ∎

Additionally we get an approximation property of this space in a standard way.

PROPOSITION 3.2. *The tangential space $M^{1,h}(\delta_{ji})$ has the following approximation property:*

$$\inf_{\psi\in M^{1,h_j}(\Gamma_{ij})} \|u - \psi\|_{L^2(\Gamma_{ij})} \preceq h_j^s |u|_{H^s(\Gamma_{ij})} \quad \forall u \in H^s(\Gamma_{ij}),$$

for $s = 0, \frac{1}{2}, 1, \frac{3}{2}, 2$.

Proof. The proof is standard for $s = 0, 1, 2$ (i.e., for $s = 1, 2$ the square of the L^2 norm over Γ_{ij} is represented as a sum of integrals over disjoint segments $[m_l, m_r]$ (notation as in the definition of $M^{1,h_j}(\delta_{ji})$)), and then using a scaling argument and a quotient space argument (cf., e.g., [16]) ends the proof of the case $s = 1, 2$. Next a Hilbertian interpolation argument yields the proper estimate for remaining cases. ∎

Remark 1. From the proof it can be seen that the statement of Proposition 3.2 can be extended to all $s \in [0.2]$, but in our paper we need only the result of that proposition.

3.2. A normal test space. As in the previous subsection, let $W^{2,h_j}(\Gamma_{ij})$ be the space of traces of normal derivatives of $X_h(\Omega_j)$ onto δ_{ji}, and let $W_0^{2,h_j}(\Gamma_{ij}) = H_0^1(\Gamma_{ij}) \cap W^{2,h_j}(\Gamma_{ij})$. That is, $W^{2,h_j}(\Gamma_{ij})$ is a space of all continuous piecewise quadratic functions on the h_j triangulation of δ_{ji}.

Let $\overline{\delta}_{ji,h/2}$ be the union of $\overline{\delta}_{ji,h}$ and of all midpoints of elements of the h_j triangulation of δ_{ji}, and let $\delta_{ji,h/2} = \overline{\delta}_{ji,h/2} \setminus \partial\delta_{ji}$; i.e., $\delta_{ji,h/2}$ is equal to $\overline{\delta}_{ji,h/2}$ minus the

ends of δ_{ji}. Then, we introduce a standard nodal basis of $W^{2,h_j}(\Gamma_{ij})$ and $W_0^{2,h_j}(\Gamma_{ij})$ as follows: Let

$$\phi_{2,p}(q) = \begin{cases} 1 & p = q, \\ 0 & p \neq q \end{cases} \tag{3.2}$$

for $p, q \in \overline{\delta}_{ji,h/2}$ and $\delta_{ji,h/2}$, respectively.

We now introduce a mortar test space $M^{2,h}(\delta_{ji}) = \mathrm{Span}\{\theta_{2,p},: p \in \delta_{ji,h/2}\}$ for each interface $\Gamma_{ij} = \gamma_{ij} = \delta_{ji}$ which should be contained in $(H^{1/2}(\Gamma_{ij}))'$. We take the one which already has been introduced in [30] where a mortar method for second order elliptic problems was considered. We define basis function of $M^{2,h_j}(\delta_{ji})$ as follows.

Let us first consider a nodal point $p \in \delta_{ji,h/2}$ which is in the closure of a left- or right-end element of δ_{ji}, respectively, denoted by $\overline{e}_l, \overline{e}_r$. Basis functions associated with such points (we have four such points) are defined in a special way: Let m_1 be a midpoint of the left-end element e_l and p_0, p_1, p_2, the three consecutive nodal points of $\delta_{ji,h}$, i.e., p_0, p_1 are the left and right neighbors of m_1, and p_0 is the left end of δ_{ji}; then

$$\theta_{2,m_1}(s) = \begin{cases} 2|p_0 - p_1|^{-1}|s - p_1| & \text{for} \quad s \in [p_0, p_1], \\ 0 & \text{for} \quad s \in \delta_{ji} \setminus [p_0, p_1], \end{cases}$$

and

$$\theta_{2,p_1}(s) = \begin{cases} |p_0 - p_1|^{-1}(2s - p_0 - p_1) & \text{for} \quad s \in [p_0, p_1], \\ \frac{1}{4}(2 + 4\phi_{2,p_1}(s) - 3\phi_{2,p_2}(s)) & \text{for} \quad s \in [p_1, p_2], \\ 0 & \text{for} \quad s \in \delta_{ji} \setminus [p_0, p_2], \end{cases}$$

respectively; cf. Figure 3.3. Here $\phi_{2,p}$ is a nodal basis function defined by (3.2). The two functions associated with two nodal points of $\delta_{ji,h/2}$ nearest to the right end of this slave are given analogously.

For the rest of the nodal points such that $p \in \delta_{ji,h}$ we set

$$\theta_{2,p}(s) = \begin{cases} \frac{1}{2}(5\phi_{2,p}(s) - 2) & \text{for} \quad s \in [p_l, p_r], \\ 0 & \text{for} \quad s \in \delta_{ji} \setminus [p_l, p_r], \end{cases}$$

and for $p \in \delta_{ji,h/2} \setminus \delta_{ji,h}$ (p is a midpoint) the function $\theta_{2,p}$ is defined by

$$\theta_{2,p}(s) = \begin{cases} \frac{1}{4}(2 - 3\phi_{2,p_l}(s) + 4\phi_{2,p_l}(s) - 3\phi_{2,p_l}(s)) & \text{for} \quad s \in [p_l, p_r], \\ 0 & \text{for} \quad s \in \delta_{ji} \setminus [p_l, p_r]. \end{cases}$$

Here $p_l, p_r \in \delta_{ji,h/2}$ are the left and right neighboring nodal points of $p \in \delta_{ji,h/2}$, respectively.

We have $\dim M^{2,h}(\delta_{ji}) = \dim W_0^{2,h_j}(\delta_{ji})$.

The two following propositions can be straightforwardly obtained from the results stated in section 1.2.4.3 in [30]; cf. also Properties (Sb), (Sd), and (Se) in section 1.2 in [30].

PROPOSITION 3.3. *The normal space $M^{2,h_j}(\Gamma_{ij})$ has the following approximation property:*

$$\inf_{\psi \in M^{2,h_j}(\Gamma_{ij})} \|u - \psi\|_{L^2(\Gamma_{ij})} \preceq h^s |u|_{H^s(\Gamma_{ij})} \ \ \forall u \in H^s(\Gamma_{ij}), \ \ s = 0, \frac{1}{2}, \frac{3}{2}.$$

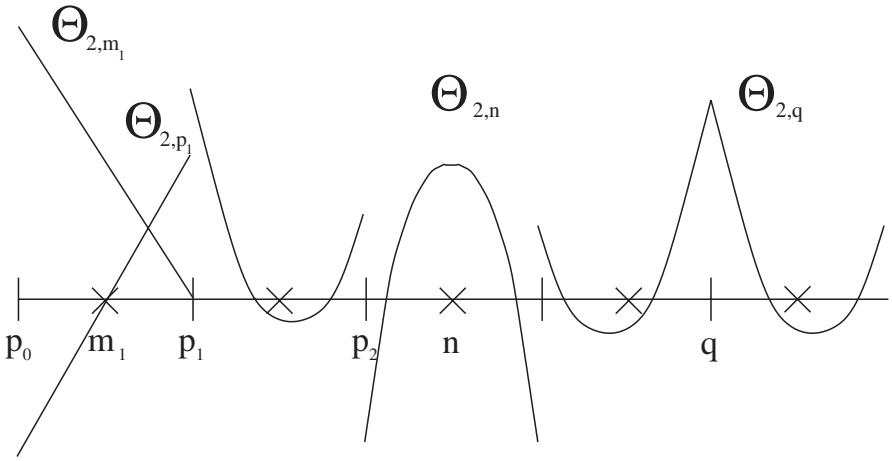


FIG. 3.3. *Normal basis functions.*

PROPOSITION 3.4. *For any* $u = \sum_{p\in\delta_{ji,h}} u(p)\phi_{2,p} \in W_0^{2,h_j}(\delta_{ji})$ *and the corresponding function* $\tilde{u} = \sum_{p\in\delta_{ji,h}} u(p)\theta_{2,p} \in M^{2,h_j}(\delta_{ji})$ *it holds that*

$$\|u\|^2_{L^2(\delta_{ji})} \asymp \|\tilde{u}\|^2_{L^2(\delta_{ji})} \asymp \int_{\delta_{ji}} u\tilde{u}\, ds.$$

3.3. Ellipticity of $a_H(\cdot,\cdot)$ in V^h. In this subsection we state an ellipticity property of $a_H(\cdot,\cdot)$ in V^h in the following proposition.

PROPOSITION 3.5. *For any* $u \in V^h$ *it holds that*

$$c\,\|u\|_{H^2_H(\Omega)} \le |u|_{H^2_H(\Omega)} \le C\,\|u\|_{H^2_H(\Omega)},$$

where c, C *are positive constants independent of* h_i *and the number of subdomains.*

Proof. Because Ω is bounded in $\mathbf{R^2}$ there exists a square $[a,b]\times[c,d]$ of the same diameter as Ω such that $\overline{\Omega} \subset [a,b]\times[c,d]$. Below we denote the extension of u to $[a,b]\times[c,d]$ by zero also by u. For any $(x_1,x_2) \in [a,b]\times[c,d]$

$$u_{x_1}(x_1,x_2) = \int_a^{x_1} u_{x_1x_1}(t,x_2)\,dt + \sum_{a_{kl}\in[a,x_1]\cap\Gamma_{kl}} [u_{x_1}](a_{kl},x_2),$$

where $[\cdot]$ denotes a jump over Γ_{kl} at point a_{kl}. Here a_{kl} is the common point of a segment $[a,x_1]$ and an interface $\Gamma_{kl} \subset \partial\Omega_k \cap \partial\Omega_l$.

By the Schwarz inequality we have

$$\int_a^{x_1} |u_{x_1x_1}(t,x_2)|\,dt \le \sqrt{\operatorname{diam}(\Omega)} \left(\int_a^b |u_{x_1x_1}(t,x_2)|^2\,dt\right)^{1/2}$$

and

$$\sum_{a_{kl}\in[a,x_1]\cap\Gamma_{kl}} |[u_{x_1}](a_{kl},x_2)| \le \left(\sum_{a_{kl}} |\Gamma_{kl}|\right)^{1/2} \left(\sum_{a_{kl}} |\Gamma_{kl}|^{-1} |[u_{x_1}](a_{kl},x_2)|^2\right)^{1/2}.$$

By the shape regularity of the division of Ω into substructures we obtain the estimate $\sum_{a_{kl}\in[a,b]\cap\Gamma_{kl}} |\Gamma_{kl}| \leq C_s\,|b-a|$, where the positive constant C_s depends only on the constant in the shape regularity condition, i.e., is independent of the number of subdomains; see Lemma 2.2 in [7].

Hence integrating over Ω, we get

$$|u|^2_{H^1_H(\Omega)} \preceq |u|^2_{H^2_H(\Omega)} + \sum_{\Gamma_{kl}\subset\Gamma} \frac{1}{|\Gamma_{kl}|} \int_{\Gamma_{kl}} [\partial_n u]^2 + [\partial_s u]^2\, ds.$$

We now have to estimate the second sum. By (2.3) the average values of $\partial_n u_k$ and $\partial_n u_l$ over interface Γ_{kl} are equal to each other. Thus the standard trace theorem and the Poincaré inequality yield that

$$\int_{\Gamma_{kl}} |\partial_n u_k - \partial_n u_l|^2\, ds \preceq \operatorname{diam}(\Omega_k)|u_k|^2_{H^2(\Omega_k)} + \operatorname{diam}(\Omega_l)|u_l|^2_{H^2(\Omega_l)}.$$

Note that $M^{1,h_l}(\delta_{lk})$ contains the space of linear polynomials. Thus by (2.2), a quotient space argument (see, e.g., Theorem 3.1.1, p. 115 in [15]), a scaling argument, and a trace theorem, we get

$$\int_{\Gamma_{kl}} |\partial_s u_k - \partial_s u_l|^2\, ds \preceq \operatorname{diam}(\Omega_k)|u_k|^2_{H^2(\Omega_k)} + \operatorname{diam}(\Omega_l)|u_l|^2_{H^2(\Omega_l)}.$$

Summing over all interfaces concludes the proof of the estimate of the H^1 seminorm. Here we used the fact that from the shape regularity assumption follows $\operatorname{diam}(\Omega_s) \preceq |\Gamma_{kl}|,\ s = k, l$.

The estimate of the L^2 norm can be proved in a similar way. ∎

As a direct consequence we have the following corollary.

COROLLARY 3.6. *The discrete problem* (2.4) *has a unique solution.*

3.4. Approximation property of V^h. In order to prove an approximation property of V^h we have to introduce two types of operators: the first one corresponds to normal parts of traces and was proposed in [30] (in more general form), and the other is associated with tangential traces.

DEFINITION 3.7. *Let* $\Pi^{(2)}_{ij} : L^2(\Gamma_{ij}) \to W_0^{2,h_j}(\delta_{ji})$ *be defined by*

$$\int_{\Gamma_{ij}} \Pi^{(2)}_{ij} uv\, ds = \int_{\Gamma_{ij}} uv\, ds \quad \forall v \in M^{2,h_j}(\delta_{ji}).$$

DEFINITION 3.8. *Let* $\Pi^{(1)}_{ij} : L^2(\Gamma_{ij}) \to W_0^{1,h_j}(\delta_{ji})$ *be defined by*

$$\int_{\Gamma_{ij}} \Pi^{(1)}_{ij} uv\, ds = \int_{\Gamma_{ij}} uv\, ds \quad \forall v \in M^{1,h_j}(\delta_{ji}).$$

We introduce $H^s_{00}(\Gamma_{ij}) = [L^2(\Gamma_{ij}), H^2_0(\Gamma_{ij})]_s$, a scale of Hilbertian interpolation spaces for $s \in [0,2]$. We note that $H^s_{00}(\Gamma_{ij}) = H^s(\Gamma_{ij})$ for $0 \leq s < 1/2$, specifically $H^0_{00}(\Gamma_{ij}) = L^2(\Gamma_{ij})$, and $H^1_{00}(\Gamma_{ij}) = H^1_0(\Gamma_{ij})$; cf. [24], [25]. The double zero in the subscript plays an important role only in the case of $s = 1/2$ and $3/2$.

We now state the most important properties of $\Pi^{(2)}_{ij}$ and $\Pi^{(1)}_{ij}$, namely, stability properties.

LEMMA 3.9. *An operator* $\Pi_{ij}^{(2)}$ *introduced in Definition* 3.7 *is well defined and satisfies*

$$\left\|\Pi_{ij}^{(2)}u\right\|_{H_{00}^s(\Gamma_{ij})} \preceq \|u\|_{H_{00}^s(\Gamma_{ij})} \quad \forall u \in H_{00}^s(\Gamma_{ij})$$

for $s \in [0,1]$.

The proof follows directly from Lemma 1.3 in [30]; cf. also section 1.2.4.1 in [30].

LEMMA 3.10. *An operator* $\Pi_{ij}^{(1)}$ *introduced in Definition* 3.8 *is well defined and, moreover, it holds that*

$$\left\|\Pi_{ij}^{(1)}u\right\|_{H_{00}^s(\Gamma_{ij})} \preceq \|u\|_{H_{00}^s(\Gamma_{ij})} \quad \forall u \in H_{00}^s(\Gamma_{ij})$$

for $s \in [0,2]$.

Proof. From Proposition 3.1 it follows that $\Pi_{ij}^{(1)}u$ is well defined for any $u \in L^2(\Gamma_{ij})$. Let $w = \Pi_{ij}^{(1)}u \in W_0^{1,h_j}(\Gamma_{ij})$ and $w = \sum_{l=0,1}\sum_{p\in\delta_{ji}} w^{(l)}\phi_{1,p}^l$ and define $\tilde{w} = \sum_{l=0,1}\sum_{p\in\delta_{ji}} w^{(l)}\theta_{1,p}^l \in M^{1,h_j}(\Gamma_{ij})$. Using Proposition 3.1 and Definition 3.8 we have

$$\begin{aligned}\|w\|_{L^2(\Gamma_{ij})}^2 &\asymp \int_{\Gamma_{ij}} w\tilde{w}\,ds = \int_{\Gamma_{ij}} u\tilde{w}\,ds \\ &\le \|u\|_{L^2(\Gamma_{ij})}\|\tilde{w}\|_{L^2(\Gamma_{ij})} \asymp \|u\|_{L^2(\Gamma_{ij})}\|w\|_{L^2(\Gamma_{ij})}.\end{aligned}$$

From this it follows that $\Pi_{ij}^{(1)}$ is well defined and stable in the L^2 norm.

We now prove the H_0^2 case. We have

$$\left|\Pi_{ij}^{(1)}u\right|_{H^2(\Gamma_{ij})} \le \left|\Pi_{ij}^{(1)}u - Q_{W_0^1}u\right|_{H^2(\Gamma_{ij})} + |Q_{W_0^1}u|_{H^2(\Gamma_{ij})},$$

where $Q_{W_0^1}$ is the standard $L^2(\Gamma_{ij})$ orthogonal projection onto $W_0^{1,h_j}(\Gamma_{ij})$ and it is a known fact that $Q_{W_0^1}$ is stable in the H^2 seminorm. Thus it suffices to estimate the first term. Utilizing an inverse inequality and the L^2 stability of $\Pi_{ij}^{(1)}$, we get

$$\begin{aligned}\left|\Pi_{ij}^{(1)}u - Q_{W_0^1}u\right|_{H^2(\Gamma_{ij})} &\preceq h_j^{-2}\left\|\Pi_{ij}^{(1)}u - Q_{W_0^1}u\right\|_{L^2(\Gamma_{ij})} \\ &= h_j^{-2}\left\|\Pi_{ij}^{(1)}(u - Q_{W_0^1}u)\right\|_{L^2(\Gamma_{ij})} \\ &\preceq h_j^{-2}\|u - Q_{W_0^1}u\|_{L^2(\Gamma_{ij})} \preceq |u|_{H^2(\Gamma_{ij})}.\end{aligned}$$

The last estimate follows from an approximation property of $Q_{W_0^1}$. We also used the the fact that $\Pi_{ij}^{(1)}$ is a projection (not orthogonal) onto $W_0^{1,h_j}(\Gamma_{ij})$. Using a Hilbertian interpolation argument ends the proof. ∎

LEMMA 3.11. *For any* $u \in H^4(\Omega) \cap H_0^2(\Omega)$, *it holds that*

$$\inf_{v\in V^h} |u - v|_{H_H^2(\Omega)}^2 \preceq \sum_{k=1}^N h_k^4 |u|_{H^4(\Omega_k)}^2.$$

The proof is similar to that of Lemma 4, section 4.1, p. 375 in [28]; therefore we omit some details.

Proof. Let $\tilde{u}$ be a function in $X_h(\Omega)$ (cf., e.g., Theorem 48.1, p. 296 in [16]), such that

$$|u-\tilde{u}_k|_{H^s(\Omega_k)} \preceq h_k^{4-s}|u|_{H^4(\Omega_k)}, \quad s=0,1,2. \tag{3.3}$$

We next define for each interface Γ_{ij} with the master γ_{ij} and respective slave δ_{ji} two functions: $w_{ji} = \Pi_{ij}^{(1)}(\tilde{u}_i - \tilde{u}_j)$ and $\partial_n w_{ji} = \Pi_{ij}^{(2)}(\partial_n \tilde{u}_i - \partial_n \tilde{u}_j)$. Here $\tilde{u}_i, \tilde{u}_j, \partial_n \tilde{u}_i, \partial_n \tilde{u}_j$ are respective traces onto the master γ_{ij} and the slave δ_{ji}. Then we define a global function $w \in X_h(\Omega)$ as follows: on a slave δ_{ji} we set $Tr\, w_{j|\delta_{ji}} = (w_{ji}, \partial_s w_{ji}, \partial_n w_{ji})$. Let Trw_j be equal to zero on all masters contained in $\partial\Omega_j$. We then define $w_j \in X_h(\Omega_j)$ as a discrete extension of Trw_j in $H^2(\Omega_j)$ such that

$$|w_j|_{H^2(\Omega_j)} \preceq |\nabla Trw_{|\partial\Omega_j}|_{H_{00}^{1/2}(\partial\Omega_j)};$$

cf. [12]. Then $v = \tilde{u} + w$ is in V^h, which follows from (2.2), (2.3), and Definitions 3.7 and 3.8. We have

$$|u-v|_{H_H^2(\Omega)} \le |u-\tilde{u}|_{H_H^2(\Omega)} + |w|_{H_H^2(\Omega)}.$$

Note that the first term was estimated by (3.3).

We now estimate the seminorm of w. We have

$$|w|^2_{H_H^2(\Omega)} \preceq \sum_{\Gamma_{ij}\subset\Gamma} \|\nabla w_j\|^2_{H_{00}^{1/2}(\delta_{ji})} \preceq \sum_{\Gamma_{ij}} \{\|\partial_s w_{ji}\|^2_{H_{00}^{1/2}(\delta_{ji})} + \|\partial_n w_{ji}\|^2_{H_{00}^{1/2}(\delta_{ji})}\}.$$

Thus by Lemmas 3.9 and 3.10 we have

$$|w|^2_{H_H^2(\Omega)} \preceq \sum_{\Gamma_{ij}\subset\Gamma} \|\nabla(\tilde{u}_i - \tilde{u}_j)\|^2_{H_{00}^{1/2}(\Gamma_{ij})} \preceq \sum_{\Gamma_{ij}\subset\Gamma} \sum_{k=i,j} \|\nabla(\tilde{u}_k - u)\|^2_{H_{00}^{1/2}(\Gamma_{ij})}.$$

Utilizing (3.3) and afterwards summing the resulting inequalities over all interfaces yields

$$|w|^2_{H_H^2(\Omega)} \preceq \sum_{\Gamma_{ij}\subset\Gamma} \left[h_i^4|u|^2_{H^4(\Omega_i)} + h_j^4|u|^2_{H^4(\Omega_j)}\right] \preceq \sum_{k=1}^{N} h_k^4|u|^2_{H^4(\Omega_k)}.$$

This concludes the proof. ∎

3.5. Consistency error. We first state two technical results.

LEMMA 3.12. *Let $Q_{1,ij}$ be the L^2 orthogonal projection onto the test space $M^{1,h_j}(\delta_{ji})$ defined on a slave $\delta_{ji} \subset \Gamma_{ij}$. Then, for $u \in H^r(\Gamma_{ij})$, it holds that*

$$h_j^s\|u - Q_{1,ij}u\|_{L^2(\Gamma_{ij})} + \|u - Q_{1,ij}u\|_{H^{-s}(\Gamma_{ij})} \preceq h_j^{s+r}|u|_{H^r(\Gamma_{ij})}, \quad s,r = \frac{1}{2}, \frac{3}{2}. \tag{3.4}$$

Proof. A proof of the estimate of the L^2 norm follows directly from Proposition 3.2. The estimates of the dual norms are proved using Proposition 3.2 and a duality trick:

$$\begin{aligned}
\|u - Q_{1,ij}u\|_{H^{-s}(\Gamma_{ij})} &= \sup_{\|\psi\|_{H^s(\Gamma_{ij})}\le 1} |(u - Q_{1,ij}u, \psi)| \\
&= \sup_{\|\psi\|_{H^s(\Gamma_{ij})}\le 1} |(u - Q_{1,ij}u, \psi - Q_{1,ij}\psi)| \\
&\le \sup_{\|\psi\|_{H^s(\Gamma_{ij})}\le 1} \|u - Q_{1,ij}u\|_{L^2(\Gamma_{ij})}\|\psi - Q_{1,ij}\psi\|_{L^2(\Gamma_{ij})} \\
&\preceq h_j^{s+r}|u|_{H^r(\Gamma_{ij})}.
\end{aligned}$$

The proof is complete. ∎

The next lemma states an analogous property of $Q_{2,ij}$, the L^2 orthogonal projection onto the normal space $M^{2,h_j}(\delta_{ji})$.

LEMMA 3.13. *Let $Q_{2,ij}$ be the L^2 orthogonal projection onto the test space $M^{2,h_j}(\delta_{ji})$ defined on a slave $\delta_{ji} \subset \Gamma_{ij}$. It holds that*

$$h_j^s \|u - Q_{2,ij}u\|_{L^2(\Gamma_{ij})} + \|u - Q_{2,ij}u\|_{H^{-s}(\Gamma_{ij})} \preceq h_j^{s+r} |u|_{H^r(\Gamma_{ij})}, \quad s, r = \frac{1}{2}, \frac{3}{2}. \tag{3.5}$$

The proof is very similar to the one of the previous lemma and follows from Proposition 3.3.

In the following lemma bounds for the consistency errors are given.

LEMMA 3.14. *Let u^*, a solution of (2.1), be in $H^4(\Omega) \cap H_0^2(\Omega)$. Then for $w \in V^h$ it holds that*

$$|a_H(u^* - u_h^*, w)| \preceq |w|_{H_H^2(\Omega)} \left(\sum_{k=1}^{N} h_k^4 |u^*|_{H^4(\Omega_k)}^2 \right)^{1/2},$$

where u_h^* is a solution of (2.4).

Proof. Using the Green's formulas (e.g., see (1.2.5) and (1.2.9), pp. 14–15 in [15]) and (2.4), we have

$$a_H(u^* - u_h^*, w) = a_H(u^*, w) - f(w) = \sum_{k=1}^{3} E_k(u^*, w),$$

where

$$E_1(u^*, w) = \int_\Gamma -\partial_n(\triangle u^*)[w]\, ds, \qquad E_2(u^*, w) = \int_\Gamma (1-\nu)\partial_n \partial_s u^* [\partial_s w]\, ds,$$

and

$$E_3(u^*, w) = \int_\Gamma (\triangle u^* - (1-\nu)\partial_s^2 u^*)[\partial_n w]\, ds.$$

We note that ∂_n, ∂_s are normal and tangential derivatives, while $[\cdot]$ is the jump across the interface Γ.

Utilizing the fact that $[w]$ is equal to zero at the ends of any interface Γ_{kl}, we obtain

$$E_2(u^*, w) = -\sum_{\Gamma_{ij}} (1-\nu) \int_{\Gamma_{ij}} \partial_s \partial_n \partial_s u^* [w]\, ds. \tag{3.6}$$

Let $E_0(u^*, w) = E_1(u^*, w) + E_2(u^*, w)$ and $E_0(u^*, w) = \int_\Gamma G_3 u^* [w]\, ds$, where $G_3 u^* = -\partial_n \triangle u^* - (1-\nu)\partial_s \partial_n \partial_s u^*$.

We now consider one interface Γ_{ij} which is equal geometrically to the mortar γ_{ij} and slave δ_{ji}. The mortar condition (2.2) yields that $Q_{1,ij} w_i = Q_{1,ij} w_j$ and

$$\begin{aligned} \int_{\Gamma_{ij}} G_3 u^* [w]\, ds &= \int_{\Gamma_{ij}} \left((I - Q_{1,ij}) G_3 u^* \right) [w]\, ds \\ &\preceq \|(I - Q_{1,ij}) G_3 u^*\|_{H^{-3/2}(\Gamma_{ij})} \left(\sum_{k=i,j} |w_k|_{H^{3/2}(\Gamma_{ij})} \right). \end{aligned}$$

Using the Schwarz inequality, Lemma 3.12, and the standard trace theorem (e.g., cf. Theorem 1.5.2.1, p. 42 in [19]), we have

$$\int_{\Gamma_{ij}} G_3 u^*[w]\, ds \preceq h_j^2 |u^*|_{H^4(\Omega_j)} (|w_i|_{H^2(\Omega_i)} + |w_j|_{H^2(\Omega_j)}).$$

Summing over all interfaces yields the estimate of $E_0(u^*, w)$.

Let $G_2 u^* = \triangle u^* - (1-\nu)(\partial_s^2 u^*)$; then proceeding similarly to above and using (2.3) instead of (2.2) and Lemma 3.13 we obtain

$$\begin{aligned}\int_{\Gamma_{ij}} G_2 u^*[\partial_n w]\, ds &\le \|(I - Q_{2,ij}) G_2 u^*\|_{H^{-1/2}(\delta_{ji})} \left(\sum_{k=i,j} |\partial_n w_k|_{H^{1/2}(\delta_{ji})} \right) \\ &\preceq h_j^2 |G_2 u^*|_{H^{3/2}(\delta_{ji})} (|\partial_n w_i|_{H^{1/2}(\delta_{ji})} + |\partial_n w_j|_{H^{1/2}(\delta_{ji})}) \\ &\preceq h_j^2 |u^*|_{H^4(\Omega_j)} (|w_i|_{H^2(\Omega_i)} + |w_j|_{H^2(\Omega_j)}).\end{aligned}$$

Summing over all interfaces yields the estimate of $E_3(u^*, w)$.

The proof is complete. ∎

As a simple consequence of the second Strang lemma (see, e.g., Lemma 8.1.9, p. 198 in [11] or [6]), Lemmas 3.11 and 3.14, and Proposition 3.5 we get an optimal error estimate that is stated in the following theorem.

THEOREM 3.15. *Let u^* and u_h^*, the solutions of* (2.1) *and of* (2.4)*, respectively, be in $H_0^2(\Omega) \cap H^4(\Omega)$. Then it holds that*

$$\|u^* - u_h^*\|_{H_H^2(\Omega)} \preceq \left(\sum_{k=1}^N h_k^4 |u^*|_{H^4(\Omega_k)}^2 \right)^{1/2}. \tag{3.7}$$

4. Inf-sup condition. In this section we prove that under our assumptions an inf-sup condition holds for (2.5). An interpretation of Lagrange multipliers is given. We follow [4] and [9] where second order elliptic problems were considered.

We first introduce a space $X_{00}(\Omega) \subset \prod_{k=1}^N H_C^2(\Omega_k)$ as follows:

$$\begin{aligned}X_{00}(\Omega) = \Bigg\{ & v \in \prod_{k=1}^N H_C^2(\Omega_k) : v \text{ continuous at crosspoints}, \\ & [v] \in H_{00}^{3/2}(\Gamma_{ij}), \text{ and } [\partial_n v] \in H_{00}^{1/2}(\Gamma_{ij}) \ \forall \Gamma_{ij} \subset \Gamma \Bigg\},\end{aligned} \tag{4.1}$$

where

$$H_C^2(\Omega_k) = \{ v \in H^2(\Omega_k) : v = \partial_n v = 0 \text{ on } \partial\Omega \cap \partial\Omega_k \}.$$

Remark 2. The space $X_{00}(\Omega)$ is not closed in the norm $\|\cdot\|_{H_H^2(\Omega)}$. The proof is very similar to that of Remark 2.1 in [9].

Therefore the space $X_{00}(\Omega)$ is endowed with the following norm:

$$\|v\|_X^2 = \|v\|_{H_H^2(\Omega)}^2 + \sum_{\Gamma_{ij} \subset \Gamma} \left\{ \|[v]\|_{H_{00}^{3/2}(\Gamma_{ij})}^2 + \|[\partial_n v]\|_{H_{00}^{1/2}(\Gamma_{ij})}^2 \right\}. \tag{4.2}$$

We have the following proposition.

PROPOSITION 4.1. *The space $X_{00}(\Omega)$ endowed with the norm $\|\cdot\|_X$ is complete; moreover, the following continuous embedding holds:*

$$H_0^2(\Omega) \subset X_{00}(\Omega) \subset \prod_{k=1}^{N} H_C^2(\Omega).$$

This proposition is analogous to Remark 2.2 in [9], where the case of H^1 space is considered. The proof is also quite similar to that in [9]; therefore it is skipped.

The space $H_0^2(\Omega)$ can be described as a subspace of $X_{00}(\Omega)$ formed by all functions for which

$$\begin{aligned} {}_{H_{00}^{-3/2}(\Gamma_{ij})}\langle \mu_1, [v] \rangle_{H_{00}^{3/2}(\Gamma_{ij})} &= 0 \quad \forall \mu_1 \in H_{00}^{-3/2}(\Gamma_{ij}), \\ {}_{H_{00}^{-1/2}(\Gamma_{ij})}\langle \mu_2, [\partial_n v] \rangle_{H_{00}^{1/2}(\Gamma_{ij})} &= 0 \quad \forall \mu_2 \in H_{00}^{-1/2}(\Gamma_{ij}). \end{aligned} \tag{4.3}$$

Here $H_{00}^{-3/2}(\Gamma_{ij})$ and $H_{00}^{-1/2}(\Gamma_{ij})$ denote the dual spaces of $H_{00}^{3/2}(\Gamma_{ij})$ and $H_{00}^{1/2}(\Gamma_{ij})$, respectively.

We introduce a space of Lagrange multipliers $M(\Gamma) = \prod_{\Gamma_{ij}\subset\Gamma} H_{00}^{-3/2}(\Gamma_{ij}) \times H_{00}^{-1/2}(\Gamma_{ij})$ endowed with the norm

$$\begin{aligned} \|\psi\|_M^2 &= \sum_{\Gamma_{ij}\subset\Gamma} \|\psi_{|\delta_{ji}}\|^2_{H_{00}^{-3/2}(\Gamma_{ij})\times H_{00}^{-1/2}(\Gamma_{ij})} \\ &= \sum_{\Gamma_{ij}\subset\Gamma} \left\{ \|\psi_1\|^2_{H_{00}^{-3/2}(\Gamma_{ij})} + \|\psi_2\|^2_{H_{00}^{-1/2}(\Gamma_{ij})} \right\}. \end{aligned} \tag{4.4}$$

Let us consider a saddle point problem: Find a pair $(u^*, \lambda^*) \in X_{00}(\Omega) \times M(\Gamma)$ such that

$$\begin{aligned} a_H(u^*, v) + b(\lambda^*, v) &= (f, v) \ \ \forall v \in X_{00}(\Gamma), \\ b(\mu, u^*) &= 0 \ \ \forall \mu \in M(\Gamma), \end{aligned} \tag{4.5}$$

where

$$b(\mu, v) = \sum_{\Gamma_{ij}\subset\Gamma} {}_{H_{00}^{-3/2}(\Gamma_{ij})}\langle \mu_1, [v] \rangle_{H_{00}^{3/2}(\Gamma_{ij})} + {}_{H_{00}^{-1/2}(\Gamma_{ij})}\langle \mu_2, [\partial_n v] \rangle_{H_{00}^{1/2}(\Gamma_{ij})}$$

Then we have the following theorem.

THEOREM 4.2. *The problem* (4.5) *has a unique solution since the following inf-sup condition holds:*

$$\inf_{\psi\in M(\Gamma)\setminus\{0\}} \sup_{u\in X_{00}(\Omega)\setminus\{0\}} \frac{b(\psi, u)}{\|\psi\|_M \|u\|_X} \geq C. \tag{4.6}$$

Moreover, the first term of the solution of (4.5) *is also a solution of* (2.1) *and if we assume that the solution of* (2.1) u^* *is in* $H_0^2 \cap H^4(\Omega)$*, then*

$$\lambda^* = (\lambda_1^*, \lambda_2^*) = (-\partial_n \triangle u^* - (1-\nu)\partial_s \partial_n \partial_s u^*, \triangle u^* - (1-\nu)\partial_s^2 u^*) \ \text{ on } \ \Gamma_{ij} \subset \Gamma.$$

Proof. The last statement of the theorem follows from the Green's integral formulas; cf., e.g., [16].

We obviously have that $b(\cdot,\cdot)$ is continuous over $X_{00}(\Omega)\times M(\Gamma)$ and that $a_H(\cdot,\cdot)$ is continuous over $X_{00}(\Omega)$ and from (4.3) we see that $a_H(\cdot,\cdot)$ is elliptic over

$$H_0^2(\Omega)=\{v\in X_{00}(\Omega): b(\mu,v)=0\ \ \forall \mu\in M(\Gamma)\}.$$

Thus there is a unique $u^*\in H_0^2(\Omega)$ which is the first part of a solution of (4.5). In order to get the existence of the second part of this solution, the Babuška–Brezzi–Ladyzhenskaya condition has to be proved.

Let $\mu\in M(\Gamma)$, $\mu=\{\mu_{ij}\}_{\Gamma_{ij}\subset\Gamma}$ for $\mu_{ij}=(\mu_1,\mu_2)\in H_{00}^{3/2}(\delta_{ji})\times H_{00}^{1/2}(\delta_{ji})$. For each interface $\Gamma_{ij}\subset\Gamma$, with jth the slave side of Γ_{ij}, we introduce a function $w_{ij}\in H^2_{0,\Gamma_{ij}}(\Omega_j)$ as a unique solution of

$$a_j(w_{ij},v)=\int_{\Gamma_{ij}}\mu_1 v\,ds+\int_{\Gamma_{ij}}\mu_2\partial_n v\,ds\quad\forall v\in H^2_{0,\delta_{ji}}(\Omega_j). \tag{4.7}$$

Here $a_j(\cdot,\cdot)=a_{|\Omega_j}(\cdot,\cdot)$ is the restriction of bilinear form $a(\cdot,\cdot)$ to the subregion Ω_j,

$$H^2_{0,\delta_{ji}}(\Omega_j)=\{u\in H^2(\Omega_j): u=\partial_n u=0 \text{ on } \partial\Omega_j\setminus\delta_{ji}\}, \tag{4.8}$$

and $\int_{\Gamma_{ij}}\mu_1 v\,ds$ and $\int_{\Gamma_{ij}}\mu_2\partial_n v\,ds$ denote the dual pairs for $H_{00}^{3/2}(\Gamma_{ij})$ and $H_{00}^{1/2}(\Gamma_{ij})$, respectively. The standard trace and extension theorems yield that

$$\|\mu_{|\Gamma_{ij}}\|_{H_{00}^{-3/2}(\delta_{ji})\times H_{00}^{-1/2}(\delta_{ji})}\asymp\sup_{|u|_{H^2(\Omega)}=1}\left(\int_{\delta_{ji}}\mu_1 u\,ds+\int_{\delta_{ji}}\mu_2\partial_n u\,ds\right),$$

where sup is taken over $H^2_{0,\delta_{ji}}(\Omega_j)$. Thus

$$\begin{aligned}\|\mu_{|\Gamma_{ij}}\|^2_{H_{00}^{-3/2}(\Gamma_{ij})\times H_{00}^{-1/2}(\Gamma_{ij})}&\asymp |w_{ij}|^2_{H^2(\Omega_j)}\\&=\int_{\Gamma_{ij}}\mu_1 w_{ij}\,ds+\int_{\Gamma_{ij}}\mu_2\partial_n w_{ij}\,ds.\end{aligned} \tag{4.9}$$

Then, we define a function $w=\{\sum_{\delta_{ji}\subset\partial\Omega_j}w_{ij}\}_{j=1}^N$. We have $w\in X_{00}(\Omega)$ as $[w]=w_{ij}$ and $[\partial_n w]=\partial_n w_{ij}$ on Γ_{ij}. By this, the Friedrich's theorem, and trace theorems we have

$$\|w\|_X^2\preceq\sum_{\Gamma_{ij}\subset\Gamma}|w_{ij}|^2_{H^2(\Omega_j)}\preceq\sum_{\Gamma_{ij}\subset\Gamma}\|\mu_{|\Gamma_{ij}}\|^2_{H_{00}^{-3/2}(\Gamma_{ij})\times H_{00}^{-1/2}(\Gamma_{ij})}=\|\mu\|_M^2.$$

Then

$$\|\mu\|_M^2=\sum_{\Gamma_{ij}\subset\Gamma}\|\mu_{|\Gamma_{ij}}\|^2_{H_{00}^{-3/2}(\Gamma_{ij})\times H_{00}^{-1/2}(\Gamma_{ij})}\preceq b(\mu,w)\preceq b(\mu,w)\frac{\|\mu\|_M}{\|w\|_X}.$$

Thus the inf-sup condition is proved. We can conclude that there is a unique solution of (4.5); see [13]. ∎

We next prove the following theorem which gives us a discrete inf-sup condition.

THEOREM 4.3. *For problem* (2.5) *the following inf-sup condition holds:*

$$\inf_{\psi\in M_h(\Gamma)\setminus\{0\}}\sup_{u\in X_h(\Omega)\setminus\{0\}}\frac{b(\psi,u)}{\|\psi\|_M\|u\|_X}\geq C, \tag{4.10}$$

where C is a constant independent of any h_k and the number of subdomains.

Proof. Let us consider a function $\psi \in M_h(\Gamma)$ and $\psi_{ij} = (\psi_1, \psi_2) \in M_h(\delta_{ji})$, its restriction to an interface $\Gamma_{ij} = \delta_{ji} = \gamma_{ij}$. Here $H^2_{0,\delta_{ji}}(\Omega_j)$ was defined in (4.8). Next, as in the proof of the previous theorem, we introduce an auxiliary function $w = \{\sum_{\delta_{ji}\subset\partial\Omega_j} w_{ij}\}_{j=1}^N$, where w_{ij} is a function defined locally on Ω_j ($\delta_{ji} \subset \partial\Omega_j$), as a unique solution of (4.7) with μ_k replaced by ψ_k for $k = 1, 2$. Again we have

$$\|\psi_{ij}\|^2_{H_{00}^{-3/2}(\delta_{ji})\times H_{00}^{-1/2}(\delta_{ji})} \asymp |w_{ij}|^2_{H^2(\Omega_j)} = \int_{\delta_{ji}} w_{ij}\psi_1\, ds + \int_{\delta_{ji}} \partial_n w_{ij}\psi_2\, ds. \tag{4.11}$$

We now construct a discrete function $w_h = \{\sum_{\delta_{ji}\subset\partial\Omega_j} w_{h,ij}\}_{j=1}^N \in X_h(\Omega)$ associated with w.

We define $w_{h,ij}$ for $\delta_{ji} \subset \partial\Omega_j$ locally over Ω_j as follows:

$$\begin{cases} a_j(w_{h,ij}, v) = 0 & \forall v \in X_{0,h}(\Omega_j), \\ Tr\, w_{h,ij} = \left(\Pi^{(1)}_{ij} w_{ij}, \partial_s\left(\Pi^{(1)}_{ij} w_{ij}\right), \Pi^{(2)}_{ij}\partial_n w_{ij}\right) & \text{on } \delta_{ji}, \\ Tr\, w_{h,ij} = 0 & \text{on } \partial\Omega_j \setminus \delta_{ji}, \end{cases} \tag{4.12}$$

where $Tr\, v = (v_{|\partial\Omega_j}, \nabla v_{|\partial\Omega_j})$ for $v, \nabla v$—the H^2 traces of $v \in H^2(\Omega_j)$, and $X_{0,h}(\Omega_j) = X_h(\Omega_j) \cap H^2_0(\Omega_j)$.

Now from the Friedrich's inequality, the discrete extension theorem (see, e.g., [23]), Lemmas 3.10 and 3.9, and the standard trace theorem we obtain

$$\begin{aligned} \|w_{h,ij}\|^2_{H^2(\Omega_j)} &\preceq \left\|\partial_s\left(\Pi^{(1)}_{ij} w_{ij}\right)\right\|^2_{H_{00}^{1/2}(\delta_{ji})} + \left\|\Pi^{(2)}_{ij}\partial_n w_{ij}\right\|^2_{H_{00}^{1/2}(\delta_{ji})} \\ &\preceq \|\partial_s w_{ij}\|^2_{H_{00}^{1/2}(\delta_{ji})} + \|\partial_n w_{ij}\|^2_{H_{00}^{1/2}(\delta_{ji})} \preceq |w_{ij}|^2_{H^2(\Omega_j)}. \end{aligned} \tag{4.13}$$

Note that

$$[w_h] = w_{h,ij} \text{ and } [\partial_n w_h] = \partial_n w_{h,ij} \text{ on } \delta_{ji} \subset \Gamma.$$

Thus

$$\begin{aligned} \|[\partial_s w_h]\|^2_{H_{00}^{1/2}(\delta_{ji})} + \|[\partial_n w_h]\|^2_{H_{00}^{1/2}(\delta_{ji})} &= \|\partial_s w_{h,ij}\|^2_{H_{00}^{1/2}(\delta_{ji})} + \|\partial_n w_{h|\delta_{ji}}\|^2_{H_{00}^{1/2}(\delta_{ji})} \\ &\preceq |w_{ij}|^2_{H^2(\Omega_j)}. \end{aligned}$$

Then summing over all interfaces we have

$$\|w_h\|^2_X \preceq \sum_{\delta_{ji}\subset\Gamma} |w_{ij}|^2_{H^2(\Omega_j)} \preceq \|\psi\|^2_M. \tag{4.14}$$

We next see that by Definitions 3.7 and 3.8

$$\begin{aligned} \|\psi\|^2_M &\asymp \sum_{\delta_{ji}\subset\Gamma} \left\{ \int_{\delta_{ji}} w_{ij}\psi_1\, ds + \int_{\delta_{ji}} \partial_n w_{ij}\psi_2\, ds \right\} \\ &= \sum_{\delta_{ji}\subset\Gamma} \left\{ \int_{\delta_{ji}} \Pi^{(1)}_{ij} w_{ij}\psi_1\, ds + \int_{\delta_{ji}} \Pi^{(2)}_{ij}\partial_n w_{ij}\psi_2 \right\} ds \end{aligned}$$

$$= \sum_{\delta_{ji}\subset\Gamma} \left\{ \int_{\delta_{ji}} w_{h,ij}\psi_1 \, ds + \int_{\delta_{ji}} \partial_n w_{h,ij}\psi_2 \right\} ds = b(\psi, w_h)$$

$$\preceq b(\psi, w_h) \frac{\|\psi\|_M}{\|w_h\|_X}.$$

The proof is completed. □

We also consider the mesh-dependent norms and introduce a series of mesh-dependent norms defined on a single slave $\delta_{ji} \subset \partial\Omega_j$ for any $\psi \in L^2(\delta_{ji})$:

$$\|\psi\|^2_{s,h,\delta_{ji}} = h_j^{-2s} \|\psi\|^2_{L^2(\delta_{ji})}, \quad s \in \mathbf{R}. \tag{4.15}$$

Certainly, we have

$$(u, v)_{L^2(\delta_{ji})} \leq \|u\|_{-s,h,\delta_{ji}} \|v\|_{s,h,\delta_{ji}}.$$

Let the mesh-dependent norm on $M_h(\Gamma)$ be defined by

$$\begin{aligned} \|\psi\|^2_{M,h} &= \sum_{\delta_{ji}\subset(\Gamma)} \{\|\psi_{t|\delta_{ji}}\|^2_{-3/2,h,\delta_{ji}} + \|\psi_{n|\delta_{ji}}\|^2_{-1/2,h,\delta_{ji}}\} \\ &= \sum_{\delta_{ji}\subset(\Gamma)} \left\{ h_j^3 \|\psi_1\|^2_{L^2(\delta_{ji})} + h_j \|\psi_2\|^2_{L^2(\delta_{ji})} \right\}, \end{aligned} \tag{4.16}$$

while on $X_h(\Omega)$ by

$$\|u\|^2_{X,h} = \|u\|^2_{H^2_H(\Omega)} + \sum_{\delta_{ji}\subset\Gamma} \{\|[u_{|\delta_{ji}}]\|^2_{3/2,h,\delta_{ji}} + \|[\partial_n u_{|\delta_{ji}}]\|^2_{1/2,h,\delta_{ji}}\}. \tag{4.17}$$

The next theorem gives us a discrete inf-sup condition in the mesh-dependent norm $\|\cdot\|_{M,h}$.

THEOREM 4.4. *For problem* (2.5) *the following inf-sup condition holds:*

$$\inf_{\psi\in M_h(\Gamma)\setminus\{0\}} \sup_{u\in X_h(\Omega)\setminus\{0\}} \frac{b(\psi, u)}{\|\psi\|_{M,h}\|u\|_{X,h}} \geq C, \tag{4.18}$$

where C is a constant independent of any h_k and the number of subdomains.

Proof. For a slave δ_{ji} and a restriction of $\mu = (\mu_1, \mu_2) \in M_h(\Gamma)$ to δ_{ji} we introduce two functions $\tilde{\mu}_1 \in W_0^{1,h_j}(\delta_{ji})$ and $\tilde{\mu}_2 \in W_0^{2,h_j}(\delta_{ji})$.

With a function $\mu_1 = \sum_{l=0,1} \sum_{p\in\delta_{ji,h}} \mu_1^{(l)}(p)\theta^l_{1,p} \in M^{1,h_j}(\Gamma_{ij})$ we associate $\tilde{\mu}_1 = \sum_{l=0,1} \sum_{p\in\delta_{ji,h}} \mu_1^{(l)}(p)\phi^l_{1,p}$ and for $\mu_2 = \sum_{p\in\delta_{ji,h/2}} \mu_2(p)\theta_{2,p} \in M^{2,h_j}(\Gamma_{ij})$ we define $\tilde{\mu}_2 = \sum_{p\in\delta_{ji,h/2}} \mu_2(p)\phi_{2,p}$. By Propositions 3.1 and 3.4 we have

$$\|\mu_k\|^2_{L^2(\delta_{ji})} \asymp \|\tilde{\mu}_k\|^2_{L^2(\delta_{ji})} \asymp \int_{\delta_{ji}} \mu_k\tilde{\mu}_k \, ds, \quad k = 1, 2. \tag{4.19}$$

We next see that

$$\|\mu_1\|_{-3/2,h,\delta_{ji}} \preceq \frac{h_j^{3/2} \int_{\delta_{ji}} \mu_1\tilde{\mu}_1 \, ds}{\|\tilde{\mu}_1\|_{0,h,\delta_{ji}}} = \frac{\int_{\delta_{ji}} \mu_1\tilde{\mu}_1 \, ds}{\|\tilde{\mu}_1\|_{3/2,h,\delta_{ji}}} = \int_{\delta_{ji}} \mu_1\hat{\mu}_1 \, ds, \tag{4.20}$$

where $\hat{\mu}_1 = \tilde{\mu}_1/\|\tilde{\mu}_1\|_{3/2,h,\delta_{ji}}$. Analogously, we get

$$\|\mu_2\|_{-1/2,h,\delta_{ji}} \preceq \int_{\delta_{ji}} \mu_2\hat{\mu}_2 \, ds, \tag{4.21}$$

where $\hat{\mu}_2 = \tilde{\mu}_2/\|\tilde{\mu}_1\|_{1/2,h,\delta_{ji}}$. Then we introduce two functions $w_{1,ij}, w_{2,ij} \in X_h(\Omega_j)$ as follows:

$$\begin{cases} a_j(w_{1,ij}, v) = 0 & \forall v \in X_{0,h}(\Omega_j), \\ Tr\, w_{1,ij} = (\hat{\mu}_1, \partial_s\hat{\mu}_1, 0) & \text{on} \quad \delta_{ji}, \\ Tr\, w_{1,ij} = 0 & \text{on} \quad \partial\Omega_j \setminus \delta_{ji}, \end{cases} \tag{4.22}$$

$$\begin{cases} a_j(w_{2,ij}, v) = 0 & \forall v \in X_{0,h}(\Omega_j), \\ Tr\, w_{2,ij} = (0, 0, \hat{\mu}_2) & \text{on} \quad \delta_{ji}, \\ Tr\, w_{2,ij} = 0 & \text{on} \quad \partial\Omega_j \setminus \delta_{ji}, \end{cases} \tag{4.23}$$

where $Tr\, v = (v_{|\partial\Omega_j}, \nabla v_{|\partial\Omega_j})$ for $v, \nabla v$—the H^2 traces of $v \in H^2(\Omega_j)$, and $X_{0,h}(\Omega_j) = X_h(\Omega_j) \cap H_0^2(\Omega_j)$. The extension of $w_{k,ij}, k = 1, 2$, by zero to Ω is denoted by the same symbols. By the Friedrich's inequality, the extension theorem of [23], and inverse inequalities we have

$$\|w_{1,ij}\|^2_{H^2(\Omega_j)} \preceq \|\partial_s\hat{\mu}_1\|^2_{H^{1/2}_{00}(\delta_{ji})} \preceq \|\hat{\mu}_1\|^2_{3/2,h,\delta_{ji}} = 1,$$
$$\|w_{2,ij}\|^2_{H^2(\Omega_j)} \preceq \|\hat{\mu}_2\|^2_{H^{1/2}_{00}(\delta_{ji})} \preceq \|\hat{\mu}_2\|^2_{1/2,h,\delta_{ji}} = 1.$$

We then define a function $w \in X_h(\Omega)$ as follows:

$$w = \sum_{k=1,2} \sum_{\delta_{ji} \subset \Gamma} b(\mu_k, w_{k,ij}) w_{k,ij} = \sum_{k=1,2} \sum_{\delta_{ji} \subset \Gamma} \int_{\delta_{ji}} \mu_k\hat{\mu}_k \, ds \, w_{k,ij}.$$

From the definition and a coloring argument we find that

$$\begin{aligned} \|w\|^2_{H^2_H(\Omega)} &\preceq \sum_{j=1}^{N} \sum_{\delta_{ji} \subset \partial\Omega_j} \sum_{k=1,2} b(\mu_k, w_{k,ij})^2 \|w_{k,ij}\|^2_{H^2(\Omega_j)} \\ &\preceq \sum_{\delta_{ji} \subset \Gamma} \{\|\mu_1\|^2_{-3/2,h,\delta_{ji}} \|w_{1,ij}\|^2_{3/2,h,\delta_{ji}} + \|\mu_1\|^2_{-1/2,h,\delta_{ji}} \|w_{2,ij}\|^2_{1/2,h,\delta_{ji}}\} \\ &= \sum_{\delta_{ji} \subset \Gamma} \{\|\mu_1\|^2_{-3/2,h,\delta_{ji}} + \|\mu_2\|^2_{-1/2,h,\delta_{ji}}\} = \|\mu\|^2_{M,h} \end{aligned} \tag{4.24}$$

and, analogously,

$$\|[w]\|_{3/2,h,\delta_{ji}} = |b(\mu_1, w_{1,ij})| \, \|w_{1,ij}\|_{3/2,h,\delta_{ji}} \preceq \|\mu_1\|_{-3/2,h,\delta_{ji}}$$

and

$$\|[\partial_n w]\|_{1/2,h,\delta_{ji}} = |b(\mu_2, w_{2,ij})| \, \|w_{2,ij}\|_{1/2,h,\delta_{ji}} \preceq \|\mu_2\|_{-1/2,h,\delta_{ji}}.$$

Summing over all $\delta_{ji} \subset \Gamma$ and utilizing (4.24) we get

$$\|w\|_{X,h} \preceq \|\mu\|_{M,h}.$$

Finally, this, (4.20), and (4.21) yield

$$\|\mu\|_{M,h}^2 \preceq \sum_{\delta_{ji}\subset\Gamma}\sum_{k=1,2} b(\mu_k, w_{k,ij}) = b(\mu, w) \preceq b(\mu, w)\frac{\|\mu\|_{M,h}}{\|w\|_{X,h}}.$$

Dividing by $\|\mu\|_{M,h}$ completes the proof. ∎

We then have a theorem which gives us estimates of the error for Lagrange multipliers in the norms $\|\cdot\|_M$ and $\|\cdot\|_{M,h}$.

THEOREM 4.5. *If we assume that u^*, a solution of* (2.1)*, belongs to $H_0^2 \cap H^4(\Omega)$, then*

$$\|\lambda^* - \lambda_h^*\|_M^2 + \|\lambda^* - \lambda_h^*\|_{M,h}^2 \preceq \sum_{k=1}^N h_k^2 |u^*|_{H^4(\Omega_k)}^2.$$

Proof. We first have for any $\mu \in M_h(\Gamma)$ (cf., e.g., [13])

$$\|\lambda^* - \lambda_h^*\|_M \le \|\lambda^* - \mu\|_M + \|\mu - \lambda_h^*\|_M \quad \forall \mu \in M_h(\Gamma).$$

Next, by Theorem 4.3 we get

$$\|\mu - \lambda_h^*\|_M \preceq \sup_{v\in X_h(\Omega)} \frac{b(\lambda_h^* - \mu, v)}{\|v\|_X},$$

and by (4.5) and (2.5)

$$b(\lambda_h^* - \mu, v) = b(\lambda_h^* - \lambda^*, v) + b(\lambda^* - \mu, v) = a_H(u^* - u_h^*, v) + b(\lambda^* - \mu, v).$$

Thus

$$\frac{b(\lambda_h^* - \mu, v)}{\|v\|_X} \preceq |u^* - u_h^*|_{H_H^2(\Omega)} + \|\lambda^* - \mu\|_M.$$

Finally, we get

$$\|\lambda^* - \lambda_h^*\|_M \preceq \inf_{\mu\in M_h(\Gamma)} \|\lambda^* - \mu\|_M + |u^* - u_h^*|_{H_H^2(\Omega)}.$$

The second term is bounded by Lemma 3.11. This estimate is also valid for the mesh-dependent norm. By continuous embeddings $H_{00}^{3/2}(\delta_{ji}) \subset H^{3/2}(\delta_{ji})$ and $H_{00}^{1/2}(\delta_{ji}) \subset H^{1/2}(\delta_{ji})$ we have

$$\|\cdot\|_{H_{00}^{-3/2}(\delta_{ji})} \preceq \|\cdot\|_{H^{-3/2}(\delta_{ji})} \text{ and } \|\cdot\|_{H_{00}^{-1/2}(\delta_{ji})} \preceq \|\cdot\|_{H^{-1/2}(\delta_{ji})}.$$

Thus Lemmas 3.12 and 3.13 and the trace theorems yield

$$\begin{aligned}\|\lambda_1^* - Q_{1,ij}\lambda_1^*\|_{H_{00}^{-3/2}(\delta_{ji})} &\le \|\lambda_1^* - Q_{1,ij}\lambda_1^*\|_{H^{-3/2}(\delta_{ji})} \preceq h_j^2 |\lambda_1^*|_{H^{1/2}(\delta_{ji})} \\ &\preceq h_j^2 |u^*|_{H^4(\Omega_j)}\end{aligned}$$

and

$$\begin{aligned}\|\lambda_2^* - Q_{2,ij}\lambda_2^*\|_{H_{00}^{-1/2}(\delta_{ji})} &\le \|\lambda_2^* - Q_{2,ij}\lambda_2^*\|_{H^{-1/2}(\delta_{ji})} \preceq h_j^2 |\lambda_2^*|_{H^{3/2}(\delta_{ji})} \\ &\preceq h_j^2 |u^*|_{H^4(\Omega_j)}.\end{aligned}$$

Summing over all $\delta_{ji} \subset \Gamma$ completes the proof.

The last part of the proof for the mesh-dependent norms $\|\cdot\|_{M,h}$ and $\|\cdot\|_{X,h}$ can be done in a very similar way, utilizing the L^2 estimates of Lemmas 3.12 and 3.13. □

Remark 3. Throughout this section the Poincaré inequalities and the trace and extension theorems are utilized frequently for the subdomains and then the respective constants depend on the diameter of the subdomain. This is taken into account implicitly in all proofs; however, it is not written explicitly. We also would like to note that the other equivalent approach is to use local norms scaled appropriately by the diameters of respective subdomains.

Acknowledgment. The author would like to thank Prof. Maksymilian Dryja for his encouragement and many helpful discussions.

REFERENCES

[1] Y. Achdou, Y. A. Kuznetsov, and O. Pironneau, *Substructuring preconditioners for the Q_1 mortar element method*, Numer. Math., 71 (1995), pp. 419–449.

[2] Y. Achdou, Y. Maday, and O. B. Widlund, *Iterative substructuring preconditioners for mortar element methods in two dimensions*, SIAM J. Numer. Anal., 36 (1999), pp. 551–580.

[3] Z. Belhachmi, *Nonconforming mortar element methods for the spectral discretization of two-dimensional fourth-order problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1545–1573.

[4] F. Ben Belgacem, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[5] F. Ben Belgacem and Y. Maday, *The mortar element method for three-dimensional finite elements*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 289–302.

[6] A. Berger, R. Scott, and G. Strang, *Approximate boundary conditions in the finite element method*, in Symposia Mathematica, Vol. X (Convegno di Analisi Numerica, INDAM, Rome, 1972), Academic Press, London, 1972, pp. 295–313.

[7] C. Bernardi and Y. Maday, *Mesh adaptivity in finite elements by the mortar mesh*, Rev. Eur. Élém. Finis, 9 (2000), pp. 451–465.

[8] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications. Collège de France Seminar, Vol. XI (Paris, 1989–1991), Pitman Res. Notes Math. Ser. 299, Longman Scientific and Technical, Harlow, UK, 1994, pp. 13–51.

[9] D. Braess, W. Dahmen, and C. Wieners, *A multigrid algorithm for the mortar finite element method*, SIAM J. Numer. Anal., 37 (1999), pp. 48–69.

[10] S. C. Brenner, *The condition number of the Schur complement in domain decomposition*, Numer. Math., 83 (1999), pp. 187–203.

[11] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, New York, 1994.

[12] S. C. Brenner and L.-Y. Sung, *Balancing domain decomposition for nonconforming plate elements*, Numer. Math., 83 (1999), pp. 25–52.

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

[14] M. A. Casarin and O. B. Widlund, *A hierarchical preconditioner for the mortar finite element method*, Electron. Trans. Numer. Anal., 4 (1996), pp. 75–88.

[15] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.

[16] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, Handb. Numer. Anal. II, North–Holland, Amsterdam, 1991, pp. 17–351.

[17] M. Dryja, *An iterative substructuring method for elliptic mortar finite element problems with a new coarse space*, East-West J. Numer. Math., 5 (1997), pp. 79–98.

[18] J. Gopalakrishnan and J. E. Pasciak, *Multigrid for the mortar finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 1029–1052.

[19] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman (Advanced Publishing Program), Boston, 1985.

[20] C. Kim, R. D. Lazarov, J. E. Pasciak, and P. S. Vassilevski, *Multiplier spaces for the mortar finite element method in three dimensions*, SIAM J. Numer. Anal., 39 (2001), pp. 519–538.

[21] C. Lacour, *Non-conforming domain decomposition method for plate and shell problems*, in Domain Decomposition Methods, Vol. 10 (Boulder, CO, 1997), Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 304–310.

[22] C. Lacour and Y. Maday, *La méthode des éléments avec joint appliquée aux méthodes d'approximations discrete Kirchhoff triangles*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1237–1242.

[23] P. Le Tallec, J. Mandel, and M. Vidrascu, *A Neumann–Neumann domain decomposition algorithm for solving plate and shell problems*, SIAM J. Numer. Anal., 35 (1998), pp. 836–867.

[24] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications,* Vol. I, Die Grundlehren der mathematischen Wissenschaften, Band 181, Springer-Verlag, New York, Heidelberg, 1972.

[25] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications,* Vol. II, Die Grundlehren der mathematischen Wissenschaften, Band 182, Springer-Verlag, New York, Heidelberg, 1972.

[26] L. Marcinkowski, *Domain decomposition methods for mortar finite element discretizations of plate problems*, SIAM J. Numer. Anal., 39 (2001), pp. 1097–1114.

[27] L. Marcinkowski, *A mortar finite element method for plate problems*, in Proceedings of the 12th International Conference on Domain Decomposition Methods, Chiba, Japan, 1999, T. Chan, T. Kako, H. Kawarada, and O. Pironneau, eds., DDM.org., Augsburg, Germany, 2001, pp. 183–190; also available online at http://www.ddm.org/DD12/index.html.

[28] L. Marcinkowski, *A mortar element method for some discretizations of a plate problem*, Numer. Math., 93 (2002), pp. 361–386.

[29] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

[30] B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Lect. Notes Comput. Sci. Eng. 17, Springer-Verlag, Berlin, 2001.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2020–2042

DISCRETIZED STABILITY AND ERROR GROWTH OF THE NONAUTONOMOUS PANTOGRAPH EQUATION*

CHENGMING HUANG† AND STEFAN VANDEWALLE‡

Abstract. This paper is concerned with the stability properties of Runge–Kutta methods for the pantograph equation, a functional differential equation with a proportional delay. The focus is on nonautonomous equations. Both linear and nonlinear cases are considered. Sufficient and necessary conditions for the asymptotic stability of the numerical solution of general neutral pantograph equations are given. An upper bound for the error growth is investigated for algebraically stable methods applied to nonneutral equations. Finally, some stability results are extended to the case of a more general class of equations.

Key words. pantograph equation, asymptotic stability, error growth, Runge–Kutta methods

AMS subject classifications. 65L06, 65L07, 65L20

DOI. 10.1137/S0036142902419296

1. Introduction.

Many real-world phenomena can be modelled by initial value problems for functional differential equations of the form

$$y'(t) = f(t, y(t), y(t-\tau(t)), y'(t-\tau(t))). \tag{1.1}$$

In recent years, the study of numerical solvers for this problem has attracted the attention of many authors. The classical case where the term $\tau(t)$ is a constant can be regarded as a representative of finite time delay and has been widely studied in the literature (see, for example, Baker [1], Bellen and Zennaro [5] and the extensive bibliography therein). Another interesting case, which can be viewed as a representative of infinite time delay, is that of the pantograph equation, where

$$\tau(t) = (1-q)t, \quad q \in (0,1).$$

For applications of this type of equation, we refer to Iserles [16].

In order to get insight into the stability of numerical methods for the pantograph equation, the scalar linear autonomous equation

$$y'(t) = ay(t) + by(qt) + cy'(qt), \quad t > 0, \tag{1.2}$$

has been used as a test problem and many interesting results have been found (cf. [2, 6, 7, 8, 9, 17, 18, 22, 24, 25]). In the early work, a constant stepsize was considered. As pointed out in Liu [22, 24], however, this kind of stepsize precludes long time integration due to computer memory restrictions. In order to overcome this difficulty, Liu [22] transformed (1.2) into a differential equation with a constant delay by a change of variable, suggested by Jackiewicz [20]. Later, Liu [24] and Bellen,

*Received by the editors December 10, 2002; accepted for publication (in revised form) December 19, 2003; published electronically February 25, 2005. This research was funded by the Research Council of the K. U. Leuven through fellowship grant F/02/060 and by the NSF of China through project 10101027.

http://www.siam.org/journals/sinum/42-5/41929.html

†Department of Mathematics, Huazhong University of Science and Technology, Wuhan 430074, China (chengming_huang@hotmail.com).

‡Katholieke Universiteit Leuven, Department of Computerscience, Celestijnenlaan 200A, B3001 Leuven, Belgium (Stefan.Vandewalle@cs.kuleuven.ac.be).

Guglielmi, and Torelli [2] proposed nonconstant stepsize strategies where the stepsizes are geometrically increasing and they investigated the stability of θ-methods. Recently, Koto [21] further studied the stability of general Runge–Kutta methods for the multidimensional system

$$u'(t) = e^t Lu(t) + e^t Mu(t + \log q) + Nu'(t + \log q), \quad t > 0, \tag{1.3}$$

which is obtained from the equation

$$y'(t) = Ly(t) + My(qt) + qNy'(qt), \quad t > 0, \tag{1.4}$$

by a change of the independent variable $u(t) = y(e^t)$, where L, M, and N are constant complex $d \times d$ matrices. In an abstract sense the geometrically increasing mesh approach and the exponential transform method may be considered to be essentially the same (cf. [21]). Relevant to the nonautonomous pantograph equation, however, only few results on numerical stability have been published. Bellen, Guglielmi, and Torelli [2], and Guglielmi and Zennaro [13] discussed the asymptotic stability of θ-methods for scalar equations with variable coefficients. Bellen, Maset, and Torelli [4] studied the "first step" integration of linear systems of neutral type and investigated the contractivity of continuous Runge–Kutta methods. Zhang and Sun [28, 29] recently obtained some stability results of Runge–Kutta methods for a class of nonlinear equations of nonneutral type (see Theorem 6.1 of this paper).

In this paper, a new approach for proving numerical stability is introduced. Sufficient and necessary conditions for asymptotic stability are derived for both linear and nonlinear problems of neutral type. Also, upper bounds for the error growth are studied for nonneutral systems and some sharper results than those published in the literature are obtained.

This paper is organized as follows. In section 2, the discrete schemes based on Runge–Kutta methods are introduced. In section 3, we focus on the asymptotic stability of the schemes for linear systems of neutral type with variable coefficients. In section 4, an upper bound for the error growth is given for a class of linear problems. In section 5, we turn our attention to nonlinear equations and an asymptotic stability result is derived. In section 6, we further investigate the error growth bound for a class of nonlinear problems. In section 7, we generalize our stability analysis to the case of a more general class of equations. Finally, in section 8, some conclusions are drawn.

2. Adaptation of Runge–Kutta methods to functional-differential equations of pantograph type. In this section, we consider the adaptation of Runge–Kutta methods to the pantograph equation

$$\begin{cases} y'(t) = f(t, y(t), y(qt), y'(qt)), \quad t > 0, \\ y(0) = y_0, \end{cases} \tag{2.1}$$

where $f : [0, +\infty) \times \mathbb{C}^d \times \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}^d$, is a given mapping and q is a constant satisfying $q \in (0, 1)$.

Let (A, b, c) denote a given Runge–Kutta method characterized by the $s \times s$ matrix $A = (a_{ij})$ and vectors $b = (b_1, \dots, b_s)^T$, $c = (c_1, \dots, c_s)^T$. In this paper we always assume that $\sum_{j=1}^{s} b_j = 1$. Let t_n, $n = 0, 1, \dots$, be grid points satisfying

$$0 = t_0 < t_1 < t_2 < \cdots < \infty, \quad \lim_{n\to\infty} t_n = \infty,$$

and $h_n = t_{n+1} - t_n$, the corresponding stepsizes. Approximations y_{n+1} to $y(t_{n+1})$ are defined by the following equations:

$$Y_i^{(n)} = y_n + h_n \sum_{j=1}^{s} a_{ij} f(t_n + c_j h_n, Y_j^{(n)}, \bar{Y}_j^{(n-m)}, \hat{Y}_j^{(n-m)}), \quad i = 1, \dots, s, \tag{2.2}$$

$$y_{n+1} = y_n + h_n \sum_{j=1}^{s} b_j f(t_n + c_j h_n, Y_j^{(n)}, \bar{Y}_j^{(n-m)}, \hat{Y}_j^{(n-m)}), \tag{2.3}$$

where each $Y_j^{(n)}$ is an approximation to $y(t_n + c_j h_n)$, the arguments $\bar{Y}_j^{(n-m)}$ and $\hat{Y}_j^{(n-m)}$ denote approximations to $y(q(t_n + c_j h_n))$ and $y'(q(t_n + c_j h_n))$, respectively, obtained by specific interpolation procedures at the point $t = q(t_n + c_j h_n)$, and m is a positive integer that will be defined later.

In this paper, we consider a nonconstant stepsize strategy where the stepsizes are geometrically increasing. This kind of grid was proposed by Liu [24], and by Bellen, Guglielmi, and Torelli [2]. As pointed out in the above references, it has two advantages. First, it can avoid the computer memory problems in the case of long-time integration. Second, an interpolation procedure is not necessary if we choose the grid such that every delayed point maps exactly onto a past grid point.

To formulate the grid, we partition the half-line $[0, +\infty)$ into a union of bounded intervals as follows:

$$[0, +\infty) = [0, h] \bigcup_{k=0}^{\infty} (q^{-k}h, q^{-k-1}h],$$

where h is an arbitrary but fixed positive number. Second, we further divide every interval $(q^{-k}h, q^{-k-1}h]$ into a fixed number m of subintervals whose length is proportionally increasing with the factor $p = q^{\frac{-1}{m}}$, i.e.,

$$(q^{-k}h, q^{-k-1}h] = \bigcup_{i=1}^{m} (q^{-k}p^{i-1}h, q^{-k}p^{i}h].$$

The first interval $[0, h]$ is divided as follows:

$$[0, h] = [0, qh] \bigcup_{i=1}^{m} (qp^{i-1}h, qp^{i}h].$$

Therefore, we obtain the global grid defined by

$$t_n = p^{n-m-1}h, \quad n = 1, 2, \dots,$$

which gives

$$h_n = t_{n+1} - t_n = p^{n-m-1}(p-1)h, \quad n = 1, 2, \dots,$$

and $h_0 = qh$. Hence, for $n = m+1, m+2, \dots,$

$$qt_n = t_{n-m} \quad \text{and} \quad qh_n = h_{n-m},$$

which shows the delayed point is just on the past grid point and

$$q(t_n + c_j h_n) = t_{n-m} + c_j h_{n-m}.$$

Therefore, we can set

$$\bar{Y}_j^{(n-m)} = Y_j^{(n-m)},$$
$$\hat{Y}_j^{(n-m)} = f(t_{n-m} + c_j h_{n-m}, Y_j^{(n-m)}, Y_j^{(n-2m)}, \hat{Y}_j^{(n-2m)}),$$

which on substitution into (2.2)–(2.3) gives

$$Y_i^{(n)} = y_n + h_n \sum_{j=1}^{s} a_{ij} \hat{Y}_j^{(n)}, \quad i = 1, \dots, s, \tag{2.4}$$

$$\hat{Y}_j^{(n)} = f(t_n + c_j h_n, Y_j^{(n)}, Y_j^{(n-m)}, \hat{Y}_j^{(n-m)}), \quad j = 1, \dots, s, \tag{2.5}$$

$$y_{n+1} = y_n + h_n \sum_{j=1}^{s} b_j \hat{Y}_j^{(n)}. \tag{2.6}$$

Remark 2.1. In Liu [24], this kind of stepsize was used in the numerical examples although a slightly more general assumption on the grid was considered in the theoretical analysis. In Bellen, Guglielmi, and Torelli [2], the stepsize strategy in the theoretical analysis is that every interval $(q^{-k}h, q^{-k-1}h]$ is divided into m intervals of the same size. The strategy of proportionally increasing stepsizes was also suggested in Remark 5.1 of their paper where it is pointed out that this choice can simplify the implementation of the method considerably and leads to a more regular behavior of the error.

Remark 2.2. Because we are interested in the stability of the numerical solution, we assume that the initial values $Y_j^{(n-1)}$, $\hat{Y}_j^{(n-1)}$, and y_n are available for $n = 1, 2, \dots, m$. For the integration of the first steps, i.e., the initializing methods, we refer to the paper by Bellen, Maset, and Torelli [4].

Now we introduce some concepts which will be used later.

DEFINITION 2.3 (see [14]). *The stability function of a Runge–Kutta method* (A, b, c) *is defined by*

$$R(z) = 1 + zb^T(I_s - zA)^{-1}e,$$

where $e = [1, \dots, 1]^T$ *and* I_s *stands for the* $s \times s$ *identity matrix.*

DEFINITION 2.4. *When* A *is nonsingular, the method is called strictly stable at infinity if*

$$|R(\infty)| = |1 - b^T A^{-1} e| < 1.$$

DEFINITION 2.5 (see [10]). *A Runge–Kutta method* (A, b, c) *is called algebraically stable if the following matrix* $\mathcal{M} = [\mathcal{M}_{ij}]$ *is nonnegative definite:*

$$\mathcal{M} = BA + A^T B - bb^T,$$

where $B = \operatorname{diag}(b_1, b_2, \dots, b_s)$.

3. Linear stability. In this section, we discuss the stability of Runge–Kutta methods for linear systems of the form

$$y'(t) = L(t)y(t) + M(t)y(qt) + N(t)y'(qt), \quad t > 0, \tag{3.1}$$

where $L(t)$, $M(t)$, and $N(t)$ are complex $d \times d$ matrices whose entries are continuous functions. First, we recall some results on the stability of the analytical solution. For

the autonomous case (1.4) of (3.1), a result obtained by Liu [23] (see also [16]) implies the following proposition.

PROPOSITION 3.1. *The zero solution of* (1.4) *is asymptotically stable if* L, M *satisfy*

$$\sigma[L] \subset \mathbb{C}^-, \qquad \rho[L^{-1}M] < 1, \tag{3.2}$$

where $\mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$, *and* $\sigma[\cdot]$ *and* $\rho[\cdot]$ *denote the spectrum and spectral radius of a matrix, respectively.*

For the asymptotic stability of the nonautonomous neutral system (3.1), a result obtained by Iserles and Terjeki [19] implies the following proposition.

PROPOSITION 3.2. *The zero solution of* (3.1) *is asymptotically stable if there exists a vector norm* $\|\cdot\|_*$ *on* $\mathbb{C}^d$, *with induced matrix norm and the corresponding logarithmic norm* $\mu[\cdot]$ *such that for all* $t \geq 0$ *the following statements hold:*

$$\mu[L(t)] \leq 0, \qquad \|N(t)\|_* \leq \xi_0 < 1, \qquad \int_0^\infty \mu[L(t)]dt = -\infty, \tag{3.3}$$

$$\max_{x\in[0,t]} \|M(x) + N(x)L(qx)\|_* + k_0\mu[L(t)](1-\xi_0) \leq 0 \text{ for some } k_0 \in (0,1). \tag{3.4}$$

Now we analyze the conditions of Proposition 3.2 in order to motivate our assumptions for the numerical stability analysis. We consider the nonneutral case, i.e., $N(t) = 0$. If $\limsup_{t\to\infty} \mu[L(t)] = 0$, from (3.4) it follows that $M(t) = 0$, $t \in [0,\infty)$, which leads to a trivial case. Therefore, we assume that for sufficiently large t, $\mu[L(t)] \leq L_0 < 0$, which implies that the matrix $L(t)$ is nonsingular. Hence, from the properties of the logarithmic norm (cf. [12]), it follows that

$$\|L^{-1}(t)\|_* \leq \frac{1}{-\mu(L(t))} \leq \frac{1}{-L_0}.$$

In addition, for every $u \in \mathbb{C}^d$,

$$-\mu[L(t)]\|u\|_* \leq \|L(t)u\|_*,$$

which guarantees that, for every $v \in \mathbb{C}^d$,

$$-\mu[L(t)]\|L^{-1}(t)M(t)v\|_* \leq \|M(t)v\|_*.$$

Therefore, we have

$$-\mu[L(t)]\|L^{-1}(t)M(t)\|_* \leq \|M(t)\|_*,$$

which, combined with (3.4), implies

$$\|L^{-1}(t)M(t)\|_* \leq k_0 < 1.$$

In our analysis of the asymptotic stability of numerical methods, we will make use of the following assumption, which is a natural extension of the conditions for scalar equations given by Guglielmi and Zennaro [13], and which can also cover condition (3.2) in the autonomous case.

Assumption $\mathcal{A}$*:* There exist a vector norm $\|\cdot\|_*$ on $\mathbb{C}^d$ and induced matrix norm such that the matrices $L(t)$, $M(t)$, and $N(t)$ satisfy for all $t > 0$

$$\|L^{-1}(t)\|_* \leq C_0, \qquad \|L^{-1}(t)N(t)\|_* \leq \hat{C}_0, \qquad \|L^{-1}(t)M(t)\|_* \leq k_0 < 1,$$

where C_0, $\hat{C}_0$ and k_0 are constants.

The application of method (2.4–2.6) to (3.1) leads to the following difference equation:

$$Y_i^{(n)} = y_n + h_n \sum_{j=1}^{s} a_{ij} \hat{Y}_j^{(n)}, \quad i = 1, \dots, s, \tag{3.5}$$

$$\begin{aligned} \hat{Y}_j^{(n)} = {} & L(t_n + c_j h_n) Y_j^{(n)} + M(t_n + c_j h_n) Y_j^{(n-m)} \\ & + N(t_n + c_j h_n) \hat{Y}_j^{(n-m)}, \quad j = 1, \dots, s, \end{aligned} \tag{3.6}$$

$$y_{n+1} = y_n + h_n \sum_{j=1}^{s} b_j \hat{Y}_j^{(n)}. \tag{3.7}$$

Now we present a preliminary result that will be used further on. For the difference equation

$$u_n = \lambda_1 u_{n-m} + \lambda_2 v_n + \lambda_3 v_{n-m}, \tag{3.8}$$

$$v_{n+1} = \lambda_4 v_n + \lambda_5 u_n, \tag{3.9}$$

with $\lambda_i \in \mathbb{C}$, we have the asymptotic stability result stated in the following lemma.

LEMMA 3.3. *The difference equation* (3.8)–(3.9) *is asymptotically stable if*

$$|\lambda_1| < 1, \quad |\lambda_4| < 1, \quad (|\lambda_2| + |\lambda_3|)|\lambda_5| < (1 - |\lambda_1|)(1 - |\lambda_4|). \tag{3.10}$$

Proof. It is easy to see that the characteristic equation of (3.8)–(3.9) is given by

$$\det \begin{bmatrix} 1 - \lambda_1 z^{-m} & -\lambda_2 - \lambda_3 z^{-m} \\ -\lambda_5 & z - \lambda_4 \end{bmatrix} = 0,$$

which gives

$$\lambda_5(\lambda_2 + \lambda_3 z^{-m}) = z(1 - \lambda_4 z^{-1})(1 - \lambda_1 z^{-m}). \tag{3.11}$$

Suppose $|\lambda_1| < 1$, $|\lambda_4| < 1$, and there exists $z \in \mathbb{C}$ satisfying (3.11) with $|z| \geq 1$. Then

$$|\lambda_5|(|\lambda_2| + |\lambda_3|) \geq (1 - |\lambda_4|)(1 - |\lambda_1|),$$

which contradicts the third inequality of (3.10). This completes the proof. ∎

Now we state and prove the main result of this section.

THEOREM 3.4. *Let Assumption* $\mathcal{A}$ *hold and the matrix* A *be nonsingular. Then the difference equation* (3.5–3.7) *is asymptotically stable if the underlying Runge–Kutta method is strictly stable at infinity.*

Proof. It follows from (3.6) that

$$\begin{aligned} L^{-1}(t_n + c_i h_n) \hat{Y}_i^{(n)} = {} & Y_i^{(n)} + L^{-1}(t_n + c_i h_n) M(t_n + c_i h_n) Y_i^{(n-m)} \\ & + L^{-1}(t_n + c_i h_n) N(t_n + c_i h_n) \hat{Y}_i^{(n-m)}, \quad i = 1, \dots, s, \end{aligned}$$

which in combination with Assumption $\mathcal{A}$ gives

$$\|Y_i^{(n)}\|_* \leq k_0 \|Y_i^{(n-m)}\|_* + C_0 \|\hat{Y}_i^{(n)}\|_* + \hat{C}_0 \|\hat{Y}_i^{(n-m)}\|_*. \tag{3.12}$$

On the other hand, (3.5) plus the nonsingularity of A implies

$$\hat{Y}_i^{(n)} = h_n^{-1} \sum_{j=1}^{s} D_{ij}(Y_j^{(n)} - y_n), \tag{3.13}$$

where $D = [D_{ij}] = A^{-1}$. Substituting (3.13) into (3.7) yields

$$y_{n+1} = R(\infty) y_n + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i D_{ij} Y_j^{(n)}. \tag{3.14}$$

Hence, there exists a constant $C_1 > 0$ such that

$$\|y_{n+1}\|_* \le |R(\infty)| \|y_n\|_* + C_1 \sum_{i=1}^{s} \|Y_i^{(n)}\|_*. \tag{3.15}$$

A combination of (3.12) and (3.13) leads to

(3.16)

$$\begin{aligned}\sum_{i=1}^{s} \|Y_i^{(n)}\|_* \le\ & k_0 \sum_{i=1}^{s} \|Y_i^{(n-m)}\|_* \\ & + \sum_{i=1}^{s}\sum_{j=1}^{s} |D_{ij}| (h_n^{-1} C_0 \|Y_j^{(n)} - y_n\|_* + h_{n-m}^{-1} \hat{C}_0 \|Y_j^{(n-m)} - y_{n-m}\|_*),\end{aligned}$$

which shows that there exists a constant $\hat{C}_1 > 0$ such that

$$\begin{aligned}\sum_{i=1}^{s} \|Y_i^{(n)}\|_* \le\ & k_0 \sum_{i=1}^{s} \|Y_i^{(n-m)}\|_* \\ & + \hat{C}_1 \left[h_n^{-1} \left(\sum_{j=1}^{s} \|Y_j^{(n)}\|_* + \|y_n\|_* \right) + h_{n-m}^{-1} \left(\sum_{j=1}^{s} \|Y_j^{(n-m)}\|_* + \|y_{n-m}\|_* \right) \right].\end{aligned}$$

Considering $h_n \to \infty$, there exist positive numbers $\mathcal{N}_0, k_1 < 1$, and $C_2 < (1-k_1)(1 - |R(\infty)|)/(2C_1)$ such that for every $n > \mathcal{N}_0$,

$$\sum_{i=1}^{s} \|Y_i^{(n)}\|_* \le k_1 \sum_{i=1}^{s} \|Y_i^{(n-m)}\|_* + C_2(\|y_n\|_* + \|y_{n-m}\|_*). \tag{3.17}$$

An application of Lemma 3.3 to (3.15) and (3.17) gives

$$\lim_{n\to\infty} \|y_n\|_* = 0 \quad \text{and} \quad \lim_{n\to\infty} \sum_{i=1}^{s} \|Y_i^{(n)}\|_* = 0.$$

Considering (3.13), we have

$$\lim_{n\to\infty} \sum_{i=1}^{s} h_n \|\hat{Y}_i^{(n)}\|_* = 0.$$

Therefore, the difference equation (3.5–3.7) is asymptotically stable. This completes the proof. ∎

Remark 3.5. In the case of constant coefficients, it is well known that the condition $\rho[L^{-1}M] < 1$ is equivalent to the condition that there exists a norm $\|\cdot\|_*$ such that $\|L^{-1}M\|_* < 1$. Therefore, specializing Theorem 3.4 to the case of autonomous equations, the obtained result is in accordance with that by Koto [21]. Here we have given a new approach to the proof which allowed us to study the variable coefficient case.

Remark 3.6. In the proof we only use the fact that $h_n \to \infty$. Hence, our result is also valid for the other grid types proposed in [2, 24]. In addition, it is easily seen from the proof that, if the condition $\|L^{-1}(t)M(t)\|_* \le k_0 < 1$ in Assumption $\mathcal{A}$ is replaced by $\lim_{t\to\infty} \|L^{-1}(t)M(t)\|_* \le k_0 < 1$, Theorem 3.4 still holds. In the one-dimensional case, the latter has been assumed for the stability analysis of θ-methods in [2]. Finally, specializing Theorem 3.4 to the nonneutral case, the induced result is also new.

Next, we show that the assumption of strict stability at infinity is also necessary for the asymptotic stability of the difference equation.

LEMMA 3.7. *Suppose the matrix A is nonsingular. Then there exists a constant $\mathcal{N}_1 > 0$ such that*

$$|R(z)| \ge |R(\infty)| - 2|z^{-1}||b^T A^{-2} e|, \quad |z| \ge \mathcal{N}_1,\ z \in \mathbb{C}. \tag{3.18}$$

Proof. Considering the nonsingularity of A and the fact

$$(A - z^{-1} I_s)^{-1} = A^{-1} + z^{-1} A^{-1} (A - z^{-1} I_s)^{-1},$$

we have

$$R(z) = R(\infty) - z^{-1} b^T A^{-1} (A - z^{-1} I_s)^{-1} e,$$

which gives

$$|R(z)| \ge |R(\infty)| - |z^{-1}||b^T A^{-1} (A - z^{-1} I_s)^{-1} e|. \tag{3.19}$$

Considering

$$\lim_{z\to\infty} |b^T A^{-1} (A - z^{-1} I_s)^{-1} e| = |b^T A^{-2} e|,$$

there exists a constant $\mathcal{N}_1$ such that for every $z \in \mathbb{C}$ with $|z| \ge \mathcal{N}_1$,

$$|b^T A^{-1} (A - z^{-1} I_s)^{-1} e| \le 2|b^T A^{-2} e|,$$

which, together with (3.19), implies the conclusion. ∎

The application of method (2.4–2.6) to the scalar equation

$$y'(t) = \lambda y(t), \quad \lambda \in \mathbb{R}, \tag{3.20}$$

leads to the difference equation

$$y_{n+1} = R(h_n \lambda) y_n. \tag{3.21}$$

THEOREM 3.8. *Suppose the matrix A is nonsingular and there exists a constant λ_0 such that the difference equation* (3.21) *is asymptotically stable for every λ satisfying $\lambda\lambda_0 > \lambda_0^2$. Then the underlying Runge–Kutta method is strictly stable at infinity.*

Proof. Suppose $|R(\infty)| \geq 1$. By the assumptions of the theorem we can choose λ such that

$$\lambda\lambda_0 > \lambda_0^2, \quad |h_i\lambda| \geq \mathcal{N}_1, \quad \text{and} \quad |h_i\lambda| > 4|b^T A^{-2} e|, \quad i = 0, 1,$$

which gives

$$|h_n\lambda| \geq \mathcal{N}_1 \quad \text{and} \quad |h_n\lambda| > 4|b^T A^{-2} e|, \quad n = 0, 1, 2, \ldots .$$

Considering Lemma 3.7, we have

$$\begin{aligned} |R(h_n\lambda)| &\geq |R(\infty)| - 2|h_n\lambda|^{-1}|b^T A^{-2} e| \\ &\geq |R(\infty)| \exp(-4|h_n\lambda|^{-1}|b^T A^{-2} e|/|R(\infty)|) \\ &= |R(\infty)| \exp(-4p^{-n+1}|h_1\lambda|^{-1}|b^T A^{-2} e|/|R(\infty)|), \end{aligned}$$

where we have used the fact that the function $1 - x - \exp(-2x)$ is positive for $x \in (0, 1/2)$. Therefore,

$$\begin{aligned} \prod_{i=1}^{n} |R(h_i\lambda)| &\geq \prod_{i=1}^{n} |R(\infty)| \exp(-4p^{-i+1}|h_1\lambda|^{-1}|b^T A^{-2} e|/|R(\infty)|) \\ &= |R(\infty)|^n \exp\left(-\frac{1 - p^{-n}}{1 - p^{-1}} 4|h_1\lambda|^{-1}|b^T A^{-2} e|/|R(\infty)|\right), \end{aligned}$$

which shows that the difference equation (3.21) is not asymptotically stable. This completes the proof. ∎

COROLLARY 3.9. *Suppose the matrix A is nonsingular and there exists a constant $\lambda_0 < 0$ such that* (3.21) *is asymptotically stable for every $\lambda \in (-\infty, \lambda_0]$. Then the underlying Runge–Kutta method is strictly stable at infinity.*

4. An upper bound of error growth for linear problems. Asymptotic stability implies that the initial error will eventually vanish for sufficiently large time points. From the viewpoint of a practical computation, it is also important to give an upper bound of error growth. This subject was studied in Koto [21], where the nonneutral pantograph equation

$$\begin{cases} y'(t) = Ly(t) + My(qt), & t > 0, \\ y(0) = y_0, \end{cases} \tag{4.1}$$

was used as a test problem and algebraically stable methods were considered. In this section we follow Koto's practice and pursue a sharper result for (4.1), which can be regarded as an error equation of a linear problem.

The application of method (2.4–2.6) to (4.1) yields

$$Y^{(n)} = (e \otimes I_d) y_n + h_n (A \otimes I_d) \hat{Y}^{(n)}, \tag{4.2}$$

$$\hat{Y}^{(n)} = (I_s \otimes L) Y^{(n)} + (I_s \otimes M) Y^{(n-m)}, \tag{4.3}$$

$$y_{n+1} = y_n + h_n (b^T \otimes I_d) \hat{Y}^{(n)}, \tag{4.4}$$

where $\otimes$ denotes the Kronecker product and

$$Y^{(n)} = \left(Y_1^{(n)^T}, Y_2^{(n)^T}, \ldots, Y_s^{(n)^T}\right)^T, \qquad \hat{Y}^{(n)} = \left(\hat{Y}_1^{(n)^T}, \hat{Y}_2^{(n)^T}, \ldots, \hat{Y}_s^{(n)^T}\right)^T.$$

The following notation is a generalization of that in [21]:

$$\mathcal{H}(\sigma) = -\begin{bmatrix} L^*G + GL + E & GM \\ M^*G & -\sigma E \end{bmatrix}, \tag{4.5}$$

where G and E are Hermitian positive definite matrices and the superscript $*$ stands for the Hermitian adjoint. It is also seen that $\mathcal{H}(\sigma_2)$ is necessarily nonnegative definite if $\mathcal{H}(\sigma_1)$ is nonnegative definite and $\sigma_2 > \sigma_1$. In the scalar case, if the complex numbers L, M satisfy

$$\sqrt{\sigma}(\mathrm{Re}\, L) + |M| < 0,$$

then $\mathcal{H}(\sigma)$ is nonnegative definite.

Throughout this section, we assume that the notations are the same as those in section 3.

LEMMA 4.1. *Suppose that the method (A, b, c) is algebraically stable and there exist a constant σ and matrices G and E such that $\mathcal{H}(\sigma)$ is nonnegative definite. Then the following inequality holds true:*

$$y_{n+1}^* G y_{n+1} \le y_n^* G y_n - h_n Y^{(n)^*}(B \otimes E) Y^{(n)} + \sigma h_n Y^{(n-m)^*}(B \otimes E) Y^{(n-m)}. \tag{4.6}$$

Proof. As in Burrage and Butcher [10], where it is proved that algebraic stability implies B-stability, we can obtain

$$\begin{aligned} &y_{n+1}^* G y_{n+1} - y_n^* G y_n - h_n Y^{(n)^*}(B \otimes G)\hat{Y}^{(n)} - h_n \hat{Y}^{(n)^*}(B \otimes G) Y^{(n)} \\ &\quad = -h_n^2 \hat{Y}^{(n)^*}(\mathcal{M} \otimes G)\hat{Y}^{(n)}. \end{aligned}$$

By using the algebraic stability of the method, we have

$$\begin{aligned} y_{n+1}^* G y_{n+1} &\le y_n^* G y_n + h_n Y^{(n)^*}(B \otimes G)\hat{Y}^{(n)} + h_n \hat{Y}^{(n)^*}(B \otimes G) Y^{(n)} \\ &= y_n^* G y_n - h_n Y^{(n)^*}(B \otimes E) Y^{(n)} + \sigma h_n Y^{(n-m)^*}(B \otimes E) Y^{(n-m)} \\ &\quad - h_n \sum_{i=1}^{s} b_i (Y_i^{(n)^*}, Y_i^{(n-m)^*}) \mathcal{H}(\sigma) (Y_i^{(n)^T}, Y_i^{(n-m)^T})^T \end{aligned}$$

which by the nonnegative definiteness of $\mathcal{H}(\sigma)$ gives (4.6). ∎

THEOREM 4.2. *Suppose that the method (A, b, c) is algebraically stable and there exist matrices G and E such that $\mathcal{H}(q)$ is nonnegative definite. Then we have that for every $n \ge m$,*

$$y_{n+1}^* G y_{n+1} + \sum_{i=n-m+1}^{n} h_i Y^{(i)^*}(B \otimes E) Y^{(i)} \le y_n^* G y_n + \sum_{i=n-m}^{n-1} h_i Y^{(i)^*}(B \otimes E) Y^{(i)}. \tag{4.7}$$

Proof. Inequality (4.7) immediately follows from Lemma 4.1 and the fact $q h_n = h_{n-m}$. ∎

Remark 4.3. By the same argument as in section 5 of Koto [21], it is seen that the functional

$$V(y(t)) = y(t)^* G y(t) + \int_{qt}^{t} y(x)^* E y(x) dx$$

is a Liapunov functional for equation (4.1) if $\mathcal{H}(q)$ is nonnegative definite. Inequality (4.7) can be regarded as a discrete analogue.

Remark 4.4. The proof of Lemma 4.1 is closely related to its counterpart in [21]. There, (4.1) was studied through an investigation of the corresponding constant delay system transformed by a change of independent variable.

In the following, we derive a result which can be applied to the more general case where $\mathcal{H}(\sigma)$ is nonnegative definite for some $\sigma \in [q, 1)$. Except for certain special statements, the following results remain valid for $\sigma \geq 1$ although they may not result in stability when $\sigma \geq 1$.

LEMMA 4.5. *Suppose that the method (A, b, c) is algebraically stable and there exist a constant $\sigma \geq q$ and matrices G and E such that $\mathcal{H}(\sigma)$ is nonnegative definite. Then we have that for every $k \geq 0$,*

$$\sum_{i=1}^{m} p^{-i} Y^{((k+2)m-i)^*} (B \otimes E) Y^{((k+2)m-i)} \leq q^2 \sigma^k \delta, \tag{4.8}$$

where

$$\delta = p h_1^{-1} y_m^* G y_m + q^{-1} \sigma \sum_{i=0}^{m-1} p^i Y^{(i)^*} (B \otimes E) Y^{(i)}. \tag{4.9}$$

Proof. It follows from Lemma 4.1 that

$$\begin{aligned} y_{n+1}^* G y_{n+1} \leq y_n^* G y_n - h_1 p^{n-1} Y^{(n)^*} (B \otimes E) Y^{(n)} \\ + \sigma q^{-1} h_1 p^{n-m-1} Y^{(n-m)^*} (B \otimes E) Y^{(n-m)}. \end{aligned}$$

By induction, one arrives at

$$\begin{aligned} y_{n+1}^* G y_{n+1} \leq y_m^* G y_m - h_1 \sum_{i=n-m+1}^{n} p^{i-1} Y^{(i)^*} (B \otimes E) Y^{(i)} \\ + (-1 + \sigma q^{-1}) h_1 \sum_{i=m}^{n-m} p^{i-1} Y^{(i)^*} (B \otimes E) Y^{(i)} \\ + \sigma q^{-1} h_1 \sum_{i=0}^{m-1} p^{i-1} Y^{(i)^*} (B \otimes E) Y^{(i)}. \end{aligned}$$

Therefore,

$$\sum_{i=n-m+1}^{n} p^i Y^{(i)^*} (B \otimes E) Y^{(i)} \leq \delta + (-1 + \sigma q^{-1}) \sum_{i=m}^{n-m} p^i Y^{(i)^*} (B \otimes E) Y^{(i)}.$$

Let $n = (k+2)m - 1$ for $k \geq 0$. We have

$$\begin{aligned} p^{(k+2)m} \sum_{i=1}^{m} p^{-i} Y^{((k+2)m-i)^*} (B \otimes E) Y^{((k+2)m-i)} \\ \leq \delta + (-1 + \sigma q^{-1}) \sum_{l=0}^{k-1} p^{(l+2)m} \sum_{i=1}^{m} p^{-i} Y^{((l+2)m-i)^*} (B \otimes E) Y^{((l+2)m-i)}, \end{aligned}$$

which gives

$$\sum_{i=1}^{m} p^{-i} Y^{((k+2)m-i)^*}(B\otimes E)Y^{((k+2)m-i)} \le q^{k+2}\delta + (-1+\sigma q^{-1})\sum_{l=1}^{k} q^l \sum_{i=1}^{m} p^{-i} Y^{((k+2-l)m-i)^*}(B\otimes E)Y^{((k+2-l)m-i)}.$$

We prove (4.8) by induction. When $k = 0$, (4.8) follows directly from the above inequality. Now we assume that (4.8) holds for every $k < j$ and show that it then also holds for $k = j$. It follows from the above inequality that

$$\begin{aligned}
&\sum_{i=1}^{m} p^{-i} Y^{((j+2)m-i)^*}(B\otimes E)Y^{((j+2)m-i)} \\
&\quad \le q^{j+2}\delta + (-1+\sigma q^{-1})\sum_{l=1}^{j} q^l \sum_{i=1}^{m} p^{-i} Y^{((j+2-l)m-i)^*}(B\otimes E)Y^{((j+2-l)m-i)} \\
&\quad \le q^{j+2}\delta + (-1+\sigma q^{-1})\sum_{l=1}^{j} q^{l+2}\sigma^{j-l}\delta \\
&\quad = q^2\sigma^j\delta.
\end{aligned}$$

Therefore, (4.8) holds for every $k \ge 0$. This completes the proof. ∎

COROLLARY 4.6. *Under the assumptions of Lemma* 4.5, (4.8) *implies that,*

$$Y^{(n)^*}(B\otimes E)Y^{(n)} \le \sigma^{n/m}\delta, \quad \textit{for every } n \ge m. \tag{4.10}$$

Proof. It follows from (4.8) that for every $k \ge 0$, $i \in \{1, \dots, m\}$,

$$Y^{((k+2)m-i)^*}(B\otimes E)Y^{((k+2)m-i)} \le p^i q^2 \sigma^k \delta = q^{2-i/m}\sigma^k\delta \le \sigma^{k+2-i/m}\delta, \tag{4.11}$$

which gives (4.10). ∎

THEOREM 4.7. *Suppose that the method (A, b, c) with a nonsingular matrix A is algebraically stable, $b_i > 0$ for all i and there exist a constant $\sigma \ge q$ and matrices G and E such that $\mathcal{H}(\sigma)$ is nonnegative definite. Then there exists a constant $\bar{C}$ such that for every $n \ge m$,*

$$\|y_{n+1}\|_* \le |R(\infty)|^{n+1-m}\|y_m\|_* + \bar{C}\sigma^{1/2}\delta^{1/2}\psi_{n+1-m}(|R(\infty)|, \sigma^{1/2m}), \tag{4.12}$$

where $\|\cdot\|_*$ denotes a norm on $\mathbb{C}^d$, and

$$\psi_n(x,y) = \begin{cases} \dfrac{x^n - y^n}{x-y}, & x \ne y, \\ nx^{n-1}, & x = y. \end{cases} \tag{4.13}$$

Proof. Since $b_i > 0$ for all i, it follows from Corollary 4.6 that there exists a constant $\bar{C}_1$ such that for every $n \ge m$

$$\sum_{j=1}^{s} \|Y_j^{(n)}\|_* \le \bar{C}_1 \sigma^{n/2m}\delta^{1/2}.$$

Considering (3.14) and the assumptions from the statement of the theorem, there exists a constant $\bar{C}_2$ such that

$$\|y_{n+1}\|_* \le |R(\infty)|\|y_n\|_* + \bar{C}_2 \sum_{j=1}^{s} \|Y_j^{(n)}\|_*.$$

By induction, we have

$$\begin{aligned}\|y_{n+1}\|_* &\le |R(\infty)|\|y_n\|_* + \bar{C}_1\bar{C}_2\sigma^{n/2m}\delta^{1/2} \\ &\le |R(\infty)|^{n+1-m}\|y_m\|_* + \bar{C}_1\bar{C}_2\sigma^{1/2}\delta^{1/2}\sum_{i=0}^{n-m}|R(\infty)|^i\sigma^{(n-m-i)/2m}.\end{aligned}$$

This implies inequality (4.12). ∎

Using the fact that $R(\infty) = 0$ for the Radau IA, Radau IIA, and Lobatto IIIC methods, and the fact that $|R(\infty)| = 1$ for the Gauss methods, we can state the following corollaries.

COROLLARY 4.8. *Suppose that there exist a constant $\sigma \ge q$ and matrices G and E such that $\mathcal{H}(\sigma)$ is nonnegative definite. Then for any Radau IA, Radau IIA, or Lobatto IIIC method, there exists a constant $\bar{C}$ such that for every $n \ge m$,*

$$\|y_{n+1}\|_* \le \bar{C}\sigma^{n/2m}\delta^{1/2}. \tag{4.14}$$

COROLLARY 4.9. *Suppose that there exist a constant $\sigma \in [q, 1)$ and matrices G and E such that $\mathcal{H}(\sigma)$ is nonnegative definite. Then for any Gauss method, there exists a constant $\bar{C}$ such that for every $n \ge m$,*

$$\|y_{n+1}\|_* \le \|y_m\|_* + \frac{\bar{C}\sigma^{1/2}\delta^{1/2}}{1-\sigma^{1/2m}}. \tag{4.15}$$

Remark 4.10. The Assumption (L) in Koto [21] is equivalent to the condition that there exist matrices G and E such that $\mathcal{H}(q)$ is nonnegative definite. Our result can be applied to the more general case $\sigma \ge q$.

5. Nonlinear stability. In this section, we derive conditions which guarantee the asymptotic stability of the numerical solution of nonlinear equations. First, we recall a result on the asymptotic stability of the analytical solution. Consider a system defined by the same function f as in (2.1) but with a different initial value,

$$\begin{cases} z'(t) = f(t, z(t), z(qt), z'(qt)), & t > 0, \\ z(0) = z_0. \end{cases} \tag{5.1}$$

Let $\langle\cdot,\cdot\rangle$ be an inner product on $\mathbb{C}^d$, let $\|\cdot\|$ be the corresponding norm, and let the function f satisfy the conditions

$$\begin{aligned}\mathrm{Re}\langle u_1 - u_2, f(t,u_1,v,\nu) - f(t,u_2,v,\nu)\rangle &\le \alpha(t)\|u_1-u_2\|^2, \\ &\text{for } t>0, u_1, u_2, v, \nu \in \mathbb{C}^d,\end{aligned} \tag{5.2}$$

$$\begin{aligned}\|f(t,u,v_1,\nu_1) - f(t,u,v_2,\nu_2)\| &\le \beta(t)\|v_1-v_2\| + \gamma(t)\|\nu_1-\nu_2\|, \\ &\text{for } t>0, u, v_1, v_2, \nu_1, \nu_2 \in \mathbb{C}^d,\end{aligned} \tag{5.3}$$

where $\alpha(t)$, $\beta(t)$, and $\gamma(t)$ are continuous functions. An application of Theorem 2.1 in Zennaro [27] (see also [5]) gives the following proposition.

PROPOSITION 5.1. *Suppose* $\gamma(t) = 0$ *and the functions* $\alpha(t)$ *and* $\beta(t)$ *satisfy*

$$\alpha(t) \le \alpha_0 < 0, \quad t > 0, \tag{5.4}$$

and, for some nonnegative real number $k_0 < 1$,

$$k_0\alpha(t) + \beta(t) \le 0, \quad t > 0. \tag{5.5}$$

Then, for the solutions $y(t)$ *and* $z(t)$ *of* (2.1) *and* (5.1), *it holds that*

$$\lim_{t\to\infty} \|y(t) - z(t)\| = 0. \tag{5.6}$$

Remark 5.2. In the literature, we have not found any stability results on general nonlinear neutral equations. Some results on equations of special form, such as separable systems and equations of Hale's form, can be found in [3, 19, 26]. We do not give the details of those results because they cannot directly be applied to equations of the form (2.1). Here, we will only use conditions (5.4) and (5.5) plus the boundedness of $\gamma(t)$ to analyze the asymptotic stability of numerical methods for nonlinear neutral equations of the form (2.1). Stability results for nonneutral equations are given in [5] (Theorem 9.7.1).

The Runge–Kutta method (A, b, c) applied to problem (5.1) leads to the following process:

$$Z_i^{(n)} = z_n + h_n \sum_{j=1}^{s} a_{ij} \hat{Z}_j^{(n)}, \quad i = 1, \dots, s, \tag{5.7}$$

$$\hat{Z}_j^{(n)} = f(t_n + c_j h_n, Z_j^{(n)}, Z_j^{(n-m)}, \hat{Z}_j^{(n-m)}), \quad j = 1, \dots, s, \tag{5.8}$$

$$z_{n+1} = z_n + h_n \sum_{j=1}^{s} b_j \hat{Z}_j^{(n)}. \tag{5.9}$$

Let

$$w_n = y_n - z_n, \quad W_j^{(n)} = Y_j^{(n)} - Z_j^{(n)}, \quad j = 1, \dots, s.$$

It follows from (2.4–2.6) and (5.7–5.9) that

$$W_i^{(n)} = w_n + h_n \sum_{j=1}^{s} a_{ij} (\hat{Y}_j^{(n)} - \hat{Z}_j^{(n)}), \quad i = 1, \dots, s, \tag{5.10}$$

$$w_{n+1} = w_n + h_n \sum_{j=1}^{s} b_j (\hat{Y}_j^{(n)} - \hat{Z}_j^{(n)}). \tag{5.11}$$

Now we are in the position to state and prove the main result of this section.

THEOREM 5.3. *Suppose that the method* (A, b, c) *with a nonsingular matrix* A *is strictly stable at infinity and that there exist positive constants* C_3, C_4, *and* k_2 *such that*

$$0 < -\alpha^{-1}(t) \le C_3, \quad |\alpha^{-1}(t)\gamma(t)| \le C_4, \quad |\alpha^{-1}(t)\beta(t)| \le k_2 < 1. \tag{5.12}$$

Then, the following results hold:

$$\lim_{n\to\infty} \|w_n\| = 0, \qquad \lim_{n\to\infty} \sum_{j=1}^{s} \|W_j^{(n)}\| = 0, \tag{5.13}$$

$$\lim_{n\to\infty} h_n \sum_{j=1}^{s} \|\hat{Y}_j^{(n)} - \hat{Z}_j^{(n)}\| = 0. \tag{5.14}$$

Proof. From (5.10) and the nonsingularity of A it follows that

$$\hat{Y}_i^{(n)} - \hat{Z}_i^{(n)} = h_n^{-1} \sum_{j=1}^{s} D_{ij}(W_j^{(n)} - w_n), \quad i = 1, \dots, s, \tag{5.15}$$

where $D = [D_{ij}] = A^{-1}$. Substituting (5.15) into (5.11) yields

$$w_{n+1} = R(\infty) w_n + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i D_{ij} W_j^{(n)}. \tag{5.16}$$

Hence, there exists a constant $C_5 > 0$ such that

$$\|w_{n+1}\| \le |R(\infty)| \|w_n\| + C_5 \sum_{i=1}^{s} \|W_i^{(n)}\|. \tag{5.17}$$

On the other hand, conditions (5.2) and (5.3) imply that $\mathrm{Re}\langle W_j^{(n)}, \hat{Y}_j^{(n)} - \hat{Z}_j^{(n)} \rangle$ can be rewritten and bounded as follows:

$$\begin{aligned}
&\mathrm{Re}\langle W_j^{(n)}, f(t_n + c_j h_n, Y_j^{(n)}, Y_j^{(n-m)}, \hat{Y}_j^{(n-m)}) - f(t_n + c_j h_n, Z_j^{(n)}, Y_j^{(n-m)}, \hat{Y}_j^{(n-m)}) \rangle \\
&\quad + \mathrm{Re}\langle W_j^{(n)}, f(t_n + c_j h_n, Z_j^{(n)}, Y_j^{(n-m)}, \hat{Y}_j^{(n-m)}) - f(t_n + c_j h_n, Z_j^{(n)}, Z_j^{(n-m)}, \hat{Z}_j^{(n-m)}) \rangle \\
&\le \alpha(t_n + c_j h_n) \|W_j^{(n)}\|^2 + \beta(t_n + c_j h_n) \|W_j^{(n)}\| \|W_j^{(n-m)}\| \\
&\quad + \gamma(t_n + c_j h_n) \|W_j^{(n)}\| \|\hat{Y}_j^{(n-m)} - \hat{Z}_j^{(n-m)}\|.
\end{aligned}$$

Considering the inequality

$$\mathrm{Re}\langle W_j^{(n)}, \hat{Y}_j^{(n)} - \hat{Z}_j^{(n)} \rangle \ge -\|W_j^{(n)}\| \|\hat{Y}_j^{(n)} - \hat{Z}_j^{(n)}\|,$$

we have that

$$\|W_j^{(n)}\| \le k_2 \|W_j^{(n-m)}\| + C_4 \|\hat{Y}_j^{(n-m)} - \hat{Z}_j^{(n-m)}\| + C_3 \|\hat{Y}_j^{(n)} - \hat{Z}_j^{(n)}\|, \tag{5.18}$$

where we have used Assumption (5.12). Using (5.15), we further obtain

$$\begin{aligned}
\sum_{i=1}^{s} \|W_i^{(n)}\| \le k_2 \sum_{i=1}^{s} \|W_i^{(n-m)}\| &+ h_{n-m}^{-1} C_4 \sum_{i=1}^{s} \sum_{j=1}^{s} |D_{ij}| \|W_j^{(n-m)} - w_{n-m}\| \\
&+ h_n^{-1} C_3 \sum_{i=1}^{s} \sum_{j=1}^{s} |D_{ij}| \|W_j^{(n)} - w_n\|.
\end{aligned} \tag{5.19}$$

Considering $h_n \to \infty$, there exist positive numbers $\mathcal{N}_2, k_3 < 1$ and $C_6 < (1 - k_3)(1 - |R(\infty)|)/(2C_5)$ such that for every $n > \mathcal{N}_2$,

$$\sum_{i=1}^{s} \|W_i^{(n)}\| \le k_3 \sum_{i=1}^{s} \|W_i^{(n-m)}\| + C_6(\|w_n\| + \|w_{n-m}\|). \tag{5.20}$$

An application of Lemma 3.3 to (5.17) and (5.20) gives (5.13). Then (5.14) follows from (5.15). This completes the proof. ∎

Remark 5.4. In the proof, we only use $h_n \to \infty$. Hence, the result is also valid for the other grid types proposed in [2, 24]. From Corollary 3.9 we can see that the assumption of strict stability at infinity is also necessary to the result.

Remark 5.5. Theorem 5.3 is different from the asymptotic stability result obtained by Zhang and Sun [28] (see Theorem 6.1 of this paper). Therefore, specializing Theorem 5.3 to the case of nonneutral equations, our result is also new. In addition, our proof is completely different from that in [28].

Remark 5.6. It should be pointed out that Theorem 5.3 cannot cover the asymptotic stability results of section 3. In fact, specializing the Assumption (5.12) in Theorem 5.3 to the case of (3.1), the induced assumptions are stronger than Assumption $\mathcal{A}$.

6. An upper bound of error growth for nonlinear problems. In this section, we investigate error growth bounds of numerical methods for nonneutral, nonlinear problems of the form

$$\begin{cases} y'(t) = f(t, y(t), y(qt)), & t > 0, \\ y(0) = y_0, \end{cases} \tag{6.1}$$

where the function f satisfies the conditions

$$\mathrm{Re}\langle u_1 - u_2, f(t, u_1, v) - f(t, u_2, v)\rangle \le \alpha \|u_1 - u_2\|^2, \quad t > 0,\ u_1, u_2, v \in \mathbb{C}^d, \tag{6.2}$$

$$\|f(t, u, v_1) - f(t, u, v_2)\| \le \beta \|v_1 - v_2\|, \quad t > 0,\ u, v_1, v_2 \in \mathbb{C}^d, \tag{6.3}$$

where α and β are constants. Throughout this section, we assume that the other notations are the same as those in section 5.

For problems (6.1–6.3), Zhang and Sun (cf. [28]) have considered a stepsize strategy where every interval $(q^{-i}h, q^{-i-1}h]$ is divided into m subintervals of the same size, and obtained the following global and asymptotic stability results for $(k, l)-$ algebraically stable Runge–Kutta methods. Here, a method (A, b, c) is said to be (k, l)-algebraically stable if there exists a nonnegative diagonal matrix D such that the matrix

$$\begin{bmatrix} k - 1 - 2le^T De & e^T D - b^T - 2le^T DA \\ De - b - 2lA^T De & DA + A^T D - bb^T - 2lA^T DA \end{bmatrix}$$

is nonnegative definite (cf. [11]).

THEOREM 6.1 (see [28]). *Suppose that the method (A, b, c) is (k, l)-algebraically stable for a nonnegative diagonal matrix $D = \mathrm{diag}(d_1, d_2, \ldots, d_s)$, where $0 < k \le 1$ and the following condition holds:*

$$q\alpha + \beta \le 0, \quad (1 - q)(q\alpha + \beta)h \le mq^2 l, \tag{6.4}$$

then

$$\|w_{n+1}\| \le \left[1 + \sqrt{(q^{-1} - 1)\beta h}\right] \max\left\{\|w_m\|, \sum_{j=1}^{s} \left(\sqrt{b_j} \max_{-m \le i \le -1} \|W_j^{(m+i)}\|\right)\right\}. \tag{6.5}$$

If it is further assumed that $k < 1$, then

$$\lim_{n \to \infty} \|y_n - z_n\| = 0.$$

Here we obtain the following results.

THEOREM 6.2. *Suppose that the method (A, b, c) is algebraically stable and that the following condition holds:*

$$\alpha + q^{-1/2}\beta \le 0. \tag{6.6}$$

Then we have that for every $n \ge 2m-1$,

$$\begin{aligned} &\|w_{n+1}\|^2 - h_1(2\alpha + q^{-1/2}\beta) \sum_{i=n-m+1}^{n} \sum_{j=1}^{s} b_j p^{i-1} \|W_j^{(i)}\|^2 \\ &\qquad \le \|w_m\|^2 + h_1 q^{-1/2}\beta \sum_{i=0}^{m-1} \sum_{j=1}^{s} b_j p^{i-1} \|W_j^{(i)}\|^2. \end{aligned} \tag{6.7}$$

Proof. It is known (see [10]) that

$$\begin{aligned} &\|w_{n+1}\|^2 - \|w_n\|^2 - 2h_n \sum_{j=1}^{s} b_j \operatorname{Re}\langle W_j^{(n)}, \hat{Y}_j^{(n)} - \hat{Z}_j^{(n)} \rangle \\ &\qquad = -h_n^2 \sum_{i=1}^{s} \sum_{j=1}^{s} \mathcal{M}_{ij} \langle \hat{Y}_i^{(n)} - \hat{Z}_i^{(n)}, \hat{Y}_j^{(n)} - \hat{Z}_j^{(n)} \rangle. \end{aligned}$$

By means of the algebraic stability of the method and by (6.2) and (6.3), we have

$$\begin{aligned} \|w_{n+1}\|^2 &\le \|w_n\|^2 + 2h_n \sum_{j=1}^{s} b_j \operatorname{Re}\langle W_j^{(n)}, \hat{Y}_j^{(n)} - \hat{Z}_j^{(n)} \rangle \\ &= \|w_n\|^2 + 2h_n \sum_{j=1}^{s} b_j \operatorname{Re}\langle W_j^{(n)}, f(t_n + c_j h_n, Y_j^{(n)}, Y_j^{(n-m)}) \\ &\qquad - f(t_n + c_j h_n, Z_j^{(n)}, Y_j^{(n-m)}) \rangle \\ &\quad + 2h_n \sum_{j=1}^{s} b_j \operatorname{Re}\langle W_j^{(n)}, f(t_n + c_j h_n, Z_j^{(n)}, Y_j^{(n-m)}) \\ &\qquad - f(t_n + c_j h_n, Z_j^{(n)}, Z_j^{(n-m)}) \rangle \\ &\le \|w_n\|^2 + h_n \sum_{j=1}^{s} b_j [(2\alpha + q^{-1/2}\beta)\|W_j^{(n)}\|^2 + q^{1/2}\beta \|W_j^{(n-m)}\|^2], \end{aligned}$$

where we have used

$$2\|W_j^{(n)}\| \|W_j^{(n-m)}\| \le q^{-1/2}\|W_j^{(n)}\|^2 + q^{1/2}\|W_j^{(n-m)}\|^2. \tag{6.8}$$

By induction, one arrives at

$$\begin{aligned} \|w_{n+1}\|^2 \le \|w_m\|^2 + h_1 \sum_{j=1}^{s} b_j \Bigg[&\sum_{i=n-m+1}^{n} p^{i-1}(2\alpha + q^{-1/2}\beta)\|W_j^{(i)}\|^2 \\ &+ 2 \sum_{i=m}^{n-m} p^{i-1}(\alpha + q^{-1/2}\beta)\|W_j^{(i)}\|^2 \\ &+ \sum_{i=0}^{m-1} p^{i-1} q^{-1/2}\beta \|W_j^{(i)}\|^2 \Bigg], \end{aligned} \tag{6.9}$$

which, combined with (6.6), gives (6.7). This completes the proof. ∎

COROLLARY 6.3. *Under the assumptions of Theorem* 6.2*, we have that for every* $n \geq 2m-1$*,*

$$\|w_{n+1}\|^2 \leq \|w_m\|^2 + hp^{-1}q^{-1/2}(1-q)\beta \max_{\substack{0\leq i\leq m-1\\ 1\leq j\leq s}} \|W_j^{(i)}\|^2. \tag{6.10}$$

Proof. The conclusion follows from the fact that

$$\sum_{i=0}^{m-1} p^i \|W_j^i\|^2 \leq \frac{p^m-1}{p-1} \max_{0\leq i\leq m-1} \|W_j^{(i)}\|^2. \qquad \blacksquare$$

Remark 6.4. The above proof procedure is closely related to its counterpart in the case of a constant delay (cf. [15]) and in the case of a proportional delay (cf. [28]). Compared to Theorem 6.1 in the case of algebraically stable methods, our result is slightly sharper because Assumption (6.4) is stronger than (6.6).

Remark 6.5. In the case of time-dependent Lipschitz constants, if (6.6) is replaced by

$$2\alpha(t) + q^{-1/2}(\beta(t) + \beta(t/q)) \leq 0, \tag{6.11}$$

we can similarly obtain that for every $n \geq 2m-1$,

$$\begin{aligned} &\|w_{n+1}\|^2 - h_1 \sum_{i=n-m+1}^{n} \sum_{j=1}^{s} b_j p^{i-1}(2\alpha(t_i + c_j h_i) + q^{-1/2}\beta(t_i + c_j h_i))\|W_j^{(i)}\|^2 \\ &\leq \|w_m\|^2 + hp^{-1}q^{-1/2}(1-q) \max_{\substack{0\leq i\leq m-1\\ 1\leq j\leq s}} \beta(t_{i+m} + c_j h_{i+m})\|W_j^{(i)}\|^2. \end{aligned} \tag{6.12}$$

Next, we derive some results which can be applied to the more general case $\alpha + \beta < 0$. We define the following two constants

$$r = \begin{cases} 0, & \text{when } 2\alpha + (1+q^{-1})\beta \leq 0, \\ \dfrac{2\alpha + (1+q^{-1})\beta}{-(2\alpha+\beta)}, & \text{when } 2\alpha + (1+q^{-1})\beta > 0, \end{cases} \tag{6.13}$$

$$\Delta = \frac{\|w_m\|^2 p/h_1 + q^{-1}\beta \sum_{j=1}^{s} b_j \sum_{i=0}^{m-1} p^i \|W_j^{(i)}\|^2}{-(2\alpha+\beta)}. \tag{6.14}$$

THEOREM 6.6. *Suppose that the method* (A, b, c) *is algebraically stable and that the following condition holds:*

$$\alpha + \beta < 0. \tag{6.15}$$

Then we have that for every $k \geq 0$*,*

$$\sum_{j=1}^{s} b_j \sum_{i=1}^{m} p^{-i} \|W_j^{((k+2)m-i)}\|^2 \leq q^2(q+qr)^k \Delta. \tag{6.16}$$

Proof. If we replace (6.8) by the inequality

$$2\|W_j^{(n)}\|\|W_j^{(n-m)}\| \leq \|W_j^{(n)}\|^2 + \|W_j^{(n-m)}\|^2,$$

we obtain

$$\|w_{n+1}\|^2 \le \|w_n\|^2 + h_n \sum_{j=1}^{s} b_j[(2\alpha+\beta)\|W_j^{(n)}\|^2 + \beta\|W_j^{(n-m)}\|^2].$$

By induction, one arrives at

$$\|w_{n+1}\|^2 \le \|w_m\|^2 + h_1 p^{-1} \sum_{j=1}^{s} b_j \left[\sum_{i=n-m+1}^{n} p^i(2\alpha+\beta)\|W_j^{(i)}\|^2 + \sum_{i=m}^{n-m} p^i(2\alpha+(1+q^{-1})\beta)\|W_j^{(i)}\|^2 + \sum_{i=0}^{m-1} p^i q^{-1}\beta\|W_j^{(i)}\|^2 \right].$$

Therefore,

$$\sum_{j=1}^{s} b_j \sum_{i=n-m+1}^{n} p^i \|W_j^{(i)}\|^2 \le r \sum_{j=1}^{s} b_j \sum_{i=m}^{n-m} p^i \|W_j^{(i)}\|^2 + \Delta.$$

Let $n = (k+2)m - 1$ for $k \ge 0$. We then have

$$\sum_{j=1}^{s} b_j \sum_{i=1}^{m} p^{-i} \|W_j^{((k+2)m-i)}\|^2 \le r \sum_{l=1}^{k} q^l \sum_{j=1}^{s} b_j \sum_{i=1}^{m} p^{-i} \|W_j^{((k-l+2)m-i)}\|^2 + q^{k+2}\Delta.$$

By induction we can prove from the above inequality that (6.16) holds for every $k \ge 0$. This completes the proof. ∎

Remark 6.7. It is easy to verify the following inequality,

$$0 < q + qr \le \max\left(q, \frac{\beta}{-(2\alpha+\beta)}\right).$$

Hence, $q + qr < 1$ if $\alpha + \beta < 0$ and the right-hand side of (6.16) goes to zero for increasing k.

Remark 6.8. If condition (6.15) is replaced by the weaker condition

$$2\alpha + \beta < 0, \tag{6.17}$$

then Theorem 6.6 is still valid. Then, however, it is not guaranteed that $q + qr < 1$.

COROLLARY 6.9. *Under the assumptions of Theorem* 6.6*, inequality* (6.16) *implies that for every* $n \ge m$

$$\sum_{j=1}^{s} b_j \|W_j^{(n)}\|^2 \le (q+qr)^{n/m}\Delta. \tag{6.18}$$

Proof. It follows from (6.16) that for every $k \ge 0, i \in \{1, \dots, m\}$,

$$\sum_{j=1}^{s} b_j \|W_j^{((k+2)m-i)}\|^2 \le p^i q^2 (q+qr)^k \Delta = q^{2-i/m}(q+qr)^k\Delta \le (q+qr)^{k+2-i/m}\Delta,$$

which gives (6.18). ∎

THEOREM 6.10. *Suppose that the method (A, b, c) with a nonsingular matrix A is algebraically stable, $b_i > 0$ for all i and* (6.15) *holds. Then there exists a constant C_7, depending only on the coefficients of the method, such that for every $n \geq m$,*

$$\|w_{n+1}\| \leq |R(\infty)|^{n+1-m}\|w_m\| + C_7(q+qr)^{1/2}\Delta^{1/2}\psi_{n+1-m}(|R(\infty)|, (q+qr)^{1/2m}),$$

where the function $\psi_n(x, y)$ is defined by (4.13).

Proof. Considering (5.16) and the assumptions of the theorem, there exists a constant C_7, depending only on the coefficients, such that

$$\|w_{n+1}\| \leq |R(\infty)|\|w_n\| + C_7\left(\sum_{j=1}^{s} b_j\|W_j^{(n)}\|^2\right)^{1/2},$$

which in combination with (6.18) gives

$$\begin{aligned}\|w_{n+1}\| &\leq |R(\infty)|\|w_n\| + C_7(q+qr)^{n/2m}\Delta^{1/2}\\ &\leq |R(\infty)|^{n+1-m}\|w_m\|\\ &\quad + C_7(q+qr)^{1/2}\Delta^{1/2}\sum_{i=0}^{n-m}|R(\infty)|^i(q+qr)^{(n-m-i)/2m}.\end{aligned}$$

This implies the result of the theorem. ∎

COROLLARY 6.11. *Suppose* (6.15) *holds. Then for any Radau IA, Radau IIA, or Lobatto IIIC method, there exists a constant C_7 depending only on the coefficients of the method, such that for every $n \geq m$,*

$$\|w_{n+1}\| \leq C_7(q+qr)^{n/2m}\Delta^{1/2}.$$

COROLLARY 6.12. *Suppose* (6.15) *holds. Then for any Gauss method there exists a constant C_7, depending only on the coefficients of the method, such that for every $n \geq m$,*

$$\|w_{n+1}\| \leq \|w_m\| + \frac{C_7(q+qr)^{1/2}\Delta^{1/2}}{1-(q+qr)^{1/2m}}.$$

Remark 6.13. It is easy to extend the results of this section to the case of (k, l)-algebraically stable methods if we impose some restrictions on stepsize similar to those in [15].

7. Extension to a more general class of equations. In this section, we generalize some stability results to the more general equations

$$\text{(7.1)}\qquad \begin{cases} y'(t) = L(t)y(t) + M(t)y(t-\tau(t)) + N(t)y'(t-\tau(t)), & t \geq t_0,\\ y(t) = g(t), & t \leq t_0, \end{cases}$$

and

$$\text{(7.2)}\qquad \begin{cases} y'(t) = f(t, y(t), y(t-\tau(t)), y'(t-\tau(t))), & t \geq t_0,\\ y(t) = g(t), & t \leq t_0. \end{cases}$$

The ideas are related to those in [13]. We assume that there exists a constrained mesh in the interval $[t_0, +\infty)$ such that

$$\text{(7.3)}\qquad t_n - \tau(t_n) = t_{n-m}, \quad n \geq m,$$

for some integer m. The corresponding discretized schemes for (7.1) and (7.2) are (3.5–3.7) and (2.4–2.6), respectively. Then, a similar analysis leads to the following general results.

THEOREM 7.1. *Suppose that the method (A, b, c) with a nonsingular matrix A is strictly stable at infinity and that the coefficient matrices of* (7.1) *satisfy*

(7.4)
$$\lim_{n\to\infty} h_n^{-1}\|L^{-1}(t_n + c_i h_n)\| = 0, \quad \lim_{n\to\infty} h_{n-m}^{-1}\|L^{-1}(t_n + c_i h_n)N(t_n + c_i h_n)\| = 0,$$

and

$$\limsup_{n\to\infty} \|L^{-1}(t_n + c_i h_n)M(t_n + c_i h_n)\| = k_0 < 1, \tag{7.5}$$

for $i = 1, 2, \ldots, s$. Then, the scheme (3.5–3.7) *is asymptotically stable for* (7.1).

THEOREM 7.2. *Suppose that the method (A, b, c) with a nonsingular matrix A is strictly stable at infinity and that the function f of* (7.2) *satisfies conditions* (5.2) *and* (5.3) *with*

$$\lim_{n\to\infty} h_n\alpha(t_n + c_i h_n) = -\infty, \qquad \lim_{n\to\infty} \frac{\gamma(t_n + c_i h_n)}{h_{n-m}\alpha(t_n + c_i h_n)} = 0, \tag{7.6}$$

and

$$\limsup_{n\to\infty} \frac{\beta(t_n + c_i h_n)}{|\alpha(t_n + c_i h_n)|} = k_2 < 1, \tag{7.7}$$

for $i = 1, 2, \ldots, s$. Then, the scheme (2.4–2.6) *is asymptotically stable for* (7.2)*, i.e.,* (5.13) *and* (5.14) *hold.*

Remark 7.3. For the constant delay system (1.3), we can consider a constant stepsize strategy, i.e., $h_n = -\log q/m$, such that (7.3), (7.4), and (7.5) hold if $\rho[L^{-1}M] < 1$.

Remark 7.4. If the delay $\tau(t)$ satisfies the conditions

$$0 < \tau_0 \le \tau(t) < t, \quad t \ge t_0, \tag{7.8}$$

$$0 < q_* \le 1 - \tau'(t) \le q^* < 1, \quad t \ge t_0, \tag{7.9}$$

then $t - \tau(t)$ is strictly increasing and $\lim_{t\to\infty} t - \tau(t) = +\infty$, which guarantees the existence of a constrained mesh (cf. [5, 13]) and

$$h_{n-m} = h_n - (\tau(t_{n+1}) - \tau(t_n)) \le q^* h_n.$$

Hence,

$$\lim_{n\to\infty} h_n = +\infty,$$

which implies (7.4) if $L^{-1}(t)$ and $N(t)$ are bounded. In the case of vanishing delays, i.e., $\tau(0) = 0$, we can choose an appropriate point $t_0 > 0$ and assume we know an approximate solution in $[0, t_0]$ such that (7.8) hold.

Remark 7.5. If $c_i \in \{0, 1\}$ or $\tau(t)$ is of the form $qt + d$, we can appropriately choose stepsize such that (7.3) implies

$$t_n + c_i h_n - \tau(t_n + c_i h_n) = t_{n-m} + c_i h_{n-m},$$

see [5, section 6.3]. In the other case, however, the above equality may no longer hold, an interpolation for the delay argument is possibly necessary and a rigorous theoretical analysis is missing (and outside the scope of the present paper).

8. Concluding remarks. In this work the stability of Runge–Kutta methods for both linear and nonlinear nonautonomous pantograph equations has been analyzed. A new approach has been introduced to derive the asymptotic stability of numerical methods and some sufficient and necessary conditions have been found. By further exploiting the special structure of the stepsize, we have also obtained some upper bounds for the error growth.

The techniques of this paper can be applied to investigate the stability of numerical methods for the constant delay system derived from the pantograph equation. They could also possess a potential applicability to integro-differential equations with proportional delays.

Acknowledgments. The authors are grateful to the anonymous referee for his valuable comments and remarks. They are also indebted to Professor Hermann Brunner for suggesting the extension of the study to a more general class of equations.

REFERENCES

[1] C. T. H. Baker, *Retarded differential equations*, J. Comput. Appl. Math., 125 (2000), pp. 309–335.
[2] A. Bellen, N. Guglielmi, and L. Torelli, *Asymptotic stability properties of θ-methods for the pantograph equation*, Appl. Numer. Math., 24 (1997), pp. 279–293.
[3] A. Bellen, N. Guglielmi, and M. Zennaro, *Numerical stability of nonlinear delay differential equations of neutral type*, J. Comput. Appl. Math., 125 (2000), pp. 251–263.
[4] A. Bellen, S. Maset, and L. Torelli, *Contractive initializing methods for the pantograph equation of neutral type*, Recent Trends in Numerical Analysis, 3 (2000), pp. 35–41.
[5] A. Bellen and M. Zennaro, *Numerical Methods for Delay Differential Equations*, Oxford University Press, Oxford, UK, 2003.
[6] M. D. Buhmann and A. Iserles, *Numerical analysis of functional equations with a variable delay*, in Numerical Analysis 1991 (Dundee, 1991), D. F. Griffiths and G. A. Watson, eds., Longman, Sci. Tech., Harlow, UK, 1992, pp. 17–33.
[7] M. D. Buhmann and A. Iserles, *On the dynamics of a discretized neutral equation*, IMA J. Numer. Anal., 12 (1992), pp. 339–363.
[8] M. D. Buhmann and A. Iserles, *Stability of the discretized pantograph differential equation*, Math. Comp., 60 (1993), pp. 575–589.
[9] M. D. Buhmann, A. Iserles, and S. P. Norsett, *Runge-Kutta methods for neutral differential equations*, in Contributions in Numerical Mathematics, World Sci. Ser. Appl. Anal. 2, World Scientific, River Edge, NJ, 1993, pp. 85–98.
[10] K. Burrage and J. C. Butcher, *Stability criteria for implicit Runge–Kutta methods*, SIAM J. Numer. Anal., 16 (1979), pp. 46–57.
[11] K. Burrage and J. C. Butcher, *Nonlinear stability of a general class of differential equation methods*, BIT, 20 (1980), pp. 185–203.
[12] K. Dekker and J. G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, CWI Monographs 2, North–Holland, Amsterdam, 1984.
[13] N. Guglielmi and M. Zennaro, *Stability of one-leg Θ-methods for the variable coefficient pantograph equation on the quasi-geometric mesh*, IMA J. Numer. Anal., 23 (2003), pp. 421–438.
[14] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations. Stiff and Differential-Algebraic Problems* II, Springer-Verlag, Berlin, 1991.
[15] C. Huang, H. Fu, S. Li, and G. Chen, *Stability analysis of Runge-Kutta methods for nonlinear delay differential equations*, BIT, 39 (1999), pp. 270–280.
[16] A. Iserles, *On the generalized pantograph functional-differential equations*, European J. Appl. Math., 4 (1993), pp. 1–38.
[17] A. Iserles, *Numerical analysis of delay differential equations with variable delays*, Ann. Numer. Math., 1 (1994), pp. 133–152.
[18] A. Iserles, *Exact and discretized stability of the pantograph equation*, Appl. Numer. Math., 24 (1997), pp. 295–308.
[19] A. Iserles and J. Terjeki, *Stability and asymptotic stability of functional-differential equations*, J. London Math. Soc. (2), 51 (1995), pp. 559–572.

[20] Z. Jackiewicz, *Asymptotic stability analysis of θ-methods for functional differential equations*, Numer. Math., 43 (1984), pp. 389–396.
[21] T. Koto, *Stability of Runge-Kutta methods for the generalized pantograph equation*, Numer. Math., 84 (1999), pp. 233–247.
[22] Y. Liu, *Stability analysis of θ-methods for neutral functional-differential equations*, Numer. Math., 70 (1995), pp. 473–485.
[23] Y. Liu, *Asymptotic behaviour of functional-differential equations with proportional time delay*, European J. Appl. Math., 7 (1996), pp. 11–30.
[24] Y. Liu, *On the θ-methods for delay differential equations with infinite lag*, J. Comput. Appl. Math., 71 (1996), pp. 177–190.
[25] Y. Liu, *Numerical investigation of the pantograph equation*, Appl. Numer. Math., 24 (1997), pp. 309–317.
[26] R. Vermiglio and L. Torelli, *A stable numerical approach for implicit non-linear neutral delay differential equations*, BIT, 43 (2003), pp. 195–C215.
[27] M. Zennaro, *Asymptotic stability analysis of Runge-Kutta methods for nonlinear systems of delay differential equations*, Numer. Math., 77 (1997), pp. 549–563.
[28] C. J. Zhang and G. Sun, *Nonlinear stability of variable stepsize Runge-Kutta methods applied to infinite delay-differential equations*, Math. Comput. Modelling, 39 (2004), pp. 495–503.
[29] C. J. Zhang and G. Sun, *The discrete dynamics of nonlinear infinite-delay-differential equations*, Appl. Math. Lett., 15 (2002), pp. 521–526.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2043–2075

ADAPTIVE OPTIMIZATION OF CONVEX FUNCTIONALS IN BANACH SPACES*

CLAUDIO CANUTO† AND KARSTEN URBAN‡

Abstract. This paper is concerned with optimization or minimization problems that are governed by operator equations, such as partial differential or integral equations, and thus are naturally formulated in an infinite dimensional function space V. We first construct a prototype algorithm of steepest descent type in V and prove its convergence. By using a Riesz basis in V we can transform the minimization problem into an equivalent one posed in a sequence space of type ℓ_p. We convert the prototype algorithm into an adaptive method in ℓ_p. This algorithm is shown to be convergent under mild conditions on the parameters that appear in the algorithm. Under more restrictive assumptions we are also able to establish the rate of convergence of our algorithm and prove that the work/accuracy balance is asymptotically optimal. Finally, we give two particular examples.

Key words. optimization, convex analysis, steepest descent method, adaptive methods, nonlinear approximation, wavelet bases

AMS subject classifications. 65K10, 65T60

DOI. 10.1137/S0036142903429730

1. Introduction.

Optimization or minimization problems arise in many areas of modern science and technology. As examples let us mention control theory, image processing and segmentation, drag reduction, and shape optimization. Constrained and nonsmooth optimization, such as in modelling American options, and elastoplastic hardening and softening, pose additional challenges and require special care. Here we are particularly interested in problems that are governed by operator equations, such as partial differential or integral equations, and thus are naturally formulated in infinite-dimensional function spaces (see, e.g., [1, 7, 24, 28]).

The numerical solution of minimization problems in function spaces is traditionally based upon the choice of a suitable discretization of the spaces, leading to similar problems in finite dimension; these are then solved by some of the available optimization algorithms in Euclidean spaces. Such an approach may incorporate an adaptive strategy, which is highly appropriate for those problems whose minimizers contain well-localized structures. With the aid of a posteriori error analysis, a refinement or derefinement (coarsening) of the current discretization can be constructed yielding to an adaptively generated sequence of finite-dimensional discretizations. Even though these kinds of adaptive methods have been successfully used in many different applications, the literature on the numerical analysis of adaptive methods for minimization problems seems fairly limited. Not much is known on the convergence of the adaptive iterations and even less on the convergence speed.

Starting from recent investigations concerning adaptive wavelet methods [10, 11, 12], it is by now clear that using as much information from the original infinite-dimensional variational problem as possible gives in particular a strong mathematical

*Received by the editors June 6, 2003; accepted for publication (in revised form) May 13, 2004; published electronically February 25, 2005. This work was partly supported by the European Commission within the Research Training Network *Breaking Complexity*, 2-2001-00574.

http://www.siam.org/journals/sinum/42-5/42973.html

†Dipartimento di Matematica, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy (ccanuto@polito.it).

‡Abteilung Numerik, University of Ulm, D-89069 Ulm, Germany (kurban@mathematik.uni-ulm.de).

tool to prove convergence and convergence rate results. The new philosophy consists of combining an infinite-dimensional iteration with an approximate finite application of the underlying exact operators. By using an analogous approach, similar results for adaptive finite element methods have been obtained [3, 23, 27]. An important ingredient in the design and analysis of such algorithms has been provided by recent results in nonlinear approximation theory [22]. They give guidelines for the definition of refinement and coarsening strategies based upon a rigorous control of the resulting errors as well as the number of degrees of freedom. They set the appropriate notion of optimality by indicating the best possible relation between accuracy and number of degrees of freedom in approximating the solution to the particular problem at hand. It turns out that the correct functional setting for this notion is provided by certain approximation spaces. In many cases, these approximation spaces are close to Besov spaces in certain scales, in which the summability index decreases as the regularity index increases (as opposed to traditional scales of Sobolev spaces for nonadaptive methods).

The purpose of the present paper is to take a first step towards the adaptive numerical treatment of infinite dimensional minimization problems following the new philosophy described above. To this end, we start with a fairly classical situation in which we want to minimize a strictly convex, Fréchet differentiable functional J defined on a reflexive Banach space V. The assumption of convexity guarantees the existence of a global minimizer and global convergence toward it, allowing us to concentrate on the central issues related to the infinite-dimensional setting. Since the ultimate goal of our investigations will be to handle a broad class of problems including constrained or nonsmooth optimization (see, e.g., [25]), the minimization strategy will be based upon a general method of steepest descent type, coupled with a line search. The advantage is that we do not require strong regularity assumptions such as existence or boundedness of the Hessian of the functional.

After making suitable assumptions on the well-posedness of the minimization problem, we construct a steepest descent algorithm in the Banach space V which is proven to be convergent. This algorithm will serve as a prototype for the adaptive algorithms constructed next. By introducing a Riesz basis in V and by considering the sequence of the expansion coefficients of a function with respect to the chosen basis, we transform the minimization problem into an equivalent one posed in a sequence space of type ℓ_p. This transformation is a crucial step toward the concrete realization of the algorithm and for the introduction of adaptive concepts.

Next, we convert the previous algorithm into a steepest descent method defined in the sequence space. While the exact minimizer is in general represented by a sequence with infinitely many nonvanishing entries, the new algorithm acts only on finite vectors (i.e., sequences having only a finite number of nonvanishing entries); hence, it is numerically feasible. Since the number of active coefficients may grow in the descent stages, we incorporate a coarsening procedure in order to remove unnecessary details from time to time. This adaptive algorithm is shown to be convergent under mild conditions on the parameters that appear in its definition.

Under more restrictive assumptions we are also able to investigate the rate of convergence of our algorithm. Precisely, we prove that the error between the exact and the approximate minimizer decays at least in a geometric manner as the number of iterations increases. The number of approximate evaluations of the gradient of the functional applied to a finite vector (which is usually the most expansive part of the algorithm) grows in an asymptotically optimal way, i.e., at most logarithmically in the accuracy. Finally, we can prove that the output of the algorithm is optimal in the sense

of nonlinear approximation theory (asymptotically optimal work/accuracy balance) under the condition that the gradient can be efficiently approximated in a sparse way and an optimal thresholding procedure is available to realize the coarsening.

The paper is organized as follows. We start from a fairly general framework of convex optimization in Banach spaces. Under these weak assumptions, we formulate the abstract algorithm in section 2 and prove its convergence. Section 3 is devoted to the transformation of the algorithm into an equivalent one set in an ℓ_p-space. This abstract setting yields a method in infinite dimension which is in general not computable. Hence, in section 4 we investigate the ingredients that are needed to define a computable adaptive version of our general algorithm and then we prove the convergence of the resulting method.

In the second part of the paper, starting with section 5, we specialize our setting to the ℓ_2-case and we assume Lipschitz continuity of the gradient. In this framework we indicate a precise choice of the parameters in our abstract algorithm and construct more efficient adaptive algorithmic ingredients in order to obtain also a rate of convergence. In section 6, we investigate the optimality of the algorithm. At last, two examples are provided in section 7.

We will frequently use the notation $A \lesssim B$, which means that there exists a constant $c >$ such that $A \leq c\,B$, uniformly in all parameters on which A and B may depend. The notation $A \sim B$ means $A \lesssim B$ and $B \lesssim A$.

2. Convex optimization in Banach spaces. Our setting is as follows. Let V be a reflexive Banach space normed by $\|\cdot\|_V$ and let us denote by $\langle\cdot,\cdot\rangle$ the duality pairing between V and V'. Let $J : V \to \mathbb{R}$ be a Fréchet differentiable functional. We consider the following minimization problem: Find $u \in V$ such that

$$J(u) = \min_{v\in V} J(v). \tag{2.1}$$

We start by collecting the general assumptions on the functional J that will hold throughout the paper. Then we will recall the concept of admissible descent directions and stepsizes adapted to the Banach space setting. Finally, we formulate our abstract algorithm and prove its convergence.

2.1. General assumptions.

ASSUMPTION 2.1. *The functional J satisfies the following conditions:*

(i) *The Fréchet derivative $J' : V \to V'$ is uniformly continuous on each bounded subset of V;*

(ii) *J is V-elliptic in the sense that there exist $c_J > 0$ and $p > 1$ such that*

$$\langle J'(w) - J'(v), w - v\rangle \geq c_J\|w - v\|_V^p$$

for all $w, v \in V$.

Let us collect some consequences of the latter assumption.

LEMMA 2.2. *Let Assumption 2.1 be satisfied. Then, the following statements hold.*

(a) *For all $w, v \in V$ we have*

$$J(v) - J(w) \geq \langle J'(w), v - w\rangle + \frac{c_J}{p}\|v - w\|_V^p. \tag{2.2}$$

(b) *J is strictly convex and bounded from below.*

(c) *Let $u^{(0)} \in V$ be arbitrary; then the set $\mathcal{R}(u^{(0)}) := \{v \in V : J(v) \leq J(u^{(0)})\}$ is bounded.*

The straightforward proof can be found in Appendix A (see also [8]). In particular, we see from (b) that our assumptions are slightly stronger than convexity.

Under Assumption 2.1, there exists a unique solution to the minimization problem (2.1) (see, e.g., [7] for a proof).

2.2. Descent directions and stepsizes. As already said, we aim at formulating an algorithm of steepest descent type that leads to the solution u of the optimization problem (2.1). Since we are working in abstract Banach spaces, we cannot expect to have orthogonal descent directions available, no matter what inner product is used. Hence, we have to specify what directions are admissible in order to yield a possible descent. We require that the direction is *not* orthogonal to the current gradient.

DEFINITION 2.3. *Given $v \in V$, we call $s \in V$ an* admissible descent direction *for v if $\|s\| = 1$ and $\langle J'(v), s\rangle < 0$.*

Once an admissible descent direction is determined, one has to find the minimum of the functional J along the search direction. Again, we cannot hope to be able to determine this minimum exactly, even though this is a one-dimensional (1D) minimization problem. In order to ensure that the line search in fact yields a smaller value of the functional then the starting one, we identify a possible range of stepsizes that guarantees this descent. We adapt the classical concept of admissible stepsizes as follows (see, e.g., [21, subsection 6.3]).

DEFINITION 2.4 (Wolfe's condition). *Let α, β be fixed constants satisfying $0 < \alpha < \beta < 1$. For any $v \in V$ and any admissible descent direction $s \in V$, define $\mathcal{A}(J; v, s)$ as the set of all those $\mu \in \mathbb{R}_+$ satisfying the following conditions:*

$$J(v + \mu s) \le J(v) + \alpha\, \mu\, \langle J'(v), s\rangle, \tag{2.3}$$

$$\langle J'(v + \mu s), s\rangle \ge \beta\, \langle J'(v), s\rangle. \tag{2.4}$$

We call $\mathcal{A}(J; v, s)$ the set of admissible stepsizes.

In order to understand the meaning of the previous conditions, it is convenient to introduce the auxiliary univariate function

$$\varphi(\mu) := J(v + \mu s), \qquad \mu \in \mathbb{R}, \tag{2.5}$$

which is a strictly convex function satisfying $\varphi(0) = J(v)$, $\varphi'(0) = \langle J'(v), s\rangle < 0$, and $\varphi(\mu) \to +\infty$ for $\mu \to +\infty$. Then, condition (2.3) reads

$$\varphi(\mu) \le \varphi(0) + \alpha\, \varphi'(0)\, \mu \tag{2.6}$$

and, due to the convexity of φ, it identifies an interval of the form $(0, \mu_{\max}]$. Conversely, condition (2.4) reads

$$\varphi'(\mu) \ge \beta \varphi'(0) \tag{2.7}$$

and identifies an interval of the form $[\mu_{\min}, +\infty)$. Then, $\mathcal{A}(J; v, s)$ is the intersection of these two intervals. It is not empty since $\alpha < \beta$; see also Figure 2.1.

2.3. A convergent steepest descent algorithm. We are now ready to introduce the steepest descent algorithm for solving the optimization problem (2.1). In this first version, we assume that all evaluations of both the functional J and its gradient J' are done exactly. Hence, we term this version the *Exact Algorithm*.

We need the following notation.

DEFINITION 2.5. *Let $R : V' \to V$ denote the* Riesz operator*; i.e., $R(f)$ is defined for $f \in V'$ as the unique element in V such that $\|R(f)\|_V = 1$ and $\|f\|_{V'} = \langle f, R(f)\rangle$.*

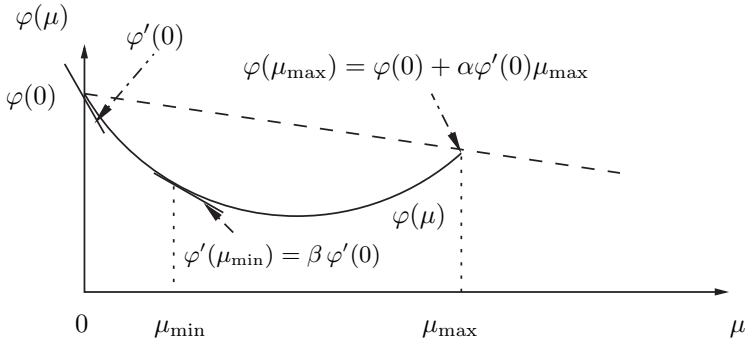


FIG. 2.1. *Conditions on the stepsize imposed by Definition* 2.4.

Obviously, $R(f)$ can also be characterized by

$$\langle f, R(f)\rangle = \sup_{v\in V} \frac{\langle f, v\rangle}{\|v\|}.$$

ALGORITHM 2.6 (Exact Algorithm). *Let $u^{(0)} \in V$ be given. Then, for $k = 0, 1, 2, \ldots$, while $J'(u^{(k)}) \neq 0$, do*

1. *choose the search direction as $s^{(k)} := -R(J'(u^{(k)}))$.*
2. *determine an admissible stepsize $\mu^{(k)} \in \mathcal{A}(J; u^{(k)}, s^{(k)})$.*
3. *update: $u^{(k+1)} := u^{(k)} + \mu^{(k)} s^{(k)}$.* □

PROPOSITION 2.7. *Under Assumption* 2.1, *the Exact Algorithm converges to u.*

Proof. Without loss of generality we assume that the algorithm produces an infinite sequence of vectors. We do the proof in several steps by adapting the convergence proof of the classical steepest descent method for a convex functional in *finite* dimension; see, e.g., [8].

(i) Using Lemma 2.2(a), we get

$$\frac{c_J}{p}\|u^{(k+1)} - u^{(k)}\|_V^p \leq [J(u^{(k+1)}) - J(u^{(k)})] - \langle J'(u^{(k)}), u^{(k+1)} - u^{(k)}\rangle.$$

Next, condition (2.3) can be equivalently written as

$$-\langle J'(u^{(k)}), u^{(k+1)} - u^{(k)}\rangle \leq \frac{1}{\alpha}[J(u^{(k)}) - J(u^{(k+1)})];$$

thus

$$\frac{c_J}{p}\|u^{(k+1)} - u^{(k)}\|_V^p \leq \left(\frac{1}{\alpha} - 1\right)[J(u^{(k)}) - J(u^{(k+1)})]. \tag{2.8}$$

(ii) By construction, we have that $\{J(u^{(k)})\}_{k\in\mathbb{N}_0}$ is monotonically decreasing and bounded from below by $J(u)$; hence

$$\lim_{k\to\infty}[J(u^{(k)}) - J(u^{(k+1)})] = 0.$$

Using (2.8), we obtain $\lim_{k\to\infty}\|u^{(k+1)} - u^{(k)}\|_V = 0$.

(iii) Now we have by condition (2.4)

$$\begin{aligned}\|J'(u^{(k)})\|_{V'} = \sup_{\|\varphi\|_V=1} \langle J'(u^{(k)}),\varphi\rangle &= -\langle J'(u^{(k)}), s^{(k)}\rangle \\ &= \langle J'(u^{(k+1)}) - J'(u^{(k)}), s^{(k)}\rangle - \langle J'(u^{(k+1)}), s^{(k)}\rangle \\ &\le \|J'(u^{(k+1)}) - J'(u^{(k)})\|_{V'}\,\|s^{(k)}\|_V - \beta\langle J'(u^{(k)}), s^{(k)}\rangle \\ &= \|J'(u^{(k+1)}) - J'(u^{(k)})\|_{V'} + \beta\,\|J'(u^{(k)})\|_{V'},\end{aligned}$$

and hence $\|J'(u^{(k)})\|_{V'} \le \frac{1}{1-\beta}\|J'(u^{(k+1)}) - J'(u^{(k)})\|_{V'}$.

(iv) The sequence $\{u^{(k)}\}_k$ lies in $\mathcal{R}(u^{(0)})$ since $J(u^{(k)}) \le J(u^{(k-1)}) \le \cdots \le J(u^{(0)})$. Hence by Assumption 2.1 and Lemma 2.2(c), we obtain

$$\|J'(u^{(k+1)}) - J'(u^{(k)})\|_{V'} \overset{k\to\infty}{\longrightarrow} 0,$$

which by (iii) implies

$$\lim_{k\to\infty} J'(u^{(k)}) = 0. \tag{2.9}$$

(v) Now using Assumption 2.1 and the fact that $J'(u) = 0$, we get

$$\begin{aligned}c_J\|u^{(k)} - u\|_V^p \le \langle J'(u^{(k)}) - J'(u), u^{(k)} - u\rangle &= \langle J'(u^{(k)}), u^{(k)} - u\rangle \\ &\le \|J'(u^{(k)})\|_{V'}\,\|u^{(k)} - u\|_V\end{aligned}$$

and finally by (2.9)

$$\|u^{(k)} - u\|_V \le \Big(\frac{1}{c_J}\|J'(u^{(k)})\|_{V'}\Big)^{1/(p-1)} \overset{k\to\infty}{\longrightarrow} 0.$$

This proves the assertion. ∎

3. Optimization in sequence spaces. So far, we considered a minimization problem in an *arbitrary* reflexive Banach space V. In many cases of interest, V is equipped with a Riesz basis Ψ; i.e., the norm in V of an expansion in Ψ is equivalent to a discrete norm of the expansion coefficients (see also (3.1) below). One may think of V being a function space and the basis being the Fourier basis. However, our considerations have been motivated by recent results in wavelet theory [9, 16]. In particular, we were driven by the results in [10, 11], where an adaptive wavelet method for solving certain operator equations was proven to be asymptotically optimally convergent. One main ingredient for defining the adaptive approximation and for analyzing it is the transformation of the operator equation into an equivalent discrete problem still on an infinite-dimensional space. We will mimic this approach for our minimization problem.

To this end, let us now consider a basis $\Psi := \{\psi_\lambda : \lambda \in \mathcal{J}\}$ in V and, for any $v = \sum_{\lambda\in\mathcal{J}} v_\lambda\psi_\lambda \in V$, let us denote by $\boldsymbol{v} = (v_\lambda)_{\lambda\in\mathcal{J}}$ the (possibly infinite) sequence of its coefficients; we will use the notation $v = \boldsymbol{v}^T\Psi$. We assume that Ψ is a Riesz basis in V, in the sense that there exists $1 < p < \infty$ and constants $0 < c_\Psi \le C_\Psi < \infty$ such that

$$c_\Psi\,\|\boldsymbol{v}\|_{\ell_p} \le \|v\|_V \le C_\Psi\,\|\boldsymbol{v}\|_{\ell_p} \quad \text{for all } v = \boldsymbol{v}^T\Psi = \sum_{\lambda\in\mathcal{J}} v_\lambda\,\psi_\lambda \in V. \tag{3.1}$$

Let us also denote by $\langle\cdot,\cdot\rangle$ the duality pairing between $\ell_p = \ell_p(\mathcal{J})$ and $\ell_{p'} = \ell_{p'}(\mathcal{J})$, where $\frac{1}{p} + \frac{1}{p'} = 1$. The isomorphism $v \mapsto \boldsymbol{v}$ induces an isomorphism $F \mapsto \boldsymbol{F}$ between V' and $\ell_{p'}$, by setting $\langle \boldsymbol{F}, \boldsymbol{v}\rangle = \langle F, v\rangle$ for all $\boldsymbol{v} \in \ell_{p'}$. The previous norm equivalences (3.1) yield

$$C_\Psi^{-1} \|\boldsymbol{F}\|_{\ell_{p'}} \le \|F\|_{V'} \le c_\Psi^{-1} \|\boldsymbol{F}\|_{\ell_{p'}} \quad \text{for all } F \in V'. \tag{3.2}$$

Thus, we can transfer the minimization in V to a minimization in the sequence space ℓ_p. To this end, let us introduce the functional $\mathbb{J} : \ell_p \to \mathbb{R}$ defined as

$$\mathbb{J}(\boldsymbol{v}) := J(\boldsymbol{v}^T\Psi). \tag{3.3}$$

We use $\mathbb{J}$ (and not $\boldsymbol{J}$) here to indicate that $\mathbb{J}(\boldsymbol{v}) \in \mathbb{R}$ is a number, whereas boldface characters always stand for sequences. Then, the minimization problem (2.1) can equivalently be formulated as follows: Find $\boldsymbol{u} \in \ell_p$ such that

$$\mathbb{J}(\boldsymbol{u}) = \min_{\boldsymbol{v}\in\ell_p} \mathbb{J}(\boldsymbol{v}). \tag{3.4}$$

Note that the solutions $u \in V$ of (2.1) and $\boldsymbol{u} \in \ell_p$ of (3.4) are related by $u = \boldsymbol{u}^T\Psi$.

The operator $\mathbb{J}$ inherits all the properties of the operator J. In particular, it satisfies the conditions in Assumption 2.1. Its derivative is given by

$$\boldsymbol{J}'(\boldsymbol{v}) = \langle J'(\boldsymbol{v}^T\Psi), \Psi\rangle,$$

where, for any $f \in V'$, we use the notation $\langle f, \Psi\rangle := (\langle f, \psi_\lambda\rangle)_{\lambda\in\mathcal{J}}$. In fact,

$$\begin{aligned}\langle \boldsymbol{J}'(\boldsymbol{v}), \boldsymbol{w}\rangle &= \lim_{t\to 0+} \frac{\mathbb{J}(\boldsymbol{v}+t\boldsymbol{w}) - \mathbb{J}(\boldsymbol{v})}{t} \;=\; \lim_{t\to 0+} \frac{J(\boldsymbol{v}^T\Psi + t\boldsymbol{w}^T\Psi) - J(\boldsymbol{v}^T\Psi)}{t} \\ &= \langle J'(\boldsymbol{v}^T\Psi), \boldsymbol{w}^T\Psi\rangle \;=\; \boldsymbol{w}^T \langle J'(\boldsymbol{v}^T\Psi), \Psi\rangle \;=\; \langle\langle J'(\boldsymbol{v}^T\Psi), \Psi\rangle, \boldsymbol{w}\rangle.\end{aligned}$$

Let us now define the operator $\boldsymbol{R} : \ell_{p'} \setminus \{0\} \to \ell_p$, $\frac{1}{p} + \frac{1}{p'} = 1$, by

$$(\boldsymbol{R}(\boldsymbol{f}))_\lambda := \frac{f_\lambda\, |f_\lambda|^{p'-2}}{\|\boldsymbol{f}\|_{\ell_{p'}}^{p'-1}}.$$

We have

$$\langle \boldsymbol{f}, \boldsymbol{R}(\boldsymbol{f})\rangle = \|\boldsymbol{f}\|_{\ell_{p'}}^{1-p'} \sum_\lambda f_\lambda^2\, |f_\lambda|^{p'-2} = \|\boldsymbol{f}\|_{\ell_{p'}};$$

since $p = \frac{p'}{p'-1}$, we also have

$$\|\boldsymbol{R}(\boldsymbol{f})\|_{\ell_p} = \|\boldsymbol{f}\|_{\ell_{p'}}^{1-p'} \Big(\sum_\lambda |f_\lambda|^{p(p'-1)}\Big)^{1/p} = \|\boldsymbol{f}\|_{\ell_{p'}}^{1-p'} \Big(\sum_\lambda |f_\lambda|^{p'}\Big)^{(p'-1)/p'} = 1.$$

Note that $\boldsymbol{R}$ coincides with the normalized Riesz operator in the Hilbert case $p = p' = 2$; otherwise the operator is nonlinear.

Now, we can formulate the discrete counterpart of Algorithm 2.6, which reads as follows.

ALGORITHM 3.1 (fully infinite-dimensional steepest descent method).
Let $\boldsymbol{u}^{(0)} \in V$ be given.
Then, for $k = 0, 1, 2, \ldots$, while $\boldsymbol{J}'(\boldsymbol{u}^{(k)}) \neq \boldsymbol{0}$, do

1. *determine the search direction $\boldsymbol{s}^{(k)}$ by $\boldsymbol{s}^{(k)} := -\boldsymbol{R}(\boldsymbol{J}'(\boldsymbol{u}^{(k)}))$;*
2. *determine an admissible stepsize $\mu^{(k)} \in \mathcal{A}(\mathbb{J}; \boldsymbol{u}^{(k)}, \boldsymbol{s}^{(k)})$;*
3. *update: $\boldsymbol{u}^{(k+1)} := \boldsymbol{u}^{(k)} + \mu^{(k)}\boldsymbol{s}^{(k)}$.*

The convergence of the latter algorithm to $\boldsymbol{u}$ follows by Proposition 2.7.

4. A general adaptive algorithm. The above fully infinite-dimensional algorithm is posed in general terms and is in general not computable. The next step is to replace all noncomputable operations by finite approximations. We start by identifying certain routines that are needed within the algorithm. To this end, for any sequence $\boldsymbol{v} \in \ell(\mathcal{J})$, we define supp $\boldsymbol{v}$ as the set of indices corresponding to the nonzero entries, i.e.,

$$\operatorname{supp} \boldsymbol{v} := \{\lambda \in \mathcal{J} : v_\lambda \neq 0\}, \tag{4.1}$$

and we call a vector *compactly (or finitely) supported* if $\#\operatorname{supp} \boldsymbol{v} < \infty$. We will often add the index Λ and use the notation $\boldsymbol{v}_\Lambda$ in order to indicate a compactly supported vector. Note that Λ sometimes (but not necessarily) coincides with supp $\boldsymbol{v}_\Lambda$. When this is the case, we will clearly state it.

There are several issues that we need to treat in order to define a computable version of the descent algorithm. First of all, even if $\boldsymbol{v}_\Lambda$ is a compactly supported vector, in general the gradient $\boldsymbol{J}'(\boldsymbol{v}_\Lambda)$ has infinitely many components, so it cannot be computed exactly. Next, once an approximate descent direction has been found, an admissible stepsize has to be computed. This may require several evaluations of the functional $\mathbb{J}$. Finally, in order to achieve an algorithm of optimal complexity, the negligible components of the current approximation of the minimizer should be removed (possibly not at each iteration), by using a coarsening procedure.

We start by assuming the availability of three basic procedures for approximating the functional and its gradient as well as for thresholding a given vector.

ASSUMPTION 4.1. *A procedure* **EVAL-GRAD:** $[\boldsymbol{v}_\Lambda, \varepsilon] \mapsto \boldsymbol{w}_\Lambda$ *is available:*
Given a compactly supported vector $\boldsymbol{v}_\Lambda$ *and a tolerance* $\varepsilon > 0$*, a compactly supported vector* $\boldsymbol{w}_\Lambda$ *is computed, such that* $\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{w}_\Lambda\|_{\ell_{p'}} \leq \varepsilon$.

ASSUMPTION 4.2. *A procedure* **EVAL-J:** $[\boldsymbol{v}_\Lambda, \varepsilon] \mapsto g$ *is available:*
Given a compactly supported vector $\boldsymbol{v}_\Lambda$ *and a tolerance* $\varepsilon > 0$*, a real number* g *is computed, such that* $|\mathbb{J}(\boldsymbol{v}_\Lambda) - g| \leq \varepsilon$.

ASSUMPTION 4.3. *A procedure* **THRESH:** $[\boldsymbol{v}_\Lambda, \varepsilon] \mapsto \boldsymbol{z}_\Lambda$ *is available:*
Given a compactly supported vector $\boldsymbol{v}_\Lambda$ *and a tolerance* $\varepsilon > 0$*, a compactly supported vector* $\boldsymbol{z}_\Lambda$ *is computed, such that* $\|\boldsymbol{v}_\Lambda - \boldsymbol{z}_\Lambda\|_{\ell_p} \leq \varepsilon$ *and* supp $\boldsymbol{z}_\Lambda \subseteq$ supp $\boldsymbol{v}_\Lambda$ *has minimal cardinality, possibly subject to certain constraints on the distribution of its entries.*

Now we indicate how to use these routines in order to realize the main ingredients of the adaptive algorithm. We use **EVAL-GRAD** to create a procedure, called **APPROX-GRAD**, which yields an admissible descent direction. Next, we combine **EVAL-GRAD** and **EVAL-J** to construct a procedure, called **LINE-SEARCH**, which defines an admissible stepsize. In turn, the routines **APPROX-GRAD** and **LINE-SEARCH** are combined to form **DESCENT** which realizes one descent step of the algorithm. On the other hand, by **THRESH** and **APPROX-GRAD** we construct a procedure **COARSE** in order to coarsen the current iterate. Finally, **DESCENT** and **COARSE** are used to define our general adaptive algorithm **MINIMIZE**. For the sake of clarity, we show the hierarchy of these routines in Figure 4.1.

4.1. Approximation of the gradient. We first note that **EVAL-GRAD** may not yield an admissible descent direction. In fact, it may very well happen that the approximate evaluation of the gradient $\boldsymbol{J}'(\boldsymbol{v}_\Lambda)$ gives $\boldsymbol{0}$ even though $\boldsymbol{v}_\Lambda$ is *not* the minimum of $\mathbb{J}$. Indeed, the approximation may discard many very small entries of $\boldsymbol{J}'(\boldsymbol{v}_\Lambda)$ so that $\boldsymbol{v}_\Lambda$ may even be far away from the minimum. Hence, we have to ensure theoretically that $\boldsymbol{G}(\boldsymbol{v}_\Lambda)$ vanishes if and only if the exact gradient vanishes.

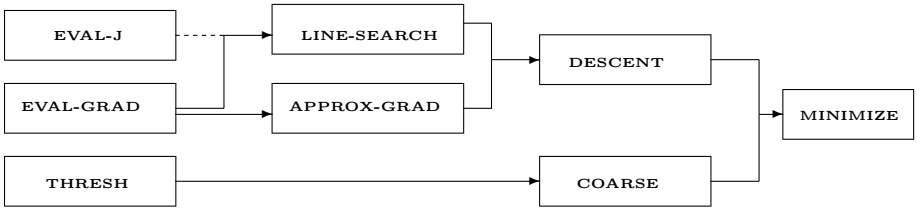


FIG. 4.1. *Hierarchy of procedures.*

This is accomplished by the procedure **APPROX-GRAD** shown below. Besides the input vector $\boldsymbol{v}_\Lambda$ and the accuracy $\varepsilon > 0$, **APPROX-GRAD** takes an extra input parameter, namely, $0 < \gamma < 1$, whose meaning will be discussed later on. Finally, $\nu \in (0,1)$ denotes any arbitrarily fixed constant.

APPROX-GRAD: $[\boldsymbol{v}_\Lambda, \varepsilon, \gamma] \mapsto [\boldsymbol{G}(\boldsymbol{v}_\Lambda), \eta]$
1. set $\eta^{(1)} := \varepsilon$;
2. for $n = 1, 2, \ldots$ do
   (a) $\boldsymbol{w}_\Lambda^{(n)} :=$ **EVAL-GRAD**$[\boldsymbol{v}_\Lambda, \eta^{(n)}]$;
   (b) if $\|\boldsymbol{w}_\Lambda^{(n)}\|_{\ell_2} \geq \frac{1+\gamma}{1-\gamma}\eta^{(n)}$, set $\boldsymbol{G}(\boldsymbol{v}_\Lambda) := \boldsymbol{w}^{(n)}$, $\eta := \eta^{(n)}$, RETURN;
   else set $\eta^{(n+1)} := \nu\,\eta^{(n)}$;
3. $\boldsymbol{G}(\boldsymbol{v}_\Lambda) := \boldsymbol{0}$, $\eta := 0$.

Note that statement 3 is reached only after an infinite loop in statement 2. Of course, in the computable version of the algorithm, we will insert a stopping criterion in statement 2; see section 5.6 below.

The following statement is an immediate consequence of the definition.

PROPOSITION 4.4. *The procedure* **APPROX-GRAD:** $[\boldsymbol{v}_\Lambda, \varepsilon, \gamma] \mapsto [\boldsymbol{G}(\boldsymbol{v}_\Lambda), \eta]$ *has the following property: Given* $0 < \gamma < 1$, $\varepsilon > 0$, *and a finitely supported vector* $\boldsymbol{v}_\Lambda$, *a vector* $\boldsymbol{G}(\boldsymbol{v}_\Lambda)$ *with finite support and a number* $\eta = \eta(\boldsymbol{v}_\Lambda) \in [0, \varepsilon]$ *are computed, such that*

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \leq \eta, \tag{4.2}$$

$$\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \geq \frac{1+\gamma}{1-\gamma}\,\eta. \tag{4.3}$$

□

Remark 4.5. Even though ε does not appear in (4.2) and (4.3), it serves as upper bound for the tolerance η determined inside the routine **APPROX-GRAD**.

The inequalities (4.2) and (4.3) show that the routine **APPROX-GRAD** defines an admissible descent direction whenever $\boldsymbol{J}'(\boldsymbol{v}_\Lambda) \neq \boldsymbol{0}$. Precisely, the following results, which will play a crucial role in the subsequent analysis, hold.

PROPOSITION 4.6. $\boldsymbol{G}(\boldsymbol{v}_\Lambda)$ *satisfies the following properties:*

(a) *The inequality*

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} - \eta \leq \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \leq \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} + \eta \tag{4.4}$$

holds.

(b) $\boldsymbol{J}'(\boldsymbol{v}_\Lambda) = \boldsymbol{0}$ *if and only if* $\boldsymbol{G}(\boldsymbol{v}_\Lambda) = \boldsymbol{0}$.

(c) *If* $\boldsymbol{J}'(\boldsymbol{v}_\Lambda) \neq \boldsymbol{0}$*, then for* $\boldsymbol{s}_\Lambda := -\boldsymbol{R}(\boldsymbol{G}(\boldsymbol{v}_\Lambda))$*, one has*

$$\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda \rangle \leq -\gamma \, \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}; \tag{4.5}$$

i.e., $\boldsymbol{s}_\Lambda$ *is an admissible descent direction.*

(d) *Finally, we have*

$$\gamma \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \leq |\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda \rangle| \leq \frac{2}{1+\gamma} \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}. \tag{4.6}$$

Proof. The first assertion is an immediate consequence of the triangle inequality. As for (b), if $\boldsymbol{J}'(\boldsymbol{v}_\Lambda) = \boldsymbol{0}$, then

$$\frac{1+\gamma}{1-\gamma}\,\eta \leq \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \leq \eta,$$

which is possible only if $\eta = 0$ and consequently $\boldsymbol{G}(\boldsymbol{v}_\Lambda) = \boldsymbol{0}$. The converse is trivial.

If $\boldsymbol{J}'(\boldsymbol{v}_\Lambda) \neq \boldsymbol{0}$, then

$$\begin{aligned}
\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda \rangle &= \langle \boldsymbol{G}(\boldsymbol{v}_\Lambda) - \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{R}(\boldsymbol{G}(\boldsymbol{v}_\Lambda)) \rangle - \langle \boldsymbol{G}(\boldsymbol{v}_\Lambda), \boldsymbol{R}(\boldsymbol{G}(\boldsymbol{v}_\Lambda)) \rangle \\
&\leq \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} - \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \\
&\leq \eta - (1+\gamma)\eta - \gamma \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \\
&= -\gamma \, (\eta + \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}) \;\leq\; -\gamma \, \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}
\end{aligned}$$

by (4.4) and Proposition 4.4, which proves (c). Note that with this choice the angle between the gradient and the descent direction is bounded away from 90°.

The first inequality in (d) follows directly from the proof of (c), taking into account that $\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda \rangle$ is negative. In fact, as above, we have

$$|\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda \rangle| \geq \gamma \, (\eta + \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}) \geq \gamma \, \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}.$$

As for the second inequality, we have

$$\begin{aligned}
|\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda \rangle| &\leq |\langle \boldsymbol{G}(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda \rangle| + |\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{G}(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda \rangle| \\
&= \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} + \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}.
\end{aligned}$$

By (4.2) and (4.3), we get

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \leq \eta \leq \frac{1-\gamma}{1+\gamma} \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}},$$

from which (d) follows. ∎

Note that **APPROX-GRAD** gives the desired result after a finite number of steps if and only if $\boldsymbol{J}'(\boldsymbol{v}_\Lambda) \neq \boldsymbol{0}$.

Remark 4.7. Proposition 4.4 shows that **APPROX-GRAD** gives an approximation of the gradient up to a *relative* accuracy:

$$\frac{\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}}{\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}} \leq \frac{1-\gamma}{2\gamma}.$$

In fact, using (4.4) as well as (4.3) gives

$$\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \leq \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} + \eta \leq \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} + \frac{1-\gamma}{1+\gamma} \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}},$$

which implies

$$\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \le \frac{1+\gamma}{2\gamma}\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}. \tag{4.7}$$

Then, we conclude by (4.2) that

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \le \eta \le \frac{1-\gamma}{1+\gamma}\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}} \le \frac{1-\gamma}{2\gamma}\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}.$$

4.2. Approximate descent step. Let $\boldsymbol{v}_\Lambda$ be a given compactly supported vector, and let $\boldsymbol{G}(\boldsymbol{v}_\Lambda)$ be an approximation of the gradient $\boldsymbol{J}'(\boldsymbol{v}_\Lambda)$ produced by **APPROX-GRAD**; let us set $\boldsymbol{s}_\Lambda = -\boldsymbol{R}(\boldsymbol{G}(\boldsymbol{v}_\Lambda))$. Hereafter, we describe how to select an admissible step size. The proposed algorithm is based on a bisection procedure, which yields the admissible step $\mu \in \mathcal{A}(\mathbb{J}; \boldsymbol{v}_\Lambda, \boldsymbol{s}_\Lambda)$ after a finite number of bisections, and requires only the approximate evaluation of the functional $\mathbb{J}$.

Setting as in (2.5) $\varphi(\mu) := \mathbb{J}(\boldsymbol{v}_\Lambda + \mu \boldsymbol{s}_\Lambda)$, we look for μ satisfying (2.6) and the similar condition

$$\varphi(\mu) \ge \varphi(0) + \beta\,\varphi'(0)\,\mu;$$

indeed, the latter condition implies (2.7), thanks to Lagrange's theorem and the monotonicity of φ'. Thus, setting

$$Q(\mu) := \frac{\varphi(\mu) - \varphi(0)}{\mu\,\varphi'(0)} = \frac{\varphi(0) - \varphi(\mu)}{\mu\,|\varphi'(0)|},$$

we look for μ satisfying

$$\alpha \le Q(\mu) \le \beta. \tag{4.8}$$

This is possible, since $\lim_{\mu\to 0+} Q(\mu) = 1$ and Q is monotonically decreasing to $-\infty$ for $\mu \to +\infty$. On the other hand, only a computable approximation of $Q(\mu)$ can be used in the search. So, we approximate $Q(\mu)$ by

$$\tilde{Q}(\mu) := \frac{\tilde{\varphi}(0) - \tilde{\varphi}(\mu)}{\mu\,\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}},$$

where $\tilde{\varphi}(0) = \textbf{EVAL-J}[\boldsymbol{v}_\Lambda, \varepsilon]$, $\tilde{\varphi}(\mu) = \textbf{EVAL-J}[\boldsymbol{v}_\Lambda + \mu\boldsymbol{s}_\Lambda, \varepsilon]$, and $\varepsilon := \varrho\,\mu\,\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_{p'}}$, $\varrho > 0$ being a constant to be determined later on. It is easily seen that

$$\frac{2}{1+\gamma}\tilde{Q}(\mu) - \frac{4\varrho}{1+\gamma} \le Q(\mu) \le \frac{1}{\gamma}\tilde{Q}(\mu) + \frac{2\varrho}{\gamma}.$$

Thus, (4.8) holds if $\tilde{Q}(\mu)$ satisfies

$$\tilde{\alpha} \le \tilde{Q}(\mu) \le \tilde{\beta} \tag{4.9}$$

with

$$\tilde{\alpha} := \frac{1+\gamma}{2}\alpha + 2\varrho \qquad \text{and} \qquad \tilde{\beta} := \gamma\beta - 2\varrho.$$

Note that these bounds are meaningful provided that α, β, and γ are chosen such that

$$\beta > \frac{1+\gamma}{2\gamma}\alpha \tag{4.10}$$

holds. In fact, then we have $\gamma\beta - \frac{1+\gamma}{2}\alpha > 0$ and we can choose $\varrho > 0$ such that

$$\varrho < \frac{1}{4}\left(\gamma\beta - \frac{1+\gamma}{2}\alpha\right) \tag{4.11}$$

and thus $\tilde{\alpha} = \frac{1+\gamma}{2}\alpha + 2\varrho < \gamma\beta - 2\varrho = \tilde{\beta}$.

Starting from any tentative stepsize and possibly halving or doubling the current stepsize a finite number of times, either we satisfy (4.9), or we find two values $\mu_0^- < \mu_0^+$ such that $\tilde{Q}(\mu_0^-) > \tilde{\beta}$ and $\tilde{Q}(\mu_0^+) < \tilde{\alpha}$. In this case, we can start the classical bisection procedure applied to $\tilde{Q}$. Let us prove that this procedure guarantees that (4.9) is satisfied after a finite number of steps. We argue by contradiction, assuming that the procedure generates an infinite sequence of values $\mu_n^- < \mu_n^+$ such that $\tilde{Q}(\mu_n^+) < \tilde{\alpha}$, $\tilde{Q}(\mu_n^-) > \tilde{\beta}$, $\lim_{n\to\infty}(\mu_n^+ - \mu_n^-) = 0$. Then, there exists $\bar{\mu}$ such that $\lim_{n\to\infty}\mu_n^+ = \lim_{n\to\infty}\mu_n^- = \bar{\mu}$. By continuity, we have $\lim_{n\to\infty}(Q(\mu_n^-) - Q(\mu_n^+)) = 0$. On the other hand,

$$Q(\mu_n^-) \geq \frac{2}{1+\gamma}\,\tilde{Q}(\mu_n^-) - \frac{4\varrho}{1+\gamma} > \frac{2\tilde{\beta} - 4\varrho}{1+\gamma}$$

and

$$Q(\mu_n^+) \leq \frac{1}{\gamma}\,\tilde{Q}(\mu_n^+) + \frac{2\varrho}{\gamma} < \frac{2\varrho + \tilde{\alpha}}{\gamma}$$

so that

$$Q(\mu_n^-) - Q(\mu_n^+) > \omega - \nu\,\varrho$$

with

$$\omega = \omega(\alpha,\beta,\gamma) := \frac{2\gamma}{1+\gamma}\,\beta - \frac{1+\gamma}{2\gamma}\,\alpha, \qquad \text{and} \qquad \nu = \nu(\gamma) := \frac{\gamma}{1+\gamma} + \frac{4}{\gamma}.$$

Choosing α, β, and γ such that in addition to (4.10) we also have

$$\beta > \frac{(1+\gamma)^2\alpha + 4\varrho(1+3\gamma)}{4\gamma^2}, \tag{4.12}$$

we obtain $\omega - \varrho\nu > 0$, i.e., a contradiction. In fact, a straightforward calculation shows that (4.12) yields

$$\varrho < \frac{4\gamma^2\beta - (1+\gamma)^2\alpha}{4(1+3\gamma)} = \frac{\omega}{\nu}.$$

We summarize the result as follows.

PROPOSITION 4.8. *Given any compactly supported vector $\boldsymbol{v}_\Lambda$, let us set $\boldsymbol{s}_\Lambda = -\boldsymbol{R}(\boldsymbol{G}(\boldsymbol{v}_\Lambda))$, where $\boldsymbol{G}(\boldsymbol{v}_\Lambda)$ is an approximation of $\boldsymbol{J}'(\boldsymbol{v}_\Lambda)$ satisfying conditions* (4.2) *and* (4.3)*. In addition, let Assumption* 4.2 *be satisfied. Then, choosing α, β, γ according to* (4.10) *and* (4.12)*, we can compute an admissible stepsize $\mu \in \mathcal{A}(\mathbb{J}; \boldsymbol{v}_\Lambda, \boldsymbol{s}_\Lambda)$.* ∎

A different algorithm for choosing the stepsize will be described under more restrictive assumptions in section 5.1 below. Both algorithms are particular realizations of a general procedure **LINE-SEARCH** which we define as follows.

DEFINITION 4.9. *We call* **LINE-SEARCH:** $[\boldsymbol{v}_\Lambda, \boldsymbol{G}(\boldsymbol{v}_\Lambda), \alpha, \beta] \mapsto \eta$ *any procedure with the following property: Given* α, β *satisfying* $0 < \alpha < \beta < 1$*, a finitely supported vector* $\boldsymbol{v}_\Lambda$*, and an approximation* $\boldsymbol{G}(\boldsymbol{v}_\Lambda)$ *of its gradient produced by* **APPROX-GRAD***, and setting* $\boldsymbol{s}_\Lambda = -\boldsymbol{R}(\boldsymbol{G}(\boldsymbol{v}_\Lambda))$*, then an admissible stepsize* $\eta \in \mathcal{A}(\mathbb{J}; \boldsymbol{v}_\Lambda, \boldsymbol{s}_\Lambda)$ *is computed.*

By the routines **APPROX-GRAD** and **LINE-SEARCH**, we perform one step of descent, which is detailed in the following routine.

DESCENT: $[\boldsymbol{v}_\Lambda, \varepsilon] \mapsto \boldsymbol{w}_\Lambda$
1. $[\boldsymbol{G}(\boldsymbol{v}_\Lambda), \eta] := \textbf{APPROX-GRAD}[\boldsymbol{v}_\Lambda, \varepsilon, \gamma]$
2. if $\eta = 0$, STOP ($\boldsymbol{v}_\Lambda = \boldsymbol{u}$); else
3. $\boldsymbol{s}_\Lambda := -\boldsymbol{R}(\boldsymbol{G}(\boldsymbol{v}_\Lambda))$;
4. $\mu := \textbf{LINE-SEARCH}[\boldsymbol{v}_\Lambda, \boldsymbol{G}(\boldsymbol{v}_\Lambda), \alpha, \beta]$;
5. $\boldsymbol{w}_\Lambda := \boldsymbol{v}_\Lambda + \mu \boldsymbol{s}_\Lambda$.

4.3. Coarsening. The problem of coarsening a vector is one of the central issues of nonlinear approximation theory (see [9, 22]). Several routines of this type are available in the literature; see, e.g., [10, 11, 12]. However, we are interested in a procedure which guarantees the possibility of significantly reducing the support of $\boldsymbol{v}_\Lambda$ while preserving the value of the functional $\mathbb{J}$ up to a small perturbation. The particular realization of the coarsening that we propose is based on the following result.

LEMMA 4.10. *Let* $\boldsymbol{v}_\Lambda$*,* $\boldsymbol{w}_\Lambda$ *be finitely supported vectors such that* $\operatorname{supp} \boldsymbol{w}_\Lambda \subseteq \operatorname{supp} \boldsymbol{v}_\Lambda =: \Lambda$*. Furthermore, let* $[\boldsymbol{G}(\boldsymbol{w}_\Lambda, \eta] = \textbf{APPROX-GRAD}[\boldsymbol{w}_\Lambda, \varepsilon, \gamma]$ *for some* $\varepsilon > 0$ *and* $0 < \gamma < 1$*. Then,*

$$\mathbb{J}(\boldsymbol{w}_\Lambda) - \mathbb{J}(\boldsymbol{v}_\Lambda) \le \Big(\|\boldsymbol{G}(\boldsymbol{w}_\Lambda)|_\Lambda\|_{\ell_{p'}} + \eta \Big) \|\boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\|_{\ell_p}.$$

Proof. By the convexity of $\mathbb{J}$, we obtain

$$\mathbb{J}(\boldsymbol{w}_\Lambda) - \mathbb{J}(\boldsymbol{v}_\Lambda) \le \langle \boldsymbol{J}'(\boldsymbol{w}_\Lambda), \boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda \rangle \le \|\boldsymbol{J}'(\boldsymbol{w})|_\Lambda\|_{\ell_{p'}} \, \|\boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\|_{\ell_p}. \tag{4.13}$$

Next, by (4.2), we have

$$\|\boldsymbol{J}'(\boldsymbol{w})|_\Lambda - \boldsymbol{G}(\boldsymbol{w})|_\Lambda\|_{\ell_{p'}} \le \|\boldsymbol{J}'(\boldsymbol{w}) - \boldsymbol{G}(\boldsymbol{w})\|_{\ell_{p'}} \le \eta,$$

whence the result immediately follows. ∎

We are ready to introduce the routine **COARSE**. It depends on a parameter $\vartheta > 0$, which bounds from above the error in the functional, as clarified in Proposition 4.11 below.

COARSE: $[\boldsymbol{v}_\Lambda, \vartheta] \to \boldsymbol{w}_\Lambda$
1. $[\boldsymbol{G}(\boldsymbol{v}_\Lambda), \eta^{(0)}] = \textbf{APPROX-GRAD}[\boldsymbol{v}_\Lambda, \vartheta, \gamma]$;
2. $A^{(0)} = \|\boldsymbol{G}(\boldsymbol{v})|_\Lambda\|_{\ell_{p'}} + \eta^{(0)}$.
3. For $k = 0, \dots$ do
   (a) $\boldsymbol{w}_\Lambda^{(k)} = \textbf{THRESH}[\boldsymbol{v}_\Lambda, \frac{\vartheta}{A^{(k)}}]$;
   (b) $[\boldsymbol{G}(\boldsymbol{w}_\Lambda^{(k)}), \eta^{(k)}] = \textbf{APPROX-GRAD}[\boldsymbol{w}_\Lambda^{(k)}, \vartheta, \gamma]$;
   (c) $B^{(k)} = \|\boldsymbol{G}(\boldsymbol{w}_\Lambda^{(k)})|_\Lambda\|_{\ell_{p'}} + \eta^{(k)}$;
   (d) if $B^{(k)} \le A^{(k)}$, set $\boldsymbol{w}_\Lambda := \boldsymbol{w}_\Lambda^{(k)}$, RETURN;
   (e) else $A^{(k+1)} = \max(B^{(k)}, 2A^{(k)})$.

PROPOSITION 4.11. *Given a finitely supported vector* $\boldsymbol{v}_\Lambda$ *and a number* $\vartheta > 0$*, the procedure* **COARSE:** $[\boldsymbol{v}_\Lambda, \vartheta] \mapsto \boldsymbol{w}_\Lambda$ *produces a finitely supported vector* $\boldsymbol{w}_\Lambda$ *obtained by thresholding* $\boldsymbol{v}_\Lambda$ *such that* $\mathbb{J}(\boldsymbol{w}_\Lambda) < \mathbb{J}(\boldsymbol{v}_\Lambda) + \vartheta$*.*

Proof. Thanks to Lemma 4.10 and the fact that the sequence $A^{(k)}$ is geometrically increasing, one sees immediately that this procedure terminates in a finite number of iterations leading to the claimed inequality. □

An improved version of **COARSE** will be given in section 5.4 under more restrictive assumptions.

4.4. The general convergent adaptive algorithm. We are now ready to define a general adaptive algorithm. It depends on the choice of various parameters. A strategy for their selection will be detailed in the next section.

ALGORITHM 4.12. **MINIMIZE**
Let $\boldsymbol{u}^{(0)} = \boldsymbol{u}_\Lambda^{(0)} \in \ell_p$ *be given. Fix constants* α*,* β*,* γ *satisfying* $0 < \alpha < \beta < 1$ *and* $0 < \gamma < 1$*, which enter into the definition of* **DESCENT***.*

For $m = 0, 1, \ldots,$ *do*
  $\boldsymbol{v}_\Lambda^{(m,0)} := \boldsymbol{u}_\Lambda^{(m)}$
  choose an integer $K^{(m)} \geq 1$
  For $k = 0, 1, \ldots, K^{(m)} - 1$*, do*
    choose $\varepsilon^{(m,k)} > 0$
    $\boldsymbol{v}_\Lambda^{(m,k+1)} := \mathbf{DESCENT}[\boldsymbol{v}_\Lambda^{(m,k)}, \varepsilon^{(m,k)}]$
  End
  choose $\vartheta^{(m)} > 0$
  $\boldsymbol{u}_\Lambda^{(m+1)} := \mathbf{COARSE}[\boldsymbol{v}_\Lambda^{(m,K^{(m)})}, \vartheta^{(m)}]$
End

THEOREM 4.13. *Let the sequence* $(K^{(m)})_m$ *be arbitrary. Assume that the sequence* $(\varepsilon^{(m,k)})_{m,k}$ *satisfies*

$$\lim_{m\to\infty} \sup_k \varepsilon^{(m,k)} = 0, \tag{4.14}$$

whereas the sequence $(\vartheta^{(m)})_m$ *satisfies*

$$\vartheta^{(m)} \leq \frac{1}{2}[\mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}) - \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})})]. \tag{4.15}$$

Then, Algorithm 4.12 **MINIMIZE** *either yields* $\boldsymbol{u}$ *after a finite number of steps or produces an infinite sequence* $(\boldsymbol{u}_\Lambda^{(m)})_m$ *which converges to* $\boldsymbol{u}$*.*

Proof. If, for some m and k, the procedure **DESCENT** stops, then $\boldsymbol{v}_\Lambda^{(m,k)} = \boldsymbol{u}$. Otherwise, the algorithm produces an infinite sequence of vectors. Assuming the latter case, we first show that the sequence $(\mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}))_m$ is strictly decreasing. Let m be fixed. By definition of **DESCENT**, we have

$$\mathbb{J}(\boldsymbol{v}_\Lambda^{(m,k+1)}) < \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,k)}), \qquad k = 0, 1, \ldots, K^{(m)} - 1, \tag{4.16}$$

whence

$$\mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) < \mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}). \tag{4.17}$$

Furthermore, by definition of **COARSE** and assumption (4.15), we get

$$\mathbb{J}(\boldsymbol{u}_\Lambda^{(m+1)}) \leq \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) + \vartheta^{(m)} \leq \frac{1}{2}[\mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) + \mathbb{J}(\boldsymbol{u}_\Lambda^{(m)})], \tag{4.18}$$

from which we obtain the desired result

$$\mathbb{J}(\boldsymbol{u}_\Lambda^{(m+1)}) - \mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}) \le \frac{1}{2}[\mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) - \mathbb{J}(\boldsymbol{u}_\Lambda^{(m)})] < 0. \tag{4.19}$$

Next, since the functional $\mathbb{J}$ is bounded from below, we deduce that

$$\lim_{m\to\infty} [\mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}) - \mathbb{J}(\boldsymbol{u}_\Lambda^{(m+1)})] = 0. \tag{4.20}$$

Using (4.18) again, we easily get $\mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}) - \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) < 2[\mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}) - \mathbb{J}(\boldsymbol{u}_\Lambda^{(m+1)})]$. Hence, we also have that

$$\lim_{m\to\infty} \vartheta^{(m)} = \lim_{m\to\infty} \left[\mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}) - \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})})\right] = 0. \tag{4.21}$$

Thanks to (4.16), we obtain

$$\lim_{m\to\infty} \sup_k [\mathbb{J}(\boldsymbol{v}_\Lambda^{(m,k)}) - \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,k+1)})] = 0.$$

Exactly as in the proof of Proposition 2.7(i), the latter result implies

$$\lim_{m\to\infty} \sup_k \|\boldsymbol{v}_\Lambda^{(m,k+1)} - \boldsymbol{v}_\Lambda^{(m,k)}\|_{\ell_p} = 0. \tag{4.22}$$

We now observe that, by (4.16) and (4.19), the sequence $(\boldsymbol{v}_\Lambda^{(m,k)})_{m,k}$ is contained in the bounded set $\mathcal{R}(u^{(0)})$. Using the uniform continuity of $\boldsymbol{J}'$ on bounded sets, we obtain from (4.22)

$$\lim_{m\to\infty} \sup_k \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k+1)}) - \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} = 0. \tag{4.23}$$

Next, setting $[\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)}), \eta^{(m,k)}] := \textbf{APPROX-GRAD}[\boldsymbol{v}_\Lambda^{(m,k)}, \varepsilon^{(m,k)}, \gamma]$ and using (4.2) as well as step 3 of **DESCENT**, we have, for $0 \le k < K^{(m)}$,

$$\begin{aligned}
\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} - \eta^{(m,k)} \le \|\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} &= -\langle \boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)}), \boldsymbol{s}_\Lambda^{(m,k)}\rangle \\
&= \langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)}) - \boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)}), \boldsymbol{s}_\Lambda^{(m,k)}\rangle \\
&\quad + \langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k+1)}) - \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)}), \boldsymbol{s}_\Lambda^{(m,k)}\rangle \\
&\quad - \langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k+1)}), \boldsymbol{s}_\Lambda^{(m,k)}\rangle.
\end{aligned}$$

Now, (2.4) and (4.2) yield

$$\begin{aligned}
&\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} - \eta^{(m,k)} \le \\
&\quad \le \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)}) - \boldsymbol{G}(\boldsymbol{v}_\Lambda^{(k)})\|_{\ell_{p'}} + \|\boldsymbol{J}'(\boldsymbol{v}^{(m,k+1)}) - \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} \\
&\qquad - \beta\, \langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)}), \boldsymbol{s}_\Lambda^{(m,k)}\rangle \\
&\quad \le \eta^{(m,k)} + \|\boldsymbol{J}'(\boldsymbol{v}^{(m,k+1)}) - \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} + \beta\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}};
\end{aligned}$$

thus

$$(1-\beta)\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} \le \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)}) - \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k+1)})\|_{\ell_{p'}} + 2\eta^{(m,k)}.$$

By definition of **APPROX-GRAD** and assumption (4.14), we get $\eta^{(m,k)} \le \varepsilon^{(m,k)} \to 0$ as $m \to \infty$, uniformly in k. This and (4.23) imply

$$\lim_{m\to\infty} \sup_k \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} = 0.$$

Since $\|\boldsymbol{v}_\Lambda^{(m,k)} - \boldsymbol{u}\|_{\ell_p} \le \left(\frac{1}{c_{J,\Psi}}\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}}\right)^{1/(p-1)}$, we conclude that

$$\lim_{m\to\infty} \sup_k \|\boldsymbol{v}_\Lambda^{(m,k)} - \boldsymbol{u}\|_{\ell_p} = 0.$$

In particular, since $\boldsymbol{v}_\Lambda^{(m,0)} = \boldsymbol{u}_\Lambda^{(m)}$, this implies the claimed result. ∎

A few remarks are in order.

Remark 4.14. (i) Fulfilling assumption (4.15) may be accomplished as follows. By Wolfe's condition, the value of the functional is decreased when going from $\boldsymbol{v}_\Lambda^{(m,k)}$ to $\boldsymbol{v}_\Lambda^{(m,k+1)}$, provided the gradient is not yet zero causing the algorithm to stop. Precisely, setting $\boldsymbol{v}_\Lambda^{(m,k+1)} = \boldsymbol{v}_\Lambda^{(m,k)} + \mu^{(k)}\boldsymbol{s}^{(k)}$, we have by (2.3), (4.5), and (4.7)

$$\begin{aligned}\boldsymbol{J}(\boldsymbol{v}_\Lambda^{(m,k)}) - \boldsymbol{J}(\boldsymbol{v}_\Lambda^{(m,k+1)}) &\ge -\alpha\mu^{(k)}\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)}), \boldsymbol{s}_\Lambda^{(m,k)}\rangle \;\ge\; \alpha\gamma\,\mu^{(k)}\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}} \\ &\ge \alpha\frac{2\gamma^2}{1+\gamma}\mu^{(k)}\|\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}}.\end{aligned}$$

Then, we obtain by using a telescopic sum

$$\begin{aligned}\mathbb{J}(\boldsymbol{u}_\Lambda^{(m)}) - \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) = \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,0)}) - \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) &= \sum_{k=0}^{K^{(m)}-1} \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,k)}) - \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,k+1)}) \\ &\ge \frac{2\alpha\gamma^2}{1+\gamma}\sum_{k=0}^{K^{(m)}-1}\mu^{(k)}\|\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}};\end{aligned}$$

i.e., we can choose

$$\vartheta^{(m)} := \frac{\alpha\gamma^2}{1+\gamma}\sum_{k=0}^{K^{(m)}-1}\mu^{(k)}\|\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_{p'}},$$

which is in fact computable.

(ii) In view of (4.21), a natural way to satisfy assumption (4.14) is to set

$$\varepsilon^{(m,k)} := \begin{cases} 1 & \text{if } m = 0, \\ \vartheta^{(m-1)} & \text{if } m \ge 1, \end{cases} \qquad k = 0, 1, \dots, K^{(m)} - 1.$$

5. An adaptive algorithm with convergence rate. From now on, we focus our analysis on the case $p = 2$. Furthermore, we assume that the Fréchet derivative J' is Lipschitz continuous on each bounded subset of V. Precisely, recalling that the set $\mathcal{R}(u^{(0)})$ (see Lemma 2.2) is bounded, we assume the existence of a constant $L_J > 0$, possibly depending on $u^{(0)}$, such that

$$\|J'(v) - J'(w)\|_{V'} \le L_J\,\|v - w\|_V, \quad v, w \in \mathcal{R}(u^{(0)}). \tag{5.1}$$

Recalling the norm equivalences (3.1) and (3.2) and setting $L_{J,\Psi} = L_J C_\Psi^2$, we obtain the discrete form of the previous inequality, namely,

$$\|\boldsymbol{J}'(\boldsymbol{v}) - \boldsymbol{J}'(\boldsymbol{w})\|_{\ell_2} \le L_{J,\Psi}\, \|\boldsymbol{v} - \boldsymbol{w}\|_{\ell_2}, \quad \boldsymbol{v}, \boldsymbol{w} \in \mathcal{R}(u^{(0)}). \tag{5.2}$$

Recalling Assumption 2.1(ii) and setting $c_{J,\Psi} = c_J c_\Psi^2$, we also get

$$c_{J,\Psi}\|\boldsymbol{v} - \boldsymbol{w}\|_{\ell_2}^2 \le \langle \boldsymbol{J}'(\boldsymbol{v}) - \boldsymbol{J}'(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle \le L_{J,\Psi}\|\boldsymbol{v} - \boldsymbol{w}\|_{\ell_2}^2, \quad \boldsymbol{v}, \boldsymbol{w} \in \mathcal{R}(u^{(0)}). \tag{5.3}$$

In this section, we present a particular realization of the abstract adaptive algorithm **MINIMIZE**, which exploits the extra properties of the functional $\mathbb{J}$ stated above. The convergence result of the algorithm will be supplemented by a precise estimate of the rate of decay of the error. This will allow us to determine the number of iterations needed to reach a target tolerance, as well as to relate the error of the algorithm to the best approximation error of the solution.

We start by describing a more efficient line-search algorithm than the one prescribed in section 4.2; it requires neither the evaluation of the functional, nor additional approximate evaluations of the gradient other than the already computed descent direction.

5.1. Line search. Given a compactly supported vector $\boldsymbol{v}_\Lambda \in \mathcal{R}(u^{(0)})$ and an approximation $\boldsymbol{G}(\boldsymbol{v}_\Lambda)$ of $\boldsymbol{J}'(\boldsymbol{v}_\Lambda)$ which satisfies conditions (4.2) and (4.3), we set $\boldsymbol{s}_\Lambda = -\boldsymbol{R}(\boldsymbol{G}(\boldsymbol{v}_\Lambda))$ and we define a closed interval with computable endpoints, contained in the interval $\mathcal{A}(\mathbb{J}; \boldsymbol{v}_\Lambda, \boldsymbol{s}_\Lambda)$. To this end, let us set $\varphi(\mu) = \mathbb{J}(\boldsymbol{v}_\Lambda + \mu \boldsymbol{s}_\Lambda)$, so that $\varphi'(\mu) = \langle \mathbb{J}(\boldsymbol{v}_\Lambda + \mu \boldsymbol{s}_\Lambda), \boldsymbol{s}_\Lambda\rangle$, and let us recall that $\mu \in \mathcal{A}(\mathbb{J}; \boldsymbol{v}_\Lambda, \boldsymbol{s}_\Lambda)$ if and only if conditions (2.6) and (2.7) are satisfied. In this subsection, we actually assume that (5.2) and, consequently, (5.3) hold indeed in a bounded set larger than $\mathcal{R}(u^{(0)})$, namely, in the neighborhood of $\mathcal{R}(u^{(0)})$ of radius $\bar{\mu} := \mathrm{diam}(\mathcal{R}(u^{(0)}))$ (in the ℓ_2-distance); obviously, this is not restrictive at all, since it amounts to properly (re-)defining the constant L_J. Thus, if we choose $\boldsymbol{v} = \boldsymbol{v}_\Lambda$ and $\boldsymbol{w} = \boldsymbol{v}_\Lambda + \mu \boldsymbol{s}_\Lambda$ in (5.3), we easily obtain

$$c_{J,\Psi}\, \mu \le \varphi'(\mu) - \varphi'(0) \le L_{J,\Psi}\, \mu, \qquad \mu \in [0, \bar{\mu}]. \tag{5.4}$$

Using the right-hand side of (5.4), we get for some $\theta \in (0,1)$

$$\varphi(\mu) = \varphi(0) + \varphi'(\theta\, \mu)\mu \le \varphi(0) + \varphi'(0)\, \mu + \theta\, L_{J,\Psi}\, \mu^2.$$

Hence, (2.6) is fulfilled if μ satisfies

$$\varphi'(0)\, \mu + L_{J,\Psi}\, \mu^2 \le \alpha\, \varphi'(0)\, \mu, \qquad \text{i.e., } L_{J,\Psi}\, \mu \le (1-\alpha)\, |\varphi'(0)|.$$

Taking into account the left-hand side of (4.6), we get the sufficient condition for the validity of (2.6)

$$\mu \le \mu^* := \gamma\, \frac{1-\alpha}{L_{J,\Psi}}\, \|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_2},$$

provided $\mu^* \le \bar{\mu}$. This is indeed the case, since by (4.7) and (5.2) we have

$$\mu^* \le \frac{1-\alpha}{L_{J,\Psi}}\, \frac{1+\gamma}{2}\, \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2} \le (1-\alpha)\, \frac{1+\gamma}{2}\, \|\boldsymbol{v}_\Lambda - \boldsymbol{u})\|_{\ell_2} < \bar{\mu}.$$

On the other hand, using the inequality on the left-hand side, condition (2.7) is fulfilled if μ satisfies

$$\varphi'(0) + c_{J,\Psi}\,\mu \geq \beta\,\varphi'(0), \qquad \text{i.e., } c_{J,\Psi}\,\mu \geq (1-\beta)\,|\varphi'(0)|.$$

By the right-hand side of (4.6), we get the sufficient condition for the validity of (2.7)

$$\mu \geq \mu_* := \frac{2}{1+\gamma}\,\frac{1-\beta}{c_{J,\Psi}}\,\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_2}.$$

Obviously, we have to require that $\mu_* \leq \mu^*$, i.e,

$$(1-\beta) \leq \frac{\gamma(\gamma+1)}{2}\,\frac{c_{J,\Psi}}{L_{J,\Psi}}(1-\alpha), \tag{5.5}$$

which can always be satisfied, e.g., by fixing α and choosing β as close to 1 as needed.

Thus, we have obtained the following result.

PROPOSITION 5.1. *For any $\alpha \in (0,1)$ and any $\gamma \in (0,1)$, there exists β^* satisfying $\alpha \leq \beta^* < 1$ such that for all $\beta \in [\beta^*, 1)$, the interval*

$$\left[\frac{2}{1+\gamma}\,\frac{1-\beta}{c_{J,\Psi}}\,\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_2}, \gamma\,\frac{1-\alpha}{L_{J,\Psi}}\,\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_2}\right] \tag{5.6}$$

is nonempty and contained in $\mathcal{A}(\mathbb{J}; \boldsymbol{v}_\Lambda, \boldsymbol{s}_\Lambda)$. □

Consequently, the output of the procedure **LINE-SEARCH**$[\boldsymbol{v}_\Lambda, \alpha, \beta]$ can be defined by picking any value in this interval.

5.2. Error reduction in one descent step. Let $\boldsymbol{v}_\Lambda \in \mathcal{R}(u^{(0)})$ be any compactly supported approximation of the exact minimizer $\boldsymbol{u}$. We apply the routine **DESCENT** to it and, assuming $\boldsymbol{v}_\Lambda \neq \boldsymbol{u}$, we get a new approximation $\boldsymbol{w}_\Lambda$. Hereafter, we are interested in studying the behavior of the approximation error in going from $\boldsymbol{v}_\Lambda$ to $\boldsymbol{w}_\Lambda$. In the analysis, the following definition will be useful.

DEFINITION 5.2. *For any compactly supported vector $\boldsymbol{v}_\Lambda$, we set*

$$\mathbb{E}(\boldsymbol{v}_\Lambda) := \mathbb{J}(\boldsymbol{v}_\Lambda) - \mathbb{J}(\boldsymbol{u}).$$

The quantity $\mathbb{E}(\boldsymbol{v}_\Lambda)$ is an a priori error bound; indeed, by (2.2) and (3.1), we get

$$\|\boldsymbol{v}_\Lambda - \boldsymbol{u}\|_{\ell_2}^2 \leq \frac{2}{c_{J,\Psi}}\mathbb{E}(\boldsymbol{v}_\Lambda). \tag{5.7}$$

In order to compare $\mathbb{E}(\boldsymbol{w}_\Lambda)$ to $\mathbb{E}(\boldsymbol{v}_\Lambda)$, we recall the definition of $\boldsymbol{w}_\Lambda = \boldsymbol{v}_\Lambda + \mu\,\boldsymbol{s}_\Lambda$ given at point 5 of **DESCENT**; from condition (2.3) and inequality (4.5), we immediately get

$$\mathbb{E}(\boldsymbol{w}_\Lambda) \leq \mathbb{E}(\boldsymbol{v}_\Lambda) - \alpha\gamma\mu\,\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2}. \tag{5.8}$$

We now establish two technical results which will be used in what follows.

LEMMA 5.3. *Let $\mu =$ **LINE-SEARCH**$[\boldsymbol{v}_\Lambda, \alpha, \beta]$. Then the following inequality holds:*

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2} \leq \frac{L_{J,\Psi}}{\gamma(1-\beta)}\,\mu.$$

Proof. We start by the trivial identity

$$-\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda\rangle = \langle \boldsymbol{J}'(\boldsymbol{w}_\Lambda) - \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda\rangle - \langle \boldsymbol{J}'(\boldsymbol{w}_\Lambda), \boldsymbol{s}_\Lambda\rangle.$$

Using (5.3) and condition (2.4), we obtain

$$-\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda\rangle \le L_{J,\Psi}\|\boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\|_{\ell_2} - \beta\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{s}_\Lambda\rangle,$$

i.e.,

$$-\langle \boldsymbol{J}'(\boldsymbol{w}_\Lambda), \boldsymbol{s}_\Lambda\rangle \le \frac{L_{J,\Psi}}{1-\beta}\,\|\boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\|_{\ell_2}.$$

We conclude by (4.5). ∎

LEMMA 5.4. *The following inequality holds:*

$$\mathbb{E}(\boldsymbol{v}_\Lambda) \le \frac{1}{2c_{J,\Psi}}\,\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|^2_{\ell_2}.$$

Proof. We again apply (2.2) and (3.1) to get

$$\begin{aligned}\mathbb{J}(\boldsymbol{u}) - \mathbb{J}(\boldsymbol{v}_\Lambda) &\ge \langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{u} - \boldsymbol{v}_\Lambda\rangle + \frac{c_{J,\Psi}}{2}\|\boldsymbol{u} - \boldsymbol{v}_\Lambda\|^2_{\ell_2}\\ &\ge \min_{\boldsymbol{z}\in\ell_2}\left\{\langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{z}\rangle + \frac{c_{J,\Psi}}{2}\|\boldsymbol{z}\|^2_{\ell_2}\right\}\\ &= \min_{\boldsymbol{z}\in\ell_2}\sum_{\lambda\in\mathcal{J}}\left\{\boldsymbol{J}'(\boldsymbol{v}_\Lambda)_\lambda\, z_\lambda + \frac{c_{J,\Psi}}{2}z_\lambda^2\right\}.\end{aligned}$$

We note that we can minimize along each component λ independently for each $\lambda \in \mathcal{J}$; the minimum is attained at $z_\lambda = -\frac{1}{c_{J,\Psi}}\,\boldsymbol{J}'(\boldsymbol{v}_\Lambda)_\lambda$. We conclude that

$$\mathbb{J}(\boldsymbol{u}) - \mathbb{J}(\boldsymbol{v}_\Lambda) \ge -\frac{1}{2c_{J,\Psi}}\,\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|^2_{\ell_2},$$

which is precisely the thesis. ∎

We are now ready to estimate the error reduction guaranteed by one application of **DESCENT**.

PROPOSITION 5.5. *Given a compactly supported vector $\boldsymbol{v}_\Lambda \in \mathcal{R}(u^{(0)})$ and any tolerance ε, define $\boldsymbol{w}_\Lambda := \mathbf{DESCENT}(\boldsymbol{v}_\Lambda, \varepsilon)$. Let α, β be any fixed constants satisfying Wolfe's condition in Definition* 2.4. *Then, setting*

$$\sigma := 1 - 2\frac{c_{J,\Psi}}{L_{J,\Psi}}\,\gamma^2\,(1-\beta)\,\alpha < 1, \tag{5.9}$$

we have

$$\mathbb{E}(\boldsymbol{w}_\Lambda) \le \sigma\,\mathbb{E}(\boldsymbol{v}_\Lambda).$$

Proof. By the two previous lemmas, we obtain

$$\mu\,\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2} \ge \frac{\gamma\,(1-\beta)}{L_{J,\Psi}}\,\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|^2_{\ell_2} \ge 2\frac{c_{J,\Psi}}{L_{J,\Psi}}\,\gamma\,(1-\beta)\,\mathbb{E}(\boldsymbol{v}_\Lambda).$$

The result follows from (5.8). ∎

Remark 5.6. The expression (5.9) for the error reduction factor σ suggests a simple strategy for selecting the two parameters α and β which appear in (2.3) and (2.4). Indeed, we observe that σ is monotonically decreasing as $1-\beta$ increases. Recalling condition (5.5), we are lead to choose

$$1-\beta = \frac{\gamma(\gamma+1)}{2}\,\frac{c_{J,\Psi}}{L_{J,\Psi}}\,(1-\alpha). \tag{5.10}$$

In this case, the interval (5.6) reduces to a point, which is the chosen stepsize; i.e., we set

$$\mu := \gamma\,\frac{1-\alpha}{L_{J,\Psi}}\,\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\ell_2}.$$

Substituting (5.10) into (5.9), we see that σ is minimized with respect to α by the choice $\alpha = \frac{1}{2}$. With this value for α, we finally obtain the error reduction rate

$$\sigma_{\mathrm{opt}} := 1 - \left(\frac{c_{J,\Psi}}{L_{J,\Psi}}\right)^2 \frac{\gamma^3\,(1+\gamma)}{4}. \tag{5.11}$$

As a by-product of the previous analysis, we obtain the next result, which describes the interplay between the available error control quantities.

PROPOSITION 5.7. *The following chain of inequalities holds:*

$$\|\boldsymbol{v}_\Lambda - \boldsymbol{u}\|_{\ell_2}^2 \le \frac{2}{c_{J,\Psi}}\,\mathbb{E}(\boldsymbol{v}_\Lambda) \le \frac{2}{c_{J,\Psi}^2}\,\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2}^2 \le 2\left(\frac{L_{J,\Psi}}{c_{J,\Psi}}\right)^2 \|\boldsymbol{v}_\Lambda - \boldsymbol{u}\|_{\ell_2}^2,$$

for all $\boldsymbol{v}_\Lambda \in \mathcal{R}(u^{(0)})$.

Proof. It is enough to combine (5.7), Lemma 5.4, and (5.2) with $\boldsymbol{v} = \boldsymbol{v}_\Lambda$ and $\boldsymbol{w} = \boldsymbol{u}$. ∎

The result means that $(\mathbb{E}(\boldsymbol{v}_\Lambda))^{1/2}$ is an a priori error bound both from above and from below; similarly $\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2}$ is an upper and lower a posteriori error estimator. Of course, in general, none of them is computable in practice. However, by using inequalities (4.4), we can easily obtain from the latter a computable a posteriori error estimator.

5.3. Improving the error estimate. In this section, we show that the application of **APPROX-GRAD** within the procedure **DESCENT** may lead to the improvement of the currently available estimate of the quantity $\mathbb{E}$. This feature will be crucial in the design of the adaptive algorithm considered in the next section.

Let $\boldsymbol{v}_\Lambda$ be any compactly supported vector and let $\mathbb{E}_\Lambda$ be any upper estimate of $\mathbb{E}(\boldsymbol{v}_\Lambda)$, i.e., $\mathbb{E}(\boldsymbol{v}_\Lambda) \le \mathbb{E}_\Lambda$. By Proposition 5.7, there exists a constant $\mathcal{C}_1$ (whose expression is easily computable) such that

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2} \le \mathcal{C}_1 \mathbb{E}(\boldsymbol{v}_\Lambda)^{1/2} \le \mathcal{C}_1 \mathbb{E}_\Lambda^{1/2}. \tag{5.12}$$

Let us enter **APPROX-GRAD**$[\boldsymbol{v}_\Lambda, \varepsilon, \gamma]$ with the choice $\varepsilon = \mathcal{C}_1 \mathbb{E}_\Lambda^{1/2}$. Observe that whenever the loop in **APPROX-GRAD** does not stop, i.e., whenever for some n we have $\|\boldsymbol{w}_\Lambda^{(n)}\|_{\ell_2} < \frac{1+\gamma}{1-\gamma}\eta^{(n)}$, then we get a new a priori information on the norm of the gradient, namely,

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2} \le \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda) - \boldsymbol{w}_\Lambda^{(n)}\|_{\ell_2} + \|\boldsymbol{w}_\Lambda^{(n)}\|_{\ell_2} < \eta^{(n)} + \frac{1+\gamma}{1-\gamma}\eta^{(n)} = \frac{2}{1-\gamma}\eta^{(n)}.$$

Thus, if $\bar{n}$ denotes the number of applications of **EVAL-GRAD** within **APPROX-GRAD**, then we know that

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2} \le \frac{2}{1-\gamma}\nu^{\bar{n}-1}\varepsilon = \mathcal{C}_1\frac{2}{1-\gamma}\nu^{\bar{n}-1}\mathbb{E}_\Lambda^{1/2}. \tag{5.13}$$

Again by Proposition 5.7 we have, for a suitable constant $\mathcal{C}_2$,

$$\mathbb{E}(\boldsymbol{v}_\Lambda) \le \mathcal{C}_2\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2}^2 \le \mathcal{C}_2\mathcal{C}_1^2\frac{4}{(1-\gamma)^2}\nu^{2\bar{n}-2}\,\mathbb{E}_\Lambda. \tag{5.14}$$

If we define n_0 as the smallest integer ≥ 1 such that

$$\mathcal{C}_2\mathcal{C}_1^2\frac{4}{(1-\gamma)^2} \le \nu^{-2(n_0-1)}, \tag{5.15}$$

then we obtain

$$\mathbb{E}(\boldsymbol{v}_\Lambda) \le \nu^{2(\bar{n}-n_0)}\,\mathbb{E}_\Lambda.$$

Obviously, we have to take the most accurate estimate between this one and the initial one $\mathbb{E}(\boldsymbol{v}_\Lambda) \le \mathbb{E}_\Lambda$. Denoting by $(z)_+$ the positive part of a number z, we conclude that at the output of an application of **APPROX-GRAD**, in which $\bar{n}$ applications of **EVAL-GRAD** have been made, we know that

$$\mathbb{E}(\boldsymbol{v}_\Lambda) \le \nu^{2(\bar{n}-n_0)_+}\,\mathbb{E}_\Lambda. \tag{5.16}$$

5.4. Coarsening. The coarsening procedure **COARSE:** $[\boldsymbol{v}_\Lambda, \vartheta] \mapsto \boldsymbol{w}_\Lambda$ as described in section 4.3 can be simplified and optimized by exploiting the assumption (5.1). Under the sole condition that an estimate of $\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2}$ is known, we can get $\boldsymbol{w}_\Lambda$ without any application of **APPROX-GRAD**, by calling once the procedure **THRESH** introduced therein. Indeed, by the convexity of $\mathbb{J}$ and by (5.1), we have for any $\boldsymbol{w}_\Lambda \in \mathcal{R}(u^{(0)})$

$$\begin{aligned}\mathbb{J}(\boldsymbol{w}_\Lambda) - \mathbb{J}(\boldsymbol{v}_\Lambda) &\le \langle \boldsymbol{J}'(\boldsymbol{w}_\Lambda), \boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\rangle \\ &= \langle \boldsymbol{J}'(\boldsymbol{w}_\Lambda) - \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\rangle + \langle \boldsymbol{J}'(\boldsymbol{v}_\Lambda), \boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\rangle \\ &\le L_{J,\Psi}\|\boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\|_{\ell_2}^2 + \|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2}\|\boldsymbol{w}_\Lambda - \boldsymbol{v}_\Lambda\|_{\ell_2}.\end{aligned}$$

Assume that $\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda)\|_{\ell_2} \le \xi$ for some $\xi > 0$. Then, setting

$$\varepsilon_\vartheta := \min\left(\left(\frac{\vartheta}{2L_{J,\Psi}}\right)^{1/2}, \frac{\vartheta}{2\xi}\right) \qquad \text{and} \qquad \boldsymbol{w}_\Lambda := \mathbf{THRESH}[\boldsymbol{v}_\Lambda, \varepsilon_\vartheta], \tag{5.17}$$

we obtain the desired coarsened vector $\boldsymbol{w}_\Lambda$.

5.5. The adaptive algorithm. Denote by $\mathbb{E}_0$ any computable estimate of the initial error $\mathbb{E}(\boldsymbol{u}_\Lambda^{(0)})$, i.e., choose $\mathbb{E}_0$ so that

$$\mathbb{E}(\boldsymbol{u}_\Lambda^{(0)}) := \mathbb{J}(\boldsymbol{u}_\Lambda^{(0)}) - \mathbb{J}(\boldsymbol{u}) \le \mathbb{E}_0. \tag{5.18}$$

Let $\nu \in (0,1)$ be a fixed constant. We now prove by recursion that the parameters appearing in Algorithm 4.12 **MINIMIZE** can be chosen in such a way that either

the exact solution $\boldsymbol{u}$ is obtained after a finite number of steps, or there exists a strictly increasing sequence of integers e_m with $e_0 = 0$ such that

$$\mathbb{E}(\boldsymbol{u}_\Lambda^{(m)}) \le \nu^{2e_m}\,\mathbb{E}_0; \tag{5.19}$$

furthermore, denoting by N_m the number of applications of **EVAL-GRAD** to get $\boldsymbol{u}_\Lambda^{(m)}$, we will relate N_m to e_m.

For $m = 0$, inequality (5.19) is precisely (5.18). By induction, we assume that (5.19) holds up to some $m \ge 0$. We set $e_{m,0} = 0$ and we prove by recursion that there exists a nondecreasing sequence of integers $e_{m,k}$ such that

$$\mathbb{E}(\boldsymbol{v}_\Lambda^{(m,k)}) \le \mathbb{E}_{m,k} \quad \text{with } \mathbb{E}_{m,k} := \sigma^k \nu^{2(e_m+e_{m,k})}\,\mathbb{E}_0, \qquad k = 0, 1, \dots. \tag{5.20}$$

By (5.19), this inequality holds for $k = 0$. By induction, assume that it holds up to some $k \ge 0$. Set

$$\varepsilon^{(m,k)} := \mathcal{C}_1 \mathbb{E}_{m,k}^{1/2} \tag{5.21}$$

and apply **DESCENT**$[\boldsymbol{v}_\Lambda^{(m,k)}, \varepsilon^{(m,k)}]$. If the output η of **APPROX-GRAD** within **DESCENT** is zero, we have found the exact solution and the algorithm stops. Otherwise, let $\bar{n}_{m,k}$ denote the number of applications of **EVAL-GRAD** inside the routine **DESCENT**$[\boldsymbol{v}_\Lambda^{(m,k)}, \varepsilon^{(m,k)}]$. Recalling (5.16) and Proposition 5.5, we obtain

$$\mathbb{E}(\boldsymbol{v}_\Lambda^{(m,k+1)}) \le \sigma \mathbb{E}(\boldsymbol{v}_\Lambda^{(m,k)}) \le \sigma \nu^{2(\bar{n}_{m,k}-n_0)_+}\,\mathbb{E}_{m,k}. \tag{5.22}$$

Setting $e_{m,k+1} := e_{m,k} + (\bar{n}_{m,k} - n_0)_+$, we obtain (5.20) with k replaced by $k+1$. This completes the inner recursion argument.

Let $K^{(m)}$ be an integer to be determined. The coarsening step **COARSE** yields

$$\mathbb{E}(\boldsymbol{u}_\Lambda^{(m+1)}) - \mathbb{E}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) = \mathbb{J}(\boldsymbol{u}_\Lambda^{(m+1)}) - \mathbb{J}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})}) \le \vartheta^{(m)}.$$

Choose

$$\vartheta^{(m)} := \bar{\mathcal{C}}\,\mathbb{E}_{m,K^{(m)}}, \tag{5.23}$$

where $\bar{\mathcal{C}} > 0$ is by now a constant which can be freely chosen, but in the next section will be chosen in order to guarantee the optimality of the thresholding procedure. Then, recalling the definition in (5.20), we obtain

$$\mathbb{E}(\boldsymbol{u}_\Lambda^{(m+1)}) \le (1+\bar{\mathcal{C}})\,\sigma^{K^{(m)}} \nu^{2(e_m+e_{m,K^{(m)}})}\,\mathbb{E}_0.$$

This suggests choosing $K^{(m)} \ge 1$ as the smallest integer for which we have

$$g_{m,K^{(m)}} \ge 1, \text{ where } g_{m,K^{(m)}} \text{ satisfies } \nu^{2g_{m,K^{(m)}}} := (1+\bar{\mathcal{C}})\,\sigma^{K^{(m)}} \nu^{2e_{m,K^{(m)}}}. \tag{5.24}$$

Note that $K^{(m)}$ is bounded by $\bar{K}$, where $\bar{K} \ge 1$ is the smallest integer satisfying $(1+\bar{\mathcal{C}})\,\sigma^{\bar{K}} \le \nu^2$. Setting

$$\bar{e}_m := [g_{m,K^{(m)}}] \ge 1 \qquad \text{and} \qquad e_{m+1} := e_m + \bar{e}_m, \tag{5.25}$$

we obtain (5.19) with m replaced by $m+1$. This completes the outer recursion argument.

We note that each coarsening step $\boldsymbol{u}_\Lambda^{(m+1)} := \mathbf{COARSE}[\boldsymbol{v}_\Lambda^{(m,K^{(m)})}, \vartheta^{(m)}]$ can be accomplished as described in section 5.4. Recalling (5.12), we have

$$\|\boldsymbol{J}'(\boldsymbol{v}_\Lambda^{(m,K^{(m)})})\|_{\ell_2} \le \mathcal{C}_1 \mathbb{E}(\boldsymbol{v}_\Lambda^{(m,K^{(m)})})^{1/2} \le \mathcal{C}_1 \mathbb{E}_{m,K^{(m)}}^{1/2} =: \xi^{(m)}.$$

Since we have chosen $\vartheta^{(m)} = \bar{\mathcal{C}}\,\mathbb{E}_{m,K^{(m)}}$, by (5.17) we obtain $\boldsymbol{u}_\Lambda^{(m+1)}$ as the output of $\mathbf{THRESH}[\boldsymbol{v}_\Lambda^{(m,K^{(m)})}, \varepsilon_\vartheta^{(m)}]$ with

$$\varepsilon_\vartheta^{(m)} := \min\left(\left(\frac{\bar{\mathcal{C}}}{2L_{J,\Psi}}\right)^{1/2}, \frac{\bar{\mathcal{C}}}{2\mathcal{C}_1}\right)\mathbb{E}_{m,K^{(m)}}^{1/2}. \tag{5.26}$$

Finally, let us count the number of applications of **EVAL-GRAD** in our algorithm. By the previous considerations, they occur only within **DESCENT**. Denote by $N_{m+1,m} = \bar{n}_{m,0} + \cdots + \bar{n}_{m,K^{(m)}-1}$ the number of applications of **EVAL-GRAD** needed to compute $\boldsymbol{u}_\Lambda^{(m+1)}$ from $\boldsymbol{u}_\Lambda^{(m)}$. We observe that surely $e_{m,K^{(m)}} \le \bar{e}_m + \delta\bar{K}$ for some $\delta \ge 1$. Thus,

$$\begin{aligned} N_{m+1,m} &= \sum_{k=0}^{K^{(m)}-1} (\bar{n}_{m,k} - n_0) + n_0 K^{(m)} \le \sum_{k=0}^{K^{(m)}-1} (\bar{n}_{m,k} - n_0)_+ + n_0\bar{K} \\ &= e_{m,K^{(m)}} + n_0\bar{K} \le \bar{e}_m + (n_0+\delta)\bar{K} \\ &\le (n_0+\delta+1)\bar{K}\bar{e}_m = q\bar{K}(e_{m+1} - e_m) \end{aligned}$$

with $q := n_0 + \delta + 1$ and n_0 defined in (5.15). We conclude that the number N_{m+1} of applications of **EVAL-GRAD** needed to compute $\boldsymbol{u}_\Lambda^{(m+1)}$ from $\boldsymbol{u}_\Lambda^{(0)}$ satisfies

$$N_{m+1} = \sum_{\mu=0}^{m} N_{\mu+1,\mu} \le q\bar{K}\sum_{\mu=0}^{m}(e_{\mu+1} - e_\mu) = q\bar{K}e_{m+1}.$$

Remark 5.8. In view of the subsequent optimality analysis, we observe that a version of the algorithm can be given in which each $\bar{n}_{m,k}$ (and hence $N_{m+1,m}$) is guaranteed to be uniformly bounded. In fact, (5.14) shows that the output $\boldsymbol{w}_\Lambda^{(n)}$ of the routine $\mathbf{EVAL-GRAD}[\boldsymbol{v}_\Lambda, \nu^{n-1}\varepsilon]$ satisfies

$$\mathbb{E}(\boldsymbol{w}_\Lambda^{(n)}) \le \mathcal{C}_2\mathcal{C}_1^2 \frac{4}{(1-\gamma)^2}\nu^{2n-2}\,\mathbb{E}_\Lambda,$$

where again $\mathbb{E}_\Lambda$ is an estimate for $\mathbb{E}(\boldsymbol{v}_\Lambda)$. Choosing $\bar{n}$ as the smallest integer satisfying

$$\mathcal{C}_2\mathcal{C}_1^2 \frac{4}{(1-\gamma)^2}\nu^{2n-2} \le \sigma,$$

we obtain $\mathbb{E}(\boldsymbol{w}_\Lambda^{(n)}) \le \sigma\,\mathbb{E}_\Lambda$. Hence, we can restrict the loop in **APPROX-GRAD** to $n = 1, 2, \ldots, \bar{n}$, with an absolute constant $\bar{n}$. If (4.3) is satisfied before the end of the loop, we leave **APPROX-GRAD** and proceed by **LINE-SEARCH** (resulting in some $\boldsymbol{v}_\Lambda$) and obtain the error reduction $\mathbb{E}(\boldsymbol{v}_\Lambda) \le \sigma\,\mathbb{E}_\Lambda$ by the above arguments. Otherwise, after at most $\bar{n}$ steps, we leave **APPROX-GRAD** and set $\boldsymbol{v}_\Lambda := \boldsymbol{w}_\Lambda^{(\bar{n})}$, avoiding the necessity of the line search in this case. This again results in the same error reduction.

Let us summarize our result in the following theorem.

THEOREM 5.9. *Let $p = 2$ and let* (5.1) *hold. We consider Algorithm* 4.12 **MINIMIZE**. *We fix a constant $\nu \in (0,1)$ and we choose a constant $\mathbb{E}_0$ satisfying* (5.18). *Then, for the choice* (5.24) *of the parameters $K^{(m)}$ (with $K^{(m)} \lesssim 1$), $\vartheta^{(m)}$ as in* (5.23) *and $\varepsilon^{(m,k)}$ as in* (5.21), *either the algorithm yields the exact solution $\boldsymbol{u}$ after a finite number of steps, or there exists a strictly increasing sequence $\{e_m\}$ of integers* (5.25), *with $e_0 = 0$, such that the sequence of approximations $\boldsymbol{u}_\Lambda^{(m)}$ produced by the algorithm satisfies*

$$\mathbb{E}(\boldsymbol{u}_\Lambda^{(m)}) \le \nu^{2e_m}\,\mathbb{E}_0, \qquad m = 0, 1, \dots.$$

Furthermore, the number N_m of applications of **EVAL-GRAD** *needed to compute $\boldsymbol{u}_\Lambda^{(m)}$ from $\boldsymbol{u}_\Lambda^{(0)}$ satisfies $N_m \lesssim e_m$; i.e., it grows at most as the logarithm of the obtained accuracy.* □

Recalling (3.1) and Proposition 5.7, we immediately obtain an error estimate in the V-norm.

COROLLARY 5.10. *Under the same conditions of the previous theorem, we have*

$$\|u_\Lambda^{(m)} - u\|_V \le \nu^{e_m}\, C_\Psi \left(\frac{2\mathbb{E}_0}{c_{J,\Psi}}\right)^{1/2}, \qquad m = 0, 1, \dots. \qquad \square$$

5.6. Stopping criteria. The previous algorithm produces arbitrarily close approximations to the exact solution u. In practice, one fixes a tolerance TOL and wishes to stop the algorithm as soon as the inequality $\|u_\Lambda^{(m)} - u\|_V < TOL$ is guaranteed to hold. According to the previous estimate, one can take as m the smallest integer for which

$$\nu^{e_m} < \frac{1}{C_\Psi}\left(\frac{c_{J,\Psi}}{2\mathbb{E}_0}\right)^{1/2} TOL.$$

On the other hand, the stopping test needs to be included also within **APPROX-GRAD**, in order to prevent an infinite loop, or simply to avoid unnecessary applications of **EVAL-GRAD** therein. Precisely, recalling (5.13), the loop in **APPROX-GRAD** is stopped if, for some n,

$$\frac{2}{1-\gamma}\nu^{n-1}\varepsilon < \omega\, TOL,$$

where $\omega \in (0,1)$ can be easily determined from Proposition 5.7 so that $\|\boldsymbol{v}_\Lambda - \boldsymbol{u}\|_{\ell_2} < TOL/(2C_\Psi)$. Setting $\boldsymbol{w}_\Lambda := \mathbf{THRESH}[\boldsymbol{v}_\Lambda, TOL/(2C_\Psi)]$, the triangle inequality yields $\|\boldsymbol{w}_\Lambda - \boldsymbol{u}\|_{\ell_2} < TOL/C_\Psi$, whence $\|w_\Lambda - u\|_V < TOL$ and the algorithm is stopped.

6. Optimality properties of the algorithm. We continue the discussion of the algorithm described in section 5. Here, we investigate how the number of active basis functions in $u_\Lambda^{(m)}$ (i.e., the cardinality of the support of $\boldsymbol{u}_\Lambda^{(m)}$) and the overall computational complexity needed to get $u_\Lambda^{(m)}$ (number of arithmetic operations and sortings) are related to the accuracy of the approximation.

The relationship between cardinality of the active degrees of freedom (i.e., the number of active basis functions) and accuracy is a central topic in nonlinear approx-

imation theory [22]. This involves the best N-term or the best η-accurate approximation to a given $\boldsymbol{v} \in \ell_2$ defined as follows. Let

$$\Sigma_N := \left\{ \boldsymbol{w} = \sum_{\lambda \in \Lambda} w_\lambda \, \psi_\lambda : \, |\Lambda| \leq N \right\}$$

(where $|\Lambda|$ denotes the cardinality of $\Lambda \subset \mathcal{J}$) be the nonlinear manifold of all linear combinations of basis functions containing at most N terms. Then a best N-term approximation $\boldsymbol{v}_N$ to $\boldsymbol{v}$ is defined by

$$\|\boldsymbol{v} - \boldsymbol{v}_N\|_{\ell_2} = \inf_{\boldsymbol{w}_N \in \Sigma_N^{\mathrm{con}}} \|\boldsymbol{v} - \boldsymbol{w}_N\|_{\ell_2} =: \varrho_N^{\mathrm{con}}(\boldsymbol{v}),$$

where $\Sigma_N^{\mathrm{con}} \subseteq \Sigma_N$ takes into account possible constraints in the choice of the active degrees of freedom (see (7.2) for an example).

Based on this, approximation spaces $\mathcal{A}^s$ are defined as the quasi-normed sequence space consisting of all those elements whose error $\varrho_N^{\mathrm{con}}(\boldsymbol{v})$ decays at least as N^{-s}. The quasi norm is defined by

$$\|\boldsymbol{v}\|_{\mathcal{A}^s} := \sup_{N>0} N^s \, \varrho_N^{\mathrm{con}}(\boldsymbol{v}).$$

On the other hand, a best η-accurate approximation $\boldsymbol{v}_\eta$ to $\boldsymbol{v}$ is a vector of smallest support compatible with the constraints, such that

$$\|\boldsymbol{v} - \boldsymbol{v}_\eta\|_{\ell_2} \leq \eta.$$

A typical result is as follows: any best η-accurate approximation $\boldsymbol{v}_\eta$ satisfies the inequality $|\operatorname{supp} \boldsymbol{v}_\eta| \lesssim \eta^{-1/s} \|\boldsymbol{v}\|_{\mathcal{A}^s}^{1/s}$ as $\eta \to 0$.

In this framework, the behavior of best N-term or best η-accurate approximations to $\boldsymbol{v}$ provides a benchmark to evaluate the quality of any compactly supported approximation of $\boldsymbol{v}$. Precisely, a family $\{\boldsymbol{v}_\Lambda\}$ of compactly supported vectors converging to $\boldsymbol{v}$ in the ℓ_2-norm is said to be *asymptotically optimal* if it satisfies $\|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s} \lesssim \|\boldsymbol{v}\|_{\mathcal{A}^s}$ and

$$\|\boldsymbol{v} - \boldsymbol{v}_\Lambda\|_{\ell_2} \lesssim |\Lambda|^{-s} \|\boldsymbol{v}\|_{\mathcal{A}^s},$$

or, equivalently,

$$|\Lambda| \lesssim \|\boldsymbol{v} - \boldsymbol{v}_\Lambda\|_{\ell_2}^{-1/s} \|\boldsymbol{v}\|_{\mathcal{A}^s}^{1/s}.$$

The latter condition is surely satisfied if for any $\boldsymbol{v}_\Lambda$ there exists $\eta = \eta_\Lambda > 0$ such that

$$\|\boldsymbol{v} - \boldsymbol{v}_\Lambda\|_{\ell_2} \leq \eta \qquad \text{and} \qquad |\Lambda| \lesssim \eta^{-1/s} \|\boldsymbol{v}\|_{\mathcal{A}^s}^{1/s}.$$

Note that in this context the subscript Λ refers to the actual support of the vector. We remark that these results are known for L_p-spaces as well; see, e.g., [14, 22].

Before we proceed, let us give two concrete examples for the space $\mathcal{A}^s$. If no constraint is imposed on the choice of the active degrees of freedom, then $\mathcal{A}^s = \ell_\tau^{\mathrm{w}}$, where ℓ_τ^{w} (with $\frac{1}{\tau} = s + \frac{1}{2}$) is the Lorentz space of sequences $\boldsymbol{v} = \{v_\lambda\} \in \ell_2$ whose nonincreasing rearrangement $|v_{\lambda_1}| \geq |v_{\lambda_2}| \geq \cdots \geq |v_{\lambda_n}| \geq \cdots$ satisfies

$$|\boldsymbol{v}|_{\ell_\tau^{\mathrm{w}}} := \sup_n n^{1/\tau} |v_{\lambda_n}| < \infty.$$

This expression can be viewed as a quantitative measurement of the sparseness of the sequence $\boldsymbol{v}$. The space ℓ_τ^{w} is equipped with the (quasi) norm $\|\boldsymbol{v}\|_{\ell_\tau^{\mathrm{w}}} := \|\boldsymbol{v}\|_{\ell_2} + |\boldsymbol{v}|_{\ell_\tau^{\mathrm{w}}}$. If V is the Besov space $B^r_{p,p}(\Omega)$ of functions defined in a domain $\Omega \subset \mathbb{R}^d$, then $\boldsymbol{v} \in \mathcal{A}^s$ is implied by the regularity condition $v = \boldsymbol{v}^T\Psi \in B^{r+ds}_{\tau,\tau}(\Omega)$, provided the basis Ψ characterizes this space (see [5, 6, 15, 17, 18, 20] for more details, where the construction is detailed in $L_2(\Omega)$ but can easily be extended to the L_p-case).

As a second example, we consider a basis Ψ of compactly supported wavelets in Ω. Suppose that the set of active degrees of freedom is constrained to have a tree-structure (i.e., $\lambda \in \operatorname{supp} \boldsymbol{v}_N$ implies $\mu \in \operatorname{supp} \boldsymbol{v}_N$ for all wavelets ψ_μ whose support contains the support of ψ_λ). Then, $\mathcal{A}^s = \mathcal{A}^{s,\mathrm{tree}}$ is the space of sequences whose best tree N-term approximation converges at a rate N^{-s}, i.e., the space of all ℓ_2-sequences $\boldsymbol{v}$ such that

$$\varrho_N^{\mathrm{tree}}(\boldsymbol{v}) \;\lesssim\; N^{-s},$$

which is a quasi-normed space under the quasi norm

$$\|\boldsymbol{v}\|_{\mathcal{A}^{s,\mathrm{tree}}} := \sup_{n\in\mathbb{N}} N^s\, \varrho_N^{\mathrm{tree}}(\boldsymbol{v}).$$

Here $\boldsymbol{v} \in \mathcal{A}^s$ holds provided $v = \boldsymbol{v}^T\Psi \in B^{r+ds}_{\tau^*,\tau^*}(\Omega)$ holds for some $\tau^* > \tau$ with τ as above. We now suppose that the procedure **THRESH** defined in section 4.3 satisfies the following condition.

ASSUMPTION 6.1. *There exists a constant $C^* \geq 1$ such that if $\boldsymbol{v} \in \mathcal{A}^s$ and if $\boldsymbol{v}_\Lambda$ satisfies $\|\boldsymbol{v} - \boldsymbol{v}_\Lambda\|_{\ell_2} \leq \varepsilon$, then $\boldsymbol{z}_\Lambda := \mathbf{THRESH}[\boldsymbol{v}_\Lambda, C^*\varepsilon]$ satisfies*

$$\|\boldsymbol{v} - \boldsymbol{z}_\Lambda\|_{\ell_2} \leq (1 + C^*)\,\varepsilon, \quad |\operatorname{supp} \boldsymbol{z}_\Lambda| \;\lesssim\; \varepsilon^{-1/s}\|\boldsymbol{v}\|_{\mathcal{A}^s}, \quad \|\boldsymbol{z}_\Lambda\|_{\mathcal{A}^s} \;\lesssim\; \|\boldsymbol{v}\|_{\mathcal{A}^s}. \tag{6.1}$$

This condition holds both for the unconstrained thresholding (e.g., with $C^* = 4$; see [10, 11]) and for the tree-thresholding (with a possibly larger C^*; see [4, 12]). Obviously, with a larger constant C^*, the results are still valid.

From now on, let us suppose that $\boldsymbol{u} \in \mathcal{A}^s$ for some $s > 0$. We recall that, for any $m \geq 0$, the vector $\boldsymbol{v}_\Lambda^{(m,K^{(m)})}$ satisfies by Proposition 5.7 and by (5.20)

$$\|\boldsymbol{v}_\Lambda^{(m,K^{(m)})} - \boldsymbol{u}\|_{\ell_2} \leq \left(\frac{2}{c_{J,\Psi}}\right)^{1/2} \mathbb{E}^{1/2}_{m,K^{(m)}} =: \bar{\varepsilon}^{(m)}. \tag{6.2}$$

Furthermore, we recall that $\boldsymbol{u}_\Lambda^{(m+1)} = \mathbf{THRESH}[\boldsymbol{v}_\Lambda^{(m,K^{(m)})}, \varepsilon_\vartheta^{(m)}]$, where $\varepsilon_\vartheta^{(m)}$ is defined in (5.26). Now we choose the constant $\bar{\mathcal{C}}$ introduced in (5.23) in such a way that $\varepsilon_\vartheta^{(m)} = C^*\bar{\varepsilon}^{(m)}$, i.e.,

$$\min\left(\left(\frac{\bar{\mathcal{C}}}{2L_{J,\Psi}}\right)^{1/2}, \frac{\bar{\mathcal{C}}}{2\mathcal{C}_1}\right) = C^*\left(\frac{2}{c_{J,\Psi}}\right)^{1/2}. \tag{6.3}$$

In this way (6.1) applies to $\boldsymbol{v} = \boldsymbol{u}$, $\boldsymbol{v}_\Lambda = \boldsymbol{v}_\Lambda^{(m,K^{(m)})}$, and $\boldsymbol{z}_\Lambda = \boldsymbol{u}_\Lambda^{(m+1)}$. Finally, observing that $\mathbb{E}^{1/2}_{m,K^{(m)}} \sim \nu^{e_{m+1}}\mathbb{E}_0^{1/2}$ and shifting m into $m-1$, we get the following result.

PROPOSITION 6.2. *Let the assumptions of Theorem* 5.9 *and Assumption* 6.1 *hold. Then, the iterates $\boldsymbol{u}_\Lambda^{(m)}$ generated by Algorithm* 4.12 **MINIMIZE** *with the choice* (6.3) *satisfy*

$$|\operatorname{supp} \boldsymbol{u}_\Lambda^{(m)}| \;\lesssim\; (\nu^{e_m}\mathbb{E}_0^{1/2})^{-1/s}\|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} \qquad \textit{and} \qquad \|\boldsymbol{u}_\Lambda^{(m)}\|_{\mathcal{A}^s} \;\lesssim\; \|\boldsymbol{u}\|_{\mathcal{A}^s},$$

for $m = 0, 1, \ldots$. ∎

Since we have $\|\boldsymbol{u} - \boldsymbol{u}_\Lambda^{(m)}\|_{\ell_2} \lesssim \nu^{e_m}\mathbb{E}_0^{1/2}$ by Theorem 5.9, this proves the optimality of the approximation as far as the number of active degrees of freedom is concerned.

In order to control the possible growth of the supports of the vectors in the intermediate stages of the algorithm, as well as the computational complexity, we suppose that the procedure **EVAL-GRAD** defined in section 4.1 satisfies the following condition.

ASSUMPTION 6.3. *Given any tolerance $\varepsilon > 0$ and any compactly supported vector $\boldsymbol{v}_\Lambda$, then the output $\boldsymbol{w}_\Lambda := \mathbf{EVAL\text{-}GRAD}[\boldsymbol{v}_\Lambda, \varepsilon]$ satisfies*

$$\begin{aligned} |\operatorname{supp} \boldsymbol{w}_\Lambda| &\lesssim \varepsilon^{-1/s}(\|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s}^{1/s} + \|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1), \\ \|\boldsymbol{w}_\Lambda\|_{\mathcal{A}^s} &\lesssim \|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s} + \|\boldsymbol{u}\|_{\mathcal{A}^s} + 1. \end{aligned}$$

The number $\operatorname{ops} \boldsymbol{w}_\Lambda$ of operations needed to compute $\boldsymbol{w}_\Lambda$ satisfies

$$\operatorname{ops} \boldsymbol{w}_\Lambda \lesssim \varepsilon^{-1/s}(\|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s}^{1/s} + \|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1) + |\operatorname{supp} \boldsymbol{v}_\Lambda|. \tag{6.4}$$

This condition is fulfilled in a number of relevant cases, as described in [11, 12]. The second term on the right-hand side of (6.4) can be neglected in certain situations (e.g., in the linear and certain nonlinear cases). When it is needed, it accounts for computing a chain of near-best trees for the given input index set. It was shown in [13, Theorem 3.4] that the number $\log_2(\|\boldsymbol{v}_\Lambda\|/\varepsilon)$ of these trees depends on only $\|\boldsymbol{v}_\Lambda\|$ and the target accuracy. Moreover, it was shown there that the overall cost to compute these trees remains proportional to $\varepsilon^{-1/s}(\|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s}^{1/s} + 1) + |\mathcal{T}(\boldsymbol{v}_\Lambda)|$, where $\mathcal{T}(\boldsymbol{v})$ denotes the smallest tree containing $\operatorname{supp} \boldsymbol{v}$. It is not restrictive to assume in the nonlinear case that the input as well as all intermediate vectors do already have tree structures. This shows that applying **EVAL-GRAD** to a sequence of decreasing tolerances for the *same* input vector simply may require computing additional near-best trees. This in turns means that $|\mathcal{T}(\boldsymbol{v}_\Lambda)| = |\operatorname{supp} \boldsymbol{v}_\Lambda|$ has to be counted only once in the operation count.

From the previous property, we deduce similar bounds for the output of the routine **APPROX-GRAD**. Precisely, recalling that **EVAL-GRAD** is recursively applied in **APPROX-GRAD** with the *same* input vector $\boldsymbol{v}_\Lambda$, we immediately deduce that

$$|\operatorname{supp} \boldsymbol{G}(\boldsymbol{v}_\Lambda)| \lesssim \eta^{-1/s}(\|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s}^{1/s} + \|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1), \tag{6.5}$$

$$\|\boldsymbol{G}(\boldsymbol{v}_\Lambda)\|_{\mathcal{A}^s} \lesssim \|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s} + \|\boldsymbol{u}\|_{\mathcal{A}^s} + 1, \tag{6.6}$$

where the constants on the right-hand side do not depend on the number of applications of **EVAL-GRAD**.

Concerning the operation count, we have to take into account that the accuracy in the calls to **EVAL-GRAD** is reduced at a geometric rate; i.e., we have $\eta^{(k)} = \varepsilon\,\nu^{k-1}$ after k calls of **EVAL-GRAD**. Furthermore, we recall that the second term on the right-hand side of (6.4) can be counted only once. Hence, denoting again by $\bar{n}$ the number of calls to **EVAL-GRAD**, and observing that $\eta = \eta^{(\bar{n})}$, we have

$$\begin{aligned} \operatorname{ops} \boldsymbol{G}(\boldsymbol{v}_\Lambda) &\lesssim \sum_{k=1}^{\bar{n}} \operatorname{ops}\left(\mathbf{EVAL\text{-}GRAD}[\boldsymbol{v}_\Lambda, \eta^{(k)}]\right) \\ &\lesssim \sum_{k=1}^{\bar{n}} (\eta^{(k)})^{-1/s}(\|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s}^{1/s} + \|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1) + |\operatorname{supp} \boldsymbol{v}_\Lambda| \\ &\lesssim \eta^{-1/s}(\|\boldsymbol{v}_\Lambda\|_{\mathcal{A}^s}^{1/s} + \|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1) + |\operatorname{supp} \boldsymbol{v}_\Lambda|, \end{aligned} \tag{6.7}$$

where again the constants on the right-hand side do not depend on $\bar{n}$. Recalling Remark 5.8, we point out that the number of applications of **EVAL-GRAD** in each iteration can be uniformly bounded. We stress the fact that the crucial point is—as already pointed out earlier—that **EVAL-GRAD** is called within **APPROX-GRAD** with the *same* input vector $\boldsymbol{v}_\Lambda$ with a series of tolerances that are geometrically decreasing. This implies, as shown in (6.7), a uniform bound on the number of operations in terms of the error which is actually reachable for the input $\boldsymbol{v}_\Lambda$.

Consequently, we obtain similar estimates if, on the left-hand sides, we replace $\boldsymbol{G}(\boldsymbol{v}_\Lambda)$ by the output $\boldsymbol{w}_\Lambda := \mathbf{DESCENT}[\boldsymbol{v}_\Lambda, \varepsilon]$; in this case, η is the quantity determined within **APPROX-GRAD**.

We now apply these results within the Algorithm **MINIMIZE**, with the choice of parameters described in sections 5 and 6. For any $m \geq 0$ and $k = 0, \dots, K^{(m)} - 1$, set $[\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)}), \eta^{(m,k)}] := \mathbf{APPROX\text{-}GRAD}[\boldsymbol{v}_\Lambda^{(m,k)}, \varepsilon^{(m,k)}]$. Recalling (5.20) and (5.21) as well as the definitions of $\bar{n}_{m,k}$ and $e_{m,k+1}$, it is straightforward to check that

$$\eta^{(m,k)} \geq \mathcal{C}_1 \sigma^{-1/2} \mathbb{E}_{m,k+1}^{1/2} \begin{cases} \nu^{n_0} & \text{if } \bar{n}_{m,k} \geq n_0, \\ 1 & \text{if } \bar{n}_{m,k} < n_0. \end{cases}$$

In both cases, $\eta^{(m,k)} \gtrsim \mathbb{E}_{m,k+1}^{1/2}$ uniformly in m and k. Thus, by (6.5)–(6.7) we obtain

$$|\operatorname{supp} \boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})| \lesssim (\mathbb{E}_{m,k+1}^{1/2})^{-1/s} (\|\boldsymbol{v}_\Lambda^{(m,k)}\|_{\mathcal{A}^s}^{1/s} + \|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1), \tag{6.8}$$

$$\|\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\mathcal{A}^s} \lesssim \|\boldsymbol{v}_\Lambda^{(m,k)}\|_{\mathcal{A}^s} + \|\boldsymbol{u}\|_{\mathcal{A}^s} + 1, \tag{6.9}$$

$$\operatorname{ops} \boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)}) \lesssim (\mathbb{E}_{m,k+1}^{1/2})^{-1/s} (\|\boldsymbol{v}_\Lambda^{(m,k)}\|_{\mathcal{A}^s}^{1/s} + \|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1) + |\operatorname{supp} \boldsymbol{v}_\Lambda^{(m,k)}|. \tag{6.10}$$

Next, from the definition of $\boldsymbol{v}_\Lambda^{(m,k+1)} = \boldsymbol{v}_\Lambda^{(m,k)} + \mu^{(m,k)} \boldsymbol{s}_\Lambda^{(m,k)}$, we deduce that

$$|\operatorname{supp} \boldsymbol{v}_\Lambda^{(m,k+1)}| \leq |\operatorname{supp} \boldsymbol{v}_\Lambda^{(m,k)}| + |\operatorname{supp} \boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})|, \tag{6.11}$$

$$\operatorname{ops}(\boldsymbol{v}_\Lambda^{(m,k+1)}, \boldsymbol{v}_\Lambda^{(m,k)}) \lesssim \operatorname{ops} \boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)}) + |\operatorname{supp} \boldsymbol{v}_\Lambda^{(m,k+1)}|, \tag{6.12}$$

where $\operatorname{ops}(\boldsymbol{w}_\Lambda, \boldsymbol{v}_\Lambda)$ means the number of operations needed to compute $\boldsymbol{w}_\Lambda$, given $\boldsymbol{v}_\Lambda$. We observe that, by the particular realization of line search we have chosen, we have

$$\mu^{(m,k)} \boldsymbol{s}_\Lambda^{(m,k)} = \zeta^{(m,k)} \|\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_2} \frac{\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})}{\|\boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})\|_{\ell_2}} = \zeta^{(m,k)} \boldsymbol{G}(\boldsymbol{v}_\Lambda^{(m,k)})$$

for some $\zeta^{(m,k)}$ uniformly bounded from above and below (see (5.6)). Thus, we have

$$\|\boldsymbol{v}_\Lambda^{(m,k+1)}\|_{\mathcal{A}^s} \lesssim \|\boldsymbol{v}_\Lambda^{(m,k)}\|_{\mathcal{A}^s} + \|\boldsymbol{u}\|_{\mathcal{A}^s} + 1. \tag{6.13}$$

Now, we apply a recursion argument; taking into account Proposition 6.2, using again the property of geometric series and the uniform boundedness of $K^{(m)}$, from (6.8)–(6.13), we deduce the estimates

$$\begin{aligned} |\operatorname{supp} \boldsymbol{v}_\Lambda^{(m,k+1)}| + \operatorname{ops}(\boldsymbol{v}_\Lambda^{(m,k+1)}, \boldsymbol{u}^{(m)}) &\lesssim (\mathbb{E}_{m,k+1}^{1/2})^{-1/s} (\|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1) \\ &\lesssim (\nu^{e_{m+1}} \mathbb{E}_0^{1/2})^{-1/s} (\|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s} + 1), \\ \|\boldsymbol{v}_\Lambda^{(m,k+1)}\|_{\mathcal{A}^s} &\lesssim \|\boldsymbol{u}\|_{\mathcal{A}^s} + 1. \end{aligned} \tag{6.14}$$

Finally, recall that also the number of sortings needed in $\mathbf{THRESH}[\boldsymbol{v}_\Lambda^{(m,K^{(m)})}, \varepsilon_\vartheta^{(m)}]$ can be made proportional to $|\operatorname{supp} \boldsymbol{v}_\Lambda^{(m,K^{(m)})}|$ (see [2]). Shifting m into $m-1$, we get the following result.

PROPOSITION 6.4. *Under the assumptions of Proposition* 6.2 *and Assumption* 6.3, *the cardinality of the supports of all vectors involved in the computation of* $\boldsymbol{u}_\Lambda^{(m)}$ *starting from* $\boldsymbol{u}_\Lambda^{(0)}$, *as well as the complexity to compute* $\boldsymbol{u}_\Lambda^{(m)}$, *is bounded by the quantity* $C(\nu^{e_m}\mathbb{E}_0^{1/2})^{-1/s}(\|\boldsymbol{u}\|_{\mathcal{A}^s}^{1/s}+1)$. ∎

This completes the assessment of the optimality of the algorithm.

7. Examples. We conclude the paper with two concrete examples.

7.1. The nonlinear Laplacian. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with Lipschitz boundary $\partial\Omega$. Given some $p > 2$, let $\mathcal{V} = W_0^{1,p}(\Omega)$ be the closed subspace of the Sobolev space $W^{1,p}(\Omega)$ of the functions vanishing on $\partial\Omega$, equipped with the norm $\|v\|_{W_0^{1,p}(\Omega)} = \left(\sum_{i=1}^n \|\frac{\partial v}{\partial x_i}\|_{L^p(\Omega)}^p\right)^{1/p}$. Let f be an element in $\mathcal{V}' = W^{-1,p'}(\Omega)$, and let $\langle f, v\rangle_{\mathcal{V}'\times\mathcal{V}}$ denote the duality pairing between $\mathcal{V}'$ and $\mathcal{V}$.

We consider the functional $J : \mathcal{V} \to \mathbb{R}$ defined as

$$J(v) = \frac{1}{p}\|v\|_{\mathcal{V}}^p - \langle f, v\rangle_{\mathcal{V}'\times\mathcal{V}}.$$

Its Fréchet derivative is given by $J'(w) = A(w) - f$, where

$$A(w) = -\sum_{i=1}^n \frac{\partial}{\partial x_i}\left(\left|\frac{\partial w}{\partial x_i}\right|^{p-2}\frac{\partial w}{\partial x_i}\right)$$

is known as the p-Laplacian. The functional J satisfies Assumption 2.1. Indeed, condition (i) can be proven by a repeated application of Hölder's inequality in L^p-spaces, whereas condition (ii) easily follows from the existence of a constant $c_p > 0$ such that

$$(|s|^{p-2}s - |t|^{p-2}t)(s-t) \geq c_p|s-t|^p \qquad \text{for all } s,t \in \mathbb{R}. \tag{7.1}$$

Thus, there exists a unique minimizer of the functional J on $\mathcal{V}$, i.e., a unique solution of the Dirichlet problem $A(u) = f$ in Ω, $u = 0$ on $\partial\Omega$ (see, e.g., [26] for more details).

In view of the numerical discretization of such a problem, we introduce a wavelet basis $\Psi^* := \{\psi_\lambda^* : \lambda \in \mathcal{J}\}$ in $L^p(\Omega)$, such that $\|\psi_\lambda^*\|_{L^p(\Omega)} \sim 1$ for all λ. Here λ is a multi-index, containing all the relevant parameters of the wavelet, including the level index $j =: |\lambda|$. Furthermore, we denote by $B_{p,p,0}^r(\Omega)$ the closure of $\mathcal{C}_0^\infty(\Omega)$ in the Besov space $B_{p,p}^r(\Omega)$, and we assume that there exists $r^* > 1$ such that, for all r satisfying $0 < r < r^*$, Ψ^* is also a basis in $B_{p,p,0}^r(\Omega)$ and the norm equivalence

$$\|v\|_{B_{p,p,0}^r(\Omega)} \sim \left(\sum_{\lambda\in\mathcal{J}} 2^{pr|\lambda|}\,|v_\lambda^*|^p\right)^{1/p} \tag{7.2}$$

holds for all $v = \sum_{\lambda\in\mathcal{J}} v_\lambda^*\psi_\lambda^* \in B_{p,p,0}^r(\Omega)$. Examples of such bases can be found, e.g., in [5, 6, 15, 17, 18, 20].

It is well known (see, e.g., [29]) that $B^r_{p,p}(\Omega) = W^{r,p}(\Omega)$ algebraically and topologically for all $r \notin \mathbb{N}$, whereas, unfortunately, $B^k_{p,p}(\Omega) \neq W^{k,p}(\Omega)$ for $k \in \mathbb{N}$; one has only $B^{k+\varepsilon}_{p,p}(\Omega) \subset W^{k,p}(\Omega) \subset B^k_{p,p}(\Omega)$ for all $\varepsilon > 0$ with strict topological inclusions. Thus, we cannot apply the results of sections 3 and 4 taking $W^{1,p}_0(\Omega)$ as V therein.

We circumvent such a drawback by resorting to a perturbation argument. To this end, we assume that $u \in B^{1+\varepsilon_0}_{p,p,0}(\Omega)$ for some $\varepsilon_0 > 0$. Then, for any fixed $\varepsilon \in (0, \varepsilon_0]$, we set $V = V_\varepsilon = B^{1+\varepsilon}_{p,p,0}(\Omega)$ and we normalize the wavelets in the V_ε-norm; i.e., we set $\psi_\lambda = 2^{-(1+\varepsilon)|\lambda|}\,\psi^*_\lambda$ for all $\lambda \in \mathcal{J}$. Then, (7.2) yields precisely (3.1). Next, we introduce the perturbed functional $J_\varepsilon : V_\varepsilon \to \mathbb{R}$, defined as

$$J_\varepsilon(v) = \frac{\varepsilon}{p}\|\boldsymbol{v}\|^p_{\ell_p} + J(v) \qquad \text{for all } v = \boldsymbol{v}^T\Psi \in V_\varepsilon.$$

Its Fréchet derivative $J'_\varepsilon : V_\varepsilon \to V'_\varepsilon$ satisfies $\langle J'_\varepsilon(v), w\rangle_{V'_\varepsilon \times V_\varepsilon} = \varepsilon \sum_{\lambda\in\mathcal{J}} |v_\lambda|^{p-2} v_\lambda w_\lambda + \langle J'(v), w\rangle_{\mathcal{V}'\times\mathcal{V}}$ for all $v, w \in V_\varepsilon$. This functional, too, satisfies Assumption 2.1. Actually, condition (i) follows from the continuous inclusion $V_\varepsilon \subset W^{1,p}_0(\Omega)$, whereas condition (ii) is satisfied since (7.1) and the similar condition for J imply for all $w, v \in V_\varepsilon$

$$\langle J'_\varepsilon(w) - J'_\varepsilon(v), w - v\rangle_{V'_\varepsilon\times V_\varepsilon} \geq \varepsilon c_p \|\boldsymbol{w} - \boldsymbol{v}\|^p_{\ell_p} + c_J \|w - v\|^p_{W^{1,p}_0(\Omega)}. \tag{7.3}$$

Using (7.2), we obtain the existence of a constant c'_J (independent of ε) such that

$$\langle J'_\varepsilon(w) - J'_\varepsilon(v), w - v\rangle_{V'_\varepsilon\times V_\varepsilon} \geq c'_J(\varepsilon\|w - v\|^p_{B^{1+\varepsilon}_{p,p,0}(\Omega)} + \|w - v\|^p_{W^{1,p}_0(\Omega)}).$$

We can also express the ellipticity bound of the functional in terms of wavelet coefficients. Indeed, since $W^{1,p}_0(\Omega) \subset B^1_{p,p,0}(\Omega)$ with continuous inclusion, by (7.2) with $r = 1$ there exists a constant c''_J (again independent of ε) such that

$$\langle J'_\varepsilon(w) - J'_\varepsilon(v), w - v\rangle_{V'_\varepsilon\times V_\varepsilon} \geq c''_J \sum_{\lambda\in\mathcal{J}} (\varepsilon + 2^{-p\varepsilon|\lambda|})|w_\lambda - v_\lambda|^p.$$

Let $u_\varepsilon \in V_\varepsilon$ be the unique minimizer of J_ε in V_ε. Since all assumptions are fulfilled, we can apply Algorithm 4.12 **MINIMIZE** described in sections 3 and 4 to the functional J_ε.

The next result proves that u_ε can be made arbitrarily close to u by choosing ε small enough.

PROPOSITION 7.1. *The following estimate holds:*

$$\|u - u_\varepsilon\|_{W^{1,p}_0(\Omega)} \lesssim \varepsilon^{1/p}, \qquad 0 < \varepsilon \leq \varepsilon_0.$$

Proof. Set $N(v) := \frac{1}{p}\|\boldsymbol{v}\|^p_{\ell_p}$ and note that $J'_\varepsilon(u_\varepsilon) = 0$ and $J'_\varepsilon(u) = J'(u) + \varepsilon N'(u) = \varepsilon N'(u)$. Applying (7.3) with $w = u$ and $v = u_\varepsilon$, we obtain

$$\begin{aligned} c_J\|u - u_\varepsilon\|^p_{W^{1,p}_0(\Omega)} + \varepsilon c_p\|\boldsymbol{u} - \boldsymbol{u}_\varepsilon\|^p_{\ell_p} &\lesssim \langle J'_\varepsilon(u) - J'_\varepsilon(u_\varepsilon), u - u_\varepsilon\rangle_{V'_\varepsilon\times V_\varepsilon} \\ &\lesssim \varepsilon\,\langle N'(u), u - u_\varepsilon\rangle_{V'_\varepsilon\times V_\varepsilon} = \varepsilon\,\langle \boldsymbol{N}'(\boldsymbol{u}), \boldsymbol{u} - \boldsymbol{u}_\varepsilon\rangle_{\ell_{p'}\times\ell_p}. \end{aligned}$$

By Hölder's inequality,

$$\langle \boldsymbol{N}'(\boldsymbol{u}), \boldsymbol{u} - \boldsymbol{u}_\varepsilon\rangle_{\ell_{p'}\times\ell_p} \leq \|\boldsymbol{N}'(\boldsymbol{u})\|_{\ell_{p'}}\|\boldsymbol{u} - \boldsymbol{u}_\varepsilon\|_{\ell_p} \leq \frac{1}{p' c_p^{p'/p}}\|\boldsymbol{N}'(\boldsymbol{u})\|^{p'}_{\ell_{p'}} + \frac{c_p}{p}\|\boldsymbol{u} - \boldsymbol{u}_\varepsilon\|^p_{\ell_p},$$

whence the result easily follows from the assumption $u \in B^{1+\varepsilon_0}_{p,p,0}(\Omega)$. ∎

The proposition provides a guideline for selecting the perturbation parameter ε. Suppose that we wish to approximate u, in the $W_0^{1,p}(\Omega)$-norm, within an accuracy of order TOL. Then, it is enough to choose $\varepsilon = (TOL)^p$ and stop Algorithm **MINIMIZE** as soon as u_ε is itself approximated with an accuracy of order TOL.

7.2. The nonlinear reaction-diffusion problem. Let $G : \mathbb{R} \to \mathbb{R}^+$ be a smooth strictly convex function, such that $G(s) \lesssim |s|^p$ for all $s \in \mathbb{R}$, where $p \geq 2$ is chosen so that $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$, $V = H_0^1(\Omega)$. We consider the functional

$$J(v) := \frac{1}{2}\int_\Omega |\nabla v|^2 + \int_\Omega G(v) - \langle f, v\rangle_{V'\times V},$$

which is strictly convex and unbounded. The Fréchet derivative is given by

$$J'(v) = -\Delta + g(v)I - f =: A(v) - f,$$

where $g(s) := G'(s)$ is strictly monotone in $\mathbb{R}$. It is readily seen that A is continuous and strictly monotone. Moreover, A is Lipschitz continuous and one can show that the following bounds hold:

$$\|v\|_V \lesssim \|A(v)\|_{V'} \lesssim \|v\|_V^{p-1}, \quad v \in V.$$

Thus, Assumption 2.1(i) is satisfied. Moreover, due to the monotonicity of g, for all $v, w \in V$, we have $(g(v) - g(w), v - w)_{0,\Omega} \geq 0$ and thus

$$\langle J'(v) - J'(w), v - w\rangle = \|\nabla(v - w)\|_{0,\Omega}^2 + (g(v) - g(w), v - w)_{0,\Omega} \gtrsim \|v - w\|_V^2,$$

so that (ii) holds for $p = 2$.

Let us take any wavelet basis Ψ of $H_0^1(\Omega)$ (see [5, 6, 15, 17, 18, 20]), so that the following norm equivalence holds:

$$\|v\|_V \sim \Big(\sum_{\lambda\in\mathcal{J}} 2^{2|\lambda|}|v_\lambda|^2\Big)^{1/2}, \qquad v = \sum_{\lambda\in\mathcal{J}} v_\lambda\,\psi_\lambda.$$

If the nonlinear function G is a global polynomial, then all required ingredients can be taken from the literature. Indeed, **EVAL-GRAD** and **EVAL-J** consist of a linear part which is described in [10, 11] and a nonlinearity of polynomial type. A corresponding evaluation scheme can be found in [13, 19]. In addition, a tree-coarsening routine is available in [12]. Finally, the gradient consists of the Laplacian and, again, a nonlinearity of polynomial type which can be realized using the methods described in [10, 11, 13, 19]. Moreover, note that if in addition the Fréchet derivative J'' of J' is well-posed in the sense that $\|J''(v)w\|_{V'} \sim \|w\|_V$ for all v in a neighborhood of the solution u of $J'(u) = 0$, then one can also use the adaptive wavelet method presented in [12] for the Euler–Lagrange equation in order to determine the minimizer u of J.

However, the function G may have a much more complicated structure, e.g., it may be defined piecewise. In this case, our general approach can also be used, provided corresponding routines for **EVAL-J**, **EVAL-GRAD**, and **THRESH** are available. Constructing and analyzing such schemes will be a subject of future research.

Appendix A. Proof of Lemma 2.2. As for Lemma 2.2(a), the proof is straightforward setting $w := u + t(v-u)$, i.e., $v - u = \frac{1}{t}(w-u)$, and using (ii) in Assumption 2.1 to obtain

$$\begin{aligned}
J(v) - J(u) &= \int_0^1 \frac{d}{dt} J(u + t(v-u))\, dt \\
&= \int_0^1 \langle J'(u + t(v-u)), v-u\rangle\, dt \\
&= \langle J'(u), v-u\rangle + \int_0^1 \langle J'(u + t(v-u)) - J'(u), v-u\rangle\, dt \\
&\geq \langle J'(u), v-u\rangle + c_J \int_0^1 \frac{1}{t} \|w-u\|_V^p\, dt \\
&= \langle J'(u), v-u\rangle + c_J \int_0^1 \frac{1}{t} \|t(u-v)\|_V^p\, dt \\
&= \langle J'(u), v-u\rangle + c_J \int_0^1 t^{p-1} \|u-v\|_V^p\, dt \\
&= \langle J'(u), v-u\rangle + \frac{c_J}{p} \|u-v\|_V^p.
\end{aligned}$$

Now we prove (b). Indeed,

$$J(v) \geq J(u) + \langle J'(u), v-u\rangle + \frac{c_J}{p}\|v-u\|^p;$$

hence, $J(v) > J(u) + \langle J'(u), v-u\rangle$ for all $u \neq v \in V$. For the boundedness, take again $u = 0$ in (a); then

$$J(v) \geq J(0) + \langle J'(0), v\rangle + \frac{c_J}{p}\|v\|_V^p \geq J(0) - \|J'(0)\|_{V'}\, \|v\|_V + \frac{c_J}{p}\|v\|_V^p,$$

which is bounded from below by the constant $J(0) - \|J'(0)\|_{V'}$.

As for (c), take $u = 0$ in (a). Then, for $v \in \mathcal{R}(u^{(0)})$, we have

$$\frac{c_J}{p}\|v\|_V^p \leq J(v) - J(0) - \langle J'(0), v\rangle \leq |J(u^{(0)})| + |J(0)| + \|J'(0)\|_{V'}\, \|v\|_V.$$

Since $p > 1$, we have $\|v\|_V \leq C(p, u^{(0)})$. ∎

Acknowledgment. The authors wish to thank one of the referees for suggesting significant improvements over the original version of the paper.

REFERENCES

[1] J.-P. Aubin, *Optima and Equilibria: An Introduction to Nonlinear Analysis*, Springer-Verlag, Berlin, 1993.

[2] A. Barinka, *Fast Computational Tools for Adaptive Wavelet Schemes*, Ph.D. thesis, RWTH Aachen, Aachen, Germany, in preparation.

[3] P. Binev, W. Dahmen, and R. A. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.

[4] P. Binev and R. A. DeVore, *Fast computation in adaptive tree approximation*, Numer. Math., 97 (2004), pp. 193–217.

[5] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method, Part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 1–52.

[6] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method, Part* II: *Realization and additional features in* 2*D and* 3*D*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 123–165.
[7] J. Céa, *Optimisation: Théorie et Algorithmes*, Dunod, Paris, 1971.
[8] P. G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimization*, Cambridge University Press, Cambridge, UK, 1988.
[9] A. Cohen, *Wavelet methods in numerical analysis*, in Handbook of Numerical Analysis, Vol. VII, Handb. Numer. Anal. VII, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 417–711.
[10] A. Cohen, W. Dahmen, and R. A. DeVore, *Adaptive wavelet schemes for elliptic operator equations: Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.
[11] A. Cohen, W. Dahmen, and R. A. DeVore, *Adaptive wavelet methods.* II. *Beyond the elliptic case*, Found. Comput. Math., 2 (2002), pp. 203–245.
[12] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet schemes for nonlinear variational problems*, SIAM J. Numer. Anal., 41 (2003), pp. 1785–1823.
[13] A. Cohen, W. Dahmen, and R. DeVore, *Sparse evaluation of compositions of functions using multiscale expansions*, SIAM J. Math. Anal., 35 (2003), pp. 279–303.
[14] A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore, *Tree approximation and optimal encoding*, Appl. Comput. Harmon. Anal., 11 (2001), pp. 192–226.
[15] A. Cohen and R. Masson, *Wavelet adaptive methods for second order elliptic problems—boundary conditions and domain decomposition*, Numer. Math., 86 (2000), pp. 193–238.
[16] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228.
[17] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., 68 (1999), pp. 1533–1567.
[18] W. Dahmen and R. Schneider, *Wavelets on manifolds* I: *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.
[19] W. Dahmen, R. Schneider, and Y. Xu, *Nonlinear functionals of wavelet expansions—adaptive reconstruction and fast evaluation*, Numer. Math., 86 (2000), pp. 49–101.
[20] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.
[21] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall Ser. Comput. Math., Prentice–Hall, Englewood Cliffs, NJ, 1983.
[22] R. A. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.
[23] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
[24] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North–Holland, Amsterdam, 1976.
[25] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.
[26] J.-L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Paris, 1969.
[27] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
[28] E. Polak, *Optimization. Algorithms and Consistent Approximations*, Appl. Math. Sci. 124, Springer-Verlag, New York, 1997.
[29] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North–Holland, Amsterdam, 1978.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2076–2097

EFFICIENT AND SAFE GLOBAL CONSTRAINTS FOR HANDLING NUMERICAL CONSTRAINT SYSTEMS*

YAHIA LEBBAH[†¶], CLAUDE MICHEL[‡¶], MICHEL RUEHER[‡¶], DAVID DANEY[§¶], AND JEAN-PIERRE MERLET[§¶]

Abstract. Numerical constraint systems are often handled by branch and prune algorithms that combine splitting techniques, local consistencies, and interval methods. This paper first recalls the principles of `Quad`, a global constraint that works on a tight and safe linear relaxation of quadratic subsystems of constraints. Then, it introduces a generalization of `Quad` to polynomial constraint systems. It also introduces a method to get safe linear relaxations and shows how to compute safe bounds of the variables of the linear constraint system. Different linearization techniques are investigated to limit the number of generated constraints. `QuadSolver`, a new branch and prune algorithm that combines `Quad`, local consistencies, and interval methods, is introduced. `QuadSolver` has been evaluated on a variety of benchmarks from kinematics, mechanics, and robotics. On these benchmarks, it outperforms classical interval methods as well as constraint satisfaction problem solvers and it compares well with state-of-the-art optimization solvers.

Key words. systems of equations and inequalities, constraint programming, reformulation linearization technique, global constraint, interval arithmetic, safe rounding

AMS subject classifications. 65H10, 65G40, 65H20, 93B18, 65G20

DOI. 10.1137/S0036142903436174

1. Introduction.

Many applications in engineering sciences require finding all isolated solutions to systems of constraints over real numbers. These systems may be nonpolynomial and are difficult to solve: the inherent computational complexity is NP-hard and numerical issues are critical in practice (e.g., it is far from being obvious to guarantee correctness and completeness as well as to ensure termination). These systems, called numerical CSP (constraint satisfaction problem) in the rest of this paper, have been approached in the past by different interesting methods:[1] interval methods [35, 24, 38, 20, 40], continuation methods [37, 2, 62], and constraint satisfaction methods [30, 6, 11, 61]. Of particular interest is the mathematical and programming simplicity of the latter approach: the general framework is a branch and prune algorithm that requires only specifying the constraints and the initial range of the variables.

The purpose of this paper is to introduce and study a new branch and bound algorithm called `QuadSolver`. The essential feature of this algorithm is a global constraint—called `Quad`—that works on a tight and safe linear relaxation of the polynomial relations of the constraint systems. More precisely, `QuadSolver` is a branch

*Received by the editors October 15, 2003; accepted for publication (in revised form) June 3, 2004; published electronically February 25, 2005.
http://www.siam.org/journals/sinum/42-5/43617.html
[†]Département Informatique, Faculté des Sciences, Université d'Oran Es-Senia, BP 1524, El-M'Naouar Oran, Algeria (Yahia.Lebbah@sophia.inria.fr).
[‡]Université de Nice-Sophia Antipolis, I3S-CNRS, 930 route des Colles, BP 145, 06903 Sophia Antipolis Cedex, France (Claude.Michel@sophia.inria.fr, rueher@essi.fr).
[§]INRIA, 2004 route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France (David.Daney@sophia.inria.fr, Jean-Pierre.Merlet@sophia.inria.fr).
[¶]COPRIN Project INRIA–I3S–CNRS, 2004 route des Luciales, BP 93, Sophia Antipolis Cedex 06903, France.
[1]Alternative methods have been proposed for solving nonlinear systems. For instance, algebraic constraints can be handled with symbolic methods [13] (e.g., Groebner basis, resultant). However, these methods can neither handle nonpolynomial systems nor deal with inequalities.

and prune algorithm that combines `Quad`, local consistencies, and interval methods. That is to say, `QuadSolver` is an attempt to merge the best interval and constraint programming techniques. `QuadSolver` has been evaluated on a variety of benchmarks from kinematics, mechanics, and robotics. On these benchmarks, it outperforms classical interval methods as well as CSP solvers and it compares well with state-of-the-art optimization solvers.

The `Quad`-filtering algorithm [27] has first been defined for quadratic constraints. The relaxation of quadratic terms is adapted from a classical linearization method, the reformulation-linearization technique (RLT) [54, 53]. The simplex algorithm is used to narrow the domain of each variable with respect to the subset of the linear set of constraints generated by the relaxation process. The coefficients of these linear constraints are updated with the new values of the bounds of the domains and the process is restarted until no more significant reduction can be done. We have demonstrated [27] that the `Quad` algorithm yields a more effective pruning of the domains than local consistency filtering algorithms (e.g., 2b-consistency [30] or box-consistency [6]). Indeed, the drawback of classical local consistencies comes from the fact that the constraints are handled independently and in a blind way.[2] That is to say, classical local consistencies do not exploit the semantic of quadratic terms; in other words, these approaches do not take advantage of the very specific semantic of quadratic constraints to reduce the domains of the variables. Conversely, linear programming techniques [1, 54, 3] do capture most of the semantics of quadratic terms, e.g., convex and concave envelopes of these particular terms.[3]

The extension of `Quad` for handling any polynomial constraint system requires replacing nonquadratic terms by new variables and adding the corresponding identities to the initial constraint system. However, a complete quadrification [58] would generate a huge number of linear constraints. Thus, we introduce here an heuristic based on a good tradeoff between a tight approximation of the nonlinear terms and the size of the generated constraint system. This heuristic works well on classical benchmarks (see section 8).

A safe rounding process is a key issue for the `Quad` framework. Let us recall that the simplex algorithm is used to narrow the domain of each variable with respect to the subset of the linear set of constraints generated by the relaxation process. The point is that most implementations of the simplex algorithm are unsafe. Moreover, the coefficients of the generated linear constraints are computed with floating point numbers. So, two problems may occur in the `Quad`-filtering process.

1. The whole linearization may become incorrect due to rounding errors when computing the coefficients of the generated linear constraints.
2. Some solutions may be lost when computing the bounds of the domains of the variables with the simplex algorithm.

We propose in this paper a safe procedure for computing the coefficients of the generated linear constraints. Neumaier and Shcherbina [42] have addressed the second

[2] 3b-consistency and kb-consistency are partial consistencies that can achieve a better pruning since they are "less local" [11]. However, they require numerous splitting steps to find the solutions of a system of quadratic constraints; so, they may become rather slow.

[3] Sherali and Tuncbilek [55] have also proposed four different filtering techniques for solving quadratic problems. Roughly speaking, the first filtering strategy performs a feasibility check on inequality constraints to discard subintervals of the domains of the variables. This strategy is very close to box-consistency filtering (see [60]). The three other techniques are based on specific properties of optimization problems with a quadratic objective function: the eigenstructure of the quadratic objective function, fathoming node, and Lagrangian dual problem. Thus, these techniques can be considered as local consistencies for optimization problems (see also [59] and Neumaier's survey [41]).

problem.[4] They have proposed a simple and cheap procedure to get a rigorous upper bound of the objective function. The incorporation of these procedures in the Quad-filtering process allows us to call the simplex algorithm without worrying about safety. So, with these two procedures, linear programming techniques can be used to tackle continuous CSPs *without losing any solution*.

The rest of this paper is organized as follows. Section 2 gives an overview of the approach whereas section 3 contains the notation. Sections 4 and 5 recall the basics of interval programming and constraint programming. Section 6 details the principle of the Quad algorithm, the linearization process, and the extension to polynomial constraints. Section 7 introduces the rounding process we propose to ensure the safe relaxations. Section 8 describes the experimental results and discusses related work. Concluding remarks are given in section 9.

2. Overview of the approach. As mentioned, QuadSolver is a branch and prune algorithm that combines Quad and a box-consistency.

Box-consistency is the most successful adaptation of arc-consistency [31] to constraints over the real numbers. The box-consistency implementation of Van-Hentenryck, McAllester, and Kapur [60] is computed on three-interval extensions of the initial constraints: the natural interval extension, the distributed interval extension, and the Taylor interval extension with a conditioning step. The leftmost and the rightmost zeros are computed using a variation of the univariate interval Newton method.

The QuadSolver we propose here combines Quad-filtering and box-consistency filtering to prune the domain of the variables of numerical constraint systems. Operationally, QuadSolver performs the following filtering processes:

1. box-consistency filtering,
2. Quad-filtering.

The box-consistency is first used to detect some inconsistencies before starting the Quad-filtering algorithm which is more costly. These two steps are wrapped into a classical fixed point algorithm which stops when the domains of the variables cannot be further reduced.[5]

To isolate the different solutions, Quad uses classical branching techniques.

Before going into the details, let us outline the advantages of our approach on a couple of small examples.

2.1. Quad-filtering. Consider the constraint system $\mathcal{C} = \{2xy+y = 1, xy = 0.2\}$ which represents two intersecting curves (see Figure 2.1). Suppose that $\mathbf{x} = [-10, +10]$ and $\mathbf{y} = [-10, +10]$ are the domains of the variables x and y. An interval $\mathbf{x} = [\underline{x}, \overline{x}]$ denotes the set of reals $\{r | \underline{x} \leq r \leq \overline{x}\}$.

The RLT (see section 6.2) yields the following constraint system:

$$\text{(a)} \qquad \begin{cases} y + 2w = 1, \quad w = 0.2, \\ \underline{y}x + \underline{x}y - w \leq \underline{x}\underline{y}, \quad \overline{y}x + \underline{x}y - w \geq \underline{x}\overline{y}, \\ \underline{y}x + \overline{x}y - w \geq \overline{x}\underline{y}, \quad \overline{y}x + \overline{x}y - w \leq \overline{x}\overline{y}, \\ x \geq \underline{x}, \ x \leq \overline{x}, \quad y \geq \underline{y}, \ y \leq \overline{y}, \end{cases}$$

where w is a new variable that stands for the product xy. Note that constraint system (a) implies that $w \in [\underline{x}, \overline{x}] * [\underline{y}, \overline{y}]$.

[4] They have also suggested a solution to the first problem though their solution is dedicated to mixed integer programming problems.

[5] In practice, the loop stops when the domain reduction is lower than a given ϵ.

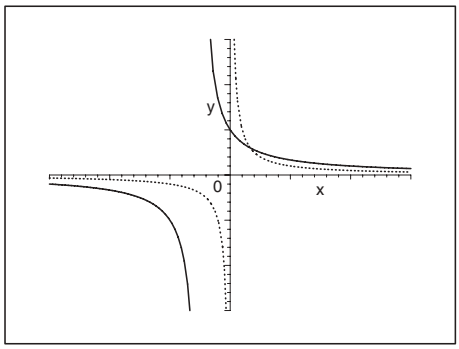


FIG. 2.1. *Geometrical representation of* $\{2xy + y = 1, xy = 0.2\}$.

Substituting $\underline{x}$, $\underline{y}$, $\overline{x}$, and $\overline{y}$ by their values and minimizing (resp., maximizing) x, y, and w with the simplex algorithm yield the following new bounds:

$$\mathbf{x} = [-9.38, 9.42], \quad \mathbf{y} = [0.6, 0.6], \quad \mathbf{w} = [0.2, 0.2].$$

By substituting the new bounds of x, y, and w in the constraint system (a), we obtain a new linear constraint system. One more minimizing (resp., maximizing) step is required to obtain tight bounds of $\mathbf{x}$. Note that numerous splitting operations are required to find the unique solution of the problem with a 3b-consistency filtering algorithm. The proposed algorithm solves the problem by generating 6 linear constraints and with 8 calls to the simplex algorithm. It finds the same solution as a solver based on 3b-consistency but without splitting and in less time.

2.2. A safe rounding procedure. Consider the constraint system

$$\mathcal{C} = \begin{cases} w_1 + w_2 = 1, \quad w_1 x_1 + w_2 x_2 = 0, \\ w_1 x_1 x_1 + w_2 x_2 x_2 = 1, \quad w_1 x_1 x_1 x_1 + w_2 x_2 x_2 x_2 = 0, \end{cases}$$

which represents a simple Gaussian quadrature formula to compute integrals [9]. Suppose that the domains of variables x_1, x_2, w_1, and w_2 are all equal to $[-1, +1]$. This system has two solutions:

- $x_1 = -1,\ x_2 = 1,\ w_1 = 0.5,\ w_2 = 0.5,$
- $x_1 = 1,\ x_2 = -1,\ w_1 = 0.5,\ w_2 = 0.5.$

A straightforward implementation of `Quad` would only find one unsafe solution with

$$x_2 \in [+0.9999\ldots944, +0.9999\ldots989].$$

Indeed, when we examine the `Quad`-filtering process, we can identify some linear programs where the simplex algorithm steps to the wrong side of the objective.

With the corrections we propose in section 7, we obtain a tight approximation of the two correct solutions (with $x_2 \in [-1.000000\ldots, -0.999999\ldots]$ and $x_2 \in [0.999999\ldots, 1.000000\ldots]$).

3. Notation and basic definitions. This paper focuses on CSPs where the domains are intervals and the constraints $C_j(x_1, \ldots, x_n)$ are n-ary relations over the reals. $\mathcal{C}$ stands for the set of constraints.

$\mathbf{x}$ or D_x denotes the domain of variable x, that is to say, the set of allowed values for x. $\mathcal{D}$ stands for the set of domains of all the variables of the considered constraint

system. $\mathbb{R}$ denotes the set of real numbers whereas $\mathbb{F}$ stands for the set of floating point numbers used in the implementation of nonlinear constraint solvers; if a is a constant in $\mathbb{F}$, a^+ (resp., a^-) corresponds to the smallest (resp., largest) number of $\mathbb{F}$ strictly greater (resp., lower) than a.

$\mathbf{x} = [\underline{x}, \overline{x}]$ is defined as the set of real numbers x verifying $\underline{x} \leq x \leq \overline{x}$. x, y denote real variables, X, Y denote vectors whereas $\mathbf{X}, \mathbf{Y}$ denote interval vectors. The *width* $w(\mathbf{x})$ of an interval $\mathbf{x}$ is the quantity $\overline{x} - \underline{x}$ while the *midpoint* $m(\mathbf{x})$ of the interval $\mathbf{x}$ is $(\overline{x} + \underline{x})/2$. A *point interval* $\mathbf{x}$ is obtained if $\underline{x} = \overline{x}$. A *box* is a set of intervals: its width is defined as the largest width of its interval members, while its center is defined as the point whose coordinates is the midpoint of the ranges. $\mathbb{IR}^n$ denotes the set of boxes and is ordered by set inclusion.

We use the RLT notation introduced in [54, 3] with slight modifications. More precisely, we will use the following notations: $[c]_L$ is the set of linear constraints generated by replacing the nonlinear terms by new variables in constraint c, and $[c]_{LI}$ denotes the set of equations that keep the link between the new variables and the nonlinear terms while $[c]_R$ contains linear inequalities that approximate the semantics of nonlinear terms of constraint c. These notations will be used indifferently whether c is a constraint or $\mathcal{C}$ is a set of constraints.

Rounding is necessary to close the operations over $\mathbb{F}$ (see [18]). A rounding function maps the result of the evaluation of an expression to available floating-point numbers. Rounding x towards $+\infty$ maps x to the least floating point number x_f such that $x \leq x_f$. $\bigtriangledown(x)$ (resp., $\bigtriangleup(x)$) denotes a rounding mode of x towards $-\infty$ (resp., $+\infty$).

4. Interval programming.

This section recalls the basic concepts of interval arithmetic that are required to understand the rest of the paper. Readers familiar with interval arithmetic may skip this section.

4.1. Interval arithmetic.

Interval arithmetic has been introduced by Moore [35]. It is based on the representation of variables as intervals.

Let f be a real-valued function of n unknowns $X = (x_1, \ldots, x_n)$. An *interval evaluation* of f for given ranges $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ for the unknowns is an interval $\mathbf{y}$ such that

$$\underline{y} \leq f(\mathbf{X}) \leq \overline{y} \quad \text{for all} \quad X = (x_1, \ldots, x_n) \in \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n). \tag{4.1}$$

In other words, $\underline{y}$ and $\overline{y}$ are lower and upper bounds for the values of f when the values of the unknowns are restricted to the box $\mathbf{X}$.

There are numerous ways to calculate an interval evaluation of a function [20, 46]. The simplest is the *natural evaluation* in which all the mathematical operators in f are substituted by their interval equivalents. Interval equivalents exist for all classical mathematical operators. Hence interval arithmetic allows us to calculate an interval evaluation for all nonlinear expressions, whether algebraic or not. For example, if $f(x) = x + \sin(x)$, then the interval evaluation of f for $x \in [1.1, 2]$ can be calculated as follows:

$$f([1.1, 2]) = [1.1, 2] + \sin([1.1, 2]) = [1.1, 2] + [0.8912, 1] = [1.9912, 3].$$

Interval arithmetic can be implemented with directed rounding to take into account round-off errors. There are numerous interval arithmetic packages implementing this property: one of the most famous library is BIAS/Profil,[6] but a promising new

[6] http://www.ti3.tu-harburg.de/Software/PROFILEnglisch.html.

package—based on the multiprecision software MPFR[7]—is MPFI [47].

The main limitation of interval arithmetic is *the overestimation of interval functions*. This is due to two well-known problems:

- the so-called *wrapping effect* [35, 39], which overestimates by a unique vector the image of an interval vector (which is in general not a vector). That is to say, $\{f(X)|X \in \mathbf{X}\}$ is contained in $f(\mathbf{X})$ but is usually not equal to $f(\mathbf{X})$;
- the so-called *dependency problem* [20], which is due to the independence of the different occurrences of some variables during the interval evaluation of an expression. In other words, during the interval evaluation process there is no correlation between the different occurrences of a same variable in an equation. For instance, consider $\mathbf{x} = [0, 10]$. $\mathbf{x} - \mathbf{x} = [\underline{x} - \overline{x}, \overline{x} - \underline{x}] = [-10, 10]$ instead of $[0, 0]$ as one could expect.

In general, it is not possible to compute the exact enclosure of the range for an arbitrary function over the real numbers [25]. Thus, Moore introduced the concept of *interval extension*: the interval extension of a function is an interval function that computes outer approximations on the range of the function over a domain [20, 36]. Two main extensions have been introduced: the natural extension and the Taylor extension [46, 20, 38].[8] Due to the properties of interval arithmetic, the evaluation of a function may yield different results according to the literal form of the equations. Thus, many literal forms may be used as, for example, factorized form (Horner for polynomial system) or distributed form [60].

Nevertheless, in general, neither the natural form nor the Taylor expansion allows us to compute the exact range of a function f. For instance, considering $f(x) = 1 - x + x^2$ and $\mathbf{x} = [0, 2]$, we have

$$\begin{aligned} f_{\text{tay}}([0,2]) &= f(x) + (2\mathbf{x} - 1)(\mathbf{x} - x) = f(1) + (2[0,2] - 1)([0,2] - 1) = [-2, 4], \\ f([0,2]) &= 1 - \mathbf{x} + \mathbf{x}^2 = 1 - [0,2] + [0,2]^2 = [-1, 5], \\ f_{\text{factor}}([0,2]) &= 1 + \mathbf{x}(\mathbf{x} - 1) = 1 + [0,2]([0,2] - 1) = [-1, 3], \end{aligned} \tag{4.2}$$

whereas the range of f over $X = [0, 2]$ is $[3/4, 3]$. In this case, this result could directly be obtained by a second form of factorization: $f_{\text{factor}_2}([0, 2]) = (\mathbf{x} - 1/2)^2 + 3/4 = ([0, 2] - 1/2)^2 + 3/4 = [3/4, 3]$.

4.2. Interval analysis methods. This section provides a short introduction to interval analysis methods (see [35, 20, 38, 40] for a more detailed introduction). We limit this overview to interval Newton-like methods for solving a multivariate system of nonlinear equations. Their use is complementary to methods provided by the constraint programming community.

The aim is to determine the zeros of a system of n equations $f_i(x_1, \ldots, x_n)$ in n unknowns x_i inside the interval vector $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with $x_i \in \mathbf{x}_i$ for $i = 1, \ldots, n$.

First, consider solving an interval linear system of equations defined as follows:

$$AX = b, \quad A \in \mathbf{A}, b \in \mathbf{b}, \tag{4.3}$$

where $\mathbf{A}$ is an interval matrix and $\mathbf{b}$ is an interval vector. Solving this linear interval system requires us to determine an interval vector $\mathbf{X}$ containing all solutions of all scalar linear systems noted $AX = b$ such that $A \in \mathbf{A}$ and $b \in \mathbf{b}$. Finding the exact value of $\mathbf{X}$ is a difficult problem, but three basic interval methods exist: Gaussian

[7] http://www.mpfr.org.

[8] $f_{\text{tay}}(\mathbf{X}) = f(X) + \mathbf{A}(\mathbf{X} - X)$, where $\mathbf{A}$ is the Jacobian or the interval slope matrix.

elimination, Gauss–Seidel iterative method, or Krawczyk method (see [24, 38, 20, 40]). They may provide an overestimated interval vector $\mathbf{X}_1$ including $\mathbf{X}$. However, in general the computed intervals are too wide and a preconditioning is required, that is to say, a multiplication of both sides of (4.3) by the inverse of a midpoint of $\mathbf{A}$. The matrix $m(\mathbf{A})^{-1}\mathbf{A}$ is then "closer" to the identity matrix and the width of $\mathbf{X}_1$ is smaller [20].

To solve nonlinear systems, an interval Newton algorithm is often used—see [20] or [38]. The basic idea is to solve iteratively a linear approximation of the nonlinear system obtained by a Taylor expansion. Many improvements [24, 19], based on variations of the resolution of the linear subsystem or the preconditioning, have been proposed. Note that many interesting properties are provided by Newton-like methods: existence and/or uniqueness of a root, convergence area/rate, ...

5. Constraint programming. This section recalls the basics of constraint programming techniques which are required to understand the rest of this paper. A detailed discussion of these concepts and techniques can be found in [6, 26].

5.1. The general framework. The constraint programming framework is based on a branch and prune scheme which was inspired by the traditional branch and bound approach used in optimization problems. That is to say, it is best viewed as an iteration of two steps [60]:

1. pruning the search space;
2. making a choice to generate two (or more) subproblems.

The pruning step ensures that some local consistency holds. In other words, the pruning step reduces an interval when it can prove that the upper bound or the lower bound does not satisfy some constraint. Informally speaking, a constraint system $\mathcal{C}$ satisfies a partial consistency property if a relaxation of $\mathcal{C}$ is consistent. For instance consider $\mathbf{x} = [\underline{x}, \overline{x}]$ and $c(x, x_1, \ldots, x_n) \in \mathcal{C}$. Whenever $c(x, x_1, \ldots, x_n)$ does not hold for any values $a \in \mathbf{x} = [\underline{x}, x']$, then $\mathbf{x}$ may be shrunk to $\mathbf{x} = [x', \overline{x}]$. Local consistencies are detailed in the next subsection. Roughly speaking, they are relaxations of arc-consistency, a notion that is well known in artificial intelligence [31, 34].

The branching step usually splits the interval associated to some variable in two intervals with the same width. However, the splitting process may generate more than two subproblems and one may split an interval at a point different from its midpoint. The choice of the variable to split is a critical issue in difficult problems. Sophisticated splitting strategies have been developed for finite domains but few results [23] are available for continuous domains.

5.2. Local consistencies [11, 26]. Local consistencies are conditions that filtering algorithms must satisfy. A filtering algorithm can be seen as a fixed point algorithm defined by the sequence $\{\mathcal{D}_k\}$ of domains generated by the iterative application of an operator $Op : \mathbb{IR}^n \longrightarrow \mathbb{IR}^n$ (see Figure 5.1).

$$\mathcal{D}_k = \begin{cases} \mathcal{D} & \text{if } k = 0 \\ Op(\mathcal{D}_{k-1}) & \text{if } k > 0 \end{cases}$$

FIG. 5.1. *Filtering algorithms as fixed point algorithms.*

The operator Op of a filtering algorithm generally satisfies the following three properties:

- $Op(\mathcal{D}) \subseteq \mathcal{D}$ (contractance);
- Op is conservative; that is, it cannot remove any solution;

- $\mathcal{D}' \subseteq \mathcal{D} \Rightarrow Op(\mathcal{D}') \subseteq Op(\mathcal{D})$ (monotonicity).

Under those conditions, the limit of the sequence $\{\mathcal{D}_k\}$, which corresponds to the greatest fixed point of the operator Op, exists and is called a *closure*. A fixed point for Op may be characterized by an lc-consistency property, called a local consistency. The algorithm achieving filtering by lc-consistency is denoted lc-filtering. A CSP is said to be lc-satisfiable if lc-filtering of this CSP does not produce an empty domain.

Consistencies used in numerical CSP solvers can be categorized in two main classes: *arc-consistency*-like consistencies and strong consistencies. Strong consistencies will not be discussed in this paper (see [30, 26] for a detailed introduction).

Most of the numerical CSP systems (for example, BNR-prolog [43], Interlog [8], CLP(BNR) [7], PrologIV [12], UniCalc [4], Ilog Solver [22], Numerica [61], and RealPaver [5]) compute an approximation of arc-consistency [31] which will be named ac-like-consistency in this paper. An *ac*-like-consistency states a local property on a constraint and on the bounds of the domains of its variables. Roughly speaking, a constraint c_j is ac-like-consistent if for any variable x_i in $var(c_j)$, the bounds $\underline{x}_i$ and $\overline{x}_i$ have a support in the domains of all other variables of c_j.

The most famous ac-like consistencies are 2b-consistency and box-consistency.

2b-consistency (also known as hull consistency) [10, 7, 28, 30] requires only to check the arc-consistency property for each bound of the intervals. The key point is that this relaxation is more easily verifiable than arc-consistency itself. Informally speaking, variable x is 2b-consistent for constraint "$f(x, x_1, \ldots, x_n) = 0$" if the lower (resp., upper) bound of the domain of x is the smallest (resp., largest) solution of $f(x, x_1, \ldots, x_n)$. The box-consistency [6, 21] is a coarser relaxation (i.e., it allows less stringent pruning) of arc-consistency than 2b-consistency. Variable x is box-consistent for constraint "$f(x, x_1, \ldots, x_n) = 0$" if the bounds of the domain of x correspond to the leftmost and rightmost zeros of the optimal interval extension of $f(x, x_1, \ldots, x_n)$. 2b-consistency algorithms actually achieve a weaker filtering (i.e., a filtering that yields bigger intervals) than box-consistency, more precisely when a variable occurs more than once in some constraint (see Proposition 6 in [11]). This is due to the fact that 2b-consistency algorithms require a decomposition of the constraints with multiple occurrences of the same variable.

2b-consistency [30] states a local property on the bounds of the domains of a variable at a single constraint level. A constraint c is 2b-consistent if, for any variable x, there exist values in the domains of all other variables which satisfy c when x is fixed to $\underline{x}$ and $\overline{x}$.

The filtering by 2b-consistency of $P = (\mathcal{D}, \mathcal{C})$ is the CSP $P' = (\mathcal{D}', \mathcal{C})$ such that

- P and P' have the same solutions;
- P' is 2b-consistent;
- $\mathcal{D}' \subseteq \mathcal{D}$ and the domains in $\mathcal{D}'$ are the largest ones for which P' is 2b-consistent.

Filtering by 2b-consistency of P always exists and is unique [30], that is to say it is a closure.

The box-consistency [6, 21] is a coarser relaxation of arc-consistency than 2b-consistency. It mainly consists of replacing every existentially quantified variable but one with its interval in the definition of 2b-consistency. Thus, box-consistency generates a system of univariate interval functions which can be tackled by numerical methods such as interval Newton. In contrast to 2b-consistency, box-consistency does not require any constraint decomposition and thus does not amplify the locality problem. Moreover, box-consistency can tackle some dependency problems when each

constraint of a CSP contains only one variable which has multiple occurrences. More formally we have the following definition.

DEFINITION 5.1 (box-consistency). *Let $(\mathcal{D}, \mathcal{C})$ be a CSP and $c \in \mathcal{C}$ a k-ary constraint over the variables $(x_1, \ldots, x_k)$. c is box-consistent if, for all x_i, the following relations hold:*

1. $c(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, [\underline{x}_i, \underline{x}_i^+), \mathbf{x}_{i+1}, \ldots, \mathbf{x}_k)$,
2. $c(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, (\overline{x}_i^-, \overline{x}_i], \mathbf{x}_{i+1}, \ldots, \mathbf{x}_k)$.

Closure by box-consistency of P is defined similarly as closure by 2b-consistency of P.

Benhamou et al. have introduced HC4 [5], an ac-like-consistency that merges 2b-consistency and box-consistency and which optimizes the computation process.

6. `Quad` basics and extensions. This section first introduces `Quad`, a global constraint that works on a tight and safe linear relaxation of quadratic subsystems of constraints. Then, it generalizes `Quad` to the polynomial part of numerical constraint systems. Different linearization techniques are investigated to limit the number of generated constraints.

6.1. The `Quad` algorithm. The `Quad`-filtering algorithm (see Algorithm 1) consists of three main steps: reformulation, linearization, and pruning.

The reformulation step generates $[\mathcal{C}]_R$, the set of implied linear constraints. More precisely, $[\mathcal{C}]_R$ contains linear inequalities that approximate the semantics of nonlinear terms of $\mathcal{C}$.

The linearization process first decomposes each nonlinear term in sums and products of univariate terms; then it replaces nonlinear terms with their associated new variables. For example, considering constraint $c : x_2x_3x_4^2(x_6 + x_7) + \sin(x_1)(x_2x_6 - x_3) = 0$, a simple linearization transformation may yield the following sets:

- $[c]_L = \{y_1 + y_3 = 0, \quad y_2 = x_6 + x_7, \quad y_4 = y_5 - x_3\}$,
- $[c]_{LI} = \{y_1 = x_2x_3x_4^2y_2, \quad y_3 = \sin(x_1)y_4, \quad y_5 = x_2x_6\}$.

$[c]_L$ is the set of linear constraints generated by replacing the nonlinear terms by new variables and $[c]_{LI}$ denotes the set of equations that keep the link between the new variables and the nonlinear terms. Note that the nonlinear terms which are not directly handled by the `Quad` are taken into account by the box-filtering process.

Finally, the linearization step computes the set of final linear inequalities and equations $LR = [\mathcal{C}]_L \cup [\mathcal{C}]_R$, the linear relaxation of the original constraints $\mathcal{C}$.

The pruning step is just a fixed point algorithm that calls iteratively a linear programming solver to reduce the upper and lower bounds of every original variable. The algorithm converges and terminates if ϵ is greater than zero.

Now we are in the position to introduce the reformulation of nonlinear terms. Section 6.2 first introduces the handling of quadratic constraints while section 6.3 extends the previous results to polynomial constraints.

6.2. Handling quadratic constraints. Quadratic constraints are approximated by linear constraints in the following way. `Quad` creates a new variable for each quadratic term: y for x^2 and $y_{i,j}$ for x_ix_j. The produced system is denoted as

$$\left[\sum_{(i,j)\in M} a_{k,i,j}x_ix_j + \sum_{i\in N} b_{k,i}x_i^2 + \sum_{i\in N} d_{k,i}x_i = b_k \right]_L .$$

.

Function Quad_filtering(IN: $\mathcal{X}$, $\mathcal{D}$, $\mathcal{C}$, ϵ) **return** $\mathcal{D}'$
% $\mathcal{X}$: initial variables; $\mathcal{D}$: input domains; $\mathcal{C}$: constraints; ϵ: minimal reduction
% $\mathcal{D}'$: output domains

1. *Reformulation*: generation of linear inequalities $[\mathcal{C}]_R$ for the nonlinear terms in $\mathcal{C}$.
2. *Linearization*: linearization of the whole system $[\mathcal{C}]_L$.
   We obtain a linear system $LR = [\mathcal{C}]_L \cup [\mathcal{C}]_R$.
3. $\mathcal{D}' := \mathcal{D}$.
4. *Pruning*:
   While the amount of reduction of some bound is greater than ϵ **and** $\emptyset \notin \mathcal{D}'$ **Do**
   (a) $\mathcal{D} \leftarrow \mathcal{D}'$.
   (b) Update the coefficients of the linearizations $[\mathcal{C}]_R$ according to the domains $\mathcal{D}'$.
   (c) Reduce the lower and upper bounds $\underline{x}'_i$ and $\overline{x}'_i$ of each *initial* variable $x_i \in \mathcal{X}$ by computing min and max of x_i subject to LR with a linear programming solver.

ALGORITHM 1
The Quad*-algorithm.*

A tight linear (convex) relaxation, or outer-approximation to the convex and concave envelope of the quadratic terms over the constrained region, is built by generating new linear inequalities.

Quad uses two tight linear relaxation classes that preserve equations $y = x^2$ and $y_{i,j} = x_i x_j$ and that provide a better approximation than interval arithmetic [27].

6.2.1. Linearization of x^2. The term x^2 with $\underline{x} \leq x \leq \overline{x}$ is approximated by the following relations:

$$[x^2]_R \;=\; \begin{cases} L1(\alpha) \equiv [(x-\alpha)^2 \geq 0]_L, \quad \text{where } \alpha \in [\underline{x}, \overline{x}], \\ L2 \equiv [(\underline{x}+\overline{x})x - y - \underline{x}\overline{x} \geq 0]_L. \end{cases} \tag{6.1}$$

Note that $[(x-\alpha_i)^2 = 0]_L$ generates the tangent line to the curve $y = x^2$ at the point $x = \alpha_i$. Actually, Quad computes only $L1(\overline{x})$ and $L1(\underline{x})$. Consider for instance the quadratic term x^2 with $x \in [-4, 5]$. Figure 6.1 displays the initial curve (i.e., D_1) and the lines corresponding to the equations generated by the relaxations: D_2 for $L1(-4) \equiv y + 8x + 16 \geq 0$, D_3 for $L1(5) \equiv y - 10x + 25 \geq 0$, and D_4 for $L2 \equiv -y + x + 20 \geq 0$.

We may note that $L1(\overline{x})$ and $L1(\underline{x})$ are underestimations of x^2 whereas $L2$ is an overestimation. $L2$ is also the concave envelope, which means that it is the optimal concave overestimation.

6.2.2. Bilinear terms. In the case of bilinear terms xy, McCormick [32] proposed the following relaxations of xy over the box $[\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}]$, stated in the equivalent RLT form [54]:

$$[xy]_R \;=\; \begin{cases} BIL1 \equiv [(x-\underline{x})(y-\underline{y}) \geq 0]_L, \\ BIL2 \equiv [(x-\underline{x})(\overline{y}-y) \geq 0]_L, \\ BIL3 \equiv [(\overline{x}-x)(y-\underline{y}) \geq 0]_L, \\ BIL4 \equiv [(\overline{x}-x)(\overline{y}-y) \geq 0]_L. \end{cases} \tag{6.2}$$

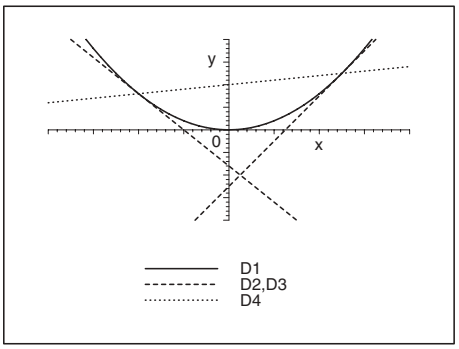


FIG. 6.1. *Approximation of* x^2.

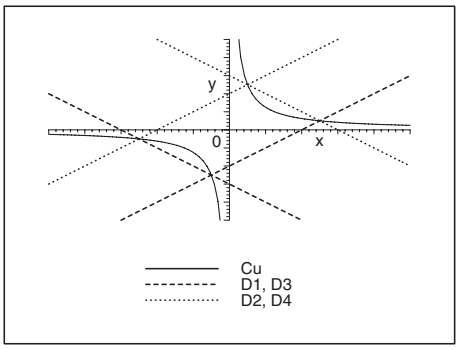


FIG. 6.2. *Illustration of xy relaxations.*

$BIL1$ and $BIL3$ define a convex envelope of xy whereas $BIL2$ and $BIL4$ define a concave envelope of xy over the box $[\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}]$. Al-Khayyal and Falk [1] showed that these relaxations are the optimal convex/concave outer-estimations of xy.

Consider for instance the quadratic term xy with $x \in [-5, 5]$ and $y \in [-5, 5]$. The work done by the linear relaxations of the three-dimensional curve $z = xy$ is well illustrated in two dimensions by fixing z. Figure 6.2 displays the two-dimensional shape, for the level $z = 5$, of the initial curve (i.e., Cu) and the lines corresponding to the equations generated by the relaxations (where $z = 5$): D_1 for $BIL1 \equiv z+5x+5y+25 \geq 0$, D_2 for $BIL2 \equiv -z+5x-5y+25 \geq 0$, D_3 for $BIL3 \equiv -z-5x+5y+25 \geq 0$, and D_4 for $BIL4 \equiv z-5x-5y+25 \geq 0$.

6.3. Extension to polynomial constraints. In this section, we show how to extend the linearization process to polynomial constraints. We first discuss the quadrification process and compare it with RLT. Then, we present the linearizations of product and power terms.

6.3.1. Transformation of nonlinear constraints into quadratic constraints. In this section, we show how to transform a polynomial constraint system into an equivalent quadratic constraint system, a process called *quadrification* [58].

For example, considering the constraint $c : x_2x_3x_4^2 + 3x_6x_7 + \sin(x_1) = 0$, the proposed transformation yields

$$\{y_1y_2 + 3y_2 + s_1 = 0, \quad y_1 = x_2x_3, \quad y_2 = x_4x_4, \quad y_3 = x_6x_7\}$$

and the set $\{y_1 = x_2x_3, \quad y_2 = x_4^2, \quad y_3 = x_6x_7, \quad s_1 = \sin(x_1)\}$ of equations that keep the link between the new variables and the nonlinear terms that cannot be further quadrified. Such a transformation is one of the possible quadrifications. It is called a *single* quadrification.

We could generate all possible single quadrifications, or all quadrifying identities, and perform a so-called *complete* quadrification. For example, the complete quadrification of $E = \{x_2x_3x_4^2 + 3x_6x_7 + \sin(x_1) = 0\}$ is

$$\begin{cases} y_1 + 3y_2 + s_1 = 0, & y_2 = x_6x_7, \\ y_1 = y_3y_4, & y_3 = x_2x_3, & y_4 = x_4^2, \\ y_1 = y_5y_6, & y_5 = x_2x_4, & y_6 = x_3x_4, \\ y_1 = x_2y_7, & y_7 = x_3y_4, & y_7 = x_4y_6, \\ y_1 = x_3y_8, & y_8 = x_2y_4, & y_8 = x_4y_5, \\ y_1 = x_4y_9, & y_9 = x_2y_6, & y_9 = x_3y_5, & y_9 = x_4y_3, \end{cases}$$

where $s_1 = \sin(x_1)$.

A quadrification for polynomial problems was introduced by Shor [58]. Sherali and Tuncbilek [57] have proposed a direct reformulation/linearization (RLT) of the whole polynomial constraints without quadrifying the constraints. They did prove the dominance of their direct reformulation/linearization technique over Shor's quadrification [56].

A complete quadrification generates as many new variables as the direct RLT. Linearizations proposed in RLT are built on every nonordered combination of δ variables, where δ is the highest polynomial degree of the constraint system.

The complete quadrification generates linearizations on every couple of nonordered combined variables $[v_i, v_j]$ where v_i (resp., v_j) is the variable that has been introduced for linearizing the nonordered combination of variables.

Complete quadrification and direct RLT yield a tighter linearization than the single quadrification but the number of generated linearizations grows in an exponential way for nontrivial polynomial constraint systems. More precisely, the number of linearizations depends directly on the number of generated new variables.

To sum up, the linearization of polynomial systems offers two main possibilities: the transformation of the initial problem into an equivalent quadratic constraint system through a process called *quadrification*, or the direct linearization of polynomial terms by means of RLT. Theoretical considerations, as well as experimentations, have been conducted to exclude as practical a complete quadrification which produces a huge amount of linear inequalities for nontrivial polynomial systems. The next two subsections present our choices for the linearization of product and power terms.

6.3.2. Product terms. For the product term

$$x_1 x_2 \dots x_n \tag{6.3}$$

we use a two-step procedure: quadrification and bilinear relaxations.

Since many single quadrifications exist, an essential point is the choice of a good heuristic that captures most of the semantics of the polynomial constraints. We use a "middle" heuristic to obtain balanced degrees on the generated terms. For instance, considering $T \equiv x_1 x_2 \dots x_n$, a monomial of degree n, the middle heuristic will identify two monomials T_1 and T_2 of highest degree such that $T = T_1 T_2$. It follows that $T_1 = x_1 x_2 \dots x_{n \div 2}$ and $T_1 = x_{n \div 2+1} \dots x_n$.

The quadrification is performed by recursively decomposing each product $x_i \dots x_j$ into two products $x_i \dots x_d$ and $x_{d+1} \dots x_j$. Of course, there are many ways to choose the position of d. Ryoo and Sahinidis [49] and Sahinidis and Twarmalani [51] use what they call `rAI`, "recursive interval arithmetic," which is a recursive quadrification where $d = j - 1$. We use the middle heuristic `Qmid`, where $d = (i + j)/2$, to obtain balanced degrees on the generated terms. Let us denote by $[E]_{RI}$ the set of equations that transforms a product terms into a set of quadratic identities.

The second step consists of a *bilinear relaxation* $[[\mathcal{C}]_{RI}]_R$ of all the quadratic identities in $[\mathcal{C}]_{RI}$ with the bilinear relaxations introduced in section 6.2.2.

Sherali and Tuncbilek [57] have proposed a promising direct reformulation/ linearization technique (RLT) of the whole polynomial constraints without quadrifying the constraints. Applying RLT on the product term $x_1 x_2 \dots x_n$ generates the

following n-ary inequalities:[9]

$$\prod_{i\in J_1}(x_i-\underline{x}_i)\prod_{i\in J_2}(\overline{x}_i-x_i)\geq 0 \quad \text{for all } J_1,J_2\subseteq\{1,\dots,n\}\colon |J_1\cup J_2|=n, \tag{6.4}$$

where $\{1,\dots,n\}$ is to be understood as a multiset and where J_1 and J_2 are multisets.

We now introduce Proposition 6.1, which states the number of new variables and relaxations, respectively, generated by the quadrification and RLT process on the product term (6.3).

PROPOSITION 6.1. *Let $T\equiv x_1x_2\dots x_n$ be some product of degree $n\geq 1$ with n distinct variables. The RLT of T will generate up to (2^n-n-1) new variables and 2^n inequalities whereas the quadrification of T will generate only $(n-1)$ new variables and $4(n-1)$ inequalities.*

Proof. The number of terms of length i is clearly the number of combinations of i elements within n elements, that is to say C_n^i. In the RLT relaxations (6.4), we generate new variables for all these combinations. Thus, the number of variables is bounded by $\sum_{i=2,\dots,n}\mathrm{C}_n^i=\sum_{i=0,\dots,n}\mathrm{C}_n^i-n-1$, that is to say 2^n-n-1 since $\sum_{i=0,\dots,n}\mathrm{C}_n^i=2^n$. In (6.4), for each variable we consider alternatively the lower bound and the upper bound: thus there are 2^n new inequalities.

For the quadrification process, the proof can be done by induction. For $n=1$, the formula is true. Now suppose that for length i (with $1\leq i<n$), $(i-1)$ new variables are generated. For $i=n$, we can split the term at the position d with $1\leq d<n$. It results from the induction hypothesis that we have $d-1$ new variables for the first part, and $n-d-1$ new variables for the second part, plus one more new variable for the whole term. So, $n-1$ new variables are generated. Bilinear terms require four relaxations, thus we get $4(n-1)$ new inequalities. ∎

Proposition 6.2 states that quadrification with bilinear relaxations provides convex and concave envelopes with any d. This property results from the proof given in [49] for the `rAI` heuristic.

PROPOSITION 6.2. *$[[x_1x_2\dots x_n]_{RI}]_R$ provides convex and concave envelopes of the product term $x_1x_2\cdots x_n$.*

Generalization for sums of products, the so-called multilinear terms

$$\sum_{i=1,\dots,t} a_i\prod_{j\in J_i}x_j,$$

have been studied recently [14, 52, 48, 49]. It is well known that finding the convex or concave envelope of a multilinear term is an NP-hard problem [14]. The most common method of linear relaxation of multilinear terms is based on the simple product term. However, it is also well known that this approach leads to a poor approximation of the linear bounding of the multilinear terms. Sherali [52] has introduced formulae for computing convex envelopes of the multilinear terms. It is based on an enumeration of vertices of a pre-specified polyhedra which is of exponential nature. Rikun [48] has given necessary and sufficient conditions for the polyhedrality of convex envelopes. He has also provided formulae of some faces of the convex envelope of a multilinear function. To summarize, it is difficult to characterize convex and concave envelopes for general multilinear terms. Conversely, the approximation of "product of variables" is an effective approach; moreover, it is easy to implement [51, 50].

[9] Linearizations proposed in RLT on the whole polynomial problem are built on every nonordered combination of δ variables, where δ is the highest polynomial degree of the constraint system.

6.3.3. Power terms. A power term of the form x^n can be approximated by $n+1$ inequalities with a procedure proposed by Sherali and Tuncbilek [57], called "bound-factor product RLT constraints." It is defined by the following formula:

$$[x^n]_R \;=\; \{[(x-\underline{x})^i(\overline{x}-x)^{n-i} \geq 0]_L, i = 0, \ldots, n\}. \tag{6.5}$$

The essential observation is that this relaxation generates tight relations between variables on their upper and lower bounds. More precisely, suppose that some original variable takes a value equal to either of its bounds. Then all the corresponding new RLT linearization variables that involve this original variable take relative values that conform with actually fixing this original variable at its particular bound in the nonlinear expressions represented by these new RLT variables [57].

Note that relaxations (6.5) of the power term x^n are expressed with x^i for all $i \leq n$, and thus provide a fruitful relationship on problems containing many power terms involving some variable.

The univariate term x^n is convex when n is even, or when n is odd and the value of x is negative; it is concave when n is odd and the value of x is positive. Sahinidis and Twarmalani [50] have introduced the convex and concave envelopes when n is odd by taking the point where the power term x^n and its underestimator have the same slope. These convex/concave relaxations on x^n are expressed with only $[x^n]_L$ and x. In other words, they do not generate any relations with x^i for $1 < i < n$.

That is why we suggest implementing the approximations defined by formulae (6.5). Note that for the case $n = 2$, (6.5) provides the concave envelope.

7. A safe rounding procedure for the `Quad`-algorithm. This section details the rounding procedure we propose to ensure the completeness of the `Quad` algorithm [33]. First, we show how to compute safe coefficients for the generated linear constraints. In the second subsection we explain how a recent result from Neumaier and Shcherbina [42] allows us to use the simplex algorithm in a safe way.

7.1. Computing safe coefficients.

(a) *Approximation of* $L1$. The linear constraint $L1(y,\alpha) \equiv y - 2\alpha x + \alpha^2 \geq 0$ approximates a term x^2 with $\alpha \in [\underline{x}, \overline{x}]$. $L1(y,\alpha)$ corresponds to the tangent lines to the curve $y = x^2$ at the point (α, α^2).

Thus, the computation over the floats of the coefficients of $L1(y,\alpha)$ may change the slope of the tangent line as well as the intersection points with the curve $y = x^2$. Consider the case where α is negative: the solutions are above the tangent line; thus we have to decrease the slope to be sure to keep all of the solutions. It follows that we have to use a rounding mode towards $+\infty$. Likewise, when α is positive, we have to set the rounding mode towards $-\infty$. More formally, we have

$$L1_{\mathbb{F}}(y,\alpha) \;\equiv\; \begin{cases} y - \bigtriangledown(2\alpha)x + \bigtriangleup(\alpha^2) \geq 0 & \text{if } \alpha \geq 0, \\ y - \bigtriangleup(2\alpha)x + \bigtriangleup(\alpha^2) \geq 0 & \text{if } \alpha < 0, \end{cases}$$

where $\bigtriangledown(x)$ (resp., $\bigtriangleup(x)$) denotes a rounding mode of x towards $-\infty$ (resp., $+\infty$).

(b) *Approximation of* $L2$. The case of $L2$ is a bit more tricky since the "rotation axis" of the line defined by $L2$ is between the extremum values of x^2 ($L2(y)$ is an overestimation of y). Thus, to keep all the solutions we have to strengthen the slope

of this line at its smallest extremum. It follows that

$$L2_{\mathbb{F}} \equiv \begin{cases} \triangle(\underline{x}+\overline{x})x - y - \triangledown(\underline{x}\overline{x}) \geq 0 & \text{if } \underline{x} \geq 0, \\ \triangledown(\underline{x}+\overline{x})x - y - \triangledown(\underline{x}\overline{x}) \geq 0 & \text{if } \overline{x} < 0, \\ \triangle(\underline{x}+\overline{x})x - y & \\ \quad -\triangledown(\underline{x}\overline{x} + Ulp(\triangle(\underline{x}+\overline{x}))\underline{x}) \geq 0 & \text{if } \overline{x} > 0,\ \underline{x} < 0,\ |\underline{x}| \leq |\overline{x}|, \\ \triangledown(\underline{x}+\overline{x})x - y & \\ \quad -\triangledown(\underline{x}\overline{x} - \triangle(Ulp(\triangle(\underline{x}+\overline{x}))\overline{x})) \geq 0 & \text{if } \overline{x} > 0,\ \underline{x} < 0,\ |\underline{x}| > |\overline{x}|, \end{cases}$$

where $Ulp(x)$ computes the distance between x and the float following x.

(c) *Approximation of $BIL1, BIL2, BIL3$, and $BIL4$.* The general form of $BIL1$, $BIL2, BIL3$, and $BIL4$ is $x_ix_j + s_1b_1x_i + s_2b_2x_j + s_3b_1b_2 \geq 0$, where b_1 and b_2 are floating point numbers corresponding to bounds of x_i and x_j whereas $s_i \in \{-1, 1\}$.

The term $s_3b_1b_2$ is the only term which results from a computation: all the other terms use constants which are not subject to round-off errors. Thus, these linear constraints can be rewritten in the following form: $Y + s_3b_1b_2$.

A rounding of $s_3b_1b_2$ towards $+\infty$ enlarges the solution space, and thus ensures that all these linear constraints are safe approximations of x^2.

It follows that $BIL\{1, \ldots, 4\}_{\mathbb{F}} \equiv Y + \triangle(s_3b_1b_2) \geq 0$.

(d) *Approximation of multivariate linearizations.* We are now in the position to introduce the corrections of multivariate linearizations as introduced for the power of x. Such linearizations could be rewritten in the following form:

$$\sum_{i=1}^{n} a_ix_i + b \geq 0,$$

where a_i denotes the expression used to compute the coefficient of variable x_i, and b is the expression used to compute the constant value. Proposition 7.1 takes advantage of interval arithmetic to compute a safe linearization with coefficients over the floating point numbers.

PROPOSITION 7.1.

$$\sum_{i=1}^{n} \overline{a}_ix_i + \sup\left(\overline{b} + \sum_{i=1}^{n} \sup(\sup(\mathbf{a}_i\underline{x}_i) - \overline{a}_i\underline{x}_i)\right) \geq \sum_{i=1}^{n} a_ix_i + b \geq 0 \quad \text{for all } x_i \in \mathbf{x}_i.$$

Proof. For all $x_i \in \mathbf{x}_i$, we have

$$\sum_{i=1}^{n} \overline{a}_ix_i + \sup\left(\overline{b} + \sum_{i=1}^{n} \sup(\sup(\mathbf{a}_i\underline{x}_i) - \overline{a}_i\underline{x}_i)\right) \geq \sum_{i=1}^{n} \overline{a}_ix_i + b + \sum_{i=1}^{n} (\sup(\mathbf{a}_i\underline{x}_i) - \overline{a}_i\underline{x}_i)$$

and

$$\sum_{i=1}^{n} \overline{a}_ix_i + b + \sum_{i=1}^{n} (\sup(\mathbf{a}_i\underline{x}_i) - \overline{a}_i\underline{x}_i) = \sum_{i=1}^{n} (\overline{a}_i(x_i - \underline{x}_i) + \sup(\mathbf{a}_i\underline{x}_i)) + b.$$

As for all $i \in \{1, \ldots, n\}$, we have $\overline{a}_i \geq a_i$, $\sup(\mathbf{a}_i\underline{x}_i) \geq a_i\underline{x}_i$, and for all $x_i \in \mathbf{x}_i$, $x_i - \underline{x}_i \geq 0$. Therefore,

$$\sum_{i=1}^{n} (\overline{a}_i(x_i - \underline{x}_i) + \sup(\mathbf{a}_i\underline{x}_i)) + b \geq \sum_{i=1}^{n} (a_i(x_i - \underline{x}_i) + a_i\underline{x}_i) + b = \sum_{i=1}^{n} a_ix_i + b. \qquad \square$$

This proposition provides a safe approximation of a multivariate linearization which holds for any a_i, x_i, and b. This result could be refined by means of the previous approximations. For instance, whenever $\underline{x_i} \geq 0$, $\overline{a}_i x_i \geq a_i x_i$. In this case, there is no need for an additional correction.

(e) *Approximation of initial constant values.* Initial constant values are real numbers that may not have an exact representation within the set of floating point numbers. Thus, a safe approximation is required.

Constant values in inequalities have to be correctly rounded according to the orientation of the inequality. The result presented in the previous paragraph sets the rounding directions which have to be used.

Equations must be transformed into inequalities when their constant values have to be approximated.

7.2. Computation of safe bounds with linear programming algorithm. Linear programming methods can solve problems of the following form:

$$\begin{array}{rl} \min & C^T X \\ \text{such that} & \underline{B} \leq AX \leq \overline{B} \\ \text{and} & \underline{X} \leq X \leq \overline{X}. \end{array} \tag{7.1}$$

The solution of such a problem is a vector $X_r \in \mathbb{R}^n$. However, the solution computed by solvers like `CPLEX` or `SOPLEX` is a vector $X_f \in \mathbb{F}^n$ that may be different from X_r due to the rounding errors. More precisely, X_f is safe for the objective only if $C^T X_r \geq C^T X_f$.

Neumaier and Shcherbina [42] provide a cheap method to obtain a rigorous bound of the objective and certificates of infeasibility. The essential observation is that the dual of (7.1) is

$$\begin{array}{rl} \max & \underline{B}^T Z' + \overline{B}^T Z'' \\ \text{such that} & A^T(Z' - Z'') = C. \end{array} \tag{7.2}$$

Let $Y = Z' - Z''$, and let the residue $R = A^T Y - C \in \mathbf{R} = [\underline{R}, \overline{R}]$. It follows that

$$C^T X = (A^T Y - R)^T X = Y^T AX - R^T X \in Y^T [\underline{B}, \overline{B}] - \mathbf{R}^T [\underline{X}, \overline{X}]$$

and the value of μ, the lower bound of the value of the objective function, is

$$\mu = \inf(Y^T \mathbf{B} - \mathbf{R}^T \mathbf{X}) = \bigtriangledown(Y^T \mathbf{B} - \mathbf{R}^T \mathbf{X}). \tag{7.3}$$

Formula (7.3) is trivially correct by construction. Note that the precision of such a safe bound depends on the width of the intervals $[\underline{X}, \overline{X}]$.

So, we just have to apply this correction before updating the lower and the upper bounds of each variable.

However, the linear program (7.1) may be infeasible. In that case, Neumaier and Shcherbina show that whenever $d = \inf(\mathbf{R'}^T \mathbf{X} - Y^T \mathbf{B}) > 0$, where $R' = A^T Y \in \mathbf{R'}$, then it is certain that no feasible point exists. However, the precision of interval arithmetic does not always allow us to get a positive value for d while the linear program is actually infeasible. In the latter case, we consider it as feasible. Note that box-consistency may be able to reject most, if not all, of the domains of such variables.

8. Experimental results. This section reports experimental results of `Quad` on a variety of twenty standard benchmarks. Benchmarks `eco6`, `katsura5`, `katsura6`, `katsura7`, `tangets2`, `ipp`, `assur44`, `cyclic5`, `tangents0`, `chemequ`, `noon5`, `geneig`, `kinema`, `reimer5`, and `camera1s` were taken from Verschelde's web site,[10] `kin2` from [60], `didrit` from [15], `lee` from [29], and finally `yama194`, `yama195`, and `yama196` from [63]. The most challenging benchmark is `stewgou40` [16]. It describes the 40 possible positions of a Gough–Stewart platform as a function of the values of the actuators. The proposed modelling of this problem consists of 9 equations with 9 variables.

The experimental results are reported in Tables 8.1 and 8.2. Column n (resp., δ) shows the number of variables (resp., the maximum polynomial degree). `BP(Φ)` stands for a *Branch and Prune* solver based on the Φ filtering algorithm, that is to say, a search-tree exploration where a filtering technique Φ is applied at each node. `quad(H)` denotes the `Quad` algorithm where bilinear terms are relaxed with formulae (6.2), power terms with formulae (6.5), and product terms with the quadrification method; `H` stands for the heuristic used for decomposing terms in the quadrification process.

The performances of the following five solvers have been investigated.

1. `RealPaver`: a free *Branch and Prune* solver[11] that dynamically combines optimized implementations of box-consistency filtering and 2b-consistency filtering algorithms [5].
2. `BP(box)`: a *Branch and Prune* solver based on the ILOG[12] commercial implementation of box-consistency.
3. `BP(box+simplex)`: a *Branch and Prune* solver based on `box` and a simple linearization of the whole system without introducing linear relaxations of the nonlinear terms.
4. `BP(box+quad(Qmid))`: a *Branch and Prune* solver which combines `box` and the `Quad` algorithm where product terms are relaxed with the `Qmid` heuristic.
5. `BP(box+quad(rAI))`: a *Branch and Prune* solver which combines `box` and the `Quad` algorithm where product terms are relaxed with the `rAI` heuristic.

Note that the `BP(box+simplex)` solver implements a strategy that is slightly different from the approach of Yamamura, Kawata, and Tokue [63].

All the solvers have been parameterized to get solutions or boxes with precision of 10^{-8}. That is to say, the width of the computed intervals is smaller than 10^{-8}. A solution is said to be *safe* if we can prove its uniqueness within the considered box. This proof is based on the well-known Brouwer fixed point theorem (see [20]) and requires just a single test.

Sols, *Ksplit*, and $T(s)$ are, respectively, the number of solutions, the number of thousands of branchings (or splittings), and the execution time in seconds. The number of solutions is followed with a number of *safe* solutions between brackets. A "–" in the column $T(s)$ means that the solver was unable to find all the solutions within eight hours. All the computations have been performed on a PC with Pentium IV processor at 2.66 GHz running Linux. The compiler was GCC 2.9.6 used with the -O6 optimization flag.

The performances of `RealPaver`, `BP(box)`, and `BP(box+quad(Qmid))` are displayed in Table 8.1. The benchmarks have been grouped into three sets. The first

[10] The database of polynomial systems is available at http://www.math.uic.edu/~jan/Demo/.
[11] See http://www.sciences.univ-nantes.fr/info/perso/permanents/granvil/realpaver/main.html.
[12] See http://www.ilog.com/products/jsolver.

TABLE 8.1
Experimental results: comparing `Quad` *and Constraint solvers.*

			`BP(box+quad(Qmid))`			`BP(box)`			*Realpaver*	
Name	n	δ	*Sols*	*Ksplits*	$T(s)$	*Sols*	*Ksplits*	$T(s)$	*Sols*	$T(s)$
cyclic5	5	5	10(10)	0.6	45.8	10(10)	13.4	26.3	10	291.6
eco6	6	3	4(4)	0.4	15.3	4(4)	1.7	3.7	4	1.3
assur44	8	3	10(10)	0.1	49.5	10(10)	15.8	72.5	10	72.6
ipp	8	2	10(10)	0.0	5.7	10(10)	4.6	14.0	10	16.8
katsura5	6	2	16(11)	0.1	9.9	41(11)	8.2	12.7	12	6.7
katsura6	7	2	60(24)	0.5	121.9	182(24)	136.6	281.4	32	191.8
kin2	8	2	10(10)	0.0	6.2	10(10)	3.5	19.3	10	2.6
noon5	5	3	11(11)	0.1	17.9	11(11)	50.2	58.7	11	39.0
tangents2	6	2	24(24)	0.1	17.5	24(24)	14.1	27.9	24	16.5
camera1s	6	2	16(16)	1.0	28.9	2(2)	11820.3	–	0	–
didrit	9	2	4(4)	0.1	14.7	4(4)	51.3	132.9	4	94.6
geneig	6	3	10(10)	0.8	39.1	10(10)	290.7	868.6	10	475.6
kinema	9	2	8(8)	0.2	19.9	15(7)	244.0	572.4	8	268.4
katsura7	8	2	58(42)	1.7	686.9	231(42)	1858.5	11104.1	44	4671.1
lee	9	2	4(4)	0.5	43.3	0(0)	8286.3	–	0	–
reimer5	5	6	24(24)	0.1	53.0	24(24)	2230.2	2892.5	24	733.9
stewgou40	9	4	40(40)	1.6	924.0	11(11)	3128.6	–	8	–
yama194	16	3	9(9)	0.0	11.1	9(8)	1842.1	–	0	–
yama195	60	3	3(3)	0.0	106.1	0(0)	19.6	–	0	–
yama196	30	1	2(1)	0.0	6.7	0(0)	816.7	–	0	–

group contains problems where the `QuadSolver` does not behave very well. These problems are quite easy to solve and the overhead of the relaxation and the calls to a linear solver does not pay off. The second group contains a set of benchmarks for which the `QuadSolver` compares well with the two other constraint solvers: the `QuadSolver` requires always much less splitting and often less time than the other solvers. In the third group, which contains difficult problems, the `QuadSolver` outperforms the two other constraint solvers. The latter were unable to solve most of these problems within eight hours whereas the `QuadSolver` managed to find all the solutions for all but two of them in less than 8 minutes. For instance, `BP(box)` requires about 74 hours to find the four solutions of the `Lee` benchmark whereas the `QuadSolver` managed to do the job in a couple of minutes. Likewise, the `QuadSolver` did find the forty safe solutions of the `stewgou40` benchmark in about 15 minutes whereas `BP(box)` required about 400 hours. The essential observation is that the `QuadSolver` spends more time in the filtering step but it performs much less splitting than classical solvers. This strategy pays off for difficult problems.

All the problems, except `cyclic5` and `reimer5`, contain many quadratic terms and some product and power terms. `cyclic5` is a pure multilinear problem that contains only sums of products of variables. The `Quad` algorithm has not been very efficient for handling this problem. Of course, one could not expect an outstanding performance on this bench since product term relaxation is a poor approximation of multilinear terms. `reimer5` is a pure power problem of degree 6 that has been well solved by the `Quad` algorithm.

Table 8.2 displays the performances of solvers combining box-consistency and three different relaxation techniques. There is no significant difference between the solver based on the `Qmid` heuristics and the solver based on the `rAI` heuristics. Indeed, both heuristics provide convex and concave envelopes of the product terms. The `QuadSolver` with relaxations outperforms the `BP(box+simplex)` approach for all

TABLE 8.2
Experimental results: comparing `Quad` *based on different relaxations.*

	`BP(box+simplex)`			`BP(box+quad(Qmid))`			`BP(box+quad(rAI))`		
Name	*Sols*	*Ksplits*	$T(s)$	*Sols*	*Ksplits*	$T(s)$	*Sols*	*Ksplits*	$T(s)$
cyclic5	10(10)	15.6	60.6	10(10)	0.6	45.8	10(10)	0.8	76.1
eco6	4(4)	1.1	7.2	4(4)	0.4	15.3	4(4)	0.4	15.3
assur44	10(10)	15.5	261.9	10(10)	0.1	49.5	10(10)	0.1	50.0
ipp	10(10)	3.2	39.7	10(10)	0.0	5.7	10(10)	0.0	5.7
katsura5	41(11)	7.7	47.8	16(11)	0.1	9.9	16(11)	0.1	9.9
katsura6	182(24)	135.2	1156.7	60(24)	0.5	121.9	60(24)	0.5	122.7
kin2	10(10)	3.4	42.5	10(10)	0.0	6.2	10(10)	0.0	6.2
noon5	11(11)	49.6	226.7	11(11)	0.1	17.9	11(11)	0.1	17.8
tangents2	24(24)	11.4	77.7	24(24)	0.1	17.5	24(24)	0.1	17.5
camera1s	4(4)	3298.6	–	16(16)	1.0	28.9	16(16)	1.0	29.9
didrit	4(4)	5.3	93.2	4(4)	0.1	14.7	4(4)	0.1	14.7
geneig	10(10)	202.8	2036.8	10(10)	0.8	39.1	10(10)	0.8	39.2
kinema	13(7)	87.0	1135.1	8(8)	0.2	19.9	8(8)	0.2	20.0
katsura7	231(42)	1867.2	21679.6	58(42)	1.7	686.9	58(42)	1.7	684.0
lee	2(2)	78.1	1791.8	2(2)	0.3	27.1	2(2)	0.3	26.5
lee2	4(4)	117.6	2687.2	4(4)	0.5	43.3	4(4)	0.5	43.3
reimer5	24(24)	2208.7	10433.5	24(24)	0.1	53.0	24(24)	0.1	53.1
stewgou40	13(13)	716.3	–	40(40)	1.6	924.0	40(40)	1.5	914.1
yama194	9(7)	442.0	–	9(9)	0.0	11.1	9(9)	0.0	11.2
yama195	3(2)	0.0	37.7	3(3)	0.0	106.1	3(3)	0.0	106.7
yama196	2(1)	0.0	6.6	2(1)	0.0	6.7	2(1)	0.0	6.7

benchmarks but `yama195`, which is a quasilinear problem. These performances on difficult problems illustrate well the capabilities of the relaxations.

Note that Verschelde's homotopy continuation system, `PHCpack` [62], required 115 s to solve `lee` and 1047 s to solve `stewgou40` on our computer. `PHCpack` is a state-of-the-art system in solving polynomial systems of equations. Unfortunately, it is limited to polynomial systems and does not handle inequalities. `PHCpack` searches for all the roots of the equations, whether real or complex, and it does not restrict its search to a given subspace. The homotopy continuation approach also suffers from an exponential growing computation time which depends on the number of nonlinear terms (`PHCpack` failed to solve `yama195` which contains 3600 nonlinear terms). In contrast to homotopy continuation methods, `QuadSolver` can easily be extended to nonpolynomial systems.

Thanks to Arnold Neumaier and Oleg Shcherbina, we had the opportunity to test BARON [50] with some of our benchmarks. `QuadSolver` compares well with this system. For example, BARON 6.0[13] and `QuadSolver` require more or less the same time to solve `camera1s`, `didrit`, `kinema`, and `lee`. BARON needs only 1.59 s to find all the solutions of `yama196` but it requires 859.6 s to solve `yama195`. Moreover, BARON loses some solutions on `reimer5` (22 solutions found) and `stewgou40` (14 solutions found) whereas it generates numerous wrong solutions for these two problems. We must also underline that BARON is a global optimization problem solver and that it has not been built to find all the solutions of a problem.

9. Conclusion. In this paper, we have exploited an RLT schema to take into account specific semantics of nonlinear terms. This relaxation process is incorporated in the *Branch and Prune* process [60] that exploits interval analysis and constraint satis-

[13]The tests were performed on an Athlon XP 1800 computer.

faction techniques to find rigorously all solutions in a given box. The reported experimental results show that this approach outperforms the classical constraint solvers.

Pesant and Boyer [44, 45] first introduced linear relaxations in a CLP language to handle geometrical constraints. However, the approximation of the constraints was rather weak. The approach introduced in this paper is also related to recent work that has been done in the interval analysis community as well as to some work achieved in the optimization community.

In the interval analysis community, Yamamura, Kawata, and Tokue [63] used a simple linear relaxation procedure where nonlinear terms are replaced by new variables to prove that some box does not contain solutions. No convex/concave outer-estimations are proposed to obtain a better approximation of the nonlinear terms. As pointed out by Yamamura, Kawata, and Tokue, this approach is well adapted to quasi-linear problems: *"This test is much more powerful than the conventional test if the system of nonlinear equations consists of many linear terms and a relatively small number of nonlinear terms"* [63].

The global optimization community also worked on solving nonlinear equation problems by transforming them into an optimization problem (see, for example, Chapter 23 in [17]). The optimization approach has the capability to take into account specific semantics of nonlinear terms by generating a tight outer-estimation of these terms. The pure optimization methods are usually not rigorous since they do not take into account rounding errors and do not prove the uniqueness of the solutions found.

Acknowledgments. We thank Arnold Neumaier for his fruitful comments on an early version of this paper. We are also grateful to Arnold Neumaier and Oleg Shcherbina for their help in testing BARON.

REFERENCES

[1] F. A. Al-Khayyal and J. E. Falk, *Jointly constrained biconvex programming*, Math. Oper. Res., 8 (1983), pp. 273–286.
[2] E. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction*, Springer-Verlag, Berlin, 1990.
[3] C. Audet, P. Hansen, B. Jaumard, and G. Savard, *Branch and cut algorithm for nonconvex quadratically constrained quadratic programming*, Math. Program., 87 (2000), pp. 131–152.
[4] A. B. Babichev, O. P. Kadyrova, T. P. Kashevarova, A. S. Leshchenko, and A. L. Semenov, *Unicalc, a novel approach to solving systems of algebraic equations*, Interval Computations, 2 (1993), pp. 29–47.
[5] F. Benhamou, F. Goualard, L. Granvilliers, and J.-F. Puget, *Revising hull and box consistency*, in Proceedings of ICLP '99, 1999, MIT Press, pp. 230–244.
[6] F. Benhamou, D. McAllester, and P. Van-Hentenryck, *CLP(intervals) revisited*, in Proceedings of the International Symposium on Logic Programming, MIT Press, Cambridge, MA, 1994, pp. 124–138.
[7] F. Benhamou and W. Older, *Applying interval arithmetic to real, integer and Boolean constraints*, J. Logic Programming, 32 (1997), pp. 1–24.
[8] B. Botella and P. Taillibert, *Interlog: Constraint logic programming on numeric intervals*, in 3rd International Workshop on Software Engineering, Artificial Intelligence and Expert Systems, Oberammergau, 1993.
[9] R. L. Burden and J. D. Faires, *Numerical Analysis*, 5th ed., PWS-KENT, Boston, MA, 1993.
[10] J. C. Cleary, *Logical arithmetic*, Future Computing Systems, 2 (1987), pp. 125–149.
[11] H. Collavizza, F. Delobel, and M. Rueher, *Comparing partial consistencies*, Reliable Computing, 5 (1999), pp. 213–228.
[12] A. Colmerauer, *Spécifications de Prolog IV*, Technical report, GIA, Faculté des Sciences de Luminy, France, 1994.
[13] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms*, 2nd ed., Springer-Verlag, New York, 1997.

[14] Y. Crama, *Recognition problems for polynomial in* 0-1 *variables*, Math. Progr., 44 (1989), pp. 139–155.
[15] O. Didrit, *Analyse par Intervalles pour L'automatique: Résolution Globale et Garantie de Problèmes Non Linéaires en Robotique et en Commande Robuste*, Ph.D. thesis, Université Parix XI Orsay, France, 1997.
[16] P. Dietmaier, *The Stewart-Gough platform of general geometry can have* 40 *real postures*, in Advances in Robot Kinematics: Analysis and Control, Kluwer, Dordrecht, 1998, pp. 1–10.
[17] C. A. Floudas, ed., *Deterministic Global Optimization: Theory, Algorithms and Applications*, Kluwer, Dordrecht, 2000.
[18] D. Goldberg, *What every computer scientist should know about floating-point arithmetic*, ACM Computing Surveys, 23 (1991), pp. 5–48.
[19] E. Hansen and S. Sengupta, *Bounding solutions of systems of equations using interval analysis*, BIT, 21 (1981), pp. 203–221.
[20] E. R. Hansen, *Global Optimization Using Interval Analysis*, Marcel Dekker, New York, 1992.
[21] H. Hong and V. Stahl, *Starting regions by fixed points and tightening*, Computing, 53 (1994), pp. 323–335.
[22] *ILOG Solver 4.0, Reference Manual*, ILOG, Mountain View, 1997.
[23] R. B. Kearfott, *Tests of generalized bisection*, ACM Trans. Math. Software, 13 (1987), pp. 197–220.
[24] R. Krawczyk, *Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken*, Computing, 4 (1969), pp. 187–201.
[25] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.
[26] Y. Lebbah and O. Lhomme, *Accelerating filtering techniques for numeric CSPs*, Artificial Intelligence, 139 (2002), pp. 109–132.
[27] Y. Lebbah, M. Rueher, and C. Michel, *A global filtering algorithm for handling systems of quadratic equations and inequations*, in Proc. of the 8th International Conference on Principles and Practice of Constraint Programming, Cornell University, New York, 2002, Lecture Notes in Comput. Sci. 2470, pp. 109–123.
[28] J. H. M. Lee and M. H. van Emden, *Interval computation as deduction in CHIP*, J. Logic Programming, 16 (1993), pp. 255–276.
[29] T.-Y. Lee and J.-K. Shim, *Elimination-based solution method for the forward kinematics of the general Stewart-Gough platform*, in Computational Kinematics, F. C. Park and C. C. Iurascu, eds., 2001, pp. 259–267.
[30] O. Lhomme, *Consistency techniques for numeric CSPs*, in Proceedings of International Joint Conference on Artificial Intelligence, Chambéry, France, 1993, pp. 232–238.
[31] A. Mackworth, *Consistency in networks of relations*, J. Artificial Intelligence, 8 (1977), pp. 99–118.
[32] G. P. McCormick, *Computability of global solutions to factorable nonconvex programs—Part* I*—Convex underestimating problems*, Math. Progr., 10 (1976), pp. 147–175.
[33] C. Michel, Y. Lebbah, and M. Rueher, *Safe embedding of the simplex algorithm in a CSP framework*, in Proc. of 5th Int. Workshop on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems CPAIOR 2003, CRT, Université de Montréal, 2003, pp. 210–220.
[34] U. Montanari, *Networks of constraints: Fundamental properties and applications to image processing*, Inform. Sci., 7 (1974), pp. 95–132.
[35] R. Moore, *Interval Analysis*, Prentice–Hall, Englewood Cliff, NJ, 1966.
[36] R. E. Moore, *Methods and Applications of Interval Analysis*, SIAM Stud. Appl. Math. 2, SIAM, Philadelphia, 1979.
[37] A. P. Morgan, *Computing all solutions to polynomial systems using homotopy continuation*, Appl. Math. Comput., 24 (1987), pp. 115–138.
[38] A. Neumaier, *Interval Methods for Systems of Equations*, Encyclopedia Math. Appl. 37, Cambridge University Press, Cambridge, 1990.
[39] A. Neumaier, *The wrapping effect, ellipsoid arithmetic, stability and confidence region*, Comput. Suppl., 9 (1993), pp. 175–190.
[40] A. Neumaier, *Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, 2001.
[41] A. Neumaier, *Complete search in continuous global optimization and constraint satisfaction*, in Acta Numerica 2004, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 2004, pp. 271–369.
[42] A. Neumaier and O. Shcherbina, *Safe bounds in linear and mixed-integer programming*, Math. Progr. A, 99 (2004), pp. 283–296.

[43] W. J. Older and A. Velino, *Extending prolog with constraint arithmetic on real intervals*, in Proc. of IEEE Canadian Conference on Electrical and Computer Engineering, IEEE Computer Society Press, 1990, pp. 14.1.1–14.1.4.
[44] G. Pesant and M. Boyer, *QUAD-CLP(R): Adding the power of quadratic constraints*, in Principles and Practice of Constraint Programming '94, Lecture Notes in Comput. Sci. 874, 1994, pp. 95–107.
[45] G. Pesant and M. Boyer, *Reasoning about solids using constraint logic programming*, J. Automat. Reason., 22 (1999), pp. 241–262.
[46] H. Ratschek and J. Rokne, *Computer Methods for the Range of Functions*, Ellis Horwood Ser. Math. Appl., Ellis Horwood, New York, 1984.
[47] N. Revol and F. Rouillier, *Motivations for an arbitrary precision interval arithmetic and the MPFI library*, in Workshop on Validated Computing, Toronto, Canada, 2002.
[48] A. Rikun, *A convex envelope formula for multilinear functions*, J. Global Optim., 10 (1997), pp. 425–437.
[49] H. S. Ryoo and N. V. Sahinidis, *Analysis of bounds for multilinear functions*, J. Global Optim., 19 (2001), pp. 403–424.
[50] N. V. Sahinidis and M. Twarmalani, *BARON 5.0: Global Optimization of Mixed-Integer Nonlinear Programs*, Technical report, Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, 2002.
[51] N. V. Sahinidis and M. Twarmalani, *Global optimization of mixed-integer programs: A theoretical and computational study*, Math. Progr., 99 (2004), pp. 563–591.
[52] H. D. Sherali, *Convex envelopes of multilinear functions over a unit hypercube and over special discrete sets*, Acta Math. Vietnam., 22 (1997), pp. 245–270.
[53] H. D. Sherali and W. P. Adams, *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*, Kluwer, Dordrecht, 1999.
[54] H. D. Sherali and C. H. Tuncbilek, *A global optimization algorithm for polynomial using a reformulation-linearization technique*, J. Global Optim., 2 (1992), pp. 101–112.
[55] H. D. Sherali and C. H. Tuncbilek, *A reformulation-convexification approach for solving nonconvex quadratic programming problems*, J. Global Optim., 7 (1995), pp. 1–31.
[56] H. D. Sherali and C. H. Tuncbilek, *A comparison of two reformulation-linearization technique based on linear programming relaxations for polynomial porgramming problems*, J. Global Optim., 10 (1997), pp. 381–390.
[57] H. D. Sherali and C. H. Tuncbilek, *New reformulation linearization/convexification relaxations for univariate and multivariate polynomial programming problems*, Oper. Res. Lett., 21 (1997), pp. 1–9.
[58] N. Z. Shor, *Dual quadratic estimates in polynomial and Boolean programming*, Ann. Oper. Res., 25 (1990), pp. 163–168.
[59] M. Tawarmalani and N. V. Sahinidis, eds., *Convexification and Global Optimization in Continuous and Mixed-Integer Non-Linear Programming*, Kluwer, Dordrecht, 2002.
[60] P. Van Hentenryck, D. McAllester, and D. Kapur, *Solving polynomial systems using a branch and prune approach*, SIAM J. Numer. Anal., 34 (1997), pp. 797–827.
[61] P. Van-Hentenryck, L. Michel, and Y. Deville, *Numerica: A Modeling Language for Global Optimization*, MIT Press, Cambridge, MA, 1997.
[62] J. Verschelde, *Algorithm* 795: *PHCPACK: A general-purpose solver for polynomial systems by homotopy continuation*, ACM Trans. Math. Software, 25 (1999), pp. 251–276.
[63] K. Yamamura, H. Kawata, and A. Tokue, *Interval solution of nonlinear equations using linear programming*, BIT, 38 (1998), pp. 186–199.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2098–2117

FAST MULTIPOLE METHOD FOR MULTIVARIABLE INTEGRALS*

OLIVIER BOKANOWSKI† AND MOHAMMED LEMOU‡

Abstract. We give a fast numerical algorithm to evaluate a class of multivariable integrals. A direct numerical evaluation of these integrals costs N^m, where m is the number of variables and N is the number of the quadrature points for each variable. For $m = 2$ and $m = 3$ and for only one-dimensional variables, we present an algorithm which is able to reduce this cost from N^m to a cost of the order of $O((-\log\epsilon)^{\mu_m} N)$, where ϵ is the desired accuracy and μ_m is a constant that depends only on m. Then, we make some comments about possible extensions of such algorithms to number of variables $m \geq 4$ and to higher dimensions. This recursive algorithm can be viewed as an extension of "fast multipole methods" to situations where the interactions between particles are more complex than the standard case of binary interactions. Numerical tests illustrating the efficiency and the limitation of this method are presented.

Key words. fast multipole method, multivariable integrals, multiparticle integrals, multidimensional integrals, multidimensional sums, correlated sums, $O(N)$ algorithm, molecular quantum physics

AMS subject classifications. 65Y20, 68Q25, 68W25, 45P05, 81Q05, 81U10

DOI. 10.1137/S0036142902409690

1. Introduction.

In this paper, we are concerned with numerical approximations of the following multivariable integrals:

$$\int_{\mathcal{C}\times\mathcal{C}\times\cdots\times\mathcal{C}} \phi_1(x_1)\phi_2(x_2)\cdots\phi_m(x_m)\left[\prod_{1\leq i<j\leq m} f_{ij}(x_i - x_j)\right] d\mu(x_1)\, d\mu(x_2)\cdots d\mu(x_m), \tag{1.1}$$

where $\mathcal{C}$ is a cube of $\mathbb{R}^d$, m is the number of variables, and μ is a positive measure. For instance, the measure μ can be the usual Lebesgue's measure or a discrete measure. The functions f_{ij} are assumed to be sufficiently regular on $\mathbb{R}^d\backslash\{0\}$. The functions ϕ_j and f_{ij} are such that the integrals (1.1) is absolutely convergent.

The basic example is the evaluation of the total interaction energy of a system of N charged particles:

$$E = \sum_{i,j=1;\ i\neq j}^{N} \frac{q_i r_j}{|x_i - x_j|}, \tag{1.2}$$

where x_i is the position of the ith particle and q_i its charge (here $r_i = q_i$), or the continuous version

$$E = \int_{\mathcal{C}\times\mathcal{C}} \frac{q(x)r(y)}{|x-y|} d\mu(x) d\mu(y). \tag{1.3}$$

*Received by the editors June 14, 2002; accepted for publication (in revised form) March 29, 2004; published electronically February 25, 2005. This research was supported by the A.C.I. Blanche (Nantes): "Modèles relativistes en mécanique quantique."
http://www.siam.org/journals/sinum/42-5/40969.html

†Laboratoire Jacques-Louis Lions, UMR 7598, Université Paris 6, 175 rue du Chevaleret, 75013 Paris, France (boka@math.jussieu.fr).

‡MIP (UMR CNRS 5640), UFR MIG, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France (lemou@mip.ups.tlse.fr).

The classical interactions described by (1.3) are binary in the sense that they correspond to the particular case $m = 2$ of (1.1). Fast computations of the integrals (1.3) for a large number of functions (q, r) must be performed in quantum chemistry, for "Hartree–Fock" models and "density functional" models involving many electrons, as explained and done in [8, 9], [18], or [19] (computation of the "J-matrix").

However, in more precise quantum models the particles are strongly correlated through a wave function $\Psi(x_1, \ldots, x_m)$, where $m \geq 3$ is the number of particles. In this case, the determination of the state of the particles leads to the problem of computing the following integrals ("Coulomb energy"):

$$\int_{\Omega^m} |\Psi(x_1, \ldots, x_m)|^2 \left(\sum_{1 \leq i < j \leq m} \frac{1}{|x_i - x_j|} \right) d\mu(x_1) \cdots d\mu(x_m), \tag{1.4}$$

with $\Omega \subset \mathbb{R}^d$, and where the function Ψ can take the following form:

$$\Psi(x_1, \ldots, x_m) := \phi_1(x_1)\phi_2(x_2) \cdots \phi_m(x_m) \prod_{1 \leq i < j \leq m} f_{ij}(x_i - x_j) \tag{1.5}$$

or

$$\Psi(x_1, \ldots, x_m) := \phi_1(x_1)\phi_2(x_2) \cdots \phi_m(x_m) \left(\sum_{1 \leq i < j \leq m} f_{ij}(x_i - x_j) \right). \tag{1.6}$$

When the variables x_i are not correlated ($f_{ij} = 1$), explicit computations for (1.4) and (1.5) can be done for some particular classes of ϕ_i's such as Gaussian functions [12, Chapters 1 and 2]. But this specific choice of Ψ is not always sufficient to describe the realistic behavior of the state of the particles [13]. On the other hand, if the particles are correlated ($f_{ij} \neq 1$), analytic computations are not possible in general, and Monte Carlo methods are usually used [10] (see also [11]). The main difficulties with these random methods, or pseudorandom methods, is to control the accuracy because of the presence of oscillations and convergence problems.

Hopefully there are some cases where partial analytic computations are possible. In the "Møller–Plesset perturbation theory of second order with linear r_{12} terms," or "MP2-R12 method" [14], the used typical functions are of the form (1.6) with $f_{ij}(x_i - x_j) = |x_i - x_j|$. In this case a simple algebraic calculation shows that the computation of the m-particle integral (1.4) reduces to the computation of $m = 2, 3$ or $m = 4$ particle integrals of the form (1.1), and of 1-particle integrals. This has been extensively used in order to compute very accurately the energy of small atoms and molecules as in Kutzelnigg and Klopper [15, 16] and Müller, Kutzelnigg, and Noga [17], or in [23]. Yet, a limitation of this approach is that a particular basis of functions ϕ_j is used in order to avoid or to approximate the computation of multiparticle integrals. It would be useful to be able to rapidly (and accurately) compute such integrals (for $m = 2, 3, 4$ only) without any particular constraint on the ϕ_j functions.

Note that for most real quantum applications, one should treat the case $d = 3$. However, quantum model problems on the line ($d = 1$) are also used [22].

In this paper we present an alternative approach. This approach is deterministic and is an extension of the well-known fast multiple method (FMM) to more complex systems of particles. The FMM method was introduced by Greengard and Rokhlin [1, 2] to compute the binary interaction (1.2) with a cost of the order of $O(N)$ instead

of the order of $O(N^2)$ (see also [5] and [6], and for instance [20] or [21] for recent developments of the FMM). Here we want to extend this technique to situations where the correlations between particles are stronger, as in multivariable integrals of the form (1.1). We give the details of the algorithm only in the case of two or three variables (i.e., $m = 2$ or $m = 3$) and for one-dimensional variables ($d = 1$). This work gives a first idea of how this algorithm could be used for more general situations ($d = 3$ for instance). However, a complete generalization is not obvious and is under investigation.

Our aim is to reduce the cost of the numerical evaluation of integrals (1.1). We use a regular mesh of $\mathcal{C}$ with N points and consider a discrete measure μ supported by the mesh. This also corresponds to a quadrature formula of the integrals (1.1), and leads to the problem of evaluating the following sums:

$$\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_m=1}^{N}\phi_1(x_{i_1})\phi_2(x_{i_2})\cdots\phi_m(x_{i_m})\left[\prod_{1\le k<\ell\le m} f_{k\ell}(x_{i_k}-x_{i_\ell})\right]. \tag{1.7}$$

A direct evaluation of the sums (1.7) requires a computational cost of the order of $O(N^m)$. The purpose of the paper is to present an algorithm reducing this cost to the order $O(\log_2(\frac{1}{\epsilon})^{\mu}\, N)$, where μ depends only on m and d and will be made precise in what follows (at least for the specific case $d = 1$ and $m \le 3$), and where ϵ is the relative error between the exact value of formula (1.7) and the result obtained by the present algorithm. The parameter ϵ is linked to the order of "multipole expansions" through the relation $p = \log_2(\frac{1}{\epsilon})$ (when $d = 1$) and does not depend on N. This result was already announced by the authors in [4] without proof.

In order to simplify the presentation of the algorithm, we shall assume that $d = 1$, $f_{ij}(r) = r^{\alpha_{ij}}$ (with $\alpha_{ij} > -1$), and $\mathcal{C} = [0, 1]$.

We also mention that the approach of [7] using wavelets, for $m = 3$ variables and with $d = 1$, can be used to compute some partially correlated sums or integrals. However, it cannot be applied here in general because the variables in the sum (1.7) are completely correlated.

We first present in section 2 the algorithm in the simple case of $m = 2$ variables. This corresponds to the case of binary interactions for which the multipole method has been widely developed by Greengard and Rokhlin [1, 2]. In sections 3 and 4, we focus on the case $m = 3$ and state our main results. We point out that this is not a straightforward generalization of the 2-variable case. In section 5, we give some numerical illustrations of the method in the case $m = 3$, and in section 6 we conclude with some remarks and extensions to the case $m \ge 4$.

2. The case of binary interactions ($m = 2$). In this section, we present a brief description of the approach in the case of binary interactions ($m = 2$). In this case, the used method is strongly related to the ideas of the classical FMM introduced by Greengard and Rokhlin [1, 2]. Even if the case of binary interactions is now classical, the presentation below is a first step in the understanding of the more complex approach for strongly correlated particles ($m = 3$ for instance). This last case is our main concern in this work and will be developed in the next section.

Consider first the case of binary interactions, i.e., the problem of approximating the following sum:

$$I_{dis} := \sum_{i,j=1}^{N}\phi_1(a_i)\phi_2(a_j)|a_i - a_j|^{\alpha}, \tag{2.1}$$

where $\alpha \geq 0$ and $a_i = \frac{1}{N}(i - \frac{1}{2})$, $i = 1, \ldots, N$, are the points of a regular mesh of $\mathcal{C}$. Note that when $-1 < \alpha < 0$ we can also consider the sum (2.1) for all pairs (i, j) with $i \neq j$. In order to simplify the presentation we restrict ourselves to the case $\alpha \geq 0$ (see Remark 3).

Note that I_{dis} is also equal to the following integral:

$$I_{dis} := \int_{\mathcal{C}\times\mathcal{C}} \phi_1(x_1)\phi_2(x_2)|x_1 - x_2|^\alpha \, d\mu(x_1)d\mu(x_2),$$

where $\mathcal{C} = [0, 1]$ and μ is the discrete measure defined on $\mathcal{C}$ by $\mu(\{x\}) = 1$ if $x = a_i$, and $\mu(\{x\}) = 0$ otherwise. To simplify, we set $d^2\mu = d\mu \otimes d\mu$ and $F(x_1, x_2) = \phi_1(x_1)\phi_2(x_2)|x_1 - x_2|^\alpha$.

To evaluate I_{dis} one needs $O(N^2)$ operations. Our purpose is to reduce this cost to the order $O(N)$. To achieve this we use a multigrid hierarchy and multipole expansions as follows.

Multigrid hierarchy. We split the interval $\mathcal{C} = \mathcal{C}_0^1$ in several parts according to the following hierarchy.

Level $k = 0$: $\mathcal{C}_0^1 = \mathcal{C}$.

Level $k = 1$: We split $\mathcal{C}$ into two equal parts (called its children) $\mathcal{C}_1^1 = [0, \frac{1}{2}]$, $\mathcal{C}_1^2 = [\frac{1}{2}, 1]$, and $\mathcal{C}_0^1$ is called their father.

Level $k = 2$: We split each $\mathcal{C}_1^i$ into two equal parts (its children). We obtain 2 intervals $\mathcal{C}_2^1, \mathcal{C}_2^2$ (when $i = 1$) or $\mathcal{C}_2^3, \mathcal{C}_2^4$ when $i = 2$.

Level k: We split each $\mathcal{C}_{k-1}^i$ into two equal parts, $\mathcal{C}_k^{2i-1}$ and $\mathcal{C}_k^{2i}$. We obtain 2^k intervals $(\mathcal{C}_k^i)_{i=1,\ldots,2^k}$.

The center of an interval $\mathcal{C}_k^i$ is $r_k^i = \frac{i-1/2}{2^k}$. We iterate this process until the finest mesh level n_g.

DEFINITION 2.1. (i) *Let $\mathcal{C}_k^i$ and $\mathcal{C}_k^j$ be two intervals of level k. We say that $\mathcal{C}_k^i$ and $\mathcal{C}_k^j$ are* well separated *if they are separated at least by one interval of the same level (i.e., $|i - j| \geq 2$).*

(ii) *We say that $\mathcal{C}_k^i$ and $\mathcal{C}_k^j$ are* neighbors *if $|i - j| \leq 1$ and then write $\mathcal{C}_k^i \in V(\mathcal{C}_k^j)$. $V(\mathcal{C}_k^i)$ denotes the set of neighbors of $\mathcal{C}_k^i$.*

(iii) *We define and denote by $Int(\mathcal{C}_k^i)$ the* interaction list *of a given interval $\mathcal{C}_k^i$ of level k, as the set of intervals $\mathcal{C}_k^j$ of the same level k, which are well separated, and whose fathers are neighbors.*

Graphic illustrations of neighboring intervals and well-separated intervals are given in Figure 2.1.

With this definition we have the following partition of $\mathcal{C} \times \mathcal{C}$:

$$\mathcal{C}\times\mathcal{C} = \left(\bigcup_{k=1,n_g} \bigcup_{\substack{i,j=1,\ldots,2^k \\ \mathcal{C}_k^j \in Int(\mathcal{C}_k^i)}} \mathcal{C}_k^i \times \mathcal{C}_k^j \right) \bigcup \left(\bigcup_{\substack{i,j=1,\ldots,2^{n_g} \\ \mathcal{C}_{n_g}^i, \mathcal{C}_{n_g}^j \text{ neighbors}}} \mathcal{C}_{n_g}^i \times \mathcal{C}_{n_g}^j \right). \tag{2.2}$$

From this remark we deduce that

$$I_{dis} = \left(\sum_{k=1,n_g} \sum_{\substack{i,j=1,\ldots,2^k \\ \mathcal{C}_k^j \in Int(\mathcal{C}_k^i)}} \int_{\mathcal{C}_k^i \times \mathcal{C}_k^j} F d^2\mu \right) + \left(\sum_{\substack{i,j=1,\ldots,2^{n_g} \\ \mathcal{C}_{n_g}^i, \mathcal{C}_{n_g}^j \text{ neighbors}}} \int_{\mathcal{C}_{n_g}^i \times \mathcal{C}_{n_g}^j} F d^2\mu \right). \tag{2.3}$$

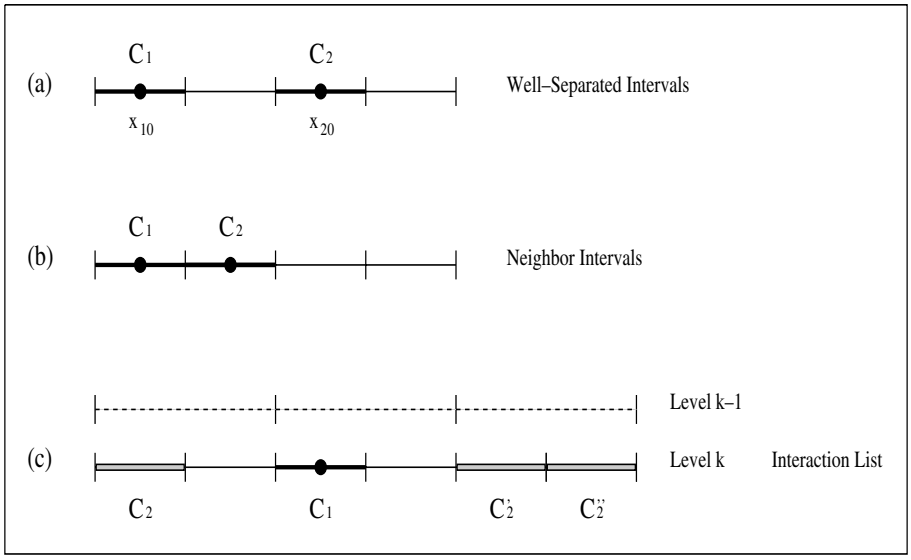


FIG. 2.1. *Example of well-separated intervals* (a) *and of neighboring intervals* (b) *at some level* k. (c) *Example of an interaction list* $Int(C_1) = (C_2, C_2', C_2'')$ *of an interval* C_1.

Thus, in order to evaluate I_{dis}, it suffices to evaluate the integrals $\int_{\mathcal{C}_k^i \times \mathcal{C}_k^j} F$ when $\mathcal{C}_k^i$ and $\mathcal{C}_k^j$ are well separated (at level k) but whose fathers are neighbors, and also the integrals $\int_{\mathcal{C}_{n_g}^i \times \mathcal{C}_{n_g}^j} F$ on neighboring intervals at the finest level n_g. Notice that in (2.2) or (2.3) we have only $O(N)$ contributions of products $\mathcal{C}_k^i \times \mathcal{C}_k^j$.

Notations. We define the moment of ϕ of order i on the interval $\mathcal{C}$ around y by

$$M_\phi(i, y, \mathcal{C}) := \int_{\mathcal{C}} \phi(x)(x-y)^i \, d\mu(x). \tag{2.4}$$

In the case where y is the center of the interval $\mathcal{C}$, we simply write

$$M_\phi(i, \mathcal{C}) := M_\phi(i, y, \mathcal{C}).$$

LEMMA 2.2 (multipole expansion). (i) *Let* $\alpha > -1$, *and let* $p \in \mathbb{N}$ *be such that* $p + 1 \geq \alpha$. *Let* $\mathcal{C}^1$ *and* $\mathcal{C}^2$ *be two intervals of the same level, which are well separated, and let* x_{10} *and* x_{20} *be their centers. Let* $u = \frac{x_{10} - x_{20}}{|x_{10} - x_{20}|}$ *(i.e.,* $u = \pm 1$*) and* $h_i = \frac{x_i - x_{i0}}{|x_{10} - x_{20}|}$. *There exist coefficients* $c_{ij}(\alpha, u)$ *such that for all* $x_1 \in \mathcal{C}^1$, *and for all* $x_2 \in \mathcal{C}^2$, *we have*

$$|x_1 - x_2|^\alpha = |x_{10} - x_{20}|^\alpha \left(\sum_{i+j \leq p} c_{ij}(\alpha, u) h_1^i h_2^j + \delta \right), \tag{2.5}$$

where $|\delta| \leq \dfrac{c_\alpha}{(p+1)^{\alpha+1} 2^p}$.

(ii) *We have the following expansion, called* multipole expansion [1]*:*

(2.6)

$$\int_{\mathcal{C}^1\times\mathcal{C}^2} F d^2\mu = |x_{10}-x_{20}|^\alpha \left(\sum_{i+j\le p} \frac{c_{ij}(\alpha,u)}{|x_{10}-x_{20}|^{i+j}} M_{\phi_1}(i,\mathcal{C}^1)\ M_{\phi_2}(j,\mathcal{C}^2) \right) + \eta,$$

where $F = F(x_1,x_2) = \phi_1(x_1)\phi_2(x_2)|x_1-x_2|^\alpha$, *and where* x_{10} *and* x_{20} *are the centers of* $\mathcal{C}^1$ *and* $\mathcal{C}^2$, *respectively. The error* η *satisfies*

$$|\eta| \le \frac{c'_\alpha}{(p+1)^{\alpha+1}2^p} \int_{\mathcal{C}^1\times\mathcal{C}^2} |F| d^2\mu,$$

where the constants c_α *and* c'_α *will be made precise in the proof.*

Note. By $\sum_{i+j\le p}$ we mean the sum for all pairs (i,j) such that $i\ge 0, j\ge 0$ and $i+j\le p$.

Proof of Lemma 2.2. We write

$$\begin{aligned} |x_1-x_2|^\alpha &= |(x_{10}-x_{20})+(x_1-x_{10})-(x_2-x_{20})|^\alpha \\ &= |x_{10}-x_{20}|^\alpha\ |u+h_1-h_2|^\alpha. \end{aligned}$$

Since C^1 and C^2 are well separated, $|h_i| \le \frac{1}{4}$. Consider the case $u=1$ (the case $u=-1$ is similar). We have $|u+h_1-h_2|^\alpha = (1+h)^\alpha$, where $h = h_1-h_2$ and $|h| \le |h_1|+|h_2| \le 2\frac{1}{4} \le \frac{1}{2}$. Then we expand $(1+h)^\alpha$ to the order p with respect to h:

$$(1+h)^\alpha = \sum_{n=0}^{p} \binom{\alpha}{n} h^n + R_{p+1}(\alpha,h),$$

where $\binom{a}{n} := \frac{a(a-1)\cdots(a-n+1)}{n!}$ and $R_p(\alpha,h) := \sum_{n=p}^\infty \binom{\alpha}{n} h^n$. Developing again $h^n = (h_1-h_2)^h = \sum_{j=1}^n \binom{j}{n} h_1^{n-j}(-h_2)^j$, we find formula (2.5) with

$$c_{ij}(u,\alpha) := (-1)^j u^{i+j} \binom{i}{i+j} \binom{\alpha}{i+j},$$

and with an error term δ that we want to bound. Let

$$q = q(\alpha) := \inf\{n\in\mathbb{N},\ n\ge\alpha\}.$$

We have $|\binom{\alpha}{n}| = A_\alpha B_\alpha(n)$ with

$$A_\alpha := \frac{1}{q!}\alpha(\alpha-1)(\alpha-2)\cdots(\alpha+1-q)$$

and

$$B_\alpha(n) := \left(1-\frac{\alpha+1}{q+1}\right)\left(1-\frac{\alpha+1}{q+2}\right)\cdots\left(1-\frac{\alpha+1}{n}\right).$$

From the definition of $q=q(\alpha)$, we have $\alpha \le q < \alpha+1$, and then $A_\alpha \le 1$. Now, using the inequality $1-x\le e^{-x}$, we obtain $0 \le B_\alpha(n) \le (\frac{q+1}{n+1})^{\alpha+1}$. Thus $|\binom{\alpha}{n}| \le (\frac{q+1}{n+1})^{\alpha+1}$. Hence, for $p\ge 1$, we have

$$
\begin{aligned}
R_p(\alpha,n) &\le (q+1)^{\alpha+1}\sum_{n\ge p}\frac{|h|^n}{(n+1)^{\alpha+1}}\\
&\le (q+1)^{\alpha+1}\frac{1}{|h|}\int_p^\infty \frac{|h|^t}{t^{\alpha+1}}dt\\
&\le (q+1)^{\alpha+1}\frac{1}{\log(1/|h|)}\frac{|h|^{p-1}}{p^{\alpha+1}}\\
&\le \frac{(q+1)^{\alpha+1}}{\log(2)}\frac{(\frac12)^{p-1}}{p^{\alpha+1}}.
\end{aligned}
$$

We then deduce that for $p\ge\alpha$, (i.e., $p\ge q$), $R_p(\alpha,n)\le c_\alpha\frac{1}{p^{\alpha+1}}(\frac12)^{p-1}$, where $c_\alpha=(\alpha+2)^{\alpha+1}/\log(2)$, and Lemma 2.2(i) follows.

(ii) Reporting this bound in the expression of I_i, and observing that $\frac23\le\frac{|x_{10}-x_{20}|}{|x_1-x_2|}\le 2$, we obtain the desired bound with $c'_\alpha:=c_\alpha\max(2^\alpha,(\frac23)^\alpha)$. □

Remark 1. We give also an estimate that will be useful in the next section. There exists a constant L_α such that, for $|h_1|\le\frac14$ and $|h_2|\le\frac14$, we have

$$\sum_{0\le i+j\le p}|c_{ij}(u,\alpha)||h_1|^i|h_2|^j\le L_\alpha,$$

where the coefficients $c_{ij}(u,\alpha)$ are defined in Lemma 2.2.

THEOREM 2.3. *Let $N=2^{n_g}s$, with s being a fixed integer. Let p be an integer such that $\bar p:=p+1\ge\alpha$. There exists an explicit algorithm that gives an approximation $I_{multipole}$ of the discretization I_{dis}, (2.1), using $O(\max\{1,\frac{\bar p}{s}\}\,\bar p\,N)$ operations with a relative error of the order of $O(1/\left(\bar p^{\alpha+1}\,2^p\right))$, i.e., $|I_{dis}-I_{multipole}|\le\frac{c_\alpha}{\bar p^{\alpha+1}\,2^p}\int_{\mathcal{C}_0\times\mathcal{C}_0}|F|$.*

Note that taking $s=\bar p$ we obtain a total cost of $O(\bar p\,N)$. Note also that s is the number of points inside each interval of the finest level n_g, and is naturally chosen to be small compared to N.

Remark 2 (elementary operations). For simplicity, we set to one all the costs of each of the following operations: one addition, one multiplication, one power (such as x^α or $|x|^\alpha$), and one evaluation of a known function. The cost of memory access will also be neglected.

Proof of Theorem 2.3. Assume that the moments $M_{\phi_j}(i,\mathcal{C})$ ($i=0,\dots,p$ and $j=1,2$) have been precalculated for all intervals $\mathcal{C}$ of each level of the multigrid hierarchy. Since an interaction list contains at most 3 intervals, it follows from Lemma 2.2 and (2.3) that the total cost of the evaluation of the integrals on well-separated intervals is of the order of $\sum_{k=1}^{n_g}3\;2^k\bar p^2\le 3\;2^{n_g+1}\bar p^2=O(p^2\frac{N}{s})$.

On the other hand, the integrals on neighboring intervals at level n_g are evaluated by direct quadrature formula (because here we cannot use a multipole expansion). The cost of this last evaluation is s^2 for each pair of intervals. There are at most 3 neighbors for each interval. Hence the total cost of these integrals is smaller than $3s^2 2^{n_g}=3sN$.

Now we present a recursive algorithm in order to precalculate the moments $M_{\phi_j}(i,\mathcal{C})$ at all levels $k=1,\dots,n_g$ and for $i=0,\dots,p$. First, we evaluate the moments at the finest mesh level n_g. This costs $s2^{n_g}(p+1)=O(\bar pN)$, since there are $\bar p=p+1$ moments for each interval, with s points per interval at the finest level. Suppose we have calculated the moments of level $k+1$. To evaluate a moment of order k, we use the decomposition of $\mathcal{C}_k^\ell$ into its two children: $\mathcal{C}_k^\ell=\mathcal{C}_{k+1}^{2\ell-1}\cup\mathcal{C}_{k+1}^{2\ell}$, and the formula

$$M_\phi(i,\mathcal{C}_k) = M_\phi(i,r_k,\mathcal{C}_k) = \sum_{j,\ \mathcal{C}_{k+1}^j \text{ child of } \mathcal{C}_k} M_\phi(i,r_k,\mathcal{C}_{k+1}^j). \tag{2.7}$$

Using the translation $x - r_k = x - r_{k+1}^j - (r_k - r_{k+1}^j)$, we obtain

$$M_\phi(i,r_k,\mathcal{C}_{k+1}^j) = \sum_{q=0}^{i} \begin{pmatrix} i \\ q \end{pmatrix} (r_k - r_{k+1}^j)^{i-q} M_\phi(q,\mathcal{C}_{k+1}^j). \tag{2.8}$$

Hence the computation of the moments $M_\phi(i,\mathcal{C}_k)$ for $i = 0,\dots,p$ requires $O(\bar{p}^2)$ operations. The total cost of the evaluation of all moments at all levels is $O(\bar{p}N) + O(\sum_{k=1}^{n_g-1} 2^k \bar{p}^2) = O(\bar{p}\,N + \frac{\bar{p}^2}{s}N)$, which concludes the proof of the first part of Theorem 2.3.

In order to obtain the error estimate of Theorem 2.3, we use the controlled multipole approximation of Lemma 2.2(ii). We obtain for any well-separated intervals $\mathcal{C}^1, \mathcal{C}^2$ of the same level

$$\left| \int_{\mathcal{C}^1 \times \mathcal{C}^2} F d^2\mu - I_{mult.}(\mathcal{C}^1 \times \mathcal{C}^2) \right| \le \frac{c'_\alpha}{\bar{p}^{\alpha+1} 2^p} \int_{\mathcal{C}^1 \times \mathcal{C}^2} |F| d^2\mu \tag{2.9}$$

with

$$I_{mult.}(\mathcal{C}^1 \times \mathcal{C}^2) := |x_{10} - x_{20}|^\alpha \left(\sum_{i+j \le p} \frac{c_{ij}(\alpha,u)}{|x_{10} - x_{20}|^{i+j}} M_{\phi_1}(i,\mathcal{C}^1)\ M_{\phi_2}(j,\mathcal{C}^2) \right).$$

Summing the above contributions (2.9) at all levels and over all products $\mathcal{C}^1 \times \mathcal{C}^2$ as in the partition (2.2), and adding the exact contributions at the finest level (for neighbor intervals), we obtain the desired error bound. This concludes the proof of Theorem 2.3. □

Remark 3. If we consider the sum (2.1) with $\alpha \in (-1,0)$ and for pairs (i,j) such that $i \ne j$, then the previous results and algorithm are unchanged except for the calculation of the sums at the finest mesh level (where we must sum only over (i,j) with $i \ne j$).

3. The case of multiple interactions ($m = 3$ variables). In this section, we consider the following three-variables integral:

$$I = \int_{\mathcal{C}_0 \times \mathcal{C}_0 \times \mathcal{C}_0} F(x_1,x_2,x_3) d^3\mu(x_1,x_2,x_3),$$

where

$$F(x_1,x_2,x_3) := \phi_1(x_1)\phi_2(x_2)\phi_3(x_3)|x_1 - x_2|^\alpha |x_2 - x_3|^\beta |x_3 - x_1|^\gamma,$$

$\mathcal{C}_0 = [0,1]$, and $\alpha,\beta,\gamma > -1$. We have also denoted $d^3\mu = d\mu \otimes d\mu \otimes d\mu$. As for the case $m = 2$, when μ is the discrete measure, this integral becomes

$$I_{dis} := \sum_{i_1,i_2,i_3=1}^{N} \phi_1(a_{i_1})\phi_2(a_{i_2})\phi_3(a_{i_3})|a_{i_1} - a_{i_2}|^\alpha |a_{i_2} - a_{i_3}|^\beta |a_{i_3} - a_{i_1}|^\gamma. \tag{3.1}$$

We shall also restrict ourselves to the case $\alpha,\beta,\gamma \ge 0$ as in the previous section in order to simplify the presentation of the result. However, as in Remark 3, it is not

difficult to extend our results to the case $\alpha, \beta, \gamma > -1$ (in which I_{dis} is defined as in (3.1) but where we sum only on $i_1 \neq i_2$, $i_2 \neq i_3$, and $i_3 \neq i_1$).

To evaluate I_{dis} one needs $O(N^3)$ operations. Our purpose is to reduce this cost to the order $O(N)$. We use the same hierarchy and the same definitions as in the previous section.

In the following theorem, we state our main result.

THEOREM 3.1. *Let $N = 2^{n_g} s$, with s being a fixed integer. Let p be an integer such that $p+1 \geq \max(\alpha, \beta, \gamma)$. Let $\bar{p} := p+1$. There exists an explicit algorithm that gives an approximation $I_{multipole}$ of the discretization I_{dis}, (3.1), using $O(\max\{\frac{\bar{p}}{s}, \frac{s^2}{\bar{p}^2}\}\bar{p}^2 N)$ operations with a relative error of the order of $O(p^{-\delta}2^{-p})$, where $\delta = \min(\alpha, \beta, \gamma)+1$, in the following sense:*

$$|I_{dis} - I_{multipole}| \leq C_{\alpha,\beta,\gamma} \frac{1}{\bar{p}^{\delta} 2^p} \int_{\mathcal{C}_0 \times \mathcal{C}_0 \times \mathcal{C}_0} |F| \, d^3\mu \ . \tag{3.2}$$

Note that taking $s = \bar{p}$ we obtain an $O(\bar{p}^2 N)$ algorithm, and taking $s = 1$ we obtain an $O(\bar{p}^3 N)$ algorithm. We recall that s is the number of particles in each interval at the finest mesh level n_g and that s is a small number compared to N.

Explicit theoretical cost bounds are given by (4.6), (4.7), and (4.8) in section 4.

Proof. To prove this theorem we follow the ideas of the previous section. We will first establish that the cost is proportional to N. We show that the proportionality factor depends only on s and $\bar{p}$, and we do not give the exact complexity in this section. In particular, we show that the cost is of the order of $O(\bar{p}^6 N)$ when $s \leq \bar{p}$. In section 4 we shall establish a more precise cost estimate using only summation techniques.

The proof of Theorem 3.1 is achieved in three steps.

3.1. Basic partition. Our first aim is to obtain a decomposition of I_{dis} in the same way as in (2.3). We then have to define a suitable notion of neighbors (for a set of three intervals), and select intervals which enter in interaction at each level.

Neighbors (or neighboring intervals). See Figure 3.1 for some examples. We say that a set of *three* intervals of a given level are *neighbors* (or *neighboring intervals*), if each of the three intervals is a neighbor of at least one of the two others. We recall that *two* intervals are neighbors if there is no interval between them. For instance,

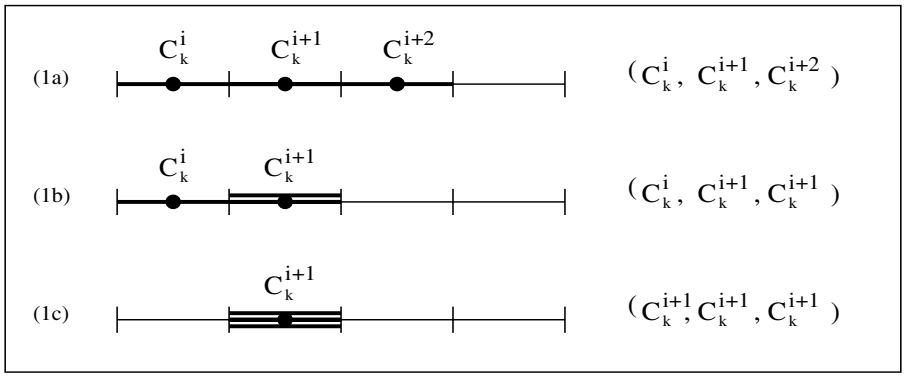


FIG. 3.1. *Example of neighbor intervals.*

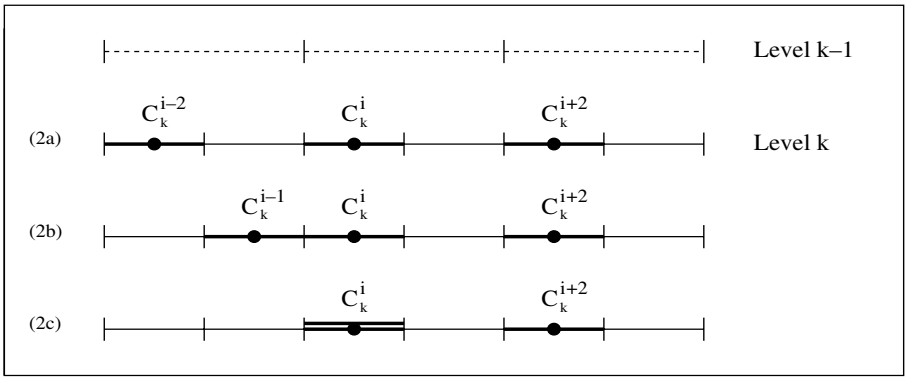


FIG. 3.2. *Example of intervals in interaction.* (a): *Case* (i); (b) *and* (c): *Case* (ii).

the *three* intervals

$$(\mathcal{C}_k^i, \mathcal{C}_k^{i+1}, \mathcal{C}_k^{i+2})$$

are neighboring intervals, even if the *two* intervals $(\mathcal{C}_k^i, \mathcal{C}_k^{i+2})$ are not neighbors.

Intervals in interaction (see Figure 3.2). We say also that *three* intervals of a given level are *in interaction* if they are *not* neighbors but their fathers are, in the sense defined above. For instance, the following intervals are always in interaction:

$$(\mathcal{C}_k^{i-2}, \mathcal{C}_k^i, \mathcal{C}_k^{i+2}), (\mathcal{C}_k^{i-1}, \mathcal{C}_k^i, \mathcal{C}_k^{i+2}). \tag{3.3}$$

Note that in this case there is at least one interval which is well separated from the two others.

Partition. Let Int_k be the product of intervals of level k which are in interaction:

$$Int_k := \bigcup_{(\mathcal{C}_k^{r_1},\ \mathcal{C}_k^{r_2},\ \mathcal{C}_k^{r_3}) \text{ in interaction}} \mathcal{C}_k^{r_1} \times \mathcal{C}_k^{r_2} \times \mathcal{C}_k^{r_3}.$$

Let also H_k be the product of intervals of level k which are neighbors:

$$H_k := \bigcup_{(\mathcal{C}_k^{r_1},\ \mathcal{C}_k^{r_2},\ \mathcal{C}_k^{r_3}) \text{ neighbors}} \mathcal{C}_k^{r_1} \times \mathcal{C}_k^{r_2} \times \mathcal{C}_k^{r_3}.$$

We note that $H_k = Int_{k+1} \cup H_{k+1}$ from the definitions, and that H_0 contains only one element: $H_0 = \mathcal{C} \times \mathcal{C} \times \mathcal{C}$ (recall that $\mathcal{C} = \mathcal{C}_0^1$). Hence we deduce the following partition of $\mathcal{C} \times \mathcal{C} \times \mathcal{C}$:

$$\mathcal{C} \times \mathcal{C} \times \mathcal{C} = \left(\bigcup_{k=2}^{n_g} Int_k \right) \bigcup H_{n_g}. \tag{3.4}$$

Note that $Int_k = \emptyset$ for $k = 1$, because intervals are neighbors.

Now, for later use, we decompose Int_k into two types of contributions. Let three intervals be in interaction. Since they are not neighbors, we have two cases.

Case (i). Each interval is well separated from the two others.

Case (ii). One interval is well separated from the two others, which are neighbors.

For instance, the first example in (3.3) corresponds to case (i), while the second corresponds to case (ii). We then have $Int_k = E_k \cup F_k$, where E_k and F_k are defined by

$$E_k := \bigcup_{(\mathcal{C}_k^{r_1},\ \mathcal{C}_k^{r_2},\ \mathcal{C}_k^{r_3}) \in Int_k,\ \text{case(i)}} \mathcal{C}_k^{r_1} \times \mathcal{C}_k^{r_2} \times \mathcal{C}_k^{r_3}$$

and

$$F_k := \bigcup_{(\mathcal{C}_k^{r_1},\ \mathcal{C}_k^{r_2},\ \mathcal{C}_k^{r_3}) \in Int_k,\ \text{case(ii)}} \mathcal{C}_k^{r_1} \times \mathcal{C}_k^{r_2} \times \mathcal{C}_k^{r_3}.$$

We finally obtain from (3.4) the following partition:

$$\mathcal{C} \times \mathcal{C} \times \mathcal{C} = \left(\bigcup_{k=2}^{n_g} E_k \cup F_k \right) \bigcup H_{n_g}. \tag{3.5}$$

Let $I_k^{r_1,r_2,r_3}$ be the integral on $\mathcal{C}_k^{r_1} \times \mathcal{C}_k^{r_2} \times \mathcal{C}_k^{r_3}$:

(3.6)
$$I_k^{r_1,r_2,r_3} = \int_{\mathcal{C}_k^{r_1} \times \mathcal{C}_k^{r_2} \times \mathcal{C}_k^{r_3}} \phi_1(x_1)\phi_2(x_2)\phi_3(x_3)|x_1 - x_2|^\alpha |x_2 - x_3|^\beta |x_3 - x_1|^\gamma d^3\mu.$$

Then, as in (3.5), we obtain I_{dis} as the sum of all these previous contributions at all levels:

$$\begin{aligned} I_{dis} = &\sum_{k=2}^{n_g} \sum_{\mathcal{C}_k^{r_1} \times \mathcal{C}_k^{r_2} \times \mathcal{C}_k^{r_3} \in E_k} I_k^{r_1,r_2,r_3} \\ &+ \sum_{k=2}^{n_g} \sum_{\mathcal{C}_k^{r_1} \times \mathcal{C}_k^{r_2} \times \mathcal{C}_k^{r_3} \in F_k} I_k^{r_1,r_2,r_3} \\ &+ \sum_{\mathcal{C}_k^{r_1}, \mathcal{C}_k^{r_2}, \mathcal{C}_k^{r_3}\ \text{neighbors}} I_{n_g}^{r_1,r_2,r_3}. \end{aligned} \tag{3.7}$$

3.2. Computation of moments and integrals: One-variable and two-variables moments. The first difference with the case of binary interactions ($m = 2$) is that we need to precalculate supplementary moments. In addition to the one-variable moments

$$M_\phi^1(i, \mathcal{C}) = \int_{\mathcal{C}} \phi(x)(x - x_0)^i d\mu(x) \tag{3.8}$$

(where x_0 is the center of $\mathcal{C}$), we define new moments (two-variables moments) as follows:

(3.9)
$$M^2_{\phi_1,\phi_2,\alpha}(i, j, \mathcal{C}^1, \mathcal{C}^2) := \int_{\mathcal{C}^1 \times \mathcal{C}^2} \phi_1(x_1)\phi_2(x_2)(x_1 - x_{10})^i (x_2 - x_{20})^j |x_1 - x_2|^\alpha d^2\mu(x_1, x_2),$$

where x_{10} and x_{20} are the centers of $\mathcal{C}^1$ and $\mathcal{C}^2$, respectively. The above integrals are sums (since μ is discrete), and for instance for M^2 we have the formula

$$M^2_{\phi_1,\phi_2,\alpha}(i,j,\mathcal{C}^1,\mathcal{C}^2) = \sum_{k,\ell=1,\dots,2^{n_g};\ a_k\in\mathcal{C}^1,\ a_\ell\in\mathcal{C}^2} \phi_1(a_k)\phi_2(a_\ell)(a_k-x_{10})^i(a_\ell-x_{20})^j|a_k-a_\ell|^\alpha$$

(we exclude the terms $k=\ell$ in the case $\alpha<0$). These moments will be used in the computation of the F_k integrals.

Developed formula with moments. We assume that all the one-variable moments have been calculated, and all the two-variables moments ($M^2_{\phi_1,\phi_2,\alpha}(i,j,\mathcal{C}^1,\mathcal{C}^2)$, $M^2_{\phi_2,\phi_3,\beta}(i,j,\mathcal{C}^1,\mathcal{C}^2)$, and $M^2_{\phi_3,\phi_1,\gamma}(i,j,\mathcal{C}^1,\mathcal{C}^2)$) have been calculated for $\mathcal{C}^1\in Int(\mathcal{C}^2)$, and at all levels. Starting from this hypothesis, we will determine the (asymptotic) cost of evaluating the total integral I_{dis}. According to the decompositions (3.5) or (3.7), we distinguish three kinds of integrals to compute.

We first consider the case of the E_k integrals. These are integrals of the form (3.6) in which the intervals $\mathcal{C}_k^{r_1}$, $\mathcal{C}_k^{r_2}$, and $\mathcal{C}_k^{r_3}$ are mutually well separated. The three variables can be mutually separated using Lemma 2.2, and we obtain an expression using only one-variable moments M^1:

$$\begin{aligned} I_k^{r_1,r_2,r_3} &= |x_{10}-x_{20}|^\alpha|x_{20}-x_{30}|^\beta|x_{30}-x_{10}|^\alpha \\ &\times\Bigg\{\sum_{i_1+j_1\le p}\ \sum_{i_2+j_2\le p}\ \sum_{i_3+j_3\le p} \frac{c_{i_1,j_1}(\alpha,u_1)\ c_{i_2,j_2}(\beta,u_2)\ c_{i_3,j_3}(\gamma,u_3)}{|x_{10}-x_{20}|^{i_1+j_1}\ |x_{20}-x_{30}|^{i_2+j_2}\ |x_{30}-x_{10}|^{i_3+j_3}} \\ &\times M^1_{\phi_1}(i_1+j_3,\mathcal{C}_k^{r_1})\ M^1_{\phi_2}(i_2+j_1,\mathcal{C}_k^{r_2})M^1_{\phi_3}(i_3+j_2,\mathcal{C}_k^{r_3})\Bigg\}+\eta_E, \end{aligned} \tag{3.10}$$

where η_E is the resulting error. Using Lemma 2.2 and Remark 1, and after easy calculations we get the following error estimate:

$$|\eta|\le C_{\alpha,\beta,\gamma}p^{-\min(\alpha,\beta,\gamma)-1}2^{-p}\int_{\mathcal{C}_k^{r_1}\times\mathcal{C}_k^{r_2}\times\mathcal{C}_k^{r_3}}|F|d^3\mu.$$

The evaluation of expression (3.10) costs $O(\bar{p}^6)$ operations. The computation of all E_k integrals then costs $O(\bar{p}^6\sum_{k=1}^{n_g}2^k)$, i.e., $O(\bar{p}^6 2^{n_g})$.

Now we consider the F_k integrals. If for instance $\mathcal{C}_k^{r_3}$ is well separated from $\mathcal{C}_k^{r_1}$ and $\mathcal{C}_k^{r_2}$, then using multipole expansions for $|x_2-x_3|^\beta$ and $|x_3-x_1|^\gamma$ (see Lemma 2.2) we obtain

$$\begin{aligned} I_k^{r_1,r_2,r_3} &= |x_{30}-x_{10}|^\gamma|x_{20}-x_{30}|^\beta\Bigg\{\sum_{i_1+j_1\le p}\ \sum_{i_2+j_2\le p}\frac{c^1_{i_1,j_1}(\gamma,u_3)\ c^2_{i_2,j_2}(\beta,u_2)}{|x_{30}-x_{10}|^{i_1+j_1}\ |x_{20}-x_{30}|^{i_2+j_2}} \\ &\times M^2_{\phi_1,\phi_2,\alpha}(i_1,i_2,\mathcal{C}_k^{r_1},\mathcal{C}_k^{r_2})\ M^1_{\phi_3}(j_1+j_2,\mathcal{C}_k^{r_3})\Bigg\}+\eta, \end{aligned} \tag{3.11}$$

where η is the error and satisfies $|\eta|\le C_{\beta,\gamma}p^{-\min(\beta,\gamma)-1}2^{-p}\int_{\mathcal{C}_k^{r_1}\times\mathcal{C}_k^{r_2}\times\mathcal{C}_k^{r_3}}|F|d^3\mu$. We point out that a fast computation of the two-variables moments again requires the use of multipole expansion (see below) and then involves an additional error. The error estimate η_F in the computation of the F_k integrals still follows the same expression

as for η_E, but the proof of such an estimate is much more technical than for the E_k integrals. Because of its limited interest, this proof is omitted.

Expression (3.11) can be calculated using the precalculated moments M^1 and M^2, and with a cost of the order of $O(\bar{p}^4)$. Then the total cost induced by the F_k integrals is of the order of $O(\bar{p}^4 \sum_{k=1}^{n_g} 2^k)$, i.e., $O(\bar{p}^4 2^{n_g})$.

Finally we consider the H_{n_g} integrals at the finest mesh level. For these, we use a direct computation of the sums of the terms of $I_{n_g}^{r_1,r_2,r_3}$. Because these integrals only concern neighboring intervals at the finest level n_g, this step costs $O(s^3 \cdot 2^{n_g})$.

Hence the total cost for the calculations of I_{dis}, using multipole expansions, is of the order of $O((s^3 + \bar{p}^6)\ 2^{n_g}) = O(\frac{s^3+\bar{p}^6}{s}\ N)$.

Computation of the moments. We first compute the moments $M^1_{\phi_j}(i, \mathcal{C})$ for the orders $i = 0, \ldots, 2p$. This costs $O(\bar{p}^2 \frac{N}{s}) + O(\bar{p}N)$ (see the previous section). We then compute the moments M^2 given by (3.9) in the only case where $\mathcal{C}^1$ and $\mathcal{C}^2$ are neighbors (other moments are not used). We proceed recursively as for the moments M^1.

First, we evaluate the moments M^2 at the finest mesh level n_g, for $i, j = 0, \ldots, p$. This costs $O(\bar{p}^2 \cdot s^2 \cdot 2^{n_g}) = O(\bar{p}^2 \cdot s \cdot N)$, where $N = 2^{n_g} \cdot s$ (since there are 2^{n_g} intervals and each interval contains s quadrature points). Now suppose we have calculated the moments M^2 at level $k+1$. To evaluate a moment M^2 at level k on neighbors $\mathcal{C}^1 \times \mathcal{C}^2$, we decompose $\mathcal{C}^1$ (resp., $\mathcal{C}^2$) into intervals of levels $k+1$: $\mathcal{C}^i_{k+1}$ (resp., $\mathcal{C}^j_{k+1}$). When $\mathcal{C}^i_{k+1}$ and $\mathcal{C}^j_{k+1}$ are well separated, we can use Lemma 2.2 in order to separate the variables and use only one-variable moments (we use also translations as in (2.8)). When $\mathcal{C}^i_{k+1}$ and $\mathcal{C}^j_{k+1}$ are neighbors, we can use the moments M^2 precalculated at level $k+1$ (we use also translations). The complexity of these last recursive operations will be determined in the next section. Roughly, the contribution corresponding to neighboring children $\mathcal{C}^i_{k+1}$ and $\mathcal{C}^j_{k+1}$ needs $O(\bar{p}^2)$ operations, because it only uses translations. The contributions of well-separated children requires both multipole expansions, and translations, and thus needs $O(\bar{p}^4)$ operations. Since we have $O(\bar{p}^2)$ moments $M^2_{\phi_1,\phi_2,\alpha}(i, j, \mathcal{C}^1, \mathcal{C}^2)$ $(i, j = 0, \ldots, p)$ to calculate, the total cost is of the order of $O(\bar{p}^6)$.

Since there are $O(2^{n_g}) = O(N/s)$ intervals, the total cost for computing the moments is of the order of $O(\bar{p}^2 sN) + O(\bar{p}^6 \frac{N}{s})$.

Total cost. Adding the previous bounds, we obtain a total cost of the order of $O(\bar{p}^2 sN) + O(\bar{p}^6 \frac{N}{s}) + O(\frac{s^3+\bar{p}^6}{s}\ N)$. This is bounded by $O(\frac{s^3+\bar{p}^6}{s}\ N)$. In particular, for $s \in [1, \bar{p}]$, we obtain a cost of the order of $O(\bar{p}^6\ N)$.

Up to now, we have shown that the total computational cost of the multipole algorithm is *asymptotically* proportional to N. In situations where N is not large enough, it is useful to know the value of the proportionality coefficient. In the next section, we give more precise expressions of the computational costs (or complexities) of the above-described algorithms. Since the proofs of such expressions are very technical and lengthy, we just present them without any proof. These reduction techniques are essentially based on reordering the terms of the various sums and/or writing them in terms of matrix-matrix products. We point out that these manipulations do not involve any additional error.

4. Exact cost. In this section, we give more precise expressions of the costs of the computations involved in the above-described multipole algorithm. These costs will be expressed as functions of the quantities p, s, and N.

Computing the one-variable moments. The computational cost of the one-variable moments, (3.8), $M_{\phi_q}(i, \mathcal{C}_k)$ for $i = 0, \ldots, 2p$, and for all levels $k = 2, 3, \ldots, n_g$, is

bounded by

$$\left(6\bar{p} + 44\frac{\bar{p}^2}{s}\right)\ N. \tag{4.1}$$

Computing the two-variables moments. As explained in the previous section, we need to compute all two-variables moments $M^2_{\phi_1,\phi_2,\alpha}(i,j,\mathcal{C}^1,\mathcal{C}^2)$ on *neighbor intervals*, at all levels, for $i,j = 0,\ldots,p$. These moments are expressed by (3.9), and their computation requires a more complicated algorithm than for the one-variable moments. The total cost for this algorithm is bounded by

$$\begin{aligned} &3 \cdot \left(18 \cdot (4\bar{p}^3 + 7\bar{p}^2) \cdot (2^{n_g} - 4) + (2s\bar{p}^2 + 12s^2\bar{p}) \cdot (3 \cdot 2^{n_g})\right) \\ &\quad \le 9 \cdot \left(6(4\bar{p}^3 + 7\bar{p}^2) + (2s\bar{p}^2 + 12s^2\bar{p})\right)\ 2^{n_g} \\ &\quad = 18\left(12\bar{p}^3 + 21\bar{p}^2 + s\bar{p}^2 + 6s^2\bar{p})\right)\ 2^{n_g} \\ &\quad = 18\left\{\frac{12\bar{p}}{s} + \frac{21}{s} + 1 + \frac{6s}{\bar{p}}\right\}\ \bar{p}^2 N, \end{aligned} \tag{4.2}$$

where the factor 3 takes the three types of two-variables moments into account, i.e., $M^2_{\phi_1,\phi_2,\alpha}$, $M^2_{\phi_2,\phi_3,\beta}$, and $M^2_{\phi_3,\phi_1,\gamma}$.

Note that for p large enough, choosing $s = \bar{p}$ (and $N = 2^{n_g}s$), this expression leads to the approximated cost bound $342\ \bar{p}^2 N$.

Computing the three-variable integrals. Using the basic partition (3.5), we know that the sum I_{dis} given by (3.1) can be split into three sets of contributions: E_k, F_k, and H_{n_g}. Here, we give the complexity of these contributions.

E_k integrals: These integrals are expressed by (3.10). The total cost of their evaluation is bounded by

$$c_E := 10\bar{p}^3 + 15\bar{p}^2 + \bar{p}. \tag{4.3}$$

Note that a direct computation of I_k would require $O(p^6)$ operations.

F_k integrals: These integrals are expressed by (3.11). The evaluation of an F_k integral needs a number of operations bounded by

$$c_F := 4\bar{p}^3 + 9\bar{p}^2.$$

H_{n_g} integrals: At level $k = n_g$, we make direct computations for neighboring intervals. The cost of the evaluation of an H_{n_g} integral is less than

$$(15s^3) \cdot 13 \cdot 2^{n_g} = 195\ s^2 N. \tag{4.4}$$

Recursion algorithm. We now proceed in the computation of the integrals corresponding to intervals which are in interaction at all levels, from level $k = n_g$ down to the level $k = 2$. The integrals of level k will be computed assuming that we have already computed the one- and two-variable moments at level k. Counting the number of occurrences of types E_k, F_k, and H_{n_g}, and using the previous costs, we get a total cost of the recursive algorithm (not including the cost induced by the computation of the moments) less than

$$456\ \left\{1 + \frac{2}{\bar{p}} + \frac{1}{19\bar{p}^2} + \frac{1}{6\bar{p}^3}\right\}\frac{\bar{p}^3}{s}\ N + 195\ s^2\ N. \tag{4.5}$$

Note that taking $s := \bar{p}$ large enough, this cost is equivalent to $651\ \bar{p}^2\ N$.

Total cost. The total cost for the evaluation of the integral I_{dis} is obtained by summing the previous costs (4.1), (4.2), and (4.5):

$$\begin{aligned} Total\ Cost \leq & \left\{ \frac{6}{\bar{p}} + \frac{44}{s} \right\}\ \bar{p}^2\ N \\ & + 18 \left\{ \frac{12\bar{p}}{s} + \frac{21}{s} + 1 + \frac{6s}{\bar{p}} \right\}\ \bar{p}^2\ N \\ & + 456\ \left\{ 1 + \frac{2}{\bar{p}} + \frac{1}{19\bar{p}^2} + \frac{1}{6\bar{p}^3} \right\} \frac{\bar{p}^3}{s}\ N + 195\ s^2\ N. \end{aligned} \tag{4.6}$$

Using the fact that $\max\{\frac{\bar{p}}{s}, 1, \frac{s}{\bar{p}}, \frac{s^2}{\bar{p}^2}\} = \max\{\frac{\bar{p}}{s}, \frac{s^2}{\bar{p}^2}\}$, and $s, \bar{p} \geq 1$, we obtain that the total cost is bounded by $O(\max\{\frac{\bar{p}}{s}, \frac{s^2}{\bar{p}^2}\} \cdot \bar{p}^2\, N)$. More precisely, we obtain the following bound:

$$Total\, Cost \leq 2800 \cdot \max \left\{ \frac{\bar{p}}{s}, \frac{s^2}{\bar{p}^2} \right\}\ \bar{p}^2\ N. \tag{4.7}$$

This concludes the proof of Theorem 3.1. □

Remark 4. For $s := \bar{p}$ we obtain from (4.6) the following asymptotic cost bound:

$$Total\, Cost \leq 993 \cdot \bar{p}^2\ N, \quad s = \bar{p} \to \infty. \tag{4.8}$$

Since a direct calculation of I_{dis} costs $15N^3$ (we count 14 elementary operations for the evaluation of the integrand term, plus 1 for the sum), in the case $s = \bar{p}$ we can predict that the FMM method is useful as soon as $993\bar{p}^2 N \leq 15N^3$, i.e., $N \geq 8.14\ \bar{p}$. For instance this gives $N \geq 82$ for $\bar{p} = 10$.

5. Numerical tests. We have chosen the functions $\phi_1(x) = \sqrt{x}$, $\phi_2(x) = \frac{x}{1+x}$, and $\phi_3(x) = \frac{x}{2+x}$, and the exponents $\alpha, \beta, \gamma = -0.2, -0.5, -0.9$. We have also taken a discretization mesh with n_g levels and $s_0 = 2$ points per interval at the finest level. Thus we have $N = 2^{n_g} s_0$ discretization points per variable. We shall also use the decomposition $N = 2^n s$ where $0 \leq n \leq n_g$ and $s = 2^{n_g - n} s_0$.

We have first tested the numerical behavior of the error with respect to the order p of the multipole expansion. In Table 5.1 and Figure 5.1 we give the absolute values of the errors $e(p) := |I_{dis} - I_{multipole}|$, versus p, in the case $n_g = 4$. The results are similar for other values of n_g. The relative error $|I_{dis} - I_{multipole}|/|I_{dis}|$ also decreases in the same way since here $I_{dis} \sim 0.86$ (with a $1/N^3$ normalization factor in the definition of I_{dis}). We note that best results are obtained for even values of p.

In particular, we observe that with $p = 10$ we have 7 correct digits, with $p = 20$ we have more than 11 correct digits, and with $p \geq 30$ we have more than 16 correct digits.

TABLE 5.1
Error versus the order p of multipole expansions (for p = 30 the error is smaller than 1E-16).

n_g	p	Error	Time
4.	0.	0.0133175	1.62
4.	5.	0.0000546	1.99
4.	10.	9.048E-08	2.60
4.	15.	1.355E-09	3.44
4.	20.	4.244E-12	4.63
4.	25.	6.661E-14	6.19
4.	30.	0	8.44

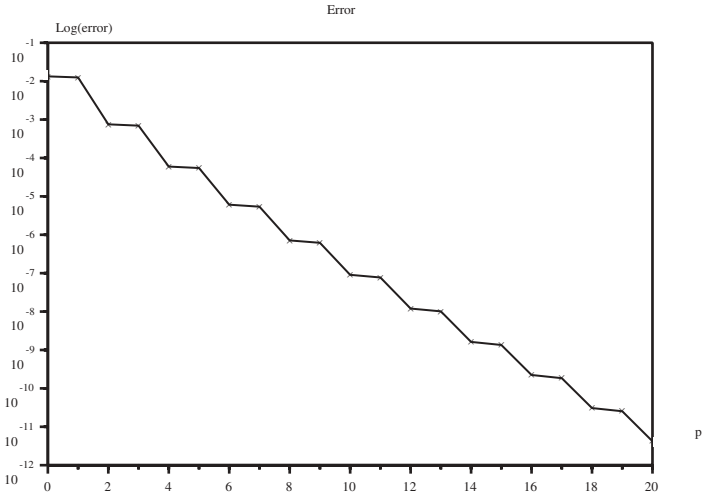


FIG. 5.1. *Error versus order p of multipole expansion.*

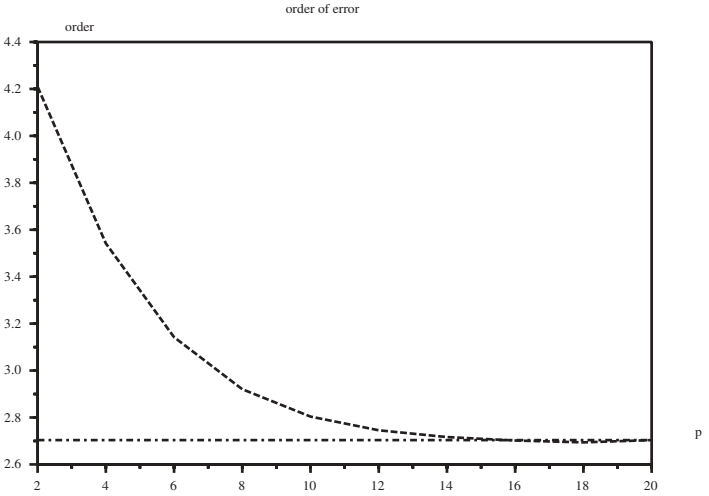


FIG. 5.2. $\sqrt{\frac{e(p-2)}{e(p)}}$ *(order of error) is plotted versus p (order of multipole expansions).*

In Figure 5.2, we have plotted $\sqrt{e(p-2)/e(p)}$, for even p, and observe that $\sqrt{e(p-2)/e(p)} \to 2.7$ for large p values. This means that the "observed" order of the method is approximately 2.7 (i.e., $e(p) \sim C/(2.7)^p$, for even p values). This is better than our theoretical bounds which predicts the order 2. The reason is that, in the theoretical analysis, the bounds for the multipole expansions are obtained in the worst case. This is a known fact in the usual FMM method [2] that can be used in order to improve the bounds.

We also mention that the CPU time increases very slowly with p (at least for small p values, $p \leq 40$). For instance with $p = 20$ it is roughly three times the CPU time of the case $p = 0$, and with $p = 10$ it is roughly 1.5 times the CPU time of $p = 0$. This is much faster than the predicted theoretical result: $O(\frac{p^3}{s}N)$ for fixed s

TABLE 5.2

CPU times. $t(n) := t_1 + t_2 + t_3$ *is the total time;* t_1, t_2 *are the computation times for the one- and two-variable moments, and* t_3 *for the three-variable integrals. Here* $p = 10$ *and* $n = n_g$*, excepted for the "optimized" multipole calculation where* $n = \max(n_g - 5, 0)$.

	$n = n_g$					$n = (n_g - 5)_+$	
n_g	t_1	t_2	t_3	$t(n)$	$\frac{t(n)}{2t(n-1)}$	$t(n)$	t_{direct}
2.	0.02	0.15	0.10	0.27		0.04	0.01
3.	0.03	0.35	0.57	0.95	1.7812	0.04	0.01
4.	0.13	0.75	1.72	2.60	1.3684	0.04	0.01
5.	0.22	1.63	4.30	6.15	1.1827	0.08	0.07
6.	0.45	3.38	9.62	13.45	1.0935	0.37	0.48
7.	0.92	6.90	20.95	28.77	1.0694	1.28	3.33
8.	1.80	13.92	42.80	58.51	1.0171	3.45	24.72
9.	3.60	28.10	89.03	120.73	1.0316	8.00	192.47
10.	7.08	57.10	180.98	245.16	1.0153	17.24	1515.90
11.	14.25	114.01	363.30	491.56	1.0025	36.08	12467.01
12.	28.35	230.09	729.77	988.21	1.0052	74.58	112829.63
13.	57.38	461.78	1478.54	1997.70	1.0108	152.12	1018518.50

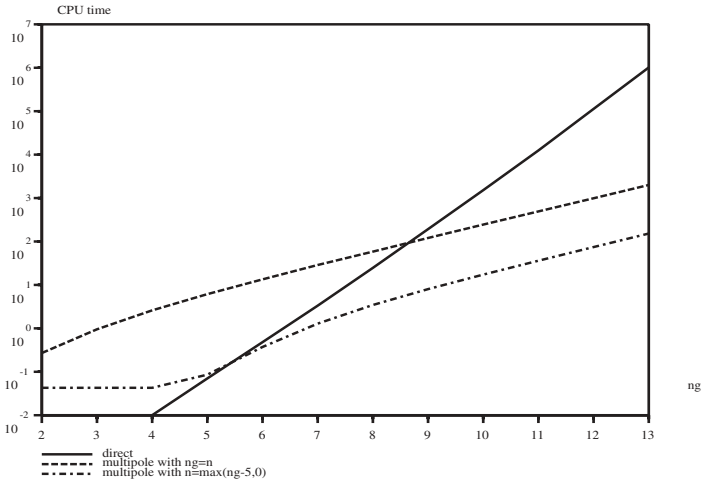


FIG. 5.3. *CPU time.*

and large p. In fact all moments have been computed using products of matrices and vectors, and thus are computed with a similar cost independently of p (for small p values, say $p \leq 30$).

In our second test we are interested in the numerical $O(N)$ behavior. In Table 5.2 and Figure 5.3, we have given the CPU time $t(n)$ versus the number of levels $n = \log_2(N/s)$, for different methods. The first method is with $n = n_g$ (and $s = s_0$); i.e., the number of levels n for the multipole method is the same as for the discretization mesh. We have fixed $p = 10$ for these calculations, since in practice the CPU time increases slowly with p as previously remarked. We see that we have $t(n)/[2t(n-1)] \sim 1$ for large n, which corresponds to an $O(N)$ behavior at this scale.

In Figure 5.3 the CPU time curves are in logarithmic scale for the time. The first curve concerns the CPU time obtained by *direct* calculations, simply doing a direct computation of the sum I_{dis} (using vectorized computations). For $n_g \geq 9$, this CPU time has been estimated (only by making a part of the calculation).

The second curve is the CPU time of our method, with $n = n_g$ and $p = 10$.

The third curve is obtained with $n = \max(n_g - 5, 0)$ and $s = 2^{n_g - n} s_0$. This means that we have used the first n levels for the multigrid hierarchy of the FMM method, and that, at level $k = n$, the moments and integrals are directly computed. At this level, there are $s = 2^{n_g - n} s_0$ points per interval (for $n_g \geq 5$, this leads to $s = 2^5 \cdot s_0 = 32 \cdot s_0$). This is like a multigrid method where a "rough" scale is used for the first levels $k \leq n$ (where we use the approximation of multipole expansions), and a refined scale is used at level $k = n$ (where exact computations are made).

We see that the multipole method (for $n = n_g$) becomes faster than the direct method for $n_g \geq 9$. The "optimized" multipole method, i.e., with $n = \max(n_g - 5, 0)$, is faster for $n_g \geq 5$. For instance, for $n_g = 10$ ($N^3 = 8 \cdot 2^{30} \sim 10^{10}$ points), the computation time is 17 s for the "optimized" FMM, which is roughly 90 times faster when compared to the direct computation, and 15 times faster than the FMM method ($n = n_g$). The fact that for small n_g values the time of the optimized multipole method is greater than the direct method is due to the initialization of one- and two-variable moments.

Note on computation and storage of moments. For each type of moments (one- and two-variable moments), and for the three-variable integrals, it is possible to use only one list of 2^{n_g} moments. Initially, the list contains the values of the moments at finest level n_g. Then, when we calculate recursively the moments of level $n_g - 1$, we can store them on the same list, using only the first $2^{n_g - 1}$ elements, and so on.

In order not to overwrite on the list of necessary moments, we proceed recursively as follows.

AT LEVEL $k = n_g$: Initialize the list of one- and second-order moments and the three-variable integrals.

FOR LEVELS $k = n_g - 1$ TO 0 STEP -1 DO:

• Computation of three-variable integrals at level k (requires at level $k + 1$ the one- and two-variable moments, and three-variable integrals). Storage in the list of three-variable integrals erases the three-variable integrals of level $k + 1$.

• Computation of two-variable moments at level k (requires one- and two-variable moments of level $k + 1$). Storage in the list of two-variable moments erases that of two-variable moments at level $k + 1$.

• Computation of one-variable moments at level k (requires one-variable moments of level $k + 1$). Storage erases the list of one-variable moments at level $k + 1$.

6. Concluding remarks. For $m \geq 2$, we have summarized in Table 6.1 some obtained results, with $N = 2^{n_g} \bar{p}$ and $\bar{p} = p + 1$. We recall that p is the order of the multipole expansions. The results between brackets are *conjectured*. (Here the dimension is $d = 1$ for each variable.) All the error bounds are of the form $O(\frac{1}{\bar{p}^\kappa 2^p})$, where $\kappa := \inf_{i<j}(\alpha_{ij}) + 1$.

TABLE 6.1
Costs and errors bounds for various numbers m of particles.

	$m = 2$	$m = 3$	$m = 4$	$m \geq 5$
Cost	$O(\bar{p}N)$	$O(\bar{p}^2 N)$	$O(\bar{p}^4 N)$	$\leq C_m \bar{p}^{\delta_m} N$
Error	$O\left(\frac{1}{\bar{p}^\kappa 2^p}\right)$	$O\left(\frac{1}{\bar{p}^\kappa 2^p}\right)$	$\left[O(\frac{1}{\bar{p}^\kappa 2^p})\right]$	$\left[O(\frac{1}{\bar{p}^\kappa 2^p})\right]$

For $m = 4$, we can prove that the cost of our FMM algorithm is $O(\bar{p}^4 N)$, using cost reduction techniques as in section 4, while the error is expected to be bounded

by $\frac{cste}{\bar{p}^\kappa 2^p}$.

For $m \geq 5$, there are $m(m-1)/2$ binary interactions in the integrand term. Again, we can prove a similar cost bound, of the form $C_m \bar{p}^{\delta_m} N$, where δ_m and C_m are some constants that depend only on m. However, we did not compute here an explicit estimate of δ_m when using cost reduction techniques.

For $d \geq 2$, i.e., for more than one dimension for the particle position, we expect a similar cost, i.e., $O(p^{\delta_{m,d}} N)$ (with $\delta_{m,d}$ being some constant). This task has not been addressed in this paper, but the simple case $d = 1$ should greatly help to treat the higher-dimensional case. In particular, one can see that the partition algorithm of the integration domain (developed in section 3.1) is independent of the variable dimension. This algorithm can be used for the three-dimensional case as it stands. However, the multipole expansions should be more technical for higher dimensions. This last point is under study and we think that this could be done using the techniques in [2] or in [3] where three-dimensional multipole expansions are developed. A complete three-dimensional algorithm is under investigation.

We also mention that it is possible to compute in $O(\bar{p}^{\mu_m} N)$ the following kind of one-particle operators:

$$v(x_1) = \int_{\mathcal{C}^{m-1}} \phi_2(x_2) \cdots \phi_m(x_m) \prod_{1 \leq i < j \leq m} |x_i - x_j|^{\alpha_{ij}} d\mu(x_2) \cdots d\mu(x_m)$$

(where x_1 takes values on a mesh with N points, and $d\mu$ is the discrete measure on the mesh). In this case, we have to use a "descending" algorithm as in the FMM of Greengard and Rokhlin [1].

Finally we mention that an alternative method based on the use of wavelet bases is the subject of a future work.

Acknowledgments. We thank P. Degond for showing us Greengard and Rokhlin's references. All tests have been done using the SCILAB programming language (http://scilab.inria.fr). We thank the SCILAB group, the *Saphir Control* team, and the *Medicis Lab* (http://www.medicis.polytechnique.fr) where part of the computations have been done.

REFERENCES

[1] L. GREENGARD AND V. ROKHLIN, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.

[2] L. GREENGARD AND V. ROKHLIN, *The rapid evaluation of potential fields in three dimensions*, in Vortex Methods, C. Anderson and C. Greengard, eds., Lecture Notes in Math. 1360, Springer-Verlag, New York, 1988.

[3] M. LEMOU, *Multipole expansions for the Fokker-Planck equation*, Numer. Math., 78 (1998), pp. 597–618.

[4] O. BOKANOWSKI AND M. LEMOU, *Fast multipole method for multidimensional integrals,* C.R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 105–110.

[5] L. GREENGARD, *The numerical solution of the N-body problem*, Comput. Phys., 4 (1990), pp. 143–152.

[6] L. GREENGARD AND J.-Y. LEE, *A direct adaptive Poisson solver of arbitrary order accuracy*, J. Comput. Phys., 125 (1996), pp. 415–424.

[7] G. BEYLKIN, R. COIFMAN, AND V. ROKHLIN, *Fast wavelet transforms and numerical algorithms* I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[8] C. A. WHITE, B. G. JOHNSON, P. M. W. GILL, AND M. HEAD-GORDON, *The continuous fast multipole method*, Chem. Phys. Lett., 230 (1994), pp. 8–16.

[9] C. A. WHITE, B. G. JOHNSON, P. M. W. GILL, AND M. HEAD-GORDON, *Linear scaling density function calculations via the continuous fast multipole method*, Chem. Phys. Lett., 253 (1996), pp. 268–278.

[10] C. J. Umrigar, K. G. Wilson, and J. W. Wilson, *Optimized trial wave-functions for quantum Monte-Carlo calculations*, Phys. Rev. Lett., 60 (1988), pp. 1719–1722.
[11] R. Assaraf and M. Caffarel, *Monte Carlo zero-variance principle for Monte Carlo algorithms*, Phys. Rev. Lett., 83 (1999), pp. 4682–4685.
[12] E. Cances, *Molecular Simulation and Environmental Effects: A Mathematical and Numerical Perspective*, Ph.D. thesis, Ecole Nationale di Ponts et Chanssées, 1998.
[13] W. Kutzelnigg and J. D. Morgan, *Rates of convergence of the partial-wave expansions of atomic correlation energies*, J. Chem. Phys., 96 (1992), pp. 4484–4508.
[14] W. Kutzelnigg and W. Klopper, *Wave functions with terms linear in the interelectronic coordinates to take care of the correlation cusp.* I. *General theory*, J. Chem. Phys., 94 (1991), pp. 1985–2001.
[15] W. Kutzelnigg and W. Klopper, *Wave functions with terms linear in the interelectronic coordinates to take care of the correlation cusp.* II. *Second-order Møller-Plesset (MP2-R12) calculations on closed-shell atoms*, J. Chem. Phys., 94 (1991), pp. 2002–2019.
[16] W. Kutzelnigg and W. Klopper, *Wave functions with terms linear in the interelectronic coordinates to take care of the correlation cusp.* III. *Second-order Møller-Plesset (MP2-R12) calculations on molecules of first row atoms*, J. Chem. Phys., 94 (1991), pp. 2020–2031.
[17] H. Müller, W. Kutzelnigg, and J. Noga, *A CCSD(T)-R12 study of the ten-electron systems* Ne, F^-, HF, H_2O, NH_3, NH_4^+ *and* CH_4, Mol. Phys., 92 (1997), pp. 535–546.
[18] E. Schwegler and M. Challacombe, *Linear scaling computation of the Fock matrix. IC. Multipole accelerated formation of the exchange matrix*, J. Chem. Phys., 111 (1999), pp. 6223–6229.
[19] J. P. Dombroski, S. W. Taylor, and P. M. W. Gill, *KWIK: Coulomb energies in* $O(N)$ *work*, J. Phys. Chem., 100 (1996), pp. 6272–6276.
[20] H. Cheng, L. Greengard, and V. Rokhlin, *A fast adaptive multipole algorithm in three dimensions*, J. Comp. Phys., 155 (1999), pp. 468–498.
[21] E. Darve, *The fast multipole method.* I: *Error analysis and asymptotic complexity*, SIAM J. Numer. Anal., 38 (2000), pp. 98–128.
[22] A. Khare and J.-M. Richard, *Testing Hall-Post inequalities with exactly solvable* N*-body problems*, J. Phys. A, 34 (2001), pp. L447–L452.
[23] S. Ten-no and O. Hino, *New transcorrelated method improving the feasibility of explicitly correlated calculations*, Int. J. Mol. Sci., 3 (2002), pp. 459–474.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2118–2135

FULLY LOCALIZED A POSTERIORI ERROR ESTIMATORS AND BARRIER SETS FOR CONTACT PROBLEMS*

RICARDO H. NOCHETTO†, KUNIBERT G. SIEBERT‡, AND ANDREAS VEESER§

Abstract. We derive novel pointwise a posteriori error estimators for elliptic obstacle problems which, except for obstacle resolution, completely vanish within the full-contact set (localization). We then construct a posteriori barrier sets for free boundaries under a natural stability (or nondegeneracy) condition. We illustrate localization properties as well as reliability and efficiency for both solutions and free boundaries via several simulations in 2 and 3 dimensions.

Key words. elliptic obstacle problem, stable free boundary, a posteriori error estimate, residual, maximum norm, maximum principle, barrier functions, barrier sets

AMS subject classifications. 65N15, 65N30, 35J85

DOI. 10.1137/S0036142903424404

1. Model problem, its discretization, and main results.

Free boundary problems are ubiquitous in applications, from nonlinear elasticity and plasticity to fluids and finance. The detection and accurate approximation of the free boundary is often a primary goal of the computation. There are, however, no results in the literature which provide a posteriori error estimates for interfaces. In case they are defined as level sets, then the mere control of the solution(s) does not yield, in general, control of the interfaces. In this paper we examine a model problem, namely the elliptic obstacle problem with Hölder obstacle, and derive novel *pointwise* a posteriori error estimates for both the solution and free boundary. The maximum norm is essential in this endeavor to convert error estimates for solutions into error estimates for interfaces. The estimators in turn exhibit complete localization (vanish within the *full-contact* set) and thus improve upon [15]. Their reliability and efficiency is assessed both theoretically and computationally herein.

We first introduce the *continuous* obstacle problem. Let Ω be a bounded, polyhedral, not necessarily convex domain in $\mathbb{R}^d$ with $d \in \{1, 2, 3\}$. Let $f \in L^\infty(\Omega)$ be a load function, $\chi \in H^1(\Omega) \cap C^{0,\alpha}(\bar{\Omega})$ be a lower obstacle, and $g \in H^1(\Omega) \cap C^{0,\alpha}(\bar{\Omega})$ be a Dirichlet boundary datum with $0 < \alpha \leq 1$. Both χ and g satisfy the compatibility condition

$$\chi \leq g \quad \text{on } \partial\Omega.$$

Let $\mathcal{K}$ be the following nonempty, closed, and convex subset of $H^1(\Omega)$:

$$\mathcal{K} := \{v \in H^1(\Omega) \mid v \geq \chi \text{ a.e. in } \Omega \text{ and } v = g \text{ on } \partial\Omega\}.$$

*Received by the editors March 6, 2003; accepted for publication (in revised form) January 6, 2004; published electronically February 25, 2005. This research was partially supported by the international cooperation NSF-DAAD grants INT-9910086 and INT-0129243 "Projektbezogene Förderung des Wissenschaftleraustauschs in den Natur-, Ingenieur- und den Sozialwissenschaften mit der NSF."
http://www.siam.org/journals/sinum/42-5/42440.html

†Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu, http://www.math.umd.edu/~rhn). The research of this author was partially supported by NFS grants DMS-9971450 and DMS-0204670.

‡Institut für Mathematik, Universität Augsburg, Universitätsstraße 14, D-86159 Augsburg, Germany (siebert@math.uni-augsburg.de, http://scicomp.math.uni-augsburg.de/Siebert/).

§Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini 50, 20133 Milano, Italy (andreas.veeser@mat.unimi.it). The research of this author was supported by Italian M.I.U.R. Cofin 2001 "Metodi numerici avanzati per equazioni alle derivate parziali di interesse applicativo."

The variational formulation of the continuous obstacle problem reads as follows:

$$u \in \mathcal{K}: \qquad \langle \nabla u, \nabla(u-v)\rangle \le \langle f, u-v\rangle \qquad \text{for all } v \in \mathcal{K}. \tag{1.1}$$

Hereafter, $\langle \varphi, \psi\rangle$ denotes the scalar product in $L^2(\Omega)$ as well as the duality pairing between $\mathring{H}^1(\Omega)$ and $H^{-1}(\Omega)$. It is well known that (1.1) admits a unique solution u [11, Theorem 6.2], [9, 17], which is also Hölder continuous [8]. The latter implies that the *contact set*

$$\Lambda := \{u = \chi\} := \{x \in \Omega \mid u(x) = \chi(x)\}$$

and the *free boundary* or *interface*

$$\mathcal{F} := \partial\{u > \chi\} \cap \Omega$$

are closed in Ω. We are interested in the numerical study of these two sets. To this end, we first approximate u by means of finite elements and, later on, we construct appropriate a posteriori *barrier sets* depending on data and the finite element solution u_h.

Given a shape-regular partition $\mathcal{T}_h$ of Ω, the set of nodes of $\mathcal{T}_h$ is denoted by $\mathcal{N}_h$, and the subset of interior nodes by $\mathring{\mathcal{N}}_h$. Let $\mathbb{V}_h$ indicate the space of continuous piecewise affine finite element functions over $\mathcal{T}_h$ and $\mathring{\mathbb{V}}_h := \mathbb{V}_h \cap \mathring{H}^1(\Omega)$. The nodal basis functions of $\mathbb{V}_h$ are given by $(\phi_z)_{z\in\mathcal{N}_h}$, and they form a *partition of unity* of Ω, that is, $\sum_{z\in\mathcal{N}_h} \phi_z = 1$. Let I_h be the Lagrange interpolation operator onto $\mathbb{V}_h$.

Let $\chi_h := I_h\chi$ be the discrete obstacle and let $I_h g$ be the discrete Dirichlet boundary datum. The discrete counterpart $\mathcal{K}_h$ of $\mathcal{K}$ is then

$$\mathcal{K}_h := \{v_h \in \mathbb{V}_h \mid v_h \ge \chi_h \text{ in } \Omega \text{ and } v_h = I_h g \text{ on } \partial\Omega\}.$$

Note that it is sufficient to check the unilateral constraint of $\mathcal{K}_h$ only at the nodes. The set $\mathcal{K}_h$ is nonempty, convex, closed but in general not a subset of $\mathcal{K}$ (nonconforming approximation). The *discrete* obstacle problem reads as follows:

$$u_h \in \mathcal{K}_h: \quad \langle \nabla u_h, \nabla(u_h - v_h)\rangle \le \langle f, u_h - v_h\rangle \qquad \text{for all } v_h \in \mathcal{K}_h. \tag{1.2}$$

Problem (1.2) admits a unique solution (use [9, 11] in the Hilbert space $\mathbb{V}_h$).

In section 2 we introduce a computable, second order estimator $\mathcal{E}_h$, and prove in particular the a posteriori error bound

$$\|u - u_h\|_{0,\infty;\Omega} \le \mathcal{E}_h. \tag{1.3}$$

Later in section 4 we couple (1.3) with the *nondegeneracy* condition $f + \Delta\chi \le -\lambda < 0$, to show an a posteriori error bound for interfaces, which roughly reads as follows:

$$\begin{aligned} &\textit{The strip } \{x \in \Omega \mid 0 < \operatorname{dist}(x, \{u_h > \chi_h + \mathcal{E}_h\}) < r_h\} \\ &\textit{of width } r_h \approx \sqrt{\mathcal{E}_h/\lambda} \textit{ contains the exact interface } \mathcal{F}. \end{aligned} \tag{1.4}$$

The reasoning behind these results is rather different from other a posteriori error analyses, except for [15]. Indeed, in section 2 we construct continuous barrier functions for the exact solution u upon correcting the discrete solution u_h via the Riesz representation of the *Galerkin functional*, a nonlinear residual appropriate in this context; cf. [15, 20]. The derivation concludes with an application of the *continuous* maximum principle, which imposes no constraint on the triangulations $\mathcal{T}_h$, in contrast to a priori analyses.

The main theoretical results, (1.3) and (1.4), exhibit the following salient features:

- *Full localization* (see also [7]): The residual estimator inside $\mathcal{E}_h$ vanishes in the discrete full-contact set where $u_h = \chi_h$ (see section 2.1 for its definition). In particular, if $\chi = \chi_h$ and $u = \chi = u_h$ on a finite element star, then the residual indicator vanishes on the star as well. This gives rise to a rare local upper bound and is an important improvement over [15]. Its computational impact is discussed and illustrated in section 3.2.
- *Reliability and efficiency*: In addition to the upper bound (1.3) (reliability), we establish local lower bounds for all estimators (efficiency); this is discussed in section 2.5 and confirmed numerically in section 3.1.
- *Partition of unity and star-based estimators*: The error analysis is based on the partition of unity $(\phi_z)_{z\in\mathcal{N}_h}$, and consequently, the residual estimators entering $\mathcal{E}_h$ are star-based; see also [7, 12].
- *Element residual oscillation*: The customary element residual $\|h^2 f\|_{0,\infty;\Omega}$ is replaced here by data oscillation on stars, namely $\max_{z\in\mathring{\mathcal{N}}_h} \|(f-\hat{f}_z)\phi_z\|_{0,\infty;\Omega}$, which is generically of higher order asymptotically. This is achieved via an additional cancellation provided directly by the partition of unity, without dealing with the element residual as in [14]; see also [7].
- *Barrier sets and interface estimates*: Two important issues must be emphasized. First, the maximum norm is the most adequate one to link solutions and interfaces. Second, the approximation of level sets is not a direct consequence of precise estimates for solutions. The missing ingredient is the *nondegeneracy* condition $f + \Delta\chi \le -\lambda < 0$ due to Caffarelli [2], which is also used in the a priori error analysis of free boundaries [1, 5, 13]. Our second main result (1.4) is a dual counterpart of the latter, and the first computable error estimate for interfaces.

This paper is organized as follows. In section 2 we introduce the concept of full-contact set along with exact and discrete multipliers associated with the unilateral constraint. We then define and study the Galerkin functional for (1.1) and use it to construct barrier functions, which eventually yield the desired pointwise a posteriori error estimates for solutions. In section 3, we present two numerical examples which document reliability, efficiency, and full localization properties of $\mathcal{E}_h$. We introduce the concept of barrier sets and derive a posteriori error estimates for free boundaries in section 4. We conclude in section 5 with two revealing numerical examples which not only corroborate the theory of section 4, but also provide support of its optimality.

2. Pointwise a posteriori error estimates. Intuitively, one expects that a discrete counterpart of the exact contact set Λ enters into the a posteriori estimate of $\|u-u_h\|_{0,\infty;\Omega}$. Crucial facts, such as the location of Λ, are encoded in the nonpositive functional $\sigma \in H^{-1}(\Omega) = \mathring{H}^1(\Omega)^*$ defined by

$$\langle \sigma, \varphi\rangle = \langle f, \varphi\rangle - \langle \nabla u, \nabla\varphi\rangle \qquad \text{for all } \varphi \in \mathring{H}^1(\Omega), \tag{2.1}$$

which plays the role of a *multiplier* for the unilateral constraint. In fact, we have $\sigma = f + \Delta\chi$ in the interior of the contact set $\Lambda = \{u = \chi\}$, where σ is typically < 0, and $\sigma = 0$ in the open *noncontact* set $\Omega \setminus \Lambda = \{u > \chi\}$. It is then not surprising that the a posteriori error analysis needs a multiplier σ_h that is associated with u_h—the discrete counterpart of σ.

2.1. Discrete full-contact set and multiplier. We first introduce some notation. Let J_h be the jumps of the normal derivatives of u_h across interior sides (nodes/edges/faces in 1 dimension/2 dimensions/3 dimensions, respectively). More

precisely, given a common side S of two different simplices T^+ and T^-, we have on S

$$J_h = [\![\partial_n u_h]\!] = \left[\nabla u_{h|T^+} - \nabla u_{h|T^-}\right] \cdot n,$$

where n is the normal of S that points from T^- to T^+. We denote the union of all interior sides (interelement boundaries) by Γ. For a node $z \in \mathcal{N}_h$, let $\omega_z = \mathrm{supp}(\phi_z)$ be the finite element star and $\gamma_z = \Gamma \cap \mathrm{int}\ \omega_z$ be the union of all interior sides in ω_z. We define

$$\mathcal{C}_h = \{z \in \mathcal{N}_h \mid u_h = \chi_h \text{ and } f \le 0 \text{ in } \omega_z,\ J_h \le 0 \text{ on } \gamma_z\}$$

to be the set of *full-contact nodes* and denote by

$$\Omega_h^0 = \left\{ x \in \Omega \mid \sum_{z \in \mathcal{C}_h} \phi_z(x) = 1 \right\}, \qquad \Omega_h^+ = \Omega \backslash \Omega_h^0$$

the discrete *full-contact set* and its complement. Furthermore, we set $\Gamma_h^0 = \Gamma \cap \Omega_h^0$ and $\Gamma_h^+ = \Gamma \cap \Omega_h^+$. We clearly have

$$z \in \mathcal{N}_h \setminus \mathcal{C}_h \implies \omega_z \subset \overline{\Omega_h^+} \text{ and } \gamma_z \subset \overline{\Gamma_h^+}. \tag{2.2}$$

Finally, let $\Pi_h : L_1(\Omega) \to \mathring{\mathbb{V}}_h$ be the interpolation operator of [3]; see also [16]. Such a Π_h is both *positivity preserving*, which helps construct $\sigma_h \le 0$, and *second order* accurate, which is crucial in dealing with the second order maximum norm error.

With these notations at hand, we define the *discrete multiplier* $\sigma_h \in H^{-1}(\Omega)$ by using the partition of unity $\langle \sigma_h, \varphi \rangle = \sum_{z \in \mathcal{N}_h} \langle \sigma_h, \varphi\, \phi_z \rangle$ and setting

$$\begin{aligned} \langle \sigma_h, \varphi\, \phi_z \rangle = &\int_{\Omega_h^0} f\, \varphi\, \phi_z + \int_{\Gamma_h^0} J_h \varphi\, \phi_z \\ &+ \int_{\Omega_h^+} f\, (\Pi_h \varphi)(z)\, \phi_z + \int_{\Gamma_h^+} J_h\, (\Pi_h \varphi)(z)\, \phi_z \end{aligned} \tag{2.3}$$

for all $z \in \mathcal{N}_h$ and $\varphi \in \mathring{H}^1(\Omega)$. Note that $\Pi_h \varphi$ is evaluated at the node z and is thus a *constant* for each $z \in \mathcal{N}_h$. Therefore, (2.2) gives

$$z \in \mathcal{N}_h \setminus \mathcal{C}_h \implies \langle \sigma_h, \varphi\, \phi_z \rangle = (\Pi_h \varphi)(z)\, s_z, \tag{2.4}$$

where s_z is a *nodal* multiplier:

$$s_z := \int_\Omega f\, \phi_z + \int_\Gamma J_h \phi_z, \qquad z \in \mathcal{N}_h. \tag{2.5}$$

It satisfies $s_z \le 0$ whenever $z \in \mathring{\mathcal{N}}_h \cup \mathcal{C}_h$. This follows from the definition of $\mathcal{C}_h$, if $z \in \mathcal{C}_h$, and from utilizing $v_h = u_h + \phi_z \in \mathcal{K}_h$ in (1.2), if $z \in \mathring{\mathcal{N}}_h$.

We point out that the definition of the discrete multiplier σ_h via (2.3) is crucially distinct from those in [7, 15], which are in turn quite different from each other. In particular, the finite element star ω_z in (2.3) is divided into two parts and thus, in contrast to [7, 15], "broken" nodal multipliers are involved.

Let us finish this section with the observation that the discrete multiplier σ_h is, as the exact multiplier σ of (2.1), nonpositive.

LEMMA 2.1 (sign of σ_h). *The discrete multiplier σ_h satisfies $\sigma_h \le 0$.*

Proof. Let $\varphi \in \mathring{H}^1(\Omega)$ be nonnegative. Since $\Pi_h\varphi \geq 0$ in Ω and $\Pi_h\varphi = 0$ on $\partial\Omega$, (2.4) implies $\langle \sigma_h,\, \varphi\, \phi_z\rangle \leq 0$ for all $z \in \mathcal{N}_h \setminus \mathcal{C}_h$. On the other hand, if $z \in \mathcal{C}_h$, then $f \leq 0$ in ω_z as well as $J_h \leq 0$ on γ_z by definition, whence $\langle \sigma_h,\, \varphi\, \phi_z\rangle \leq 0$ follows from (2.3). ∎

The multiplier σ_h is *not* a discrete function and is thus noncomputable. In evaluating our error estimator, we will only make use of the nodal multipliers s_z for $z \in \mathring{\mathcal{N}}_h \cup \mathcal{C}_h$. The properties of these computable multipliers are closely related to properties of σ_h (see Proposition 2.5 below).

2.2. Galerkin functional: Definition and properties. We are now in the position to define the Galerkin functional $\mathcal{G}_h \in H^{-1}(\Omega)$, which plays the role of the residual for (unconstrained) equations; see [15, 20]:

$$\begin{aligned}\langle \mathcal{G}_h,\, \varphi\rangle &:= \langle \nabla(u - u_h),\, \nabla\varphi\rangle + \langle \sigma - \sigma_h,\, \varphi\rangle \\ &= -\langle \nabla u_h,\, \nabla\varphi\rangle + \langle f - \sigma_h,\, \varphi\rangle \qquad \text{for all } \varphi \in \mathring{H}^1(\Omega).\end{aligned} \tag{2.6}$$

Integrating by parts and employing the partition of unity $(\phi_z)_{z\in\mathcal{N}_h}$, we obtain

$$\begin{aligned}\langle \mathcal{G}_h,\, \varphi\rangle &= \int_\Omega f\,\varphi - \int_\Omega \nabla u_h \nabla\varphi - \langle \sigma_h,\, \varphi\rangle = \int_\Omega f\,\varphi + \int_\Gamma J_h\,\varphi - \langle \sigma_h,\, \varphi\rangle \\ &= \sum_{z\in\mathcal{N}_h} \Bigg\{ \int_\Omega f\,\varphi\,\phi_z + \int_\Gamma J_h\,\varphi\,\phi_z - \int_{\Omega_h^0} f\,\varphi\,\phi_z - \int_{\Gamma_h^0} J_h\,\varphi\,\phi_z \\ &\qquad\qquad - \int_{\Omega_h^+} f\,(\Pi_h\varphi)(z)\,\phi_z - \int_{\Gamma_h^+} J_h\,(\Pi_h\varphi)(z)\,\phi_z \Bigg\} \\ &= \sum_{z\in\mathcal{N}_h} \left\{ \int_{\Omega_h^+} f\,[\varphi - (\Pi_h\varphi)(z)]\,\phi_z + \int_{\Gamma_h^+} J_h\,[\varphi - (\Pi_h\varphi)(z)]\,\phi_z \right\} \\ &= \int_{\Omega_h^+} f\left[\varphi - \sum_{z\in\mathcal{N}_h} (\Pi_h\varphi)(z)\,\phi_z\right] + \int_{\Gamma_h^+} J_h\left[\varphi - \sum_{z\in\mathcal{N}_h} (\Pi_h\varphi)(z)\,\phi_z\right] \\ &= \int_{\Omega_h^+} f\,[\varphi - \Pi_h\varphi] + \int_{\Gamma_h^+} J_h\,[\varphi - \Pi_h\varphi].\end{aligned}$$

This expression shows the effect of *full localization* of the Galerkin functional to the set Ω_h^+. The construction of $\tilde\sigma_h$ in [15] leads merely to a *partial* localization. The full localization of σ_h defined by (2.3) is due to the notion of full-contact nodes, which was introduced in [7] so as to achieve full localization in the context of a first order estimator. In [7, Remark 4.5] one finds also an argument that the sign conditions on f and J_h in the definition of $\mathcal{C}_h$ are crucial.

To exploit further cancellation properties, we introduce the constant values

$$\bar\psi_z = \begin{cases} \left(\int_{\Omega_h^+} \phi_z\right)^{-1} \int_{\Omega_h^+} \psi\,\phi_z & \text{if } \rho_z = 0, \\ 0 & \text{else} \end{cases}$$

for all $z \in \mathcal{N}_h$ and $\psi \in L_1(\Omega)$, where

$$\rho_z := \int_{\Omega_h^+} f\,\phi_z + \int_{\Gamma_h^+} J_h\,\phi_z.$$

Note that, in view of (1.2) and (2.2), we certainly have $\rho_z = 0$ if $u_h(z) > \chi_h(z)$ and perhaps by chance otherwise. Setting $\psi := \varphi - \Pi_h\varphi$, employing again the partition of unity, and the fact that $\bar{\psi}_z\, \rho_z = 0$ we can rewrite $\langle \mathcal{G}_h,\, \varphi\rangle$ as follows:

$$\langle \mathcal{G}_h,\, \varphi\rangle = \sum_{z\in\mathcal{N}_h} \int_{\Omega_h^+} f\,[\psi - \bar{\psi}_z]\phi_z + \int_{\Gamma_h^+} J_h\,[\psi - \bar{\psi}_z]\,\phi_z.$$

For nodes $z \in \mathcal{N}_h$ with $\rho_z = 0$, the value $\bar{\psi}_z$ is the weighted L^2-projection of ψ to the constant functions on $\omega_z \cap \Omega_h^+$. Hence, we can subtract a constant from the element residual at these nodes without altering the expression. In particular, we can write

$$\langle \mathcal{G}_h,\, \varphi\rangle = \sum_{z\in\mathcal{N}_h} \int_{\omega_z^+} [f - \hat{f}_z]\,[\psi - \bar{\psi}_z]\phi_z + \int_{\gamma_z^+} J_h\,[\psi - \bar{\psi}_z]\,\phi_z, \tag{2.7}$$

where

$$\omega_z^+ = \omega_z \cap \Omega_h^+, \qquad \gamma_z^+ = \gamma_z \cap \Omega_h^+, \tag{2.8}$$

and

$$\hat{f}_z = \begin{cases} \frac{1}{2}\big(\min_{\omega_z^+} f + \max_{\omega_z^+} f\big) & \text{if } \rho_z = 0,\\ 0 & \text{else.}\end{cases} \tag{2.9}$$

This shows that only the *oscillation* $f - \hat{f}_z$ of the interior residual f enters in the estimators on all stars with $\rho_z = 0$ and not f itself.

2.3. Galerkin functional: Estimates. In order to bound the pointwise error $\|u - u_h\|_{0,\infty;\Omega}$, we shall need an estimate of $\mathcal{G}_h$ in the dual norm

$$\|\mathcal{G}_h\|_{-2,\infty;\Omega} := \sup\big\{\langle \mathcal{G}_h,\, \varphi\rangle \mid \varphi \in \mathring{H}^1(\Omega) \cap W_1^2(\Omega) \text{ with } \|D^2\varphi\|_{0,1;\Omega} \le 1\big\}. \tag{2.10}$$

In what follows, the symbol "$\preccurlyeq$" stands for "$\le C$," where the generic constant C may depend on the shape-regularity of the partition $\mathcal{T}_h$, the domain Ω, and its dimension d. The starting point for estimating (2.10) is (2.7). For $\varphi \in \mathring{H}^1(\Omega) \cap W_1^2(\Omega)$ and $\psi = \varphi - \Pi_h\varphi$, the Bramble–Hilbert lemma and *second order* interpolation estimates for Π_h provide

$$\|\psi - \bar{\psi}_z\|_{0,1;\omega_z} \preccurlyeq h_z\|\nabla\psi\|_{0,1;\omega_z} = h_z\|\nabla(\varphi - \Pi_h\varphi)\|_{0,1;\omega_z} \preccurlyeq h_z^2\|D^2\varphi\|_{0,1;\mathcal{U}_h(\omega_z)}$$

and, with the additional use of a scaled trace theorem,

$$\|\psi - \bar{\psi}_z\|_{0,1;\gamma_z} \preccurlyeq h_z^{-1}\|\psi - \bar{\psi}_z\|_{0,1;\omega_z} + \|\nabla\psi\|_{0,1;\omega_z} \preccurlyeq h_z\|D^2\varphi\|_{0,1;\mathcal{U}_h(\omega_z)},$$

where $\mathcal{U}_h(\omega_z)$ is the union of all simplices of $\mathcal{T}_h$ having nonempty intersection with the star ω_z. In view of (2.7), the last two estimates yield the following one:

$$|\langle \mathcal{G}_h,\, \varphi\rangle| \preccurlyeq \max_{z\in\mathcal{N}_h} \Big(h_z^2\|(f - \hat{f}_z)\,\phi_z\|_{0,\infty;\omega_z^+} + h_z\|J_h\,\phi_z\|_{0,\infty;\gamma_z^+}\Big)\,\|D^2\varphi\|_{0,1;\Omega}. \tag{2.11}$$

Analogously, one obtains

$$|\langle \mathcal{G}_h,\, \varphi\rangle| \preccurlyeq \left(\sum_{z\in\mathcal{N}_h} h_z^p\|(f - \hat{f}_z)\,\phi_z\|_{0,p;\omega_z^+}^p + h_z\|J_h\,\phi_z\|_{0,p;\gamma_z^+}^p\right)^{1/p} \|\nabla\varphi\|_{0,p';\Omega}, \tag{2.12}$$

where $p' = p/(p-1)$ is the dual exponent of $p \in [1,\infty)$. The proof of (2.11)–(2.12) is straightforward and thus omitted; we refer to [15] for similar bounds for a slightly different Galerkin functional $\mathcal{G}_h$. Let $w \in \mathring{H}^1(\Omega)$ be the Riesz representation of $\mathcal{G}_h$,

$$w \in \mathring{H}^1(\Omega): \qquad \int_\Omega \nabla w \cdot \nabla \varphi = \langle \mathcal{G}_h, \varphi \rangle \quad \text{for all } \varphi \in \mathring{H}^1(\Omega). \tag{2.13}$$

The following estimate will be instrumental in section 2.5 and, compared with [15], exhibits extra localization and cancellation of the element residual. Since the argument is similar to that in [15], which in turn is based on linear theory [4, 14], we only sketch it here for completeness.

LEMMA 2.2 (properties of w). *The function w is Hölder continuous and satisfies*

$$\|w\|_{0,\infty;\Omega} \le C^* |\log h_{\min}|^2 \max_{z\in\mathcal{N}_h} \eta_z, \tag{2.14}$$

where $C^* > 0$ is an interpolation constant solely depending on mesh regularity, $h_{\min} := \min_{z\in\mathcal{N}_h} h_z$, and η_z is the star-based residual indicator

$$\eta_z := h_z^2 \|(f - \hat{f}_z)\,\phi_z\|_{0,\infty;\omega_z^+} + h_z\,\|J_h\,\phi_z\|_{0,\infty;\gamma_z^+}, \tag{2.15}$$

with ω_z^+, γ_z^+, and $\hat{f}_z$ defined in (2.8) *and* (2.9).

Proof. We first apply the classical Hölder estimate of De Giorgi and Nash to deduce that $w \in C^{0,\alpha}(\Omega)$ for $\alpha = 1 - d/p > 0$ and $\|w\|_{C^{0,\alpha}(\Omega)} \preccurlyeq \|\mathcal{G}_h\|_{-1,p;\Omega}$. Consequently, (2.12) yields $\|w\|_{C^{0,\alpha}(\Omega)} \preccurlyeq \big(\sum_{z\in\mathcal{N}_h} \zeta_z^p\big)^{1/p}$ with

$$\zeta_z := h_z\,\|(f - \hat{f}_z)\,\phi_z\|_{0,p;\omega_z^+} + h_z^{1/p}\,\|J_h\,\phi_z\|_{0,p;\gamma_z^+}.$$

Hence

$$|w(x_0) - w(x_1)| \preccurlyeq |x_0 - x_1|^\alpha \|w\|_{C^{0,\alpha}(\Omega)} \preccurlyeq |x_0 - x_1|^\alpha \left(\sum_{z\in\mathcal{N}_h} \zeta_z^p\right)^{1/p}. \tag{2.16}$$

To prove a bound for $|w(x_0)| = \|w\|_{0,\infty;\Omega}$, we first invoke the uniform cone property of Ω and find a ball $B \subset \Omega$ of radius $\rho = h_{\min}^\beta$ ($\beta \ge 1$ to be determined) such that $\operatorname{dist}(x_0, B) \preccurlyeq \rho$. We then introduce a regularized delta function δ supported in B and corresponding regularized Green's function $G \in \mathring{H}^1(\Omega)$ satisfying $-\Delta G = \delta$. As in [4, 14], we get

$$\|D^2 G\|_{0,1;\Omega} \preccurlyeq |\log h_{\min}|^2,$$

whence, for some $x_1 \in B$ and with the help of (2.11),

$$w(x_1) = \langle w, \delta\rangle = \langle \nabla w, \nabla G\rangle = \langle \mathcal{G}_h, G\rangle \preccurlyeq |\log h_{\min}|^2 \max_{z\in\mathcal{N}_h} \eta_z. \tag{2.17}$$

Fixing $p > d$ and choosing $\beta = \alpha^{-1}$, we deduce that

$$h_{\min}^{\alpha\beta} \zeta_z \preccurlyeq h_z \zeta_z \preccurlyeq \eta_z |\omega_z^+|^{1/p} \qquad \text{for all } z \in \mathcal{N}_h. \tag{2.18}$$

Since $|x_0 - x_1|^\alpha \preccurlyeq \rho^\alpha = h_{\min}$, combining (2.16) and (2.17) leads to (2.14). ∎

Remark 2.3. We point out that (2.14) improves upon estimates in [4, 14] for linear elliptic PDEs (corresponding to the situation when $u > \chi$ and u_h/χ_h) in four respects.

First, there is no structural assumption $h_{\max} \preccurlyeq h_{\min}^{\gamma}$ ($\gamma \geq 1$) on the partition $\mathcal{T}_h$. Second, the assumption $\|w\|_{0,\infty;\Omega} \succcurlyeq h_{\max}^2$ is totally circumvented via (2.16) and (2.18). Third, the residual f need not be globally continuous—these three assumptions may not be valid in the present context (see section 3.2 below). Fourth, the usual interior residual $h_z\|f\phi_z\|_{0,\infty;\omega_z}$ is *directly* replaced by data oscillation $h_z\|(f-\hat{f}_z)\phi_z\|_{0,\infty;\omega_z}$, which is asymptotically smaller for $f \in C^0(\Omega)$; this is in the spirit of [14, Corollary 7.2].

Remark 2.4. The estimate (2.14) stems from *linear* theory and the constant C^* is the usual interpolation constant of residual-type estimators, which is not explicity known but can be estimated with the help of numerical experiments. It is conceivable that a sharp constant-free estimator in the maximum norm could be designed, but we are unaware of such an undertaking. Note also that w is an auxiliary function, never to be approximated and thus only accessible with additional work.

2.4. Barrier functions. We now introduce the continuous barriers u_* (lower) and u^* (upper), and derive a posteriori comparison estimates via the *continuous* maximum principle, thereby imposing no geometric constraints on the mesh. This is in striking contrast to existing a priori error analyses.

Given a function v, let $v^+ = \max(v, 0)$ denote its nonnegative part.

PROPOSITION 2.5 (lower barrier). *Let u_* be the function*

$$u_* := u_h + w - \|w\|_{0,\infty;\Omega} - \|g - I_h g\|_{0,\infty;\partial\Omega} - \|(u_h - \chi)^+\|_{0,\infty;\Lambda_h}, \tag{2.19}$$

where Λ_h is the contact set

$$\Lambda_h := \bigcup \{\omega_z : z \in \mathring{\mathcal{N}}_h \cup (\mathcal{C}_h \cap \partial\Omega) \text{ and } s_z < 0\} \tag{2.20}$$

with s_z defined in (2.5). *Then u_* satisfies*

$$u_* \leq u \qquad \text{in } \Omega.$$

Proof. We split the proof into four steps.

1. Since

$$(u_* - u)_{|\partial\Omega} \;\leq\; (u_h - u)_{|\partial\Omega} - \|g - I_h g\|_{0,\infty;\partial\Omega} \;\leq\; 0,$$

the function $v := (u_* - u)^+$ satisfies

$$v_{|\partial\Omega} = 0. \tag{2.21}$$

We want to show that $\|\nabla v\|_{0,2;\Omega} = 0$ and then use (2.21) to conclude that $v = 0$.

2. In view of (2.19), (2.13), (2.6), and $\sigma \leq 0$, we can write

$$\begin{aligned} \|\nabla v\|_{0,2;\Omega}^2 &= \int_\Omega \nabla(u_* - u)\cdot\nabla v = \int_\Omega \nabla(u_h - u)\cdot\nabla v + \int_\Omega \nabla w\cdot\nabla v \\ &= \langle \sigma - \sigma_h,\, v\rangle \leq -\langle \sigma_h,\, v\rangle . \end{aligned} \tag{2.22}$$

It thus remains to show $\langle \sigma_h, v\rangle = 0$, i.e., $\langle \sigma_h, v\,\phi_z\rangle = 0$ for all $z \in \mathcal{N}_h$.

3. We now show that

$$s_z = 0 \text{ or } z \in (\mathcal{N}_h \cap \partial\Omega) \setminus \mathcal{C}_h \implies \langle \sigma_h, v\,\phi_z\rangle = 0.$$

First, consider $z \in \mathcal{C}_h$ with $s_z = 0$. By definition of $\mathcal{C}_h$, we have $J_h \le 0$ on γ_z and $f \le 0$ in ω_z. Hence,

$$0 = s_z = \int_{\omega_z} f\,\phi_z + \int_{\gamma_z} J_h\,\phi_z$$

implies in fact $J_h = 0$ on γ_z and $f = 0$ in ω_z. This yields

$$\begin{aligned}\langle \sigma_h,\, v\,\phi_z\rangle = &\int_{\omega_z\cap\Omega_h^0} f\,v\,\phi_z + \int_{\gamma_z\cap\Omega_h^0} J_h\,v\,\phi_z \\ &+ (\Pi_h v)(z)\left[\int_{\omega_z\setminus\Omega_h^0} f\,\phi_z + \int_{\gamma_z\setminus\Omega_h^0} J_h\phi_z\right] = 0.\end{aligned}$$

Next, for $z \in \mathcal{N}_h\backslash\mathcal{C}_h$ with $s_z = 0$, we directly obtain $\langle \sigma_h,\, v\,\phi_z\rangle = 0$ by (2.4). Finally, if $z \in (\mathcal{N}_h \cap \partial\Omega) \setminus \mathcal{C}_h$ is a boundary node not being in full contact, then (2.4) and $(\Pi_h v)(z) = 0$ give $\langle \sigma_h,\, v\,\phi_z\rangle = 0$.

4. It remains to show that there is no node $z \in \mathring{\mathcal{N}}_h \cup \mathcal{C}_h$ with $\langle \sigma_h,\, v\,\phi_z\rangle < 0$ and $s_z < 0$. Suppose that z were such a node. Then there would exist an $x \in \omega_z$ with $v(x) > 0$, whence the definitions of u_* and Λ_h give

$$u_h(x) > u(x) + \|(u_h - \chi)^+\|_{0,\infty;\Lambda_h} \ge \chi(x) + \|(u_h - \chi)^+\|_{0,\infty;\omega_z} \ge u_h(x).$$

This contradiction concludes the proof. ∎

PROPOSITION 2.6 (upper barrier). *The function*

$$u^* := u_h + w + \|w\|_{0,\infty;\Omega} + \|g - I_h g\|_{0,\infty;\partial\Omega} + \|(\chi - u_h)^+\|_{0,\infty;\Omega} \tag{2.23}$$

satisfies

$$u \le u^* \qquad \text{in } \Omega.$$

Proof. We proceed as in Proposition 2.5, dealing with $v := (u - u^*)^+ \in \mathring{H}^1(\Omega)$ and using $\sigma_h \le 0$ from Lemma 2.1. The crucial property $\langle \sigma,\, (u - u^*)^+\rangle = 0$ follows easily as in [15, Proposition 4.1]. ∎

2.5. Upper and lower bounds. Combining the results of sections 2.3 and 2.4, we can now establish an upper a posteriori error estimate.

THEOREM 2.7 (reliability). *Let (u, σ) be the continuous solution satisfying* (1.1) *and* (2.1)*, and let (u_h, σ_h) be the discrete solution satisfying* (1.2) *and* (2.3)*, respectively. Then the following global a posteriori upper bound holds:*

$$\max\left\{\|u - u_h\|_{0,\infty;\Omega}, \|\sigma - \sigma_h\|_{-2,\infty;\Omega}\right\} \le \mathcal{E}_h, \tag{2.24}$$

where $\|\cdot\|_{-2,\infty;\Omega}$ is defined in (2.10)*, the error estimator $\mathcal{E}_h$ is given by*

$$\begin{aligned}\mathcal{E}_h := {} & C_*|\log h_{\min}|^2\ \max_{z\in\mathcal{N}_h}\eta_z && \textit{localized residual}\\ & + \|(\chi - u_h)^+\|_{0,\infty;\Omega} + \|(u_h - \chi)^+\|_{0,\infty;\Lambda_h} && \textit{localized obstacle approx.}\\ & + \|g - I_h g\|_{0,\infty;\partial\Omega} && \textit{boundary datum approx.}\end{aligned}$$

C_* is twice the geometric constant C^* in (2.14)*, solely depending on mesh regularity, η_z is the star-based indicator defined in* (2.15)*, and Λ_h is defined in* (2.20)*.*

Proof. Combining Propositions 2.5 and 2.6, we obtain $u_* \le u \le u^*$, whence

$$\begin{aligned}\|u-u_h\|_{0,\infty;\Omega} \le 2\|w\|_{0,\infty;\Omega} + \|(\chi-u_h)^+\|_{0,\infty;\Omega} \\ + \|(u_h-\chi)^+\|_{0,\infty;\Lambda_h} + \|g-I_hg\|_{0,\infty;\partial\Omega}.\end{aligned}$$

Lemma 2.2 then yields (2.24) for $u-u_h$. Finally, we resort to (2.6), namely,

$$\langle \sigma-\sigma_h,\,\varphi\rangle = \langle \mathcal{G}_h,\,\varphi\rangle + \langle u-u_h,\,\Delta\varphi\rangle \qquad \text{for all } \varphi\in \mathring{H}^1(\Omega)\cap W_1^2(\Omega),$$

and make use of (2.11) in conjunction with the bound above for $\|u-u_h\|_{0,\infty;\Omega}$ to derive the remaining estimate for $\sigma-\sigma_h$. ∎

The latter observation is important for the establishment of lower bounds and underlines the significance of the Galerkin functional. The global upper bound of Theorem 2.7 is of *optimal order* because the computable quantities therein are (locally) bounded by the combined error $\|u-u_h\|_{0,\infty;\Omega} + \|\sigma-\sigma_h\|_{-2,\infty;\Omega}$ and data approximation as follows.

Remark 2.8. The only constant in (2.24), C_* in $\mathcal{E}_h$, comes from linear theory (see Remark 2.4). The nonlinear theory of sections 2.4 and 2.5, which accounts for the highly nonlinear effects of the unilateral constraint, is thus constant-free.

THEOREM 2.9 (efficiency). *The following local lower bounds hold for any $z\in\mathcal{N}_h$ and $T\in\mathcal{T}_h$:*

$$\begin{aligned} h_z\|J_h\phi_z\|_{0,\infty;\gamma_z^+} &\preccurlyeq \|u-u_h\|_{0,\infty;\omega_z^+} + \|\sigma-\sigma_h\|_{-2,\infty;\omega_z^+} + h_z^2\|f-\hat f_z\|_{0,\infty;\omega_z^+}, \\ \|I_hg-g\|_{0,\infty;T\cap\partial\Omega} &\le \|u-u_h\|_{0,\infty;T}, \qquad \|(\chi-u_h)^+\|_{0,\infty;T} \le \|u-u_h\|_{0,\infty;T}, \end{aligned}$$

and, if $\omega_z\subset\Lambda_h$,

$$\begin{aligned}\|(u_h-\chi)^+\|_{0,\infty;\omega_z} \preccurlyeq \|u-u_h\|_{0,\infty;\omega_z} + \|\sigma-\sigma_h\|_{-2,\infty;\omega_z} + h_z^2\|f-\hat f_z\|_{0,\infty;\omega_z} \\ + \|(\chi_h-\chi)^+\|_{0,\infty;\omega_z} + \|\,[\![\partial_n\chi_h]\!]\,\|_{0,\infty;\gamma_z}.\end{aligned}$$

The proofs of these estimates are very similar to those of the corresponding lower bounds in [15, section 6] and are therefore omitted. The efficiency predicted by these estimates is corroborated computationally in section 3.

3. Numerical experiments I: Pointwise error. In this section we present a couple of insightful examples computed with the finite element toolbox ALBERT of Schmidt and Siebert [18, 19]. This code implements a bisection algorithm for refinement and thus guarantees mesh regularity. In each iteration of the adaptive algorithm, the solver for the resulting complementary problem is the projective nonlinear SOR analyzed in [6].

The factor $C_*|\log h_{\min}|$ of Theorem 2.7 is, in practice, replaced by $C^* = 0.02$. This choice is consistent with (2.24) for meshes with reasonable shape-regularity and moderate $h_{\min}$. For the computation of the maximum norm, functions are evaluated at the element Lagrange nodes corresponding to polynomials of degree 7. The marking strategy for refinement is based on the maximum norm criterion.

3.1. Madonna's obstacle: Reliability and efficiency. Let $\Omega := (-1,1)^2$ and the obstacle χ be the upward cone with tip at $x_0 = (\frac{3}{8},\frac{3}{8})$ and slope $m = 1.8$:

$$\chi(x) = 1 - m|x-x_0|.$$

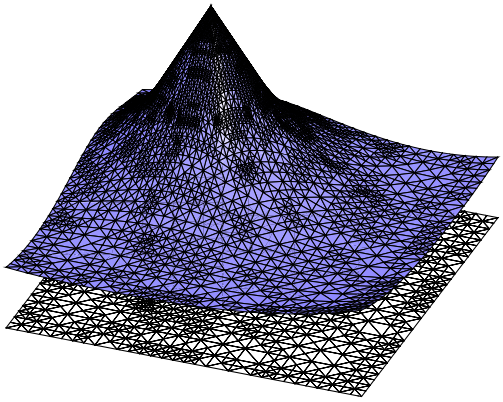


FIG. 3.1. *Madonna's obstacle: Graph and grid of the discrete solution for adaptive iteration* 14.

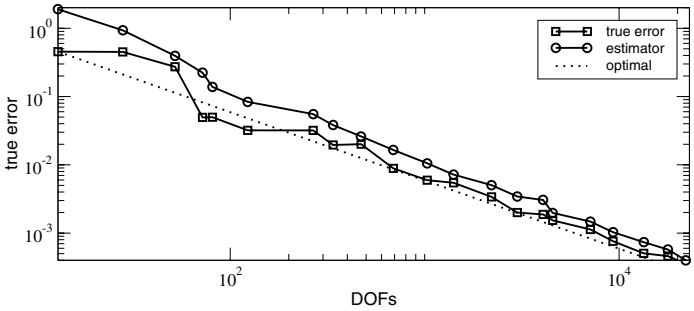


FIG. 3.2. *Madonna's obstacle: Equivalence of estimator* $\mathcal{E}_h$ *and true pointwise error* $\|u - u_h\|_{0,\infty;\Omega}$. *The optimal decay is indicated by the dotted line with slope* -1.

The exact solution is radially symmetric with respect to x_0, vanishes at $|x - x_0| = \frac{1}{m}$ and has a first order contact with the obstacle at $|x - x_0| = \frac{1}{2m}$; this corresponds to height $\frac{1}{2}$ (see Figure 3.1). The obstacle is thus singular within the contact set, due to the upward tip, which leads to local refinement. Several meshes are displayed in Figure 5.2 below.

Since we know the exact solution u, this example allows for a precise computational study of the estimator $\mathcal{E}_h$. Figure 3.2 displays both $\mathcal{E}_h$ and $\|u - u_h\|_{0,\infty;\Omega}$ versus the number of degrees of freedom (DOFs), and clearly demonstrates the equivalence between them. This result is consistent with Theorems 2.7 (reliability) and 2.9 (efficiency) and confirms their optimality.

3.2. Pyramid obstacle: Full localization. We now consider the same pyramid obstacle as in [15, section 7.4], namely

$$\chi(x) := \operatorname{dist}(x, \partial\Omega) - \tfrac{1}{5},$$

$f = -5$ and $g = 0$ on the square domain $\Omega := \{x \mid |x|_1 < 1\}$; see Figure 3.3. We show the dramatic effect of full localization of $\mathcal{E}_h$ in Figure 3.4, which exhibits coarse meshes within the full contact set (bottom row) in striking contrast to recent results

from [15] (top row). In addition, the new estimator is sharper with respect to the maximum norm than that in [15] and thus yields much fewer DOFs for about the same accuracy.

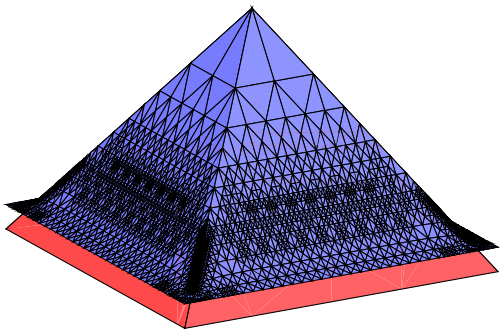


FIG. 3.3. *Pyramid obstacle: Graph with grid of the discrete solution over the obstacle for adaptive iteration* 10*, displaying lack of refinement along the diagonals inside the full-contact set (effect of full localization).*

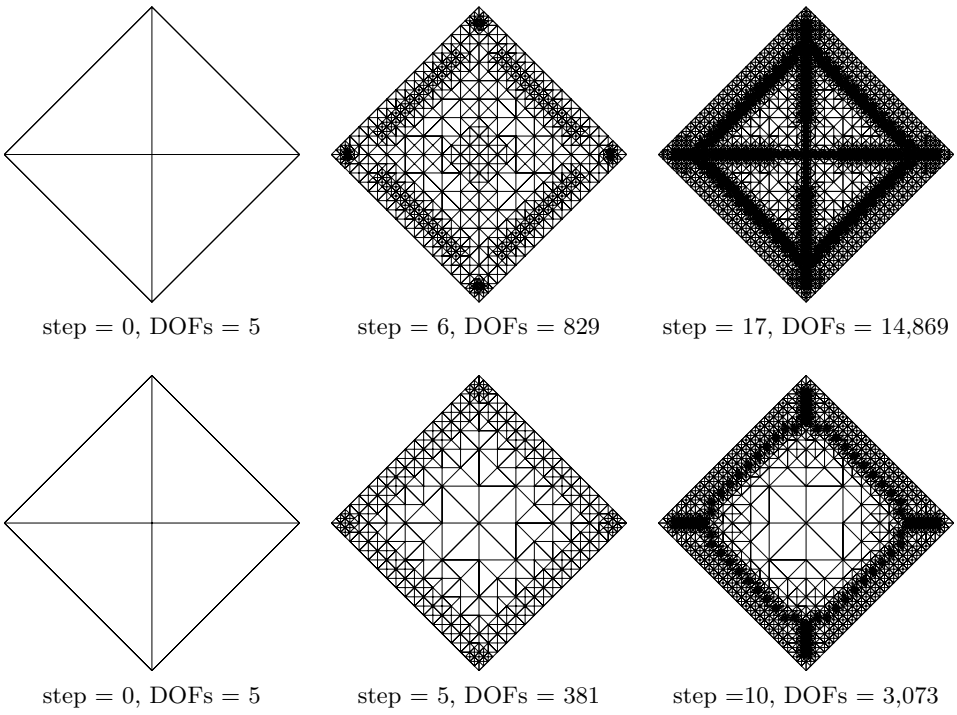


FIG. 3.4. *Pyramid obstacle: Comparison of grids obtained with the partially localized estimator of* [15] *(top) and the fully localized estimator (bottom). The meshes on the same column correspond to about the same value of the estimator, whereas the number of DOFs are much reduced with the new approach. The benefits of full localization are apparent since the refinement on the diagonals in contact is avoided.*

4. A posteriori barrier sets. The error in the approximation of σ is related to some "weak distance" of the exact contact set Λ and an appropriate approximation;

cf. [20, Remark 3.2]. However, the fact that the estimator $\mathcal{E}_h$ controls the pointwise error $\|u-u_h\|_{0,\infty;\Omega}$ allows, in certain situations, for more accurate a posteriori information on Λ and also on the exact free boundary (or interface) $\mathcal{F}$. This topic is the main concern of this section.

We consider the nondegenerate situation when one knows $\lambda > 0$ such that

$$\langle f,\, \varphi\rangle - \langle \nabla\chi,\, \nabla\varphi\rangle \;\le\; -\lambda \int_\Omega \varphi \tag{4.1}$$

for all $\varphi \in \mathring{H}^1(\Omega)$ with $\varphi \ge 0$. Condition (4.1) guarantees *stability* of the exact free boundary $\mathcal{F}$ and is due to Caffarelli [2]; see also [9, section 2.10]. Moreover, (4.1) implies

$$\sup_{B(x;r)} (u-\chi) \ge u(x) - \chi(x) + \frac{\lambda r^2}{2d} \tag{4.2}$$

for any $x \in \overline{\{u>\chi\}}$ and any $r>0$ such that $B(x;r) \subset \Omega$. Its proof proceeds along the same lines as that of Lemma 3.1 in [9, Chapter 2].

Let us define $K := \{\mathrm{dist}(\cdot, \partial\Omega) \ge r_h\}$ and the *barrier sets*

$$\Lambda^* := \{u_h \le \chi + \mathcal{E}_h\}, \qquad \Lambda_* := \Big\{\mathrm{dist}\big(\cdot, \{u_h \ge \chi_h + \mathcal{E}_h\}\big) \ge r_h\Big\}, \tag{4.3}$$

where

$$r_h^2 := \frac{2d}{\lambda}\big(2\mathcal{E}_h + \|(\chi_h - \chi)^+\|_{0,\infty;\{u_h \le \chi_h + \mathcal{E}_h\}}\big). \tag{4.4}$$

The following result, based on Theorem 2.7 and (4.2), locates the exact contact set Λ and the free boundary $\mathcal{F}$ a posteriori.

THEOREM 4.1 (a posteriori control of contact set and interface). *The set Λ^* is an upper barrier set for the exact contact set $\Lambda = \{u=\chi\}$, i.e., $\Lambda \subset \Lambda^*$.*

Moreover, if the stability condition (4.1) *holds, then the set Λ_* is a lower barrier set for Λ in the sense that $\Lambda_* \cap K \subset \Lambda \cap K$, whence*

$$\mathcal{F} \cap K \;\subset\; \big(\Lambda^* \cap K\big) \setminus \mathrm{int}\big(\Lambda_* \cap K\big).$$

Remark 4.2 (conditioning). In the light of (4.2), λ dictates the quadratic growth of $u-\chi$ in the noncontact set away from the free boundary $\mathcal{F}$, and so the larger λ, the more stable $\mathcal{F}$; that is, λ acts as a measure of conditioning of the free boundary. Correspondingly, the thicknesses of the strips $\Omega\setminus K$ and $\big(\Lambda^*\cap K\big)\setminus \mathrm{int}\big(\Lambda_*\cap K\big)$ depend inversely on λ.

Remark 4.3 (existence of exact interface). Suppose that condition (4.1) holds. Then $\Lambda_* \cap K \ne \emptyset$ implies $\Lambda \ne \emptyset$. Moreover, $\Lambda_* \cap K \ne \emptyset$ and $\Omega \setminus \Lambda^* \ne \emptyset$ imply $\mathcal{F} \ne \emptyset$.

Proof of Theorem 4.1. We first prove $\Lambda \subset \Lambda^*$. We use Theorem 2.7 with $x \in \Lambda$

$$u_h(x) = u(x) + \big[u_h(x) - u(x)\big] \le \chi(x) + \mathcal{E}_h$$

to deduce $x \in \Lambda^*$. We next prove $\Lambda_* \cap K \subset \Lambda \cap K$ provided that (4.1) holds. Let $x \in \Lambda_* \cap K$ and suppose that

$$u(x) > \chi(x). \tag{4.5}$$

Then, the definition of Λ_* in (4.3) implies that

$$u_h \le \chi_h + \mathcal{E}_h \qquad \text{in } \overline{B(x;r_h)} \tag{4.6}$$

holds and (4.2) yields

$$\sup_{B(x;r_h)} (u-\chi) > \frac{\lambda r_h^2}{2d}. \tag{4.7}$$

Consequently, Theorem 2.7, (4.4), and (4.7) give for some point $y \in \overline{B(x;r_h)}$:

$$\begin{aligned} u_h(y) &= u(y) + \big[u_h(y) - u(y)\big] > \chi(y) + \frac{\lambda r_h^2}{2d} - \mathcal{E}_h \\ &= \chi(y) + 2\mathcal{E}_h + \|(\chi_h - \chi)^+\|_{0,\infty;\{u_h \le \chi_h + \mathcal{E}_h\}} - \mathcal{E}_h \ \ge\ \chi_h(y) + \mathcal{E}_h. \end{aligned}$$

This contradicts (4.6) and so (4.5) is false. Consequently, $x \in \Lambda$ as asserted. ∎

Remark 4.4 (estimate in distance). We stress that Theorem 4.1 relies solely on (4.2) and not on estimating the measure of $\{0 < u - \chi < \epsilon\}$, the so-called *nondegeneracy property* of Caffarelli [2]. This leads, in the a priori error analysis for $\chi = \chi_h$, to estimates in measure for the discrete free boundary relative to $\mathcal{F}$ [1, 13]. Bounds in distance require regularity of $\mathcal{F}$ [1, 5, 13]. We locate here $\mathcal{F}$ relative to

$$\mathcal{F}_h = \partial\{u_h > \chi_h + \mathcal{E}_h\} \cap \Omega.$$

This dual approach yields estimates *in distance* without regularity assumptions on the exact free boundary $\mathcal{F}$.

Remark 4.5 (computation of effective condition number). Statement (4.6) reveals that (4.1) is needed in the proof of Theorem 4.1 only for positive test functions φ with $\operatorname{supp}\varphi \subset \{u_h \le \chi_h + \mathcal{E}_h\}$. Therefore, if $\chi \in H^2(\Omega)$, one can adaptively compute the condition number λ by

$$\lambda = - \sup_{\{u_h \le \chi_h + \mathcal{E}_h\}} (f + \Delta\chi).$$

Remark 4.6 (Computation of barrier sets). *The argument for Theorem* 4.1 itself is "constant-free." Therefore, the only not explicitly known quantity entering the definition of the barrier sets Λ^* and Λ_* is the constant C^* mentioned in Remark 2.4.

5. Numerical experiments II: Free boundaries. In this section we present several numerical experiments illustrating the impact of the a posteriori barrier sets in section 4 on the numerical study of exact free boundaries.

5.1. Madonna's obstacle: Reliability and efficiency. Let us reconsider the example from section 3.1, this time focusing on the approximation of the exact free boundary $\mathcal{F} = \{x \in \Omega \mid |x - x_0| = \frac{1}{2m}\}$. The condition number λ which enters the definition of r_h in (4.4), and thus the one of Λ_*, is computed according to Remark 4.5. Figure 5.1 depicts the true distance $\operatorname{dist}(\mathcal{F}, \mathcal{F}_h)$ between $\mathcal{F}$ and $\mathcal{F}_h = \partial\{u_h > \chi_h + \mathcal{E}_h\}$ together with r_h versus the number of DOFs; the number r_h essentially measures the gap between the two barrier sets. Both quantities decay with optimal order. Their behavior corroborates the reliability statement of Theorem 4.1 and, furthermore, reveals nice efficiency properties of r_h, which are not explained by the theory of section 4. Note also that, for the final computations the two barrier sets are quite close, i.e.,

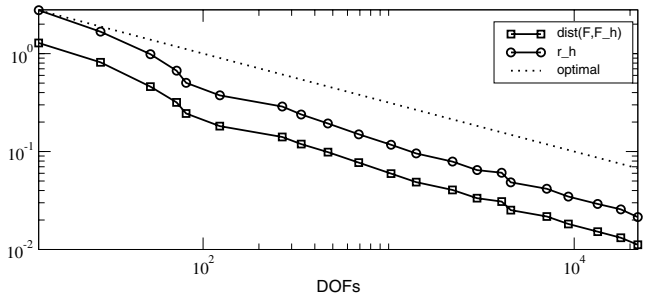


FIG. 5.1. *Madonna's obstacle: Equivalence of dist$(\mathcal{F}, \mathcal{F}_h)$ and the distance r_h of the barriers. The optimal decay is indicated by the dotted line with slope $-1/2$.*

$r_h \approx 0.02$. The grids and interface barriers in Figure 5.2 illustrate different stages in the information about the exact free boundary: the very coarse grid of the first column only indicates a possible exact free boundary; the still quite coarse grid of the second column assures the existence of the free boundary within the a posteriori annulus $\Lambda^* \backslash \Lambda_*$ (see Remark 4.3) and suggests that it might be a circle; the latter is further confirmed by the finer grid of the third column and corresponding better interface resolution.

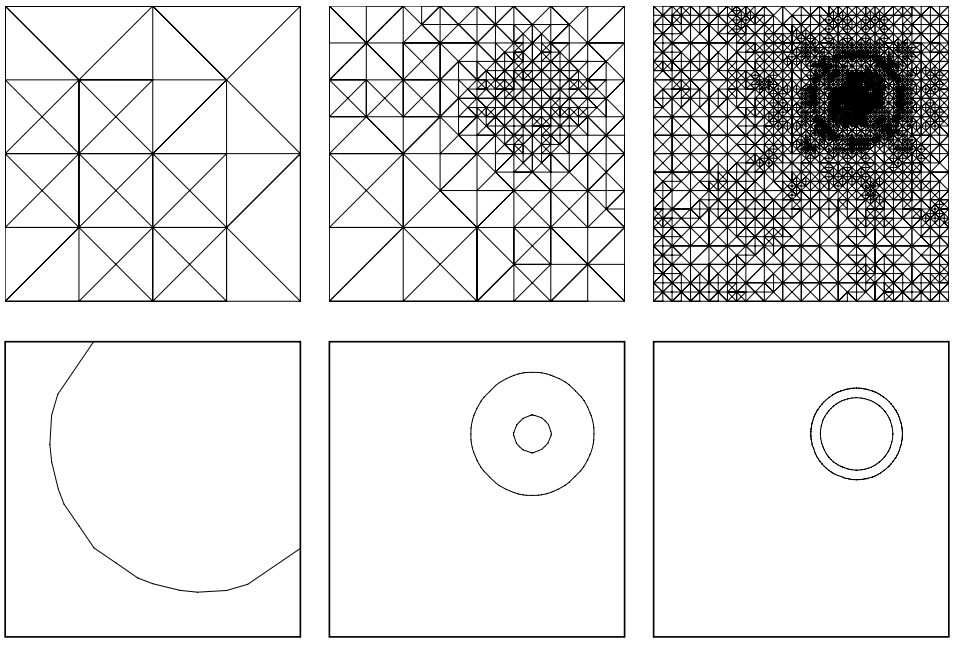


FIG. 5.2. *Madonna's obstacle: grids and interface barriers obtained by the adaptive algorithm in steps* 1, 6, *and* 13.

5.2. From balls to bones. We consider the domain $\Omega = (-2, 2) \times (-1, 1)^{d-1}$, boundary value $g = 0$, several constant loads f, and the smooth obstacle

$$\chi(x) = \alpha - \beta(x_1^2 - 1)^2 - \gamma(|x|^2 - x_1^2) \tag{5.1}$$

with α=10, β=6, γ=20 in 2 dimensions, and α=5, β=6, and γ=30 in 3 dimensions. In 2 dimensions the graph of the obstacle consists of two hills connected by a saddle.

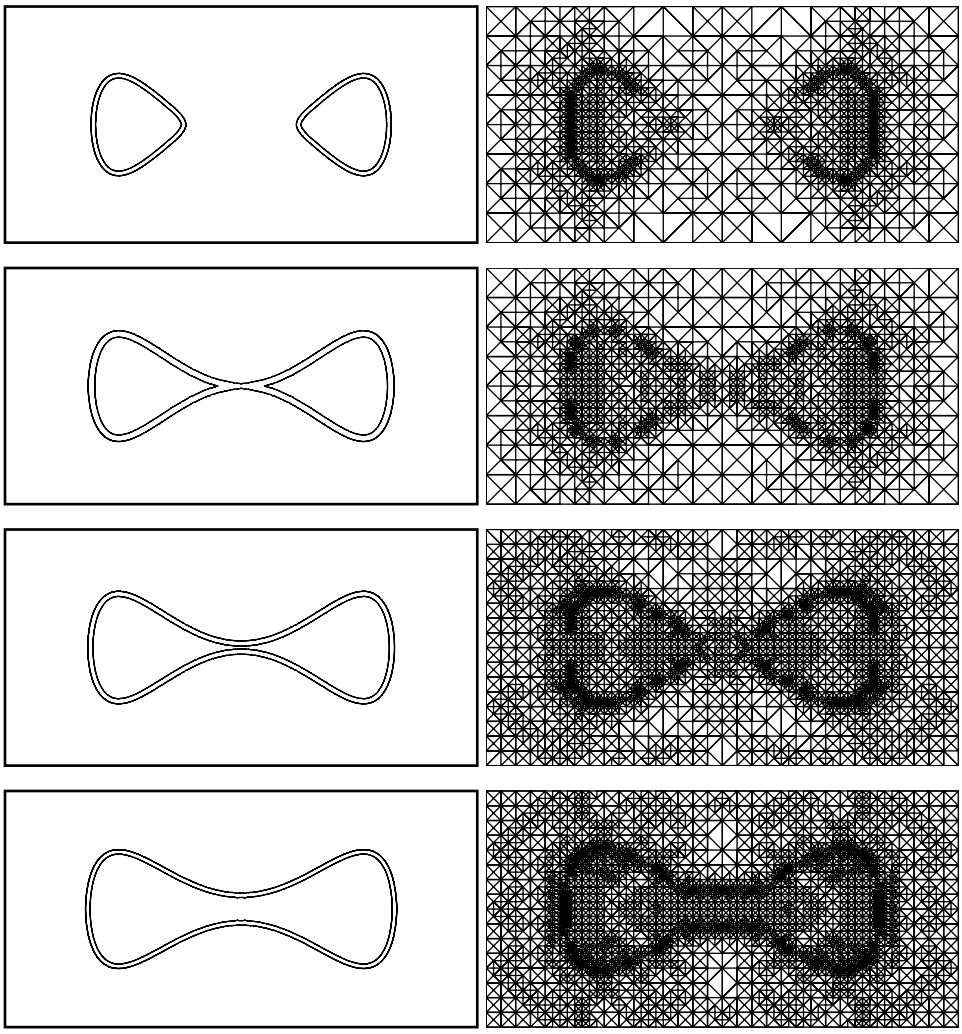


FIG. 5.3. *From balls to bones: Interface barriers for tolerance* ≈ 0.01 *(left) and adaptive grids for tolerance* ≈ 0.15 *(right) in* 2 *dimensions for forcing term* $f = 0, -5.9, -8.1, -15$ *(from top to bottom). The distance of the barriers is* ≈ 0.05 *for all four forces.*

In what follows, "barrier sets for tolerance $\approx \tau$" (> 0) denote those barrier sets which are constructed in the first adaptive iteration with $\mathcal{E}_h \leq \tau$. The left column in Figure 5.3 illustrates the interface barriers for four constant loads $f = 0$, $-5.9, -8.1, -15$ in 2 dimensions for about the same tolerance $\tau \approx 0.01$; the exterior curves correspond to $\partial\Lambda^*$ whereas the interior curves display $\partial\Lambda_*$. For $f = 0$, the contact set does not contain the saddle, whereas, for $f = -15$, it does. This happens because the solution, being pushed down by f, adheres longer to the obstacle. During the transition between these two extreme cases, the free boundary has a singular point, namely a "double-cusp" at the origin, for some critical value f_{crit}. The barrier sets constructed in section 4 from the discrete solution and the estimator give a reli-

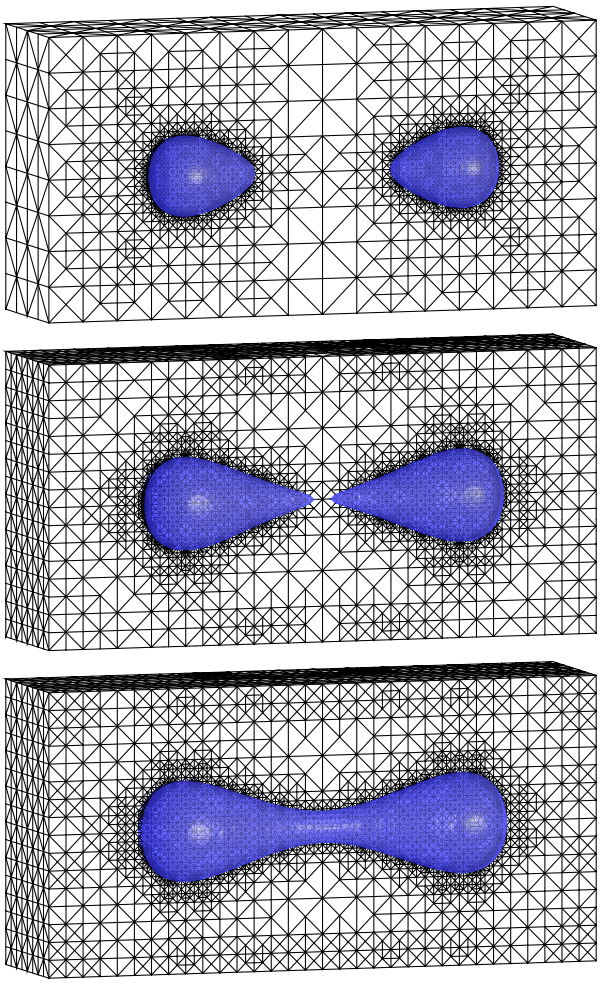


FIG. 5.4. *From balls to bones: Upper barrier* $\partial\Lambda^* \cap \Omega$ *and adaptive grids in* 3 *dimensions for tolerance* ≈ 0.25 *and* $f = 0, -8.5, -15.1$. *The distance of the barriers is* ≈ 0.1 *for all three forces.*

able range for f_{crit}: as long as Λ^* does not contain the saddle and $\Lambda_* \neq \emptyset$, the true contact set Λ exists and does not contain the saddle; this happens for $0 \geq f > -5.9$. For $f < -8.1$, the lower barrier Λ_* contains the saddle and exhibits a dumbbell shape, and so does Λ; hence, $f_{\text{crit}} \in (-5.9, -8.1)$. The size of this interval depends on the size of the estimator $\mathcal{E}_h$ and decreases for smaller values of $\mathcal{E}_h$, as documented in Table 5.1. Although the true interface develops a singularity, it is worth noticing that $u \in W^2_\infty(\Omega)$ and thus no special refinement is needed to approximate either u (or σ). Moreover, $f + \Delta\chi \leq -16$ in Ω for the 4 loads, which shows that the double-cusp is not due to lack of stability. The interface estimate of Theorem 4.1 thus applies and provides a posteriori error control of the entire free boundary including the double-cusp.

A similar situation occurs in 3 dimensions, as depicted in Figure 5.4, for tolerance ≈ 0.25 and values $f = 0, -8.5, -15.1$. These pictures, as well as Figure 3.1, were created using the graphics package GRAPE [10]. For tolerance ≈ 0.1, we can predict that a double-cusp forms for $f_{\text{crit}} \in (-9.3, -13.9)$. The interval for other tolerances is shown in Table 5.1.

TABLE 5.1

From balls to bones: A posteriori control of the interval containing f_{crit} for different tolerances in 2 dimensions (left) and 3 dimensions (right).

Tolerance	Interval for f_{crit}
$\tau \approx 0.5$	$(-3.3, -17.0)$
$\tau \approx 0.1$	$(-5.1, -9.5)$
$\tau \approx 0.05$	$(-5.5, -8.8)$
$\tau \approx 0.01$	$(-5.9, -8.1)$
$\tau \approx 0.005$	$(-6.0, -7.3)$
$\tau \approx 0.001$	$(-6.5, -6.9)$

Tolerance	Interval for f_{crit}
$\tau \approx 0.5$	$(-8.0, -21.0)$
$\tau \approx 0.25$	$(-8.5, -15.1)$
$\tau \approx 0.1$	$(-9.3, -13.9)$

REFERENCES

[1] F. BREZZI AND L. A. CAFFARELLI, *Convergence of the discrete free boundaries for finite element approximations*, RAIRO Anal. Numer., 17 (1983), pp. 385–395.

[2] L. A. CAFFARELLI, *A remark on the Hausdorff measure of a free boundary, and the convergence of coincidence set*, Boll. Un. Mat. Ital., A (5), 18 (1981), pp. 109–113.

[3] Z. CHEN AND R. H. NOCHETTO, *Residual type a posteriori error estimates for elliptic obstacle problems*, Numer. Math., 84 (2000), pp. 527–548.

[4] E. DARI, R. G. DURÁN, AND C. PADRA, *Maximum norm error estimators for three-dimensional elliptic problems*, SIAM J. Numer. Anal., 37 (2000), pp. 683–700.

[5] K. DECKELNICK AND K. G. SIEBERT, *$W^{1,\infty}$-convergence of the discrete free boundary for obstacle problems*, IMA J. Numer. Anal., 20 (2000), pp. 481–498.

[6] C. M. ELLIOTT, *On the finite element approximation of an elliptic variational inequality arising from an implicit time discretization of the Stefan Problem*, IMA J. Numer. Anal., 1 (1981), pp. 115–125.

[7] F. FIERRO AND A. VEESER, *A posteriori error estimators for regularized total variation of characteristic functions*, SIAM J. Numer. Anal., 41 (2003), pp. 2032–2055.

[8] J. FREHSE AND U. MOSCO, *Irregular obstacles and quasivariational inequalities of stochastic impulse control*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 9 (1982), pp. 105–157.

[9] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, Pure Appl. Math., John Wiley, New York, 1982.

[10] *GRAPE—GRAphics Programming Environment Manual, Version* 5, SFB 256, University of Bonn, 1995.

[11] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and their Applications*, Pure Appl. Math. 88, Academic Press, New York, 1980.

[12] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Local problems on stars: A posteriori error estimation, convergence, and performance*, Math. Comp. 72 (2003), pp. 1067–1097.

[13] R. H. NOCHETTO, *A note on the approximation of free boundaries by finite elements*, RAIRO Model. Math. Anal. Numer., 20 (1986), pp. 355–368.

[14] R. H. NOCHETTO, *Pointwise a posteriori error estimates for elliptic problems on highly graded meshes*, Math. Comp., 64 (1995), pp. 1–22.

[15] R. H. NOCHETTO, K. G. SIEBERT, AND A. VEESER, *Pointwise a posteriori error estimates for elliptic variational inequalities*, Numer. Math., 95 (2003), pp. 163–195.

[16] R. H. NOCHETTO AND L. B. WAHLBIN, *Positivity preserving finite element approximation*, Math. Comp. 71 (2002), 1405–1419.

[17] J.-F. RODRIGUES, *Obstacle Problems in Mathematical Physics*, North-Holland Math. Stud. 134, North-Holland, Amsterdam, 1987.

[18] A. SCHMIDT AND K. G. SIEBERT, *ALBERT—Software for scientific computations and applications*, Acta Math. Univ. Comenian. 70 (2001), pp. 105–122.

[19] A. SCHMIDT AND K. G. SIEBERT, *ALBERT: An adaptive hierarchical finite element toolbox*, Documentation, Preprint 06/2000 Universität Freiburg, p. 244.

[20] A. VEESER, *Efficient and reliable a posteriori error estimators for elliptic obstacle problems*, SIAM J. Numer. Anal., 39 (2001), pp. 146–167.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2136–2147

ESTIMATION OF HIGHER SOBOLEV NORM FROM LOWER ORDER APPROXIMATION*

CARSTEN CARSTENSEN†

Abstract. Given an accurate piecewise constant flux approximation q of a smooth (but unknown) exact flux p, how can one compute a Sobolev norm of p over a small domain ω? This question arises in a duality argument of goal-oriented finite element a posteriori error analysis. Two equivalent practical estimators η_M and $\eta_{\mathcal{E}}$ are presented for an approximation $|||q|||$ of $|p|_{H^1(\omega)}$ and model situations are addressed where $|||q|||$ is asymptotically an upper and +lower bound of $|p|_{H^1(\Omega)}$.

Key words. a posteriori error estimates, approximation of higher order Sobolev norms, reliability, efficiency, finite element method

AMS subject classifications. Primary, 65N30; Secondary, 65N15

DOI. 10.1137/S0036142902413615

1. Introduction.

Given a known approximation $q \in L^2(\Omega)$ to some unknown $p \in H^2(\Omega)$ and a parameter $s > 0$, the expression

$$|||q||| := \sup_{r \in H^1_0(\Omega)^n \setminus \{0\}} \left(\int_\Omega q \ \mathrm{div}\, r \, dx \right) / (\|r\|_{L^2(\Omega)} + s|r|_{H^1(\Omega)}) \tag{1.1}$$

is discussed as a computable approximation to $|p|_{H^1(\Omega)}$. (L^2, H^m, H^m_0 follow standard notation for Lebesgue and Sobolev spaces with their respective norms $\|\cdot\|_{L^2}, \|\cdot\|_{H^m}$ and seminorm $|\cdot|_{H^m}$.) Indeed, for $p \in H^2_0(\Omega)$, i.e., if p and ∇p vanish on the boundary $\partial\Omega$ of the bounded Lipschitz domain $\Omega \Subset \mathbb{R}^n$, there holds

$$\begin{aligned} &(\max\{0, |||q||| - \sqrt{n}s^{-1}\|p-q\|_{L^2(\Omega)}\})^2 \le |p|^2_{H^1(\Omega)} \\ &\le |||q|||^2 + 2|p|_{H^2(\Omega)}(2\sqrt{n}\|p-q\|_{L^2(\Omega)} + s|p|_{H^1(\Omega)}). \end{aligned} \tag{1.2}$$

The point is that q exclusively belongs to $L^2(\Omega)$ and is assumed to be close to the smoother $p \in H^2_0(\Omega)$. Provided

$$\lim_{(s, \|p-q\|_{L^2(\Omega)}) \to 0} \|p-q\|_{L^2(\Omega)}/s = 0,$$

the established estimate guarantees (for fixed p)

$$\lim_{(s, \|p-q\|_{L^2(\Omega)}) \to 0} |||q||| = |p|_{H^1(\Omega)}.$$

In other words, under some conditions to be discussed below, $|||q|||$ appears to be a highly accurate approximation to $|p|_{H^1(\Omega)}$. (Notice carefully that the limit for $s \to 0$ and $q \to p$ is subject to the above side condition $\|p-q\|_{L^2(\Omega)} = o(s)$ and so neither $s = 0$ nor $s = 1$ is representative of the results of this paper.)

*Received by the editors August 29, 2002; accepted for publication (in revised form) December 22, 2003; published electronically February 25, 2005.
http://www.siam.org/journals/sinum/42-5/41361.html
†Institute of Mathematics, Hunboldt University of Berlin, Unter den Linden 6, D-10099, Berlin, Germany (cc@mathematik.hu-berlin.de).

The simple proof of (1.2) utilizes the direct consequence

$$|||p-q||| \le \sqrt{n}s^{-1}\|p-q\|_{L^2(\Omega)} \quad \text{and} \quad |||p||| \le |p|_{H^1(\Omega)} \tag{1.3}$$

of the definition (1.1) to show the lower bound in (1.2) via

$$|||q||| \le |||p||| + |||p-q||| \le |p|_{H^1(\Omega)} + \sqrt{n}s^{-1}\|p-q\|_{L^2(\Omega)}. \tag{1.4}$$

An integration by parts and the Cauchy and Young inequality lead to

$$\begin{aligned}
|p|^2_{H^1(\Omega)} &= \int_\Omega \nabla p \cdot \nabla p \, dx = -\int_\Omega p\,\Delta p \, dx \\
&= \int_\Omega (q-p)\Delta p \, dx - \int_\Omega q \,\mathrm{div}\,\nabla p \, dx \\
&\le \sqrt{n}\|q-p\|_{L^2(\Omega)}|p|_{H^2(\Omega)} + |||q|||(|p|_{H^1(\Omega)} + s|p|_{H^2(\Omega)}) \\
&\le |p|_{H^2(\Omega)}(\sqrt{n}\|p-q\|_{L^2(\Omega)} + s|||q|||) + \frac{1}{2}|||q|||^2 + \frac{1}{2}|p|^2_{H^1(\Omega)}
\end{aligned}$$

and so to the upper bound in (1.2) after $s|||q|||$ is estimated with (1.4).

The indicated proof of (1.2) uses integration by parts at some stage and so $p = 0$ and $\nabla p = 0$ on $\partial\Omega$ to avoid any boundary contribution. This paper generalizes (1.2) for the case $p \in H^2(\omega)$ without any restriction of p on the boundary $\partial\omega$ of a Lipschitz domain ω. One additional argument for that is based on a deeper remark due to Tartar on some interpolation spaces [Ta].

The task to compute an approximation to $|p|_{H^1(\omega)}$ arises naturally in the a posteriori error analysis as reported in section 2 below and is (partly) justified in this paper. There, $p = \partial z/\partial x_j$ is one component of the gradient of an unknown dual solution z. A computed approximation z_h to z provides a (known) piecewise constant $q = \partial z_h/\partial x_j \in L^2(\omega)\backslash H^1(\omega)$. Up to a weight $C\,\mathrm{diam}(\omega)$, the interpolation error by nodal interpolation $z - Iz$ is then bounded by (the sum over all $j = 1, \ldots, n$ of) $|p|_{H^1(\omega)} = |\partial z/\partial x_j|_{H^1(\omega)}$ approximated by $|||q|||$. Two computable estimates $\eta_M \approx \eta_{\mathcal{E}}$ of $|||q|||$ will be studied in section 4 below. The localized version is discussed in section 5 for a model scenario where strict efficiency and reliability of $|||q|||$ imply that a mesh much finer than $h_\omega = \mathrm{diam}(\omega)$ is required to compute q as an accurate approximation to p on ω.

2. Motivation and main application. Given a Lipschitz domain Ω and a Sobolev space

$$H := H^1_0(\Omega)^m := \{v \in H^1(\Omega)^m : v = 0 \text{ on } \partial\Omega\}$$

we consider the exact solution $u \in H$ of

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

and its Galerkin approximation $u_h \in H_h \subset H$ with

$$a(u_h, v_h) = b(v_h) \quad \text{for all } v_h \in H_h. \tag{2.2}$$

Here, $a : H \times H \to \mathbb{R}$ is a bounded bilinear form and $b : H \to \mathbb{R}$ is a bounded linear form of the weak form of an elliptic PDE. The error $e := u - u_h$ is monitored with

respect to a particular norm or goal functional. The latter is a linear and bounded functional $J : H \to \mathbb{R}$. To bound or approximate $J(e)$ one considers the dual problem

$$a(v, z) = J(v) \quad \text{for all } v \in H \tag{2.3}$$

with an exact solution $z \in H$. Based on the Galerkin orthogonality

$$a(e, v_h) = 0 \quad \text{for all } v_h \in H_h \tag{2.4}$$

(which is an immediate consequence of (2.1)–(2.2)), one infers

$$J(e) = a(e, z) = a(e, z - z_h) \tag{2.5}$$

for any $z_h \in H_h$. Typically, the residual $a(e, \cdot)$ assumes a representation

$$a(e, v) = \int_\Omega R_\Omega \cdot v \, dx + \int_{\cup\mathcal{E}} R_\mathcal{E} \cdot v \, ds = \mathrm{Res}(v) \tag{2.6}$$

with volume residuals $R_\Omega \in L^2(\Omega)^m$ and edge residuals $R_\mathcal{E} \in L^2(\cup\mathcal{E})^m$ on the skeleton $\cup\mathcal{E}$ of all inner-element edges with respect to a regular triangulation $\mathcal{T}$ of Ω [AO, BaS]. If a is coercive,

$$\|z\|_H \le C_{stab} \|J\|_{H^*} \tag{2.7}$$

for some stability constant $C_{stab} > 0$. Then, standard a posteriori estimates of the given linear functional Res (with $\mathrm{Res}(z_h) = 0$) prove

$$\begin{aligned} a(e, z - z_h) = \mathrm{Res}(z) &\le \| \mathrm{Res} \|_{H^*} \, \|z\|_H \\ &\le \|J\|_{H^*} \, C_{stab} \, \| \mathrm{Res} \|_{H^*} \end{aligned} \tag{2.8}$$

and so (combining (2.5) and (2.8))

$$|J(e)| \le \|J\|_{H^*} \, C_{stab} \| \mathrm{Res} \|_{H^*}. \tag{2.9}$$

The main disadvantage of this a posteriori estimate is that the local influence of J and z is not reflected in the global constant $\|J\|_{H^*} \, C_{stab}$, i.e., the error estimation is essentially that of the energy error $\|e\|_H \approx \| \mathrm{Res} \|_{H^*}$. Any local information needs a computation of z or of an approximation z_h. We refer to [AO, Chapter 8] and [BaS, section 6.2] for guaranteed lower and upper bounds of $J(e)$ in terms of a parallelogram identity and global errors $\|\alpha e \pm \alpha^{-1}(z - z_h)\|^2$ for some parameter $\alpha > 0$. Becker and Rannacher [BR] advertised a different and local approach where, given $z \in H^2(\Omega)$, one selects $z_h := Iz$ as the (nodal) interpolant of z and obtains, with some constant $C_I > 0$,

$$h_T^2 \|z - Iz\|_{L^2(T)} + h_T^{3/2} \|z - Iz\|_{L^2(\partial T)} \le C_I |z|_{H^2(T)} \tag{2.10}$$

for any element $T \in \mathcal{T}$. Then (2.5)–(2.6) and (2.10) yield

$$\begin{aligned} |J(e)| &\le \sum_{T \in \mathcal{T}} \left(\|R_\Omega\|_{L^2(T)} \, \|z - Iz\|_{L^2(T)} + \|R_\mathcal{E}\|_{L^2(\partial T)} \, \|z - Iz\|_{L^2(\partial T)} \right) \\ &\le \sum_{T \in \mathcal{T}} C_I \left(h_T^2 \|R_\Omega\|_{L^2(T)} + h_T^{3/2} \|R_\mathcal{E}\|_{L^2(\partial T)} \right) |z|_{H^2(T)}. \end{aligned} \tag{2.11}$$

The point is that the upper bound involves the unknown H^2 seminorm $|z|_{H^2(T)}$ which is to be replaced by some discrete analogue based on a computed approximation z_h.

The main result of this paper, Theorem 3.2, allows for $z \in H^3(\omega)$ an estimate

$$|z|_{H^2(\omega)} \le c_1\eta + c_2\|\nabla z - q\|^{1/2}_{L^2(\omega)}, \tag{2.12}$$

where the constant c_1 only depends on the shape of elements (aspect ratio, maximal angle, etc.) and c_2 on ω and $\|z\|_{H^3(\omega)}$. The quantity $q = \nabla z_h$ is a computed approximation of $p = \nabla z$ on which the calculation of η is based. We suppose that $\|\nabla z - q\|_{L^2(\omega)} \to 0$ as the maximal mesh-size h tends to zero. Notice that, here, ω is fixed and the mesh that covers ω becomes finer and finer. Then, the left-hand side of (2.12) stays bounded as well as the constants c_1 and c_2 while $\|\nabla z - q\|^{1/2}_{L^2(\omega)} \to 0$ (as a consequence of energy norm convergence of the Galerkin approximation z_h to the exact solution z as $h \to 0$).

As an application of (2.12) to the goal-oriented a posteriori error estimate (2.11), suppose that (2.12) holds for $\eta_\omega := \eta$ and any domain ω from a partition $\mathcal{P}$ of Ω subordinated to $\mathcal{T}$ (i.e., each domain $\omega \in \mathcal{P}$ is the interior of a union $\cup\mathcal{T}_\omega$ of elements in $\mathcal{T}_\omega \subset \mathcal{T}$ such that $\overline{\cup\mathcal{P}} = \overline{\Omega}$ and two distinct ω_1 and ω_2 in $\mathcal{P}$ are disjoint). Then, (2.11)–(2.12) yield

$$\begin{aligned} |J(e)| &\le \sum_{\omega\in\mathcal{P}} C_I\,(c_1\eta_\omega + \text{h.o.t.}) \\ &\quad\times \left(\sum_{T\in\mathcal{T}_\omega} \left(h_T^2\,\|R_\Omega\|_{L^2(T)} + h_T^{3/2}\,\|R_\mathcal{E}\|_{L^2(\partial T)}\right)^2\right)^{1/2} \end{aligned} \tag{2.13}$$

with higher order terms (h.o.t.) which tend to zero as $\|\nabla z - q\|_{L^2(\omega)}$ does with $h \to 0$. Notice that, up to h.o.t., the right-hand side of (2.13) is fully computable.

Two examples for the computable quantity $\eta = \eta_\omega$ are suggested in section 4, namely (with the $\mathcal{T}$-piecewise affine and globally continuous functions $\mathcal{S}^1(\mathcal{T})$, with the jumps $[q]$ across an inner element edge, and with the skeleton $\cup\mathcal{E}_\omega$ of such edges)

$$\eta = \min_{Q\in\mathcal{S}^1(\mathcal{T})} \|q - Q\|_{L^2(\omega)} \quad \text{or} \quad \eta = \|h_\mathcal{E}^{1/2}[q]\|_{L^2(\cup\mathcal{E}_\omega)}$$

for a piecewise constant $q = \nabla z_h$. Some remarks on the efficiency and on the situation with domains ω that shrink with $h \to 0$ follow in section 5.

With the above result (2.13), the approach from [BR] is justified for a fixed partition $\mathcal{P}$ and finer and finer triangulations $\mathcal{T}$.

3. Main result. The announced local version of (1.2) aims the estimation of $|p|_{H^1(\omega)}$ for a subdomain $\omega \subset \Omega$. Then, $p \not\equiv 0$ on $\partial\omega$ causes technical difficulties. To overcome them we work with Besov spaces that do not see boundary effects and employ the following known result.

THEOREM 3.1 (Tartar, Brezzi–Marini–Süli). *Given a Lipschitz domain $\omega \Subset \mathbb{R}^d$ of diameter h_ω there exists a constant $c(\omega) > 0$ such that*

$$\begin{aligned} &\forall t > 0\,\forall v \in H^1(\omega)\,\exists a \in L^2(\omega)\,\exists b \in H^1_0(\omega), v = a + b \text{ and} \\ &t^{-1/2}\|a\|_{L^2(\omega)} + t^{1/2}|b|_{H^1(\omega)} \\ &\quad\le c(\omega)\|v\|^{1/2}_{L^2(\omega)}\left(|v|^2_{H^1(\omega)} + \|v\|^2_{L^2(\omega)}/h_\omega^2\right)^{1/4}. \end{aligned}$$

The constant $c(\omega)$ depends only on the shape of ω and not on h_ω.

Sketch of the proof. The assertion follows from the analysis in [BMS, section 3] and hence we give solely an introduction to the arguments for a convenient reading. Two interpolation spaces,

$$X(\infty,\omega) = [L^2(\omega), H^1_0(\omega)]_{1/2,\infty} = \mathring{B}^{1/2}_{2,\infty}(\omega),$$
$$Y(1,\omega) = [L^2(\omega), H^1(\omega)]_{1/2,1} = B^{1/2}_{2,1}(\omega)$$

(identified with certain Besov spaces), are compared in a result due to Tartar [Ta], namely, for all $v \in H^1(\omega)$,

$$\|v\|_{X(\infty,\omega)} \le c_1(\omega)\|v\|_{Y(1,\omega)}.$$

The two norms are defined by real interpolation of Sobolev spaces with slightly different norms

$$\|v\|_{X(\infty,\omega)} = \operatorname{ess}\sup_{0<t<\infty} \inf\{t^{-1/2}\|a\|_{L^2(\omega)} + t^{1/2}|b|_{H^1(\omega)} : v = a+b, a \in L^2(\omega), b \in H^1_0(\omega)\},$$
$$\|v\|_{Y(1,\omega)} = \int_0^\infty \inf\{t^{-3/2}\|a\|_{L^2(\omega)} + t^{-1/2}(|b|^2_{H^1(\omega)} + \|b\|^2_{L^2(\omega)}/h^2_\omega)^{1/2} : v = a+b, a \in L^2(\omega), b \in H^1(\omega)\}\, dt.$$

The point is that the function b in $\|v\|_{X(\infty,\omega)}$ vanishes on $\partial\omega$ (and so only the H^1 seminorm arises) while there is no such restriction on b in $\|v\|_{Y(1,\omega)}$ (and hence the scaled L^2 norm arises). Given $v \in H^1(\omega)$ and $t > 0$, there exist $a \in L^2(\omega)$ and $b \in H^1_0(\omega)$ such that $v = a + b$ and

$$t^{-1/2}\|a\|_{L^2(\omega)} + t^{1/2}|b|_{H^1(\omega)} \le 2\|v\|_{X(\infty,\omega)}.$$

(This follows immediately from the definition of $\|v\|_{X(\infty,\omega)}$.) The norm $\|v\|_{Y(1,\omega)}$ is bounded from above by the interpolation estimate

$$\|v\|_{Y(1,\omega)} \le c_2(\omega)\|v\|^{1/2}_{L^2(\omega)}(|v|^2_{H^1(\omega)} + \|v\|^2_{L^2(\omega)}/h^2_\omega)^{1/4}$$

in Besov spaces [Tr]. A scaling argument justifies that $c_2(\omega)$ is h_ω independent [BMS]. The combination of the three estimates proves the theorem with $c(\omega) = 2c_1(\omega)c_2(\omega)$. □

The main result of this paper concerns a smooth function p and its approximation q with a seminorm $|||q|||$ such that

$$|p|_{H^1(\omega)} \le |||q||| + \text{h.o.t.},$$

where h.o.t. $\to 0$ as $\|p-q\|_{L^2(\omega)} + s|||q||| \to 0$ for fixed $p \in H^2(\omega)$. (Arguing as in (1.4), the latter follows merely from $s \to 0$ and $\|p-q\|_{L^2(\omega)} \to 0$ for fixed $p \in H^1(\omega)$.)

THEOREM 3.2. *Suppose $p \in H^2(\omega)$ and $q \in L^2(\omega)$ for a bounded Lipschitz domain ω in $\mathbb{R}^n$. Let $s > 0$ and let $|||q|||$ be given by*

$$|||q||| := \sup_{r \in H^1_0(\omega)^d \setminus \{0\}} \int_\omega q \operatorname{div} r\, dx / (\|r\|_{L^2(\omega)} + s|r|_{H^1(\omega)}). \tag{3.1}$$

Let $c(\omega)$ denote the h_ω-independent constant from Theorem 3.1. *Then there holds*

$$|p|_{H^1(\omega)} \le |||q||| + c(\omega)\sqrt{8}\left(|p|^2_{H^1(\omega)}/h^2_\omega + |p|^2_{H^2(\omega)}\right)^{1/4} \times \left(\sqrt{n}\|p-q\|_{L^2(\omega)} + s|||q|||\right)^{1/2}. \tag{3.2}$$

Proof. Set $t := (\sqrt{n}\|p-q\|_{L^2(\omega)} + s|||q|||)/|p|_{H^1(\omega)} > 0$ and choose $a_1, \dots, a_n \in L^2(\omega)$ and $b_1, \dots, b_n \in H^1_0(\omega)$ for each $j = 1, \dots, n$ from Theorem 3.1 with $a_j + b_j = \partial p/\partial x_j$ and

$$\begin{aligned} &t^{-1/2}\|a_j\|_{L^2(\omega)} + t^{1/2}|b_j|_{H^1(\omega)} \\ &\le c(\omega)\|\partial p/\partial x_j\|^{1/2}_{L^2(\omega)}\left(|\partial p/\partial x_j|^2_{H^1(\omega)} + \|\partial p/\partial x_j\|^2_{L^2(\omega)}/h^2_\omega\right)^{1/4}. \end{aligned}$$

Write $a = (a_1, \dots, a_n)$ and b as vectors. Cauchy inequalities verify

$$\begin{aligned} &1/2\left(t^{-1/2}\|a\|_{L^2(\omega)} + t^{1/2}|b|_{H^1(\omega)}\right)^2 \le t^{-1}\|a\|^2_{L^2(\omega)} + t|b|^2_{H^1(\omega)} \\ &\le c(\omega)^2 \sum_{j=1}^d \|\partial p/\partial x_j\|_{L^2(\omega)}\left(|\partial p/\partial x_j|^2_{H^1(\omega)} + \|\partial p/\partial x_j\|^2_{L^2(\omega)}/h^2_\omega\right)^{1/2} \\ &\le c(\omega)^2|p|_{H^1(\omega)}(|p|^2_{H^2(\omega)} + |p|^2_{H^1(\omega)}/h^2_\omega)^{1/2}. \end{aligned}$$

This and the definition of t yield

$$\begin{aligned} \|a\|_{L^2(\omega)} + t|b|_{H^1(\omega)} &\le \sqrt{2t}\, c(\omega)|p|^{1/2}_{H^1(\omega)}(|p|^2_{H^2(\omega)} + |p|^2_{H^1(\omega)}/h^2_\omega)^{1/4} \\ &\le \sqrt{2}\, c(\omega)\left(\sqrt{n}\|p-q\|_{L^2(\omega)} + s|||q|||\right)^{1/2} \\ &\quad \times \left(|p|^2_{H^2(\omega)} + |p|^2_{H^1(\omega)}/h^2_\omega\right)^{1/4}. \end{aligned} \tag{3.3}$$

On the other hand, $Dp = a + b$, a Cauchy inequality, and an integration by parts (notice $b = 0$ on $\partial\omega$) lead to

$$\begin{aligned} |p|^2_{H^1(\omega)} &= \int_\omega a \cdot Dp\, dx + \int_\omega b \cdot Dp\, dx \\ &\le \|a\|_{L^2(\omega)}\, |p|_{H^1(\omega)} - \int_\omega p \operatorname{div} b\, dx. \end{aligned} \tag{3.4}$$

The definition of $|||q|||$ in (3.1) shows

$$\begin{aligned} -\int_\omega p \operatorname{div} b\, dx &= -\int_\omega q \operatorname{div} b\, dx + \int_\omega (q-p)\operatorname{div} b\, dx \\ &\le |||q|||(\|b\|_{L^2(\omega)} + s|b|_{H^1(\omega)}) + \|p-q\|_{L^2(\omega)}\sqrt{n}\, |b|_{H^1(\omega)}. \end{aligned}$$

This, the definition of t, and $\|b\|_{L^2(\omega)} \le \|a\|_{L^2(\omega)} + |p|_{H^1(\omega)}$ result in

$$-\int_\omega p \operatorname{div} b\, dx \le |||q|||(\|a\|_{L^2(\omega)} + |p|_{H^1(\omega)}) + t|b|_{H^1(\omega)}|p|_{H^1(\omega)}. \tag{3.5}$$

The combination of (3.4)–(3.5) reads

$$|p|^2_{H^1(\omega)} \le |p|_{H^1(\omega)}(\|a\|_{L^2(\omega)} + t|b|_{H^1(\omega)} + |||q|||) + \|a\|_{L^2(\omega)}|||q|||,$$

and so with $\alpha := |p|_{H^1(\omega)}$, $\beta := |||q|||$, $\gamma := \|a\|_{L^2(\omega)} + t|b|_{H^1(\omega)}$,

$$\alpha^2 \leq \alpha(\beta+\gamma) + \beta\gamma = \alpha(\beta+2\gamma) + \beta\gamma - \alpha\gamma.$$

This is recast into

$$\alpha(\alpha+\gamma) \leq \alpha(\beta+2\gamma) + \beta\gamma \leq (\alpha+\gamma)(\beta+2\gamma)$$

and hence shows $\alpha \leq \beta + 2\gamma$, i.e.,

$$|p|_{H^1(\omega)} \leq |||q||| + 2(\|a\|_{L^2(\omega)} + t|b|_{H^1(\omega)}).$$

This and (3.3) conclude the proof. □

4. Computable bounds for $|||q|||$. This section concerns the estimation of $|||q|||$ by $\eta_{\mathcal{E}}$ and η_M for piecewise constant q. Throughout this section and subsequently, $A \lesssim B$ abbreviates an estimate $A \leq CB$ with a generic constant C (referred to as the multiplicative constant hidden in $A \lesssim B$) that depends on the dimension n and on the shape (not the size h_ω) of the domain ω through $c(\omega)$ of Theorem 3.1. The gradient of a scalar is written ∇ and the functional matrix of a vector valued function is written D.

DEFINITION 4.1. *Let ω be a bounded Lipschitz domain covered by the regular triangulation $\mathcal{T}_\omega$. Let $\mathcal{E}_\omega$ denote the set of edges inside $\cup\mathcal{T}_\omega \supset \overline{\omega}$, i.e., for any $E \in \mathcal{E}_\omega$ there exist $T_1, T_2 \in \mathcal{T}_\omega$ with $E = T_1 \cap T_2$. Given a triangulation $\mathcal{T}$ (resp., a set $\mathcal{E}$ of edges), let $\mathcal{L}^k(\mathcal{T})$ (resp., $\mathcal{L}^k(\mathcal{E})$) denote the $\mathcal{T}$-piecewise (resp., $\mathcal{E}$-piecewise) polynomials of degree $\leq k$ and set $\mathcal{S}^1(\mathcal{T}) := \mathcal{L}^1(\mathcal{T}) \cap C(\cup\mathcal{T})$.*

Given $q \in \mathcal{L}^0(\mathcal{T}_\omega)$, let $[q] \in \mathcal{L}^0(\mathcal{E}_\omega)$ denote its $\mathcal{E}_\omega$-piecewise constant jumps across the skeleton $\cup\mathcal{E}_\omega = \{x \in \overline{\Omega} : x \in E \in \mathcal{E}_\omega\}$, i.e., $[q]|_E := q|_{T_2} - q|_{T_1}$, where the normal ν_E along E points into T_2. Let $h_{\mathcal{E}} \in \mathcal{L}^0(\mathcal{E}_\omega)$ denote the $\mathcal{E}_\omega$-piecewise constant edge size, $h_{\mathcal{E}}|_E := h_E := \operatorname{diam}(E)$ for any $E \in \mathcal{E}_\omega$; recall that $h_{\mathcal{T}} \in \mathcal{L}^0(\mathcal{T}_\omega)$ denotes the $\mathcal{T}_\omega$-piecewise constant element size, $h_{\mathcal{T}}|_T := h_T := \operatorname{diam}(T)$ for any $T \in \mathcal{T}_\omega$. Then, let $\hat{\omega} := \operatorname{int}(\cup\mathcal{T}_\omega)$ and set

$$\eta_{\mathcal{E}} := \|h_{\mathcal{E}}^{-1/2}[q]\|_{L^2(\cup\mathcal{E}_\omega)} := \left(\sum_{E\in\mathcal{E}_\omega} h_E^{-1}\|[q]\|^2_{L^2(E)}\right)^{1/2}, \tag{4.1}$$

$$\eta_M := \min_{Q\in\mathcal{S}^1(\mathcal{T}_\omega)} \|(q-Q)/h_{\mathcal{T}}\|_{L^2(\hat{\omega})}. \tag{4.2}$$

The following theorem essentially asserts equivalence of $|||q||| \approx \eta_M \approx \eta_{\mathcal{E}}$ whenever the mesh-size $h_{\mathcal{T}} \approx s$ on ω is equivalent to the parameter $s > 0$.

THEOREM 4.2. *Given any $q \in \mathcal{L}^0(\mathcal{T}_\omega)$, define η_M and $\eta_{\mathcal{E}}$ by (4.1) and (4.2), respectively. Then there holds*

$$\sup_{r\in H^1_0(\omega)^n\setminus\{0\}} \int_\omega q \operatorname{div} r\, dx/(\|r\|_{L^2(\omega)} + \|h_{\mathcal{T}}Dr\|_{L^2(\omega)}) \lesssim \eta_M$$
$$\lesssim \eta_{\mathcal{E}} \lesssim \sup_{\hat{r}\in H^1_0(\hat{\omega})^n\setminus\{0\}} \int_{\hat{\omega}} q \operatorname{div} \hat{r}\, dx/(\|\hat{r}\|_{L^2(\hat{\omega})} + \|h_{\mathcal{T}}D\hat{r}\|_{L^2(\hat{\omega})}).$$

Proof. Given any $Q \in \mathcal{S}^1(\mathcal{T}_\omega)$ and $r \in H^1_0(\omega)^n$ (extended by zero outside ω) an

integration by parts and a Cauchy inequality show

$$\int_\omega q \operatorname{div} r\,dx = \int_\omega Q \operatorname{div} r\,dx + \int_\omega (q-Q)\operatorname{div} r\,dx \leq -\int_\omega r\cdot DQ\,dx + \sqrt{n}\|h_{\mathcal{T}} Dr\|_{L^2(\omega)}\|(q-Q)/h_{\mathcal{T}}\|_{L^2(\omega)}. \tag{4.3}$$

Let $D_{\mathcal{T}}$ denote the $\mathcal{T}_\omega$-piecewise action of the gradient operator D, e.g., $D_{\mathcal{T}} q = 0$ almost everywhere in ω. Then, a Cauchy inequality and an elementwise inverse estimate, i.e., $\|D(Q-q)\|_{L^2(T)} \lesssim \|q-Q\|_{L^2(T)}/h_T$, lead to

$$\begin{aligned} -\int_\omega r\cdot DQ\,dx &\leq \|r\|_{L^2(\omega)}\|D_{\mathcal{T}}(q-Q)\|_{L^2(\omega)} \\ &\lesssim \|r\|_{L^2(\omega)}\|(q-Q)/h_{\mathcal{T}}\|_{L^2(\cup\mathcal{T}_\omega)}. \end{aligned} \tag{4.4}$$

(Notice that $\cup\mathcal{T}_\omega$ may be larger than ω and, for simplicity, we employ the inverse estimate on complete elements.) The combination of (4.3)–(4.4) shows

$$\sup_{r\in H^1_0(\omega)^n\setminus\{0\}} \int_\omega q\operatorname{div} r\,dx/(\|r\|_{L^2(\omega)} + \|h_{\mathcal{T}} Dr\|_{L^2(\omega)}) \lesssim \eta_M.$$

The proof of $\eta_M \lesssim \eta_{\mathcal{E}}$ follows ideas from [C2, CB]. Let $(\varphi_z : z\in\mathcal{N})$ be a nodal basis of $\mathcal{S}^1(\mathcal{T}_\omega)$. Then $\sum_{z\in\mathcal{N}}\varphi_z = 1$ almost everywhere on $\hat\omega := \cup\mathcal{T}_\omega$ and so, for some $Q = \sum_{z\in\mathcal{N}} q_z\varphi_z$,

$$\begin{aligned} \eta_M^2 = \|(q-Q)/h_{\mathcal{T}}\|^2_{L^2(\hat\omega)} &= \Big\|\sum_{z\in\mathcal{N}}(q-q_z)\varphi_z/h_{\mathcal{T}}\Big\|^2_{L^2(\hat\omega)} \\ &= \sum_{T\in\mathcal{T}}\Big\|\sum_{z\in\mathcal{N}(T)}(q-q_z)\varphi_z\Big\|^2_{L^2(T)}/h_T^2. \end{aligned}$$

Since $\mathcal{N}(T) := \{z\in\mathcal{N} : z\in T\}$ has a bounded cardinality,

$$\Big\|\sum_{z\in\mathcal{N}(T)}(q-q_z)\varphi_z\Big\|^2_{L^2(T)} \lesssim \sum_{z\in\mathcal{T}}\|(q-q_z)\varphi_z\|^2_{L^2(T)}.$$

With $h_T \approx h_z$ for $z\in\mathcal{N}(T)$, $T\in\mathcal{T}$, we have shown

$$\begin{aligned} \eta_M^2 &\lesssim \min_{(q_z:\, z\in\mathcal{N})}\sum_{T\in\mathcal{T}}\sum_{z\in\mathcal{N}(T)}\|(q-q_z)\varphi_z\|^2_{L^2(T)}/h_z^2 \\ &= \min_{(q_z:\, z\in\mathcal{N})}\sum_{z\in\mathcal{N}}\|(q-q_z)\varphi_z\|^2_{L^2(\hat\omega)}/h_z^2 \\ &= \sum_{z\in\mathcal{N}}\min_{q_z\in\mathbb{R}}\|(q-q_z)\varphi_z\|^2_{L^2(\omega_z)}/h_z^2. \end{aligned} \tag{4.5}$$

For each node $z\in\mathcal{N}$ with patch $\omega_z := \operatorname{int}(\cup\{T\in\mathcal{T}_\omega : z\in T\})$, $q|_{\omega_z}$ belongs to a finite dimensional space $\mathcal{L}^0(\mathcal{T}_\omega|_{\omega_z}) := \{q|_{\omega_z} : q\in\mathcal{T}_\omega)\}$. Notice that

$$\begin{aligned} |\!|\!|q|\!|\!|_{z,1} &:= \min_{q_z\in\mathbb{R}}\|(q-q_z)\varphi_z\|_{L^2(\omega_z)}/h_z, \\ |\!|\!|q|\!|\!|_{z,2} &:= \|h_{\mathcal{E}}^{-1/2}[q]\|_{L^2(\cup\mathcal{E}_z)} \quad \text{with } \mathcal{E}_z := \{E\in\mathcal{E}_\omega : z\in E\}, \end{aligned}$$

define two seminorms on $\mathcal{L}^0(\mathcal{T}_\omega|_{\omega_z})$ with zero-set $\mathcal{P}_0(\omega_z)$ (i.e., $|||q|||_{z,j} = 0$ implies q is constant on ω_z). Hence $|||\cdot|||_{z,1}$ and $|||\cdot|||_{z,2}$ are norms on the quotient space $\mathcal{L}^0(\mathcal{T}_\omega|_{\omega_z})/\mathcal{P}_0(\omega_z)$ of a finite dimension $\leq \mathrm{card}(\mathcal{T}_\omega) \lesssim 1$. Since two norms on a finite dimensional space are equivalent we have

$$|||q|||_{z,1} \approx |||q|||_{z,2} \quad \text{for all } q \in \mathcal{L}^0(\mathcal{T}_\omega). \tag{4.6}$$

A scaling argument shows that the equivalence constants in (4.6) are h_z independent. Therefore, we deduce in (4.5) that

$$n_M^2 \lesssim \sum_{z\in\mathcal{N}} \|h_{\mathcal{E}}^{-1/2}[q]\|^2_{L^2(\cup\mathcal{E}_z)} \lesssim \eta_{\mathcal{E}}^2$$

(since the overlap of $(\cup\mathcal{E}_z : z \in \mathcal{N})$ is finite). It remains for us to prove the last asserted inequality with the inverse estimate technique from [V]. For $E = T_1 \cap T_2 \in \mathcal{E}_\omega$ let b_E denote the edge bubble defined through $b_E|_E \in \mathcal{P}_d(E)$, $b_E|_{T_j}$ is the product of all barycentric coordinates λ_ℓ on T_j with respect to a vertex of E. We regard $b_E \in H^1_0(\omega_E)$ as a function in $H^1(\mathbb{R}^d)$ which vanishes outside ω_E. Then,

$$0 \leq b_E \leq 1, \quad \|b_E\|_{L^2(E)} \approx |E|^{1/2},$$
$$\int_E b_E\,ds \approx |E|, \quad |b_E|_{H^1(\omega_E)} \approx |\omega_E|^{1/2}/h_E.$$

Set $J_E := [q]|_E \in \mathbb{R}$ and estimate

$$h_E^{-1}\|[q]\|^2_{L^2(E)} \approx J_E/h_E \int_E b_E[q]\,ds = J_E/h_E \int_E (b_E\nu_E)\cdot[q\,\nu_E]\,ds$$

for the unit normal vector ν_E along E. The elementwise application of the Gaussian divergence theorem to $q\,Db_E$ yields

$$h_E^{-1}\|[q]\|^2_{L^2(E)} \approx J_E/h_E \int_{\omega_E} (Db_E\cdot\nu_E)q\,dx.$$

The sum over all $E \in \mathcal{E}_\omega$ verifies

$$\eta_{\mathcal{E}}^2 \approx \int_{\hat\omega} fq\,dx \quad \text{for } f := \sum_{E\in\mathcal{E}_\omega} J_E(Db_E\cdot\nu_E)/h_E.$$

Since ν_E is constant, $Db_E\cdot\nu_E = \mathrm{div}(b_E\nu_E)$ and so

$$\eta_{\mathcal{E}}^2 \lesssim \int_{\hat\omega} q\,\mathrm{div}\,\hat r\,dx \quad \text{for } \hat r := \sum_{E\in\mathcal{E}_\omega} J_E(b_E/h_E)\,\nu_E \in H^1_0(\hat\omega)^n.$$

It therefore remains to verify

$$\|\hat r\|_{L^2(\hat\omega)} + \|h_{\mathcal{T}} D\hat r\|_{L^2(\hat\omega)} \lesssim \eta_{\mathcal{E}}. \tag{4.7}$$

For each $E \in \mathcal{E}_\omega$ we have

$$\begin{aligned}
&\|J_E b_E\,\nu_E/h_E\|^2_{L^2(\hat\omega)} + h_E^2\|J_E\,Db_E\otimes\nu_E/h_E\|^2_{L^2(\omega_E)}\\
&\approx |J_E|^2(\|b_E\|^2_{L^2(\omega_E)}/h_E^2 + |b_E|^2_{H^1(\omega_E)})\\
&\approx |J_E|^2|\omega_E|/h_E^2 \approx h_E\|[q]\|^2_{L^2(E)}.
\end{aligned}$$

Since $(\omega_E : E \in \mathcal{E}_\omega)$ has finite overlap, the sum of the last equivalences over all $E \in E_\omega$ verifies (4.7). This concludes the proof. ∎

Remark 4.1. With arguments of [C1, C2, V] it can be shown that η_M is equivalent to

$$\eta_A := \|(q - Aq)/h_{\mathcal{T}}\|_{L^2(\hat{\omega})}, \text{ where } Aq = \sum_{z \in \mathcal{N}} (Aq)(z)\, \varphi_z$$

with the local integral mean $(Aq)(z) := \int_{\omega_z} q\, dx/|\omega_z|$ of q over the patch ω_z of a node $z \in \mathcal{N}$.

5. Discussion. This section is devoted to a discussion about the implications and limitations of the usage of $|||q|||$ (or some computable approximation thereof) as an estimator for $|p|_{H^1(\omega)}$ for a fixed $p \in H^2(\Omega)$ and varying parameters h_ω (the diameter of $\omega \subset \Omega$) or mesh-size $h_{\min,\omega}$ and $h_{\max,\Omega}$ for the minimal and maximal mesh-size of the mesh $\mathcal{T}$ (used for the computation of q) on ω and Ω, respectively. Recall that $\|\!|q|\!\|$ defined in (3.1) involves a parameter $s > 0$ to be chosen properly.

5.1. The concepts of (strict) efficiency and reliability. The estimator $\|\!|q|\!\|$ is called *efficient* if

$$\text{(E)} \qquad \big(s^{-1}\|p - q\|_{L^2(\omega)} \lesssim |p|_{H^1(\omega)}\big)$$

(since then $\|\!|q|\!\| \lesssim |p|_{H^1(\omega)}$ by the arguments of (1.4)) and called *strict efficient* if

$$\text{(SE)} \qquad \big(s^{-1}\|p - q\|_{L^2(\omega)} \ll |p|_{H^1(\Omega)}\big)$$

(since then $\|\!|q|\!\| \le |p|_{H^1(\omega)} + o(1)$). The notation $A \ll B$ means that A/B is very small for small $h_{\max,\Omega}$ in the sense that $\lim_{h_{\max,\Omega} \to 0} A/\, B = 0$. The estimator $\|\!|q|\!\|$ is called *reliable* if

$$\text{(R)} \qquad \big(|p|_{H^1(\omega)}/h_\omega + |p|_{H^2(\omega)}\big)\, \big(\|p - q\|_{L^2(\omega)} + s\|\!|q|\!\|\big) \lesssim \|\!|q|\!\|^2$$

(since then $|p|_{H^1(\omega)} \lesssim \|\!|q|\!\|$ by Theorem 3.2) and called *strict reliable* if

$$\text{(SR)} \qquad \big(|p|_{H^1(\omega)}/h_\omega + |p|_{H^2(\omega)}\big)\, \big(\|p - q\|_{L^2(\omega)} + s\|\!|q|\!\|\big) \ll |p|^2_{H^1(\omega)}$$

(since then $|p|_{H^1(\omega)} \le \|\!|q|\!\| + o(1)|p|_{H^1(\omega)}$ and so $|p|_{H^1(\omega)} \le \|\!|q|\!\|(1 + o(1))$ for sufficiently small $h_{\max,\Omega}$).

Notice carefully that (each of) the concepts (E), (SE), (R), (SR) based on (1.4) and (3.2) are sufficient (but possibly not necessary) conditions for a justification of $\|\!|q|\!\| \approx |p|_{H^1(\omega)}$.

5.2. Implications of (E) and (R). This subsection is devoted to discussing necessary conditions for (E) and (R). If (E) holds, then

$$\|p - q\|_{L^2(\omega)} \lesssim s|p|_{H^1(\omega)}. \tag{5.1}$$

Assuming that $q = \nabla z_h$ is a first-order finite element approximation of $p = \nabla z$, the nodal interpolation error $|z - Iz|_{H^1(\omega)}$ is generically of size $\|hD^2 z\|_{L^2(\omega)}$ and a lower bound of $\|\nabla z - \nabla z_h\|_{L^2(\omega)}$. For the purpose of this discussion we will therefore suppose the estimate

$$h_{\min,\omega}|p|_{H^1(\omega)} \lesssim \|p - q\|_{L^2(\omega)}. \tag{5.2}$$

This hypothesis is untrue in very particular situations and we thereby exclude it from this discussion. Then, (5.1)–(5.2) imply

$$h_{\min,\omega} \lesssim s. \tag{5.3}$$

From (R) and $|||q||| \lesssim |p|_{H^1(\omega)}$ (a consequence of (E)) one deduces

$$\big(|p|_{H^1(\omega)}/h_\omega\big)\, s|||q||| \lesssim |||q|||\, |p|_{H^1(\omega)}$$

and so (assuming $0 < |p|_{H^1(\omega)}\, |||q|||$)

$$s \lesssim h_\omega. \tag{5.4}$$

In particular, $h_{\min,\omega} \lesssim s \lesssim h_\omega$. Thus, h_ω may *not* be much smaller than the mesh-size $h_{\min,\omega}$ on ω. Moreover, the aforementioned arguments suggest under the hypothesis (SE), (SR), and (5.2) that

$$h_{\min,\omega} \ll s \ll h_\omega. \tag{5.5}$$

This clearly enforces an approximation of q on a mesh much finer than $h_\omega = \operatorname{diam}(\omega)$. The analysis of this paper guarantees $|||q||| = |p|_{H^1(\omega)} + o(1)$ only if (5.5) holds.

5.3. Theoretical choice of s. The choice of $s := \|p-q\|_{L^2(\omega)}/|p|_{H^1(\omega)}$ immediately implies (E). Furthermore, this and

$$\|p-q\|_{L^2(\omega)} \big(|p|_{H^1(\omega)}/h_\omega + |p|_{H^2(\omega)}\big) \lesssim |||q|||^2 \tag{5.6}$$

imply (R). Hence, given a fixed $p \in H^2(\Omega)$ and a fixed domain ω, the condition

$$\|p-q\|_{L^2(\omega)} \to 0 \quad \text{as } h_{\max,\Omega} \to 0$$

and the above choice of s imply (E) and (R).

Proof. The left-hand side of (5.6) tends to zero and so either there holds (R) for sufficiently small $h_{\max,\Omega}$ or $|||q||| \to 0$ as $h_{\max,\Omega} \to 0$. The latter option is excluded for fixed $|p|_{H^1(\omega)} > 0$. ∎

The disadvantage of this choice of s is that it is unknown but required in the computation of $|||q|||$ in (3.1).

5.4. Asymptotic regime for $(h_\omega, h_{\max,\Omega}) \to 0$. To analyze sufficient conditions for (E) and (R) in the limit for $(h_\omega, h_{\max,\Omega}) \to 0$ we suppose that

$$H := h_{\max,\Omega}, \quad h_\omega = H^\alpha, \quad s = H^\beta, \quad \text{and } h_{\min,\omega} = H^\gamma$$

for some convergence rates $\alpha, \beta, \gamma > 0$. For instance, a uniform mesh leads to $\gamma = 1$ and $\alpha < 1$ indicates that the mesh is finer and finer than the size of the domain ω. From (5.2)–(5.4) we obtain $\alpha \le \beta \le \gamma$ as necessary conditions for (E) and (R) and $\alpha < \beta < \gamma$ as necessary conditions for (SE) and (SR). Sufficient conditions include the inverse assumption

$$|p|_{H^1(\omega)} + h_\omega\, |p|_{H^2(\omega)} \lesssim |p|_{H^1(\omega)} \tag{5.7}$$

(generically implied for a fixed $p \in H^2(\Omega)$ as ω varies) and the approximation estimate

$$\|p-q\|_{L^2(\omega)}/|p|_{H^1(\omega)} \lesssim H^\beta. \tag{5.8}$$

Then (5.7)–(5.8) and $\alpha < \beta$ imply (E) and (SR). (The proofs are immediate with the arguments already mentioned.)

It seems to be an open question to design weaker conditions sufficient for (weaker versions of) reliability. Theorem 3.2 suggests looking at the constants in (3.2) which, under the assumptions (5.7)–(5.8), leads to

$$|p|_{H^1(\omega)} \leq |||q||| + cH^{(\beta-\alpha)/2}\, |p|^{1/2}_{H^1(\omega)} \left(|p|_{H^1(\omega)} + |||q|||\right)^{1/2}. \tag{5.9}$$

In fact, that constant $c > 0$ depends on n, $c(\omega)$, and the constants involved in (5.7)–(5.8). For instance, if $\alpha < \beta$ and H is small or if $\alpha = \beta$ and $c < 1$ then (5.9) yields $|p|_{H^1(\omega)} \lesssim |||q|||$.

Acknowledgments. The research was initiated when the author visited the Max Plank Institute in the Sciences of Leipzig, Germany, in 2000 and finished while he visited the Graduiertenkolleg 357 "Effiziente Algorithmen und Mehrskalenmethoden" in 2002 at the University of Kiel, Germany. The support and hospitality of these institutions are thankfully acknowledged.

REFERENCES

[AO] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley-Interscience, New York, 2000.

[BaS] I. Babuška and T. Strouboulis, *The Finite Element Method and Its Reliability*, The Clarendon Press, Oxford University Press, New York, 2001.

[BR] R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer. 10 (2001), 2001, pp. 1–102.

[BMS] F. Brezzi, D. Marini, and E. Süli, *Residual-free bubbles for advection-diffusion problems: The general error analysis*, Numer. Math., 85 (2000), pp. 31–47.

[C1] C. Carstensen, *On the history and future of averaging techniques in a posteriori FE error analysis*, ZAMM 84 (2004), pp. 3–21.

[C2] C. Carstensen, *All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable*, Math. Comp., 73 (2004), pp. 1153–1165.

[CB] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids.* I. *Low order conforming, nonconforming, and mixed FEM*, Math. Comp, 71 (2002), pp. 945–969.

[Ta] L. Tartar, *Remarks on some interpolation spaces*, in Boundary Value Problems for Partial Differential Equations and Applications, Masson, Paris, 1993, pp. 229–252.

[Tr] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., Johann Ambrosius Barth, Heidelberg, 1995.

[V] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Stuttgart, 1996.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2148–2158

ORTHOGONAL HESSENBERG REDUCTION AND ORTHOGONAL KRYLOV SUBSPACE BASES*

JÖRG LIESEN† AND PAUL E. SAYLOR‡

Abstract. We study necessary and sufficient conditions that a nonsingular matrix A can be B-orthogonally reduced to upper Hessenberg form with small bandwidth. By this we mean the existence of a decomposition $AV = VH$, where H is upper Hessenberg with few nonzero bands, and the columns of V are orthogonal in an inner product generated by a hermitian positive definite matrix B. The classical example for such a decomposition is the matrix tridiagonalization performed by the hermitian Lanczos algorithm, also called the orthogonal reduction to tridiagonal form. Does there exist such a decomposition when A is nonhermitian? In this paper we completely answer this question. The related (but not equivalent) question of necessary and sufficient conditions on A for the existence of short-term recurrences for computing B-orthogonal Krylov subspace bases was completely answered by the fundamental theorem of Faber and Manteuffel [*SIAM J. Numer. Anal.*, 21 (1984), pp. 352–362]. We give a detailed analysis of B-normality, the central condition in both the Faber–Manteuffel theorem and our main theorem, and show how the two theorems are related. Our approach uses only elementary linear algebra tools. We thereby provide new insights into the principles behind Krylov subspace methods, that are not provided when more sophisticated tools are employed.

Key words. linear systems, Krylov subspace methods, Hessenberg reduction, matrix decomposition, short-term recurrences, normal matrices, B-normality

AMS subject classifications. 15A21, 15A23, 65F10, 65F25

DOI. 10.1137/S0036142903393372

1. Introduction.

The decompositional approach to matrix computations, formalized by Householder in the 1950s, is counted among the "Top 10" algorithmic ideas of the 20th century [2]. One of the best known of these decompositions is the tridiagonalization of a nonsingular hermitian matrix A; see, e.g., [7, Chapter 9.1.2]. In a nutshell (and without specifying the respective matrix dimensions), for each nonzero vector v there exists a matrix V with first column v, and a square tridiagonal matrix H, such that $AV = VH$, and the columns of V are mutually orthogonal in the Euclidean inner product. This decomposition is computed by the hermitian Lanczos algorithm, and is sometimes called the *orthogonal reduction to tridiagonal form*. It is easy to see, by comparing columns in the matrix equation $AV = VH$, that the first j columns of V form a basis of $\mathcal{K}_j(A, v) \equiv \mathrm{span}\{v, \ldots, A^{j-1}v\}$, the jth Krylov subspace

*Received by the editors September 29, 2003; accepted for publication (in revised form) May 26, 2004; published electronically February 25, 2005.
http://www.siam.org/journals/sinum/42-5/39337.html

†Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de). The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft. Part of this work was done while the author was a postdoctoral research assistant at the Center for Simulation of Advanced Rockets, University of Illinois, Urbana, IL 61801, supported by the U.S. Department of Energy under grant DOE LLNL B341494.

‡Department of Computer Science, University of Illinois, Urbana, IL, 61801 (saylor@cs.uiuc.edu). The work of this author was supported by NASA NCC 5-615 (1-5-21639) and by the U.S. Department of Energy under grant DOE LLNL B341494 at the University of Illinois. The contributions of SCI Dat 1528576 at the University of Illinois are acknowledged, a part of the Terascale Supernova Initiative. During a portion of the preparation of this paper, this author was a Program Director in the Division of Mathematical Sciences at the National Science Foundation. The opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not reflect views either of the sponsoring agencies or the employers.

generated by A and v. The importance of this reduction from a theoretical as well as from a practical point of view can hardly be underestimated.

When A is nonhermitian, we are naturally led to ask for generalizations of the orthogonal reduction to tridiagonal form. Specifically, we ask if there exists a hermitian positive definite (HPD) matrix B such that a nonsingular nonhermitian A can still be B-orthogonally reduced to an upper Hessenberg matrix *with small bandwidth*. B-orthogonally here means that the columns of V are orthogonal in the B-inner product.

This paper studies necessary and sufficient conditions on A that guarantee the existence of such a B-orthogonal reduction. Our subject seems to be elementary, and one might suspect that it is covered in many textbooks on numerical linear algebra. However, while it appears to be common knowledge that the orthogonal reduction to tridiagonal form does not exist in general, see, e.g., [7, p. 499], we are not aware of any publication where the potential for generalizations has been thoroughly studied.

On the other hand, the related question of necessary and sufficient conditions on A for the existence of a short-term recurrence for computing B-orthogonal Krylov subspace basis vectors was completely solved by the fundamental theorem of Faber and Manteuffel [4]. Denoting the columns of V by v_j, we say that these vectors can be computed by an $(s+2)$-term recurrence, when only the previous $s+1$ vectors, $v_{j-s}, \ldots, v_j$, are required to compute v_{j+1}. For example, if the matrix H in the decomposition $AV = VH$ is tridiagonal, then the vectors v_j are computed by a 3-term recurrence. This is a key recurrence in many algorithms, including the famous conjugate gradient method. One immediately expects that for a given matrix A the existence of an $(s+2)$-term recurrence for computing a B-orthogonal Krylov subspace basis is equivalent to B-reducibility of A to upper Hessenberg form with bandwidth $s+2$. However, due to intricate details that are easily overlooked, this expectation is in general false.

Our paper has the following goals. First, we give a thorough analysis of the B-reducibility of a nonsingular matrix A to upper Hessenberg form with small bandwidth. This is an interesting matrix property that apparently was not studied previously. Despite common belief, the necessary and sufficient conditions so that A has this property are not the same as the necessary and sufficient conditions in the Faber–Manteuffel theorem. This situation deserves to be clarified. Second, the proofs in this paper use standard tools of linear algebra only. We thereby hope to provide some additional insight into the necessity of the conditions in the Faber–Manteuffel theorem, for which no elementary (linear algebra based) proof is known. Third, we intend to improve the understanding of B-normality, the central necessary and sufficient condition in our context, by completely characterizing the set of HPD matrices B with respect to which a given matrix A is B-normal. Finally, our goal is to help in the general understanding of the foundations of and principles behind Krylov subspace methods.

The paper is organized as follows. In section 2 we discuss the basic algorithm for B-orthogonal Hessenberg reduction of a matrix and for computing B-orthogonal Krylov subspace bases. In section 3 we explain the sufficiency of B-normality in our context, and study this important concept in detail. In section 4 we discuss the necessity of B-normality. In section 5 we relate our results to the Faber–Manteuffel theorem and the existence of short-term recurrences for computing B-orthogonal Krylov subspace bases. Concluding remarks in section 6 close the paper.

Throughout the paper we assume exact arithmetic. In particular, the word

"computation" in this paper does not refer to a finite precision computation.

2. *B*-orthogonal reduction to upper Hessenberg form. Let A be any nonsingular N by N matrix, let v_1 be any nonzero N-vector (v_1 is assumed to be nonzero to exclude trivialities), and let

$$\mathcal{K}_j(A, v_1) \equiv \operatorname{span}\{v_1, Av_1, \ldots, A^{j-1}v_1\} \quad \text{for} \quad j = 1, 2, \ldots, \tag{2.1}$$

denote the jth Krylov subspace generated by A and v_1. It is well known that the $\mathcal{K}_j(A, v_1)$ form a nested sequence of subspaces of increasing dimension, and that there exists an index

$$d = d(A, v_1) \equiv \dim \mathcal{K}_N(A, v_1), \tag{2.2}$$

which is often called the grade of v_1 with respect to A, for which

$$\mathcal{K}_1(A, v_1) \subset \cdots \subset \mathcal{K}_{d-1}(A, v_1) \subset \mathcal{K}_d(A, v_1) = \mathcal{K}_{d+1}(A, v_1) = \cdots = \mathcal{K}_N(A, v_1).$$

Furthermore, for each v_1, $d \le d(A)$, where $d(A)$ denotes the degree of the minimal polynomial of A.

For any N by N HPD matrix B, the function $\langle \cdot, \cdot \rangle_B$, defined by $\langle x, y \rangle_B \equiv y^* B x$ for N-vectors x and y, is a positive definite inner product. Suppose that, for a given nonsingular matrix A, vector v_1, and HPD matrix B, we want to compute bases of the Krylov subspaces $\mathcal{K}_j(A, v_1)$, for $j = 1, 2, \ldots, d$, that are orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle_B$ (B-orthogonal). In other words, we want to compute vectors $v_1, v_2, \ldots, v_d$ such that

$$\operatorname{span}\{v_1, \ldots, v_j\} = \mathcal{K}_j(A, v_1), \qquad j = 1, \ldots, d, \tag{2.3}$$

$$\langle v_j, v_k \rangle_B = 0, \qquad j \ne k, j = 1, \ldots, d, k = 1, \ldots, d. \tag{2.4}$$

Starting from v_1, this familiar and important task is performed by the following basic algorithm:

$$v_{k+1} = Av_k - \sum_{j=1}^{k} h_{jk} v_j, \qquad k = 1, \ldots, d, \tag{2.5}$$

where

$$h_{jk} = \frac{\langle Av_k, v_j \rangle_B}{\langle v_j, v_j \rangle_B}. \tag{2.6}$$

Apparently, this algorithm is nothing but the classical Gram–Schmidt implementation of Arnoldi's method; see, e.g., [7, Chapter 9.4.1].

Rewriting (2.5) in the form

$$Av_k = v_{k+1} + \sum_{j=1}^{k} h_{jk} v_j, \qquad k = 1, \ldots, d,$$

yields the matrix representation

$$A\,[v_1, \ldots, v_k] = [v_1, \ldots, v_{k+1}] \begin{bmatrix} h_{11} & \cdots & h_{1k} \\ 1 & \ddots & \vdots \\ & \ddots & h_{kk} \\ & & 1 \end{bmatrix}, \tag{2.7}$$

or

$$AV_k = V_{k+1}H_{k+1,k}\,, \qquad k = 1, \ldots, d-1\,, \tag{2.8}$$

where $H_{k+1,k}$ is a $(k+1)$ by k unreduced upper Hessenberg matrix. Since $v_{d+1} = 0$, the matrix representation for $k = d$ may be written as

$$AV_d \;=\; V_d \begin{bmatrix} h_{11} & \cdots & \cdots & h_{1d} \\ 1 & \ddots & & \vdots \\ & \ddots & \ddots & \\ & & 1 & h_{dd} \end{bmatrix} \;=\; V_d H_d\,, \tag{2.9}$$

where H_d is a d by d unreduced upper Hessenberg matrix. The B-orthogonality of the basis vectors, cf. (2.4), in this notation means that $V_d^* B V_d$ is a diagonal matrix.

For given A, v_1, and B, the decomposition (2.9) always exists and conditions (2.3) and (2.4) define it uniquely up to scaling of the columns of V_d. There are in fact several different algorithms that realize conditions (2.3) and (2.4). But in exact arithmetic all these algorithms lead to a decomposition of the form

$$A(V_d D) \;=\; (V_d D)\,(D^{-1} H_d D)\,,$$

where V_d and H_d are as in (2.9), and D is a nonsingular diagonal matrix. Clearly, the nonzero pattern of H_d is invariant under diagonal similarity transformation, and thus, for given A, v_1, and B, conditions (2.3) and (2.4) lead to a uniquely defined nonzero pattern of H_d.

In the following we will be mostly interested in this pattern, particularly in the upper bandwidth of H_d. We call an upper Hessenberg matrix $(s+2)$-*band Hessenberg*, if it has no nonzero entries above its sth superdiagonal. (Here the 0th superdiagonal is the diagonal.) This gives rise to the following definition.

DEFINITION 2.1. *The nonsingular matrix A is B-reducible to $(s+2)$-band Hessenberg form if there exists an HPD matrix B such that for each v_1, either H_d in the decomposition* (2.9) *is $(s+2)$-band Hessenberg, or $d \le s+1$.*

Note that for each nonsingular matrix A, $s \ge 0$ in Definition 2.1, since a 0- or 1-band Hessenberg matrix H_d is singular, which contradicts the nonsingularity of A. Hence B-reducibility to 2-band Hessenberg (lower bidiagonal) form is the one with smallest possible bandwidth that may occur. On the other hand, $s \le d(A) - 1$ always holds as well, since no matrix H_d in (2.9) can possibly have more than $d(A) - 1$ superdiagonals. The condition $d \le s+1$ in Definition 2.1 covers the trivial cases in which H_d has at most $s+2$ bands simply because it is of size at most $s+1$ by $s+1$.

The classical example is the one for $s = 1$, namely the reduction to tridiagonal form (tridiagonalization) with respect to the Euclidean inner product (i.e., $B = I$). For each nonsingular hermitian matrix A and each v_1 the decomposition (2.9) exists, where H_d is a tridiagonal (3-band Hessenberg) matrix and $V_d^* V_d$ is diagonal; see, e.g., [7, Chapter 9.1.2].

In the following sections we will study sufficient (section 3) and necessary (section 4) conditions that A is B-reducible to $(s+2)$-band Hessenberg form. We will then relate our result to the Faber–Manteuffel theorem, which gives necessary and sufficient conditions on A so that for each v_1, a B-orthogonal Krylov subspace basis can be computed by an $(s+2)$-term recurrence (section 5).

3. Sufficiency and characterization of B-normality. Let us consider the sufficient conditions that a given nonsingular matrix A is B-reducible to $(s+2)$-band Hessenberg form. A trivial sufficient condition is that the minimal polynomial of A has degree $d(A) \leq s+1$. Then, for each v_1 we obtain $d \leq d(A) \leq s+1$, and the second sufficient condition in Definition 2.1 is always satisfied. Thus, if $d(A) \leq s+1$, then A is B-reducible to $(s+2)$-band Hessenberg form for each HPD matrix B.

If $d(A) > s+1$, we require that there exists an HPD matrix B so that for each v_1 with $d > s+1$ the matrix H_d in (2.9) is $(s+2)$-band Hessenberg, i.e., that

$$h_{jk} \;=\; 0 \qquad \text{for} \quad j+s+1 \leq k \leq d\,. \tag{3.1}$$

From (2.6) it follows that $h_{jk} = 0$ if and only if

$$0 \;=\; \langle Av_k, v_j\rangle_B \;=\; \langle v_k, A^+ v_j\rangle_B\,,$$

where $A^+ \;\equiv\; B^{-1}A^*B$ is the B-adjoint of A.

Now suppose that $A^+ = p_s(A)$ for a polynomial p_s of degree s. Then, since each v_j is of the form $v_j = p_{j-1}(A)v_1$,

$$A^+ v_j \;=\; p_s(A)p_{j-1}(A)v_1 \;\in\; \mathcal{K}_{j+s}(A, v_1)\,.$$

But then B-orthogonality of v_k to $\operatorname{span}\{v_1, \ldots, v_{k-1}\} = \mathcal{K}_{k-1}(A, v_1)$ for all $k \geq 2$ shows that (3.1) indeed holds for $j+s+1 \leq k \leq d$. We formally state the nontrivial sufficient condition for B-reducibility of A to $(s+2)$-band Hessenberg form in the following definition.

DEFINITION 3.1. *If there exists an HPD matrix B such that*

$$A^+ \;\equiv\; B^{-1}A^*B \;=\; p_s(A)$$

for a polynomial p_s of degree s, then the matrix A is called normal of degree s with respect to B, or short B-normal(s).

Using this definition we can state our main theorem.

THEOREM 3.2. *The nonsingular matrix A is B-reducible to $(s+2)$-band Hessenberg form if and only if either A is B-normal(s), or $d(A) \leq s+1$.*

Above we have shown sufficiency. Before we continue with necessity we will study the important concept of B-normality in more detail. We start with a collection of equivalent characterizations.

THEOREM 3.3. *For any matrix A the following are equivalent:*

1. *There exists an HPD matrix B such that $A^+ = p(A)$ for a polynomial p.*
2. *There exists an HPD matrix B such that $AA^+ = A^+A$.*
3. *A is normalizable (similar to a normal matrix).*
4. *A is diagonalizable.*
5. *There exists an HPD matrix B such that A and A^+ (for this B) have the same complete set of B-orthogonal eigenvectors.*

Proof. (1) $\Rightarrow$ (2). Obviously, $p(A)A = Ap(A)$ for each polynomial p.

(2) $\Rightarrow$ (3). Assume (2) and define the matrix $\mathcal{A} \equiv B^{1/2}AB^{-1/2}$, to which A is similar. Then $\mathcal{A}\mathcal{A}^* = \mathcal{A}^*\mathcal{A}$, i.e., $\mathcal{A}$ is normal.

(3) $\Leftrightarrow$ (4). Suppose that A is normalizable, $A = S^{-1}MS$ with M normal. Since M is diagonalizable [8, Condition 11], A is diagonalizable as well. If A is diagonalizable, $A = WDW^{-1}$ with D diagonal, it is obviously similar to the normal matrix D.

(3) $\Rightarrow$ (1). Again assume that $A = S^{-1}MS$ with M normal. Define $B = S^*S$, which is HPD. By [8, Condition 17], $M^* = p(M)$ for some polynomial p, which implies $A^+ = (S^*S)^{-1}A^*(S^*S) = S^{-1}M^*S = S^{-1}p(M)S = p(A)$.

(3) $\Rightarrow$ (5). If $A = S^{-1}MS$ with M normal, define $B = S^*S$, which is HPD. By [8, Condition 11], there exist a unitary matrix U and a diagonal matrix D such that $M = U^*DU$. Hence, $A = (US)^{-1}D(US)$, where the columns of the matrix $(US)^{-1}$ form a complete set of eigenvectors of A. It is easy to see that these are B-orthogonal. In addition, A^+ for this B is of the form $A^+ = (US)^{-1}D^*(US)$, and hence has the same set of eigenvectors as A.

(5) $\Rightarrow$ (2). If there exists an HPD matrix B such that $A^+ = W\Lambda W^{-1}$ and $A = WDW^{-1}$, where Λ and D are diagonal, then $AA^+ = WD\Lambda W^{-1} = W\Lambda DW^{-1} = A^+A$. ∎

The implications in Theorem 3.3 have appeared in the literature before. In particular, the equivalence of (1), (2), (4), and (5) was proven in [5, Theorem 5 and Corollary 6], and later cited and used, for example, in [9, Theorem 4.4] and [1, p. 772]. However, the proofs in [5] are different from ours, as they do not directly make use of the fact that B-normality is equivalent to normalizability. Because we use this equivalence our proofs appear to be almost trivial, and the list of conditions in Theorem 3.3 can be easily extended by exploiting the lists of equivalent conditions of normality [3, 8], and rephrasing each such condition in terms of normalizability.

We will now characterize the HPD matrices B with respect to which a given (diagonalizable) matrix is normal. Clearly, the matrix B might not be uniquely defined. For example, if A itself is HPD, then it is normal with respect to $B = I$ and $B = A$.

As shown in Theorem 3.3, A is normal with respect to an HPD matrix B if and only if it has a complete set of B-orthogonal eigenvectors. Let these eigenvectors w_i be scaled to have B-norm one, i.e. $w_i^*Bw_i = 1$, and suppose the w_i are the columns of the matrix W. Then $W^*BW = \operatorname{diag}(w_i^*Bw_i) = I$. This is equivalent to $B = (WW^*)^{-1}$. On the other hand, let $B = (WW^*)^{-1}$, where the columns of W form any complete set of eigenvectors of A. Then an easy calculation shows that $AA^+ = A^+A$, i.e., that A is normal with respect to B. We thus have proven the following theorem.

THEOREM 3.4. *Suppose that the matrix A is diagonalizable. Then the set of all HPD matrices B with respect to which A is normal is given by*

$$\{\,(WW^*)^{-1} \;:\; W \text{ is an eigenvector matrix of } A\,\}. \tag{3.2}$$

The characterization (3.2) allows us to derive an expression for the unique B-adjoint of A in case A is normal with respect to B.

COROLLARY 3.5. *Suppose that the matrix A is diagonalizable, $A = W\Lambda W^{-1}$, and that it is normal with respect to an HPD matrix B. Then the B-adjoint of A corresponding to this B is given by*

$$A^+ = W\,\Lambda^*\,W^{-1}. \tag{3.3}$$

In particular, the B-adjoint of A is unique for all HPD matrices B with respect to which A is normal. Moreover, A is B-normal(s) if and only if A is diagonalizable and $p_s(\Lambda) = \Lambda^*$ for a polynomial of degree s.

Proof. Each HPD matrix B with respect to which A is normal is of the form (3.2). A direct computation shows that $A^+ = B^{-1}A^*B$ has the form (3.3), which is unique since it does not depend on the particular choice of W (similarly to A itself). For the second part of the corollary, suppose that A is B-normal(s), i.e., $A^+ = p_s(A)$

for a polynomial of degree s. Then A must be diagonalizable, $A = W\Lambda W^{-1}$, and (3.3) shows that $A^+ = W\Lambda^* W^{-1} = p_s(A) = W p_s(\Lambda) W^{-1}$, which yields $p_s(\Lambda) = \Lambda^*$.

The proof of the reverse implication is similar. ∎

Remark 3.6. The second part of this corollary was also derived in [9, Theorem B.1]. There the authors used a proof different from ours, and do not comment on the uniqueness of the B-adjoint in general.

As shown by Corollary 3.5, the B-normal degree of a diagonalizable matrix is determined by the location of its eigenvalues. It is well known that the B-normal(1) matrices are precisely the diagonalizable matrices that have all eigenvalues on a straight line in the complex plane [4, Lemma 3]. By sufficiency, each such matrix A is B-reducible to 3-band Hessenberg (tridiagonal) form. The standard examples are the hermitian and skew-hermitian matrices, that are all I-normal(1). Rare practical examples of B-normal(1) matrices that are normal with respect to an HPD matrix $B \neq I$ were derived in [6].

When the eigenvalues of the diagonalizable matrix A do not lie on a line, A must have B-normal degree $s > 1$. The question then arises about the lowest degree polynomial p_s for which $p_s(\Lambda) = \Lambda^*$. A recent result of Khavinson and Świątek [10] shows that each harmonic polynomial of the form $p_s(z) - \bar{z}$, where p_s is a polynomial of degree $s > 1$, has at most $3s-2$ complex zeros. Consequently, the class of B-normal(s) matrices with $s > 1$ contains diagonalizable matrices that may have at most $3s - 2$ distinct eigenvalues. This shows that the maximal size of the B-normal(s) matrices for small $s > 1$ is severely limited. We illustrate the results of this section by an example for $s = 3$.

Example 3.7. Consider the third degree harmonic polynomial

$$-\frac{1}{8}\, z(z^2 - 9) - \bar{z}\,, \quad \text{which has the 7 roots} \quad 0,\ \pm 1,\ \frac{\pm 5 \pm \sqrt{-7}}{2}\,.$$

We use the nonzero roots to define the diagonal nonsingular 6 by 6 matrix

$$A \;=\; \operatorname{diag}\left(\ \pm 1,\ \frac{1}{2}(\pm 5 \pm \sqrt{-7})\ \right)\,.$$

By the second part of Corollary 3.5, this matrix is B-normal(3), and by (3.3) its unique B-adjoint is given by $A^+ = A^* = A$. Theorem 3.4 shows that A is normal with respect to all diagonal HPD matrices B. If we use any such matrix B and any v_1 with $d = 6$ in the basic algorithm (2.5)–(2.6), then sufficiency in Theorem 3.2 shows that the resulting matrix H_6 in (2.9) is 5-band Hessenberg. In fact, A is B-reducible to 5-band Hessenberg form for any diagonal HPD matrix B.

4. Necessary conditions. In this section we will show that the conditions in Theorem 3.2 are necessary. To avoid confusion, we will in this section denote the grade of the vector v_1 with respect to A by $d(A, v_1)$, cf. (2.2). Furthermore, we assume that the given nonsingular N by N matrix A is nonderogatory, i.e., that $d(A) = N$. This assumption is made for notational convenience, and without loss of generality. In case A is derogatory, we may in our derivation restrict to starting vectors v_1 with $d(A, v_1) = d(A)$ and all results will then hold for N replaced by $d(A)$. We start our discussion with proving an essential technical lemma.

LEMMA 4.1. *For a nonsingular and nonderogatory N by N matrix A there exists an HPD matrix B such that*

$$A^+ v_1 \;\in\; \mathcal{K}_{s+1}(A, v_1) \qquad \textit{for all } v_1 \textit{ with} \quad d(A, v_1) \;=\; N\,, \tag{4.1}$$

if and only if either A is B-normal(s) for $0 \le s \le N-2$, or $s = N-1$.

Proof. We first prove sufficiency. Suppose that $s = N-1$. Then for all v_1 with $d(A, v_1) = N$, $\mathcal{K}_{s+1}(A, v_1) = \mathcal{K}_N(A, v_1)$ is equal to the whole (real or complex) N-dimensional space and thus $A^+ v_1 \in \mathcal{K}_N(A, v_1)$ holds for all HPD matrices B. Now consider that $0 \le s \le N-2$ and that A is B-normal(s). This means there exists an HPD matrix B for which $A^+ = p_s(A)$ for a polynomial p_s of degree s. Hence for each v_1, $A^+ v_1 = p_s(A) v_1 \in \mathcal{K}_{s+1}(A, v_1)$.

We next prove necessity, the harder part. Suppose that there exists an HPD matrix B such that (4.1) holds for some s with $0 \le s \le N-2$ (in case $s = N-1$ we are done). We need to show that A is B-normal(s).

Let $A = [X_1, \ldots, X_l] \operatorname{diag}(J_1, \ldots, J_l) [X_1, \ldots, X_l]^{-1}$ denote the Jordan canonical form of A, with the distinct eigenvalues $\lambda_1, \ldots, \lambda_l$. Since A is nonderogatory, only one Jordan block corresponds to each eigenvalue and, hence, each X_m, $m = 1, \ldots, l$, is a Jordan chain. Then it is easy to show that the vectors v that satisfy $d(A, v) = N$ are precisely the vectors that have a nonzero component corresponding to the last vector of *each* Jordan chain X_m. As a consequence, when we choose *any* v with $d(A, v) = N$, and *any* nonzero scalar γ with $\gamma \ne -\lambda_m$ for $m = 1, \ldots, l$, then the vector $w \equiv \gamma v + Av$ will satisfy $d(A, v) = d(A, Av) = d(A, w) = N$. In particular, we can find N linearly independent vectors v, such that v, Av, and $w \equiv \gamma v + Av$ for each nonzero $\gamma \ne -\lambda_m$, $m = 1, \ldots, l$, satisfy $d(A, v) = d(A, Av) = d(A, w) = N$.

Suppose that one such vector v is chosen, and let $w \equiv \gamma v + Av$ for some fixed nonzero $\gamma \ne -\lambda_m$, $m = 1, \ldots, l$. Then $d(A, w) = N$, so that by (4.1), $A^+ w \in \mathcal{K}_{s+1}(A, w)$. Hence there exist coefficients $\alpha_j^{(w)}$, $j = 0, \ldots, s$, such that

$$\begin{aligned} A^+ w &= \sum_{j=0}^{s} \alpha_j^{(w)} A^j w \\ &= \gamma \sum_{j=0}^{s} \alpha_j^{(w)} A^j v + \sum_{j=0}^{s} \alpha_j^{(w)} A^{j+1} v \\ &= \gamma \alpha_0^{(w)} v + \sum_{j=1}^{s} (\gamma \alpha_j^{(w)} + \alpha_{j-1}^{(w)}) A^j v + \alpha_s^{(w)} A^{s+1} v \,. \end{aligned} \tag{4.2}$$

Similarly, there exist coefficients $\alpha_j^{(v)}$ and $\alpha_j^{(Av)}$, $j = 0, \ldots, s$, such that

$$\begin{aligned} A^+ w &= \gamma A^+ v + A^+ (Av) \\ &= \gamma \sum_{j=0}^{s} \alpha_j^{(v)} A^j v + \sum_{j=0}^{s} \alpha_j^{(Av)} A^{j+1} v \\ &= \gamma \alpha_0^{(v)} v + \sum_{j=1}^{s} (\gamma \alpha_j^{(v)} + \alpha_{j-1}^{(Av)}) A^j v + \alpha_s^{(Av)} A^{s+1} v \,. \end{aligned} \tag{4.3}$$

Now note that since $d(A, v) = N$ and $0 \le s \le N-2$, the vectors $v, \ldots, A^{s+1} v$ are linearly independent. Thus, the equality of (4.2) and (4.3) implies that

$$\alpha_0^{(v)} = \alpha_0^{(w)} \,, \tag{4.4}$$

$$\gamma \alpha_j^{(v)} + \alpha_{j-1}^{(Av)} = \gamma \alpha_j^{(w)} + \alpha_{j-1}^{(w)} \,, \qquad j = 1, \ldots, s \,, \tag{4.5}$$

$$\alpha_s^{(Av)} = \alpha_s^{(w)} \,. \tag{4.6}$$

Let us define $\eta_j \equiv \alpha_j^{(Av)} - \alpha_j^{(v)}$ for $j = 0, \ldots, s$. By construction, the η_j do not depend on γ. Moreover, we claim that $\eta_j = 0$ for $j = 0, \ldots, s$.

If $s = 0$, then the set of conditions (4.5) is empty and our claim follows directly from comparing (4.4) and (4.6). To show our claim for $1 \leq s \leq N-2$, we rewrite (4.5) in the equivalent form

$$\alpha_j^{(w)} \;=\; \alpha_j^{(v)} \;+\; \frac{1}{\gamma}\,(\alpha_{j-1}^{(Av)} - \alpha_{j-1}^{(w)})\,, \qquad j = 1, \ldots, s\,. \tag{4.7}$$

Then (4.6) and (4.7) for $j = s$ yield

$$\eta_s \;=\; \frac{1}{\gamma}\,(\alpha_{s-1}^{(Av)} - \alpha_{s-1}^{(w)})\,.$$

In this formula we can replace $\alpha_{s-1}^{(w)}$ by the right-hand side of (4.7) for $j = s-1$,

$$\eta_s \;=\; \frac{1}{\gamma}\left(\alpha_{s-1}^{(Av)} - \alpha_{s-1}^{(v)} - \frac{1}{\gamma}\,(\alpha_{s-2}^{(Av)} - \alpha_{s-2}^{(w)})\right) \;=\; \frac{1}{\gamma}\left(\eta_{s-1} - \frac{1}{\gamma}\,(\alpha_{s-2}^{(Av)} - \alpha_{s-2}^{(w)})\right).$$

In the same way we now exploit (4.7) for $j = s-2, \ldots, 1$, and finally use (4.4) to replace $\alpha_0^{(w)}$ by $\alpha_0^{(v)}$. The result of this process is equivalent to the relation

$$\sum_{j=0}^{s} (-1)^{s-j}\,\eta_j\,\gamma^j \;=\; 0\,. \tag{4.8}$$

The coefficients η_j do not depend on γ, so that the left-hand side of (4.8) is a polynomial in γ of degree at most s. Since γ is allowed to vary almost freely without violating the assumption $d(A, v) = d(A, Av) = d(A, w) = N$ (see above), but on the other hand (4.8) must always hold, we conclude that $\eta_j = 0$ for $j = 0, \ldots, s$.

To summarize, since $d(A, v) = d(A, Av) = N$, (4.1) implies that $A^+ v = p_s(A)v$ and $A^+ Av = q_s(A)Av$ for two polynomials p_s and q_s of degree at most s, respectively. But since we have just shown that $p_s = q_s$, we receive

$$A^+ Av \;=\; q_s(A)Av \;=\; p_s(A)Av \;=\; Ap_s(A)v \;=\; AA^+ v\,.$$

Since we can find N linearly independent vectors v for which this is true, we conclude that $A^+ A = AA^+$, and indeed A must be B-normal(s). ∎

Remark 4.2. This lemma represents a strengthened version of a result of Faber and Manteuffel [4, Lemma 2]. In their result, (4.1) is replaced by "$A^+ v_1 \in \mathcal{K}_{s+1}(A, v_1)$ for all v_1," and their proof of necessity uses an eigenvector v_1 of A, i.e., a vector v_1 with $d(A, v_1) = 1$. Our proof is inspired by an idea of Voevodin and Tyrtyshnikov [12].

Necessity in Theorem 3.2. We now come to the main goal of this section, namely the proof of necessity in Theorem 3.2. Our supposedly necessary condition reads *either A is B-normal(s), or $N - 1 \leq s$*. We will prove this by assuming the opposite and then showing that A is *not* B-reducible to $(s+2)$-band Hessenberg form.

The opposite of our necessary condition is that *A is not B-normal(s) and* $0 \leq s \leq N-2$. But this is precisely the opposite of the nontrivial necessary condition in Lemma 4.1. Therefore, Lemma 4.1 implies that for each HPD matrix B there exists at least one vector v_1 with $d(A, v_1) = N$ such that $A^+ v_1 \notin \mathcal{K}_{s+1}(A, v_1)$. On the other hand, since $d(A, v_1) = N$, $A^+ v_1 \in \mathcal{K}_N(A, v_1)$. Hence $A^+ v_1$ is a linear combination of the basis vectors $v_1, \ldots, v_N$ computed by (2.5),

$$A^+ v_1 = \sum_{j=1}^{N} \beta_j v_j\,,$$

where (at least) one of the coefficients β_j, $s+2 \le j \le N$, is nonzero. Let this be the coefficient with index k. Then, according to (2.6) and the B-orthogonality conditions (2.4),

$$h_{1k} \;=\; \frac{\langle Av_k, v_1\rangle_B}{\langle v_1, v_1\rangle_B} \;=\; \frac{\langle v_k, A^+v_1\rangle_B}{\langle v_1, v_1\rangle_B} \;=\; \frac{\overline{\beta}_k \langle v_k, v_k\rangle_B}{\langle v_1, v_1\rangle_B} \;\neq\; 0\,, \tag{4.9}$$

for an index k with $s+2 \le k \le N$. Consequently, A is *not* B-reducible to $(s+2)$-band Hessenberg form, which completes the proof of necessity.

5. The existence of $(s+2)$-term recurrences. We next relate Theorem 3.2 to the existence of an $(s+2)$-term recurrence for computing B-orthogonal Krylov subspace bases.

Suppose that, for a given nonsingular matrix A, vector v_1, and HPD matrix B, only the previous $s+1$ vectors $v_k, v_{k-1}, \ldots, v_{k-s}$ are required to compute v_{k+1}, $k = 1, \ldots, d-1$, in (2.5). Then we say that the B-orthogonal basis $v_1, \ldots, v_d$ of $\mathcal{K}_d(A, v_1)$ is computed by an $(s+2)$-term recurrence.

Note that the basic algorithm (2.5)–(2.6) has computed all basis vectors $v_1, \ldots, v_d$ in step $k = d-1$. Therefore, in terms of the matrix representation (2.8), the B-orthogonal basis vectors are computed by an $(s+2)$-term recurrence, if $H_{d,d-1}$ is $(s+2)$-band Hessenberg. We stress that, unlike for the B-reducibility to $(s+2)$-band Hessenberg form, we here use $H_{d,d-1}$ and not H_d. It is a subtle and easily overlooked fact that the last column of H_d plays no role for the computation of the basis vectors $v_1, \ldots, v_d$. This column indeed may be full, and still the basis vectors are computed by an $(s+2)$-term recurrence.

DEFINITION 5.1. *The nonsingular matrix A admits an $(s+2)$-term recurrence, if there exists an HPD matrix B so that for each v_1, either $H_{d,d-1}$ in the decomposition* (2.8) *is $(s+2)$-band Hessenberg, or $d \le s+2$.*

The difference in the trivial conditions between this definition and Definition 2.1 ($d \le s+2$ versus $d \le s+1$) precisely corresponds to the different roles of the matrices $H_{d,d-1}$ and H_d. We immediately realize that if a nonsingular matrix A is B-reducible to $(s+2)$-band Hessenberg form, then it also admits an $(s+2)$-term recurrence. The reverse implication, however, does not hold. In other words, B-reducibility to $(s+2)$-band Hessenberg form and admissibility of an $(s+2)$-term recurrence are in general *not equivalent*. As an example, consider any nonsingular 3 by 3 matrix A with $d(A) = 3$. Then each v_1 with $d = 3$ leads to an H_3 of the form

$$H_3 \;=\; \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ 1 & h_{22} & h_{23} \\ 0 & 1 & h_{33} \end{bmatrix}.$$

Trivially A admits a 3-term recurrence as each $H_{3,2}$ has only 3 nonzero bands (here $s = 1$). But B-reducibility to 3-band Hessenberg form requires that $h_{13} = 0$ for all v_1. Since $d(A) = 3 > s+1$, this holds by Theorem 3.2 only when A is B-normal(1).

Necessary and sufficient conditions for admissibility of an $(s+2)$-term recurrence were proven in the fundamental paper of Faber and Manteuffel [4].

THEOREM 5.2 (Faber–Manteuffel theorem). *The nonsingular matrix A admits an $(s+2)$-term recurrence if and only if either A is B-normal(s), or $d(A) \le s+2$.*

The proof of this result is based on a highly nontrivial and clever construction, which, unfortunately, provides little insight into the necessity of B-normality. A similar result was announced by Voevodin [11], but its proof by Voevodin and Tyrtyshnikov [12] is difficult to understand and appears to be unknown even to many

specialists in the field. Comparing Theorem 3.2 with the Faber–Manteuffel theorem yields the following important observation.

COROLLARY 5.3. *The nonsingular matrix A with $d(A) > s+2$ is B-reducible to $(s+2)$-band Hessenberg form if and only if it admits an $(s+2)$-term recurrence.*

In other words, in all cases of practical interest; i.e., when $d(A)$ is "large" and s is "small," the two matrix properties studied in this paper in fact are equivalent.

6. Concluding remarks. The reader may now ask if we successfully tried to prove necessity in the Faber–Manteuffel theorem using similar elementary (linear algebra) means as for our Theorem 3.2. The answer is yes, we tried, but no, we were unsuccessful. To explain the main difficulty, at least in our opinion, consider our proof of necessity in section 4. We assume the opposite of the necessary conditions and construct a certain nonzero entry h_{1k}, $s+2 \le k \le N$, in the first row of H_N, cf. (4.9). Hence H_N cannot be $(s+2)$-band Hessenberg, which leads to a contradiction showing that the conditions are indeed necessary. Except for the range $s+2 \le k \le N$, we have no information about the location of the nonzero entry h_{1k}. If we could show that indeed $h_{1k} \neq 0$ for a k in the range $s+2 \le k \le N-1$, then $H_{N,N-1}$ cannot be $(s+2)$-band Hessenberg either. This could subsequently be used to show necessity in the Faber–Manteuffel theorem. However, it is apparently quite difficult to "fix" k inside the range $s+2 \le k \le N-1$.

Acknowledgments. We thank Tom Manteuffel for sharing his insights in a thorough discussion of the subject that took place during the Copper Mountain Conference on Iterative Methods in March 2002. We also thank Zdeněk Strakoš, Petr Tichý, and an anonymous referee for many helpful and constructive comments.

REFERENCES

[1] T. BARTH AND T. MANTEUFFEL, *Multiple recursion conjugate gradient algorithms Part* I: *Sufficient conditions*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 768–796.

[2] B. A. CIPRA, *The Best of the* 20*th Century: Editors Name Top* 10 *Algorithms*, SIAM News, 33 (2000).

[3] L. ELSNER AND K. D. IKRAMOV, *Normal matrices: An update*, Linear Algebra Appl., 285 (1998), pp. 291–303.

[4] V. FABER AND T. MANTEUFFEL, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[5] V. FABER AND T. A. MANTEUFFEL, *Orthogonal error methods*, SIAM J. Numer. Anal., 24 (1987), pp. 170–187.

[6] B. FISCHER, A. RAMAGE, D. J. SILVESTER, AND A. J. WATHEN, *Minimum residual methods for augmented systems*, BIT, 38 (1998), pp. 527–543.

[7] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[8] R. GRONE, C. R. JOHNSON, E. M. DE SÁ, AND H. WOLKOWICZ, *Normal matrices*, Linear Algebra Appl., 87 (1987), pp. 213–225.

[9] W. D. JOUBERT AND D. M. YOUNG, *Necessary and sufficient conditions for the simplification of generalized conjugate-gradient algorithms*, Linear Algebra Appl., 88/89 (1987), pp. 449–485.

[10] D. KHAVINSON AND G. ŚWIĄTEK, *On the number of zeros of certain harmonic polynomials*, Proc. Amer. Math. Soc., 131 (2003), pp. 409–414.

[11] V. V. VOEVODIN, *The question of non-self-adjoint extension of the conjugate gradients method is closed*, U.S.S.R. Comput. Math. and Math. Phys., 23 (1983), pp. 143–144.

[12] V. V. VOEVODIN AND E. E. TYRTYSHNIKOV, *On generalization of conjugate direction methods*, in Numerical Methods of Algebra (Chislennye Metody Algebry), Moscow State University Press, Moscow, 1981, pp. 3–9. English translation provided by E. E. Tyrtyshnikov.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2159–2175

A PRECONDITIONER FOR THE FETI-DP FORMULATION WITH MORTAR METHODS IN TWO DIMENSIONS*

HYEA HYUN KIM† AND CHANG-OCK LEE†

Abstract. In this paper, we consider a dual-primal FETI (FETI-DP) method for elliptic problems on nonmatching grids. The FETI-DP method is a domain decomposition method that uses Lagrange multipliers to match solutions continuously across subdomain boundaries in the sense of dual-primal variables. We use the mortar matching condition as the continuity constraints for the FETI-DP formulation. We construct a preconditioner for the FETI-DP operator and show that the condition number of the preconditioned FETI-DP operator is bounded by

$$C \max_{i=1,\dots,N}\{(1+\log(H_i/h_i))^2\},$$

where H_i and h_i are sizes of domain and mesh for each subdomain, respectively, and C is a constant independent of H_i's and h_i's. We allow jumps of coefficients of elliptic problems across subdomain boundaries. Numerical results are included.

Key words. FETI-DP, nonmatching grids, mortar methods, preconditioner

AMS subject classifications. 65N30, 65N55

DOI. 10.1137/S0036142903423381

1. Introduction.

This paper is concerned with preconditioners for an iterative method for the parallel solution of symmetric, positive definite systems of linear equations that arise from elliptic boundary value problems discretized by finite elements on nonconforming meshes. Nonconforming discretizations are important for multiphysics simulations, contact-impact problems, the generation of meshes and partitions aligned with jumps in diffusion coefficients, *hp*-adaptive methods, and special discretizations in the neighborhood of singularities (corners or joints).

Of the many methods for nonmatching meshes, including [3] and [13], we consider the mortar method [1, 2, 17, 18]. In mortar methods, orthogonality relations between the jumps in the traces across subdomain interfaces are satisfied using a discrete Lagrange multiplier space. The sparse linear systems that arise in mortar methods are similar to the systems solved by an iterative substructuring method with Lagrange multipliers developed for conforming discretizations (see [6, 8, 12, 14] for an introduction).

Recently the dual-primal FETI (FETI-DP) method introduced by Farhat, Lesoinne, and Pierson [7] has been applied to mortar finite elements methods [4, 5, 15]. The primary contribution of our work is using the framework of parallel subspace correction methods [19] to better formulate the FETI-DP preconditioner for the mortar matching condition.

The FETI-DP method enforces the continuity of the solution at cross points directly in the formulation of the dual problem: the degrees of freedom (d.o.f.) at a cross point remain common to all subdomains sharing the cross point and the continuity of the remaining d.o.f. on the interfaces are enforced by Lagrange multipliers [10]. The

*Received by the editors February 20, 2003; accepted for publication (in revised form) May 27, 2004; published electronically February 25, 2005. This work was partially supported by KRF-2001-041-D00038.

http://www.siam.org/journals/sinum/42-5/42338.html

†Division of Applied Mathematics, KAIST, Daejeon 305-701, Korea (mashy@amath.kaist.ac.kr, colee@amath.kaist.ac.kr).

d.o.f. are then eliminated and the resulting dual problem for the Lagrange multipliers is solved by preconditioned conjugate gradients (CGs).

For FETI-DP methods on nonmatching grids, Dryja and Widlund [4] proposed a preconditioner, the so-called Dirichlet preconditioner, which gives a condition number bound $C(1+\log(H/h))^2$ with the Neumann–Dirichlet ordering of substructures, where H and h denote the maximum diameter of subdomains and minimum size of meshes of all subdomains, respectively. In general cases, that is, without considering ordered substructures, they obtained $C(1+\log(H/h))^4$ for the condition number bound. Moreover, in [5], they proposed a different preconditioner, which is similar to the one in [9], and proved the condition number bound $C(1+\log(H/h))^2$. However, the constant C in the condition number bound depends on the ratio of meshes between neighboring subdomains. This restriction is impractical when the coefficients of elliptic problems are highly discontinuous between subdomains (see Wohlmuth [18]).

In this paper, we formulate an FETI-DP operator in a different way from that of Dryja and Widlund [4, 5] and propose a Neumann–Dirichlet preconditioner which gives the condition number bound $C(1+\log(H/h))^2$ with the constant C not depending on the ratio of meshes between neighboring subdomains. The proposed preconditioner is easy to implement and the operator from the nodal values on the interfaces of subdomains to the Lagrange multiplier space requires only the nodal values on the slave side. Hence, the cost for multiplying the operator to a vector is reduced by half compared with preconditioners developed elsewhere (see, e.g., [4, 5]). For the elliptic problems with heterogeneous coefficients, with careful choices of slave and master sides according to the magnitude of coefficients, the preconditioner gives the same condition number bound, which does not depend on the coefficients.

In the mortar matching condition, we consider a standard Lagrange multiplier space introduced by [2]. In the condition number analysis, we use the continuity of the mortar projection operator in an $H_{00}^{1/2}$-norm. Hence, our result can be extended to Lagrange multiplier spaces with this property. A few such Lagrange multiplier spaces are developed by Wohlmuth [17, 18].

This paper is organized as follows. In section 2, we introduce finite element spaces and norms and, in section 3, we derive the FETI-DP operator using the mortar matching condition as continuity constraints and propose a preconditioner. In section 4, we analyze the condition number bound of the preconditioned FETI-DP operator. Numerical results are provided in section 5. In the numerical tests, we compare the proposed preconditioner with that of Dryja and Widlund [5] for solving elliptic problems with highly discontinuous coefficients on noncomparable meshes.

2. Finite element spaces and norms.

2.1. A model problem and Sobolev spaces. Let Ω be a bounded polygonal domain in $\mathbb{R}^2$ and $L^2(\Omega)$ be the space of square integrable functions defined in Ω equipped with the norm $\|\cdot\|_{0,\Omega}$:

$$\|v\|_{0,\Omega}^2 := \int_\Omega v^2\,dx.$$

The space $H^1(\Omega)$ is the set of functions, which are square integrable up to the first weak derivatives, and the norm is given by

$$\|v\|_{1,\Omega} := \left(\int_\Omega \nabla v\cdot\nabla v\,dx + \frac{1}{d_\Omega^2}\int_\Omega v^2\,dx\right)^{1/2},$$

where d_Ω denotes the diameter of Ω.

We consider an FETI-DP method on nonmatching grids for the following elliptic problem: For $f \in L^2(\Omega)$, find $u \in H^1(\Omega)$ such that

$$\begin{aligned} -\nabla \cdot (A(x)\nabla u(x)) + \beta(x)u(x) &= f(x) \ \text{ in } \Omega, \\ u(x) &= 0 \ \text{ on } \Gamma_D, \\ \boldsymbol{n} \cdot (A(x)\nabla u(x)) &= 0 \ \text{ on } \Gamma_N. \end{aligned} \tag{2.1}$$

Here, $A(x) = (\alpha_{ij}(x))$ for $i, j = 1, 2$ and $\boldsymbol{n}$ is the outward unit vector normal to Γ_N. We assume that $\alpha_{ij}(x), \beta(x) \in L^\infty(\Omega)$, $A(x)$ is uniformly elliptic, $\beta(x) \geq 0$ for all $x \in \Omega$, and $|\Gamma_D| \neq 0$, where $|\Gamma_D|$ denotes the measure of Γ_D.

Let Ω be partitioned into nonoverlapping polygonal subdomains $\{\Omega_i\}_{i=1}^N$. We assume that the partition is geometrically conforming, which means that the subdomains intersect with neighboring subdomains on the whole of an edge or at a vertex. Ω_i^h denotes a quasi-uniform triangulation of the subdomain Ω_i. The quasi uniformity means that there exist constants γ and σ such that $\gamma h_i \leq d_\tau \leq \sigma \rho_\tau$ for all $\tau \in \Omega_i^h$, where ρ_τ is the diameter of the circle inscribed in τ, d_τ is the diameter of τ, and $h_i = \max_{\tau \in \Omega_i^h} d_\tau$. We note that the meshes need not match across the subdomain interfaces.

For each subdomain Ω_i, we introduce a finite element space

$$X_i := \{v \in H_D^1(\Omega_i) \ : \ v|_\tau \in P_1(\tau), \ \tau \in \Omega_i^h\},$$

where $H_D^1(\Omega_i) := \{v \in H^1(\Omega_i) \ : \ v = 0 \text{ on } \Gamma_D \cap \partial\Omega_i\}$ and $P_1(\tau)$ is a set of polynomials of degree ≤ 1 in τ. For $(u_i, v_i) \in X_i \times X_i$, define a bilinear form

$$a_i(u_i, v_i) := \int_{\Omega_i} A(x)\nabla u_i \cdot \nabla v_i \, dx + \int_{\Omega_i} \beta(x) u_i v_i dx.$$

To get the FETI-DP formulation, we need a finite element space in Ω as follows:

$$X := \left\{ v \in \prod_{i=1}^N X_i \ : \ v \text{ is continuous at subdomain vertices} \right\}.$$

By restricting the space X_i on the boundary of the subdomain Ω_i, we define

$$W_i := X_i|_{\partial\Omega_i} \ \forall i = 1, \ldots, N.$$

Then we let

$$W := \left\{ w \in \prod_{i=1}^N W_i \ : \ w \text{ is continuous at subdomain vertices} \right\}. \tag{2.2}$$

In this paper, we will use the same notation for finite element functions and the corresponding vectors of nodal values. For example, w_i is used to denote a finite element function or the vector of nodal values of that function. The same applies to the notation for function spaces such as W_i, X, W, etc.

We define S^i as the Schur complement matrix obtained from the bilinear form $a_i(\cdot, \cdot)$ over the finite elements X_i (see page 50 in [11]). Using this operator, a seminorm is defined for $w_i \in W_i$:

$$|w_i|_{S^i}^2 := \langle S^i w_i, w_i \rangle,$$

where $\langle\cdot,\cdot\rangle$ is the l^2-inner product of vectors. For $w \in W$, since w is continuous at subdomain vertices, by summing up these seminorms, we define a norm

$$\|w\|_W^2 := \sum_{i=1}^{N} |w_i|_{S^i}^2, \quad w_i = w|_{\partial\Omega_i}. \tag{2.3}$$

Moreover, we define a subspace of W

$$W_r := \{w \in W \ : \ w \text{ vanishes at subdomain vertices}\}. \tag{2.4}$$

Now, we introduce Sobolev spaces defined on the boundaries of subdomains. The space $H^{1/2}(\partial\Omega_i)$ is the trace space of $H^1(\Omega_i)$ equipped with the norm

$$\|w_i\|_{1/2,\partial\Omega_i}^2 := |w_i|_{1/2,\partial\Omega_i}^2 + \frac{1}{d_{\Omega_i}}\|w_i\|_{0,\partial\Omega_i}^2,$$

where

$$|w_i|_{1/2,\partial\Omega_i}^2 := \int_{\partial\Omega_i}\int_{\partial\Omega_i} \frac{|w_i(x) - w_i(y)|^2}{|x-y|^2}\, ds(x)\, ds(y).$$

For any $\Gamma_{ij} \in \partial\Omega_i$, $H_{00}^{1/2}(\Gamma_{ij})$ is the set of functions in $L^2(\Gamma_{ij})$ such that the zero extension of the function into $\partial\Omega_i$ is contained in $H^{1/2}(\partial\Omega_i)$. For $v \in H_{00}^{1/2}(\Gamma_{ij})$, let

$$|v|_{H_{00}^{1/2}(\Gamma_{ij})}^2 := |v|_{1/2,\Gamma_{ij}}^2 + \int_{\Gamma_{ij}} \frac{v^2(x)}{\operatorname{dist}(x,\partial\Gamma_{ij})}\, ds,$$

and the norm is given by

$$\|v\|_{H_{00}^{1/2}(\Gamma_{ij})} := \left(|v|_{H_{00}^{1/2}(\Gamma_{ij})}^2 + \frac{1}{d_{\Omega_i}}\|v\|_{0,\Gamma_{ij}}^2\right)^{1/2}.$$

From section 4.1 in [19], for $v \in H_{00}^{1/2}(\Gamma_{ij})$ we have the following relation:

$$C_1\|\widetilde{v}\|_{1/2,\partial\Omega_i} \le \|v\|_{H_{00}^{1/2}(\Gamma_{ij})} \le C_2\|\widetilde{v}\|_{1/2,\partial\Omega_i}, \tag{2.5}$$

where the constants C_1 and C_2 are independent of d_{Ω_i} and $\widetilde{v}$ denotes the zero extension of v into $\partial\Omega_i$.

2.2. Mortar matching conditions. We note that the space X is not contained in $H^1(\Omega)$. To approximate the solution of the problem (2.1) in X, we use the mortar matching condition. More precisely, we construct the Lagrange multiplier space as follows.

First, let $\Gamma_{ij} := \partial\Omega_i \cap \partial\Omega_j$. For Γ_{ij} such that $|\Gamma_{ij}| \neq 0$, we distinguish $\Omega_i^h|_{\Gamma_{ij}}$ and $\Omega_j^h|_{\Gamma_{ij}}$, as in Figure 1. We assume that both sides have more than three nodal points including end points. Then we choose one as a slave side and the other as a master side and define

$$m_i := \{j \ : \ |\Gamma_{ij}| \neq 0,\ \Omega_j^h|_{\Gamma_{ij}} \text{ is a master side of } \Gamma_{ij}\},$$
$$s_i := \{j \ : \ |\Gamma_{ij}| \neq 0,\ \Omega_j^h|_{\Gamma_{ij}} \text{ is a slave side of } \Gamma_{ij}\}.$$

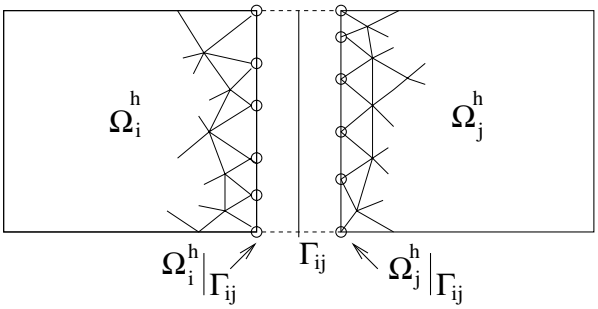


FIG. 1. *Master and slave sides of* Γ_{ij}.

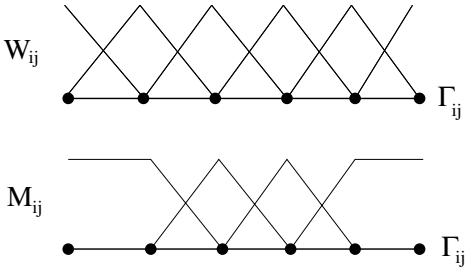


FIG. 2. *Basis functions for* W_{ij} *and* M_{ij}.

For $j \in m_i$, $\Omega_i^h|_{\Gamma_{ij}}$ is the slave side of Γ_{ij} and from the finite elements on the slave side, we get

$$W_{ij} := \{v|_{\Gamma_{ij}} \ : \ v \in X_i\} \ \ \forall\, j \in m_i.$$

Next, let

$$\{\phi_0^{ij},\ \phi_1^{ij}, \ldots, \phi_{N_{ij}}^{ij},\ \phi_{N_{ij}+1}^{ij}\}$$

be nodal basis functions for W_{ij}. Moreover, we assume that the basis functions are sequentially ordered according to the location of nodes on Γ_{ij}. Let (see Figure 2)

$$M_{ij} := \text{span}\{\phi_0^{ij} + \phi_1^{ij},\ \phi_2^{ij}, \ldots, \phi_{N_{ij}-1}^{ij}, \phi_{N_{ij}}^{ij} + \phi_{N_{ij}+1}^{ij}\}.$$

Then we take the Lagrange multiplier space

$$M := \prod_{i=1}^{N} \prod_{j \in m_i} M_{ij}.$$

Bernardi, Maday, and Patera [2] first introduced this type of Lagrange multiplier space. They imposed the following mortar matching condition on X, i.e., $v \in X$ satisfies

$$\int_{\Gamma_{ij}} (v_i - v_j)\lambda_{ij}\, ds = 0 \ \ \forall\ \lambda_{ij} \in M_{ij},\ i = 1, \ldots, N,\ j \in m_i. \tag{2.6}$$

In our FETI-DP formulation, we use (2.6) as continuity constraints and define a bilinear form $b(\cdot,\cdot) : W \times M \to \mathbb{R}$ as

$$b(w,\mu) := \sum_{i=1}^{N} \sum_{j\in m_i} \int_{\Gamma_{ij}} (w_i - w_j)\,\mu_{ij}\,ds \quad \forall (w,\mu) \in W \times M. \tag{2.7}$$

For $|\partial\Omega_i \cap \partial\Omega_j| \neq 0$, we denote $\partial\Omega_i \cap \partial\Omega_j$ as Γ_{ij} if $\Omega_i^h|_{\Gamma_{ij}}$ is a slave side and as Γ_{ji} otherwise. We assume that $\Omega_i^h|_{\Gamma_{ij}}$ is the slave side and $\Omega_j^h|_{\Gamma_{ij}}$ is the master side of Γ_{ij}. Denote the basis for M_{ij} by $\{\xi_k^{ij}\}_{k=1}^{N_{ij}}$ and let $\{\phi_k^{ji}\}_{k=0}^{N_{ji}+1}$ be the basis functions for $W_j|_{\Gamma_{ij}}$. Define matrices B_i^{ij} and B_j^{ij} with entries

$$\left(B_i^{ij}\right)_{lk} = \int_{\Gamma_{ij}} \xi_l^{ij}\phi_k^{ij}\,ds,\ l = 1,\dots,N_{ij},\ k = 0,\dots,N_{ij}+1,$$
$$\left(B_j^{ij}\right)_{lk} = -\int_{\Gamma_{ij}} \xi_l^{ij}\phi_k^{ji}\,ds,\ l = 1,\dots,N_{ij},\ k = 0,\dots,N_{ji}+1.$$

Then we rewrite (2.6) as

$$B_i^{ij} w_i^{ij} + B_j^{ij} w_j^{ij} = 0,$$

where $w_i^{ij} = v_i|_{\Gamma_{ij}}$ and $w_j^{ij} = v_j|_{\Gamma_{ij}}$.

Now define $E_{ij} : M_{ij} \to M$, an extension operator from M_{ij} to M by zero, and $R_{ij}^l : W_l \to W_l|_{\Gamma_{ij}}$ for $l = i, j$, a restriction operator. Let

$$B_i = \sum_{j\in m_i} E_{ij} B_i^{ij} R_{ij}^i + \sum_{j\in s_i} E_{ji} B_i^{ji} R_{ji}^i.$$

Then the mortar matching condition (2.6) becomes

$$\sum_{i=1}^{N} B_i w_i = 0,$$

where $w_i = v_i|_{\partial\Omega_i}$.

Define

$$W_{ij}^0 := \{v \in W_{ij} \ :\ v = 0 \text{ at the end points of } \Gamma_{ij}\}$$

and let

$$W^0 = \prod_{i=1}^{N} \prod_{j\in m_i} W_{ij}^0.$$

For $w_{ij} \in W_{ij}^0$, we define $\widetilde{w}_{ij} \in W_i$ by the zero extension of w_{ij} into $\partial\Omega_i$. Let $\widetilde{w}_i = \sum_{j\in m_i} \widetilde{w}_{ij}$ and $\widetilde{w} = (\widetilde{w}_1,\dots,\widetilde{w}_N)$. Since $\widetilde{w}$ is continuous at subdomain vertices, $\widetilde{w} \in W$. Hence for $w \in W^0$, we define a norm by

$$\|w\|_{W^0} := \|\widetilde{w}\|_W. \tag{2.8}$$

Let $\langle\cdot,\cdot\rangle_m$ be a duality pairing between M and W^0 such that

$$\langle\lambda, w\rangle_m := \sum_{i=1}^{N} \sum_{j\in m_i} \int_{\Gamma_{ij}} \lambda_{ij} w_{ij}\,ds \ \ \forall\,(\lambda, w) \in M \times W^0. \tag{2.9}$$

Using this, we define a dual norm on M by

$$\|\lambda\|_{(W^0)'} := \max_{w \in W^0 \setminus \{0\}} \frac{\langle \lambda, w \rangle_m}{\|w\|_{W^0}}. \tag{2.10}$$

3. FETI-DP formulation.

3.1. FETI-DP operator. In this section, we construct the FETI-DP operator for the problem (2.1) with the mortar matching condition as constraints. The derivation of the FETI-DP equation for the Lagrange multipliers follows [10]. However, the FETI-DP operator with mortar matching condition is new. Dryja and Widlund [4, 5] eliminate unknowns on both interior and vertex nodal points and impose a mortar matching condition over W_r in (2.4). Hence, the resulting solution u does not satisfy the mortar matching condition (2.6). We eliminate only interior nodal points and impose the mortar matching condition on the function over W in (2.2).

For $w_i \in W_i$ we write

$$w_i = \begin{pmatrix} w_r^i \\ w_c^i \end{pmatrix},$$

where r and c stand for the nodal values on the edges and vertices. From now on, we use the subscript symbol r and c to represent the d.o.f. corresponding to nodes on the edges and at the vertices, respectively.

Define W_c as the set of vectors which have d.o.f. corresponding to the union of subdomain vertices, that is, global corner points. For $w = (w_1, \dots, w_N) \in W$, since w is continuous at subdomain vertices, there exists $w_c \in W_c$ such that $L_c^i w_c = w_c^i$ for all $i = 1, \dots, N$, where the matrix L_c^i consists of 0 and 1 and restricts the value of w_c on the vertices of subdomain Ω_i. Hence, for $w = (w_1, \dots, w_N) \in W$, we write

$$w_i = \begin{pmatrix} w_r^i \\ L_c^i w_c \end{pmatrix} \ \forall i \ \text{ for some } w_c \in W_c.$$

Recall that S^i is the Schur complement matrix obtained from the bilinear form $a_i(\cdot,\cdot)$ and let g^i be the Schur complement forcing vector obtained from $\int_{\Omega_i} f v_i \, dx$. The matrix S^i and vector g^i are ordered in the following way:

$$S^i = \begin{pmatrix} S_{rr}^i & S_{rc}^i \\ S_{cr}^i & S_{cc}^i \end{pmatrix}, \quad g^i = \begin{pmatrix} g_r^i \\ g_c^i \end{pmatrix}.$$

Let $B_{i,r}$ and $B_{i,c}$ be matrices that consist of the columns of B_i corresponding to the nodal points on the edges and at the vertices, respectively.

Then the problem (2.1) becomes the following: Find $(w_r, w_c, \lambda) \in W_r \times W_c \times M$ such that

$$S_{rr} w_r + S_{rc} w_c + B_r^t \lambda = g_r, \tag{3.1}$$

$$S_{cr} w_r + S_{cc} w_c + B_c^t \lambda = g_c, \tag{3.2}$$

$$B_r w_r + B_c w_c = 0, \tag{3.3}$$

where

$$S_{rr} = \mathrm{diag}_{i=1,\dots,N} \left(S_{rr}^i \right),$$

$$S_{rc} = \begin{pmatrix} S_{rc}^1 L_c^1 \\ \vdots \\ S_{rc}^N L_c^N \end{pmatrix},$$

$$
\begin{aligned}
&S_{cr} = S_{rc}^t,\\
&S_{cc} = \sum_{i=1}^{N} (L_c^i)^t S_{cc}^i L_c^i,\\
&B_r = (B_{1,r}, \dots, B_{N,r}), \quad B_c = \sum_{i=1}^{N} B_{i,c} L_c^i,\\
&g_r = \begin{pmatrix} g_r^1 \\ \vdots \\ g_r^N \end{pmatrix}, \; g_c = \sum_{i=1}^{N} (L_c^i)^t \; g_c^i, \; w_r = \begin{pmatrix} w_r^1 \\ \vdots \\ w_r^N \end{pmatrix}.
\end{aligned}
$$

Since S_{rr} is invertible, we solve (3.1) for w_r to get

$$w_r = S_{rr}^{-1} \left(g_r - S_{rc} w_c - B_r^t \lambda \right).$$

After substituting w_r into (3.3) and (3.2), we obtain

$$
\begin{aligned}
B_r S_{rr}^{-1} B_r^t \lambda + \left(B_r S_{rr}^{-1} S_{rc} - B_c \right) w_c &= B_r S_{rr}^{-1} g_r,\\
\left(S_{cr} S_{rr}^{-1} B_r^t - B_c^t \right) \lambda - \left(S_{cc} - S_{cr} S_{rr}^{-1} S_{rc} \right) w_c &= - \left(g_c - S_{cr} S_{rr}^{-1} g_r \right).
\end{aligned}
$$

Let

$$
\begin{aligned}
&F_{I_{rr}} = B_r S_{rr}^{-1} B_r^t,\\
&F_{I_{rc}} = B_r S_{rr}^{-1} S_{rc} - B_c,\\
&F_{I_{cr}} = S_{cr} S_{rr}^{-1} B_r^t - B_c^t \left(= F_{I_{rc}}^t \right),\\
&F_{I_{cc}} = S_{cc} - S_{cr} S_{rr}^{-1} S_{rc},\\
&d_r = B_r S_{rr}^{-1} g_r,\\
&d_c = g_c - S_{cr} S_{rr}^{-1} g_r.
\end{aligned}
\tag{3.4}
$$

Then (λ, w_c) satisfies

$$\begin{pmatrix} F_{I_{rr}} & F_{I_{rc}} \\ F_{I_{cr}} & -F_{I_{cc}} \end{pmatrix} \begin{pmatrix} \lambda \\ w_c \end{pmatrix} = \begin{pmatrix} d_r \\ -d_c \end{pmatrix}.$$

Eliminating w_c in the above equation, we obtain

$$\left(F_{I_{rr}} + F_{I_{rc}} F_{I_{cc}}^{-1} F_{I_{cr}} \right) \lambda = d_r - F_{I_{rc}} F_{I_{cc}}^{-1} d_c.$$

Here, $F_{DP} = F_{I_{rr}} + F_{I_{rc}} F_{I_{cc}}^{-1} F_{I_{cr}}$ is called the FETI-DP operator for the problem (2.1).

3.2. Preconditioner. From now on, we will propose $\widehat{F}_{DP}^{-1}$, a preconditioner for F_{DP}, which is derived from the dual norm on the Lagrange multiplier space M in the following sense:

$$\langle \widehat{F}_{DP} \lambda, \lambda \rangle = \|\lambda\|_{(W^0)'}^2. \tag{3.5}$$

We will derive the matrix form of $\widehat{F}_{DP}$ from the above relation. Define $E_{ij}^i : W_{ij}^0 \to W_i$ as the extension operator by 0. Then we get

$$\widetilde{w}_{ij} = E_{ij}^i w_{ij} \quad \text{for} \quad w_{ij} \in W_{ij}^0.$$

Define $R_{ij} : W^0 \to W^0_{ij}$ to be a restriction operator. For $w \in W^0$, we let $w_{ij} = R_{ij}w$. Hence by (2.8) and (2.3), we get

$$\|w\|^2_{W^0} = \sum_{i=1}^{N} \left\langle S^i \left(\sum_{j\in m_i} E^i_{ij} R_{ij} w \right), \sum_{j\in m_i} E^i_{ij} R_{ij} w \right\rangle.$$

Let $E^i = \sum_{j\in m_i} E^i_{ij} R_{ij}$; then we have

$$\|w\|^2_{W^0} = \langle \widehat{S} w, w\rangle \quad \text{with} \quad \widehat{S} = \sum_{i=1}^{N} (E^i)^t S^i E^i. \tag{3.6}$$

Recall that

$$(B^{ij}_i)_{lk} = \int_{\Gamma_{ij}} \xi^{ij}_l \phi^{ij}_k \, ds, \ l = 1, \dots, N_{ij}, \ k = 0, 1, \dots, N_{ij} + 1,$$

and take $(B^{ij}_{i,r})_{lk} = (B^{ij}_i)_{lk}$ for $l, k = 1, \dots, N_{ij}$. Since $w_{ij} \in W^0_{ij}$, we have

$$\lambda^t_{ij} B^{ij}_{i,r} w_{ij} = \int_{\Gamma_{ij}} \lambda_{ij} w_{ij} \, ds.$$

Let

$$\widehat{B} = \mathrm{diag}_{i=1,\dots,N} \left(\mathrm{diag}_{j\in m_i} \left(B^{ij}_{i,r} \right) \right).$$

Then, for $(w, \lambda) \in W^0 \times M$, we get

$$\lambda^t \widehat{B} w = \sum_{i=1}^{N} \sum_{j\in m_i} \int_{\Gamma_{ij}} \lambda_{ij} w_{ij} \, ds, \tag{3.7}$$

where $\lambda_{ij} = \lambda|_{\Gamma_{ij}}$ and $w_{ij} = w|_{\Gamma_{ij}}$.

From the definition of the dual norm (2.10), (2.9), (3.7), and (3.6), we obtain

$$\|\lambda\|^2_{(W^0)'} = \max_{w\in W^0\setminus\{0\}} \frac{\langle \lambda, \widehat{B} w\rangle^2}{\langle \widehat{S} w, w\rangle}.$$

Since $\widehat{S}$ is symmetric and positive definite on W^0, in the above equation the maximum occurs when $\widehat{B}^t \lambda = \widehat{S} w$. Therefore, we have

$$\|\lambda\|^2_{(W^0)'} = \langle \widehat{B} \widehat{S}^{-1} \widehat{B}^t \lambda, \lambda\rangle$$

and let $\widehat{F}_{DP} = \widehat{B} \widehat{S}^{-1} \widehat{B}^t$. Then we take $\widehat{F}^{-1}_{DP} = (\widehat{B} \widehat{S}^{-1} \widehat{B}^t)^{-1}$ as a preconditioner for F_{DP} and call it a Neumann–Dirichlet preconditioner.

Note that $\widehat{F}^{-1}_{DP} = (\widehat{B}^t)^{-1} \widehat{S} \widehat{B}^{-1}$ is easy to implement due to the block diagonal structure of $\widehat{B}$ and $\widehat{B}^t = \widehat{B}$. Therefore, we have

$$\widehat{F}^{-1}_{DP} = \sum_{i=1}^{N} \left(\sum_{j\in m_i} R^t_{ij} (B^{ij}_{i,r})^{-1} (E^i_{ij})^t \right) S^i \left(\sum_{j\in m_i} E^i_{ij} (B^{ij}_{i,r})^{-1} R_{ij} \right) \tag{3.8}$$

so that the work can be done in parallel in each subdomain. Let

$$\widehat{B}_i = \sum_{j \in m_i} R_{ij}^t (B_{i,r}^{ij})^{-1} (E_{ij}^i)^t.$$

Moreover, from the operator $\widehat{B}_i$, we can see that the preconditioner $\widehat{F}_{DP}^{-1}$ is different from the preconditioners in [4, 5, 8, 9, 10]. Only on the slave sides of interfaces are the function values transferred between the spaces W_i and M. Hence, the cost needed to compute $\widehat{B}_i w_i$ and $\widehat{B}_i^t \lambda$ is reduced by half compared with other FETI (-DP) preconditioners.

4. Condition number estimation for the preconditioned FETI-DP operator. The following well-known result is given when $a_i(u,v) = \int_{\Omega_i} \nabla u \cdot \nabla v \, dx$ (see Theorem 4.1.3 in [11]). With slight modification, we can obtain a similar result for a general case.

LEMMA 4.1. *For $w_i \in W_i$, we have*

$$C_1 |w_i|^2_{1/2,\partial\Omega_i} \le \langle S^i w_i, w_i \rangle \le C_2 \|w_i\|^2_{1/2,\partial\Omega_i},$$

where C_1 and C_2 are constants depending on $A(x)$ and $\beta(x)$, but independent of H_i and h_i.

In the following, we obtain a formula that is useful for analyzing the condition number bound and the result is the same as Lemma 4.3 of Mandel and Tezaur [10]. However, in our formulation, the continuity constraints are imposed on $w \in W$, that is, the d.o.f. on edges and global corners (see (3.3)). The proof can be done similarly as in Lemma 37 of Tezaur [16].

LEMMA 4.2. *For $\lambda \in M$, we have*

$$\max_{w \in W \setminus \{0\}} \frac{b(w,\lambda)^2}{\|w\|_W^2} = \langle F_{DP} \lambda, \lambda \rangle.$$

Now, we estimate the lower bound of the condition number for the operator $\widehat{F}_{DP}^{-1} F_{DP}$.

LEMMA 4.3. *For any $\lambda \in M$, we have*

$$\max_{w \in W \setminus \{0\}} \frac{b(w,\lambda)^2}{\|w\|_W^2} \ge \|\lambda\|_{(W^0)'}.$$

Proof. For $w \in W^0$, let $\widetilde{w}_i = \sum_{j \in m_i} \widetilde{w}_{ij}$ and $\widetilde{w} = (\widetilde{w}_1, \dots, \widetilde{w}_N)$. Then we have $\widetilde{w} \in W$. Hence it follows that

$$\max_{w \in W \setminus \{0\}} \frac{b(w,\lambda)^2}{\|w\|_W^2} \ge \max_{w \in W^0 \setminus \{0\}} \frac{b(\widetilde{w},\lambda)^2}{\|\widetilde{w}\|_W^2}. \tag{4.1}$$

Since $\widetilde{w}_j = 0$ on Γ_{ij} for $j \in m_i$, we have

$$b(\widetilde{w},\lambda) = \sum_{i=1}^N \sum_{j \in m_i} \int_{\Gamma_{ij}} w_{ij} \lambda_{ij} \, ds = \langle \lambda, w \rangle_m. \tag{4.2}$$

Combining (4.2), (2.8), and (2.10), we obtain

$$\max_{w \in W^0 \setminus \{0\}} \frac{b(\widetilde{w},\lambda)^2}{\|\widetilde{w}\|_W^2} = \max_{w \in W^0 \setminus \{0\}} \frac{\langle \lambda, w \rangle_m^2}{\|w\|_{W^0}^2} = \|\lambda\|^2_{(W^0)'}. \tag{4.3}$$

From (4.1) and (4.3), we complete the proof. ∎

To estimate the upper bound for $\langle F_{DP}\lambda, \lambda\rangle$, we need the following estimate for $\|w_i - w_j\|^2_{H_{00}^{1/2}(\Gamma_{ij})}$.

LEMMA 4.4. *For $w \in W$, let $w_i = w|_{\partial\Omega_i}$ and $w_j = w|_{\partial\Omega_j}$. Then we have*

$$\|w_i - w_j\|^2_{H_{00}^{1/2}(\Gamma_{ij})} \le C \max_{l\in\{i,j\}} \left\{ \left(1 + \log \frac{H_l}{h_l}\right)^2 \right\} \left(|w_i|^2_{1/2,\partial\Omega_i} + |w_j|^2_{1/2,\partial\Omega_j} \right),$$

where C is a constant independent of h_i's and H_i's.

Proof. Let $I^H w$ be a linear function on Γ_{ij} that has the same value as w at the end points of Γ_{ij}. From Lemma 5.1 in [10], we have

$$|w_l - I^H w_l|_{H_{00}^{1/2}(\Gamma_{ij})} \le C \left(1 + \log \frac{H_l}{h_l}\right) |w_l|_{1/2,\partial\Omega_l} \text{ for } l = i, j.$$

Using the above bound and the equivalence of $|\cdot|_{H_{00}^{1/2}(\Gamma_{ij})}$ and $\|\cdot\|_{H_{00}^{1/2}(\Gamma_{ij})}$, the result follows. ∎

DEFINITION 4.5. *We define a projection $\pi_{ij} : H_{00}^{1/2}(\Gamma_{ij}) \to W_{ij}^0$ for $v \in H_{00}^{1/2}(\Gamma_{ij})$ by*

$$\int_{\Gamma_{ij}} (v - \pi_{ij} v)\lambda_{ij}\, ds \ = 0 \ \ \forall \lambda_{ij} \in M_{ij}.$$

From Lemma 2.2 in [1], π_{ij} is a continuous operator on $H_{00}^{1/2}(\Gamma_{ij})$, i.e., there exists a constant C such that

$$\|\pi_{ij} v\|_{H_{00}^{1/2}(\Gamma_{ij})} \le C \|v\|_{H_{00}^{1/2}(\Gamma_{ij})} \ \ \forall v \in H_{00}^{1/2}(\Gamma_{ij}). \tag{4.4}$$

We note that the constant C is independent of H_i's and h_i's.

Now, we estimate the upper bound for the operator $\widehat{F}_{DP}^{-1} F_{DP}$.

LEMMA 4.6. *For $\lambda \in M$, we have*

$$\max_{w\in W\setminus\{0\}} \frac{b(w,\lambda)^2}{\|w\|^2_W} \le C \max_{i=1,\dots,N} \left\{ \left(1 + \log \frac{H_i}{h_i}\right)^2 \right\} \|\lambda\|^2_{(W^0)'},$$

where C is a constant depending on $A(x)$ and $\beta(x)$, but independent of h_i's and H_i's.

Proof. From the definitions of $b(w,\lambda)$ in (2.7) and π_{ij}, we have

$$b(w,\lambda)^2 = \left(\sum_{i=1}^N \sum_{j\in m_i} \int_{\Gamma_{ij}} \pi_{ij}(w_i - w_j)\lambda_{ij}\, ds \right)^2.$$

We let $z \in W^0$ such that $z|_{\Gamma_{ij}} = \pi_{ij}(w_i - w_j)$. Then the above equation is the duality pairing between λ and z. Hence, using the definition of dual norm on λ, we get

$$b(w,\lambda)^2 \le \|\lambda\|^2_{(W^0)'} \|z\|^2_{W^0}. \tag{4.5}$$

Let $\widetilde{z} = (\widetilde{z}_1, \dots, \widetilde{z}_N) \in W$ be the zero extension of z. Then from (2.8), (2.3), Lemma 4.1, (2.5), (4.4), and Lemma 4.4,

$$
\begin{aligned}
\|z\|_{W^0}^2 &= \sum_{i=1}^{N} \langle S^i \widetilde{z}_i, \widetilde{z}_i \rangle \\
&\le C \sum_{i=1}^{N} \|\widetilde{z}_i\|_{1/2,\partial\Omega_i}^2 \\
&\le C \sum_{i=1}^{N} \sum_{j \in m_i} \|\pi_{ij}(w_i - w_j)\|_{H_{00}^{1/2}(\Gamma_{ij})}^2 \\
&\le C \sum_{i=1}^{N} \sum_{j \in m_i} \|w_i - w_j\|_{H_{00}^{1/2}(\Gamma_{ij})}^2 \\
&\le C \max_{i=1,\cdots,N} \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\} \sum_{i=1}^{N} |w_i|_{1/2,\partial\Omega_i}^2 \\
&\le C \max_{i=1,\cdots,N} \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\} \|w\|_W^2.
\end{aligned}
\tag{4.6}
$$

Here, C denotes a generic constant independent of h_i's and H_i's, which may vary from occurrence to occurrence. Combining (4.5) and (4.6), we complete the proof. □

Since the preconditioner $\widehat{F}_{DP}^{-1}$ follows from the dual norm of $\lambda \in M$ (see (3.5)), combining Lemmas 4.2, 4.3, and 4.6, we obtain the following estimate.

THEOREM 4.7. *For $\lambda \in M$, we have*

$$
\langle \widehat{F}_{DP}\lambda, \lambda \rangle \le \langle F_{DP}\lambda, \lambda \rangle \le C \max_{i=1,\dots,N} \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\} \langle \widehat{F}_{DP}\lambda, \lambda \rangle,
$$

where C is a constant depending on $A(x)$ and $\beta(x)$, but independent of H_i's and h_i's.

COROLLARY 4.8. *We have the following condition number estimate:*

$$
\kappa \left(\widehat{F}_{DP}^{-1} F_{DP} \right) \le C \max_{i=1,\dots,N} \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\},
$$

where C is a constant depending on $A(x)$ and $\beta(x)$, but independent of H_i's and h_i's.

Remark 4.1. On each Γ_{ij}, the choices of master and slave sides are arbitrary.

Remark 4.2. In Corollary 4.8, the condition number depends on $A(x)$ and $\beta(x)$. Now we consider the problem

$$
\begin{aligned}
-\nabla \cdot (\alpha(x) \nabla u(x)) &= f(x) \ \text{ in } \ \Omega, \\
u &= 0 \ \text{ on } \ \partial\Omega,
\end{aligned}
$$

where $\alpha(x)$ is a piecewise constant and has jumps across the subdomain boundaries, i.e., $\alpha(x) = \rho_i$ for all $x \in \Omega_i$ for some constant $\rho_i > 0$. On Γ_{ij}, we choose $\Omega_i^h|_{\Gamma_{ij}}$ as the slave side if $\rho_i \le \rho_j$. Otherwise, we choose $\Omega_i^h|_{\Gamma_{ij}}$ as the master side. Then we have

$$
C_1 \rho_i |w_i|_{1/2,\partial\Omega_i}^2 \le \langle S^i w_i, w_i \rangle \le C_2 \rho_i \|w_i\|_{1/2,\partial\Omega_i}^2,
$$

where C_1 and C_2 are constants independent of ρ_i's, h_i's, and H_i's. Following the proof of Lemma 4.6 and using the above inequalities instead of Lemma 4.1, we obtain

$$\begin{aligned}
\|z\|^2_{W^0} &\le C\sum_{i=1}^{N}\sum_{j\in m_i}\rho_i\|w_i-w_j\|^2_{H^{1/2}_{00}(\Gamma_{ij})}\\
&\le C\sum_{i=1}^{N}\sum_{j\in m_i}\left\{\max_{l\in\{i,j\}}\left\{\left(1+\log\frac{H_l}{h_l}\right)^2\right\}\right.\\
&\qquad\left.\times\left(\rho_i|w_i|^2_{1/2,\partial\Omega_i}+\rho_i|w_j|^2_{1/2,\partial\Omega_j}\right)\right\}\\
&\le C\sum_{i=1}^{N}\sum_{j\in m_i}\left\{\max_{l\in\{i,j\}}\left\{\left(1+\log\frac{H_l}{h_l}\right)^2\right\}\right.\\
&\qquad\left.\times\left(\langle S^i w_i,w_i\rangle+\frac{\rho_i}{\rho_j}\langle S^j w_j,w_j\rangle\right)\right\},
\end{aligned}$$

where C is a generic constant independent of ρ_i's, H_i's, and h_i's. Since $\rho_i \le \rho_j$, we can see that the constant C in Lemma 4.6 is bounded independently of the coefficients. Hence, the condition number bound is independent of ρ_i's.

5. Numerical results. In this section, we provide numerical tests for the FETI-DP formulation developed in this paper. Let $\Omega = [0,1]\times[0,1] \subset \mathbb{R}^2$ and consider the following model problem:

$$\begin{aligned}
-\nabla\cdot(\alpha(x,y)\nabla u) &= f \ \text{ in } \ \Omega,\\
u &= 0 \ \text{ on } \ \partial\Omega.
\end{aligned}\tag{5.1}$$

We compare the proposed preconditioner (3.8) with the preconditioner of Dryja and Widlund [5] for the cases when $\alpha(x,y)=1$ and mesh sizes are comparable between neighboring subdomains, and when $\alpha(x,y)$ are highly discontinuous across subdomain interfaces and mesh sizes are not comparable. In the following, we use the notation $\widehat{F}^{-1}_{KL}$ for the preconditioner (3.8) and use $\widehat{F}^{-1}_{DW}$ for Dryja and Widlund's. The preconditioner $\widehat{F}^{-1}_{DW}$ is

$$\widehat{F}^{-1}_{DW} = (B_r\widetilde{B}^t_r)^{-1}\widetilde{B}_r S_{rr}\widetilde{B}^t_r(\widetilde{B}_r B^t_r)^{-1},$$

where $\widetilde{B}_r$ is the scaled matrix of B_r divided by the mesh parameters of slave and master sides (see (3.13) in [5]).

First, we compare these two preconditioners for the same problem with nonmatching discretizations. We take $\alpha(x,y)=1$ and the exact solution $u(x,y) = y(1-y)\sin\pi x$. The CG iteration continues until the relative residual norm is less than 10^{-6}. We use n to denote the number of nodes on edges, including end points, and use N to denote the number of subdomains. In this problem, we use the same n for all subdomains, divide Ω into rectangular subdomains, as in Figure 3, and denote each subdomain by Ω_{ij}.

To make nonmatching grids across subdomain interfaces, we generate triangulations in each subdomain in the following way: For each subdomain, we have chosen n random quasi-uniform nodes on each horizontal and vertical edge. Using these

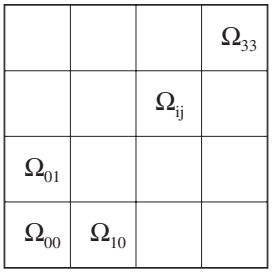


FIG. 3. *Partition of subdomains when* $N = 4 \times 4$.

TABLE 1

Comparison between the proposed preconditioner F_{KL}^{-1} *and the Dryja–Widlund preconditioner* F_{DW}^{-1} *on nonmatching grids when n increases with* $N = 4 \times 4$*: Iter (number of CG iteration), Cond (condition number of the preconditioned FETI-DP operator).*

$n-1$	L^2-error	H^1-error	$\widehat{F}_{KL}^{-1}$		$\widehat{F}_{DW}^{-1}$	
			Iter	Cond	Iter	Cond
4	5.0850e-4	6.0126e-2	10	3.07	7	1.94
8	1.2865e-4	3.0128e-2	13	5.67	8	2.68
16	3.2235e-5	1.5072e-2	15	7.68	10	3.69
32	8.0627e-6	7.5374e-3	16	9.99	10	4.80
64	2.0163e-6	3.7688e-3	17	12.6	11	6.14

nodes, we generate nonuniform structured grids in each subdomain. Since we use the same n for all subdomains, the sizes of meshes between neighboring subdomains are comparable.

In Table 1, we divide Ω into $N = 4 \times 4$ subdomains (see Figure 3), increase the number of nodes n, and compute L^2- and H^1-errors, the number of CG iterations and condition numbers for those preconditioners. For the H^1-error, we compute the broken H^1-norm of errors over all subdomains. Table 2 shows the numerical results when we fix $n - 1 = 4$ and increase the number of subdomains N. For the cases $N = 8 \times 8$, 16×16, and 32×32, we divide Ω into subdomains in the same manner as $N = 4 \times 4$. Here, we used the FETI-DP formulation developed in this paper. From Tables 1 and 2, we can see that our FETI-DP formulation gives $O(h^2)$ and $O(h)$ convergences for L^2- and H^1-errors, respectively. Furthermore, we can see that both preconditioners seem to give the $\log^2$-growth of the condition number bound and that the CG iteration of $\widehat{F}_{DW}^{-1}$ is smaller than $\widehat{F}_{KL}^{-1}$.

Now, we consider the problem (5.1) when $\alpha(x, y)$ is highly discontinuous across subdomain interfaces and the mesh sizes between subdomains are not comparable. In this situation, we will compare two preconditioners $\widehat{F}_{KL}^{-1}$ and $\widehat{F}_{DW}^{-1}$.

We consider the cases of $N = 2 \times 2$, 4×4, 8×8 subdomains. For each subdomain Ω_{ij}, we choose the coefficient $\alpha(x, y)$ in the following way:

$$\alpha(x,y) = \begin{cases} 1 & \text{if both } i \text{ and } j \text{ are even,} \\ 250 & \text{if } i \text{ is odd and } j \text{ is even,} \\ 5000 & \text{if } i \text{ is even and } j \text{ is odd,} \\ 10 & \text{if both } i \text{ and } j \text{ are odd,} \end{cases}$$

and denote them by ρ_{ij}. In addition, we consider the exact solution $u(x, y)$, which

TABLE 2

Comparison between the proposed preconditioner F_{KL}^{-1} and the Dryja–Widlund preconditioner F_{DW}^{-1} on nonmatching grids when N increases with $n-1=4$: Iter (number of CG iteration), Cond (condition number of the preconditioned FETI-DP operator).

$N \times N$	L^2-error	H^1-error	$\hat{F}_{KL}^{-1}$		$\hat{F}_{DW}^{-1}$	
			Iter	Cond	Iter	Cond
4×4	5.0850e-4	6.0126e-2	10	3.07	7	1.94
8×8	1.1744e-4	2.9900e-2	11	3.22	8	2.13
16×16	2.9743e-5	1.4980e-2	12	3.39	8	2.11
32×32	7.4318e-6	7.4917e-3	12	3.51	8	2.10

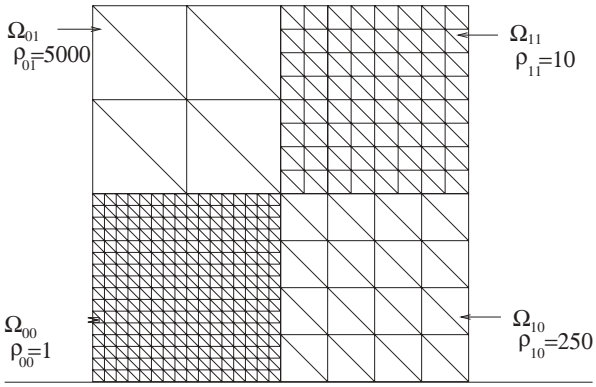


FIG. 4. *Triangulations for the case $N = 2 \times 2$ and $\max(H_{ij}/h_{ij}) = 16$.*

belongs to $H^1(\Omega)$, according to the partition of the domain:

$$u(x,y) = \begin{cases} p_1(x,y)\sin(\pi x)\sin(\pi y)/\alpha(x,y) & \text{when } N = 2 \times 2, \\ p_2(x,y)\sin(2\pi x)\sin(2\pi y)/\alpha(x,y) & \text{when } N = 4 \times 4, \\ \sin(8\pi x)\sin(8\pi y)/\alpha(x,y) & \text{when } N = 8 \times 8, \end{cases}$$

where

$$p_1(x,y) = (x-1/2)(y-1/2),$$
$$p_2(x,y) = (x-1/4)(x-3/4)(y-1/4)(y-3/4).$$

Following [18, section 1.5.3], we have chosen a different mesh size in each subdomain according to the ratio of coefficients between neighboring subdomains, that is,

$$\frac{h_{ij}}{h_{kl}} \simeq \sqrt[4]{\frac{\rho_{ij}}{\rho_{kl}}},$$

where h_{ij} is the mesh size of the subdomain Ω_{ij}, and we use H_{ij} to denote the size of the subdomain Ω_{ij}. Using the mesh sizes of these ratios, we divide each subdomain into uniform meshes. When $N = 2 \times 2$ and $\max(H_{ij}/h_{ij}) = 16$, we

TABLE 3
Comparison between the proposed preconditioner $\hat{F}_{KL}^{-1}$ and the Dryja–Widlund preconditioner $\hat{F}_{DW}^{-1}$ for the problem of highly discontinuous coefficients: Iter (number of GC iteration).

N	$\max(H_{ij}/h_{ij})$	L^2-error	H^1-error	$\hat{F}_{KL}^{-1}$	$\hat{F}_{DW}^{-1}$
				Iter	Iter
2×2	16	3.0571e-5	7.6362e-3	3	17
	32	7.8276e-6	3.8249e-3	3	26
	64	1.9747e-6	1.9133e-3	4	39
	128	4.9571e-7	9.5675e-4	4	50
	256	1.2421e-7	4.7839e-4	4	60
4×4	16	2.1574e-6	1.0939e-3	4	75
	32	5.4460e-7	5.4805e-4	4	81
	64	1.3799e-7	2.7415e-4	4	111
	128	3.4810e-8	1.3709e-4	4	130
8×8	16	1.0262e-3	8.8753e-1	3	113
	32	2.4870e-4	4.4462e-1	4	136
	64	6.4579e-5	2.2240e-1	4	168

obtain triangulations as in Figure 4 and the triangulations are not comparable between neighboring subdomains.

In section 1.5.3 of [18], it was shown that a good approximation of the solution is obtained when the slave side is chosen to give a Lagrange multiplier space of higher dimension. Hence, choosing the subdomain with smaller h_{ij} (smaller ρ_{ij}) as the slave side, we can approximate the exact solution more accurately. This observation coincides with the choices of master and slave sides in Remark 4.2.

Table 3 shows L^2- and H^1-errors and CG iterations with $\widehat{F}_{KL}^{-1}$ and $\widehat{F}_{DW}^{-1}$ as preconditioners. In CG iteration, we use the same stopping criterion 10^{-6} as before. Increasing $\max(H_{ij}/h_{ij})$, we observe the $O(h^2)$ and $O(h)$ convergences of L^2- and H^1-errors, respectively, for all cases of N. Furthermore, we see that the CG iteration of $\widehat{F}_{KL}^{-1}$ is much smaller than $\widehat{F}_{DW}^{-1}$. Since, the condition number bound of the preconditioner $\widehat{F}_{DW}^{-1}$ depends on the ratio of meshes between neighboring subdomains, the preconditioner works inefficiently for these problems with noncomparable meshes.

From our numerical results, we conclude that our formulation gives the correct approximation of the model problem with nonmatching grids. For the case of continuous coefficients and comparable meshes between subdomain interfaces, the preconditioner $\widehat{F}_{DW}^{-1}$ by Dryja and Widlund gives a smaller number of iterations than our preconditioner $\widehat{F}_{KL}^{-1}$. However, our preconditioner $\widehat{F}_{KL}^{-1}$ turns out to be much more efficient than $\widehat{F}_{DW}^{-1}$ for the problem of highly discontinuous coefficients on noncomparable meshes.

REFERENCES

[1] F. B. BELGACEM, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[2] C. BERNARDI, Y. MADAY, AND A. T. PATERA, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. XI (Paris, 1989–1991), Pitman Res. Notes Math. Ser. 299, Longman Scientific and Technical, Harlow, 1994, pp. 13–51.

[3] B. COCKBURN AND C.-W. SHU, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[4] M. DRYJA AND O. B. WIDLUND, *A FETI-DP method for a mortar discretization of elliptic problems*, in Recent Developments in Domain Decomposition Methods (Zürich, 2001),

Lecture Notes in Comput. Sci. Engrg. 23, Springer, Berlin, 2002, pp. 41–52.

[5] M. Dryja and O. B. Widlund, *A generalized FETI-DP method for a mortar discretization of elliptic problems*, in Domain Decomposition Methods in Science and Engineering (Cocoyoc, Mexico, 2002), UNAM, Mexico City, 2003, pp. 27–38.

[6] C. Farhat, *A saddle-point principle domain decomposition method for the solution of solid mechanics problems*, in Proceedings of the 5th International Symposium on Domain Decomposition Methods for Partial Differential Equations (Norfolk, VA, 1991), D. E. Keyes, T. F. Chan, G. Meurant, J. S. Scroggs, and R. G. Voigt, eds., SIAM, Philadelphia, PA, 1992, pp. 271–292.

[7] C. Farhat, M. Lesoinne, and K. Pierson, *A scalable dual-primal domain decomposition method*, Numer. Linear Algebra Appl., 7 (2000), pp. 687–714.

[8] C. Farhat and F.-X. Roux, *A method of finite element tearing and interconnecting and its parallel solution algorithm*, Int. J. Numer. Methods Engrg., 32 (1991), pp. 1205–1227.

[9] A. Klawonn and O. B. Widlund, *FETI and Neumann-Neumann iterative substructuring methods: Connections and new results*, Comm. Pure Appl. Math., 54 (2001), pp. 57–90.

[10] J. Mandel and R. Tezaur, *On the convergence of a dual-primal substructuring method*, Numer. Math., 88 (2001), pp. 543–558.

[11] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Numerical Mathematics and Scientific Computation, The Clarendon Press, Oxford University Press, New York, 1999.

[12] F. Rapetti and A. Toselli, *A FETI preconditioner for two dimensional edge element approximations of Maxwell's equations on nonmatching grids*, SIAM J. Sci. Comput., 23 (2001), pp. 92–108.

[13] D. J. Rixen, *Extended preconditioners for the FETI method applied to constrained problems*, Int. J. Numer. Methods Engrg., 54 (2002), pp. 1–26.

[14] D. Stefanica, *A numerical study of FETI algorithms for mortar finite element methods*, SIAM J. Sci. Comput., 23 (2001), pp. 1135–1160.

[15] D. Stefanica, *FETI and FETI-DP methods for spectral and mortar spectral elements: A performance comparison*, J. Sci. Comput., 17, (2002), pp. 629–638.

[16] R. Tezaur, *Analysis of Lagrange Multiplier Based Domain Decompoistion*, Ph.D. thesis, University of Colorado at Denver, 1998.

[17] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

[18] B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Lecture Notes in Comput. Sci. Engrg. 17, Springer-Verlag, Berlin, 2001.

[19] J. Xu and J. Zou, *Some nonoverlapping domain decomposition methods*, SIAM Rev., 40 (1998), pp. 857–914.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2176–2187

THE ACCURACY OF THE CHEBYSHEV DIFFERENCING METHOD FOR ANALYTIC FUNCTIONS*

S. C. REDDY† AND J. A. C. WEIDEMAN‡

Abstract. The Chebyshev spectral collocation method is one of the most powerful tools for numerical differentiation, particularly when the function under consideration is smooth. An upper bound on the error, in the discrete maximum norm, is derived here when the function is analytic. Two model functions are analyzed in detail.

Key words. pseudospectral, spectral collocation, numerical differentiation, Chebyshev methods, interpolation

AMS subject classifications. 65D25, 65D05

DOI. 10.1137/040603280

1. Introduction. Let $f(x)$ be a differentiable function defined on a bounded interval, say $[-1,1]$. In this interval, choose a discrete grid $\{x_j\}$ consisting of $N+1$ distinct points. How does one compute approximations to the derivative values $\{f'(x_j)\}$? One of the most powerful procedures for calculating this is also conceptually the simplest: Fit a polynomial of degree N, say $p_N(x)$, through the data values $\{(x_j, f(x_j))\}$. Then differentiate this polynomial (analytically) and evaluate the derivative at the gridpoints, i.e., $p'_N(x_j) \approx f'(x_j)$.

Approximation theory dictates that $\{x_j\}$ should not be just any set of points. Recall, e.g., the Runge phenomenon associated with an equidistant point distribution: those wild oscillations near the endpoints of the interval when a seemingly innocuous function such as $f(x) = 1/(1+25x^2)$ is interpolated on $[-1,1]$; see, for example, [16, p. 44]. Good sets of points are the roots or extrema of the Chebyshev polynomials. First, these points have precisely the right density distribution for avoiding the Runge phenomenon [16, p. 45]. Second, interpolants based on these sets of points are to within a relatively small factor of the best min-max polynomial approximation of $f(x)$ on $[-1,1]$ [11, p. 160]. Third, these interpolants may be evaluated quickly and efficiently with the aid of the fast Fourier transform (FFT) [11, p. 91].

The differentiation procedure described above is at the heart of what has become known as the Chebyshev pseudospectral method or, equivalently, the Chebyshev spectral collocation process. Since its introduction in the 1970s and early 1980s, this idea has been used in practice to solve a variety of differential equations; case studies are cited in the survey papers [7], [9], [17] and the monographs [3], [4], [6], [8], [10], [12], [16].

The Chebyshev differencing method works best when the function is smooth, and particularly when $f(x)$ can be continued into the complex plane as a function $f(z)$

*Received by the editors January 16, 2004; accepted for publication (in revised form) June 22, 2004; published electronically March 25, 2005.
http://www.siam.org/journals/sinum/42-5/60328.html
†Quadrus Financial Technologies Inc., 15th Floor, 999 West Hastings St., Vancouver, BC, Canada V6C 2W2 (satishr@quadrusfinancial.com).
‡Department of Applied Mathematics, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa (weideman@dip.sun.ac.za). The research of this author was supported by National Research Foundation in South Africa grant NRF5289.

which is analytic in an open neighborhood of $[-1,1]$. In this case, the error

$$E_N(f) = \max_{0\le j\le N} |f'(x_j) - p'_N(x_j)|$$

decays at least exponentially fast as $N \to \infty$, a fact mentioned (but not proved) in the monographs and survey papers listed above. To see a proof, we had to go to the research literature, and even then we could find only a single reference, namely that of Tadmor [15].

Tadmor gives very general results, including Sobolev estimates for the error when one assumes less regularity than analyticity. To get these estimates, Tadmor works in the space of Chebyshev coefficients. We present a fundamentally different approach here, which is based in the function space and, therefore, is more in the "spirit" of the pseudospectral (collocation) process. In the process, the error estimate in [15] is improved by a factor of $O(N^{3/2})$. On the negative side, our approach is strictly limited to functions that are analytic in some open neighborhood of $[-1,1]$.

2. General error bounds. Let $T_N(x)$ denote the Chebyshev polynomial of degree N. The differentiation process, described in the first paragraph of section 1, is typically implemented on the set of zeros $\{s_j\}$ of $T_{N+1}(x)$ or on the set of extrema $\{t_j\}$ of $T_N(x)$, including the endpoint extrema at $x = \pm 1$. That is, we define for $j = 0, 1, \ldots, N$,

$$\textit{Zeros:} \qquad s_j = \cos\left(\frac{(2j+1)\pi}{2N+2}\right), \tag{2.1}$$

$$\textit{Extrema:} \qquad t_j = \cos\left(\frac{j\pi}{N}\right). \tag{2.2}$$

Sometimes referred to as the Chebyshev points of the second kind, the second set of nodes is often preferable since it includes the endpoints $x = \pm 1$. In the solution of differential equations, this enables one to incorporate boundary conditions.

These two sets of points share the attractive feature that the derivative of the interpolant, $p'_N(x)$, can be evaluated at the $N+1$ nodes $\{s_j\}$ (resp., $\{t_j\}$) via the FFT in only $O(N \log N)$ operations. This stands in contrast to the direct implementation that requires $O(N^2)$ operations. For details, see [4], [6], [16], [17].

We record the following preliminary facts. Assuming $N+1$ distinct nodes $\{x_j\}$ in $[-1,1]$, the error in polynomial interpolation of the data $\{(x_j, f(x_j))\}$ is given by Hermite's contour integral [5, p. 68]

$$f(x) - p_N(x) = \frac{1}{2\pi i}\int_\Gamma \frac{\omega_{N+1}(x)}{\omega_{N+1}(z)}\,\frac{f(z)}{(z-x)}\,dz. \tag{2.3}$$

Here Γ is a simple closed positively oriented contour in the complex plane that (a) contains the point x and encloses $[-1,1]$, and (b) lies in some simply connected region in which $f(z)$ is analytic. The polynomial $\omega_{N+1}(x)$ is defined as

$$\omega_{N+1}(x) = c\prod_{j=0}^{N}(x - x_j),$$

where the normalization factor c is immaterial, as it cancels in the Hermite formula. For the two sets of interpolation nodes (2.1) and (2.2), we use, respectively,

$$\textit{Zeros:} \qquad \omega_{N+1}(x) = T_{N+1}(x), \tag{2.4}$$

$$\textit{Extrema:} \qquad \omega_{N+1}(x) = T_{N+1}(x) - T_{N-1}(x). \tag{2.5}$$

In the latter case the expression $\omega_{N+1}(x) = (1-x^2)T_N'(x)$ is commonly used in the literature but, owing to the identity [11, p. 35]

$$(1-x^2)T_N'(x) = \frac{1}{2}N(T_{N-1}(x) - T_{N+1}(x)),$$

we may use expression (2.5) instead, which is more useful for our purposes here.

By direct differentiation of the Hermite formula (2.3), one obtains an integral formula for the pointwise error

$$f'(x) - p_N'(x) = \frac{1}{2\pi i}\int_\Gamma \left(\frac{\omega_{N+1}'(x)}{(z-x)} + \frac{\omega_{N+1}(x)}{(z-x)^2}\right)\frac{f(z)}{\omega_{N+1}(z)}\,dz. \tag{2.6}$$

Evaluating this at $x = x_j$ and using the fact that $\omega_{N+1}(x_j) = 0$, one gets

$$f'(x_j) - p_N'(x_j) = \frac{1}{2\pi i}\int_\Gamma \frac{\omega_{N+1}'(x_j)}{\omega_{N+1}(z)}\frac{f(z)}{(z-x_j)}\,dz, \qquad j = 0,\ldots,N. \tag{2.7}$$

An appropriate contour Γ for estimating this integral is the ellipse E_ϱ, with foci at ± 1 and the sum of its semimajor and semiminor axes equal to $\varrho\,(>1)$, i.e.,

$$E_\varrho: \quad z = \frac{1}{2}(\varrho\, e^{i\theta} + \varrho^{-1}\, e^{-i\theta}), \qquad 0 \le \theta \le 2\pi. \tag{2.8}$$

The reason for choosing this particular contour is that the factor $\omega_{N+1}(z)$ appearing in the integrand of (2.3) and (2.7) is of nearly constant magnitude on the ellipse E_ϱ when N is large.

To show this, we follow the analysis of [11, p. 15]. The starting point is the following representation formula for the Chebyshev polynomials,

$$T_N(z) = \frac{1}{2}\left[\left(z+\sqrt{z^2-1}\right)^N + \left(z-\sqrt{z^2-1}\right)^N\right],$$

which implies that if $z \in E_\varrho$, then

$$\begin{aligned} T_N(z) &= \frac{1}{2}\left[(\varrho\, e^{i\theta})^N + (\varrho^{-1}\, e^{-i\theta})^N\right] \\ &= \frac{1}{2}\left[(\varrho^N + \varrho^{-N})\cos N\theta + i(\varrho^N - \varrho^{-N})\sin N\theta\right]. \end{aligned} \tag{2.9}$$

Hence

$$|T_N(z)| = \frac{1}{2}\sqrt{\varrho^{2N} + \varrho^{-2N} + 2\cos 2N\theta},$$

and therefore

$$\frac{1}{2}(\varrho^N - \varrho^{-N}) \le |T_N(z)| \le \frac{1}{2}(\varrho^N + \varrho^{-N}), \qquad z \in E_\varrho.$$

By introducing the positive number η through $\varrho = e^\eta$, we see that these inequalities become

$$\sinh(\eta N) \le |T_N(z)| \le \cosh(\eta N), \qquad z \in E_\varrho. \tag{2.10}$$

It is now possible to estimate the quantity $\omega_{N+1}(z)$ that appears in (2.7). In the case of Chebyshev zeros (cf. (2.4)), one simply replaces N by $N+1$ in (2.10) to get

$$\text{(2.11)} \qquad \textit{Zeros:} \qquad \sinh(\eta(N+1)) \;\le\; |\omega_{N+1}(z)| \;\le\; \cosh(\eta(N+1)).$$

In the case of Chebyshev extrema (cf. (2.5)), one proceeds as follows. From (2.9),

$$\omega_{N+1}(z) = T_{N+1}(z) - T_{N-1}(z) = \frac{1}{2}(\varrho\, e^{i\theta} - \varrho^{-1}\, e^{-i\theta})\,(\varrho^N\, e^{iN\theta} - \varrho^{-N}\, e^{-iN\theta})$$

when $z \in E_\varrho$, and therefore

$$|\omega_{N+1}(z)| = \frac{1}{2}\,\sqrt{\varrho^2 + \varrho^{-2} - 2\cos 2\theta}\,\sqrt{\varrho^{2N} + \varrho^{-2N} - 2\cos 2N\theta}.$$

The minimum value of this quantity occurs when $\theta = 0$, which gives the lower bound $\frac{1}{2}(\varrho - \varrho^{-1})(\varrho^N - \varrho^{-N})$. When N is odd, the maximum value occurs at $\theta = \pi/2$, which gives an upper bound $\frac{1}{2}(\varrho + \varrho^{-1})(\varrho^N + \varrho^{-N})$. When N is even, the same upper bound is obtained by maximizing the two terms on the right individually. We have therefore proved that if $z \in E_\varrho$, then

$$\text{(2.12)} \quad \textit{Extrema:} \quad 2\sinh(\eta)\sinh(\eta N) \;\le\; |\omega_{N+1}(z)| \;\le\; 2\cosh(\eta)\cosh(\eta N).$$

Since the upper and lower bounds in both (2.11) and (2.12) approach one another as $N \to \infty$, our assertion that $|\omega_{N+1}(z)|$ is nearly constant on E_ϱ for large N is affirmed.

Bounds on the quantities $\omega'_{N+1}(x_j)$ that appear in (2.7) are also required. By using elementary calculus and standard properties of the Chebyshev polynomials, it is possible to show that for each $N = 1, 2, \ldots$,

$$\text{(2.13)} \qquad \textit{Zeros:} \qquad \max_{0\le j\le N} |\omega'_{N+1}(s_j)| = (N+1)\csc\left(\frac{\pi}{2N+2}\right) \;\le\; (N+1)^2,$$

$$\text{(2.14)} \quad \textit{Extrema:} \qquad \max_{0\le j\le N} |\omega'_{N+1}(t_j)| = 4N.$$

In each case the maximum is achieved at the extreme two nodes s_0 and s_N (resp., t_0 and t_N). The bound on the right of (2.13) is the quantity $\max_{-1\le x\le 1} |\omega'_{N+1}(x)|$, which will be used in Remark 6 below (following Theorem 2.1). No such bound is provided for (2.14), as in fact $\max_{-1\le x\le 1} |\omega'_{N+1}(x)| = 4N$ in the case of extrema.

Finally, we record the following properties of the ellipse E_ϱ. Its length, say L_ϱ, can of course be expressed as an elliptic function, but as a more easily computable bound we use Euler's estimate

$$\text{(2.15)} \qquad L_\varrho \le \pi\sqrt{\varrho^2 + \varrho^{-2}},$$

which overestimates the perimeter by less than 12 percent. (Better bounds for the perimeter of an ellipse can be found in [2].) We shall also need the distance from E_ϱ to the interval $[-1, 1]$, say D_ϱ, which a quick calculation shows to be

$$\text{(2.16)} \qquad D_\varrho = \frac{1}{2}(\varrho + \varrho^{-1}) - 1.$$

The ratio between the perimeter and distance will feature in the error estimate below, and so we define, for $\varrho > 1$,

$$\text{(2.17)} \qquad \phi(\varrho) = \frac{\pi\sqrt{\varrho^2 + \varrho^{-2}}}{\frac{1}{2}(\varrho + \varrho^{-1}) - 1} = 2\pi\frac{\sqrt{\varrho^4 + 1}}{(\varrho - 1)^2}.$$

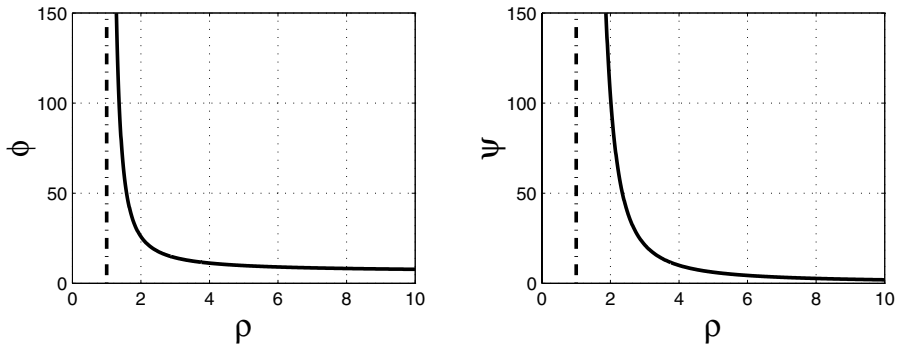


FIG. 1. *Graphs of the functions $\phi(\varrho)$ and $\psi(\varrho)$ defined, respectively, by* (2.17) *and* (4.4).

The graph of this function is shown in Figure 1. A strictly decreasing function on $(1,\infty)$, ϕ assumes values less than 100 for all $\varrho \geq 1.36$ and less than 10 for all $\varrho \geq 4.84$, with asymptotic value 2π as $\varrho \to \infty$.

We are now in a position to state the theorem.

THEOREM 2.1. *Let $p_N(x)$ be the polynomial interpolant of $f(z)$ at the set of Chebyshev zeros $\{s_j\}$ defined by* (2.1) *or extrema $\{t_j\}$ defined by* (2.2). *Suppose $f(z)$ is analytic in some ellipse E_ϱ, defined by* (2.8), *with $\varrho = e^\eta$, $\eta > 0$. Then, for each $N \geq 1$, we have*

$$\text{(2.18)} \qquad \textit{Zeros:} \quad \max_{0\leq j\leq N} |f'(s_j) - p_N'(s_j)| \leq \frac{1}{2\pi}\, C_\varrho\, \phi(\varrho)\, \frac{(N+1)}{\sin\left(\frac{\pi}{2N+2}\right)}\, \frac{1}{\sinh(\eta(N+1))},$$

$$\text{(2.19)} \qquad \textit{Extrema:} \quad \max_{0\leq j\leq N} |f'(t_j) - p_N'(t_j)| \leq \frac{1}{\pi}\, C_\varrho\, \phi(\varrho)\, N\, \frac{1}{\sinh(\eta)\sinh(\eta N)}.$$

In both cases, $\phi(\varrho)$ is defined by (2.17), *and*

$$\text{(2.20)} \qquad C_\varrho = \max_{z\in E_\varrho} |f(z)|.$$

Proof. Taking absolute values in the error integral (2.7), with $\Gamma = E_\varrho$, yields

$$\begin{aligned} |f'(x_j) - p_N'(x_j)| &\leq \frac{1}{2\pi}|\omega_{N+1}'(x_j)| \int_{E_\varrho} \frac{|f(z)|}{|\omega_{N+1}(z)||z-x_j|}|dz| \\ &\leq \frac{1}{2\pi}|\omega_{N+1}'(x_j)| \frac{\max_{z\in E_\varrho}|f(z)|}{\min_{z\in E_\varrho}|\omega_{N+1}(z)|} \int_{E_\varrho} \frac{|dz|}{|z-x_j|}. \end{aligned}$$

Taking the maximum over all nodes yields

$$\max_{0\leq j\leq N} |f'(x_j) - p_N'(x_j)| \leq \frac{1}{2\pi} \max_{0\leq j\leq N} |\omega_{N+1}'(x_j)| \frac{\max_{z\in E_\varrho}|f(z)|}{\min_{z\in E_\varrho}|\omega_{N+1}(z)|} \frac{L_\varrho}{D_\varrho}.$$

The error bounds (2.18)–(2.19) now follow from (2.11)–(2.17). ∎

Remark 1. The exponential convergence of the spectral method comes from the hyperbolic sine term in the denominators of (2.18)–(2.19). This can be made explicit by using the bound

$$\frac{1}{\sinh(\eta N)} \leq 2 \coth(\eta N)\, e^{-\eta N},$$

which is sharp to within a factor of 2, and an asymptotic equality in the limit $N \to \infty$, with η fixed. A large ellipse (i.e., a large region about $[-1, 1]$ in which the function is analytic) implies large values of ρ and hence η, which means quick convergence.

Remark 2. When the function is entire, i.e., free of singularities in the finite complex plane, the size of the ellipse E_ϱ is not limited by singularities but by the growth rate of the factor C_ϱ as ϱ increases. In this case convergence is typically faster than exponential, a situation referred to in [3, p. 26] as supergeometric convergence.

Remark 3. The equalities (2.13) and (2.14) give rise to algebraic factors of $O(N)$ and $O(N^2)$, respectively, in the error bounds (2.18) and (2.19). This shows that the Chebyshev extreme points are slightly better for differentiation than the Chebyshev zeros, as will also be observed empirically in the numerical experiment below.

Remark 4. Continuing the idea of Remark 3, the algebraic factor in the error estimate can be minimized by choosing the nodes x_j such that the maximum of $|\omega'_{N+1}(x)|$ on $[-1, 1]$ is minimized. Using the well-known min-max property of the Chebyshev polynomials, we therefore propose that

$$\omega'_{N+1}(x) = T_N(x).$$

Integration yields

$$\omega_{N+1}(x) = \frac{1}{2}\left(\frac{T_{N+1}(x)}{N+1} - \frac{T_{N-1}(x)}{N-1}\right) + c,$$

where the constant c may be chosen to make $\omega_{N+1}(1) = 0$ or $\omega_{N+1}(-1) = 0$ (depending on where a boundary condition needs to be enforced). If N is odd, both of these conditions can, owing to symmetry, be satisfied simultaneously. In addition, if N is odd, the roots of $\omega_{N+1}(x)$ lie in $[-1, 1]$, which is not the case for N even. With these roots as the nodes of the spectral collocation process, one should obtain an error estimate similar to (2.18)–(2.19), but with the leading algebraic term of size $O(1)$ instead of $O(N)$ or $O(N^2)$.

In practice this new set of points does indeed seem to give marginally better accuracy than the other two node sets analyzed here. With $N = 15$, for example, we computed the error in the numerical differentiation of the test function (2.21) below, using $a = 2$. The maximum error at the grid points were 2.3×10^{-7} and 9.1×10^{-8}, respectively, with Chebyshev zeros and Chebyshev extrema as nodes. The new set of nodes proposed here yielded an error 4.9×10^{-8}.

Remark 5. The algebraic factor in (2.19) is $O(N)$, which improves on the $O(N^{2.5})$ factor in [15, eq. (4.18)]. The exponential factors are identical.

Remark 6. It is possible to extend Theorem 2.1 with minimal effort to the continuous rather than the discrete maximum norm. First one applies the triangle inequality to the integral (2.6). Then one uses elementary bounds such as $\max_{-1\leq x\leq 1} |\omega_{N+1}(x)| \leq 1$ (zeros), $\max_{-1\leq x\leq 1} |\omega_{N+1}(x)| \leq 2$ (extrema), and the inequality (2.13) to obtain

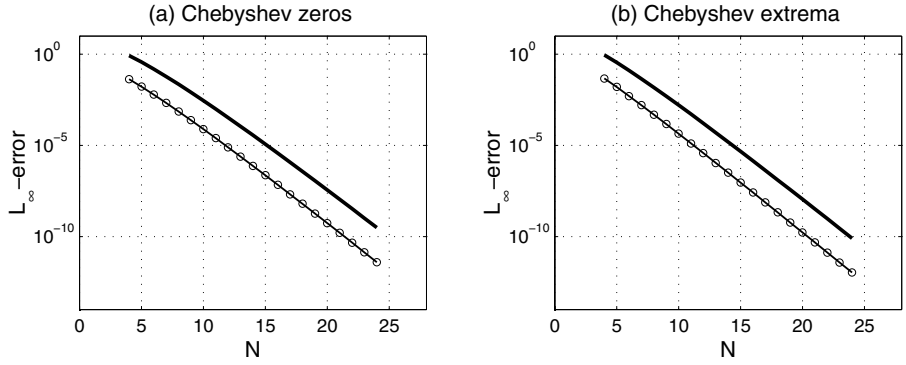


FIG. 2. *Actual error (dotted curve) and theoretical error bound (thicker curve) in the Chebyshev derivative of* $f(x) = 1/(x-a), a = 2$.

$$\textit{Zeros:} \quad \max_{-1\le x\le 1} |f'(x) - p_N'(x)| \le \frac{1}{2\pi}\, C_\varrho\, L_\varrho \left(\frac{(N+1)^2}{D_\varrho} + \frac{1}{D_\varrho^2}\right) \frac{1}{\sinh(\eta(N+1))},$$

$$\textit{Extrema:} \quad \max_{-1\le x\le 1} |f'(x) - p_N'(x)| \le \frac{1}{2\pi}\, C_\varrho\, L_\varrho \left(\frac{2N}{D_\varrho} + \frac{1}{D_\varrho^2}\right) \frac{1}{\sinh(\eta)\sinh(\eta N)}.$$

In each case, the asymptotic behavior of the error bound, as $N \to \infty$, is the same as in Theorem 2.1.

Numerical experiments. We checked the tightness of the bounds in Theorem 2.1 numerically, using

$$f(x) = \frac{1}{x-a}, \qquad a > 1, \tag{2.21}$$

as test function. In Figure 2 we show, as the thinner, dotted lines, the computed error $\max_{0\le j\le N} |f'(x_j) - p_N'(x_j)|$ as a function of N, for the case $a = 2$. Also shown, as the thicker curves, are the theoretical error bounds (2.18)–(2.19). In order to compute the bounds, one is free to choose the ellipse E_ϱ, the only constraint being that it intersects the real axis between $x = 1$ and the pole at $x = a$, i.e.,

$$1 < \frac{1}{2}\left(\varrho + \varrho^{-1}\right) < a \quad \Longrightarrow \quad 1 < \varrho < a + \sqrt{a^2 - 1}.$$

For each value of N we computed the right-hand sides of (2.18)–(2.19), for a thousand values of ϱ in the above interval, and picked the minimum value to determine the best upper bound.

We note from the figure that for $a = 2$ the theoretical bound overestimates the actual error by about one order of magnitude. If we decrease (resp., increase) the value of a, this overestimation factor increases (resp., decreases).

As a second example, we consider

$$f(x) = \sqrt{a - x}, \qquad a > 1, \tag{2.22}$$

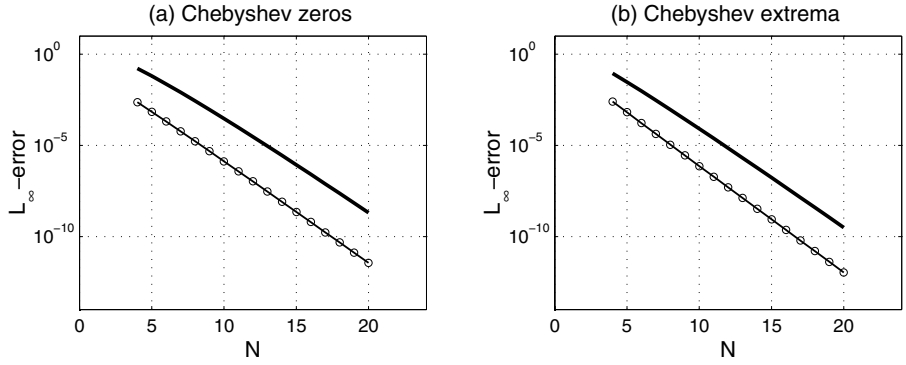


FIG. 3. *Same as Figure* 2, *but the function is* $f(x) = \sqrt{a-x}, a = 2$.

with corresponding error curves shown in Figure 3. Compared to the first example, one notices (a) that the actual errors in Figure 3 are smaller than in Figure 2 (note that the scale on the horizontal axis is different in the two figures), and (b) that the gap between the actual errors and the theoretical error bounds is slightly wider. Both of these observations will be explained in the next section.

The reason for the gap between the actual errors and the theoretical error bounds is the fact that the ellipse (2.8) is a general, all-purpose contour that has to cope with all possible singularities of the function. When specific assumptions regarding the singularities are made, different contours may yield more precise estimates. This will be done in the next section, where we investigate these numerical results more carefully. In fact, we shall obtain an explicit representation for the error in (2.21) and accurate estimates for the error in (2.22).

3. Analysis of two model functions. The test functions (2.21) and (2.22), representative of functions with poles and branch-point singularities, respectively, will be analyzed here with the aid of the contour shown in Figure 4. It consists of a large circle, centered at the origin and with radius R, and a small circle, centered at the singularity $z = a$ with radius ϵ. The larger circle is positively traversed, i.e., runs counterclockwise, and the small one runs clockwise. The two circles are connected by two line segments on the real interval $[a + \epsilon, R]$, as indicated by the arrows in the figure.

We start with the model function (2.21) and use Chebyshev zeros as collocation points. The error integral (2.7) thus becomes

$$f'(s_j) - p'_N(s_j) = \frac{T'_{N+1}(s_j)}{2\pi i} \int_\Gamma \frac{dz}{T_{N+1}(z)(z - s_j)(z - a)}.$$

Taking Γ to be the contour in Figure 4, the various contributions to the integral can be estimated using standard results from complex variable theory; see for example [14, Ch. 6]. The contribution on the large circle, $|z - a| = R$, vanishes in the limit $R \to \infty$, owing to the polynomial term in the denominator. The contribution along the connecting line segments cancels, and the contribution on the small circle, $|z - a| = \epsilon$, can be evaluated as a residue. The result is the explicit expression for the error that was announced at the end of section 2, namely

$$f'(s_j) - p'_N(s_j) = \frac{T'_{N+1}(s_j)}{T_{N+1}(a)} \frac{1}{s_j - a}, \qquad j = 0, 1, \ldots, N. \tag{3.1}$$

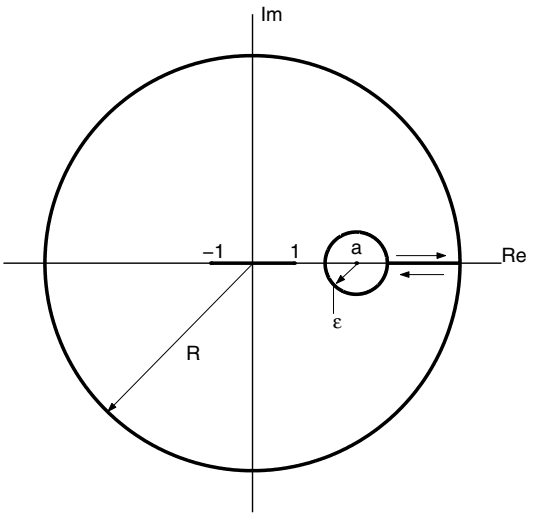


FIG. 4. *Contour used in the analysis of the model functions.*

A uniform bound on the right-hand side of (3.1) can be obtained by using (2.13) and the inequality

$$T_{N+1}(x) > \frac{1}{2}\Big(x + \sqrt{x^2-1}\Big)^{N+1}, \quad x > 1. \tag{3.2}$$

This yields

$$\max_{0\le j\le N} |f'(s_j) - p_N'(s_j)| \le \frac{2}{a-1}\frac{N+1}{\sin(\frac{\pi}{2N+2})}\Big(a + \sqrt{a^2-1}\Big)^{-(N+1)}, \tag{3.3}$$

a bound that is virtually indistinguishable from the actual error curve in Figure 2(a).

Comparing this tight bound with the general bound (2.18), one concludes that the latter formula captures the dependence on N perfectly. The gap between the actual error and the theoretical error bounds observed in Figure 2 is therefore entirely due to overestimating the constant factor.

Turning to the second model problem (2.21), we choose the branch-cut of $\sqrt{a-z}$ to be the real interval $[a,\infty)$. In this case, therefore, the contributions of the real line segments do not cancel, as one segment is "above" the branch-cut and the other "below" [14, Sect. 6.6]. In fact, these contributions represent the error completely, as the contribution on the small circle, $|z-a| = \epsilon$, vanishes in the limit $\epsilon \to 0$, and so does the contribution on the large circle, $|z| = R$, as $R \to \infty$.

Adding the contributions on $[a,\infty)$, one obtains in the case of Chebyshev zeros

$$f'(s_j) - p_N'(s_j) = \frac{T_{N+1}'(s_j)}{\pi}\int_a^\infty \frac{\sqrt{x-a}}{s_j - x}\frac{dx}{T_{N+1}(x)}, \quad j = 0,1,\dots,N. \tag{3.4}$$

We have not succeeded in finding a closed form expression for the integral on the right, but a relatively sharp bound can be obtained as follows.

Note that for $a > 1$ and $|s| \le 1$,

$$\int_a^\infty \frac{\sqrt{x-a}}{x-s}\frac{dx}{T_{N+1}(x)} \le \frac{1}{a-s}\int_a^\infty \frac{\sqrt{x-a}}{T_{N+1}(x)}\,dx \le \frac{2}{a-s}\int_a^\infty \frac{\sqrt{x-a}}{(x+\sqrt{x^2-1})^{N+1}}\,dx,$$

where (3.2) was used. The latter integral can be evaluated explicitly in terms of the beta function, B, and the hypergeometric function, ${}_2F_1$; see [13, entry 7, p. 314]. The result is

$$\int_a^\infty \frac{\sqrt{x-a}}{(x+\sqrt{x^2-1})^{N+1}}\,dx = \frac{\sqrt{a}}{2}(2a)^{-N}B\left(\frac{3}{2},N-\frac{1}{2}\right){}_2F_1\left(\frac{1}{2}N+\frac{1}{4},\frac{1}{2}N-\frac{1}{4};2+N;\frac{1}{a^2}\right).$$

In order to obtain a form of the ${}_2F_1$ function that can be expressed in terms of elementary functions, we replace the third parameter $2+N$ by $\frac{1}{2}+N$, which causes a relatively small overestimate when N is large. Using entry 15.1.13 from [1, p. 556], one consequently obtains

$$\int_a^\infty \frac{\sqrt{x-a}}{(x+\sqrt{x^2-1})^{N+1}}\,dx \le \frac{1}{2\sqrt{2}}B\left(\frac{3}{2},N-\frac{1}{2}\right)\left(a+\sqrt{a^2-1}\right)^{-N+1/2}.$$

Approximating the B function by Stirling's formula yields $B(\frac{3}{2},N-\frac{1}{2}) \sim (\sqrt{\pi}/2)N^{-(3/2)}$, $N \to \infty$. This approximation and the above inequalities are now inserted into (3.4) to obtain an error estimate similar to (3.3):

$$\max_{0\le j\le N}|f'(s_j)-p_N'(s_j)| \le \frac{1}{\sqrt{\pi}(a-1)(2N)^{3/2}}\frac{N+1}{\sin(\frac{\pi}{2N+2})}\left(a+\sqrt{a^2-1}\right)^{-N+1/2}. \tag{3.5}$$

Numerical verification confirms that this bound is tight. When compared to the data of Figure 3(a), the gap between this bound and the actual error curve is tiny. In fact, with $a = 2$ the right side of (3.5) overestimates the left side by less than a factor of 2 for all $N \ge 3$.

By first comparing (3.5) to the analogous expression for the first model function, equation (3.3), one sees that the errors in the second model function are smaller owing to the $N^{3/2}$ factor in the denominator. Next, by comparing (3.5) to the general error bound (2.18), one notices that the exponential dependence on N is the same. The leading algebraic factor in (2.18) is, however, overestimated by a factor of $O(N^{3/2})$, and this accounts for the widening gap between the actual error and the theoretical error bound in Figure 3(a).

The results in this section were presented for Chebyshev zeros only, but similar results can be derived for the extreme points. We omit the details.

4. Second derivatives. The results of section 2 can be generalized in a straightforward manner to second derivatives. Differentiating (2.3) twice with respect to x, and again using the fact that $\omega_{N+1}(x_j) = 0$ at the discretization points, we obtain for $j = 0,\dots,N$

$$f''(x_j)-p_N''(x_j) = \frac{1}{2\pi i}\int_\Gamma\left(\frac{2\omega_{N+1}'(x_j)}{(z-x_j)^2}+\frac{\omega_{N+1}''(x_j)}{(z-x_j)}\right)\frac{f(z)}{\omega_{N+1}(z)}\,dz. \tag{4.1}$$

The factor $|\omega''(x_j)|$ that appears in this formula can be estimated similarly to the estimation of (2.13)–(2.14), and one obtains for $N = 1, 2, \ldots,$

$$\textit{Zeros:} \quad \max_{0\le j\le N} |\omega''_{N+1}(s_j)| = (N+1)\cos\left(\frac{\pi}{2N+2}\right)\csc^3\left(\frac{\pi}{2N+2}\right), \\ \le \frac{1}{3}N(N+2)(N+1)^2,$$

$$\textit{Extrema:} \quad \max_{0\le j\le N} |\omega''_{N+1}(t_j)| = \frac{4}{3}N(2N^2+1).$$

The error bounds are as follows.

THEOREM 4.1. *Let $p_N(x)$ be the polynomial interpolant of $f(z)$ at the set of Chebyshev zeros $\{s_j\}$ defined by* (2.1) *or extrema $\{t_j\}$ defined by* (2.2). *Suppose $f(z)$ is analytic in some ellipse E_ϱ, defined by* (2.8), *with $\varrho = e^\eta$, $\eta > 0$. Then, for each $N \ge 2$,*

$$\textit{Zeros:} \quad \max_{0\le j\le N} |f''(s_j) - p''_N(s_j)| \\ \le \frac{1}{2\pi} C_\varrho \left(\alpha_N \psi(\varrho) + \beta_N \phi(\varrho)\right) \frac{1}{\sinh(\eta(N+1))}, \tag{4.2}$$

$$\textit{Extrema:} \quad \max_{0\le j\le N} |f''(t_j) - p''_N(t_j)| \\ \le \frac{1}{2\pi} C_\varrho \left(\gamma_N \psi(\varrho) + \delta_N \phi(\varrho)\right) \frac{1}{\sinh(\eta)\sinh(\eta N)}, \tag{4.3}$$

where C_ϱ and $\phi(\varrho)$ are defined by (2.20) *and* (2.17), *respectively, and*

$$\psi(\varrho) = 4\pi \frac{\varrho\sqrt{\varrho^4+1}}{(\varrho-1)^4}. \tag{4.4}$$

In addition,

$$\alpha_N = 2(N+1)\csc\left(\frac{\pi}{2N+2}\right), \quad \beta_N = (N+1)\cos\left(\frac{\pi}{2N+2}\right)\csc^3\left(\frac{\pi}{2N+2}\right), \\ \gamma_N = 4N, \qquad\qquad \delta_N = \frac{2}{3}N(2N^2+1).$$

Bounds on errors in higher derivatives can also be obtained. For the pth derivative, the leading algebraic factor scales like $O(N^{2p})$ for the Chebyshev zeros and like $O(N^{2p-1})$ for the Chebyshev extreme points.

Acknowledgments. The authors are indebted to Dave Sloan and two anonymous referees for a careful reading of the manuscript, and for many good suggestions. Milton Maritz provided help with the software package Mathematica.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, Inc., New York, 1992.

[2] R. W. BARNARD, K. PEARCE, AND L. SCHOVANEC, *Inequalities for the perimeter of an ellipse*, J. Math. Anal. Appl., 260 (2001), pp. 295–306.

[3] J. P. BOYD, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover Publications Inc., Mineola, NY, 2001.

[4] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer Ser. Comput. Phys., Springer-Verlag, New York, 1988.
[5] P. J. Davis, *Interpolation and Approximation*, Dover Publications Inc., New York, 1975.
[6] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge Monographs on Applied and Computational Mathematics 1, Cambridge University Press, Cambridge, UK, 1996.
[7] B. Fornberg and D. M. Sloan, *A review of pseudospectral methods for solving partial differential equations*, in Acta Numerica, Cambridge University Press, Cambridge, UK, 1994, pp. 203–267.
[8] D. Funaro, *Polynomial Approximation of Differential Equations*, Lecture Notes in Physics. New Series m: Monographs 8, Springer-Verlag, Berlin, 1992.
[9] D. Gottlieb, M. Y. Hussaini, and S. A. Orszag, *Theory and applications of spectral methods*, in Spectral Methods for Partial Differential Equations, SIAM, Philadelphia, 1984, pp. 1–54.
[10] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.
[11] J. C. Mason and D. C. Handscomb, *Chebyshev Polynomials*, Chapman & Hall/CRC Press, Boca Raton, FL, 2003.
[12] B. Mercier, *An Introduction to the Numerical Analysis of Spectral Methods*, Lecture Notes in Phys. 318, Springer-Verlag, Berlin, 1989.
[13] A. P. Prudnikov, Y. A. Brychkov, and O. I. Marichev, *Integrals and Series. Vol.* 1, Gordon & Breach Science Publishers, New York, 1986.
[14] E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis for Mathematics, Science, and Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1976.
[15] E. Tadmor, *The exponential accuracy of Fourier and Chebyshev differencing methods*, SIAM J. Numer. Anal., 23 (1986), pp. 1–10.
[16] L. N. Trefethen, *Spectral Methods in MATLAB*, SIAM, Philadelphia, 2000.
[17] J. A. C. Weideman and S. C. Reddy, *A MATLAB differentiation matrix suite*, ACM Trans. Math. Software, 26 (2000), pp. 465–519.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2188–2217

AN OPTIMAL ADAPTIVE FINITE ELEMENT METHOD*

ROB STEVENSON†

Abstract. Although existing adaptive finite element methods for solving second order elliptic equations often perform better in practical computations than nonadaptive ones, usually they are not even proven to converge. Only recently in the work of Dörfler [*SIAM J. Numer. Anal.*, 33 (1996), pp. 1106–1124] and that of Morin, Nochetto, and Siebert [*SIAM J. Numer. Anal.*, 38 (2000), pp. 466–488], adaptive methods were constructed for which convergence could be demonstrated. However, convergence alone does not imply that the method is more efficient than its nonadaptive counterpart. In [*Numer. Math.*, 97 (2004), pp. 219–268], Binev, Dahmen, and DeVore added a coarsening step to the routine of Morin, Nochetto, and Siebert, and proved that the resulting method is quasi-optimal in the following sense: If the solution is such that for some $s > 0$, the error in energy norm of the best continuous piecewise linear approximations subordinate to any partition with n triangles is $\mathcal{O}(n^{-s})$, then given an $\varepsilon > 0$, the adaptive method produces an approximation with an error less than ε subordinate to a partition with $\mathcal{O}(\varepsilon^{-1/s})$ triangles, in only $\mathcal{O}(\varepsilon^{-1/s})$ operations.

In this paper, employing a different type of adaptive partition, we develop an adaptive method with properties similar to those of Binev, Dahmen, and DeVore's method, but unlike their method, our coarsening routine will be based on a transformation to a wavelet basis, and we expect it to have better quantitative properties. Furthermore, all our results are valid uniformly in the size of possible jumps of the diffusion coefficients. Since the algorithm uses solely approximations of the right-hand side, we can even allow right-hand sides in $H^{-1}(\Omega)$ that lie outside $L_2(\Omega)$, at least when they can be sufficiently well approximated by piecewise constants. In our final adaptive algorithm, all tolerances depend on an a posteriori estimate of the current error instead of an a priori one; this can be expected to provide quantitative advantages.

Key words. adaptive finite element method, optimal computational complexity, wavelet basis, a posteriori error estimator, discontinuous coefficients

AMS subject classifications. 65N30, 65N50, 65N15, 65Y20, 65T60, 41A25

DOI. 10.1137/S0036142903425082

1. Introduction.

For solving elliptic boundary value problems for which the solution has singularities, the use of adaptive finite element methods potentially has the advantage of a significant reduction of the computational cost, compared to nonadaptive methods. Although the adaptive methods that can be found in the literature often exhibit such a reduction, they are, usually, not even proven to converge, let alone shown to outperform nonadaptive methods. Only quite recently, in the work of Dörfler [11], which was later extended by Morin, Nochetto, and Siebert in [15], adaptive methods were constructed that were proven to converge. These methods are based on an adaptive refinement strategy that guarantees the so-called saturation property, namely that the difference between the solutions on two consecutive partitions is greater than some multiple of the error in the solution on the first partition. Exploiting Galerkin orthogonality, convergence then easily follows.

In [3], Binev, Dahmen, and DeVore added a coarsening step to the method from [15]. Basically the idea of such a step, which has to be applied after each fixed number of refinement steps, is to undo refinements that ultimately hardly contribute

*Received by the editors March 19, 2003; accepted for publication (in revised form) January 9, 2004; published electronically March 25, 2005. This work was supported by the Netherlands Organization for Scientific Research and by the European Community's Human Potential Programme under contract HPRN-CT-2002-00286.

http://www.siam.org/journals/sinum/42-5/42508.html

†Department of Mathematics, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands (stevenson@math.uu.nl).

to a better approximation. Thanks to this coarsening, under some conditions on the right-hand side, the resulting adaptive method was proven to be quasi-optimal in the following sense: if the solution is such that, for some $s > 0$, the error in energy norm of the best continuous piecewise linear approximations subordinate to any partition with n triangles is $\mathcal{O}(n^{-s})$, then given an $\varepsilon > 0$, the adaptive method produces an approximation with an error less than ε subordinate to a partition with $\mathcal{O}(\varepsilon^{-1/s})$ triangles, in only $\mathcal{O}(\varepsilon^{-1/s})$ operations.

In this paper, we consider a method, as developed in [15] and extended with a coarsening in [3], in a slightly different context: instead of working with conforming partitions produced with the so-called newest vertex bisection method, we consider generally nonconforming partitions produced by only "red-refinement" steps, i.e., splittings of triangles into four congruent subtriangles. In this setting we extend the findings from [15] and [3] on some points.

- Following [15], we consider the model problem of Poisson's equation on a two-dimensional domain, generalized in the sense that a piecewise constant diffusion tensor is allowed. Assuming a so-called quasi-monotone distribution of possible jumps that this tensor may have, all results from this paper will be proven to hold uniformly in the size of such jumps.
- With an adaptive method, the right-hand side and the discrete Galerkin solution subordinate to the current partition together determine the next partition via an a posteriori error estimator. Aiming at proving optimal computational complexity, following [3] we consider inexact solutions of the discrete systems. In addition to that, we allow inexact right-hand sides to be used for both setting up the discrete systems and for the evaluation of the a posteriori error estimator. This generalization can be used to model the effect of the application of quadrature. Furthermore, it will allow us to prove quasi-optimality of the adaptive method even for right-hand sides in $H^{-1}(\Omega)$ outside $L_2(\Omega)$, at least when they can be sufficiently well approximated by piecewise constants.
- We introduce a new coarsening procedure that, unlike the procedure from [3], is based on a transformation to a wavelet basis for the space of continuous piecewise linears subordinate to the adaptively refined partition. We expect our procedure to have better quantitative properties, although admittedly a final answer can only be given after performing numerical tests. Both the coarsening from [3] and our coarsening rely on an adaptive tree approximation algorithm developed by Binev and DeVore in [5].
- The adaptive finite element method from [3] and our first routine **SOLVE1** require as input an a priori upper bound μ of the convergence rate of the algorithm without coarsening. When one supplies a μ that is too small, quasi-optimality as a consequence of the coarsening is not guaranteed. On the other hand, taking a μ that is unnecessarily close to 1 will result in a quantitatively less attractive algorithm, since, due to the coarsening, the convergence rate will be limited by this μ. In this paper, we develop a second routine **SOLVE2** in which the tolerances allowed in the inexact Galerkin solutions and in the approximations of the right-hand side, and those required in the coarsening, are all some fixed multiples of an a posteriori estimate of the current error. Apart from the fact that this releases the user from the task of supplying this most critical parameter, the new algorithm benefits from a better convergence rate than what might appear to be the case from an a priori worst case analysis.

Finally, let us comment on the necessity of applying a coarsening routine. In the numerical experiments reported in [15], the partitions, although produced without

coarsening, already seem to have a quasi-optimal cardinality. Of course, this does not exclude the possibility that there are other examples for which coarsening is necessary. On the other hand, it is also possible that coarsening is not a necessary ingredient of a quasi-optimal adaptive algorithm for solving these elliptic problems, but that a proof of such a fact is eluding us. In any case, we do not consider our construction of a coarsening routine as being relevant for theoretical purposes only. Instead, we expect that coarsening will be very useful inside adaptive routines for solving nonstationary problems.

This paper is organized as follows: In section 2 our model boundary value problem is described.

In section 3, we introduce the class of admissible partitions, which is the subclass of all partitions that can be generated by red-refinements for which the generations of neighboring triangles differ at most by one. We show that any partition can be refined to an admissible one by increasing the number of triangles by, at most, a constant factor.

In section 4, we introduce a wavelet basis for the space of continuous piecewise linears subordinate to any admissible partition. We show that both the basis transformation from wavelet to nodal basis and its inverse can be performed in optimal computational complexity.

Our coarsening routine is defined in section 5. It is based on a transformation to wavelet basis, an application of the adaptive tree approximation routine from [5], and, finally, a construction of a reduced partition subordinate to which the remaining terms in the wavelet expansion are continuous piecewise linear functions.

In section 6, an a posteriori error estimator is derived. A refinement strategy is developed that is shown to be convergent also for inexact, but sufficiently accurate, right-hand sides and discrete solutions.

In section 7, the coarsening routine and the convergent adaptive refinement strategy are combined into an optimal adaptive finite element method.

Finally, in section 8 we derive an optimal adaptive finite element method in which the tolerances for the errors in the right-hand side and in the discrete solution as well as for the coarsening routine are determined by an a posteriori estimate of the current error.

In order to avoid the repeated use of generic but unspecified constants, in this paper by $C \lesssim D$ we mean that C can be bounded by a multiple of D, independent of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

2. Boundary value problem. Let Ω be a polygonal bounded domain in $\mathbb{R}^2$. We consider the following model boundary value problem in variational form: Given $f \in H^{-1}(\Omega)$, find $u \in H^1_0(\Omega)$ such that

$$a(u,w) := \int_\Omega \mathbf{A}\nabla u \cdot \nabla w = f(w), \quad (w \in H^1_0(\Omega)), \tag{2.1}$$

where $\mathbf{A} \in L_\infty(\Omega)$ is a symmetric 2×2 matrix with $\operatorname{ess\,inf}_{x\in\Omega}\lambda_{\min}(\mathbf{A}(x)) > 0$. Further assumptions on $\mathbf{A}$ are collected in the forthcoming Assumption 3.8. Defining $L : H^1_0(\Omega) \to H^{-1}(\Omega) = (H^1_0(\Omega))'$ by $(Lu)(w) = a(u,w)$, (2.1) can be rewritten as

$$Lu = f.$$

In some places we will assume that the right-hand side $f \in L_2(\Omega)$, in which case $f(w)$ should be interpreted as $\int_\Omega fw$.

Aiming at results that hold uniformly in the size of variations that the spectral radius $\rho := \rho(\mathbf{A})$ may have, we introduce a weighted $L_2(\Omega)$-scalar product

$$\langle u, w\rangle_0 = \int_\Omega \rho uw,$$

and define the weighted norms

$$\|w\|_0 = \langle w, w\rangle_0^{\frac{1}{2}}, \quad |w|_1 = a(w,w)^{\frac{1}{2}}, \quad \|g\|_{-1} = \sup_{0\neq w\in H_0^1(\Omega)} \frac{|g(w)|}{|w|_1}$$

on $L_2(\Omega)$, $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, respectively. Equipped with these norms, L is an isomorphism between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$.

For $\Sigma \subset \Omega$, we define $|w|_{1,\Sigma} := (\int_\Sigma \mathbf{A}|\nabla w|^2)^{\frac{1}{2}}$.

3. Partitions of Ω. We shall approximate the solution of (2.1) by continuous piecewise linear functions subordinate to a partition of Ω into triangles. In this subsection we precisely describe the type of partitions to be considered.

We will use P to denote a *partition* of Ω, defined as a collection of closed triangles $\triangle$ such that $\overline{\Omega} = \cup_{\triangle\in P}\triangle$ and $\text{meas}(\triangle \cap \tilde{\triangle}) = 0$ for any two different $\triangle, \tilde{\triangle} \in P$. When $\triangle \cap \tilde{\triangle} \neq \emptyset$, such triangles will be called *neighbors*. A partition $\tilde{P}$ is called a *refinement* of P, when $\tilde{P}$ can be constructed by, for zero or more $\triangle \in P$, replacing $\triangle$ by the four subtriangles created by connecting the midpoints of edges of $\triangle$ or by a recursive application of this elementary "red" refinement step. The above $\triangle$ will be referred to as being the *parent* of its four subtriangles, called *children* of $\triangle$. As expected, children of children of $\triangle$ are called *grandchildren* of $\triangle$.

Throughout this paper we consider only partitions P that are refinements of some *fixed initial partition* P_0 of Ω. Many of our statements will involve constants that actually depend on P_0. However, since P_0 is assumed to be fixed, for ease of presentation we ignore these dependencies. Clearly, any $\triangle \in P$ is *similar* to a triangle from P_0. For $\triangle \in P$, $\text{gen}(\triangle)$ will denote the number of elementary refinement steps needed to create $\triangle$ starting from some $\tilde{\triangle} \in P_0$, where $\text{gen}(\tilde{\triangle}) := 0$.

We call v a *vertex of* P, when there exists a $\triangle \in P$ such that v is a vertex of $\triangle$. A vertex v of P is called *nonhanging* when it is a vertex of all $\triangle \in P$ that contain v, otherwise it is called a *hanging* vertex of P. With $\bar{V}_P$ or V_P we will denote the set of all *nonhanging* vertices of P or all *nonhanging, interior* vertices of P, respectively. We assume that P_0 is *conforming*, i.e., all its vertices are nonhanging.

A vertex v of P is called *regular* when for all $\triangle \in P$ that contain v, $\text{gen}(\triangle)$ has the same value; see Figure 3.1. Note that a regular vertex is nonhanging.

For a vertex v of P, the number of $\triangle \in P$ that contain v is called the *valence* of v in P. The valence of any v of P is less than or equal to the maximum of 6 and the maximum valence of all vertices of P_0. If for a $\triangle \in P$, $\text{gen}(\triangle) = \max_{\tilde{\triangle}\in P} \text{gen}(\tilde{\triangle})$, then its edges cannot contain hanging nodes. As a consequence, for such $\triangle$, the number of its neighbors in P is given by the sum of the valences of its vertices minus the sum of 6 and the number of edges of $\triangle$ on $\partial\Omega$; this shows, in particular, that the number of neighbors is uniformly bounded.

PROPOSITION 3.1. *For any partition P of the type we consider, there exists a unique sequence of partitions*

$$P_0, P_1, \ldots, P_n$$

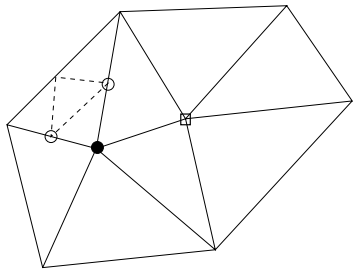


FIG. 3.1. *Regular (□), nonhanging but nonregular (•), and hanging vertices (○).*

with $\max_{\triangle\in P_i}\operatorname{gen}(\triangle)=i$, $P_n=P$, *and where* P_{i+1} *is created from* P_i *by refining some* $\triangle\in P_i$ *with* $\operatorname{gen}(\triangle)=i$. *For convenience, we set* $P_{-1}=\emptyset$ *and so* $V_{P_{-1}}=\emptyset$. *The following properties are valid:*

(i) $V_{P_{i-1}}\subset V_{P_i}$ *and so* $V_P=\cup_{i=0}^n V_{P_i}\backslash V_{P_{i-1}}$ *with empty mutual intersections;*

(ii) *a* $v\in V_{P_i}\backslash V_{P_{i-1}}$ *is not a vertex of* P_{i-1}, *and so it is a regular vertex of* P_i.

Proof. The existence and uniqueness of the sequence $(P_i)_i$ and also (i) are obvious.

Suppose that for some $1\le i\le n$, $v\in V_{P_i}\backslash V_{P_{i-1}}$ is a vertex of P_{i-1}; then it is a hanging vertex of P_{i-1}. So there is a $\triangle\in P_{i-1}$ with $\operatorname{gen}(\triangle)<i-1$ that contains v. However, by definition of the sequence $(P_i)_i$, such a $\triangle$ is not refined when going to P_i, meaning that v is also a hanging vertex of P_i, which gives a contradiction. □

NOTATION 3.2. *Throughout this paper, for any partition* P *(or* $\hat P$, $\tilde P$, *etc.), by* $(P_i)_i$ *(or* $(\hat P_i)_i$, $(\tilde P_i)_i$, *etc.) we will always mean the corresponding sequence, as in Proposition* 3.1. *When we write* $P=P_n$, *we mean that* P *is a partition with* $\max_{\triangle\in P}\operatorname{gen}(\triangle)=n$.

In view of the forthcoming discussion on adaptively refined partitions, we emphasize here that given any partition P, the definition of the corresponding sequence $(P_i)_i$ is independent of the way P has been constructed.

DEFINITION 3.3. *A partition* P^a *is called* admissible *when for all neighbors* $\triangle,\tilde\triangle\in P^a$, $|\operatorname{gen}(\triangle)-\operatorname{gen}(\tilde\triangle)|\le 1$.

As will turn out later, the reason to consider this restricted class of admissible partitions is given by the following proposition.

PROPOSITION 3.4. *Let* P^a *be admissible. For any* $\triangle\in P^a$ *with* $i:=\operatorname{gen}(\triangle)>0$ *the vertices of the parent* $\tilde\triangle\in P^a_{i-1}$ *of* $\triangle$ *are regular vertices of* P^a_{i-1}.

Proof. For $i=1$ the statement is true. Now let $i>1$. Suppose that some vertex v of $\tilde\triangle$ is not a regular vertex of P^a_{i-1}. Then there exists a $\hat\triangle\in P^a_{i-1}$ with $v\in\hat\triangle$ and $\operatorname{gen}(\hat\triangle)<i-1$. Since, by definition of the sequence $(P^a_i)_i$, this $\hat\triangle$ will never be refined, we get a contradiction with the fact that P^a is admissible. □

PROPOSITION 3.5. *If* $P^a=P^a_n$ *is admissible, then for any* $0\le i\le n$, P^a_i *is also admissible.*

Proof. Suppose P^a_i is not admissible; then it contains neighbors $\hat\triangle,\triangle$ with $\operatorname{gen}(\hat\triangle)<\operatorname{gen}(\triangle)-1$. Since $\operatorname{gen}(\hat\triangle)<i$, it will never be refined and so P^a cannot be admissible. □

Although not all partitions are admissible, any partition has an admissible refinement with a number of triangles which is at most a constant factor larger. Indeed, given a partition $P=P_n$, consider the following algorithm to compute the partition $P^a=P^a_n$.

ALGORITHM 3.6.
$P_0^a := P_0$
`for` $i = 0, \dots, n$ `do`
define P_{i+1}^a *as the union of* P_{i+1} *and, when* $i \leq n-2$*, the collection children of those* $\triangle \in P_i^a$ *that have a neighbor in* P_i *with grandchildren in* P_{i+2}
`od`

PROPOSITION 3.7. *The partition* P^a *produced by Algorithm* 3.6 *is an admissible refinement of* P *with* $\#P^a \lesssim \#P$.

Proof. The criterion to add children of a $\triangle \in P_i^a$ to P_{i+1} with the construction of P_{i+1}^a can only be fulfilled when $\triangle$ is a neighbor of a $\tilde{\triangle} \in P_i$, which was refined when going to P_{i+1}, and thus with $\mathrm{gen}(\tilde{\triangle}) = i$. As we have seen, since $\max_{\triangle \in P_i^a} \mathrm{gen}(\triangle) = i$, the number of neighbors in P_i^a of such a $\tilde{\triangle}$ is uniformly bounded. So defining λ_i or λ_i^a as the number of triangles that were refined when going from P_i to P_{i+1} or from P_i^a to P_{i+1}^a, respectively, we have $\lambda_i^a \lesssim \lambda_i$.

Note that $(P_i^a)_{0 \leq i \leq n}$ corresponds to P^a in the sense of Proposition 3.1. Since each time a triangle in a partition is refined the number of triangles is increased by 3, we conclude that

$$\#P^a = \#P_0 + 3\sum_{i=0}^{n-1} \lambda_i^a \lesssim \#P_0 + 3\sum_{i=0}^{n-1} \lambda_i = \#P.$$

What is left to show is that P^a is admissible. Obviously, the partitions P_0^a and P_1^a are admissible. Suppose that there exists an $1 \leq i \leq n-1$ such that P_i^a is admissible, whereas P_{i+1}^a is not. Then, there exist neighbors $\triangle, \tilde{\triangle} \in P_{i+1}^a$ with $\mathrm{gen}(\triangle) - \mathrm{gen}(\tilde{\triangle}) > 1$. Since P_i^a is admissible, necessarily $\mathrm{gen}(\triangle) = i+1$ and $\mathrm{gen}(\tilde{\triangle}) = i-1$.

If $\triangle \in P_{i+1}$, then $\tilde{\triangle} \in P_{i-1}^a$ has a neighbor in P_{i-1} with grandchildren in P_{i+1}. So, by construction, $\tilde{\triangle}$ would have been refined when going to P_i^a, which gives a contradiction with the assumption that $\tilde{\triangle} \in P_{i-1}^a$.

If $\triangle \in P_{i+1}^a \backslash P_{i+1}$, then by construction, its parent $\triangle_f \in P_i^a$ has a neighbor $\hat{\triangle} \in P_i$ with grandchildren in P_{i+2}, whereas obviously also $\triangle_f \in P_i^a$ and $\tilde{\triangle}$ are neighbors. Let $\triangle_{ff} \in P_{i-1}^a$ and $\hat{\triangle}_f \in P_{i-1}$ denote the parents of $\triangle_f$ and $\hat{\triangle}$, respectively. We are going to show that $\hat{\triangle}_f$ and $\tilde{\triangle}$ are neighbors in P_{i-1}^a, meaning, because $\hat{\triangle}_f$ has grandchildren in P_{i+1}, that $\tilde{\triangle}$ must have been refined when going to P_i^a, resulting in a contradiction with the assumption that $\tilde{\triangle} \in P_{i+1}^a$. We have to distinguish between two cases: If $\triangle_f$ is the central subtriangle of $\triangle_{ff}$, then both $\triangle_{ff}$ and $\hat{\triangle}_f$, and $\triangle_{ff}$ and $\tilde{\triangle}$ share an edge, and so $\hat{\triangle}_f$ and $\tilde{\triangle}$ are neighbors (cf. left picture in Figure 3.2). If $\triangle_f$ is a corner subtriangle of $\triangle_{ff}$, i.e, $\triangle_f$ and $\triangle_{ff}$ share a vertex v, then v is also a vertex of both $\hat{\triangle}_f$ and $\tilde{\triangle}$, again showing that they are neighbors (cf. right picture in Figure 3.2). ∎

With $P_0^* = P_0$, by induction on i we construct the partition P_i^* from P_{i-1}^* by applying a red refinement step to *all* $\triangle \in P_{i-1}^*$, i.e., P_i^* is the result of applying recursively i *uniform refinement steps* to P_0. Note that these definitions are in accordance with Notation 3.2. We define

$$V_* = \cup_{i \geq 0} V_{P_i^*} \backslash V_{P_{i-1}^*};$$

this set contains V_P for any partition $P = P_n$. Obviously, for $0 \leq i \leq n$, $V_{P_i} \subset V_{P_i^*}$, and Proposition 3.1(ii) shows that $V_{P_i} \backslash V_{P_{i-1}} \subset V_{P_i^*} \backslash V_{P_{i-1}^*}$.

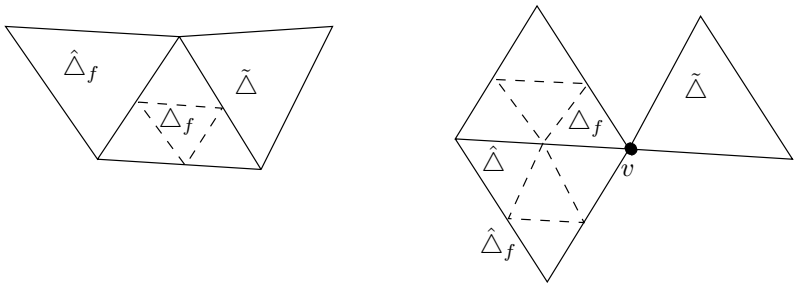


FIG. 3.2. *Illustration with the proof of Proposition* 3.7.

Finally, in this subsection, having defined the initial partition P_0, we are able to formulate all assumptions on the coefficient matrix $\mathbf{A}$ that we will need.

Assumption 3.8. In addition to assuming that $\mathbf{A} \in L_\infty(\Omega)$ is a symmetric 2×2 matrix with $\operatorname{ess\,inf}_{x\in\Omega}\lambda_{\min}(\mathbf{A}(x)) > 0$, we assume that $\mathbf{A} \eqsim \rho(\mathbf{A})\mathbf{id}$, uniformly over the domain (*isotropic diffusion*), and that $\mathbf{A}$ is *piecewise constant with respect to* P_0. Further, following [12], we assume that $\rho = \rho(\mathbf{A})$ is *quasi-monotone* with respect to P_0. That is, defining for $v \in \bar{V}_{P_0}$, $P_0(v) = \{\triangle \in P_0 : v \in \triangle\}$ and $\triangle(v) = \operatorname{argmax}\{\rho|_\triangle : \triangle \in P_0(v)\}$, we assume that for some absolute constant $c > 0$, for all $v \in \bar{V}_{P_0}$ and $\triangle \in P_0(v)$ there exist $\triangle = \triangle_1, \dots, \triangle_m = \triangle(v)$ such that $\triangle_i$ shares an edge with $\triangle_{i+1}$ and $\rho|_{\triangle_i} \le c\rho|_{\triangle_{i+1}}$. Moreover, if $v \in \bar{V}_{P_0}\backslash V_{P^0}$, then in addition we assume that there exists a $\triangle \in P_0(v)$ having an edge on $\partial\Omega$ such that $\rho|_{\triangle(v)} \le c\rho|_\triangle$. Under these assumptions, all of our results that depend on $\mathbf{A}$ should be interpreted as ones that hold *uniformly in* $(\rho|_\triangle)_{\triangle\in P_0}$.

4. Finite element spaces and bases, and Galerkin approximations. For a given partition P, let $\mathcal{S}_P \subset H^1_0(\Omega)$ denote the space of continuous, piecewise linear functions subordinate to P which vanish at $\partial\Omega$. The solution $u_P \in \mathcal{S}_P$ of

$$a(u_P, w_P) = f(w_P) \quad (w_P \in \mathcal{S}_P), \tag{4.1}$$

is called the *Galerkin approximation* of the solution u of (2.1). Defining $L_P : \mathcal{S}_P \to (\mathcal{S}_P)' \supset H^{-1}(\Omega)$ by $(L_P u_P)(w_P) = a(u_P, w_P)$, the solution of (4.1) is $L_P^{-1} f$.

On some places we will replace the right-hand side f by some approximation from $\mathcal{S}^0_P$, being defined as the space of functions that are piecewise constant with respect to P.

If $\tilde{P}$ is a refinement of P, then $\mathcal{S}_P \subset \mathcal{S}_{\tilde{P}}$ and $\mathcal{S}^0_P \subset \mathcal{S}^0_{\tilde{P}}$. Each $u \in \mathcal{S}_P$ is uniquely determined by its values on V_P, and so, in particular, $\#V_P = \dim\mathcal{S}_P$. Defining, for $v \in V_P$, $\phi^v_P \in \mathcal{S}_P$ by

$$\phi^v_P(\tilde{v}) = \begin{cases} 1, & v = \tilde{v}, \\ 0, & v \ne \tilde{v} \in V_P, \end{cases}$$

the set

$$\{\phi^v_P : v \in V_P\}$$

is a basis for $\mathcal{S}_P$, called the *nodal basis*.

One easily verifies the following result.

LEMMA 4.1. *Let v be a regular vertex of a partition P. Then, with $i := \mathrm{gen}(\triangle)$ for any (and thus all) $\triangle \in P$ that contain v, we have that $\phi^v_{P^*_i} \in \mathcal{S}_P$.*

PROPOSITION 4.2. *For any partition $P = P_n$, $\cup_{i=0}^n \{\phi^v_{P^*_i} : v \in V_{P_i} \backslash V_{P_{i-1}}\}$ is a basis for $\mathcal{S}_P$, called* hierarchical basis.

Proof. Proposition 3.1(ii) and Lemma 4.1 show that for $v \in V_{P_i} \backslash V_{P_{i-1}}$, $\phi^v_{P^*_i} \in \mathcal{S}_{P_i} \subset \mathcal{S}_P$. Since $\phi^v_{P^*_i}$ vanishes on $V_{P_{i-1}}$, by induction on i we conclude that for given scalars $(d_v)_{v \in V_P}$ the interpolation problem of finding scalars $(c_v)_{v \in V_P}$ with $\sum_{i=0}^n \sum_{v \in V_{P_i} \backslash V_{P_{i-1}}} c_v \phi^v_{P^*_i}(\tilde{v}) = d_{\tilde{v}}$ $(\tilde{v} \in V_P)$ has a unique solution. Since $\dim \mathcal{S}_P = \#V_P = \sum_{i=0}^n \#(V_{P_i} \backslash V_{P_{i-1}})$ by Proposition 3.1(i), the proof is completed. ∎

Besides the nodal and hierarchical bases, for admissible partitions P^a we introduce another basis for $\mathcal{S}_{P^a}$, which, as we shall see, is appropriately called a *wavelet basis*.

Let $v \in V_*$; then there exists a unique $i \in \mathbb{N}$ such that $v \in V_{P^*_i} \backslash V_{P^*_{i-1}}$. When $i > 0$, v is the midpoint of the common edge of two triangles $\triangle_1, \triangle_2 \in P^*_{i-1}$. Let us denote by $v_1(v), \ldots, v_4(v)$ the vertices of these $\triangle_1, \triangle_2$, with $v_2(v), v_3(v)$ being the vertices on the edge containing v; see Figure 4.1. For some scalars $\mu_{v,j}$ which will be specified later on, with $\mu_{v,j} := 0$ when $v_j(v) \in \partial\Omega$, we define

$$\psi^v := \phi^v_{P^*_i} - \sum_{j=1}^4 \mu_{v,j} \phi^{v_j(v)}_{P^*_{i-1}}, \tag{4.2}$$

and, for convenience, for $v \in V_{P^*_0}$ we set $\psi^v = \phi^v_{P^*_0}$.

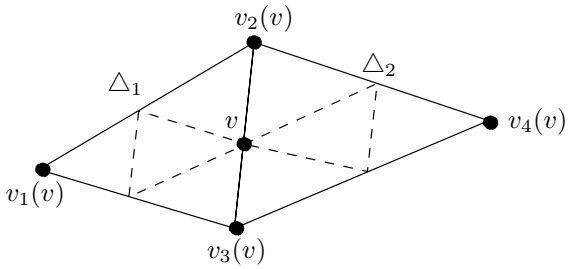


FIG. 4.1. *Definition of $v_j(v)$.*

PROPOSITION 4.3. *If P^a is* admissible, *then*

$$\{\psi^v : v \in V_{P^a}\} \tag{4.3}$$

is a basis for $\mathcal{S}_{P^a}$.

Proof. Obviously $\{\psi^v : v \in V_{P_0}\}$ is a basis for $\mathcal{S}_{P_0}$. Assuming that $\{\psi^v : v \in V_{P^a_{n-1}}\}$ is a basis for $\mathcal{S}_{P^a_{n-1}}$, the same argument as the one that was applied in the proof of Proposition 4.2 shows that $\{\psi^v : v \in V_{P^a_{n-1}}\} \cup \{\phi^v_{P^*_n} : v \in V_{P^a_n} \backslash V_{P^a_{n-1}}\}$ is a basis for $\mathcal{S}_{P^a_n}$. Proposition 3.1(ii) shows that each $v \in V_{P^a_n} \backslash V_{P^a_{n-1}}$ is the midpoint of $\triangle_1, \triangle_2 \in P^a_{n-1}$ with $\mathrm{gen}(\triangle_1) = \mathrm{gen}(\triangle_2) = n-1$, both of which are refined in the transition to P^a_n. Proposition 3.4 shows that each of $v_1(v), \ldots, v_4(v)$, which are the vertices of $\triangle_1$ and $\triangle_2$, is a regular vertex of P^a_{n-1}, obviously with $\mathrm{gen}(\hat{\triangle}) = n-1$ for any and thus all $\hat{\triangle} \in P^a_{n-1}$ which contain this vertex. From Lemma 4.1 we now conclude that $\sum_{j=1}^4 \mu_{v,j} \phi^{v_j(v)}_{P^*_{i-1}} \in \mathcal{S}_{P^a_{n-1}}$, from which it follows that also $\{\psi^v : v \in V_{P^a_n}\}$ is a basis for $\mathcal{S}_{P^a_n}$. ∎

For $P^a = P^a_n$ and $w_{P^a} \in \mathcal{S}_{P^a}$ given by its values $(w_{P^a}(v))_{v \in \bar{V}_{P^a}}$, i.e., the coefficients of its representation with respect to the nodal basis, where $w_{P^a}(v) := 0$ when

$v \in \bar{V}_{P^a} \backslash V_{P^a}$, an application of the following routine yields the coefficients $(c_v)_{v \in V_{P^a}}$ of its representation with respect to the wavelet basis (4.3).

ALGORITHM 4.4.

$c_v := w_{P^a}(v)$ *($v \in \bar{V}_{P^a}$)*
`for` $i = n, \dots, 1$ `do`
  $c_v := c_v - \frac{1}{2}(c_{v_2(v)} + c_{v_3(v)})$ *($v \in V_{P^a_i} \backslash V_{P^a_{i-1}}$)*
  $c_{v_j(v)} := c_{v_j(v)} + c_v \mu_{v,j}$ *($v \in V_{P^a_i} \backslash V_{P^a_{i-1}}$, $1 \le j \le 4$)*
`od`

Conversely, if $w_{P^a} \in \mathcal{S}_{P^a}$ is given by its coefficients $(c_v)_{v \in \bar{V}_{P^a}}$ with respect to the wavelet basis (4.3), where $c_v := 0$ when $v \in \bar{V}_{P^a} \backslash V_{P^a}$, then the values $(w_{P^a}(v))_{v \in V_{P^a}}$ are obtained from the following routine.

ALGORITHM 4.5.

`for` $i = 1, \dots, n$ `do`
  $c_{v_j(v)} := c_{v_j(v)} - c_v \mu_{v,j}$ *($v \in V_{P^a_i} \backslash V_{P^a_{i-1}}$, $1 \le j \le 4$)*
  $c_v := c_v + \frac{1}{2}(c_{v_2(v)} + c_{v_3(v)})$ *($v \in V_{P^a_i} \backslash V_{P^a_{i-1}}$)*
`od`
$w_{P^a}(v) := c_v$ *($v \in V_{P^a}$)*

One directly infers the following result.

PROPOSITION 4.6. *Both of the above algorithms for switching between the representation of a $w \in \mathcal{S}_{P^a}$ in terms of the nodal basis to its representation in terms of* (4.3)*, and vice versa, take $\mathcal{O}(\dim \mathcal{S}_{P^a})$ operations.*

Now we come to the specification of the coefficients $\mu_{v,j}$ from (4.2). We take

$$\mu_{v,j} = \frac{3(\rho|_{\triangle_1}\mathrm{meas}(\triangle_1) + \rho|_{\triangle_2}\mathrm{meas}(\triangle_2))}{8\sum_{\{\triangle \in P^*_{i-1}, v_j(v) \in \triangle\}} \rho|_{\triangle}\mathrm{meas}(\triangle)} \text{ when } j \in \{2,3\} \text{ and } v_j(v) \notin \partial\Omega, \tag{4.4}$$

and $\mu_{v,j} = 0$ otherwise. An alternative choice of the coefficients will be discussed later, in Remark 4.9. When both $v_2(v), v_3(v) \notin \partial\Omega$, a simple calculation reveals that, for $i \ge 1$ and $v \in V_{P^*_i} \backslash V_{P^*_{i-1}}$, $\int_\Omega \rho \psi^v = 0$, so that it is appropriate to call ψ^v a *wavelet*.

For coefficient matrices $\mathbf{A}$ that satisfy Assumption 3.8, a combination of results from [19, 18, 12] shows the following result.

THEOREM 4.7. *With $\bar{\psi}^v := \psi^v / |\psi^v|_1$,*

$$\{\bar{\psi}^v : v \in V_*\} \tag{4.5}$$

is a Riesz basis for $H^1_0(\Omega)$ equipped with $|\cdot|_1$. For the sake of completeness we emphasize that this result is valid uniformly in $(\rho|_{\triangle})_{\triangle \in P_0}$.

Defining $\|\!|w|\!\|_1 = (\sum_{v \in V_*} \bar{c}_v^2)^{\frac{1}{2}}$, where $w = \sum_{v \in V_*} \bar{c}_v \bar{\psi}^v$ is the unique expansion of $w \in H^1_0(\Omega)$, with $\lambda_{\bar{\Psi}}, \Lambda_{\bar{\Psi}} > 0$ we will denote the largest or smallest constant for which

$$\lambda_{\bar{\Psi}} \|\!| \cdot |\!\|_1^2 \le |\cdot|_1^2 \le \Lambda_{\bar{\Psi}} \|\!| \cdot |\!\|_1^2 \tag{4.6}$$

and $\kappa_{\bar{\Psi}} := \frac{\Lambda_{\bar{\Psi}}}{\lambda_{\bar{\Psi}}}$. We note that this equivalence between the $|\cdot|_1$-norm of a function in $H^1_0(\Omega)$ and the ℓ_2-norm of its, generally infinite, coefficient vector holds, in particular, for functions from $\mathcal{S}_{P^a}$ for admissible P^a, which by Proposition 4.3 have a finite wavelet expansion.

Remark 4.8. Since the proof of Theorem 4.7 can only be deduced by combining results from different papers, we briefly comment on its derivation. For any $\triangle \in \cup_{i \ge 0} P^*_i$, let $I_\triangle : C(\triangle) \to P_1(\triangle)$ be the nodal value interpolant, consider the bilinear

form $\langle\!\langle u, w\rangle\!\rangle_\triangle := \frac{1}{3}\mathrm{meas}(\triangle)\cdot\sum_{v \text{ vertex of } \triangle} u(v)w(v)$, and let

$$\langle u, w\rangle_\triangle := \langle\!\langle I_\triangle u, I_\triangle w\rangle\!\rangle_\triangle + \sum_{k=1}^{4}\left[\langle\!\langle u, w\rangle\!\rangle_{\triangle_k} - \langle\!\langle I_\triangle u, I_\triangle w\rangle\!\rangle_{\triangle_k}\right],$$

where $\triangle_1, \ldots, \triangle_4$ are the children of $\triangle$. Note that for $u, w \in P_1(\triangle)$, $\langle u, w\rangle_\triangle = \langle\!\langle u, w\rangle\!\rangle_\triangle$ which is a scalar product on $P_1(\triangle)$. Let $Y_\triangle : C(\triangle)\cap\prod_{k=1}^4 P_1(\triangle_k) \to P_1(\triangle)$ be the orthogonal projector with respect to $\langle\cdot,\cdot\rangle_\triangle$. Finally, for $i \geq 1$, on $\mathcal{S}_{P_i^*}\times\mathcal{S}_{P_i^*}$ let

$$\langle u, w\rangle_{\mathcal{S}_{P_i^*}} := \sum_{\triangle\in P_{i-1}^*} \rho|_\triangle \langle u, w\rangle_\triangle.$$

Using that $\{\phi^v_{P_{i-1}^*} : v \in V_{P_{i-1}^*}\}$ is an orthogonal set with respect to $\langle\cdot,\cdot\rangle_{\mathcal{S}_{P_i^*}}$, in [18, section 6] it was shown that for $i \geq 1$ the sets

$$\{\psi^v/\|\psi^v\|_0 : v \in V_{P_i^*}\backslash V_{P_{i-1}^*}\},$$

defined by (4.2), (4.4), are uniform Riesz bases for $\mathcal{S}_{P_i^*}\cap\mathcal{S}_{P_{i-1}^*}^{\perp_{\langle\,,\,\rangle_{\mathcal{S}_{P_i^*}}}}$ equipped with $\|\cdot\|_0$, where "uniform" refers both to the parameter i and to $(\rho|_\triangle)_{\triangle\in P_0}$.

With

$$t := \sup_{0\neq w\in C(\triangle)\cap\prod_{k=1}^4 P_1(\triangle_k)} \frac{\langle w, w\rangle_\triangle - \langle (I-Y_\triangle)w, (I-Y_\triangle)w\rangle_\triangle}{\sum_{k=1}^4 \langle\!\langle w, w\rangle\!\rangle_{\triangle_k}},$$

whose value is independent of the triangle $\triangle$, it follows from [19, Thm. 3.1 and (4.7)] that in case of *constant* $\mathbf{A} = \mathbf{id}$, for $\frac{3}{2} > s > \log_2\sqrt{t}$ the infinite collection

$$\cup_{i\geq 0}\{2^{(s-1)i}\psi^v : v \in V_{P_i^*}\backslash V_{P_{i-1}^*}\}$$

is a Riesz basis for $H_0^s(\Omega)$. Some calculations show that $t = \frac{181}{64}$, and so $\sqrt{t} \approx .7499$, meaning in particular that (4.5) is a Riesz basis for $H_0^1(\Omega)$.

With $Q_i : L_2(\Omega) \to \mathcal{S}_{P_i^*}$ being the $\langle\,,\,\rangle_0$-orthogonal projector onto $\mathcal{S}_{P_i^*}$, and $Q_{-1} := 0$, for coefficient matrices $\mathbf{A}$ that satisfy Assumption 3.8, it was shown in [12] that

$$\|w\|_1^2 \eqsim \sum_{i=0}^{\infty} 4^i\|(Q_i - Q_{i-1})w\|_0^2 \quad (w \in H_0^1(\Omega)),$$

uniformly in $(\rho|_\triangle)_{\triangle\in P_0}$. As was shown in [19], from this result and the fact that $t < 1$ it even follows that (4.5) is a Riesz basis for $H_0^1(\Omega)$ equipped with $|\cdot|_1$, uniformly in $(\rho|_\triangle)_{\triangle\in P_0}$.

Remark 4.9. For *constant* $\mathbf{A} = \mathbf{id}$, in [8] other values for the coefficients $\mu_{v,1}, \ldots, \mu_{v,4}$ from (4.2) were proposed; generally, all four of these coefficients are nonzero. Although uniform refinements of an arbitrary initial partition are considered, just as outlined above, for admissible $P^a = P_n^a$ a subset of the wavelet basis for $\mathcal{S}_{P_n^*}$ spans $\mathcal{S}_{P^a}$. For some $\tilde{s} > 0$ and $s \in (-\tilde{s}, \frac{3}{2})$, the infinite collection of properly scaled wavelets from [8] is shown to generate a Riesz basis for

$$\mathcal{H}^s(\Omega) = \begin{cases} H_0^s(\Omega) & \text{when } s \geq 0, \\ (H_0^{-s}(\Omega))' & \text{when } s < 0. \end{cases}$$

On regular "type-I triangulations" of the whole of $\mathbb{R}^2$, the wavelet proposals from [18, sect. 6] or [8] reduce to the so-called coarse-grid stabilized HB-systems from [13] with parameters $a = \frac{1}{8}$ or $a = -\frac{3}{16}$, respectively. There, it is shown that for these uniform triangulations the exact $H^s(\mathbb{R}^2)$ stability ranges are $s \in (0.022818, \frac{3}{2})$ and $s \in (-0.440765, \frac{3}{2})$, respectively (cf. also [9]).

Finally, numerical results [13, Table 1.2] show that both wavelet bases are also quantitatively well conditioned ($\kappa_{\tilde{\Psi}}$ is approximately 16 or 10, respectively, for $\mathbf{A} = \mathbf{id}$ and P_0 being the standard regular partition of the unit square into 8 triangles, so that $\#V_{P^0} = 1$).

5. A coarsening algorithm. In [6], a coarsening step was introduced into an adaptive algorithm in the framework of a wavelet method. The idea of such a step, which has to be applied after each fixed number of iterations that produce increasingly more accurate approximations, is to remove a possibly large number of small terms in the current approximation that hardly contribute to its quality but which, because of their number, spoil the complexity. With wavelet methods such "small terms" stand for terms in a wavelet expansion with small coefficients, and in our finite element setting they correspond to a representation of the approximation as a piecewise linear function subordinate to a locally fine partition, whereas it is close to being linear on the union of these triangles.

Given some current approximation defined on some partition, in order to find a more efficient representation without increasing the error too much, one cannot simply join arbitrary collections of triangles since, generally, their union will not be a triangle. Instead one can only join groups of all siblings of one parent, that is, one has to respect the underlying tree structure. In view of this, in [3], for each triangle in the tree associated to the partition, an error functional was defined. It was shown that, for any subtree, the ℓ_2-norm of these error functionals over the leaves is bounded by some multiple of the error of the best continuous piecewise linear approximation subordinate to the partition defined by this subtree. Giving a tolerance that one allows to be added to the current error, a tree-coarsening algorithm from [5] was run, which, modulo some constant factor, yields the smallest subtree for which the above ℓ_2-norm is less than this tolerance. The ℓ_2-norm could not be shown to be equivalent to the $|\cdot|_1$-norm of the error in the best approximation. Therefore, this procedure had to be completed by a number of uniform refinement steps.

In this paper, for the different type of partitions we consider, based on ideas from [3, 5] an alternative coarsening procedure is developed, which we hope is more attractive for practical computations. Given a current approximation from $\mathcal{S}_P$, in case P is not admissible, we will first embed it into $\mathcal{S}_{P^a}$, where P^a is constructed using Algorithm 3.6. Next, we determine its finite set of wavelet coefficients. Now using the norm equivalence (4.6), an obvious coarsening procedure would be just to order these coefficients by their modulus, and then to remove coefficients, starting with the smallest one, until the tolerance is met. Yet, the task is not to find an approximation with a minimum number of wavelet coefficients, but to find an approximation from a finite element space subordinate to a partition that has, modulo some constant factor, a minimum number of triangles, and the suggested procedure will generally fail to deliver this. Therefore we will equip the infinite index set V_* of all wavelets with a tree structure, and run the algorithm from [5] to find a subtree approximation at a distance less than or equal to the tolerance, which has, modulo some constant factor, a minimum number of terms. Since the tree structure will be designed so that for any subtree $\tilde{V}$ we can construct a partition P such that $\operatorname{span}\{\psi^v : v \in \tilde{V}\} \subset \mathcal{S}_P$, where

$\#P \lesssim \#\tilde{V}$, we will be able to conclude that we found an approximation subordinate to a partition that has, modulo some constant factor, a minimum number of triangles. An advantage of our coarsening procedure will be that it not only gives a (quasi-) optimal partition, but that, at the same time, it also yields a (quasi-) optimal continuous piecewise linear approximation subordinate to this partition.

The tree structure with which we equip V_* is defined as follows. The vertices from $V_{P_0^*}$ are the roots of the tree. For $i = 1$ and $v \in V_{P_i^*} \backslash V_{P_{i-1}^*}$, we assume that at least one of the vertices $v_1(v), \ldots, v_4(v)$ is not on $\partial\Omega$, and we just pick one of them to be the parent of v. Now let $i > 1$ and $v \in V_{P_i^*} \backslash V_{P_{i-1}^*}$. At least one of $v_1(v), \ldots, v_4(v)$ is in $V_{P_{i-1}^*} \backslash V_{P_{i-2}^*}$, and we may pick just one of them to be the parent of v. In the case of multiple choices, a deterministic rule to make the selection is given, for example, by the following: If one of $v_2(v)$ or $v_3(v)$ is in $\bar{V}_{P_{i-2}^*}$, then the other is in $V_{P_{i-1}^*} \backslash V_{P_{i-2}^*}$, and we define it to be the parent of v. Otherwise, if both $v_2(v), v_3(v) \in \bar{V}_{P_{i-1}^*} \backslash \bar{V}_{P_{i-2}^*}$, then one of the remaining $v_1(v), v_4(v)$ is in $\bar{V}_{P_{i-2}^*}$, and we call it $v_1(v)$, whereas the other, thus called $v_4(v)$, is in $\bar{V}_{P_{i-1}^*} \backslash \bar{V}_{P_{i-2}^*}$. After numbering $v_1(v), v_2(v), v_3(v)$ in a clockwise direction, we select the first of $v_2(v), v_3(v), v_4(v) \in V_{P_{i-1}^*} \backslash V_{P_{i-2}^*}$ to be the parent of v. The number of children of any parent in this tree is uniformly bounded and is only dependent on P_0.

For $w \in H_0^1(\Omega)$, let $w = \sum_{v \in V_*} \bar{c}_v \bar{\psi}^v$ be its expansion with respect to (4.5). Obviously, for any $\tilde{V} \subset V_*$, its best approximation with respect to $\|\!|\cdot|\!\|_1$ from $\operatorname{span}\{\bar{\psi}^v : v \in \tilde{V}\}$ is $\sum_{v \in \tilde{V}} \bar{c}_v \bar{\psi}^v$. The squared error $E(\tilde{V})$ of this approximation with respect to $\|\!|\cdot|\!\|_1$ is $E(\tilde{V}) = \sum_{v \in V_* \backslash \tilde{V}} |\bar{c}_v|^2$.

We will call $\tilde{V} \subset V_*$ a *subtree* when it contains all roots $V_{P_0^*}$, and when, for any $v \in \tilde{V}$, all its ancestors and all its siblings, i.e., those $w \in V_*$ that have the same parent as v, are also in $\tilde{V}$. The set of leaves $\mathcal{L}(\tilde{V})$ is defined as the set of those $v \in \tilde{V}$ which have no children in $\tilde{V}$. Defining for $v \in V_*$ the error functional $e(v) := \sum_{\bar{v} \text{ a descendant of } v} |\bar{c}_{\bar{v}}|^2$, we have $E(\tilde{V}) = \sum_{v \in \mathcal{L}(\tilde{V})} e(v)$.

Following [5], we define a modified error functional $\tilde{e}(v)$ for $v \in V_*$ as follows. For the roots $v \in V_{P_0^*}$, $\tilde{e}(v) := e(v)$. Assuming that $\tilde{e}(v)$ has been defined, then for all of its children $v_1, \ldots, v_m$,

$$\tilde{e}(v_j) := \frac{\sum_{i=1}^m e(v_i)}{e(v) + \tilde{e}(v)} \tilde{e}(v).$$

Now given a $w \in H_0^1(\Omega)$ and a tolerance $\varepsilon > 0$, the *thresholding second algorithm* from [5] for determining a quasi-optimal *subtree approximation* runs as follows.

ALGORITHM 5.1.

$\tilde{V} := V_{P_0^*}$
`while` $E(\tilde{V}) > \varepsilon^2$ `do`
    compute $\rho = \max_{v \in \mathcal{L}(\tilde{V})} \tilde{e}(v)$
    `forall` $v \in \mathcal{L}(\tilde{V})$ *with* $\tilde{e}(v) = \rho$ `do` *add all children of* v *to* $\tilde{V}$ `od`
`od`

Remark 5.2. During the evaluation of Algorithm 5.1, the values $\tilde{e}(v)$ for the current leaves should be stored as an ordered list. As a consequence, with $\tilde{V}$ being the subtree at termination, the operation count of Algorithm 5.1 will contain a term $\mathcal{O}(\#\tilde{V} \log((\#\tilde{V}))$ due to the insertions of $\tilde{e}(v)$ for newly created leaves in this list. Since the other costs of the algorithm are $\mathcal{O}(\#\tilde{V})$, asymptotically the cost of these insertions will dominate. Although it seems unlikely that this will happen with practical problem sizes, for mathematical completeness we sketch here a modification with

which the log-factor is avoided (see [1, 14, 20] for solutions of a similar problem in a wavelet context).

Noting that $\tilde{e}(v_j) \le \sum_{i=1}^m e(v_i) \le e(v) \le \|w\|_1^2$, we may store the current leaves v in binary bins $V_0, \dots, V_q$, where for $0 \le i \le q-1$, V_i contains those v with $\tilde{e}(v) \in (2^{-(i+1)}\|w\|_1^2, 2^{-i}\|w\|_1^2]$, and the remaining v, thus with $\tilde{e}(v) \le 2^{-q}\|w\|_1^2$, are put into V_q. Instead of replacing all $v \in \mathcal{L}(\tilde{V})$ with maximal $\tilde{e}(v)$ by their children, in each iteration of the while-loop we replace just one v taken from the first nonempty bin by its children. Again, from $\sum_{i=1}^m e(v_i) \le e(v)$ we have $\tilde{e}(v_j) \le \tilde{e}(v)$ meaning that for any i, once $V_0, \dots, V_i$ got empty they will remain empty.

If q is chosen so that during the iteration we only extract v from bins V_i with $i < q$, then the corresponding $\tilde{e}(v)$ will be at most a factor 2 smaller than the current maximal value of $\tilde{e}$. As a consequence, one may verify that, with the exception of the operation-counts, all results proven in [5] about Algorithm 5.1 are also valid for this modified version (making use of the property $\sum_{i=1}^m e(v_i) \le e(v)$, which is stronger than the assumption made in [5]; only (5.13) of [5] has to be adapted).

With $\tilde{V}$ being the subtree at termination, the number of operations required by this modified Algorithm 5.1 is $\lesssim \#\tilde{V} + q$, where q is the maximum number of bins that have to be generated or inspected for containing leaves. Thinking of the situation when in the course of the iteration the maximum value of $\tilde{e}$ over the leaves varies largely in size, note that, generally, q cannot be bounded in terms of $\#\tilde{V}$.

We will apply the modified Algorithm 5.1 only in the situation when there exists a finite subtree $\bar{V} \subset V_*$ such that

$$e(v) = 0 \qquad (v \in \mathcal{L}(\bar{V})), \tag{5.1}$$

which allows us to make a suitable choice for q. Note that the subtree $\tilde{V}$ at termination satisfies $\tilde{V} \subset \bar{V}$. The definition of $\tilde{e}(v)$ shows that $\sum_{j=1}^m \frac{e(v_j)}{\tilde{e}(v_j)} = 1 + \frac{e(v)}{\tilde{e}(v)}$. A recursive application of this formula gives $\sum_{v \in \mathcal{L}(\tilde{V})} \frac{e(v)}{\tilde{e}(v)} = \#(\tilde{V} \backslash \mathcal{L}(\tilde{V})) + \#V_{P_0^*} \le \#\tilde{V} \le \#\bar{V}$, and so $E(\tilde{V}) \le \#\bar{V} \max_{v \in \mathcal{L}(\tilde{V})} \tilde{e}(v)$. We conclude that the modified Algorithm 5.1 terminates before any leaf v with $\tilde{e}(v) \le \varepsilon^2/\#\bar{V}$ is replaced by its children. Solving for the smallest $q \in \mathbb{N}_0$ with $2^{-q}\|w\|_1^2 = \varepsilon^2/\#\bar{V}$ yields $q = \max\{0, \lceil \log_2(\varepsilon^{-2}\|w\|_1^2 \#\bar{V})\rceil\}$; such a q thus satisfies the assumption we made earlier.

The analysis of Algorithm 5.1 from [5], together with the additions from the above remark concerning our slight modification, yields the following result.

PROPOSITION 5.3 (see [5, Corollary 5.3]). *The subtree $\tilde{V}$ yielded by (the modified) Algorithm* 5.1 *satisfies $E(\tilde{V}) \le \varepsilon^2$. There exists absolute constants $t_1, T_2 > 0$, necessary with $t_1 \le 1 \le T_2$, such that if $\hat{V}$ is a subtree with $E(\hat{V}) \le t_1\varepsilon^2$, then $\#\tilde{V} \le T_2\#\hat{V}$. The number of evaluations of e and the number of additional arithmetic operations required by the modified Algorithm* 5.1 *are $\lesssim \#\tilde{V} + \max\{0, \log(\varepsilon^{-2}\|w\|_1^2\#\bar{V})\}$, with $\bar{V}$ being any subtree satisfying* (5.1).

We are now almost ready to define our coarsening routine. Inside this routine, for some admissible partition P^a we will apply the (modified) Algorithm 5.1 to a $w_{P^a} \in \mathcal{S}_{P^a}$. Such a w_{P^a} has an expansion $w_{P^a} = \sum_{v \in V_{P^a}} \bar{c}_v \bar{\psi}^v$, i.e., $\bar{c}_v = 0$ for $v \in V_* \backslash V_{P^a}$. Although V_{P^a} is nearly a subtree since it contains the roots V_{P_0} as well as all ancestors of any $v \in V_{P^a}$, it may contain $v \in V_{P^a}$ with siblings outside V_{P^a}. As a consequence, the (modified) Algorithm 5.1 may output a subtree $\tilde{V}$ containing such siblings. Yet, since the corresponding wavelet coefficients are zero, these siblings do not contribute to the approximation and can therefore be discarded. An efficient way of implementing this is to call the algorithm pretending that V_* is V_{P^a} so that

siblings outside V_{P^a} will never be created. The set $\bar{V}$ from (5.1) can be taken to be equal to V_{P^a}.

COARSE$[P, w_P, \varepsilon] \to [\tilde{P}^a, w_{\tilde{P}^a}]$:

% $P = P_n$ is some admissible partition, $w_P \in \mathcal{S}_P$ is given by its values $(w_P(v))_{v\in V_P}$
% and $\varepsilon > 0$. The output $\tilde{P}^a$ is an admissible partition, and $w_{\tilde{P}^a} \in \mathcal{S}_{\tilde{P}^a}$ is given by
% its values $(w_{\tilde{P}^a}(v))_{v\in V_{\tilde{P}^a}}$.

(i) *When P is admissible, let $P^a = P$; otherwise compute an admissible refinement P^a of P by applying Algorithm* 3.6. *Compute the values $(w_{P^a}(v))_{v\in V_{P^a}}$ of $w_{P^a} := w_P$.*

(ii) *Compute the wavelet coefficients $(\bar{c}_v)_{v\in V_{P^a}}$ of w_{P^a} using Algorithm* 4.4.

(iii) *Compute recursively $(e(v))_{v\in V_{P^a}}$ starting with each $v \in V_{P^0}$ as follows:*
`if` *v has no children in V_{P^a}* `then` $e(v) := 0$
`else` $e(v) := \sum_{\{\tilde{v}\in V_{P^a}:\tilde{v} \text{ is a child of } v\}} e(\tilde{v}) + \bar{c}_{\tilde{v}}^2$ `fi`

(iv) *Apply the modified Algorithm* 5.1, *with* $q := \max\{0, \lceil \log_2(\varepsilon^{-2}\|w_{P^a}\|_1^2 \# V_{P^a})\rceil\}$, *yielding a set $\tilde{V} \subset V_{P^a}$.*

(v) *Determine a partition $\tilde{P}$ as follows:*
$\tilde{P}_0 := P_0$, $i := 1$
`while` $\tilde{V} \cap (V_{P_i^*} \backslash V_{P_{i-1}^*}) \neq \emptyset$ `do`
construct $\tilde{P}_i$ from $\tilde{P}_{i-1}$ by refining those $\triangle \in \tilde{P}_{i-1}$ that contain a $v \in \tilde{V} \cap (V_{P_i^*} \backslash V_{P_{i-1}^*})$, $i{:=}i{+}1$
`od`

(vi) *Apply Algorithm* 3.6 *to determine an admissible refinement $\tilde{P}^a$ of $\tilde{P}$. Note that $\tilde{V} \subset V_{\tilde{P}} \subset V_{\tilde{P}^a}$, and thus that $w_{\tilde{P}^a} := \sum_{v\in\tilde{V}} \bar{c}_v \bar{\psi}^v \in \mathcal{S}_{\tilde{P}^a}$.*

(vii) *Apply Algorithm* 4.5 *to compute $(w_{\tilde{P}^a}(v))_{v\in V_{\tilde{P}^a}}$.*

THEOREM 5.4. (a) $[\tilde{P}^a, w_{\tilde{P}^a}] :=$ **COARSE**$[P, w_P, \varepsilon]$ *satisfies $\|w_P - w_{\tilde{P}^a}\|_1 \le \varepsilon$. There exists an absolute constant $D > 0$, such that for any partition $\hat{P}$ for which there exists a $w_{\hat{P}} \in \mathcal{S}_{\hat{P}}$ with $\|w_P - w_{\hat{P}}\|_1 \le t_1^{\frac{1}{2}}\varepsilon$, we have that $\#\tilde{P}^a \le D\#\hat{P}$.*

(b) *The call requires $\lesssim \#P + \max\{0, \log(\varepsilon^{-1}\|w_P\|_1)\}$ arithmetic operations.*

Proof. (a) The first statement follows by construction. Let $(\bar{c}_v)_{v\in V_*}$ be the wavelet coefficients in the expansion $w_P = \sum_{v\in V_*} \bar{c}_v\bar{\psi}_v$. Let $\hat{P}$, $w_{\hat{P}} \in \mathcal{S}_{\hat{P}}$ with $\|w_P - w_{\hat{P}}\|_1 \le t_1^{\frac{1}{2}}\varepsilon$. Let $\hat{P}^a$ be the admissible refinement of $\hat{P}$ constructed by applying Algorithm 3.6, and let $(\bar{c}_v)_{v\in V_{\hat{P}^a}}$ be the wavelet coefficients of $w_{\hat{P}} \in \mathcal{S}_{\hat{P}} \subset \mathcal{S}_{\hat{P}^a}$. Let $\hat{V}$ be the enlargement of $V_{\hat{P}^a}$ by adding all siblings of all $v \in V_{\hat{P}^a}$ to this set. Then $\hat{V}$ is a subtree with $\#\hat{V} \lesssim \#V_{\hat{P}^a} \lesssim \#\hat{P}^a \lesssim \#\hat{P}$. Because of

$$E(\hat{V}) = \sum_{v\in V_*\backslash\hat{V}} |\bar{c}_v|^2 \le \sum_{v\in V_*} |\bar{c}_v - \bar{c}_v|^2 = \|w_P - \hat{w}_P\|_1^2 \le t_1\varepsilon^2,$$

an application of Proposition 5.3 shows that $\tilde{V}$ constructed in (iv) satisfies $\#\tilde{V} \le T_2\#\hat{V}$. The proof is completed by noting that the partitions $\tilde{P}$ and $\tilde{P}^a$ constructed in (v) and (vi) satisfy $\#\tilde{P}^a \lesssim \#\tilde{P} \lesssim \#\tilde{V}$.

(b) The number of arithmetic operations required by (i)–(iii) is $\lesssim \#P^a \lesssim \#P$. Since $\#V_{P^a} \lesssim \#P^a \lesssim \#P$, Proposition 5.3 shows that (iv) requires $\lesssim \#\tilde{V} + \max\{0, \log(\varepsilon^{-2}\|w_P\|_1^2\#V_{P^a})\} \lesssim \#P + \max\{0, \log(\varepsilon^{-1}\|w_P\|_1)\}$ arithmetic operations. The number of arithmetic operations required by (v)–(vii) is $\lesssim \#\tilde{V} \lesssim \#P$. ∎

Had the unmodified Algorithm 5.1 been applied inside **COARSE**, the required number of arithmetic operations would have been $\lesssim \#P\log(\#P)$. In contrast with such a log-factor, for our application it will turn out that the log-*term* from

Theorem 5.4(b) is completely harmless. The next corollary shows that if for $u \in H^1_0(\Omega)$ and $s > 0$ the errors of the best approximations from any $\mathcal{S}_P$ with $\#P \leq n$ are $\mathcal{O}(n^{-s})$, then given any $\varepsilon > 0$, a partition P and a $w_P \in \mathcal{S}_P$ with $|u - w_P|_1 \leq \varepsilon$, by allowing the tolerance to increase by some suitable, sufficiently large constant factor, the coarsening procedure yields an (admissible) partition $\tilde{P}^a$ and a $w_{\tilde{P}^a} \in \mathcal{S}_{\tilde{P}^a}$ with $|u - w_{\tilde{P}^a}|_1 \lesssim \varepsilon$ and $\#\tilde{P}^a \lesssim \varepsilon^{-1/s}$, which, in view of the assumption, is the smallest size, modulo some constant factor, one can generally expect for an approximation with this accuracy. The short proof of this corollary is based on an argument taken from [3, proof of Theorem 4.9].

COROLLARY 5.5. *Let $\gamma > t_1^{-\frac{1}{2}}$. Then, for any $\varepsilon > 0$, $u \in H^1_0(\Omega)$, a partition P, $w_P \in \mathcal{S}_P$ with $\|\!|u - w_P|\!\|_1 \leq \varepsilon$, for $[\tilde{P}^a, w_{\tilde{P}^a}] := \mathbf{COARSE}[P, w_P, \gamma\varepsilon]$ we have that $\|\!|u - w_{\tilde{P}^a}|\!\|_1 \leq (1+\gamma)\varepsilon$, and*

$$\tilde{P}^a \leq D\#\hat{P}$$

for any partition $\hat{P}$ with $\inf_{w_{\hat{P}} \in \mathcal{S}_{\hat{P}}} \|\!|u - w_{\hat{P}}|\!\|_1 \leq (t_1^{\frac{1}{2}}\gamma - 1)\varepsilon$.

Proof. The first statement is an obvious consequence of Theorem 5.4. The second one also follows from this theorem using that

$$\inf_{w_{\hat{P}} \in \mathcal{S}_{\hat{P}}} \|\!|w_P - w_{\hat{P}}|\!\|_1 \leq \|\!|u - w_P|\!\|_1 + \inf_{w_{\hat{P}} \in \mathcal{S}_{\hat{P}}} \|\!|w_{\hat{P}} - u|\!\|_1 \leq \varepsilon + (t_1^{\frac{1}{2}}\gamma - 1)\varepsilon = t_1^{\frac{1}{2}}\gamma\varepsilon. \qquad \square$$

6. A convergent adaptive refinement strategy. We derive an a posteriori estimate of the error in the Galerkin approximation (4.1) of the solution of the boundary value problem (2.1) where we temporarily assume that the right-hand side $f \in L_2(\Omega)$. Second, under the assumption that f is piecewise constant with respect to the current partition, we derive a refinement strategy which guarantees that the difference between the Galerkin solutions on the new and old partition is greater than some fixed multiple of the error estimate for the solution on the old partition. Exploiting Galerkin orthogonality, we can therefore conclude that the error in the new solution is less than some absolute constant times the error in the old solution.

This section is largely based on ideas from [15] by Morin, Nochetto, and Siebert on the construction of an adaptive finite element method that can be proven to converge. The following are new aspects.

- We consider a different type of partitions (nonconforming ones generated by red-refinements vs. conforming ones generated by newest vertex bisection).
- Under Assumption 3.8 our results are valid uniformly in the size of jumps of $\rho = \rho(\mathbf{A})$.
- We use approximations of the right-hand side for setting up the discrete systems, with which the application of quadrature can be modeled, and, as in [3], we solve these systems inexactly. We evaluate the a posteriori error estimator using the inexact discrete solution and the approximate right-hand side. Because of the latter we can allow right-hand sides in $H^{-1}(\Omega)$ that lie outside $L_2(\Omega)$, at least when they can be sufficiently well approximated by piecewise constants.

We start by introducing some notation. We call e *an edge of a partition* P, when e is an edge of some $\triangle \in P$ and e connects two vertices from $\bar{V}_P$. Note that since we allow nonconforming partitions, not all edges of $\triangle \in P$ are edges of P. With $\bar{E}_P$ or E_P, respectively, we denote the set of all edges of P or all edges of P which are not part of $\partial\Omega$. Note that for an admissible partition P^a, any edge $e \in E_{P^a}$ is either the

common edge of $\triangle_1, \triangle_2 \in P^a$, or it is the common edge of $\triangle_1, \hat{\triangle}$, where $\triangle_1 \in P^a$ and $\hat{\triangle}$ is the parent of four triangles in P^a.

For $e \in E_P$ and $u \in \mathcal{S}_P$, we set

$$\eta_e(u) := \frac{\operatorname{diam}(e)}{\max\{\rho|_{e^-}, \rho|_{e^+}\}} \|[\mathbf{A}\nabla u]_e \cdot \mathbf{n}_e\|^2_{L_2(e)},$$

where $\mathbf{n}_e$ is a unit vector orthogonal to e, $[\mathbf{A}\nabla u]_e$ denotes the jump of $\mathbf{A}\nabla u$ in the direction of $\mathbf{n}_e$, and $\rho|_{e^\pm} = \rho(x \pm \delta \mathbf{n}_e)$ for arbitrary $x \in e$ and $\delta > 0$ small enough. For $\triangle \in P$ and $f \in L_2(\Omega)$, we set

$$\zeta_\triangle(f) := \frac{\operatorname{diam}(\triangle)^2}{\rho|_\triangle} \|f\|^2_{L_2(\triangle)},$$

and finally,

$$\mathcal{E}(P, f, u) := \left[\sum_{\triangle \in P} \zeta_\triangle(f) + \sum_{e \in E_P} \eta_e(u) \right]^{\frac{1}{2}}.$$

The a posteriori error estimate given in the following theorem extends results from [2, 17] for conforming partitions to admissible nonconforming partitions.

THEOREM 6.1. *There exists an absolute constant C_1, such that for any $f \in L_2(\Omega)$ and an admissible partition P^a, with $u := L^{-1}f$ and $u_{P^a} := L^{-1}_{P^a} f$, we have that*

$$|u - u_{P^a}|_1 \le C_1 \mathcal{E}(P^a, f, u_{P^a}).$$

Proof. For any $w \in H^1_0(\Omega)$, $w_{P^a} \in \mathcal{S}_{P^a}$, by the Galerkin orthogonality and because $\mathbf{A}$ is piecewise constant, integration by parts shows that

$$\begin{aligned} a(u - u_{P^a}, w) &= a(u - u_{P^a}, w - w_{P^a}) = \int_\Omega f(w - w_{P^a}) - a(u_{P^a}, w - w_{P^a}) \\ &= \sum_{\triangle \in P^a} \int_\triangle f(w - w_{P^a}) - \sum_{e \in E_{P^a}} \int_e ([\mathbf{A}\nabla u_{P^a}]_e \cdot \mathbf{n}_e)(w - w_{P^a}), \end{aligned} \tag{6.1}$$

where $\mathbf{n}_\triangle$ denotes the unit exterior normal to $\triangle$.

It was shown in [16, pp. 17–18] that for any triangle $\triangle$ and any of its vertices v, there exists a $\varphi(\triangle, v) \in L_\infty(\triangle)$ such that $\int_\triangle \varphi(\triangle, v)p = p(v)$ for all polynomials p of degree 1, and $\|\varphi(\triangle, v)\|_{L_\infty} \lesssim \operatorname{meas}(\triangle)^{-1}$ independently of $\triangle$. For each $v \in V_{P^a}$ we now select $\triangle_v = \operatorname{argmax}\{\rho|_\triangle : \triangle \in P^a, v \in \triangle\}$, and define $w_{P^a} \in \mathcal{S}_{P^a}$ by $w_{P^a}(v) = \int_{\triangle_v} \varphi(\triangle_v, v)w$. For $v \in \bar{V}_{P^a} \backslash V_{P^a}$, we set $\triangle_v = \emptyset$.

We start by estimating the first sum from (6.1). Let $\triangle \in P^a$, say $\triangle \in P^a_i$. The triangle $\triangle$ has between 0 and 3 hanging vertices. Since P^a is admissible, by Proposition 3.4 each of these hanging vertices is the midpoint of an edge connecting two vertices from $\bar{V}_{P^a_{i-1}}$. In the following we consider the case when $\triangle$ has one hanging vertex, and so in particular $i > 0$, but the other cases can be treated similarly.

Let $\tilde{\triangle} \in P^a_{i-1}$ be the parent of $\triangle$, v_1, v_2 the vertices of $\tilde{\triangle}$ such that the hanging vertex of $\triangle$ is the midpoint of the edge connecting v_1 and v_2, and let v_3 and v_4 be the nonhanging vertices of $\triangle$; see Figure 6.1. The linear function $w_{P^a}|_\triangle$ is the solution of the elementary interpolation problem with data $w_{P^a}(v_i)$ $(1 \le i \le 4)$, and

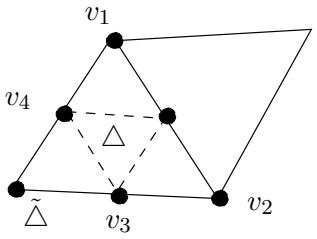


FIG. 6.1. *Illustration with the proof of Theorem* 6.1.

it is easily seen that $\|w_{P^a}\|_{L_2(\triangle)} \lesssim \operatorname{diam}(\triangle)\max_{1\le i\le 4}|w_{P^a}(v_i)|$. By construction, $|w_{P^a}(v_i)| \lesssim \operatorname{diam}(\triangle_{v_i})^{-1}\|w\|_{L_2(\triangle_{v_i})}$, or, when $v_i \in \partial\Omega$, $w_{P^a}(v_i) = 0$.

Since ρ is quasi-monotone (see Assumption 3.8), there exists a uniformly bounded collection $P^a_\triangle \subset P^a$ such that $\triangle, \triangle_{v_1}, \dots, \triangle_{v_4} \subset P^a_\triangle$, $\Omega_\triangle := \cup_{\tilde\triangle\subset P^a_\triangle}\tilde\triangle$ is a simply connected uniformly Lipschitz domain, $\rho|_\triangle \lesssim \inf_{\tilde\triangle\subset P^a_\triangle}\rho|_{\tilde\triangle}$, and, when one of $v_1,\dots,v_4$ is on $\partial\Omega$, $P^a_\triangle$ contains a $\tilde\triangle$ having an edge on $\partial\Omega$. From $\|w-w_{P^a}\|_{L_2(\triangle)} \lesssim \|w\|_{L_2(\Omega_\triangle)}$, a homogeneity argument, and either the fact that our interpolation operator reproduces any polynomial of first degree, and, in particular, any constant, together with the Bramble–Hilbert lemma, or, when one of $v_1,\dots,v_4$ is on $\partial\Omega$, the Poincaré–Friedrichs inequality, we infer that

$$\|w-w_{P^a}\|^2_{L_2(\triangle)} \lesssim \operatorname{diam}(\Omega_\triangle)^2|w|^2_{H^1(\Omega_\triangle)} \lesssim \frac{\operatorname{diam}(\triangle)^2}{\rho|_\triangle}\sum_{\tilde\triangle\in P^a_\triangle}\rho|_{\tilde\triangle}|w|^2_{H^1(\triangle)}.$$

By applying the Cauchy–Schwarz inequality, we deduce that

$$\left|\sum_{\triangle\in P^a}\int_\triangle f(w-w_{P^a})\right|^2 \lesssim \sum_{\triangle\in P^a}\frac{\operatorname{diam}(\triangle)^2}{\rho|_\triangle}\|f\|^2_{L_2(\triangle)}\sum_{\triangle\in P^a}\rho|_\triangle|w|^2_{H^1(\triangle)}. \tag{6.2}$$

Now we estimate the second sum from (6.1). Let $P^a = P^a_n$ and $e \in E_{P^a}$, then $e = \triangle_1\cap\triangle_2$ for some $\triangle_1,\triangle_2 \in \cup_{i=0}^n P^a_i$. Let us assume that $\rho|_{\triangle_1} \ge \rho|_{\triangle_2}$. Note that either $\triangle_1 \in P^a$ or it is the parent of four triangles from P^a. From the trace theorem we have $\|w\|_{L_2(e)} \lesssim \operatorname{diam}(e)^{-\frac12}\|w\|_{L_2(\triangle_1)} + \operatorname{diam}(e)^{\frac12}|w|_{H^1(\triangle_1)}$. With $v_1, v_2 \in \bar V_{P^a}$ being the endpoints of e, we have $\|w_{P^a}\|_{L_2(e)} \le \operatorname{diam}(e)^{\frac12}\max\{|w_{P^a}(v_1)|, |w_{P^a}(v_2)|\}$. From $|w_{P^a}(v_i)| \lesssim \operatorname{diam}(\triangle_{v_i})^{-1}\|w\|_{L_2(\triangle_{v_i})}$, or, when $v_i \in \partial\Omega$, $w_{P^a}(v_i) = 0$, we find that $\|w-w_{P^a}\|_{L_2(e)} \lesssim \operatorname{diam}(e)^{-\frac12}\|w\|_{L_2(\triangle_1\cup\triangle_{v_1}\cup\triangle_{v_2})} + \operatorname{diam}(e)^{\frac12}|w|_{H^1(\triangle\cup\triangle_1\cup\triangle_2)}$. As above, we can extend $\triangle_1, \triangle_{v_1}, \triangle_{v_2}$ to a uniformly bounded collection $P^a_e \subset P^a$, such that $\Omega_e := \cup_{\tilde\triangle\subset P^a_e}\tilde\triangle$ is a simply connected uniformly Lipschitz domain, $\rho|_{\triangle_1} \lesssim \inf_{\tilde\triangle\subset P^a_e}\rho|_{\tilde\triangle}$, and, when v_1 or v_2 is on $\partial\Omega$, P^a_e contains a $\tilde\triangle$ having an edge on $\partial\Omega$. Using the same arguments as above, we infer that

$$\|w-w_{P^a}\|^2_{L_2(e)} \lesssim \operatorname{diam}(e)|w|^2_{H^1(\Omega_e)} \lesssim \frac{\operatorname{diam}(e)}{\max\{\rho|_{e^-},\rho|_{e^+}\}}\sum_{\tilde\triangle\in P^a_e}\rho|_{\tilde\triangle}|w|^2_{H^1(\tilde\triangle)}.$$

By applying the Cauchy–Schwarz inequality, we deduce that

$$\left| \sum_{e\in E_{P^a}} \int_e ([\mathbf{A}\nabla u_{P^a}]_e \cdot \mathbf{n}_e)(w - w_{P^a}) \right|^2$$

$$\lesssim \sum_{e\in E_{P^a}} \frac{\operatorname{diam}(e)}{\max\{\rho|_{e^-}, \rho|_{e^+}\}} \|[\mathbf{A}\nabla u_{P^a}]_e \cdot \mathbf{n}_e\|^2_{L_2(e)} \sum_{\triangle\in P^a} \rho|_{\triangle} |w|^2_{H^1(\triangle)}. \tag{6.3}$$

On noting that $\sum_{\triangle\in P^a} \rho|_{\triangle} |w|^2_{H^1(\triangle)} \eqsim |w|^2_1$ and by substituting $w = u - u_{P^a}$, the proof follows from (6.1), (6.2), (6.3). ∎

The next lemma, which is based on [15, Lemma 4.2], gives local lower bounds on the difference between the Galerkin solution on some partition and a refinement of this partition. The differences with [15] are that we consider a different class of partitions, and that our results hold uniformly in the size of jumps of ρ. Furthermore, we simply assume that the right-hand side f is piecewise constant with respect to the first partition; for the moment we postpone the analysis in the case of more general f.

LEMMA 6.2. *Let P^a be an admissible partition and $\tilde{P}$ a refinement of P^a. Let $f_{P^a} \in L_2(\Omega)$ be* piecewise constant with respect to P^a, *i.e., $f_{P^a} \in \mathcal{S}^0_{P^a}$, and let $u_{P^a} = L^{-1}_{P^a} f_{P^a}, u_{\tilde{P}} = L^{-1}_{\tilde{P}} f_{P^a}$ be the corresponding Galerkin solutions.*

(a) *Let $\triangle_1, \triangle_2 \in P^a$ such that $e := \triangle_1 \cap \triangle_2 \in E_{P^a}$. Assume that $V_{\tilde{P}}$ contains points interior to $\triangle_1$, $\triangle_2$, and e; see Figure* 6.2. *Then*

$$|u_{\tilde{P}} - u_{P^a}|^2_{1,\triangle_1\cup\triangle_2} \gtrsim \eta_e(u_{P^a}) + \sum_{i=1}^{2} \zeta_{\triangle_i}(f_{P^a}).$$

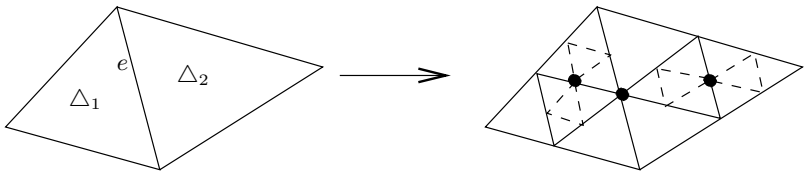


FIG. 6.2. *Illustration of Lemma* 6.2(a).

(b) *Let $\triangle_1, \hat{\triangle}$ such that $e := \triangle_1 \cap \hat{\triangle} \in E_{P^a}$, $\triangle_1 \in P^a$ and $\hat{\triangle}$ is the parent of four triangles $\triangle_2, \ldots, \triangle_5 \in P^a$, numbered such that $\triangle_2$, $\triangle_3$ have an edge e_2, e_3 on e. Assume that $V_{\tilde{P}}$ contains points interior to $\triangle_1$, $\triangle_2$, $\triangle_3$, e_2, and e_3; see Figure* 6.3. *Then*

$$|u_{\tilde{P}} - u_{P^a}|^2_{1,\triangle_1\cup\triangle_2\cup\triangle_3} \gtrsim \eta_e(u_{P^a}) + \sum_{i=1}^{3} \zeta_{\triangle_i}(f_{P^a}).$$

(c) *Assume that $V_{\tilde{P}}$ contains a point interior to $\triangle \in P^a$. Then*

$$|u_{\tilde{P}} - u_{P^a}|^2_{1,\triangle} \gtrsim \zeta_{\triangle}(f_{P^a}).$$

Proof. (a) By assumption, there exist $\varphi_1, \varphi_2, \varphi_3 \in H^1_0(\triangle_1 \cup \triangle_2) \cap \mathcal{S}_{\tilde{P}}$ with $|\varphi_i|_1 = 1$, and such that for $i \in \{1, 2\}$,

$$\operatorname{supp}(\varphi_i) \subset \triangle_i,\ \int_{\triangle_i} \varphi_i \eqsim \frac{\operatorname{meas}(\triangle_i)}{\rho|^{\frac{1}{2}}_{\triangle_i}},\ \int_{\triangle_i} \varphi_3 \eqsim \frac{\operatorname{meas}(\triangle_i)}{\max\{\rho|_{e^-}, \rho|_{e^+}\}^{\frac{1}{2}}},\ \int_e \varphi_3 \eqsim \frac{\operatorname{meas}(e)}{\max\{\rho|_{e^-}, \rho|_{e^+}\}^{\frac{1}{2}}}.$$

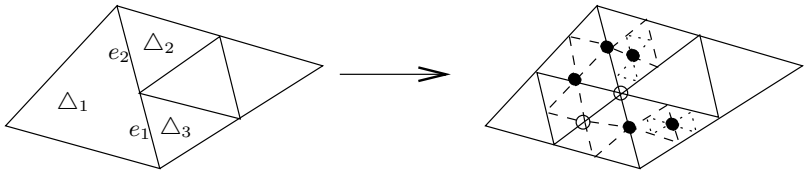


FIG. 6.3. *Illustration of Lemma* 6.2(b). *Open circles correspond to degrees of freedom that are not used.*

For any $\varphi = \sum_{i=1}^3 \mathbf{c}_i \varphi_i$, integration by parts shows that

$$\int_{\triangle_1 \cup \triangle_2} f_{P^a}\varphi - \int_e ([\mathbf{A}\nabla u_{P^a}]_e \cdot \mathbf{n}_e)\varphi = a(u_{\tilde{P}} - u_{P^a}, \varphi) \leq |u_{\tilde{P}} - u_{P^a}|_{1,\triangle_1\cup\triangle_2}|\varphi|_1 \lesssim |u_{\tilde{P}} - u_{P^a}|_{1,\triangle_1\cup\triangle_2}\|\mathbf{c}\|. \tag{6.4}$$

Let $g_j \in H^1_0(\triangle_1 \cup \triangle_2)'$ be defined by $g_j(\varphi) = \frac{\rho|^{\frac{1}{2}}_{\triangle_j}}{\mathrm{diam}(\triangle_j)\mathrm{meas}(\triangle_j)^{\frac{1}{2}}} \int_{\triangle_j} \varphi$ when $j = 1$ or 2, and $g_3(\varphi) = \frac{\max\{\rho|_{e^-}, \rho|_{e^+}\}^{\frac{1}{2}}}{(\mathrm{diam}(e)\mathrm{meas}(e))^{\frac{1}{2}}} \int_e \varphi$, and let $\mathbf{B}_{ij} := g_j(\varphi_i)$. Then with $\mathbf{d}_j = \frac{\mathrm{diam}(\triangle_j)}{\rho|^{\frac{1}{2}}_{\triangle_j}} \|f_{P^a}\|_{L_2(\triangle_j)}$ when $j = 1$ or 2, and $\mathbf{d}_3 = \frac{\mathrm{diam}(e)^{\frac{1}{2}}}{\max\{\rho|_{e^-}, \rho|_{e^+}\}^{\frac{1}{2}}} \|[\mathbf{A}\nabla u]_e \cdot \mathbf{n}_e\|_{L_2(e)}$, the left-hand side of (6.4) is $\sum_{i,j=1}^3 \mathbf{d}_j \mathbf{c}_i \mathbf{B}_{ij}$. Using (6.4), the statement of the lemma reduces to the question as to whether $\sup_{\mathbf{c}} \frac{\langle \mathbf{Bd}, \mathbf{c}\rangle}{\|\mathbf{c}\|} \gtrsim \|\mathbf{d}\|$ or $\|\mathbf{B}^{-1}\| \lesssim 1$. Since $\mathbf{B}_{ii} \eqsim 1$, and $\mathbf{B}_{31}, \mathbf{B}_{32} \lesssim 1$, whereas the other coefficients of $\mathbf{B}$ are zero, the proof of (a) is completed.

Part (b) can be proven similarly to (a). Note that, generally, $[\mathbf{A}\nabla u_{P^a}]_e$ has different values on e_2 and e_3. The proof of (c) poses no additional difficulties. ∎

As an immediate consequence we have the following result.

COROLLARY 6.3. *Let P^a be an admissible partition, $f_{P^a} \in \mathcal{S}^0_{P^a}$, and let $\tilde{P}$ be a refinement of P^a, such that for some $G \subset P^a$, $F \subset E_{P^a}$, for all $\triangle \in G$, $V_{\tilde{P}}$ satisfies the conditions from Lemma* 6.2(c), *and for all $e \in F$, $V_{\tilde{P}}$ satisfies the conditions from either Lemma* 6.2(a) *or Lemma* 6.2(b) *(it is sufficient that $\tilde{P}$ contains all grandchildren of all $\triangle \in P^a$ which either are in G or have an edge on an $e \in F$). Then, with $u_{P^a} = L^{-1}_{P^a} f_{P^a}, u_{\tilde{P}} = L^{-1}_{\tilde{P}} f_{P^a}$ denoting the corresponding Galerkin solutions, we have that*

$$|u_{\tilde{P}} - u_{P^a}|^2_1 \geq c_2^2 \left\{ \sum_{\triangle\in F} \zeta_\triangle(f_{P^a}) + \sum_{e\in G} \eta_e(u_{P^a}) \right\}$$

for some absolute constant $c_2 > 0$.

The way to obtain a convergent adaptive refinement strategy is to select the sets F and G such that $\sum_{\triangle\in F} \zeta_\triangle(f_{P^a}) + \sum_{e\in G} \eta_e(u_{P^a})$ is bounded from below by some multiple of $\sum_{\triangle\in P^a} \zeta_\triangle(f_{P^a}) + \sum_{e\in E_{P^a}} \eta_e(u_{P^a}) = \mathcal{E}(P^a, f_{P^a}, u_{P^a})^2$. Then, convergence follows by combining Theorem 6.1 and Corollary 6.3.

Since Lemma 6.2 also applies when $u_{\tilde{P}} = L^{-1}_{\tilde{P}} f_{P^a}$ is replaced by $u = L^{-1} f_{P^a}$, for later use we state the following result.

COROLLARY 6.4. *Let P^a be an admissible partition and let $f_{P^a} \in \mathcal{S}^0_{P^a}$. With $u := L^{-1} f_{P^a}$, $u_{P^a} := L^{-1}_{P^a} f_{P^a}$ and c_2 the constant from Corollary* 6.3, *we have that*

$$|u - u_{P^a}|_1 \geq c_2 \mathcal{E}(P^a, f_{P^a}, u_{P^a}).$$

In Corollary 6.3 it was assumed that the right-hand side f is piecewise constant with respect to the current partition P^a, and that the discrete system is solved exactly. In the remainder of this section we will relax both of these assumptions.

LEMMA 6.5. *There exists an absolute constant $C_3 > 0$ such that for any partition P, $f \in L_2(\Omega)$, $u_P, \tilde{u}_P \in \mathcal{S}_P$,*

$$|\mathcal{E}(P, f, u_P) - \mathcal{E}(P, f, \tilde{u}_P)| \leq C_3 |u_P - \tilde{u}_P|_1.$$

Proof. We have that

$$\begin{aligned} |\mathcal{E}(P, f, u_P) - \mathcal{E}(P, f, \tilde{u}_P)| &= \left[\sum_{\triangle \in P} \zeta_\triangle(f) + \sum_{e \in E_P} \eta_e(u_P) \right]^{\frac{1}{2}} \\ &\quad - \left[\sum_{\triangle \in P} \zeta_\triangle(f) + \sum_{e \in E_P} \eta_e(\tilde{u}_P) \right]^{\frac{1}{2}} \\ &\leq \left[\sum_{e \in E_P} (\eta_e(u_P)^{\frac{1}{2}} - \eta_e(\tilde{u})^{\frac{1}{2}})^2 \right]^{\frac{1}{2}}. \end{aligned}$$

For any $e \in E_P$,

$$\begin{aligned} |\eta_e(u_P)^{\frac{1}{2}} - \eta_e(\tilde{u}_P)^{\frac{1}{2}}| &= \frac{\operatorname{diam}(e)^{\frac{1}{2}}}{\max\{\rho|_{e^-}, \rho|_{e^+}\}^{\frac{1}{2}}} \left| \|[\mathbf{A}\nabla u_P]_e \cdot \mathbf{n}_e\|_{L_2(e)} - \|[\mathbf{A}\nabla \tilde{u}_P]_e \cdot \mathbf{n}_e\|_{L_2(e)} \right| \\ &\leq \frac{\operatorname{diam}(e)^{\frac{1}{2}}}{\max\{\rho|_{e^-}, \rho|_{e^+}\}^{\frac{1}{2}}} \|[\mathbf{A}\nabla(u_P - \tilde{u}_P)]_e \cdot \mathbf{n}_e\|_{L_2(e)}. \end{aligned}$$

The proof is completed by the observation that, for any edge e_i of a $\triangle \in P$ and any $w \in P_1(\triangle)$ and unit vector $\mathbf{n}$, by a homogeneity argument we have that

$$\operatorname{diam}(e_i)^{\frac{1}{2}} \|\mathbf{A}\nabla w \cdot \mathbf{n}\|_{L_2(e_i)} \lesssim \rho|_\triangle |w|_{H^1(\triangle)} \eqsim \rho|_\triangle^{\frac{1}{2}} |w|_{1,\triangle}. \qquad \square$$

For some *fixed constant* $\theta \in (0, 1]$, we consider the following refinement procedure.
REFINE$[P^a, f_{P^a}, w_{P^a}] \to \tilde{P}$
% P^a is an admissible partition, $f_{P^a} \in \mathcal{S}^0_{P^a}$, and $w_{P^a} \in \mathcal{S}_{P^a}$.
Select $F \subset P^a$, $G \subset E_{P^a}$ such that

$$\sum_{\triangle \in F} \zeta_\triangle(f_{P^a}) + \sum_{e \in G} \eta_e(w_{P^a}) \geq \theta^2 \, \mathcal{E}(P^a, f_{P^a}, w_{P^a})^2.$$

Determine a refinement $\tilde{P}$ of P^a such that for all $\triangle \in F$, $V_{\tilde{P}}$ satisfies the conditions from Lemma 6.2(c), and for all $e \in G$, $V_{\tilde{P}}$ satisfies the conditions from either Lemma 6.2(a) or Lemma 6.2(b), where at the same time each $\tilde{\triangle} \in \tilde{P}$ is either in P^a or it is a child or a grandchild of a $\triangle \in P^a$.

As long as the selection of F and G is organized so that it does not involve the exact ordering of all $\zeta_\triangle(f_{P^a})$ and $\eta_e(w_{P^a})$ by their moduli (cf. discussion from Remark 5.2), we have the following result.

PROPOSITION 6.6. *The call* **REFINE**$[P^a, f_{P^a}, w_{P^a}]$ *requires a number of arithmetic operations $\lesssim \# P^a$.*

As stated before, we would like to consider $f \in H^{-1}(\Omega)$, possibly outside $L_2(\Omega)$, and we will use approximate right-hand sides both for setting up the discrete systems as well as for the evaluation of the a posteriori error estimator. In fact, since the terms $\zeta_\triangle(f)$ are only defined for $f \in L_2(\Omega)$, generally we will need approximate right-hand sides different from the exact one. The following theorem shows that if **REFINE** is called with a sufficiently accurate piecewise constant approximation for the right-hand side and a sufficiently accurate approximation of the discrete solution, then the solution on the new partition has an error that is less than the error in the solution on the previous partition. Together with the coarsening routine from section 5, in the next two subsections this theorem will be the basis for constructing adaptive finite element methods that converge with optimal rates.

THEOREM 6.7. *Let $f \in H^{-1}(\Omega)$, $u = L^{-1}f$, let P^a be an admissible partition, $f_{P^a} \in \mathcal{S}^0_{P^a}$, $u_{P^a} = L^{-1}_{P^a} f_{P^a}$, $\bar{u}_{P^a} \in \mathcal{S}_{P^a}$, $\tilde{P} = \mathbf{REFINE}[P^a, f_{P^a}, \bar{u}_{P^a}]$ or a refinement of it, $f_{\tilde{P}} \in H^{-1}(\Omega)$ and $u_{\tilde{P}} = L^{-1}_{\tilde{P}} f_{\tilde{P}}$. Then*

$$|u - u_{\tilde{P}}|_1 \le \left[1 - \frac{1}{2}\left(\frac{c_2\theta}{C_1}\right)^2\right]^{\frac{1}{2}} |u - u_{P^a}|_1 + 2c_2C_3|u_{P^a} - \bar{u}_{P^a}|_1 + 3\|f - f_{P^a}\|_{-1} + \|f - f_{\tilde{P}}\|_{-1}. \tag{6.5}$$

Proof. Let $\hat{u} = L^{-1}f_{P^a}$ and $\hat{u}_{\tilde{P}} = L^{-1}_{\tilde{P}} f_{P^a}$. With $F \subset P^a$, $G \subset E_{P^a}$ as determined in the call $\mathbf{REFINE}[P^a, f_{P^a}, \bar{u}_{P^a}]$, Corollary 6.3, two applications of Lemma 6.5, and Theorem 6.1 show that

$$\begin{aligned}
|\hat{u}_{\tilde{P}} - u_{P^a}|_1 &\ge c_2\left[\left\{\sum_{\triangle\in F}\zeta_\triangle(f_{P^a}) + \sum_{e\in G}\eta_e(u_{P^a})\right\}\right]^{\frac{1}{2}} \\
&\ge c_2\left[\left\{\sum_{\triangle\in F}\zeta_\triangle(f_{P^a}) + \sum_{e\in G}\eta_e(\bar{u}_{P^a})\right\}^{\frac{1}{2}} - C_3|u_{P^a} - \bar{u}_{P^a}|_1\right] \\
&\ge c_2[\theta\mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) - C_3|u_{P^a} - \bar{u}_{P^a}|_1] \\
&\ge c_2[\theta\mathcal{E}(P^a, f_{P^a}, u_{P^a}) - 2C_3|u_{P^a} - \bar{u}_{P^a}|_1] \\
&\ge c_2\left[\frac{\theta}{C_1}|\hat{u} - u_{P^a}|_1 - 2C_3|u_{P^a} - \bar{u}_{P^a}|_1\right].
\end{aligned}$$

Since for any scalars a, b, $(a-b)^2 \ge \frac{1}{2}a^2 - b^2$, we infer that

$$|\hat{u}_{\tilde{P}} - u_{P^a}|_1^2 \ge \frac{1}{2}\left(\frac{c_2\theta}{C_1}\right)^2 |\hat{u} - u_{P^a}|_1^2 - 4c_2^2C_3^2|u_{P^a} - \bar{u}_{P^a}|_1^2.$$

Since $\hat{u}_{\tilde{P}} \in \mathcal{S}_{\tilde{P}}$ is the Galerkin approximation of $\hat{u}$ from $\mathcal{S}_{\tilde{P}}$ and $u_{P^a} - \hat{u}_{\tilde{P}} \in \mathcal{S}_{\tilde{P}}$, we have

$$\begin{aligned}
|\hat{u} - \hat{u}_{\tilde{P}}|_1^2 &= |\hat{u} - u_{P^a}|_1^2 - |u_{P^a} - \hat{u}_{\tilde{P}}|_1^2 \\
&\le \left[1 - \frac{1}{2}\left(\frac{c_2\theta}{C_1}\right)^2\right]|\hat{u} - u_{P^a}|_1^2 + 4c_2^2C_3^2|u_{P^a} - \bar{u}_{P^a}|_1^2 \\
&\le \left\{\left[1 - \frac{1}{2}\left(\frac{c_2\theta}{C_1}\right)^2\right]^{\frac{1}{2}}|\hat{u} - u_{P^a}|_1 + 2c_2C_3|u_{P^a} - \bar{u}_{P^a}|_1\right\}^2,
\end{aligned}$$

where we have used that $c_2 \le C_1$ and thus that $1 - \frac{1}{2}(\frac{c_2\theta}{C_1})^2 > 0$. The proof is completed by observing that $|u - \hat{u}|_1 = \|f - f_{P^a}\|_{-1}$ and $|u_{\tilde{P}} - \hat{u}_{\tilde{P}}|_1 \le \|f_{\tilde{P}} - f_{P^a}\|_{-1} \le \|f_{\tilde{P}} - f\|_{-1} + \|f - f_{P^a}\|_{-1}$. ∎

7. A first optimal adaptive finite element method. We start with a corollary that is an easy consequence of Theorem 6.7. It shows that the reduction of the error in the exact discrete solutions as a result of refinement extends to a similar reduction of the error in sufficiently accurate approximations of these discrete solutions.

COROLLARY 7.1. *For any* $\mu \in ([1-\frac{1}{2}(\frac{c_2\theta}{C_1})^2]^{\frac{1}{2}},1)$*, there exists a sufficiently small constant* $\delta>0$ *such that if for* $f\in H^{-1}(\Omega)$*, an admissible partition* P^a*,* $\bar{u}_{P^a}\in\mathcal{S}_{P^a}$*,* $f_{P^a}\in\mathcal{S}^0_{P^a}$*,* $\tilde{P}=\mathbf{REFINE}[P^a,f_{P^a},\bar{u}_{P^a}]$ *or a refinement of it,* $\bar{u}_{\tilde{P}}\in\mathcal{S}_{\tilde{P}}$*,* $f_{\tilde{P}}\in H^{-1}(\Omega)$ *and* $\varepsilon>0$*, with* $u=L^{-1}f$*,* $u_{P^a}=L^{-1}_{P^a}f_{P^a}$ *and* $u_{\tilde{P}}=L^{-1}_{\tilde{P}}f_{\tilde{P}}$*, we have that* $|u-\bar{u}_{P^a}|_1\le\varepsilon$ *and*

$$|u_{P^a}-\bar{u}_{P^a}|_1+\|f-f_{P^a}\|_{-1}+|u_{\tilde{P}}-\bar{u}_{\tilde{P}}|_1+\|f-f_{\tilde{P}}\|_{-1}\le 2(1+\mu)\delta\varepsilon$$

then $|u-\bar{u}_{\tilde{P}}|_1\le\mu\varepsilon$*.*

Proof. The proof is an easy consequence of Theorem 6.7, $|u-\bar{u}_{\tilde{P}}|_1\le|u-u_{\tilde{P}}|_1+|u_{\tilde{P}}-\bar{u}_{\tilde{P}}|_1$ and $|u-u_{P^a}|_1\le|u-\bar{u}_{P^a}|_1+|u_{P^a}-\bar{u}_{P^a}|_1$. ∎

We assume the availability of the following routine.

GALSOLVE$[P^a,f_{P^a},u^{(0)}_{P^a},\varepsilon]\to\bar{u}_{P^a}$

% P^a is an admissible partition, $f_{P^a}\in(\mathcal{S}_{P^a})'$, and $u^{(0)}_{P^a}\in\mathcal{S}_{P^a}$. With $u_{P^a}:=L^{-1}_{P^a}f_{P^a}$,
% the output $\bar{u}_{P^a}\in\mathcal{S}_{P^a}$ satisfies

$$|u_{P^a}-\bar{u}_{P^a}|_1\le\varepsilon.$$

% The call requires $\lesssim\max\{1,\log(\varepsilon^{-1}|u_{P^a}-u^{(0)}_{P^a}|_1|)\}\#P^a$ arithmetic operations.

Thus not only do we assume that we have an iterative solver at our disposal that converges with a rate independent of the problem size, but, in accordance with the idea of an adaptive solver, we additionally assume that we have an efficient and reliable control of the algebraic error. As a consequence the number of iterations to be performed does not depend on a possibly pessimistic a priori bound on the initial error. Two possible realizations of **GALSOLVE** are discussed in the next remark.

Remark 7.2. One can apply conjugate gradients, starting with $u^0_{P^a}$, to the representation of $L_{P^a}u_{P^a}=f_{P^a}$ with respect to $\{\bar{\psi}^v:v\in V_{P^a}\}$. In each iteration, that takes $\eqsim\#P^a$ operations, the $|\cdot|_1$-norm of the error is multiplied by a factor less than or equal to some constant $\tau<1$ only dependent on $\kappa_{\bar{\Psi}}$, meaning that after $\lceil\log_\tau(\varepsilon|u_{P^a}-u^{(0)}_{P^a}|_1^{-1})\rceil$ iterations the $|\cdot|_1$-norm of the error is $\le\varepsilon$. The $|\cdot|_1$-norm of the error in an approximation for u_{P^a} from $\mathcal{S}_{P^a}$ is less (respectively, greater) than or equal to $\lambda_{\bar{\Psi}}^{-\frac{1}{2}}$ (respectively, $\Lambda_{\bar{\Psi}}^{-\frac{1}{2}}$) times the Euclidean norm of the corresponding residual vector. So if one stops the iteration as soon as the latter norm is $\le\lambda_{\bar{\Psi}}^{\frac{1}{2}}\varepsilon$, then the $|\cdot|_1$-norm of the error is $\le\varepsilon$, whereas the number of iterations is bounded by

$$\lceil\log_\tau(\kappa_{\bar{\Psi}}^{-\frac{1}{2}}\varepsilon|u_{P^a}-u^{(0)}_{P^a}|_1^{-1})\rceil\lesssim\max\{1,\log(\varepsilon^{-1}|u_{P^a}-u^{(0)}_{P^a}|_1)\},$$

showing that this approach results in a valid routine **GALSOLVE**.

Alternatively, one may apply the conjugate gradient method to the representation of $L_{P^a}u_{P^a}=f_{P^a}$ with respect to the nodal basis $\{\phi^v_{P^a}:v\in V_{P^a}\}$ using a BPX preconditioner where, similarly as above, the Euclidean norm of the residual of the preconditioned system may serve to develop a stopping criterion. Indeed, when $P^a=P^a_n$, for $0\le i\le n$ one can select $\tilde{V}_{P^a_i}\subset V_{P^a_i}$ such that both $\operatorname{span}\{\bar{\psi}^v:v\in V_{P^a_i}\backslash V_{P^a_{i-1}}\}\subset\operatorname{span}\{\phi^v_{P^a_i}:v\in\tilde{V}_{P^a_i}\}$ and $\#\tilde{V}_{P^a_i}\lesssim\#(P^a_i\backslash P^a_{i-1})$. Using (4.6), it can then be proven that on $\mathcal{S}_{P^a}$, $\inf_{u=\sum_{i=0}^n\sum_{v\in\tilde{V}_{P^a_i}}c_i^v\phi^v_{P^a_i}}|c_i^v|^2|\phi^v_{P^a_i}|_1^2\eqsim|u|_1^2$,

showing that the resulting BPX preconditioner or, more precisely, in view of possible jumps of ρ the MDS preconditioner (cf. [16]) gives rise to uniformly well-conditioned systems, whereas it can be implemented in $\lesssim \#P^a$ operations.

Before continuing, let us first explain what we mean by an optimal method for solving the boundary value problem (2.1). Following [3], we say that a method is optimal if whenever the solution u is such that for some $s > 0$ the error of the best approximation from any $\mathcal{S}_P$ with $\#P \leq n$ is $\lesssim n^{-s}$, then for any $\varepsilon > 0$, the method yields a partition P and a $w_P \in \mathcal{S}_P$ with $|u - w_P|_1 \leq \varepsilon$ taking only $\lesssim \#P$ operations where $\#P \lesssim \varepsilon^{-1/s}$. Indeed, note that in view of the assumption on u, the smallest partition P for which there exists such a $w_P \in \mathcal{S}_P$ generally has cardinality $\eqsim \varepsilon^{-1/s}$. A definition of the class of functions $u \in H^1_0(\Omega)$ for which for some $s > 0$ the errors of the best approximations decay as indicated above is given by

$$\mathcal{A}^s = \left\{ u \in H^1_0(\Omega) : \sup_{n \geq 0} n^s \inf_{\#P \leq n} \inf_{u_P \in \mathcal{S}_P} |u - u_P|_1 < \infty \right\},$$

where P is any partition of the type we consider.

It is well known that for $s \leq \frac{1}{2}$, $H^1_0(\Omega) \cap H^{1+2s}(\Omega) \subset \mathcal{A}^s$. Indeed, for functions in $H^1_0(\Omega) \cap H^{1+2s}(\Omega)$, the sequence of errors of the best continuous piecewise linear approximations subordinate to the sequence of uniform refinements P_i^* of P_0 already exhibits a decay of $\lesssim (\#P_i^*)^{-s}$. Obviously, the class $\mathcal{A}^s$ contains many more functions than only those in $H^1_0(\Omega) \cap H^{1+2s}(\Omega)$, which is the reason to consider adaptive methods anyway. Although the class $\mathcal{A}^s$ is nontrivial for any $s > 0$, as it contains all $u \in \mathcal{S}_P$ for any partition P, since we are approximating with piecewise linears only for $s \leq \frac{1}{2}$ membership of $\mathcal{A}^s$ can be guaranteed by imposing suitable smoothness conditions. For partitions generated by the so-called newest vertex bisection, a characterization of $\mathcal{A}^s$ for $s \leq \frac{1}{2}$ in terms of Besov spaces can be found in [4]. We expect the same results to be valid for the type of partitions considered here. The characterization via Besov spaces, together with regularity results, as in [10], allows one to obtain a priori knowledge about the class $\mathcal{A}^s$ to which the solution u of the boundary value problem (2.1) belongs.

With the adaptive refinement strategy from Theorem 6.7 and Corollary 7.1, convergence can be guaranteed only when sufficiently accurate piecewise constant approximations to the right-hand side f are available. We assume the availability of the following routine **RHS** that, as the routine **REFINE**, may involve a refinement of the current partition.

RHS$[P, f, \varepsilon] \to [P^a, f_{P^a}]$
% P is a partition, $f \in H^{-1}(\Omega)$ and $\varepsilon > 0$. The output consists of an admissible
% refinement P^a of P, and an $f_{P^a} \in \mathcal{S}^0_{P^a}$ with $\|f - f_{P^a}\|_{-1} \leq \varepsilon$.

Assuming that the solution $u \in \mathcal{A}^s$ for some $s > 0$, the cost of approximating the right-hand side f using a routine **RHS** will generally not dominate the other costs of our adaptive method only if there is some constant c_f such that for any $\varepsilon > 0$ and any partition P, for $[P^a, f_{P^a}] := \mathbf{RHS}[P, f, \varepsilon]$ both $\#P^a$ and the number of arithmetic operations required by this call are $\lesssim \#P + c_f^{1/s} \varepsilon^{-1/s}$. We will call such a pair $(f, \mathbf{RHS})$ *s-optimal*. Obviously, given s, such a pair only exists when $f \in \bar{\mathcal{A}}^s$, defined by

$$\bar{\mathcal{A}}^s = \left\{ f \in H^{-1}(\Omega) : \sup_{n \geq 0} n^s \inf_{\#P \leq n} \inf_{f_P \in \mathcal{S}^0_P} \|f - f_P\|_{-1} < \infty \right\},$$

where P is any partition of the type we consider.

When $f \in L_2(\Omega)$ with $\|\rho^{-\frac{1}{2}} f\|_{L_2(\Omega)} \lesssim 1$, the routine **RHS** can be based on uniform refinements. Indeed, for some integer i to be determined below, let $\tilde{P}$ denote the smallest common refinement of the given P and P_i^*, let $\tilde{P}^a$ be its admissible refinement as a result of applying Algorithm 3.6, and with $Q^0_{\tilde{P}^a} : L_2(\Omega) \to \mathcal{S}^0_{\tilde{P}^a}$ denoting the $L_2(\Omega)$-orthogonal projector onto $\mathcal{S}^0_{\tilde{P}^a}$, let $f_{\tilde{P}^a} = Q^0_{\tilde{P}^a} f$. For any $w \in H^1_0(\Omega)$, we have

$$\left| \int_\Omega (f - f_{\tilde{P}^a}) w \right| = \left| \int_\Omega f(w - Q^0_{\tilde{P}^a} w) \right| \le C \sum_{\triangle \in \tilde{P}^a} \|f\|_{L_2(\triangle)} \mathrm{diam}(\triangle) \|\nabla w\|_{L_2(\triangle)}$$
$$\le C 2^{-i} \|\rho^{-\frac{1}{2}} f\|_{L_2(\Omega)} |w|_1,$$

where $C > 0$ is some absolute constant. By taking i to be the smallest integer such that $2^{-i} C \|\rho^{-\frac{1}{2}} f\|_{L_2(\Omega)} \le \varepsilon$, we have $\|f - f_{\tilde{P}^a}\|_{-1} \le \varepsilon$, and

$$\#\tilde{P}^a \lesssim \#P + \#P_i^* \lesssim \#P + (2^i)^2 \lesssim \#P + \varepsilon^{-2} \|\rho^{-\frac{1}{2}} f\|^2_{L_2(\Omega)}.$$

Thus, when for any $\triangle \in \cup_{j \ge 0} P_j^*$ the evaluation of $\int_\triangle f$ takes $\mathcal{O}(1)$ operations, we may conclude that for **RHS** based on this procedure $(f, \mathbf{RHS})$ is s-optimal with $s = \frac{1}{2}$, which, as we have seen, covers the range of main interest. Alternatively, instead of assuming the exact evaluation of $\int_\triangle f$, one easily infers that it also suffices to approximate the integral with an error $\lesssim \rho^{\frac{1}{2}} \mathrm{diam}(\triangle)$, which, in any case, is possible to accomplish in $\mathcal{O}(1)$ operations when $\rho^{-\frac{1}{2}} f$ has some piecewise smoothness with respect to P_0.

Obviously, the class $\bar{\mathcal{A}}^s$ is much larger than $L_2(\Omega)$. Yet, for $f \notin L_2(\Omega)$ the realization of a suitable routine **RHS** has to depend on the right-hand side at hand. We give one example.

Example 7.3. Let $\rho = 1$, and for a sufficiently smooth curve K in Ω, let $f \in H^{-1}(\Omega)$ be defined by $f(w) = \int_K w$. We define $\mathbf{RHS}[P, f, \varepsilon]$ by the following steps. Recursively refine all $\triangle \in P$ that have nonempty intersection with K, until all those $\triangle$ satisfy $\mathrm{diam}(\triangle) \le C\varepsilon^2$ for some constant $C > 0$. Let P^a be the admissible refinement of the obtained partition. Define $f_{P^a} \in \mathcal{S}^0_{P^a}$ by $f_{P^a}|_\triangle = \frac{\mathrm{length}(K \cap \triangle)}{\mathrm{vol}(\triangle)}$ $(\triangle \in P^a)$. Then one may verify that by choosing C suitably, we have that $\|f - f_{P^a}\|_{-1} \le \varepsilon$ and that both the number of arithmetic operations required by this call of **RHS** and $\#P^a$ are $\lesssim \#P + \varepsilon^{-2}$, showing that this $(f, \mathbf{RHS})$ pair is s-optimal with $s = \frac{1}{2}$. Note that although, depending on the curve, the exact evaluation of f applied to any $v_P \in \mathcal{S}_P$ might not pose any problems, nevertheless the piecewise constant approximations are required because of the evaluation of the a posteriori error estimator.

We are ready to give our first adaptive finite element method for solving (2.1). As expected, it is based on the repeated application of the triple **REFINE**, **RHS** and **GALSOLVE**, the latter two with suitable tolerances, which by Corollary 7.1 give rise to linearly convergent approximations. To obtain an optimal work-accuracy balance, **COARSE** is applied after every M iteration of the above triple, with M being some fixed constant.

SOLVE1$[f, \varepsilon, \bar{u}_{P_0}, \varepsilon_0] \to [P^a, \bar{u}_{P^a}]$:
% The following constants are fixed: $\delta < \frac{1}{3}$ *and it is small enough so that it corresponds % to a* $\mu < 1$ *as in Corollary* 7.1*;* $\gamma > t_1^{-\frac{1}{2}}$ *with* t_1 *as in Proposition* 5.3*;* $M \in \mathbb{N}$ *% such that* $\frac{(1+\gamma)\kappa_{\mathcal{V}}^{\frac{1}{2}} \mu^M}{1-3\delta} < 1$.

% The input must satisfy $f \in H^{-1}(\Omega)$ such that a valid routine **RHS** *is available,*
% $\varepsilon > 0$, $\bar{u}_{P_0} \in \mathcal{S}_{P_0}$ and $\varepsilon_0 \geq |u - \bar{u}_{P_0}|_1$.
$P^a := P_0$, $f_{P^a} := f_{P_0}$, $\bar{u}_{P^a} := \bar{u}_{P_0}$
`if` $\varepsilon \geq \varepsilon_0$ `then` $N := 0$
`else` $N \geq 1$ *is the smallest integer with* $(\frac{\mu^M}{1-3\delta})^N((1+\gamma)\kappa_{\bar{\Psi}}^{\frac{1}{2}})^{N-1}\varepsilon_0 \leq \varepsilon$ `fi`
`for` $i = 1, \ldots, N$ `do`
  `if` $i = 1$ `then` $\varepsilon_1 := \frac{\varepsilon_0}{1-3\delta}$
  `else` $[P^a, \bar{u}_{P^a}] := \mathbf{COARSE}[P^a, \bar{u}_{P^a}, \gamma\lambda_{\bar{\Psi}}^{-\frac{1}{2}}\mu^M\varepsilon_{i-1}]$, $\varepsilon_i := \frac{(1+\gamma)\kappa_{\bar{\Psi}}^{\frac{1}{2}}\mu^M}{1-3\delta}\varepsilon_{i-1}$ `fi`
  $[P^a, f_{P^a}] := \mathbf{RHS}[P^a, f, \delta\varepsilon_i]$
  $\bar{u}_{P^a} := \mathbf{GALSOLVE}[P^a, f_{P^a}, \bar{u}_{P^a}, \delta\varepsilon_i]$
  `for` $j = 1, \ldots, M$ `do`
    $P := \mathbf{REFINE}[P^a, f_{P^a}, \bar{u}_{P^a}]$
    $[P^a, f_{P^a}] := \mathbf{RHS}[P, f, \delta\mu^j\varepsilon_i]$
    $\bar{u}_{P^a} := \mathbf{GALSOLVE}[P^a, f_{P^a}, \bar{u}_{P^a}, \delta\mu^j\varepsilon_i]$
  `od`
`od`

The next theorem shows that **SOLVE1** is an optimal method whenever this is allowed by the $(f, \mathbf{RHS})$ pair.

THEOREM 7.4. *$[P^a, \bar{u}_{P^a}] := \mathbf{SOLVE1}[f, \varepsilon, \bar{u}_{P_0}, \varepsilon_0]$ satisfies $|u - \bar{u}_{P^a}|_1 \leq \varepsilon$. Assuming $\varepsilon_0 \lesssim |u|_1$, if for some $s > 0$, $u \in \mathcal{A}^s$ and $(f, \mathbf{RHS})$ is s-optimal, then both $\#P^a$ and the number of arithmetic operations required by this call are $\lesssim \max\{1, \varepsilon^{-1/s}(c_f^{1/s} + |u|_{\mathcal{A}^s}^{1/s})\}$.*

Proof. For $\varepsilon \geq \varepsilon_0$ there is nothing to prove, so let us assume that $\varepsilon < \varepsilon_0$. By induction on i we prove that at termination of the if-then-else-fi clause inside the loop over i,

$$|u - \bar{u}_{P^a}|_1 \leq (1 - 3\delta)\varepsilon_i. \tag{7.1}$$

For $i = 1$, this follows from the input condition on ε_0. Let us now assume (7.1) for some $i \geq 1$. Then, after the call $[P^a, f_{P^a}] := \mathbf{RHS}[P^a, f, \delta\varepsilon_i]$, by definition

$$\|f - f_{P^a}\|_{-1} \leq \delta\varepsilon_i. \tag{7.2}$$

So, for $u_{P^a} := L_{P^a}^{-1} f_{P^a}$ we have

$$\begin{aligned} |u - u_{P^a}|_1 &\leq |u - L^{-1}f_{P^a}|_1 + |L^{-1}f_{P^a} - u_{P^a}|_1 \leq |u - L^{-1}f_{P^a}|_1 + |L^{-1}f_{P^a} - \bar{u}_{P^a}|_1 \\ &\leq 2|u - L^{-1}f_{P^a}|_1 + |u - \bar{u}_{P^a}|_1 = 2\|f - f_{P^a}\|_{-1} + |u - \bar{u}_{P^a}|_1 \\ &\leq 2\delta\varepsilon_i + (1 - 3\delta)\varepsilon_i = (1 - \delta)\varepsilon_i, \end{aligned} \tag{7.3}$$

where for the second inequality we have used that $\bar{u}_{P^a} \in \mathcal{S}_{P^a}$ and that u_{P^a} is the best approximation with respect to $|\cdot|_1$ of $L^{-1}f_{P^a}$ from $\mathcal{S}_{P^a}$. We conclude that after the update of $\bar{u}_{P^a}$ by the call of **GALSOLVE**,

$$|u_{P^a} - \bar{u}_{P^a}|_1 \leq \delta\varepsilon_i \text{ and so } |u - \bar{u}_{P^a}|_1 \leq |u - u_{P^a}|_1 + |u_{P^a} - \bar{u}_{P^a}|_1 \leq \varepsilon_i. \tag{7.4}$$

After the first calls of **REFINE**, **RHS**, and **GALSOLVE** in the inner loop, i.e., when $j = 1$, for the new P^a, f_{P^a}, $\bar{u}_{P^a}$, $u_{P^a} := L_{P^a}^{-1}f_{P^a}$ we have $\|f - f_{P^a}\|_{-1} \leq \delta\mu\varepsilon_i$ and $|u_{P^a} - \bar{u}_{P^a}|_1 \leq \delta\mu\varepsilon_i$, and so by (7.2), (7.4), Corollary 7.1 shows that $|u - \bar{u}_{P^a}|_1 \leq \mu\varepsilon_i$. Repeating this argument for $j = 2, \ldots, M$ shows that at termination of the inner loop

over j, we have that $|u-\bar{u}_{P^a}|_1 \leq \mu^M \varepsilon_i$. In particular, when $i=N$, we have that $|u-\bar{u}_{P^a}|_1 \leq \mu^M \varepsilon_N = (\frac{\mu^M}{1-3\delta})^N((1+\gamma)\kappa_{\bar{\Psi}}^{\frac{1}{2}})^{N-1}\varepsilon_0 \leq \varepsilon$ by definition of N. Otherwise, if $i<N$, then in the next iteration, thus after increasing i by one, just before the call of **COARSE**, we have that $\|\!|u-\bar{u}_{P^a}|\!\|_1 \leq \lambda_{\bar{\Psi}}^{-\frac{1}{2}}|u-\bar{u}_{P^a}|_1 \leq \lambda_{\bar{\Psi}}^{-\frac{1}{2}}\mu^M\varepsilon_{i-1}$. By Corollary 5.5, after this call we have

$$|u-\bar{u}_{P^a}|_1 \leq \Lambda_{\bar{\Psi}}^{\frac{1}{2}}\|\!|u-\bar{u}_{P^a}|\!\|_1 \leq \Lambda_{\bar{\Psi}}^{\frac{1}{2}}(1+\gamma)\lambda_{\bar{\Psi}}^{-\frac{1}{2}}\mu^M\varepsilon_{i-1} = (1-3\delta)\varepsilon_i, \tag{7.5}$$

which completes the proof of (7.1), and thus that of $|u-\bar{u}_{P^a}|_1 \leq \varepsilon$ at termination of **SOLVE1**.

Now we will prove that for any $i=1,\ldots,N$, both $\#P^a$ at the end of the outer cycle for this i and the cost of this cycle excluding, for $i>1$, the cost of the **COARSE**, but including, for $i<N$, the cost of the **COARSE** in the next cycle, are $\lesssim \varepsilon_i^{-1/s}(c_f^{1/s}+|u|_{\mathcal{A}^s}^{1/s})$. Because of $\sum_{i=1}^N \varepsilon_i^{-1/s} \lesssim \varepsilon_N^{-1/s} \leq \varepsilon^{-1/s}$ this will prove the statement about the complexity.

At the start of the outer cycle for $i=1$, we have $\#P^a \lesssim \varepsilon_i^{-1/s}|u|_{\mathcal{A}^s}^{1/s}$, which follows from $\#P^a = \#P_0 \lesssim 1$ and the assumption that $\varepsilon_0 \lesssim |u|_1$. Since, as we have seen, for $i>1$ before the call of **COARSE**, we have that $\|\!|u-\bar{u}_{P^a}|\!\|_1 \leq \lambda_{\bar{\Psi}}^{-\frac{1}{2}}\mu^M\varepsilon_{i-1}$, Corollary 5.5 shows that after this call, $\#P^a \leq D\#\hat{P}$ for any partition $\hat{P}$ with $\inf_{u_{\hat{P}}\in\mathcal{S}_{\hat{P}}}\|\!|u-u_{\hat{P}}|\!\|_1 \leq (t_1^{\frac{1}{2}}\gamma-1)\lambda_{\bar{\Psi}}^{-\frac{1}{2}}\mu^M\varepsilon_{i-1}$. From $\|\!|\cdot|\!\|_1 \leq \lambda_{\bar{\Psi}}^{-\frac{1}{2}}|\cdot|_1$ and $u\in\mathcal{A}^s$, we find that

$$\#P^a \leq D(\#P_0 + [(t_1^{\frac{1}{2}}\gamma-1)\mu^M\varepsilon_{i-1}]^{-1/s}|u|_{\mathcal{A}^s}^{1/s}) \lesssim \varepsilon_i^{-1/s}|u|_{\mathcal{A}^s}^{1/s}.$$

Since $(f,\mathbf{RHS})$ is s-optimal and M is a fixed constant, from the properties of **RHS** and **REFINE** we conclude that for any i, at the end of the outer cycle

$$\#P^a \lesssim \varepsilon_i^{-1/s}(c_f^{1/s}+|u|_{\mathcal{A}^s}^{1/s}), \tag{7.6}$$

whereas the cost of all calls of **RHS** and **REFINE** inside this cycle are also $\lesssim \varepsilon_i^{-1/s}(c_f^{1/s}+|u|_{\mathcal{A}^s}^{1/s})$. Furthermore, for $i<N-1$, Theorem 5.4(b) shows that the cost of **COARSE** in the next iteration is $\lesssim \#P^a + \max\{0,\log(\varepsilon_i^{-1}\|\!|\bar{u}_{P^a}|\!\|_1)\}$. From $\log(\varepsilon_i^{-1}\|\!|\bar{u}_{P^a}|\!\|_1) \leq \varepsilon_i^{-1/s}\|\!|\bar{u}_{P^a}|\!\|_1^{1/s}$, $\|\!|\bar{u}_{P^a}|\!\|_1 \lesssim |\bar{u}_{P^a}|_1 \leq |u|_1+\varepsilon_0 \lesssim |u|_1 \leq |u|_{\mathcal{A}^s}$, and (7.6), we conclude also that these costs are $\lesssim \varepsilon_i^{-1/s}(c_f^{1/s}+|u|_{\mathcal{A}^s}^{1/s})$.

What is left is to bound the cost of the applications of **GALSOLVE**. As we have seen, just before the call **GALSOLVE**$[P^a, f_{P^a}, \bar{u}_{P^a}, \delta\varepsilon_i]$ outside the inner loop over j, we have that $|u-\bar{u}_{P^a}|_1 \leq (1-3\delta)\varepsilon_i$, and with $u_{P^a} := L_{P^a}^{-1}f_{P^a}$, $|u-u_{P^a}|_1 \leq (1-\delta)\varepsilon_i$, and so $|u_{P^a}-\bar{u}_{P^a}|_1 \leq 2(1-2\delta)\varepsilon_i$. Since $\frac{2(1-2\delta)\varepsilon_i}{\delta\varepsilon_i}$ is a constant, we conclude that the cost of this call is $\lesssim \#P^a$.

Let us now consider a call **GALSOLVE**$[P^a, f_{P^a}, \bar{u}_{P^a}, \delta\mu^j\varepsilon_i]$ inside the loop over j. Just before this call we have that $|u-\bar{u}_{P^a}|_1 \leq \mu^{j-1}\varepsilon_i$ and $\|f-f_{P^a}\|_{-1} \leq \delta\mu^j\varepsilon_i$. As in (7.3), for $u_{P^a} := L_{P^a}^{-1}f_{P^a}$ we have

$$|u-u_{P^a}|_1 \leq 2\|f-f_{P^a}\|_{-1} + |u-\bar{u}_{P^a}|_1 \leq (2\delta\mu^j+\mu^{j-1})\varepsilon_i,$$

and so $|u_{P^a}-\bar{u}_{P^a}|_1 \leq 2(\delta\mu^j+\mu^{j-1})\varepsilon_i$. Since $\frac{(2\delta\mu^j+\mu^{j-1})\varepsilon_i}{\delta\mu^j\varepsilon_i}$ is a constant, we conclude that the cost of this call is $\lesssim \#P^a$, which completes the proof. ∎

8. An optimal adaptive finite element method with a posteriori error control. It follows from the proof of Theorem 7.4 that the approximations $\bar{u}_{P^a}$ on the sequence of partitions produced by **SOLVE1** converge with an asymptotic rate $\leq (\frac{(1+\gamma)\kappa_{\bar{\Psi}}^{\frac{1}{2}}\mu^M}{1-3\delta})^{1/(M+1)}$, which is close to μ when M is not too small. Because of the application of a coarsening, the asymptotic rate is generally even equal to the above number. Indeed, after the evaluation of $[P^a, \bar{u}_{P^a}] := \mathbf{COARSE}[P^a, `\bar{u}_{P^a}^{old}', \gamma\lambda_{\bar{\Psi}}^{-\frac{1}{2}}\mu^M\varepsilon_{i-1}]$, we have that

$$|u-\bar{u}_{P^a}|_1 \geq \lambda_{\bar{\Psi}}^{\frac{1}{2}}\|\!|u-\bar{u}_{P^a}|\!\|_1 \geq \lambda_{\bar{\Psi}}^{\frac{1}{2}}\left|\|\!|\bar{u}_{P^a}-\bar{u}_{P^a}^{old}|\!\|_1 - \|\!|u-\bar{u}_{P^a}^{old}|\!\|_1\right| \geq ((\gamma-\eta)-1)\mu^M\varepsilon_{i-1},$$

where η can be arbitrary small, so that this lower bound is only by a constant factor smaller than the upper bound for $|u-\bar{u}_{P^a}|_1$ from (7.5). The value μ has to be supplied by the user. It should be large enough to ensure that indeed $\lambda_{\bar{\Psi}}^{-\frac{1}{2}}\mu^M\varepsilon_{i-1}$ is an upper bound for $\|\!|u-\bar{u}_{P^a}^{old}|\!\|_1$, so that the quasi-optimality of the partition after **COARSE** is guaranteed by Corollary 5.5. A safe choice of μ will be the result of a worst-case analysis, and so likely it will be unnecessarily close to 1, resulting in a quantitatively less attractive algorithm. All adaptive finite element or wavelet methods based on coarsening introduced so far share this drawback that a judicious choice of such a parameter μ has to be made.

In this final subsection, we develop a modified routine **SOLVE2** in which the tolerances used in the routines **COARSE**, **RHS**, and **GALSOLVE** will depend on an a posteriori estimate of the error, instead of an a priori one. Moreover, as is also suggested in [7], instead of performing a fixed number of iterations of the **REFINE**, **RHS**, **GALSOLVE** triple between two applications of **COARSE**, in **SOLVE2** the iteration is stopped as soon as sufficient reduction of the error has taken place as indicated by the a posteriori error estimator. We will use our estimator $\mathcal{E}$, extended with some terms to incorporate the error in the right-hand side and that as a consequence of the inexact solution of the discrete system; cf. (8.1). One problem is that a monotone decrease of the a posteriori error estimates cannot be guaranteed. In Proposition 8.1, however, we show that the error estimates are equivalent to quantities that do decrease linearly; cf. (8.2) and (8.4).

PROPOSITION 8.1. *Let $C_4 := 1 + C_1C_3$, where $C_1, C_3 > 0$ are the constants from Theorem* 6.1 *and Lemma* 6.5. *For $f \in H^{-1}(\Omega)$, any admissible partition P^a, $\bar{u}_{P^a} \in \mathcal{S}_{P^a}$, and $f_{P^a} \in L_2(\Omega)$, with $u = L^{-1}f$, $u_{P^a} = L_{P^a}^{-1}f_{P^a}$ we have*

$$|u-\bar{u}_{P^a}|_1 \leq C_1\mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + \|f-f_{P^a}\|_{-1} + C_4|u_{P^a}-\bar{u}_{P^a}|_1. \tag{8.1}$$

With $f_{P^a} \in \mathcal{S}_{P^a}^0$, and ζ_{P^a} being an upper bound for $\|f-f_{P^a}\|_{-1}+C_4|u_{P^a}-\bar{u}_{P^a}|_1$, we have that

$$C_1\mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + \zeta_{P^a} \eqsim |u-\bar{u}_{P^a}|_1 + \zeta_{P^a}. \tag{8.2}$$

Finally, for any $\mu \in ([1-\frac{1}{2}(\frac{c_2\theta}{C_1})^2]^{\frac{1}{2}}, 1)$, there exists $\bar{\delta} > 0$ small enough and $C_5 > 0$ large enough, such that if $\tilde{P} = \mathbf{REFINE}[P^a, f_{P^a}, \bar{u}_{P^a}]$ or a refinement of it, $\bar{u}_{\tilde{P}} \in \mathcal{S}_{\tilde{P}}$, $f_{\tilde{P}} \in H^{-1}(\Omega)$, $u_{\tilde{P}} := L_{\tilde{P}}^{-1}f_{\tilde{P}}$, and

$$\zeta_{\tilde{P}} \leq \bar{\delta}(1+C_4)[C_1\mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + \zeta_{P^a}], \tag{8.3}$$

where thus $\zeta_{\tilde{P}}$ denotes an upper bound on $\|f-f_{\tilde{P}}\|_{-1} + C_4|u_{\tilde{P}}-\bar{u}_{\tilde{P}}|_1$; then

$$|u-\bar{u}_{\tilde{P}}|_1 + C_5\zeta_{\tilde{P}} \leq \mu[|u-\bar{u}_{P^a}|_1 + C_5\zeta_{P^a}]. \tag{8.4}$$

Proof. With $\hat{u} := L^{-1} f_{P^a}$, (8.1) follows from Theorem 6.1 and Lemma 6.5 by

$$\begin{aligned}|u - \bar{u}_{P^a}|_1 &\le |u - \hat{u}|_1 + |\hat{u} - u_{P^a}|_1 + |u_{P^a} - \bar{u}_{P^a}|_1 \\ &\le \|f - f_{P^a}\|_{-1} + C_1 \mathcal{E}(P^a, f_{P^a}, u_{P^a}) + |u_{P^a} - \bar{u}_{P^a}|_1 \\ &\le \|f - f_{P^a}\|_{-1} + C_1 \mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + (1 + C_1 C_3)|u_{P^a} - \bar{u}_{P^a}|_1.\end{aligned}$$

In one direction, (8.2) follows immediately from (8.1), whereas in the other direction it is a consequence of Corollary 6.4 in combination with Lemma 6.5.

Theorem 6.7 shows that with $C_6 := \max\{\frac{1+2c_2C_3}{C_4}, 3\}$,

$$|u - \bar{u}_{\tilde{P}}|_1 + C_5 \zeta_{\tilde{P}} \le \left[1 - \frac{1}{2}\left(\frac{c_2\theta}{C_1}\right)^2\right]^{\frac{1}{2}} |u - \bar{u}_{P^a}|_1 + C_6(\zeta_{P^a} + \zeta_{\tilde{P}}) + C_5 \zeta_{\tilde{P}},$$

which, given a $\mu \in ([1 - \frac{1}{2}(\frac{c_2\theta}{C_1})^2]^{\frac{1}{2}}, 1)$, is less than or equal to $\mu[|u - \bar{u}_{P^a}|_1 + C_5 \zeta_{P^a}]$ if and only if

$$(C_6 + C_5)\zeta_{\tilde{P}} \le \left(\mu - \left[1 - \frac{1}{2}\left(\frac{c_2\theta}{C_1}\right)^2\right]^{\frac{1}{2}}\right)|u - \bar{u}_{P^a}|_1 + (\mu C_5 - C_6)\zeta_{P^a}.$$

Hence, by selecting the constant $C_5 > \frac{C_6}{\mu}$, the proof of (8.4) is completed by observing that for $\bar{\delta}$ small enough,

$$\bar{\delta}(1+C_4)[C_1 \mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + \zeta_{P^a}] \le \frac{(\mu - [1 - \frac{1}{2}(\frac{c_2\theta}{C_1})^2]^{\frac{1}{2}})|u - \bar{u}_{P^a}|_1 + (\mu C_5 - C_6)\zeta_{P^a}}{C_6 + C_5},$$

which is a consequence of (8.2). □

We are ready to formulate the adaptive finite element method **SOLVE2** in which the tolerances are controlled by the a posteriori error estimator. Any faster convergence of the approximations produced by the **REFINE**, **RHS**, **GALSOLVE** triple than appears from a priori estimates can be expected to lead to better quantitative properties for **SOLVE2** than for **SOLVE1**.

SOLVE2$[f, \varepsilon, \bar{u}_{P_0}, \varepsilon_0] \to [P^a, \bar{u}_{P^a}]$:
% The following constants are fixed: $\bar{\delta}$ is small enough so that it corresponds to a
% $\mu < 1$ as in Proposition 8.1; $\gamma > t_1^{-\frac{1}{2}}$ with t_1 as in Proposition 5.3; and $\sigma \in (0,1)$.
% The input must satisfy $f \in H^{-1}(\Omega)$ such that a valid routine **RHS** *is available,*
% $\varepsilon > 0$, $\bar{u}_{P_0} \in \mathcal{S}_{P_0}$ and $\varepsilon_0 \ge |u - \bar{u}_{P_0}|_1$.
$P^a := P_0$, $f_{P^a} := f_{P_0}$, $\bar{u}_{P^a} := \bar{u}_{P_0}$, $\bar{e} := \tilde{e} := \varepsilon_0$
`while` $\bar{e} > \varepsilon$ `do`
  `if` *not first iteration* `then`
    $[P^a, \bar{u}_{P^a}] := \mathbf{COARSE}[P^a, \bar{u}_{P^a}, \gamma \lambda_{\bar{\Psi}}^{-\frac{1}{2}} \bar{e}]$, $\tilde{e} := (1+\gamma)\kappa_{\bar{\Psi}}^{\frac{1}{2}} \bar{e}$
  `fi`
  $[P^a, f_{P^a}] := \mathbf{RHS}[P^a, f, \bar{\delta}\tilde{e}]$
  $\bar{u}_{P^a} := \mathbf{GALSOLVE}[P^a, f_{P^a}, \bar{u}_{P^a}, \bar{\delta}\tilde{e}]$
  $\tilde{e} := C_1 \mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + (1 + C_4)\bar{\delta}\tilde{e}$
  `while` $\tilde{e} > \sigma \bar{e}$ `do`
    $P := \mathbf{REFINE}[P^a, f_{P^a}, \bar{u}_{P^a}]$
    $[P^a, f_{P^a}] := \mathbf{RHS}[P, f, \bar{\delta}\tilde{e}]$
    $\bar{u}_{P^a} := \mathbf{GALSOLVE}[P^a, f_{P^a}, \bar{u}_{P^a}, \bar{\delta}\tilde{e}]$
    $\tilde{e} := C_1 \mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + (1 + C_4)\bar{\delta}\tilde{e}$
  `od`
  $\bar{e} := \tilde{e}$
`od`

The next theorem shows that **SOLVE2** is an optimal method whenever this is allowed by the $(f, \mathbf{RHS})$ pair.

THEOREM 8.2. *$[P^a, \bar{u}_{P^a}] := \mathbf{SOLVE1}[f, \varepsilon, \bar{u}_{P_0}, \varepsilon_0]$ satisfies $|u - \bar{u}_{P^a}|_1 \leq \varepsilon$. Assuming $\varepsilon_0 \lesssim |u|_1$, if for some $s > 0$, $u \in \mathcal{A}^s$ and $(f, \mathbf{RHS})$ is s-optimal, then both $\#P^a$ and the number of arithmetic operations required by this call are $\lesssim \max\{1, \varepsilon^{-1/s}(c_f^{1/s} + |u|_{\mathcal{A}^s}^{1/s})\}$.*

Proof. At the beginning of a cycle of the outer while-loop, we have that $|u - \bar{u}_{P^a}|_1 \leq \bar{e}$, which for a cycle other than the first one is a consequence of Proposition 8.1. In particular, when the outer loop terminates, we have $|u - \bar{u}_{P^a}|_1 \leq \varepsilon$.

After the if-then-fi clause, we have that $|u - \bar{u}_{P^a}|_1 \leq \tilde{e}$, where $\tilde{e} = \bar{e}$ for the first iteration, and $\tilde{e} = (1+\gamma)\kappa_{\Psi}^{\frac{1}{2}}\bar{e}$ otherwise (apply Corollary 5.5 and (4.6)). After the call of **RHS**, by definition we have $\|f - f_{P^a}\|_{-1} \leq \bar{\delta}\tilde{e}$, and so as in (7.3), for $u_{P^a} := L_{P^a}^{-1} f_{P^a}$ we have $|u - u_{P^a}|_1 \leq 2\|f - f_{P^a}\|_{-1} + |u - \bar{u}_{P^a}|_1 \leq (2\bar{\delta} + 1)\tilde{e}$. After the call of **GALSOLVE**, by definition we have $|u_{P^a} - \bar{u}_{P^a}|_1 \leq \bar{\delta}\tilde{e}$, and so $|u - \bar{u}_{P^a}|_1 \leq (3\bar{\delta} + 1)\tilde{e}$. By applying (8.2), these estimates show that just before starting the inner while-loop, the new $\tilde{e}$ satisfies

$$\tilde{e} \leq C\bar{e} \tag{8.5}$$

for some absolute constant $C > 0$.

Let us now consider any newly computed $\bar{u}_{P^a}$ in the inner while-loop, and let us denote by τ_1, τ_2 the tolerances that were used in the corresponding calls of **RHS** and **GALSOLVE** and let $\zeta_{P^a} = \tau_1 + C_4\tau_2$. We have that ζ_{P^a} is equal to $(1 + C_4)\bar{\delta}[C_1\mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + \zeta_{P^a}]$, where in the latter expression P^a, f_{P^a}, $\bar{u}_{P^a}$, and ζ_{P^a} refer to the previous partition, right-hand side, approximate solution, and ζ_{P^a}, respectively. Since by assumption $\bar{\delta}$ corresponds to a $\mu < 1$ as in Proposition 8.1, formula (8.4) shows that in each iteration of the inner loop $|u - \bar{u}_{P^a}|_1 + C_5\zeta_{P^a}$ is multiplied by a factor $\leq \mu$. Since by (8.2), $C_1\mathcal{E}(P^a, f_{P^a}, \bar{u}_{P^a}) + \zeta_{P^a} \eqsim |u - \bar{u}_{P^a}|_1 + C_5\zeta_{P^a}$, the geometric decrease of $|u - \bar{u}_{P^a}|_1 + C_5\zeta_{P^a}$ together with (8.5) shows that the inner while-loop terminates within an (absolute) constant number of iterations. After termination of the inner while-loop, the new $\bar{e}$ will be less than or equal to $\sigma < 1$ times the previous $\bar{e}$, showing that **SOLVE2** terminates, with, as we have seen, $|u - \bar{u}_{P^a}|_1 \leq \varepsilon$.

Let us consider any cycle of the outer loop with $\bar{e} > \varepsilon$ being the value at the beginning of this cycle. After the if-then-fi clause, we have that $\#P^a \lesssim \bar{e}^{-1/s}|u|_{\mathcal{A}^s}^{1/s}$, which for the first iteration follows from $\#P_0 \lesssim 1$ and $\varepsilon_0 \lesssim |u|_1$ by assumption, and which for any other cycle follows from Corollary 5.5 analogously as in the proof of Theorem 7.4. The properties of **RHS** and **REFINE** and the fact that the inner while-loop terminates within a fixed number of iterations show that at termination of this outer cycle, $\#P^a \lesssim \bar{e}^{-1/s}(c_f^{1/s} + |u|_{\mathcal{A}^s}^{1/s})$, which, in particular, proves the statement about $\#P^a$ at termination of **SOLVE2**. Furthermore, the cost of all calls of **RHS** and **REFINE**, as well as the cost of **COARSE**, in the possibly next cycle are $\lesssim \bar{e}^{-1/s}(c_f^{1/s} + |u|_{\mathcal{A}^s}^{1/s})$. Assuming, for the moment, that the cost of any application of **GALSOLVE** in **SOLVE2** on a partition P^a is $\lesssim \#P^a$, by the geometric decrease of the values of $\bar{e}$ at the beginning of the outer while-loop we conclude that the total cost of **SOLVE2** is $\lesssim \varepsilon^{-1/s}(c_f^{1/s} + |u|_{\mathcal{A}^s}^{1/s})$.

As we have seen, just before the evaluation of $\mathbf{GALSOLVE}[P^a, f_{P^a}, \bar{u}_{P^a}, \bar{\delta}\tilde{e}]$ outside the inner while-loop, we have $|u - \bar{u}_{P^a}|_1 \leq \tilde{e}$, and with $u_{P^a} := L_{P^a}^{-1} f_{P^a}$, $|u - u_{P^a}|_1 \leq (2\bar{\delta} + 1)\tilde{e}$ and so $|u_{P^a} - \bar{u}_{P^a}|_1 \leq 2(\bar{\delta} + 1)\tilde{e}$. We conclude that the cost of this call is $\lesssim \#P^a$.

Analogously, just before an evaluation of **GALSOLVE**$[P^a, f_{P^a}, \bar{u}_{P^a}, \bar{\delta}\tilde{e}]$ inside the inner while-loop, we have $|u - \bar{u}_{P^a}|_1 \leq \tilde{e}$, $\|f - f_{P^a}\|_{-1} \leq \bar{\delta}\tilde{e}$, and so, as in (7.3), with $u_{P^a} := L_{P^a}^{-1} f_{P^a}$, $|u - u_{P^a}|_1 \leq 2\bar{\delta}\tilde{e} + \tilde{e}$ and so $|u_{P^a} - \bar{u}_{P^a}|_1 \leq 2(\bar{\delta}\tilde{e} + \tilde{e})$. We conclude that the cost of such a call is also $\lesssim \#P^a$. □

Acknowledgments. The author would like to thank Peter Binev (University of South Carolina) for his explanations concerning adaptive tree approximations and both referees for numerous valuable comments.

REFERENCES

[1] A. Barinka, *Fast Evaluation Tools for Adaptive Wavelet Schemes*, Ph.D. thesis, RTWH Aachen, 2004.

[2] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Numer. Math., 85 (2000), pp. 579–608.

[3] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.

[4] P. Binev, W. Dahmen, R. DeVore, and P. Petruchev, *Approximation classes for adaptive methods*, Serdica Math. J., 28 (2002), pp. 391–416.

[5] P. Binev and R. DeVore, *Fast computation in adaptive tree approximation*, Numer. Math., 97 (2004), pp. 193–217.

[6] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations: Convergence rates*, Math. Comp, 70 (2001), pp. 27–75.

[7] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet schemes for nonlinear variational problems*, SIAM J. Numer. Anal., 41 (2003), pp. 1785–1823.

[8] A. Cohen, L. Echeverry, and Q. Sun, *Finite Element Wavelets*, Technical report, Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Paris, France, 2000.

[9] A. Cohen and J. Schlenker, *Compactly supported bidimensional wavelet bases with hexagonal symmetry*, Constr. Approx., 9 (1993), pp. 209–236.

[10] S. Dahlke, *Besov regularity for elliptic boundary value problems in polygonal domains*, Appl. Math. Lett., 12 (1999), pp. 31–36.

[11] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[12] M. Dryja, M. V. Sarkis, and O. B. Widlund, *Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions*, Numer. Math., 72 (1996), pp. 313–348.

[13] R. Lorentz and P. Oswald, *Multilevel finite element Riesz bases in Sobolev spaces*, in Proceedings of the 9th Symposium on Domain Decomposition Methods, P. Bjørstad, M. Espedal, and D. Keyes, eds., John Wiley & Sons, 1996.

[14] A. Metselaar, *Handling Wavelet Expansions in Numerical Methods*, Ph.D. thesis, University of Twente, Enschede, The Netherland, 2002.

[15] P. Morin, R. Nochetto, and K. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[16] P. Oswald, *Multilevel Finite Element Approximation: Theory and Applications*, B.G. Teubner, Stuttgart, 1994.

[17] M. Petzoldt, *A posteriori error estimators for elliptic equations with discontinuous coefficients*, Adv. Comput. Math., 16 (2002), pp. 47–75.

[18] R. Stevenson, *Piecewise linear (pre-)wavelets on non-uniform meshes*, in Multigrid Methods V, Proceedings of the Fifth European Multigrid Conference held in Stuttgart, Germany, October 1–4, 1996, H. W. and G. Wittum, eds., Lect. Notes Comput. Sci. Eng. 3, Springer-Verlag, Heidelberg, 1998, pp. 306–319.

[19] R. Stevenson, *Stable three-point wavelet bases on general meshes*, Numer. Math., 80 (1998), pp. 131–158.

[20] R. Stevenson, *Adaptive solution of operator equations using wavelet frames*, SIAM J. Numer. Anal., 41 (2003), pp. 1074–1100.

SIAM J. NUMER. ANAL.
Vol. 42, No. 5, pp. 2218–2256

EFFICIENT COMPUTATION OF THE MATRIX EXPONENTIAL BY GENERALIZED POLAR DECOMPOSITIONS*

ARIEH ISERLES† AND ANTONELLA ZANNA‡

Abstract. In this paper we explore the computation of the matrix exponential in a manner that is consistent with Lie-group structure. Our point of departure is the method of generalized polar decompositions, which we modify and combine with similarity transformations that bring the underlying matrix to a form more amenable to efficient computation. We develop techniques valid for a range of Lie groups: the orthogonal group, the symplectic group, Lorentz, isotropy, and scaling groups. However, the GPD approach is equally promising in a more general context. Even when Lie-group structure is not at issue, our algorithm is more efficient in many settings than classical methods for the computation of the matrix exponential.

Key words. matrix exponential, Lie group, Lie algebra, generalized polar decomposition

AMS subject classifications. 65F30

DOI. 10.1137/S0036142902415936

1. Introduction.

The approximation of the matrix exponential is among the oldest and most extensively researched problems in numerical mathematics. Yet, nineteen dubious ways [15] and many efficient algorithms (cf., for example, [8]) later, the problem is far from being satisfactorily solved and many challenges remain. This is true in particular when we wish to approximate an exponential of a matrix Z, say, which resides in a *Lie algebra.* This is a central problem in *geometric integration,* which arises once we wish to discretize systems of differential equations evolving in *Lie groups* (smooth manifolds with group structure) and in *homogeneous manifolds* (smooth manifolds which are subjected to transitive group action).

While referring the reader to [10] for a substantive survey of Lie-group methods and their applications, and to section 2 for formal definitions, it is important to mention informally a number of salient features of such methods, since they motivate much of the work of the present paper.

- The tangent space $\mathrm{T}_x G$, where G is a Lie group and $x \in G$, is $\{Zx : Z \in \mathfrak{g}\}$, where $\mathfrak{g} = \mathrm{T}_I G$ and I is the identity of G. Therefore, once we know $\mathfrak{g}$, we can describe all vector fields (hence all differential equations) on G.
- The linear space $\mathfrak{g}$ is a Lie algebra: it is closed under an antisymmetric binary operation of *commutation.*
- The *exponential map* takes the Lie algebra to "its" Lie group, $\exp \mathfrak{g} \subseteq G$.
- Most finite-dimensional Lie groups in practical applications are comprised of matrices. Familiar examples are the general linear group $\mathrm{GL}(\mathbb{R}, n)$ ($n \times n$ nonsingular real matrices), the special linear group $\mathrm{SL}(\mathbb{R}, n)$ ($n \times n$ real matrices with unit determinant), and the orthogonal group $\mathrm{O}(\mathbb{R}, n)$ ($n \times n$ real orthogonal matrices).

*Received by the editors October 8, 2002; accepted for publication (in revised form) January 11, 2004; published electronically March 25, 2005.
http://www.siam.org/journals/sinum/42-5/41593.html
†Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Silver Street, Cambridge CB3 9EW, UK (A.Iserles@damtp.cam.ac.uk).
‡Department of Informatics, University of Bergen, Høyteknologisenteret, N-5020 Bergen, Norway (Antonella.Zanna@ii.uib.no).

• All finite-dimensional Lie algebras are isomorphic to Lie algebras of matrices. In particular, the Lie algebras corresponding to the three Lie groups above are $\mathfrak{gl}(\mathbb{R}, n)$ (the $n \times n$ real matrices), $\mathfrak{sl}(\mathbb{R}, n)$ ($n \times n$ real matrices with zero trace), and $\mathfrak{so}(\mathbb{R}, n)$ ($n \times n$ real skew-symmetric matrices), respectively.

• If G is a matrix group (hence $\mathfrak{g}$ is a matrix algebra) the operations of commutation and exponentiation are the familiar matricial commutator and exponent, respectively.

Therefore, once a differential equation evolves in a matrix Lie group, it can be always written in the form

$$y' = F(t, y)y, \quad t \geq 0, \qquad y(0) \in G, \tag{1.1}$$

where $F : \mathbb{R}_+ \times G \to \mathfrak{g}$. Moreover, its solution can be represented (subject to the usual caveats of convergence) in the form $y(t) = \exp(\Omega(t))y(0)$, where Ω evolves in the Lie algebra $\mathfrak{g}$. It is possible to replace (1.1) by an equation for Ω, which evolves in $\mathfrak{g}$ [10], and there are important benefits in solving the latter, returning to the Lie group in every time step by means of the exponential map. The main advantage is that $\mathfrak{g}$ is a linear space and, as long as we discretize equations therein employing exclusively linear-space operations and commutators, we can be assured that the numerical solution stays in $\mathfrak{g}$. Thus, once exponentiated, we obtain a numerical solution that evolves in the Lie group; this is important in the many instances when the preservation of Lie-group structure is important and in variance with most numerical methods applied directly in G [10].

The above argument is at the heart of many Lie-group methods (Runge–Kutta–Munthe-Kaas schemes, Magnus expansions). Other methods, based on different premises (e.g., Crouch–Grossman methods, Fer expansions and methods based upon canonical coordinates of the second kind) also require the computation (or approximation) of the matrix exponential. However, standard methods for the approximation of the matrix exponential, e.g., Padé approximations and Krylov subspace techniques, are not guaranteed to map elements from $\mathfrak{g}$ to G. Thus, having gone to a great length to respect Lie-algebraic structure, we might well loose the fruits of this endeavor while computing the exponential! On the positive side, diagonal Padé approximations map some Lie algebras ("quadratic" algebras: $\mathfrak{so}(\mathbb{R}, n)$, the symplectic algebra, the Lorentz algebra) to the underlying group. However, it is possible to show that the only analytic function f, that maps $\mathfrak{sl}(\mathbb{R}, n)$ into $\mathrm{SL}(\mathbb{R}, n)$ consistently with the exponential function (i.e., $f(z) = 1 + z + \mathcal{O}(z^2)$) is the exponential itself [11]. Also other classical methods for the approximation of the exponential fail in that case, and this motivates the development of new breeds of approximation algorithms.

Early inroads into the approximation of the exponential in a Lie-algebraic setting have been made in [2], using the splitting approach,

$$\mathrm{e}^{tZ} \approx \mathrm{e}^{tV_1}\mathrm{e}^{tV_2} \cdots \mathrm{e}^{tV_m},$$

where each V_k resides in $\mathfrak{g}$ and the computation of its exponential is easy. The latter is true when the V_ks are of low rank and this, indeed, was the approach introduced in [2].

Suppose that $\dim \mathfrak{g} = s$ and let $\boldsymbol{X} = \{X_1, X_2, \ldots, X_s\}$ be a basis of $\mathfrak{g}$. In that case it is possible to represent $\exp(tZ)$ for $Z \in \mathfrak{g}$ and sufficiently small $|t|$ in *canonical coordinates of the second kind,*

$$\mathrm{e}^{tZ} = \mathrm{e}^{g_1(t)X_1}\mathrm{e}^{g_2(t)X_2} \cdots \mathrm{e}^{g_s(t)X_s},$$

where the scalar functions g_k are analytic at the origin. Although the g_ks are implicitly defined, it is possible to approximate their truncated Taylor expansion, an approach adopted in [3]. A naive procedure of this kind might be excessively expensive, but the cost can be reduced by several orders of magnitude by a clever choice of the basis $\boldsymbol{X}$, exploiting the Lie-algebraic structure.

The work underlying the approach of the present paper, *generalized polar decompositions (GPD),* has been introduced in [16] and further elaborated on in [20, 21]. In section 2 we present a brief review of such methods. It suffices to state here that, while building upon former work in this area, they establish a general framework which leads to robust and affordable algorithms. Having said this, such algorithms can be fairly expensive when the required order is high, in particular when they need to be computed, perhaps repeatedly, in each time step. The purpose of this paper is to bring together generalized polar decompositions with techniques from numerical linear algebra, thereby leading to more efficient and cheaper algorithms.

At a conceptual level, we are attempting to marry two types of structures, which are often incompatible. For example, a viable approach to compute $\exp tZ$ for $Z \in \mathfrak{gl}(\mathbb{R}, n)$ is to represent $Z = VHV^*$, where V is a product of Householder reflections and H is upper Hessenberg. Since this is a similarity transformation, it is true that $\mathrm{e}^{tZ} = V\mathrm{e}^{tH}V^*$ and we need to compute an exponential of an upper-Hessenberg matrix. As we show in what follows, this can be done very efficiently indeed by a modification of the GPD technique. Unfortunately, this approach cannot be extended to other Lie algebras. Thus, suppose that Z resides in the *symplectic algebra*

$$\mathfrak{sp}(n) = \{Y \in \mathfrak{gl}(\mathbb{R}, 2n) : YJ + JY^\top = O\}, \quad \text{where } J = \begin{bmatrix} O & I \\ -I & O \end{bmatrix}.$$

In that case, in general, $H \notin \mathfrak{sp}(n)$: a Hessenberg form and symplecticity are incompatible! This is an illustration of a more general state of affairs, when numerical-algebraic and Lie-algebraic structures clash. In this paper we present numerical-algebraic structures which are compatible with a long list of matrix Lie algebras that occur in applications. Moreover, in each case we need to modify and fine tune the GPD algorithm to reduce its cost and improve its efficiency.

This is the time to mention that the GPD approach, combined with an upper-Hessenberg form and the "peel-up" technique, result in an algorithm that compares favorably, in terms of both cost and accuracy, with classical methods to compute the exponential of a matrix. Thus, a procedure motivated by retention of specialized differential-geometric structure, and based on mathematics which might be unfamiliar to many numerical analysts, is very valuable also in a general context, where Lie-group structure is not at issue.

The plan of this paper is as follows. In section 2 we consider in greater detail Lie groups and Lie algebras, introducing requisite theory and notation. This is followed by a brief review of generalized polar decompositions by means of involutory automorphisms. In section 3 we debate the computation, using GPD, of exponentials of tridiagonal matrices. We introduce a new approach (the "peel-up" algorithm) which renders the GPD method substantially more efficient in this setting. The theme of section 4 is how to bring matrices to an upper-Hessenberg form, or alternative forms that lend themselves to our approach, by means of similarity transformations. Thus, for example, symplectic matrices are converted into a so-called butterfly form. We

discuss the implementation and cost of the "peel-up" technique in all these settings. Section 5 is devoted to a divide-and-conquer strategy, which, approximating the exponential of a matrix by computations in lower-dimensional spaces, which can be performed in unison, lends itself to implementation in parallel architectures. This strategy is fully compatible with GPD and the retention of Lie-group structure and it again displays the merits of the "peel-up" approach. Finally, in section 6 we discuss the calculation of the exponential by GPD and the "peel-up" technique for a range of more "exotic" Lie groups: the Lorentz group, the isotropy group and the scaling group. The paper concludes with an appendix, to which we have relegated some of the more technical calculations.

The issue of stability and conditioning is outside the scope of this paper. Although much of the underlying framework, based upon similarity transformations by orthogonal matrices, is consistent with good conditioning, stability might become an issue. We plan to return to this subject area in a subsequent paper, where we will explore in detail the stability of the GPD technique.

Let us conclude this section anticipating that the proposed method has a complexity of $\mathcal{O}(n^3)$, both when the exponential is applied to a vector and a matrix. This might seem quite expensive especially in the case when one needs to compute $\exp(A)\mathbf{v}$ if compared to other techniques like Krylov subspace methods, which require generally $\mathcal{O}(n^2)$ operations instead [5, 8]. However, let us remind the reader that Krylov subspace methods also require about $18n^3$ operations when the exponential is applied to a matrix. In this case, our methods achieve at least a 50% improvement on the execution time, depending on the algebra under consideration.

Another important difference with Krylov methods is that while Krylov methods approximate the exponential to machine precision, the proposed method approximates the exponential to a given *order* of accuracy. This makes them particularly suited to exponential approximations within numerical integrators for ODEs, since the error is subsumed in that of the integration method and can be controlled by standard error control techniques for ODEs. For other applications, there exist global error bounds that depend on the scaling of the matrix [19]. We speculate that these bounds can be used efficiently in tandem with scaling and squaring to achieve a given accuracy and we plan to further explore this approach in the next future.

2. Background theory. The natural setting of GPD is Lie-group and Lie-algebra theory; therefore it is convenient to present the background theory in the language of differential geometry. To distinguish between group and algebra elements, it is usual in differential geometry to denote Lie-group elements with lower-case letters and Lie-algebra elements with upper-case letters, whether they represent matrices, vectors or scalars [7]. An arbitrary Lie group will be denoted by G and the corresponding Lie algebra by $\mathfrak{g}$. Subspaces of $\mathfrak{g}$ are also usually denoted by Gothic letters. We adopt this convention throughout this subsection. Later on, when most of the computations take place at the algebra level, we will revert to a language that is more familiar to the numerical analysis community and matrices (except when we want to emphasize the Lie-group context) will be denoted as usual with capital letters.

Let $G \subseteq \mathrm{GL}(\mathbb{R}, n)$ be a matrix Lie group with Lie algebra $\mathfrak{g}$. Given an involutive automorphism σ of G, i.e., a one-to-one map $G \to G$ such that

$$\sigma(x \cdot y) = \sigma(x) \cdot \sigma(y) \quad \forall x, y \in G,$$
$$\sigma(\sigma(x)) = x \quad \forall x \in G,\ \sigma \neq \mathrm{id},$$

it is possible to show that, for t sufficiently small, every element $z = \exp(tZ) \in G$,

$Z \in \mathfrak{g}$, can be factorized in the form

$$z = xy, \tag{2.1}$$

where $\sigma(y) = y$ and $\sigma(x) = x^{-1}$ [12, 16]. The decomposition (2.1) is called the GPD of z, in analogy with the case of real matrices with the special choice $\sigma(z) = z^{-\top}$, when it reduces to the familiar polar decomposition.

The automorphism σ induces in a natural manner an involutive automorphism $\mathrm{d}\sigma$ on the Lie algebra $\mathfrak{g}$,

$$\mathrm{d}\sigma(Z) = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \sigma(\exp(tZ)), \tag{2.2}$$

which defines a splitting of $\mathfrak{g}$ into the direct sum of two linear spaces,

$$\mathfrak{g} = \mathfrak{p} \oplus \mathfrak{k}, \tag{2.3}$$

where $\mathfrak{k} = \{Z \in \mathfrak{g} : \mathrm{d}\sigma(Z) = Z\}$ is a subalgebra of $\mathfrak{g}$, while the set $\mathfrak{p} = \{Z \in \mathfrak{g} : \mathrm{d}\sigma(Z) = -Z\}$ has the structure of a Lie triple system, a linear space closed under the double commutator,

$$A, B, C \in \mathfrak{p} \qquad \Longrightarrow \qquad [A, [B, C]] \in \mathfrak{p},$$

where the bracket $[A, B] = AB - BA$ is the standard matrix commutator.

To show that (2.3) is true, denote by $\Pi_{\mathfrak{p}} : \mathfrak{g} \to \mathfrak{p}$ the canonical projection of $\mathfrak{g}$ onto the subspace $\mathfrak{p}$ and by $\Pi_{\mathfrak{k}} : \mathfrak{g} \to \mathfrak{k}$ its projection onto $\mathfrak{k}$. Set

$$P = \Pi_{\mathfrak{p}}(Z), \qquad K = \Pi_{\mathfrak{k}}(Z).$$

It is easily verified by direct computation that every element Z can be written in a unique manner as $Z = P + K$, where

$$P = \Pi_{\mathfrak{p}}(Z) = \frac{1}{2}(Z - \mathrm{d}\sigma(Z)), \qquad K = \Pi_{\mathfrak{k}}(Z) = \frac{1}{2}(Z + \mathrm{d}\sigma(Z)).$$

To keep our presentation relevant to the subject matter of this paper, we refer the reader to [16, 20] and references therein for a more extensive treatment of such decompositions. However, it is of fundamental importance to note that the sets $\mathfrak{k}$ and $\mathfrak{p}$ possess the following properties:

$$[\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}, \qquad [\mathfrak{k}, \mathfrak{p}], [\mathfrak{p}, \mathfrak{k}] \subseteq \mathfrak{p}, \qquad [\mathfrak{p}, \mathfrak{p}] \subseteq \mathfrak{k}, \tag{2.4}$$

meaning that for all $K_1, K_2 \in \mathfrak{k}$ and $P_1, P_2 \in \mathfrak{p}$, it is true that $[K_1, K_2] \in \mathfrak{k}$, $[K_1, P_1], [P_2, K_2] \in \mathfrak{p}$, and $[P_1, P_2] \in \mathfrak{k}$.

How does the splitting of $Z = P + K$ relate to the factorization (2.1)? It is possible to show that, for t sufficiently small, the factors x and y in (2.1) are of the form $x = \exp(X(t))$ and $y = \exp(Y(t))$, where $X(t) \in \mathfrak{p}$ and $Y(t) \in \mathfrak{k}$, for all $t \in [0, t_0]$. Moreover they can be expanded in series

$$X(t) = \sum_{i=1}^{\infty} X_i t^i, \qquad Y(t) = \sum_{i=1}^{\infty} Y_i t^i,$$

where the coefficients X_i and Y_i can be calculated by means of explicit recurrence relations from the matrices P and K [20]. The first terms in the expansions of $X(t)$ and $Y(t)$ are

$$\begin{aligned} X = {} & Pt - \frac{1}{2}[P,K]t^2 - \frac{1}{6}[K,[P,K]]t^3 \\ & + \left(\frac{1}{24}[P,[P,[P,K]]] - \frac{1}{24}[K,[K,[P,K]]]\right)t^4 \\ & + \left(\frac{7}{360}[K,[P,[P,[P,K]]]] - \frac{1}{120}[K,[K,[K,[P,K]]]] \right. \\ & \left. - \frac{1}{180}[[P,K],[P,[P,K]]]\right)t^5 + \mathcal{O}\big(t^6\big), \\ Y = {} & Kt - \frac{1}{12}[P,[P,K]]t^3 + \left(\frac{1}{120}[P,[P,[P,[P,K]]]] + \frac{1}{720}[K,[K,[P,[P,K]]]] \right. \\ & \left. - \frac{1}{240}[[P,K],[K,[P,K]]]\right)t^5 + \mathcal{O}\big(t^7\big). \end{aligned} \tag{2.5}$$

Since $X(t)$ and $Y(t)$ and their truncations live in $\mathfrak{p}$ and $\mathfrak{k}$, respectively, it is clearly desirable to choose automorphisms σ such that exponentials of elements in $\mathfrak{p}$ (and eventually $\mathfrak{k}$) and repeated commutators of P and K are easy to compute.

Assume next that $\sigma_1, \sigma_2, \ldots, \sigma_m$ is a sequence of involutive automorphisms on G that satisfies the above conditions. Then, taking $\sigma \equiv \sigma_1$, we partition $\mathfrak{g} = \mathfrak{p}_1 \oplus \mathfrak{k}_1$, and approximate

$$\exp(tZ) \approx \exp(X^{[1]}(t)) \exp(Y^{[1]}(t)), \tag{2.6}$$

where $X^{[1]}$ and $Y^{[1]}$ are truncations of (2.5) of suitable order.

By the same token, $\mathfrak{k}_1$ is partitioned as $\mathfrak{p}_2 \oplus \mathfrak{k}_2$ by means of the automorphism σ_2, and

$$\exp(Y^{[1]}(t)) \approx \exp(X^{[2]}(t)) \exp(Y^{[2]}(t)),$$

where again $X^{[2]}$ and $Y^{[2]}$ are truncations of (2.5) of suitable order.

The procedure is iterated for m steps, say, so that $\mathfrak{k}_m$ is of low dimension and, therefore, exponentials of its elements are easy to compute exactly.

This algorithm approximates $\exp(tZ)$ to a given order of accuracy. In these circumstances, (2.6) will read

$$\exp(tZ) \approx F(t,Z) = \exp(X^{[1]}(t)) \cdots \exp(X^{[m]}(t)) \exp(Y^{[m]}(t)) \tag{2.7}$$

and it corresponds to the algebra direct-sum decomposition

$$\mathfrak{g} = \mathfrak{p}_1 \oplus \cdots \oplus \mathfrak{p}_m \oplus \mathfrak{k}_m. \tag{2.8}$$

In some circumstances, it might be more convenient to use a mirrored form of (2.7),

$$\exp(tZ) \approx \exp(\tilde{Y}^{[m]}(t)) \exp(\tilde{X}^{[m]}(t)) \cdots \exp(\tilde{X}^{[1]}(t)). \tag{2.9}$$

Clearly, the functions $\tilde{Y}(t)$ and $\tilde{X}(t)$ and their truncations are related to $Y(t)$ and $X(t)$. Indeed, in [16] it is easily verified that

$$\tilde{Y}(t) = Y(t), \qquad \tilde{X}(t) = -X(-t).$$

2.1. On the choice of the order of approximation. As we have mentioned earlier, one of the core applications of the splitting methods proposed in this paper is the numerical solution of ODEs on Lie groups by means of Lie-group methods [10] using exponentials. If, in this context, a numerical integrator of order p is used, it is reasonable to use an exponential approximation of the same order since the truncation error of the exponential approximation is subsumed in the truncation error of the numerical integrator. This is accomplished by truncating the expansions (2.5) to include all the terms up to $\mathcal{O}(t^p)$. For instance, if using a numerical integrator of order four, a reasonable approximation is

$$\begin{aligned} X(t) &\approx X_1 t + X_2 t^2 + X_3 t^3 + X_4 t^4 \\ &= Pt - \frac{1}{2}[P,K]t^2 - \frac{1}{6}[K,[P,K]]t^3 + \left(\frac{1}{24}[P,[P,[P,K]]] - \frac{1}{24}[K,[K,[P,K]]]\right) t^4 \end{aligned}$$

for $X(t)$, and, similarly,

$$Y(t) \approx Y_1 t + Y_3 t^3 = Kt - \frac{1}{12}[P,[P,K]]t^3$$

for $Y(t)$.

In other cases, for example, when the given matrix Z has large entries or is badly scaled, the methods we propose in this paper should be combined with other procedures like error control and scaling and squaring.

2.2. On the choice of automorphisms. From this point onward, we are mostly interested in the algebraic setting; therefore, we revert our notation to the more familiar in numerical analysis. Matrices (otherwise specified) will be denoted by capital letters, vectors by boldface letters, et cetera.

To obtain the algebra splitting (2.8), Zanna et al. [21] suggested using automorphisms of the type

$$\sigma(z) = \mathrm{Ad}_S Z = SzS, \quad z \in G, \tag{2.10}$$

(inner automorphisms), where $S \in \mathrm{O}(n) \cap G$ is a suitable involutory matrix (i.e., $S^2 = I$) such that $SzS \in G$. Note that, at the algebra level, (2.2) implies that

$$\mathrm{d}\sigma(Z) = \mathrm{Ad}_S Z = SZS,$$

in other words, $\mathrm{d}\sigma$ and σ are essentially of the same form.

Consider next an inner automorphism Ad_S, where

$$S = \mathrm{diag}(-1)^{\mathbf{s}} = \mathrm{diag}[(-1)^{s_1}, (-1)^{s_2}, \dots, (-1)^{s_n}] \qquad s_i \in \{0,1\}. \tag{2.11}$$

Given an arbitrary matrix Z, one has

$$(SZS)_{i,j} = (-1)^{s_i+s_j} Z_{i,j}$$

consequently,

$$P_{i,j} = \frac{1}{2}[1 - (-1)^{s_i+s_j}]Z_{i,j}, \qquad K_{i,j} = \frac{1}{2}[1 + (-1)^{s_i+s_j}]Z_{i,j}.$$

Hence, choosing appropriately the vector $\mathbf{s}$, it is possible to dispatch selected rows and columns of Z to the different subspaces.

For instance, choosing $s_i = 0$, $i \neq k$, and $s_k = 1$, we obtain subspaces $\mathfrak{p}$ and $\mathfrak{k}$ with the sparsity structure

$$\mathfrak{p} \ni \begin{bmatrix} 0 & \cdots & 0 & \times & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \times & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & \times & 0 & 0 & \cdots & 0 \\ \times & \times & \times & 0 & \times & \times & \times & \times \\ 0 & \cdots & 0 & \times & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \times & 0 & 0 & & 0 \\ \vdots & & \vdots & \times & \vdots & & & \vdots \\ 0 & \cdots & 0 & \times & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \mathfrak{k} \ni \begin{bmatrix} \times & \cdots & \times & 0 & \times & \times & \cdots & \times \\ \vdots & & \vdots & 0 & \vdots & \vdots & & \vdots \\ \times & \cdots & \times & 0 & \times & \times & \cdots & \times \\ 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 \\ \times & \cdots & \times & 0 & \times & \times & \cdots & \times \\ \times & \cdots & \times & 0 & \times & \times & \cdots & \times \\ \vdots & & \vdots & 0 & \vdots & \vdots & & \vdots \\ \times & \cdots & \times & 0 & \times & \times & \cdots & \times \end{bmatrix},$$

respectively.

3. Tridiagonal matrices. Let us assume that Z is a tridiagonal matrix,

$$Z = \begin{bmatrix} \alpha_1 & \gamma_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \gamma_{n-1} \\ & & \beta_{n-1} & \alpha_n \end{bmatrix}, \tag{3.1}$$

and denote by $\mathbf{e}_i$ the ith unit vector in $\mathbb{R}^n$.

DEFINITION 3.1. *We say that the sequence of automorphisms*

$$\mathrm{Ad}_{S_i}, \quad S_i = \mathrm{diag}(-1)^{\mathbf{s}_i}, \quad i = 1, \dots, n-1, \tag{3.2}$$

constitutes a peel-down approach *if* $\mathbf{s}_1 = \mathbf{e}_1 = [1, 0, 0, \dots, 0]^\top$, $\mathbf{s}_2 = \mathbf{e}_2 = [0, 1, 0, \dots, 0]^\top$, $\dots$, $\mathbf{s}_{n-1} = \mathbf{e}_{n-1} = [0, 0, 0, \dots, 1, 0]^\top$. *The choice* $\mathbf{s}_1 = \mathbf{e}_n$, $\mathbf{s}_2 = \mathbf{e}_{n-1}, \dots, \mathbf{s}_{n-1} = \mathbf{e}_2$ *constitutes a* peel-up approach.

The above definition is motivated by the fact that the peel-down approach leads to the splitting in bordered matrices proposed in [21]; the bordered matrices are obtained by "peeling" the matrix Z from the top-left corner downwards. In the peel-up approach, we target rows and columns of Z starting from the bottom-right corner instead and proceed upwards.

In what follows, we apply a peel-up approach to (3.1) and discuss in detail the first stage, corresponding to the automorphism Ad_{S_1}, $\mathbf{s}_1 = \mathbf{e}_n$. The remaining stages of the peel-up approach share very similar features.

Using $\mathrm{Ad}_{\mathrm{diag}(-1)^{\mathbf{e}_n}}$ to perform the first algebra splitting we obtain $Z = P + K$, where

$$P = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 0 & 0 & \gamma_{n-1} \\ 0 & \cdots & 0 & \beta_{n-1} & 0 \end{bmatrix}, \quad K = \begin{bmatrix} \alpha_1 & \gamma_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \gamma_{n-2} & 0 \\ \vdots & \ddots & \beta_{n-2} & \alpha_{n-1} & 0 \\ 0 & \cdots & 0 & 0 & \alpha_n \end{bmatrix}. \tag{3.3}$$

Next, we start computing commutators. We write

$$P = \gamma_{n-1}\mathbf{e}_{n-1}\mathbf{e}_n^\top + \beta_{n-1}\mathbf{e}_n\mathbf{e}_{n-1}^\top,$$

and commence our computations with

$$\begin{aligned}[P,K] = &-\gamma_{n-2}\gamma_{n-1}\mathbf{e}_{n-2}\mathbf{e}_n^\top + \beta_{n-2}\beta_{n-1}\mathbf{e}_n\mathbf{e}_{n-2}^\top\\ &+\gamma_{n-1}(\alpha_n-\alpha_{n-1})\mathbf{e}_{n-1}\mathbf{e}_n^\top - \beta_{n-1}(\alpha_n-\alpha_{n-1})\mathbf{e}_n\mathbf{e}_{n-1}^\top.\end{aligned} \tag{3.4}$$

It is immediate to observe that there appears a fill-in in the tridiagonal structure of $X(t) = Pt - \frac{1}{2}t^2[P,K] + \mathcal{O}(t^3)$. In general, the more commutators we take, the greater the fill-in: two extra nonzero elements (one in the nth row and one in the nth column) for every extra power of t. In principle, the whole nth row and column would eventually be filled in. But this is not such bad news as it might appear at a first glance. Below we explain the reason for this.

3.1. Dealing with fill-in. An important observation is that fill-in of $X(t)$ propagates only in the Lie triple-system $\mathfrak{p}$, which consists of rank-2 matrices of the form

$$\left[\begin{array}{c|c} O & \mathbf{a} \\ \hline \mathbf{b}^\top & 0 \end{array}\right], \quad \mathbf{a},\mathbf{b}\in\mathbb{R}^m. \tag{3.5}$$

Therefore, once $X(t)$ is approximated, its exponential can be computed exactly by means of an expression analogous to the Euler–Rodrigues formula for the exponential of a skew-symmetric matrix. Assume that $A\in\mathfrak{p}$ is of the form (3.5). Then,

$$\exp(A) = \begin{cases} I + \dfrac{\sinh\theta}{\theta}A + \dfrac{1}{2}\left[\dfrac{\sinh(\theta/2)}{\theta/2}\right]^2 A^2, & \mathbf{a}^\top\mathbf{b} > 0,\ \theta = \sqrt{\mathbf{a}^\top\mathbf{b}},\\ I + A + \dfrac{1}{2}A^2, & \mathbf{a}^\top\mathbf{b} = 0,\\ I + \dfrac{\sin\theta}{\theta}A + \dfrac{1}{2}\left[\dfrac{\sin(\theta/2)}{\theta/2}\right]^2 A^2, & \mathbf{a}^\top\mathbf{b} < 0,\ \theta = \sqrt{-\mathbf{a}^\top\mathbf{b}}\end{cases} \tag{3.6}$$

[21], where

$$A^2 = \left[\begin{array}{c|c} \mathbf{a}\mathbf{b}^\top & \mathbf{0} \\ \hline \mathbf{0}^\top & \mathbf{a}^\top\mathbf{b} \end{array}\right].$$

Setting to η_1, η_2 the coefficients of (3.6), application to a vector yields

$$\exp(A)\mathbf{v} = \mathbf{v} + \eta_1\begin{bmatrix} O & \mathbf{a}\\ \mathbf{b}^\top & 0\end{bmatrix}\mathbf{v} + \eta_2\begin{bmatrix}\mathbf{a}\mathbf{b}^\top & \mathbf{0}\\ \mathbf{0}^\top & \pm\theta^2\end{bmatrix}\mathbf{v},$$

where the sign of θ^2 is chosen according to (3.6). Assume next that $\mathbf{w},\mathbf{v}\in\mathbb{R}^{k+1}$, $\mathbf{a},\mathbf{b}\in\mathbb{R}^k$. Writing $\mathbf{v} = [\mathbf{v}_k, v]^\top$, $\mathbf{w} = [\mathbf{w}_k, w]^\top$, a direct computation reveals that

$$\begin{bmatrix}\mathbf{w}_k\\ w\end{bmatrix} = \exp(A)\mathbf{v} = \begin{bmatrix}\mathbf{v}_k + \zeta_1\mathbf{a}\\ \zeta_2\end{bmatrix}, \tag{3.7}$$

where

$$\begin{aligned}\zeta_1 &= [\eta_1 v + \eta_2(\mathbf{b}^\top\mathbf{v}_k)],\\ \zeta_2 &= (1\pm\eta_2\theta^2)v + \eta_1(\mathbf{b}^\top\mathbf{v}_k).\end{aligned}$$

TABLE 3.1

Cost of the computation (including both addition and multiplication) of the exponential (3.6). *The* (k,k) *column corresponds to the case when* $\mathbf{a},\mathbf{b}$ *are full, the* (k,p) *corresponds to the case when* $\mathbf{a}$ *is full while only the last* p *components of* $\mathbf{b}$ *are nonzero and, finally, the* (p,p) *column corresponds to both* $\mathbf{a}$ *and* $\mathbf{b}$ *having only the last* p *components nonzero.*

Cost of $\exp(A)$	(k,k)	(k,p)	(p,p)
$\mathbf{a}^\top\mathbf{b}$	$2k$	$2p$	$2p$
$\mathbf{b}^\top\mathbf{v}_k$	$2k$	$2p$	$2p$
$\zeta_1\mathbf{a}$	k	k	p
$\mathbf{w}_k$	k	k	p
total, stage k	$6k$	$2k+4p$	$6p$
total, summing $1\le k\le n$ (vector)	$3n^2$	n^2+4pn	$6pn$
matrix (n vectors)	$2n^3$	n^3+2pn^2	$4pn^2$

When the exponential is applied to a matrix, we apply (3.7) to each column vector. Note, however, that the scalar product $\mathbf{a}^\top\mathbf{b}$ need be computed only once, even if the matrix columns are distinct.

In passing, we mention that there exists another formula for the exact computation of the exponential of a matrix as in (3.5), due to Celledoni et al. [2],

$$\exp(A) = I + [\mathbf{e}_k, \mathbf{k}]\varphi(D)\begin{bmatrix}\mathbf{l}^\top\\ \mathbf{e}_k^\top\end{bmatrix}, \tag{3.8}$$

where

$$\mathbf{k} = \begin{bmatrix}\mathbf{a}\\ 0\end{bmatrix}, \qquad \boldsymbol{l} = \begin{bmatrix}\mathbf{b}\\ 0\end{bmatrix}, \qquad D = \begin{pmatrix}0 & \mathbf{b}^\top\mathbf{a}\\ 1 & 0\end{pmatrix},$$

$\mathbf{e}_k$ is the vector $[0,0,\dots,0,1]^\top \in \mathbb{R}^k$ and, finally, $\varphi(z) = (\mathrm{e}^z-1)/z$. This formula can be shown to have the same computational cost as (3.7).

If $\mathbf{a},\mathbf{b}$ have $p \ll n$ nonzero elements only, say $a_{n-p},\dots,a_{n-1},b_{n-p}\dots,b_{n-1}$, it is clear that the computation of exponentials of elements in $\mathfrak{p}$ requires just $\mathcal{O}(pn)$ operations, when the exponentials are multiplied by a vector, and $\mathcal{O}(pn^2)$ operations when a multiplication by a matrix is required (see Table 3.1). Therefore, the fill-in in the $\mathfrak{p}$ part is inconvenient (in principle one would have preferred to preserve the neater tridiagonal structure), but it is not dangerous insofar as the increase of the computational cost is concerned.

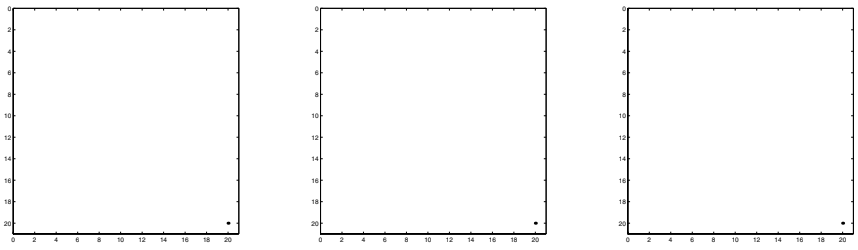


FIG. 3.1. *Fill-in in the tridiagonal structure at step one, two, and three in the peel-up procedure.*

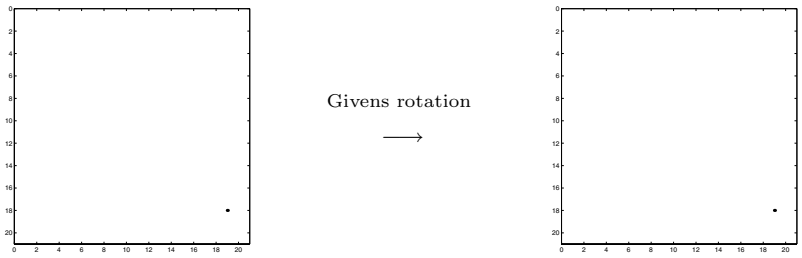


FIG. 3.2. *In the peel-up approach for skew-symmetric matrices, using a Givens rotation to annihilate sub-diagonal fill-in does not cause further fill-in.*

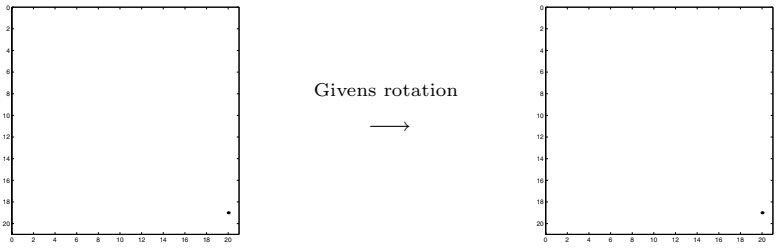


FIG. 3.3. *In the peel-down approach for skew-symmetric matrices, using a Givens rotation to annihilate sub-diagonal fill-in does cause fill-in that propagates downwards. If the rotations are chosen to annihilate* super-diagonal *fill-in instead, there is no fill-in propagation.*

More subtle is the case when the fill-in appears in $\mathfrak{k} \ni Y(t) = Kt - \frac{1}{12}t^3[P,[P,K]] +$ higher order terms. If such fill-in occurs and is not suitably dealt with, it will propagate further and the tridiagonal structure of the sub-matrices will be lost. The matrices then become increasingly full, and the cost of the splitting becomes $\mathcal{O}(n^3)$, as discussed in [21]. Such an instance is displayed below in Figure 3.1, in which we perform three steps of the peel-up procedure for a tridiagonal symmetric matrix. We compute the function $Y(t)$ given in (2.5) and truncate the expansion to order six. It is clearly observed that in each step the number of nonzero elements increases and that the sub-matrices have a tendency to become full. This is clearly not desirable, since we do not want to lose the benefits of the original tridiagonal form!

However, two important observations are the following:

- The fill-in in the $Y(t)$ part appears at order five only—truncations of order one to four are still tridiagonal.
- The number of fill-in elements is usually very small.

Hence it is reasonable to eliminate the fill-in at each step by means of similarity orthogonal transformations (for instance, Givens rotations). At each step this contributes only $\mathcal{O}(1)$ to the cost of the splitting, which is negligible compared to the total cost of the approximation.

If Z is a skew-symmetric matrix, the fill-in in the $Y(t)$ part is chess board-like. Givens rotations can be used to eliminate the sub-diagonal fill-in and their transpose takes care of the super-diagonal fill-in. This is precisely the point when our choice of the peel-up approach, in preference to peel-down, starts to pay dividends. Using a peel-down approach, Givens rotations that eliminate the sub-diagonal fill-in cause

further fill-in that is propagated downwards. However, no fill-in is caused if Givens rotations are targeted to annihilate super-diagonal fill-ins instead (see and compare Figures 3.2 and 3.3). For this reason we will restrict our attention to the peel-up approach instead of the peel-down approach of [21].

3.2. On the computation of commutators. We return to the computation of commutators and denote by $\mathrm{ad}_B = [B, \cdot]$ the operator that performs commutation with B, i.e., $\mathrm{ad}_B C = [B, C]$, $\mathrm{ad}_B^2 C = [B, [B, C]]$, etc.

Our first observation is that the involutions S are usually chosen so that $P = \Pi_{\mathfrak{p}}(Z)$ has low rank, hence only just a few nonzero eigenvalues. Thus, we can use the theory of minimal polynomials of matrices [9] so that few commutators need be computed. All the remaining commutators with P can be obtained as linear combinations of those.

LEMMA 3.2. *Consider the matrix A of the form* (3.5). *If $\mathbf{a}\mathbf{b}^\top \neq O$, then the minimal polynomial of* ad_A *is*

$$\begin{aligned} p(\lambda) &= \lambda(\lambda - 2\theta)(\lambda + 2\theta)(\lambda - \theta)(\lambda + \theta) \\ &= \lambda^5 - 5\mathbf{b}^\top\mathbf{a}\lambda^3 + 4(\mathbf{b}^\top\mathbf{a})^2\lambda, \end{aligned} \tag{3.9}$$

where $\theta = \sqrt{\mathbf{b}^\top\mathbf{a}}$.

If either $\mathbf{a}$ or $\mathbf{b}$ is zero, then the minimal polynomial is

$$p(\lambda) = \lambda^3. \tag{3.10}$$

Proof. Recall that if A has distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ with algebraic multiplicities $r_1, r_2, \ldots, r_m$, respectively, the minimal polynomial of A has the form

$$q(\lambda) = \prod_{i=1}^{m} (\lambda - \lambda_i)^{g_i},$$

where g_i is the order of the largest Jordan block of A corresponding to the eigenvalue λ_i [9].

Let us assume first that $\mathbf{b}^\top\mathbf{a} \neq 0$. Imposing $A\mathbf{v} = \lambda\mathbf{v}$, we deduce immediately that the eigenvalues of A are $\lambda = \pm\theta = \pm\sqrt{\mathbf{b}^\top\mathbf{a}}$ and $\lambda = 0$ with algebraic multiplicities one, one, and $n - 2$, respectively. It is easily verified that these are also their geometric multiplicities: for $\lambda = \pm\theta$, eigenvectors are of the form $[\mathbf{a}, \pm 1]^\top$; for the zero eigenvalues, eigenvectors are of the form $[\mathbf{v}_1, 0]^\top$, $\mathbf{0} \neq \mathbf{v}_1 \in \mathbb{R}^{n-1}$, satisfying $\mathbf{b}^\top\mathbf{v}_1 = 0$, furthermore, it is possible to find $n - 2$ of those that are linearly independent.

Since the eigenvalues and eigenvectors of ad_A are the form $\lambda_i - \lambda_j$ and $\mathbf{y}_i^\top\mathbf{x}_j$, respectively, the λ_is being eigenvalues of A with left and right eigenvector $\mathbf{y}_i$ and $\mathbf{x}_i$, respectively, we deduce that ad_A has eigenvalues

$$\lambda = \pm 2\theta, \qquad \lambda = \pm\theta$$

with algebraic/geometric multiplicities one each, and

$$\lambda = 0$$

with algebraic and geometric multiplicity $n^2 - 4$. This implies that all Jordan blocks have size one, from which it follows directly that the minimal polynomial of ad_A is of the form åq:minpoly.

Next, if $\theta = 0$ but $\mathbf{a}\mathbf{b}^\top \neq O$ (namely $\mathbf{a}, \mathbf{b} \neq \mathbf{0}$) the eigenvectors of A, that we write as $[\mathbf{v}_1, v_2]^\top$, must obey the conditions

$$\begin{aligned} \mathbf{a}v_2 &= \mathbf{0}, \\ \mathbf{b}^\top \mathbf{v}_1 &= 0. \end{aligned}$$

Since $\mathbf{a} \neq 0$, it must necessarily be $v_2 = 0$. Therefore, eigenvectors must be of the form $[\mathbf{v}_1, 0]$. Recall that $\mathbf{v}_1$ has $n-1$ entries ($n-1$ free parameters) while the second equation $\mathbf{b}^\top \mathbf{v}_1 = 0$ gives only a linear constraint. This means that we can find only $n-2$ linearly independent eigenvectors and two further linearly independent generalized eigenvectors. In terms of Jordan blocks, this means that A has a Jordan block of the form

$$J(0) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

or two Jordan blocks of the form

$$J(0) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

A simple discussion eliminates the second possibility: if there were two such Jordan blocks, the minimal polynomial would be λ^2, i.e., $A^2 = O$, which is clearly not the case by direct computation, being $\mathbf{a}\mathbf{b}^\top \neq O$. Hence, only the first 3×3 Jordan block is admissible. Therefore λ^3 is the minimal polynomial of A and, as a consequence, $A^3 = O$. Passing to the adjoint operator ad_A, recall that, for an arbitrary matrix C,

$$\mathrm{ad}_A^k C = \sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} A^i C A^{k-i}, \quad k = 1, 2, \ldots . \tag{3.11}$$

Clearly, $\mathrm{ad}_A^5 C = O$ since in all terms there appears a power A^i with $i \geq 3$. For lower order powers, there are always terms of the type $A^i C A^{k-i}$, where $i, k-i \leq 2$. This means that it is always possible to find a matrix C for which at least one of terms does not vanish. Hence the minimal polynomial of ad_A is

$$p(\lambda) = \lambda^5.$$

Finally, in the case when either $\mathbf{a} = \mathbf{0}$ or $\mathbf{b} = \mathbf{0}$, by direct computation,

$$A^2 = O,$$

corresponding to a Jordan block $J(0) = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$. Hence the minimal polynomial of A is λ^2. Insofar as ad_A is concerned, the first power to vanish in (3.11) is ad_A^3, and no lower power vanishes for arbitrary matrices C. Hence the minimal polynomial is

$$p(\lambda) = \lambda^3.$$

This completes the proof of the lemma. ∎

Trivially, in the case when both $\mathbf{a}$ and $\mathbf{b}$ are zero, both $A = O$ and $\mathrm{ad}_A = O$, hence their minimal polynomial is $p(\lambda) = \lambda$.

THEOREM 3.3. *Assume that the matrix A is of the form* (3.5). *Then, for every $k = 1, 2, \dots$, commutators by A can be computed as*

$$\mathrm{ad}_A^k = [C_1 + (-1)^k C_2] 2^k \theta^k + [C_3 + (-1)^k C_4] \theta^k, \quad k = 1, 2, \dots, \tag{3.12}$$

when $\theta = \sqrt{\mathbf{b}^\top \mathbf{a}} \neq 0$, and

$$\begin{aligned} C_1 - C_2 &= \frac{1}{6}\left(-\frac{\mathrm{ad}_A}{\theta} + \frac{\mathrm{ad}_A^3}{\theta^3}\right), & C_3 - C_4 &= \frac{1}{3}\left(\frac{4\mathrm{ad}_A}{\theta} - \frac{\mathrm{ad}_A^3}{\theta^3}\right), \\ C_1 + C_2 &= \frac{1}{12}\left(-\frac{\mathrm{ad}_A^2}{\theta^2} + \frac{\mathrm{ad}_A^4}{\theta^4}\right), & C_3 + C_4 &= \frac{1}{3}\left(\frac{4\mathrm{ad}_A^2}{\theta^2} - \frac{\mathrm{ad}_A^4}{\theta^4}\right). \end{aligned} \tag{3.13}$$

If $\theta = 0$ but $\mathbf{a}, \mathbf{b} \neq O$, then

$$\mathrm{ad}_A^k = O, \quad k = 5, 6, 7, \dots .$$

If either $\mathbf{a}$ or $\mathbf{b}$ is a zero vector, then

$$\mathrm{ad}_A^k = O, \quad k = 3, 4, 5, \dots .$$

Proof. Recall that the minimal polynomial is the least degree monic polynomial such that

$$p(\mathrm{ad}_A) = 0,$$

hence

$$\mathrm{ad}_A^5 - 5(\mathbf{b}^\top \mathbf{a})\mathrm{ad}_A^3 + 4(\mathbf{b}^\top \mathbf{a})^2 \mathrm{ad}_A = O.$$

Multiplying by $\mathrm{ad}_A, \mathrm{ad}_A^2, \dots$, we obtain the recurrence relation

$$\mathrm{ad}_A^{k+2} - 5(\mathbf{b}^\top \mathbf{a})\mathrm{ad}_A^k + 4(\mathbf{b}^\top \mathbf{a})^2 \mathrm{ad}_A^{k-2} = O, \quad k = 3, 4, 5, \dots,$$

whose general solution is (3.12). The unknowns C_1, C_2, C_3, C_4 are obtained by requiring that the formula (3.12) is correct for $k = 1, 2, 3, 4$. We obtain

$$\begin{aligned} (C_1 - C_2) + \frac{1}{2}(C_3 - C_4) &= \frac{1}{2\theta}\mathrm{ad}_A, \\ (C_1 + C_2) + \frac{1}{4}(C_3 + C_4) &= \frac{1}{4\theta^2}\mathrm{ad}_A^2, \\ (C_1 - C_2) + \frac{1}{8}(C_3 - C_4) &= \frac{1}{8\theta^3}\mathrm{ad}_A^3, \\ (C_1 + C_2) + \frac{1}{16}(C_3 + C_4) &= \frac{1}{16\theta^4}\mathrm{ad}_A^4, \end{aligned}$$

and (3.13) follows by direct computation. Thus, (3.12) is determined for both odd and even values of k. ∎

In other words, given an arbitrary matrix B, the commutator $\mathrm{ad}_A^k B$, $k \geq 5$, can be obtained as a mere linear combination of $B, \mathrm{ad}_A B, \dots, \mathrm{ad}_A^4 B$. In our case, taking $A \equiv P$ and taking into account the sparsity of P, the computation of ad_P^k, $k = 1, 2, 3, 4$, is particularly simple. Setting $E_{i,j} = \mathbf{e}_i \mathbf{e}_j^\top$, a matrix with 1 in the (i, j)

position and 0 otherwise, and applying ad_P to K (whose elements are as in (3.3)), we have

$$
\begin{aligned}
[P,K] &= c_{n-2,n}E_{n-2,n} + c_{n,n-2}E_{n,n-2} \\
&\quad + c_{n-1,n}E_{n-1,n} + c_{n,n-1}E_{n,n-1}, \\
[P,[P,K]] &= d_{n-2,n-1}E_{n-2,n-1} + d_{n-1,n-2}E_{n-1,n-2} \\
&\quad + d_{n-1,n-1}E_{n-1,n-1} + d_{n,n}E_{n,n}, \\
[P,[P,[P,K]]] &= e_{n-2,n}E_{n-2,n} + e_{n,n-2}E_{n,n-2} \\
&\quad + e_{n-1,n}E_{n-1,n} + e_{n,n-1}E_{n,n-1}, \\
[P,[P,[P,[P,K]]]] &= f_{n-2,n-1}E_{n-2,n-1} + f_{n-1,n-2}E_{n-1,n-2} \\
&\quad + f_{n-1,n-1}E_{n-1,n-1} + f_{n,n}E_{n,n}.
\end{aligned} \tag{3.14}
$$

The nonzero coefficients $c_{i,j}$, $d_{i,j}$, $e_{i,j}$, $f_{i,j}$ are given by

$$
\begin{aligned}
c_{n-2,n} &= -\gamma_{n-2}\gamma_{n-1}, & d_{n-2,n-1} &= -\beta_{n-1}c_{n-2,n}, \\
c_{n,n-2} &= \beta_{n-2}\beta_{n-1}, & d_{n-1,n-2} &= \gamma_{n-1}c_{n,n-2}, \\
c_{n-1,n} &= \gamma_{n-1}(\alpha_n - \alpha_{n-1}), & d_{n-1,n-1} &= \gamma_{n-1}c_{n,n-1} - \beta_{n-1}c_{n-1,n}, \\
c_{n,n-1} &= \beta_{n-1}(\alpha_n - \alpha_{n-1}), & d_{n,n} &= -d_{n-1,n-1}, \\
e_{n-2,n} &= -\gamma_{n-1}d_{n-2,n-1}, & f_{n-2,n-1} &= -\beta_{n-1}e_{n-2,n}, \\
e_{n,n-2} &= \beta_{n-1}d_{n-1,n-2}, & f_{n-1,n-2} &= \gamma_{n-1}e_{n,n-2}, \\
e_{n-1,n} &= 2\gamma_{n-1}d_{n,n}, & f_{n-1,n-1} &= \gamma_{n-1}e_{n,n-1} - \beta_{n-1}e_{n-1,n}, \\
e_{n,n-1} &= -2\beta_{n-1}d_{n-1,n-1}, & f_{n,n} &= -f_{n-1,n-1},
\end{aligned} \tag{3.15}
$$

where β_ks and γ_ks originate in (3.3).

Unfortunately, the theory of minimal polynomials is not equally insightful insofar as commutators with K are concerned. Instead, we have computed the first few such terms explicitly and they are also in a form that renders their evaluation cheap,

$$
\begin{aligned}
[K,[P,K]] &= g_{n,n-3}E_{n,n-3} + g_{n,n-2}E_{n,n-2} + g_{n,n-1}E_{n,n-1} \\
&\quad + g_{n-3,n}E_{n-3,n} + g_{n-2,n}E_{n-2,n} + g_{n-1,n}E_{n-1,n}, \\
[K,[K,[P,K]]] &= h_{n,n-4}E_{n,n-4} + h_{n,n-3}E_{n,n-3} + h_{n,n-2}E_{n,n-2} \\
&\quad + h_{n,n-1}E_{n,n-1} + h_{n-4,n}E_{n-4,n} + h_{n-3,n}E_{n-3,n} \\
&\quad + h_{n-2,n}E_{n-2,n} + h_{n-1,n}E_{n-1,n}, \\
[K,[P,[P,[P,K]]]] &= i_{n,n-3}E_{n,n-3} + i_{n,n-2}E_{n,n-2} + i_{n,n-1}E_{n,n-1} \\
&\quad + i_{n-3,n}E_{n-3,n} + i_{n-2,n}E_{n-2,n} + i_{n-1,n}E_{n-1,n}, \\
[K,[K,[K,[P,K]]]] &= j_{n,n-5}E_{n,n-5} + j_{n,n-5}E_{n,n-4} + j_{n,n-3}E_{n,n-3} \\
&\quad + j_{n,n-2}E_{n,n-2} + j_{n,n-1}E_{n,n-1} \\
&\quad + j_{n-5,n}E_{n-5,n} + j_{n-4,n}E_{n-4,n} + j_{n-3,n}E_{n-3,n} \\
&\quad + j_{n-2,n}E_{n-2,n} + j_{n-1,n}E_{n-1,n}, \\
[[P,K],[P,[P,K]]] &= k_{n,n-2}E_{n,n-2} + k_{n,n-1}E_{n,n-1} \\
&\quad + k_{n-2,n}E_{n-2,n} + k_{n-1,n}E_{n-1,n}, \\
[K,[K,[P,[P,K]]]] &= l_{n-1,n-4}E_{n-1,n-4} + l_{n-1,n-3}E_{n-1,n-3} \\
&\quad + l_{n-1,n-2}E_{n-1,n-2} + l_{n-1,n-1}E_{n-1,n-1} \\
&\quad + l_{n-2,n-3}E_{n-2,n-3} + l_{n-2,n-2}E_{n-2,n-2}
\end{aligned} \tag{3.16}
$$

$$
\begin{aligned}
&+ l_{n-2,n-1}E_{n-2,n-1} + l_{n-3,n-2}E_{n-3,n-2}\\
&+ l_{n-3,n-1}E_{n-3,n-1} + l_{n-4,n-1}E_{n-4,n-1},\\
[[P,K],[K,[P,K]]] &= m_{n,n}E_{n,n} + m_{n-1,n-3}E_{n-1,n-3}\\
&+ m_{n-1,n-2}E_{n-1,n-2} + m_{n-1,n-1}E_{n-1,n-1}\\
&+ m_{n-2,n-3}E_{n-2,n-3} + m_{n-2,n-2}E_{n-2,n-2}\\
&+ m_{n-2,n-1}E_{n-2,n-1} + m_{n-3,n-2}E_{n-3,n-2}\\
&+ m_{n-3,n-1}E_{n-3,n-1}.
\end{aligned}
$$

The nonzero coefficients of (3.16) are reported in the appendix.

Note that, when Z in (3.1) is symmetric or skew-symmetric, only about half of the coefficients in (3.14) and (3.16) need be computed. To see this, assume that Z is symmetric as well as P and K. Then, so is $[P,[P,K]],[K,[P,K]],\dots$, and in general all terms which include an even number of commutators. Those including an odd number of commutators, like $[P,K],[K,[K,[P,K]]],\dots$, are instead skew-symmetric. Symmetry and skew-symmetry can be transparently taken into account when computing the coefficients (3.15) and (A.1)–(A.6).

By the same token, when Z is skew-symmetric, so are P and K, and *all* their commutators. Again, this means that we only need to compute just the under-diagonal (or over-diagonal) coefficients in (3.14) and (3.16).

3.3. The reduction to a tridiagonal form: Symmetric and skew-symmetric matrices. For symmetric and skew-symmetric matrices, the problem of reduction to a tridiagonal form is classical, and we briefly review well-known techniques based on Householder reflections and Lanczos tridiagonalization.

Symmetric and skew-symmetric matrices can be reduced to a tridiagonal form by means of Householder reflections,

$$H = I - \beta \mathbf{v}\mathbf{v}^\top, \quad \beta = \frac{2}{\|\mathbf{v}\|^2},$$

which are orthogonal transformations. Since $H = H^\top$, it is easily verified that HZH is symmetric/skew-symmetric if Z is also. The above transformation can be computed very effectively for symmetric and skew-symmetric matrices thanks to an algorithm due to Wilkinson, which amounts to n^3 operations approximatively (counting both additions and multiplications) [6]. If counting only multiplications (or only *flops*, i.e., operations of the form $av + b$), the count reduces to $\frac{2}{3}n^3$, consistently with [6].

When Z is sparse and very large, a possible alternative is to use the Lanczos method, which is particularly attractive when it is cheap to form products of the form $Z\mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^n$ [6, 18]. Denote by $Q = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n]$ an orthogonal matrix that tridiagonalizes Z, i.e., $Q^\top ZQ = T$, where T is symmetric or skew-symmetric depending on Z, and tridiagonal. From

$$ZQ = QT,$$

one readily obtains the recurrences relations for the unknowns of the problem. The procedure may break down when, in exact arithmetics, T is reducible. In floating-point arithmetics it is also possible that the vectors $\mathbf{q}_j$ might become progressively

TABLE 3.2
Comparison of cost of the approximation of the exponential without (ZMK) and with reduction to tridiagonal form (IZ) for splittings of order $2, 3, 4$. *Only dominant terms are reported.*

Order 2	ZMK		IZ	
	Vector	Matrix	Vector	Matrix
Tridiag.	–	–	n^3	n^3
Order cond.	$\frac{2}{3}n^3$	$\frac{2}{3}n^3$	$\mathcal{O}(n)$	$\mathcal{O}(n)$
Assembly exp	$3n^2$	$2n^3$	$6pn$	$4pn^2$
Total	$\frac{2}{3}n^3$	$2\frac{2}{3}n^3$	$n^3+\mathcal{O}(pn)$	n^3+4pn^2

Order 3	ZMK		IZ	
	Vector	Matrix	Vector	Matrix
Tridiag.	–	–	n^3	n^3
Order cond.	$2\frac{1}{2}n^3$	$2\frac{1}{2}n^3$	$\mathcal{O}(n)$	$\mathcal{O}(n)$
Assembly exp	$3n^2$	$2n^3$	$6pn$	$4pn^2$
Total	$2\frac{1}{2}n^3$	$4\frac{1}{2}n^3$	$n^3+\mathcal{O}(pn)$	n^3+4pn^2

Order 4	ZMK		IZ	
	Vector	Matrix	Vector	Matrix
Tridiag.	–	–	n^3	n^3
Order cond.	$4n^3$	$4n^3$	$\mathcal{O}(n)$	$\mathcal{O}(n)$
Assembly exp	$3n^2$	$2n^3$	$6pn$	$4pn^2$
Total	$4n^3$	$6n^3$	$n^3+\mathcal{O}(pn)$	n^3+4pn^2

less orthogonal, and, in such cases, a restart process is recommended. We refer the reader to [6, 18] for further details on the algorithm.

4. Other matrices. When Z is neither symmetric nor skew-symmetric, and, in particular, when it is not normal, the tridiagonalization process (nonsymmetric tridiagonalization, i.e., tridiagonalization by similarity transform, not necessarily orthonormal) might be either unstable or it might destroy the underlying algebraic structure [6]. For instance, the tridiagonalization of a matrix in the symplectic algebra $\mathfrak{sp}(n) := \{Z : ZJ = -JZ^\top\}$, where

$$J = \begin{bmatrix} O & I \\ -I & O \end{bmatrix}, \tag{4.1}$$

might not produce an output in $\mathfrak{sp}(n)$, and this is not desirable in many applications, for instance when conservation of Lie-group structure is important. To force the group structure, it might be more appropriate to look for other sparsity patterns that are (a) compatible with the algebra structure and (b) retained under commutation.

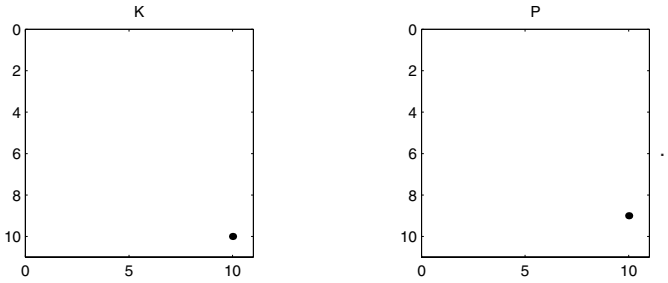
For matrices in $\mathfrak{gl}(n), \mathfrak{sl}(n)$ which do not strictly belong to other subalgebras, it is more convenient to reduce to an *upper-Hessenberg form*, by means of orthogonal transformations (e.g., Householder reflections). This is a more stable process

than nonsymmetric tridiagonalization [6]. For generic matrices, reduction to upper-Hessenberg form costs about $3\frac{1}{3}n^3$ operations [6].

For matrices belonging to other subalgebras, like the symplectic algebra, Lorentz-type algebras, quadratic algebras, etc., it is possible to consider other specific transformations that preserve the algebraic structure. These will be described at length in what follows.

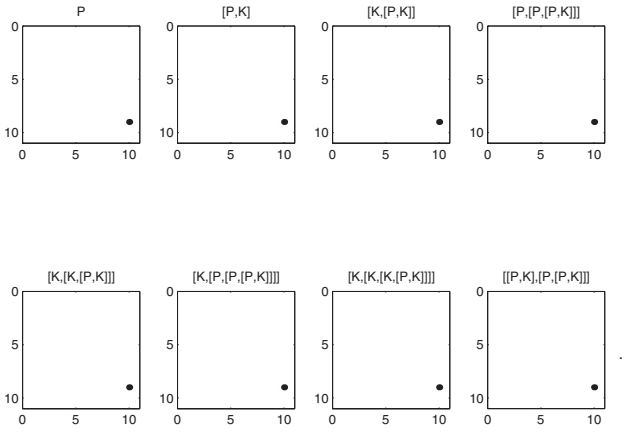
4.1. Upper Hessenberg matrices. In this subsection we analyze the first step of a peel-up approach, as in Definition 3.1, corresponding to the automorphism Ad_{S_1}, where $S_1 = \mathrm{diag}(-1)^{\mathbf{s}_1}$, $\mathbf{s}_1 = \mathbf{e}_n$. The remaining steps, corresponding to $\mathrm{Ad}_{S_2}, \mathrm{Ad}_{S_3}, \ldots$, corresponding to $\mathbf{s}_2 = \mathbf{e}_{n-1}$, $\mathbf{s}_3 = \mathbf{e}_{n-2}, \ldots$, share similar features. Therefore, Y_6 has the sparsity pattern displayed in Figure 4.2.

Assume that Z in $\mathfrak{gl}(n)$ or in $\mathfrak{sl}(n)$ is in an upper-Hessenberg form. The matrices P, K corresponding to the splitting induced by $\mathrm{Ad}_{\mathrm{diag}(-1)^{\mathbf{e}_n}}$ have the sparsity pattern



In Figure 4.1 we display the terms required in the generation of Y up to $\mathcal{O}(t^6)$, with just three filled-in entries, at the $(n-2, n-4)$, $(n-1, n-4)$ and $(n-1, n-3)$ positions. It thus takes just three Givens rotations to bring Y_6 to the same sparsity pattern as K via similarity transformations.

Next, we proceed to generate X. All the terms up to $\mathcal{O}(t^5)$ are



Therefore, the sparsity patterns of the truncations X_2, X_3, X_4, X_5 are

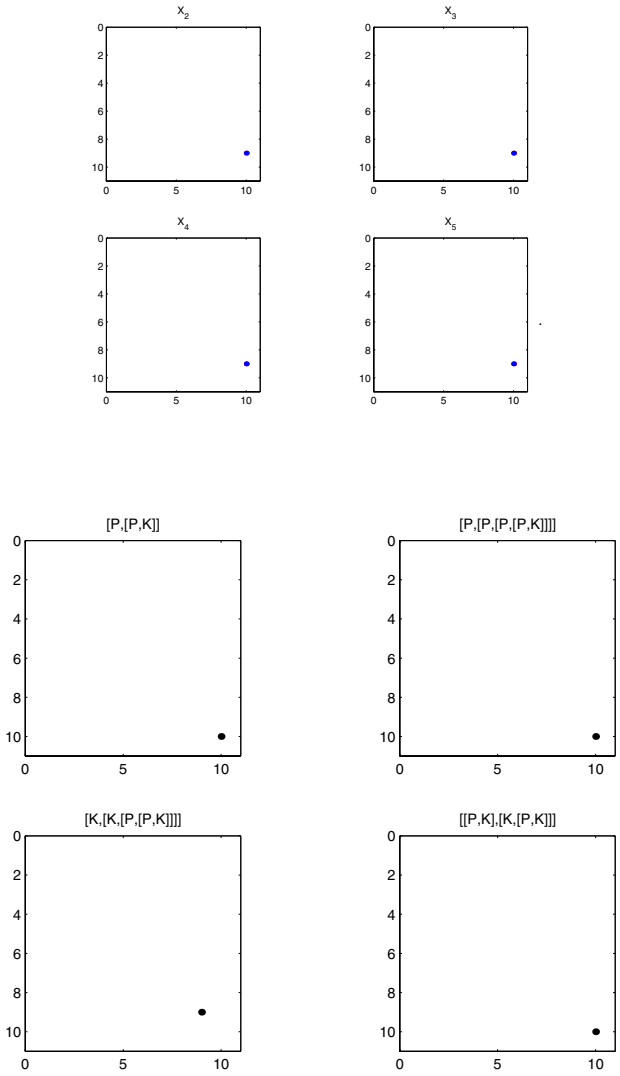


FIG. 4.1. *Commutators required for the generation of Y_6 for a matrix in a Hessenberg form.*

What about the exponential of X_p? Each such matrix is again of the form (3.5) and (3.6) still holds. Assume now that **a** has n nonzero elements, while **b** has only p nonzero elements, say $b_{n-p}, \ldots, b_{n-1}$. As displayed in Table 3.1, multiplying a vector by $n-1$ exponentials of matrices of decreasingly small dimension costs just $\mathcal{O}\big(pn^2\big)$ flops, while multiplying a matrix in a similar fashion carries the price tag of $\frac{1}{2}n^3 + \mathcal{O}\big(n^2\big)$ flops.

What is the cost of computing the commutators? Clearly, one has to take advantage of the sparsity of the matrices under consideration.

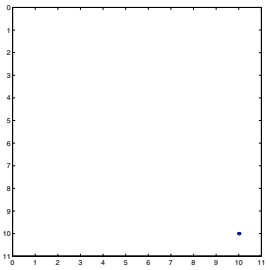


FIG. 4.2. *Sparsity pattern of Y_6 for a matrix in a Hessenberg form.*

We commence with the analysis of commutators of the type $[A, B]$, where $A \in \mathfrak{p}$ and $B \in \mathfrak{k}$. We write

$$A = \begin{bmatrix} O & \mathbf{a} \\ \mathbf{b}^\top & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 & \mathbf{0} \\ \mathbf{0}^\top & B_2 \end{bmatrix},$$

where $\mathbf{b}^\top = [0, 0, \ldots, b_{n-q}, \ldots, b_{n-1}]$, has only q nonzero elements, B_1 is $(n-1) \times (n-1)$ and in a Hessenberg form, while B_2 has a single nonzero entry. We have

$$[A, B] = \begin{bmatrix} O & \mathbf{a}B_2 - B_1\mathbf{a} \\ \mathbf{b}^\top B_1 - B_2\mathbf{b}^\top & 0 \end{bmatrix}.$$

Thus, computing $[A, B]$ amounts to the following:

- about $n(n-1)$ operations for the computation of $B_1\mathbf{a}$, since B_1 is in a Hessenberg form. Note that this reduces to about $2qn$ operations if only the last q columns of B_1 are nonzero.
- $(n-1)$ operations for $\mathbf{a}B_2$ and further $n-1$ operations to compute $\mathbf{a}B_2 - B_1\mathbf{a}$,
- about $q(q-1)$ operations for $\mathbf{b}^\top B_1$, q to compute $B_2\mathbf{b}$ and further q to compute $\mathbf{b}^\top B_1 - B_2\mathbf{b}^\top$.

In total, we require about n^2 operations (assuming that $q \ll n$). Running the commutators over matrices of decreasing dimension, we have in total

$$\sum_{j=1}^{n-1} j^2 \approx \frac{1}{3}n^3$$

operations. This is the cost of the commutators $[P, K], [K, [P, K]], [K, [K, [P, K]]], \ldots$, namely of $\mathfrak{p}$-elements with K.

Next, we consider commutators of the form $[A_1, A_2]$, where $A_1, A_2 \in \mathfrak{p}$ are bordered matrices. Set

$$A_1 = \begin{bmatrix} O & \mathbf{a} \\ \mathbf{b}^\top & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} O & \mathbf{c} \\ \mathbf{d}^\top & 0 \end{bmatrix},$$

where only the last q elements of $\mathbf{b}$, $\mathbf{d}$ are nonzero. Since

$$[A_1, A_2] = \begin{bmatrix} \mathbf{a}\mathbf{c}^\top - \mathbf{d}\mathbf{b}^\top & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{b}^\top\mathbf{d} - \mathbf{c}^\top\mathbf{a} \end{bmatrix},$$

it is evident that the computation of these commutators costs about $3q(n-1)$ operations, contributing a total of $\frac{3}{2}qn^2$ to the total cost (i.e., summing the contribution of similar terms over matrices of decreasing dimension). This is the cost of the commutators $[P,[P,K]],[P,[P,[P,[P,K]]]]$, etc. Finally, if $C_1, C_2 \in \mathfrak{k}$, with

$$C_1 = \begin{bmatrix} B_1 & \mathbf{0} \\ \mathbf{0}^\top & b_1 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} B_2 & \mathbf{0} \\ \mathbf{0}^\top & b_2 \end{bmatrix},$$

with B_1, B_2 of dimension n, one has

$$[C_1, C_2] = \begin{bmatrix} [B_1, B_2] & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{bmatrix}.$$

If only the last p (resp., r) columns of B_1, (resp., B_2) are nonzero, setting $r = \min\{p,q\}+1$, we deduce that the commutators $[C_1, C_2]$ cost about $6rn$ operations—an $\mathcal{O}\big(n^2\big)$ contribution when the count is carried over matrices of decreasing dimension.

Putting all the bricks together,

- Terms for order 2:

$$[P,K] \to \frac{1}{3}n^3.$$

- Order 3:

$$\begin{aligned} [K,[P,K]] &\to \frac{1}{3}n^3, \\ [P,[P,K]] &\to \mathcal{O}\big(n^2\big). \end{aligned}$$

- Order 4:

$$\begin{aligned} [P,[P,[P,K]]] &\to \mathcal{O}\big(n^2\big), \\ [K,[K,[P,K]]] &\to \frac{1}{3}n^3. \end{aligned}$$

- Order 5:

$$\begin{aligned} [K,[P,[P,[P,K]]]] &\to \frac{1}{3}n^3, \\ [K,[K,[K,[P,K]]]] &\to \frac{2}{3}n^3, \\ [[P,K],[P,[P,K]]] &\to \mathcal{O}\big(n^2\big), \\ [P,[P,[P,[P,K]]]] &\to \mathcal{O}\big(n^2\big), \\ [[P,K],[K,[P,K]]] &\to \mathcal{O}\big(n^2\big), \\ [K,[K,[P,[P,K]]]] &\to \mathcal{O}\big(n^2\big). \end{aligned}$$

In Table 4.1 we summarize the cost for the various stages of the exponential approximation of a generic matrix $Z \in \mathfrak{gl}(n)$ for orders $2, 3$, and 4 and compare our new algorithms with those proposed in [21]. For full matrices, it is evident that the benefits of our approach appear for orders greater than two. For order 4 the new

TABLE 4.1
Comparison of cost of the approximation of the exponential without (ZMK) and with reduction to Hessenberg form (IZ) for splittings of order $2, 3, 4$. *Only dominant terms are reported.*

Order	ZMK		IZ	
2	Vector	Matrix	Vector	Matrix
Hessenberg	–	–	$3\frac{1}{3}n^3$	$3\frac{1}{3}n^3$
Order cond.	$1\frac{1}{3}n^3$	$1\frac{1}{3}n^3$	$\frac{1}{3}n^3$	$\frac{1}{3}n^3$
Assembly exp	$3n^2$	$2n^3$	n^2	n^3
Total	$1\frac{1}{3}n^3$	$2\frac{1}{3}n^3$	$3\frac{2}{3}n^3$	$4\frac{2}{3}n^3$

Order	ZMK		IZ	
3	Vector	Matrix	Vector	Matrix
Hessenberg	–	–	$3\frac{1}{3}n^3$	$3\frac{1}{3}n^3$
Order cond.	$5n^3$	$5n^3$	$\frac{2}{3}n^3$	$\frac{2}{3}n^3$
Assembly exp	$3n^2$	$2n^3$	n^2	n^3
Total	$5n^3$	$7n^3$	$4n^3$	$5n^3$

Order	ZMK		IZ	
4	Vector	Matrix	Vector	Matrix
Hessenberg	–	–	$3\frac{1}{3}n^3$	$3\frac{1}{3}n^3$
Order cond.	$7n^3$	$7n^3$	n^3	n^3
Assembly exp	$3n^2$	$2n^3$	n^2	n^3
Total	$7n^3$	$9n^3$	$4\frac{1}{3}n^3$	$5\frac{1}{3}n^3$

algorithm is almost 40% faster than the one without transformation to an upper-Hessenberg form.

4.2. Symplectic matrices. The symplectic group of matrices $\mathrm{Sp}(n)$ is the set of all invertible matrices M of dimension $2n$ such that $MJM^\top = J$ where J is as in (4.1). The symplectic algebra $\mathfrak{sp}(n)$ is the set of $2n \times 2n$ matrices such that $NJ = -JN^\top$.

A symplectic matrix $M \in \mathrm{Sp}(n)$ is said to be in a *butterfly form* if it can be written as

$$M = \begin{bmatrix} D_1 & B_1 \\ D_2 & B_2 \end{bmatrix},$$

where D_1, D_2 are $n \times n$ diagonal matrices and B_1, B_2 are $n \times n$ tridiagonal matrices. This butterfly form was considered by Benner et al. [1] as a starting point of a QR-type algorithm (the SR-algorithm) to compute the eigenvalues of a symplectic matrix so that the transformed matrix remains symplectic at each step.

The transformation of a given symplectic matrix to a butterfly form can be performed by the use of three different types of similarity mappings with the following

matrices:

- *Symplectic Givens transformations:*

$$G = \left[\begin{array}{cc|cc} I_{k-1} & & & \\ & c & & s \\ & & I_{n-k} & \\ \hline & & I_{k-1} & \\ & -s & & c \\ & & & & I_{n-k} \end{array}\right].$$

- *Symplectic Householder transformations:*

$$H = \left[\begin{array}{cc|cc} I_{k-1} & & & \\ & Q & & \\ \hline & & I_{k-1} & \\ & & & Q \end{array}\right], \quad Q = I_{n-k+1} - \beta \mathbf{v}\mathbf{v}^\top, \quad \beta = \frac{2}{\|\mathbf{v}\|^2}.$$

- *Symplectic Gauss transformations:*

$$L = \left[\begin{array}{ccc|ccc} I_{k-2} & & & & & \\ & c & & & d & \\ & & c & & & \\ & & & I_{n-k} & & \\ \hline & & & I_{k-2} & & \\ & & & & c^{-1} & \\ & & & & & c^{-1} \\ & & & & & & I_{n-k} \end{array}\right].$$

The following algorithm for reduction to a butterfly form is described in more detail in [4].

Given a $2n \times 2n$ symplectic matrix M, compute its reduction to a butterfly form. M will be overwritten by its butterfly form.

```
for j = 1 : n − 1
  for k = n : −1 : j + 1
    compute G_k such that (G_k M)_{k+n,j} = 0
    M = G_k M G_k^⊤
  end
  if j < n − 1 then
    compute H_j such that (H_j M)_{j+2:n,j} = 0
    M = H_j M H_j^⊤
  end
  compute L_{j+1} such that (L_{j+1} M)_{j+1,j} = 0
  M = L_{j+1} M L_{j+1}^{−1}
  for k = n : −1 : j + 1
    compute G_k such that (M G_k)_{j,k} = 0
    M = G_k^⊤ M G_k
  end
  if j < n − 1 then
    compute H_j such that (M H_j)_{j,j+2+n:2n} = 0
    M = H_j^⊤ M H_j
  end
end
```

The algorithm introduces zeros in the rows by applying one of the above-mentioned transformations from the right, while zeros in the columns are obtained by applying the transformations from the left. To maintain similarity, the inverse of each transformation is applied also on the other side. The basic idea of the algorithm is, at each step j, (i) to bring the jth column of M into the desired form; (ii) to bring the $(n+j)$th row of M into the desired form. For more details, see [4].

Clearly, symplectic transformations can also be used at the algebra level. Our idea is to reduce a symplectic matrix $A \in \mathfrak{sp}(n)$ to a butterfly form, using the same algorithm as above. Note that

$$A \in \mathfrak{sp}(n) \Leftrightarrow A = \begin{bmatrix} A_1 & A_2 \\ A_3 & -A_1^\top \end{bmatrix},$$

where A_1, A_2, A_3 are $n \times n$ matrices and A_2, A_3 are symmetric; therefore, a butterfly matrix $B \in \mathfrak{sp}(n)$ must be of the form

$$B = \begin{bmatrix} B_1 & B_2 \\ B_3 & -B_1^\top \end{bmatrix},$$

where B_1 and B_3 are now diagonal, and B_2 is tridiagonal and symmetric.

Assume next that $B \in \mathfrak{sp}(n)$ is in a butterfly form. To approximate $\exp(tB)$ we use again a peel-up approach. However, the matrices S_i need be appropriately modified to preserve the $\mathfrak{sp}(n)$ structure. Set

$$\tilde{S}_i = \operatorname{diag}(-1)^{\tilde{\mathbf{s}}_i}, \quad \tilde{\mathbf{s}}_i = [s_{i;1}, \dots, s_{i;n}, s_{i;1}, \dots, s_{i;n}]. \tag{4.2}$$

In other words, the $\tilde{S}_i$s can be taken as direct products $S_i \times S_i$ of the matrices S_i in (2.11), and they act in the same manner on the first and the second n rows and columns of the matrix B.

To have a mental picture of the splitting induced by the automorphisms $\mathrm{Ad}_{\tilde{S}_i}$, we set $i = n$ and obtain subspaces $\mathfrak{p}$ and $\mathfrak{k}$ with the sparsity structure

$$\mathfrak{p} \ni \begin{bmatrix} 0 & \cdots & 0 & \times & 0 & \cdots & 0 & \times \\ \vdots & & \vdots & \times & \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \times & 0 & \cdots & 0 & \times \\ \times & \times & \times & 0 & \times & \times & \times & 0 \\ 0 & \cdots & 0 & \times & 0 & \cdots & 0 & \times \\ \vdots & & \vdots & \times & \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \times & 0 & \cdots & 0 & \times \\ \times & \times & \times & 0 & \times & \times & \times & 0 \end{bmatrix}, \quad \mathfrak{k} \ni \begin{bmatrix} \times & \cdots & \times & 0 & \times & \cdots & \times & 0 \\ \vdots & & \vdots & 0 & \vdots & & \vdots & \vdots \\ \times & \cdots & \times & 0 & \times & \cdots & \times & 0 \\ 0 & 0 & 0 & \times & 0 & 0 & 0 & \times \\ \times & \cdots & \times & 0 & \times & \cdots & \times & 0 \\ \vdots & & \vdots & 0 & \vdots & & \vdots & \vdots \\ \times & \cdots & \times & 0 & \times & \cdots & \times & 0 \\ 0 & 0 & 0 & \times & 0 & 0 & 0 & \times \end{bmatrix},$$

respectively. Observe that the automorphism targets the $2n$th and nth rows and columns.

The sparsity pattern of the computed terms in the generation of $X(t)$ up to $\mathcal{O}(t^5)$ is displayed in Figure 4.3 below.

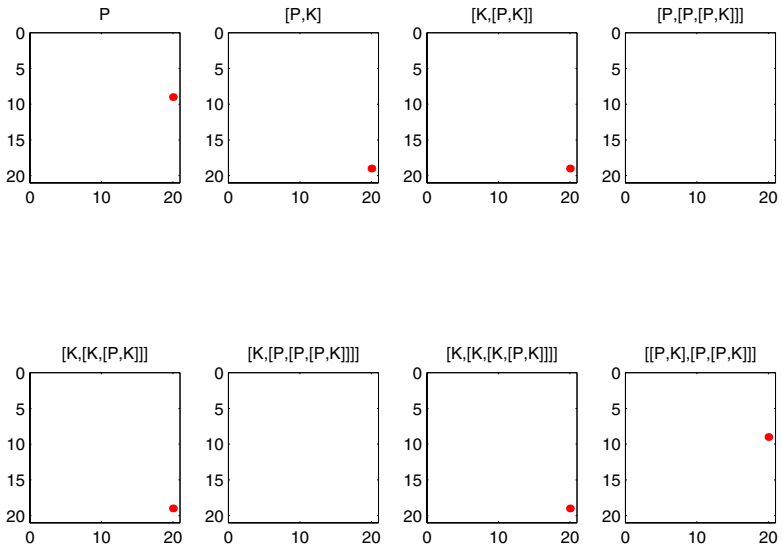


FIG. 4.3. *The sparsity pattern in the generation of* $X(t)$.

It is trivial to observe that $[P,[P,[P,K]]] = O$. This is not just a consequence of the reduction to a butterfly form, but of a more general result.

PROPOSITION 4.1. *Let* A *be a* $2n \times 2n$ *matrix, partitioned in* $n \times n$ *blocks, and assume that* A *is of the form*

$$\begin{bmatrix} O & A_{1,2} \\ O & O \end{bmatrix}.$$

Then, for any matrix $C \in M_{2n,2n}$, *it is true that*

$$[A,[A,[A,C]]] = O.$$

Proof. Let us partition the matrix C in blocks of the same size of those of A,

$$C = \begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix}.$$

By direct computation, we observe that

$$[A,C] = \begin{bmatrix} D_{1,1} & D_{1,2} \\ O & D_{2,2} \end{bmatrix},$$

where $D_{1,1} = A_{1,2}C_{2,1}$, $D_{1,2} = A_{1,2}C_{2,2} - C_{1,1}A_{1,2}$, and $D_{2,2} = -C_{2,1}A_{1,2}$. Similarly,

$$[A,[A,C]] = \begin{bmatrix} O & F_{1,2} \\ O & O \end{bmatrix},$$

where $F_{1,2} = -2A_{1,2}C_{1,2}A_{1,2}$. Finally, it is immediate to check that $[A,[A,[A,C]]] = O$. ∎

Note that the lemma is valid in the more general case, when A is $n \times n$, $A1,2$ is $n_1 \times n_2$, and $n_1 + n_2 = n$. The (trivial) extension of the proof is left to the reader.

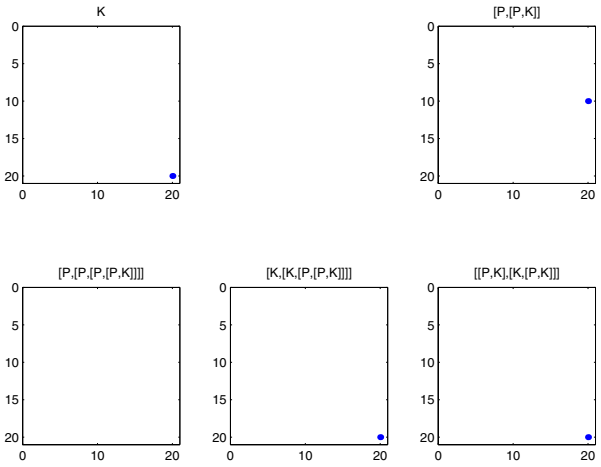


FIG. 4.4. *Commutators in the expansion of* $Y(t)$.

The terms required for the generation of $Y(t)$ are displayed in Figure 4.4 and the superposition of Y_5 (darker shade) and X_5 (lighter shade) is displayed in Figure 4.5 (for convenience, we have plotted X_5 with $\times$).

In Figure 4.5 we observe that no fill-in is introduced at least up to order 6. This is most welcome news because we do not need to use extra computation to annihilate further entries.

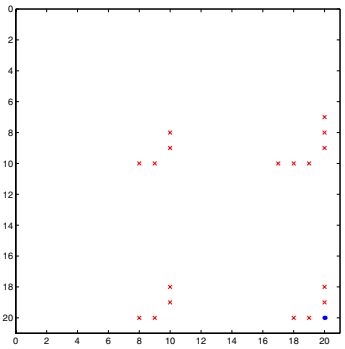


FIG. 4.5. *Superposition of* X_5 *(crosses) and* Y_5 *(dots). No fill-in is introduced in the* $\mathfrak{k}$ *subalgebra.*

4.3. The cost of reduction to a butterfly form. Assume that we have already partially reduced the matrix M to a butterfly form, where the blocks L_k, M_k, N_k have dimension k and N_k, M_k are symmetric (see Figure 4.6). To set to zero the terms in the leading column of N_k we apply symplectic Givens rotations from the left. Each of these rotations requires about $6k$ operations (multiplications and additions) for the update of L_k, M_k, and $3k$ operations for the update of N_k, for a total of $9k$ operations. When we apply their transpose from the right, we can take into account the symmetry of the $(1,2)$ and $(2,1)$ blocks, so that only L_k needs be updated, hence

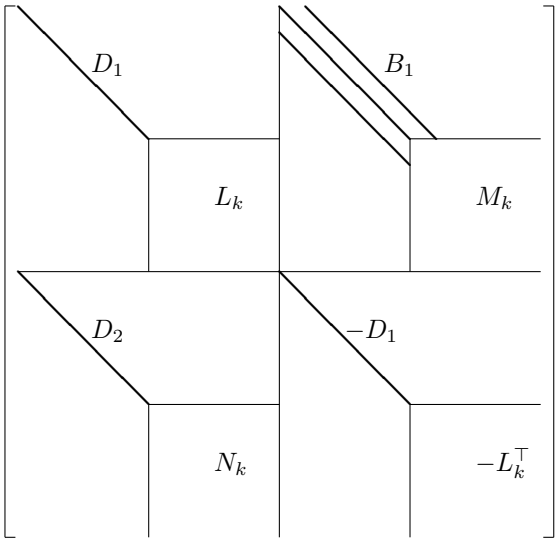


FIG. 4.6. *A matrix in* $\mathfrak{sp}(n)$ *partially reduced to a butterfly form.*

further $6k$ operations, for a total of $15k$ operations. A similar count holds for the Givens rotations that are applied to annihilate the top-row elements of L_k, hence *in toto* Givens rotations account for $30k$. Since for each column/row there are k of those Givens rotation, for matrices of decreasing dimension, neglecting lower order terms we have

$$\text{total cost of Givens rotations} \approx 30 \sum_{k=1}^{n} k^2 \approx 10n^3.$$

The application of Householder symplectic reflections reduces to the application of standard Householder to the blocks L_k, M_k, N_k, and costs $3\frac{1}{3}n^3$, n^3, and n^3, respectively, since the two latter blocks are symmetric. We need to apply two sets of such Householder reflections for a total of

$$\text{total cost of Householder reflections} \approx 10\frac{2}{3}n^3 \text{ operations.}$$

Since the cost of Gauss transformations is of a lower order of magnitude, the total cost of reduction to a butterfly form is

$$20\frac{2}{3}n^3 \text{ operations,}$$

which is not really prohibitive, given that the matrix has dimension $2n$.

5. A divide-and-conquer strategy. In what follows, we shall introduce an alternative approach that can be particularly useful in the context of very large n and parallel computing. The main idea is to choose an automorphism Ad_S so that the matrix Y_p is reducible, hence the exponential of each submatrix can be computed separately and possibly in parallel.

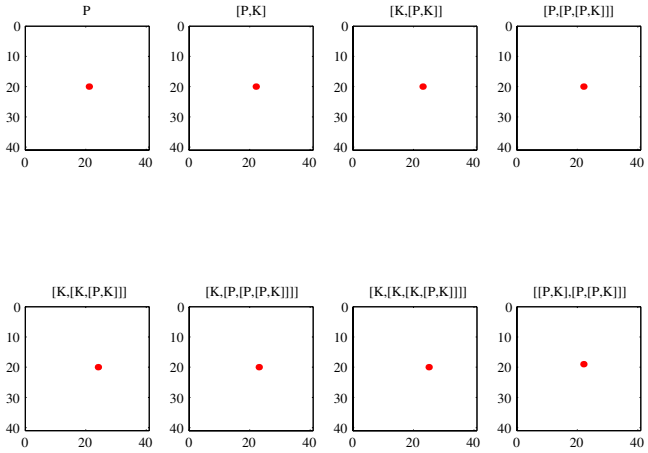


FIG. 5.1. *Sparsity pattern of matrices in* $\mathfrak{p}$ *in the divide-and-conquer approach for skew-symmetric matrices.*

5.1. Skew-symmetric matrices. We again commence our exposition with $Z \in \mathfrak{so}(n)$ and assume that it is already in tridiagonal form. Our point of departure is to consider an inner automorphism Ad_G where

$$G = \begin{bmatrix} I_{n_1 \times n_1} & O_{n_1 \times n_2} \\ O_{n_2 \times n_1} & -I_{n_2 \times n_2} \end{bmatrix},$$

where $n_1 + n_2 = n$: an obvious choice is $n_1 = \lfloor n/2 \rfloor$, however, other choices are possible, e.g., the index corresponding to the least off-diagonal element. Then,

$$\mathfrak{k} \ni \begin{bmatrix} K^{(1)}_{n_1 \times n_1} & O_{n_1 \times n_2} \\ O_{n_2 \times n_1} & K^{(2)}_{n_2 \times n_2} \end{bmatrix}, \qquad \mathfrak{p} \ni \begin{bmatrix} O_{n_1 \times n_1} & P^{(1)}_{n_1 \times n_2} \\ P^{(2)}_{n_2 \times n_1} & O_{n_2 \times n_2} \end{bmatrix}.$$

Therefore,

$$K = \begin{bmatrix} K_1 & O \\ O & K_2 \end{bmatrix}, \qquad P = \begin{bmatrix} O & P_1 \\ P_2 & O \end{bmatrix}.$$

We stress that both K_1 and K_2 are tridiagonal while P_1 and P_2 have a single nonzero entry, in the lower left and upper right corner, respectively. We write $P_1 = c_1 \mathbf{e}_{n_1,n_1} \mathbf{e}^\top_{n_2,1}$, $P_2 = c_2 \mathbf{e}_{n_2,1} \mathbf{e}^\top_{n_1,n_1}$, where $\mathbf{e}_{m,k} \in \mathbb{R}^m$ is the kth unit vector.

In Figures 5.1 and 5.2 we display the sparsity pattern of elements in $\mathfrak{p}$ and in $\mathfrak{k}$, respectively, while in Figure 5.3 the matrices X and Y for different orders are superposed (truncations of X are denoted in lighter shade, while the darker shade corresponds to truncations of Y).

The following observations form the basis for an efficient divide-and-conquer algorithm to compute the exponential function in $\mathfrak{so}(n)$:

- All the commutators, hence also X and Y (up to the requisite order), can be evaluated in $\mathcal{O}(1)$ flops.

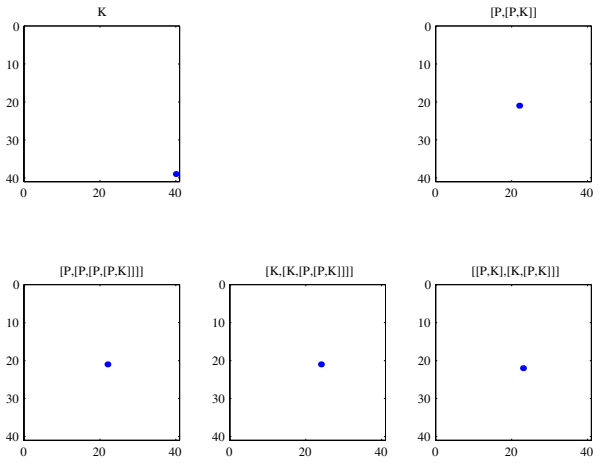


FIG. 5.2. *Sparsity pattern of matrices in* $\mathfrak{k}$ *in the divide-and-conquer approach for skew-symmetric matrices.*

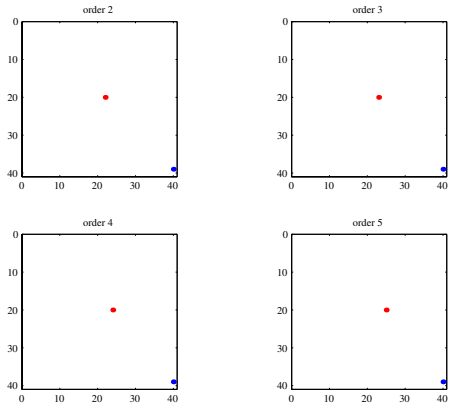


FIG. 5.3. *Superposition of truncations of* $X(t)$ *(lighter shade) and* $Y(t)$ *(darker shade) of order* $2, 3, 4, 5$ *in the divide-and-conquer approach for skew-symmetric matrices.*

- The exact exponential of X reduces to that of a small matrix, hence can be evaluated in $\mathcal{O}(1)$ flops.
- Y is a reducible matrix.
- The departure of Y from tridiagonal can be corrected in a small number of Givens rotations. Because of reducibility, we can act *separately* on each of the two components, hence the outcome is two tridiagonal matrices of size $n_1 \times n_1$ and $n_2 \times n_2$, respectively. Again, the cost is $\mathcal{O}(1)$ flops.

We can now continue with the two pieces of Y in a similar vain, splitting them into progressively smaller pieces. All this is similar to many familiar techniques in numerical linear algebra, not least domain decomposition. Altogether, we require

$\log_2 n$ stages to parcel out the exponential of a tridiagonal $Z \in \mathfrak{so}(n)$ into a product of rank-2 orthogonal matrices, although in practice this divide-and-conquer technique can terminate with matrices of higher rank.

5.2. General matrices. Let $Z \in \mathfrak{gl}(n)$ or $\mathfrak{sl}(n)$ and suppose that we have already brought it to an upper-Hessenberg form. Proceeding as before, P_1 is a dense matrix, while $P_2 = c\mathbf{e}_{n_2,1}\mathbf{e}_{n_1,n_1}^\top$.

Again, we can always bring Y into an upper-Hessenberg form in $\mathcal{O}(1)$ Givens rotations. More interesting is the evaluation of $\exp X$. Note that

$$X = \begin{bmatrix} O & X_1 \\ X_2 & O \end{bmatrix},$$

where X_1 is $n_1 \times n_2$ and dense, while X_2 is $n_2 \times n_1$ and zero except for a $q \times q$ block in the upper right corner, where $q \geq 1$. Let

$$V = X_1 X_2, \qquad W = X_2 X_1.$$

Then

$$X^{2m} = \begin{bmatrix} V^m & O \\ O & W^m \end{bmatrix}, \qquad X^{2m+1} = \begin{bmatrix} V^m & O \\ O & W^m \end{bmatrix} \begin{bmatrix} O & X_1 \\ X_2 & O \end{bmatrix}, \quad m \in \mathbb{Z}_+,$$

therefore,

$$\exp(X) = \begin{bmatrix} C(V) & S(V)X_1 \\ S(W)X_2 & C(W) \end{bmatrix},$$

where

$$C(Z) = \sum_{m=0}^{\infty} \frac{Z^m}{(2m)!} = \cosh Z^{1/2}, \qquad S(Z) = \sum_{m=0}^{\infty} \frac{Z^m}{(2m+1)!} = \sinh Z^{1/2}.$$

We write X_1 and X_2 in a compound form,

$$X_1 = \begin{bmatrix} T_{1,1} & T_{1,2} \\ T_{2,1} & T_{2,2} \end{bmatrix}, \qquad X_2 = \begin{bmatrix} O & R \\ O & O \end{bmatrix},$$

where

$$T_{1,1} \in \mathrm{M}_{(n_1-p)\times p}, \quad T_{1,2} \in \mathrm{M}_{(n_1-p)\times(n_2-p)}, \quad T_{2,1} \in \mathrm{M}_{p\times p}, \quad T_{2,2} \in \mathrm{M}_{p\times(n_2-p)}$$

and $R \in \mathrm{M}_{p\times p}$, where $\mathrm{M}_{n\times m}$ denotes the set of $n \times m$ matrices. Therefore,

$$V = \begin{bmatrix} O & V_1 \\ O & V_2 \end{bmatrix}, \qquad W = \begin{bmatrix} W_1 & W_2 \\ O & O \end{bmatrix},$$

where

$$\begin{aligned} V_1 &= T_{1,1}R \in \mathrm{M}_{(n_1-p)\times p}, \qquad V_2 = T_{2,1}R \in \mathrm{M}_{p\times p}, \\ W_1 &= RT_{2,1} \in \mathrm{M}_{p\times p}, \qquad W_2 = RT_{2,2} \in \mathrm{M}_{p\times(n_2-p)}. \end{aligned}$$

In particular, note that $V_2, W_1 \in \mathrm{M}_{p\times p}$, hence they are square and *small!*
We can easily prove by induction that

$$V^m = \begin{bmatrix} O & V_1 V_2^{m-1} \\ O & V_2^m \end{bmatrix}, \qquad W^m = \begin{bmatrix} W_1^m & W_1^{m-1} W_2 \\ O & O \end{bmatrix}, \quad m \in \mathbb{Z}.$$

Therefore, simple calculation affirms that

$$\begin{aligned} C(V) &= \begin{bmatrix} I & V_1 V_2^{-1}[C(V_2) - I] \\ O & C(V_2) \end{bmatrix}, \\ C(W) &= \begin{bmatrix} C(W_1) & [C(W_1) - I] W_1^{-1} W_2 \\ O & I \end{bmatrix}, \\ S(V)X_1 &= \begin{bmatrix} I & V_1[S(V_2) - I]V_2^{-1} \\ O & S(V_2) \end{bmatrix} \begin{bmatrix} T_{1,1} & T_{1,2} \\ T_{2,1} & T_{2,2} \end{bmatrix}, \\ S(W)X_2 &= \begin{bmatrix} O & S(W_1)R \\ O & O \end{bmatrix}. \end{aligned}$$

Given that $p \ll n_1, n_2$ and $n_1 + n_2 = n$, we have the following cost (disregarding lower order terms),

1. Computing $C(V)$: $n_1 p^2$ flops;
2. Computing $C(W)$: $n_2 p^2$ flops;
3. Computing $S(V)$: $n_1 p^2$ flops;
4. Multiplying $S(V)X_1$: $n_1 n_2 p + n p^2$;
5. Computing $S(W)X_2$: p^3 flops.[1]

Hence, altogether the cost is $n_1 n_2 p$: if $n_1 = n_2 = n/2$ then the entire cost of computing the exponential *exactly* is just $\frac{1}{4} n^2 p$.

Suppose that $n = 2^s$ and $n_1 = n_2 = 2^{s-1}$, whence the cost is $\approx p 2^{2s-2}$. Moreover, we continue with the divide-and-conquer technique. In the next stage we have two $2^{s-1} \times 2^{s-1}$ matrices, then four $2^{s-2} \times 2^{s-2}$ matrices and so on. The entire cost of computing all the exponentials then becomes

$$p \sum_{r=1}^{s} 2^{2s-r-1} = \frac{1}{2} pn(n-1) \approx \frac{1}{2} pn^2.$$

The above is true on a serial machine. If the calculations for different pieces of the matrix are performed in parallel, we have instead just a single $2^{s+1-r} \times 2^{s+1-r}$ matrix to deal with in the rth stage and the overall cost is

$$p \sum_{r=1}^{s} 2^{s+1-r} \approx 2p(n-1).$$

5.3. The cost of computing commutators. Assume that P and K are as in Figures 5.4 and 5.5,

$$P = \begin{bmatrix} O & P_1 \\ P_2 & O \end{bmatrix}, \qquad K = \begin{bmatrix} K_1 & O \\ O & K_2 \end{bmatrix}, \tag{5.1}$$

[1] Note that we do not need to compute $S(W)$ first—only $S(W_1)$.

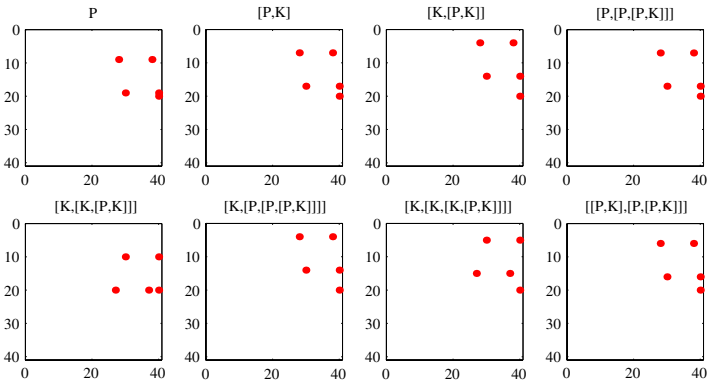


FIG. 5.4. *Sparsity structure for the first terms in* $\mathfrak{p}$ *in the divide-and-conquer approach (upper Hessenberg matrices).*

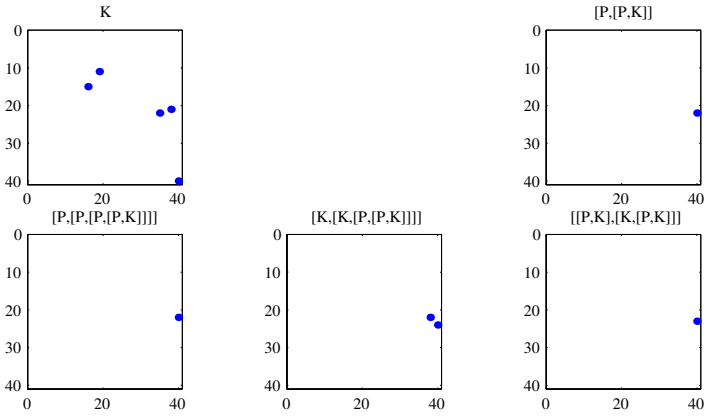


FIG. 5.5. *Sparsity structure for the first terms in* $\mathfrak{k}$ *in the divide-and-conquer approach (upper Hessenberg matrices).*

where the blocks P_i and K_i have dimension k. Note that

$$\begin{aligned}\left[\begin{bmatrix} O & P_1 \\ P_2 & O \end{bmatrix}, \begin{bmatrix} O & R_1 \\ R_2 & O \end{bmatrix}\right] &= \begin{bmatrix} O & P_1R_2 - R_1P_2 \\ P_2R_1 - R_2P_1 & O \end{bmatrix}, \\ \left[\begin{bmatrix} O & P_1 \\ P_2 & O \end{bmatrix}, \begin{bmatrix} K_1 & O \\ O & K_2 \end{bmatrix}\right] &= \begin{bmatrix} O & P_1K_2 - K_1P_1 \\ P_2K_1 - K_2P_1 & O \end{bmatrix}.\end{aligned} \tag{5.2}$$

From (5.2) we conclude that the most expensive commutators are those of the form $[P, K]$, for which we need to compute $P_1K_2 - K_1P_1$, amounting to about $2k^3$ operations (counting addition and multiplication). All the remaining commutators are of lower complexity, which we ignore. For a splitting of order 5 (as depicted in Figure 5.6), we need to compute seven such commutators, amounting to $14k^3$. Next, assume that $k = 2^s$ and that there are $2^{\log_2 n - s}$ such blocks. For order five, the total cost of

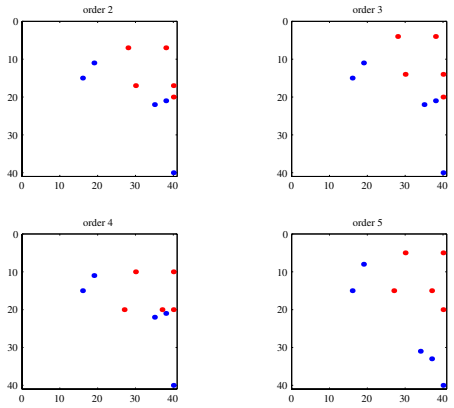


FIG. 5.6. *A superposition of X_5 and Y_5 in the divide-and-conquer approach (upper Hessenberg matrices).*

commutators (disregarding lower order terms) amounts to

$$14n \sum_{s=1}^{\log_2 n} 2^{2s} \approx 5n^3$$

operations on a serial machine. When implementing this in parallel on $\log_2 n$ processors (that is, when the commutators of blocks of dimension 2^s are evaluated simultaneously), the cost reduces to

$$14 \sum_{s=1}^{\log_2 n} 2^{3s} \approx 2n^3.$$

6. Other groups. In this section we discuss briefly GPD for a number of more unusual Lie groups which, nonetheless, feature in applications.

6.1. Lorentz-type groups SO(p, q). Let

$$J = \begin{pmatrix} I_p & O \\ O & -I_q \end{pmatrix},$$

and consider the group $\mathrm{SO}(p,q) = \{x : xJx^\top = J\}$. As is well known, the corresponding algebra is $\mathfrak{so}(p,q) = \{Z : ZJ = -JZ^\top\}$. The block form of Z is

$$Z = \begin{pmatrix} Z_1 & Z_2 \\ Z_2^\top & Z_3 \end{pmatrix},$$

where Z_1 and Z_3 are skew-symmetric. The most widespread Lorentz-type groups in applications are $\mathrm{SO}(3,1)$ and $\mathrm{SO}(5,2)$, for which it is not too costly to compute the exponential exactly given the low dimension. Algorithms for computing the exact exponential of these matrices have been proposed in [13].

In what follows, we focus instead on the less ordinary case when $p + q = n$ is large, yet $p \ll q$. The basic idea consists of splitting

$$Z = \begin{pmatrix} O & Z_2 \\ Z_2^\top & O \end{pmatrix} + \begin{pmatrix} Z_1 & O \\ O^\top & Z_3 \end{pmatrix} = P + K$$

so that the problem is reduced to computing the exponential of skew-symmetric matrices and that of a (symmetric) bordered matrix.

The only issue that can cause complications is the computation of commutators with P, K (especially if we desire high order). To do this in a cheaper manner, we consider the matrix

$$H = \begin{pmatrix} I & O \\ O^\top & H_2 \end{pmatrix},$$

where H_2 is the matrix that QR factorizes P (a product of p elementary Householder reflections). Computing commutators with P (ad_P) now costs $2p^2q$. Commutators with K are more expensive—at present we do not see how this can be avoided but perhaps one can exploit skew symmetry.

However, these order conditions are used only once. Once we have split into the symmetric and skew-symmetric parts, the problem reduces to the approximation of exponentials of skew-symmetric matrices, which has been described at length earlier in this paper.

6.2. Isotropy groups. In this section we consider the computation of the exponential in isotropy groups. Recall that the (left) isotropy group G_V at $V \in \mathrm{M}_{n\times m}$ is the group of matrices x that leaves V fixed under left multiplication,

$$G_V = \{x \in \mathrm{GL}(n) : xV = V\}$$

(see, for instance, [17]). The corresponding algebra can be easily computed,

$$\mathfrak{g}_V = \{X \in \mathfrak{gl}(n) : XV = V\}.$$

Let us assume that $m \le n$ and that V has rank m (if V has rank less then m then it is possible to ignore some of its columns so that the resulting matrix has full rank). In that case, the problem essentially reduces to the isotropy group (isomorphic to G_V)

$$G_{\tilde R} = \{y : y\tilde R = \tilde R\}, \quad \tilde R = \begin{bmatrix} R \\ O \end{bmatrix},$$

where R is $m \times m$ upper triangular and $G_{\tilde R} = Q^\top G_V Q$, Q being the orthogonal matrix that performs the QR factorization of V, i.e., $V = Q\tilde R$. In what follows, we abuse notation and write G_R instead of $G_{\tilde R}$, hoping that this does not cause confusion of types.

Let us study in greater detail the elements of G_R. Assume that

$$y = \begin{bmatrix} y_{1,1} & y_{1,2} \\ y_{2,1} & y_{2,2} \end{bmatrix},$$

where $y_{1,1}$ is $m \times m$, $y_{1,2}$ and $y_{2,1}^\top$ are $m \times (n-m)$, and $y_{2,2}$ is $(n-m) \times (n-m)$. Imposing $y\tilde R = \tilde R$ we obtain the following conditions:

$$\begin{aligned} y_{1,1}R = R \quad &\Rightarrow \quad y_{1,1} = I_{m\times m} \\ y_{2,1}R = O_{(n-m)\times m} \quad &\Rightarrow \quad y_{2,1} = O_{(n-m)\times m} \\ y_{1,2}, y_{2,2} \quad &\text{arbitrary} \end{aligned}$$

(recall that R has full rank, hence it is invertible).

Correspondingly, at the algebra level, we set

$$Y = \begin{bmatrix} Y_{1,1} & Y_{1,2} \\ Y_{2,1} & Y_{2,2} \end{bmatrix},$$

which, in tandem with the algebra conditions, implies

$$\begin{aligned} Y_{1,1} &= O_{m\times m}, \\ Y_{2,1} &= O_{(n-m)\times m}, \\ Y_{1,2}, Y_{2,2} &\quad \text{arbitrary.} \end{aligned}$$

Note that $Y_{1,2}$, $Y_{2,2}$ can be considered as free parameters. Although their action does not change $\tilde{R}$, they do move points around $\tilde{R}$. As a possible application, in [14] isotropy at a point is used to improve retention of qualitative features by numerical integrators for ODEs.

The exponential of Y can be evaluated exactly by the formula

$$\exp(tY) = \begin{bmatrix} I_{m\times m} & Y_{1,2}\phi(Y_{2,2}) \\ O & \exp(tY_{2,2}) \end{bmatrix},$$

where

$$\phi(Z) = Z^{-1}[\exp(Z) - I] = \sum_{k=0}^{\infty} \frac{1}{(k+1)!} Z^k.$$

Thus, the problem reduces to computing the exponential of $Y_{2,2}$, which is of dimension $(n-m)$. Now, $Y_{2,2}$ can be reduced to Hessenberg form and its exponential computed as in the general $\mathrm{GL}(n)$ case.

6.3. Scaling groups. We commence as in the case of the isotropy groups. A one-parameter curve in the scaling group G_V^{sc} of $V = (\mathbf{v}_1, \dots, \mathbf{v}_m) \in \mathbb{R}^{n\times m}$,

$$G_V^{\mathrm{sc}} = \{x : \exists \lambda_1, \lambda_2, \dots, \lambda_m \text{ such that } x\mathbf{v}_i = \lambda_i \mathbf{v}_i,\ i = 1, \dots, m\},$$

satisfies

$$x(t)V = V\Lambda(t),$$

where $\Lambda(t)$ is a smooth diagonal matrix function [17]. Again, performing a QR decomposition of V, we find that G_V^{sc} is isomorphic to $Q^\top G_V^{\mathrm{sc}} Q = G_{\tilde{R}}^{\mathrm{sc}}$, the scaling group of the upper triangular matrix $\tilde{R}$, where $Q\tilde{R} = V$.

At the algebra level $\mathfrak{g}_R^{\mathrm{sc}}$, we set

$$Y = \begin{bmatrix} Y_{1,1} & Y_{1,2} \\ Y_{2,1} & Y_{2,2} \end{bmatrix},$$

which, in tandem with the algebra condition

$$Y\tilde{R} = \tilde{R}\Lambda'(t)$$

implies that

$$\begin{aligned} Y_{1,1} &= R\Lambda'(t)R^{-1}, \\ Y_{2,1} &= O_{(n-m)\times m}, \\ Y_{1,2}, Y_{2,2} &\quad \text{arbitrary.} \end{aligned}$$

Again, $Y_{1,2}, Y_{2,2}$ can be considered as free parameters: their action does not move/scale $\tilde{R}$ but only the points in its neighborhood. It is also useful to note that $Y_{1,1}$ is upper triangular.

Thus, the problem reduces to computing exponentials of block-upper triangular matrices of the form

$$Y = \begin{bmatrix} Y_{1,1} & Y_{1,2} \\ O & Y_{2,2} \end{bmatrix},$$

where $Y_{1,1}$ is upper triangular.

In general, the exponential of a block triangular matrix can be evaluated exactly by the formula

$$\exp\left(t \begin{bmatrix} A & B \\ O & C \end{bmatrix}\right) = \begin{bmatrix} \exp(tA) & \int_0^t \mathrm{e}^{(t-\tau)A} B \mathrm{e}^{\tau C} \mathrm{d}\tau \\ 0 & \exp(tC) \end{bmatrix}, \tag{6.1}$$

however, the integral might be a difficult to compute exactly. It could be approximated by quadrature formulae, but this will require the computation of roots of matrices, adding an extra layer of complexity.

We could again use the *divide-and-conquer* approach, split

$$Y = K + P, \quad K = \begin{bmatrix} Y_{1,1} & O \\ O & Y_{2,2} \end{bmatrix}, \quad P = \begin{bmatrix} O & Y_{2,1} \\ O & O \end{bmatrix}.$$

It can be observed that Y is block upper triangular and so are K and P. Since triangular (and block triangular) matrices form subalgebras, $\mathfrak{k}$ and $\mathfrak{p}$ also consist of block triangular matrices (in other words, the lower (2,1) block never fills in the $\mathfrak{p}$ part).

Now,

$$[P, K] = \begin{bmatrix} O & Y_{1,2}Y_{2,2} - Y_{1,1}Y_{1,2} \\ O & O \end{bmatrix},$$

moreover, if $P_1, P_2 \in \mathfrak{p}$, it is easily verified that

$$O = [P_1, P_2] \in \mathfrak{k}.$$

In other words, K *does not need any order corrections* and the only nonzero commutators are those of the form

$$[K, [K, [\ldots, [K, [P, K]]]]]$$

in the $\mathfrak{p}$ part.

Moreover, observe that

$$\begin{bmatrix} O & B \\ O & O \end{bmatrix}^2 = O,$$

therefore,

$$\exp(t\tilde{P}) = I + t\tilde{P}, \quad \tilde{P} \in \mathfrak{p},$$

while, for matrix $\tilde{K} = \left[\begin{smallmatrix} A & O \\ O & C \end{smallmatrix}\right] \in \mathfrak{k}$, one has

$$\exp(t\tilde{K}) = \begin{bmatrix} \exp(tA) & O \\ O & \exp(tC) \end{bmatrix}.$$

An alternative is to expand the integral in (6.1) in Taylor series and truncate to appropriate order.

Appendix A. Herewith, we report the coefficients for the commutators in (3.16).

$$
\begin{aligned}
g_{n,n-3} &= -\beta_{n-3}c_{n,n-2},\\
g_{n,n-2} &= (\alpha_n-\alpha_{n-2})c_{n,n-2}-\beta_{n-2}c_{n,n-1},\\
g_{n,n-1} &= (\alpha_n-\alpha_{n-1})c_{n,n-1}-\gamma_{n-2}c_{n,n-2},\\
g_{n-3,n} &= \gamma_{n-3}c_{n-2,n},\\
g_{n-2,n} &= -(\alpha_n-\alpha_{n-2})c_{n-2,n}+\gamma_{n-2}c_{n-1,n},\\
g_{n-1,n} &= -(\alpha_n-\alpha_{n-1})c_{n-1,n}-\beta_{n-2}c_{n-2,n},
\end{aligned}
\tag{A.1}
$$

$$
\begin{aligned}
h_{n,n-4} &= -\beta_{n-4}g_{n,n-3},\\
h_{n,n-3} &= (\alpha_n-\alpha_{n-3})g_{n,n-3}-\beta_{n-3}g_{n,n-2},\\
h_{n,n-2} &= (\alpha_n-\alpha_{n-2})g_{n,n-2}-\gamma_{n-3}g_{n,n-3}-\beta_{n-2}g_{n,n-1},\\
h_{n,n-1} &= (\alpha_n-\alpha_{n-1})g_{n,n-1}-\gamma_{n-2}g_{n,n-2},\\
h_{n-4,n} &= \gamma_{n-4}g_{n-3,n},\\
h_{n-3,n} &= -(\alpha_n-\alpha_{n-3})g_{n-3,n}+\gamma_{n-3}g_{n-2,n},\\
h_{n-2,n} &= -(\alpha_n-\alpha_{n-2})g_{n-2,n}+\beta_{n-3}g_{n-3,n}+\gamma_{n-2}g_{n-1,n},\\
h_{n-1,n} &= -(\alpha_n-\alpha_{n-1})g_{n-1,n}+\beta_{n-2}g_{n-2,n},
\end{aligned}
\tag{A.2}
$$

$$
\begin{aligned}
j_{n,n-5} &= \beta_{n-5}h_{n,n-4},\\
j_{n,n-4} &= (\alpha_n-\alpha_{n-4})h_{n,n-4}-\beta_{n-4}h_{n,n-2},\\
j_{n,n-3} &= -\gamma_{n-4}h_{n,n-4}+(\alpha_n-\alpha_{n-3})h_{n,n-3}-\beta_{n-3}h_{n,n-1},\\
j_{n,n-2} &= -\gamma_{n-3}h_{n,n-4}+(\alpha_n-\alpha_{n-2})h_{n,n-2}-\beta_{n,2}h_{n,n-1},\\
j_{n,n-1} &= -\gamma_{n-2}h_{n,n-2}+(\alpha_n-\alpha_{n-1})h_{n,n-1},\\
j_{n-5,n} &= \gamma_{n-5}h_{n-3,n},\\
j_{n-4,n} &= -(\alpha_n-\alpha_{n-4})h_{n-4,n}+\gamma_{n-4}h_{n-3,n},\\
j_{n-3,n} &= \beta_{n-4}h_{n-4,n}-(\alpha_n-\alpha_{n-3})h_{n-4,n}+\gamma_{n-3}h_{n-2,n},\\
j_{n-2,n} &= \beta_{n-3}h_{n-3,n}-(\alpha_n-\alpha_{n-2})h_{n-2,n}+\gamma_{n-2}h_{n-1,n},\\
j_{n-1,n} &= \beta_{n-2}h_{n-2,n}-(\alpha_n-\alpha_{n-1})h_{n-1,n},
\end{aligned}
\tag{A.3}
$$

$$
\begin{aligned}
k_{n,n-2} &= c_{n,n-1}d_{n-1,n-2}-c_{n,n-2}d_{n,n},\\
k_{n,n-1} &= c_{n,n-2}d_{n-2,n-1}-c_{n,n-1}(d_{n,n}-d_{n-1,n-1}),\\
k_{n-2,n} &= c_{n-2,n}d_{n,n}-d_{n-2,n-1}c_{n-1,n},\\
k_{n-1,n} &= (d_{n,n}-d_{n-1,n-1})c_{n-1,n}-d_{n-1,n-2}c_{n-2,n},
\end{aligned}
\tag{A.4}
$$

$$
\begin{aligned}
l_{n-1,n-4} &= \beta_{n-3}\beta_{n-4}d_{n-1,n-2},\\
l_{n-1,n-3} &= -\beta_{n-3}[(2\alpha_{n-1}-\alpha_{n-2}-\alpha_{n-3})d_{n-1,n-2}-d_{n-1,n-1}\beta_{n-2}],\\
l_{n-1,n-2} &= [(\alpha_{n-1}-\alpha_{n-2})^2+2\beta_{n-2}\gamma_{n-2}+\beta_{n-3}\gamma_{n-3}]d_{n-1,n-2}\\
&\quad -(\alpha_{n-1}-\alpha_{n-2})\beta_{n-2}d_{n-1,n-1}-2\beta_{n-1}^2d_{n-2,n-1},\\
l_{n-1,n-1} &= -l_{n-2,n-2},\\
l_{n-2,n-3} &= \beta_{n-3}(-2\gamma_{n-2}d_{n-1,n-2}+\beta_{n-2}d_{n-2,n-1}),\\
l_{n-2,n-2} &= (\alpha_{n-1}-\alpha_{n-2})(\gamma_{n-2}d_{n-1,n-2}+\beta_{n-2}d_{n-2,n-1})\\
&\quad -2\gamma_{n-2}\beta_{n-2}d_{n-1,n-1},\\
l_{n-2,n-1} &= [(\alpha_{n-1}-\alpha_{n-2})^2+2\beta_{n-2}\gamma_{n-2}+\gamma_{n-3}\beta_{n-3}]d_{n-2,n-1}\\
&\quad -(\alpha_{n-1}-\alpha_{n-2})\gamma_{n-2}d_{n-1,n-1},
\end{aligned}
\tag{A.5}
$$

$$
\begin{aligned}
l_{n-3,n-2} &= \gamma_{n-3}(\gamma_{n-2}d_{n-1,n-2} - 2\beta_{n-2}d_{n-2,n-1}),\\
l_{n-3,n-1} &= -\gamma_{n-3}[(2\alpha_{n-1} - \alpha_{n-2} - \alpha_{n-3})d_{n-2,n-1} - \gamma_{n-2}d_{n-1,n-1}],\\
l_{n-4,n-1} &= \gamma_{n-3}\gamma_{n-4}d_{n-2,n-1},
\end{aligned}
$$

$$
\begin{aligned}
m_{n-1,n-3} &= c_{n-1,n}g_{n,n-3},\\
m_{n-1,n-2} &= c_{n-1,n}g_{n,n-2} - c_{n,n-2}g_{n-1,n},\\
m_{n-1,n-1} &= c_{n-1,n}g_{n,n-1} - c_{n,n-1}g_{n-1,n},\\
m_{n-2,n-3} &= c_{n-2,n}g_{n,n-3},\\
m_{n-2,n-2} &= c_{n-2,n}g_{n,n-2} - c_{n,n-2}g_{n-2,n},\\
m_{n-2,n-3} &= c_{n-2,n}g_{n,n-1} - c_{n,n-1}g_{n-2,n},\\
m_{n-3,n-2} &= -c_{n,n-2}g_{n-3,n},\\
m_{n-3,n-1} &= -c_{n,n-1}g_{n-3,n},\\
m_{n,n} &= -(m_{n-2,n-2} + m_{n-1,n-1}).
\end{aligned}
\tag{A.6}
$$

Acknowledgments. This work was completed while both authors were visiting the Centre for Advanced Studies, Oslo, Norway. They both wish to thank the staff and the members for making their stay so pleasant and productive. Thanks also to Brad Baxter, Stein Krogstad, Hans Z. Munthe-Kaas, and Beresford Parlett for useful discussion and suggestions.

REFERENCES

[1] P. Benner, H. Fassbender, and D. S. Watkins, *SR and SZ algorithms for the symplectic (butterfly) eigenproblem*, Linear Algebra Appl., 287 (1999), pp. 41–76.

[2] E. Celledoni and A. Iserles, *Approximating the exponential from a Lie algebra to a Lie group*, Math. Comp., 69 (2000), pp. 1457–1480.

[3] E. Celledoni and A. Iserles, *Methods for the approximation of the matrix exponential in a Lie-algebraic setting*, IMA J. Numer. Anal., 21 (2001), pp. 463–488.

[4] H. Fassbender, *The parameterized SR algorithm for symplectic (butterfly) matrices*, Math. Comp., 70 (2001), pp. 1515–1541.

[5] E. Gallopoulos and Y. Saad, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1236–1264.

[6] G. H. Golub and C. F. van Loan, *Matrix Computations*, 2nd ed., John Hopkins University Press, Baltimore, 1989.

[7] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, New York, 1978.

[8] M. Hochbruck, C. Lubich, and H. Selhofer, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[10] A. Iserles, H. Munthe-Kaas, S. P. Nørsett, and A. Zanna, *Lie-group methods*, Acta Numerica, 9 (2000), pp. 215–365.

[11] F. Kang and Z.-J. Shang, *Volume-preserving algorithms for source-free dynamical systems*, Numer. Math., 71 (1995), pp. 451–463.

[12] J. D. Lawson, *Polar and Ol'shanskii decompositions*, J. Reine Angew. Math., 448 (1994), pp. 191–219.

[13] F. S. Leite and P. Crouch, *Closed forms for the exponential mapping on matrix Lie groups based on Putzer's method*, J. Math. Phys., 40 (1999), pp. 3561–3568.

[14] D. Lewis and P. J. Olver, *Geometric integration algorithms on homogeneous manifolds*, Found. Comp. Math., 2 (2002), pp. 363–392.

[15] C. B. Moler and C. F. van Loan, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836.

[16] H. Munthe-Kaas, R. G. W. Quispel, and A. Zanna, *Generalized polar decompositions on Lie groups with involutive automorphisms*, Found. Comp. Math., 1 (2001), pp. 297–324.

[17] P. J. Olver, *Applications of Lie Groups to Differential Equations*, Grad. Texts in Math. 107, 2nd ed., Springer-Verlag, New York, 1993.
[18] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.
[19] A. Zanna, *Error Analysis for Exponential Splittings Based on Generalized Polar Decompositions* I*: Local and Global Bounds*, Tech. Report in Informatics 220, University of Bergen, Norway, 2001.
[20] A. Zanna, *Recurrence Relation for the Factors in the Polar Decomposition on Lie Groups*, Tech. Report no. 192, Department of Informatics, University of Bergen, Norway, May 2000, to appear in Math. Comp.
[21] A. Zanna and H. Z. Munthe-Kaas, *Generalized polar decompositions for the approximation of the matrix exponential*, SIAM J. Matrix Anal. Appl., 23 (2001/2002), pp. 840–862.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2257–2276

NUMERICAL ANALYSIS OF STOCHASTIC SCHEMES IN GEOPHYSICS*

BRIAN D. EWALD† AND ROGER TÉMAM‡

Abstract. We present and study the stability, convergence, and order of convergence of a numerical scheme used in geophysics, namely, the stochastic version of a deterministic "implicit leapfrog" scheme which has been developed for the approximation of the so-called barotropic vorticity model. Two other schemes which might be useful in the context of geophysical applications are also introduced and discussed.

Key words. numerical methods, stochastic differential equations, leapfrog scheme, Adams–Bashforth scheme, geophysical fluid dynamics

AMS subject classifications. 65C30, 60H35, 65N12, 86A10

DOI. 10.1137/S0036142902418333

1. Introduction.

Much effort has been invested in studying numerical schemes for stochastic differential equations of the form

$$dU_t = a(U_t)\,dt + b(U_t)\,dW_t, \tag{1.1}$$

where $U_t \in \mathbb{R}^d$, a is a function from $\mathbb{R}^d$ into itself, W is a Wiener process on $\mathbb{R}^m$, and b is a function from $\mathbb{R}^d$ into $\mathbb{R}^{d\times m}$.

For the so-called weak approximation of (1.1), in which the approximation of the expectation of functions of U is considered, extensive work is due, for example, to Talay and his collaborators, work relying on probabilistic methods more involved than those used in this article (see, e.g., [1], [2], [12] and the references therein).

The question of strong approximation of (1.1), in which the approximation of sample paths of U is desired, has also been much studied. Mil'shtein, in [8], introduced the scheme

$$\begin{aligned} U_{n+1}^k = U_n^k &+ \sum_{j=1}^{d} b^{k,j}(U_n)\Delta W_n^j + a^k(U_n)\Delta t \\ &+ \sum_{j_1,j_2,\ell=1}^{d} b^{\ell,j_1}(U_n)\frac{\partial b^{k,j_2}}{\partial x^\ell}(U_n)\int_{t_n}^{t_{n+1}} (W_s^{j_1} - W_{t_n}^{j_1})\,dW_s^{j_2}, \end{aligned} \tag{1.2}$$

which converges to U to the order of Δt in mean-square error. His method involved the consideration of a functional analytic Taylor series for the infinitesimal generator of a semigroup corresponding to U and W. Rümelin later investigated a stochastic analogue of Runge–Kutta (RK) schemes in [10], in which he compared them to Mil'shtein's scheme. The RK schemes which he derives can be arranged to converge

*Received by the editors November 20, 2002; accepted for publication (in revised form) September 25, 2003; published electronically March 31, 2005. This work was supported in part by a grant from the National Science Foundation (NSF-DMS 0074334) and by the Research Fund of Indiana University.

http://www.siam.org/journals/sinum/42-6/41833.html

†Department of Mathematics, Texas A&M University, College Station, TX 77843 (ewald@math.tamu.edu).

‡Department of Mathematics, Institute for Scientific Computing and Applied Mathematics, Indiana University, Bloomington, IN 47405 (temam@indiana.edu).

to U when the stochastic integral is interpreted in the sense of Itô, Stratonovich, or in fact for any stochastic calculus whatsoever. However, the issue of the accuracy of these RK schemes (which is not the same as in the deterministic case) is not fully addressed in [10] and is mostly unresolved. In fact, there are indications that these stochastic RK schemes are of significantly lower orders of accuracy than their deterministic counterparts (see [3]).

The book by Kloeden and Platen [6] and the companion volume by Kloeden, Platen, and Schurz [7] offer a systematic investigation of numerical schemes for (1.1) in both the sense of Itô and of Stratonovich, the two stochastic calculi which in applications are by far the most useful. Their methods are analytic and are applicable to proving the convergence of a wide range of numerical schemes, and they derive a very general scheme (formula (12.6.2) of [6]) which, for various choices of parameters, includes stochastic analogues of such deterministic schemes as the explicit and implicit Euler schemes, the Crank–Nicholson scheme, and the leapfrog scheme.

In the geophysics community, an enormous amount of work has been spent in developing large, complex numerical models of the oceans and atmosphere. The questions therefore arise: Is it possible to add stochastic numerical noise to these already existing models in such a way that it is known to what the scheme converges (e.g., to the Itô or Stratonovich solution of some stochastic differential equation), to what order they may be expected to converge, etc.? While we certainly do not answer these complex questions here, we consider a simple "implicit leapfrog" scheme for a barotropic model (supplied to us by Cecile Penland and Prashant Sardeshmukh) and demonstrate one way of adding stochastic noise to it so that these questions can be answered for the resulting stochastic scheme (section 4).

This scheme and the scheme in section 5 have been applied in investigating El Niño (see [4]). In this paper, the schemes were used for the numerical timestepping to determine if a linear inverse model of El Niño (see [9]) can be reconciled with the observed skew toward warm events in the Pacific. It was found that the observed skew is well within the range predicted by the model, although the observed trend is not.

We also propose a stochastic analogue for the deterministic Adams–Bashforth scheme, using methods similar to those of [6], as an attempt to produce alternate schemes which are higher order in time (studied in section 3, following the preliminary results in section 2).

Last, we examine the derivatives of a and b which occur naturally in the above schemes, and which can prove to be troublesome in certain applications in which these functions, especially b, are given by physical parametrizations (i.e., by "tables") and not by analytic expressions. We consider how these derivatives can be replaced by finite differences derived from space-discretization while still maintaining the existing rate of convergence (section 5).

We realize that the results of this article, while very useful in our opinion, are just some small contribution to an outstanding problem, namely, the numerical analysis of stochastic differential equations which raise—with more difficulty—the same issues as in the deterministic case: consistency, convergence, and accuracy. All of these issues—partly due to the form of the stochastic Taylor formula—are considerably more difficult than in the deterministic case; in particular, consistency includes here the issue of the type of stochastic calculus (Itô, Stratonovich, or otherwise) to which the scheme converges.

In the case of the geoscience scheme, the scheme that we study in section 4 is the closest we could get, at this time, to a scheme actually used in the geo-

sciences, without any prior information on its consistency, convergence, and accuracy properties. The version of the Adams–Bashforth scheme studied in section 3 has given very good numerical results in simulations for simple (one-dimensional) stochastic differential equations; see section 6 and forthcoming articles. The numerical example in section 6 is actually based on a mistake: one of the stochastic processes involved in the scheme was mistakenly believed to be (and treated as) a Gaussian process; this did not affect the accuracy of the scheme, which remains at order two as predicted by the theory. This raises perhaps an interesting probabilistic problem about the approximation of certain non-Gaussian processes used in numerical schemes. Another issue of probabilistic nature is raised by the Adams–Bashforth-type scheme studied in section 3: in several or in high dimensions, a large number of stochastic processes need to be simulated, which could make the cost prohibitive. It is not excluded that future probabilistic developments will improve this situation. In particular, the first author, using some ideas of Gaines and Lyons [5], is trying at this time to develop methods of generating the needed stochastic increments.

As we have said, there are, of course, a great many mathematical difficulties which this paper does not address. However, methods involving stochastic noise are already in common use in numerical simulations for the geosciences and turbulence (and, no doubt, many other areas in science). As mathematicians, we can attempt to help these scientists develop the necessary numerical tools, or watch as they do it themselves.

2. Preliminary results. We consider a stochastic differential equation

$$dU_t = a(t, U_t)dt + b(t, U_t)dW_t \tag{2.1}$$

for $U = (u_1, \dots, u_d) \in \mathbb{R}^d$, where $a : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$, $b : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$, and W is a Wiener process in $\mathbb{R}^m$ adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$.

We then have the Itô formula, which states that if $F : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^{\hat{d}}$, then $F_t = F(t, U_t)$ satisfies the stochastic differential equation

$$dF_t = \left[\frac{\partial F}{\partial t} + a^k(F_t)\frac{\partial F}{\partial u^k} + \frac{1}{2}b^{ij}(F_t)b^{kj}(F_t)\frac{\partial^2 F}{\partial u^i \partial u^k}\right]dt + b^{ij}(F_t)\frac{\partial F}{\partial u^i}dW_t^j; \tag{2.2}$$

here we use the Einstein convention for repeated indices.

We use the following notation from [6]: We call a row vector $\alpha = (j_1, j_2, \dots, j_l)$, where each $j_i \in \{0, 1, \dots, m\}$, a multi-index of length $l = \ell(\alpha) \in \{1, 2, \dots\}$. We also use ν to denote the multi-index of length 0, i.e., $\ell(\nu) = 0$. We define $n(\alpha)$ to be the number of entries of α which are 0. For adapted, right-continuous functions f, and stopping times ρ, τ such that $0 \leq \rho \leq \tau \leq T$ almost surely, we define

$$I_\alpha[f(\cdot)]_{\rho,\tau} = \begin{cases} f(\tau) & \text{if } \ell(\alpha) = 0, \\ \int_\rho^\tau I_{\alpha-}[f(\cdot)]_{\rho,s}\, ds & \text{if } \ell(\alpha) \geq 1,\ j_{\ell(\alpha)} = 0, \\ \int_p^\tau I_{\alpha-}[f(\cdot)]_{\rho,s}\, dW_s^{j_{\ell(\alpha)}} & \text{if } \ell(\alpha) \geq 1,\ j_{\ell(\alpha)} \neq 0. \end{cases} \tag{2.3}$$

Here $\alpha-$ is α with its final component removed.

We define the spaces $\mathcal{H}_\alpha$ as follows.

First, $\mathcal{H}_\nu$ is the space of adapted right-continuous stochastic processes f with left limits such that $|f(t)|$ is almost surely finite for each $t \geq 0$. Next, $\mathcal{H}_{(0)}$ contains those elements of $\mathcal{H}_\nu$ such that

$$\int_0^t |f(s)|\, ds < \infty \tag{2.4}$$

almost surely for each $t \geq 0$; and $\mathcal{H}_{(j)}$ for $j \neq 0$ contains those elements of $\mathcal{H}_\nu$ such that

$$\int_0^t |f(s)|^2 \, ds < \infty \tag{2.5}$$

almost surely for each $t \geq 0$. Finally, if $\ell(\alpha) \geq 2$, we define $\mathcal{H}_\alpha$ recursively as those elements of $\mathcal{H}_\nu$ that satisfy

$$I_{\alpha-}[f(\cdot)]_{0,t} \in \mathcal{H}_{(j_{\ell(\alpha)})} \tag{2.6}$$

almost surely for all $t \geq 0$.

We also define the operators

$$L^0 = \frac{\partial}{\partial t} + a^k \frac{\partial}{\partial u^k} + \frac{1}{2} b^{kj} b^{lj} \frac{\partial^2}{\partial u^k \partial u^l}, \tag{2.7}$$

$$L^j = b^{kj} \frac{\partial}{\partial u^k}, \tag{2.8}$$

and, if $f \in C^h(\mathbb{R}^+ \times \mathbb{R}^d, \mathbb{R})$, where $h \geq \ell(\alpha) + n(\alpha)$, we set

$$f_\alpha = \begin{cases} f & \text{if } \ell(\alpha) = 0, \\ L^{j_1} f_{-\alpha} & \text{if } \ell(\alpha) \geq 1. \end{cases} \tag{2.9}$$

Here $-\alpha$ is α with its first component removed.

We note that if $f(t,u) \equiv u$, then $f_{(0)} = a, f_{(j)} = b^j$, etc. In what follows, unless explicitly stated otherwise, we will assume that f is this identity function.

A set, $\mathcal{A}$, of multi-indices is said to be a hierarchical set if $\mathcal{A} \neq \emptyset$, $\sup_{\alpha \in \mathcal{A}} \ell(\alpha) < \infty$, and $-\alpha \in \mathcal{A}$ whenever $\alpha \in \mathcal{A} - \{\nu\}$. We then define the remainder set $\mathcal{B}(\mathcal{A})$ of $\mathcal{A}$ by $\mathcal{B}(\mathcal{A}) = \{\alpha \mid \alpha \notin \mathcal{A} \text{ and } -\alpha \in \mathcal{A}\}$. We can now provide a stochastic Taylor expansion for U satisfying (2.1): If $f : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}$, then, provided the derivatives and integrals exist,

$$f(\tau, U_\tau) = \sum_{\alpha \in \mathcal{A}} I_\alpha[f_\alpha(\rho, U_\rho)]_{\rho,\tau} + \sum_{\alpha \in \mathcal{B}(\mathcal{A})} I_\alpha[f_\alpha(\cdot, U_\cdot)]_{\rho,\tau}, \tag{2.10}$$

where $\mathcal{A}$ is some hierarchical set.

Now, for $\gamma = 0.5, 1.0, 1.5, \ldots$, we set

$$\mathcal{A}_\gamma = \left\{ \alpha \,\middle|\, \ell(\alpha) + n(\alpha) \leq 2\gamma \text{ or } \ell(\alpha) = n(\alpha) = \gamma + \frac{1}{2} \right\}. \tag{2.11}$$

We call the stochastic Taylor expansion with $\mathcal{A} = \mathcal{A}_\gamma$ the stochastic Taylor expansion to order γ.

We will make use of the following lemmas in the succeeding sections. In each of them, U is the solution to (2.1), and $t_k = k\Delta t$ for $k = 0, 1, \ldots, N$ is an equipartition of $[0,T]$, so that $t_N = T$; we partly rely on [6] for the proofs.

LEMMA 2.1. *Suppose Y_n is a stochastic process adapted to the filtration $\mathcal{F}_t$ at the equipartition (i.e., Y_n is $\mathcal{F}_{t_n}$-measurable), the function f satisfies $|f(t,x) - f(t,y)| \leq K|x-y|$ for all $t \in [0,T]$ and $x, y \in \mathbb{R}$, and α is a multi-index with $\ell(\alpha) \geq 1$. Then*

$$\begin{aligned} &\mathbb{E} \sup_{0 \leq m \leq n} \left| \sum_{k=0}^{m-1} I_\alpha[f(t_k, U_{t_k}) - f(t_k, Y_k)]_{t_k, t_{k+1}} \right|^2 \\ &\qquad \leq C\Delta t \sum_{k=0}^{n-1} \mathbb{E} \sup_{0 \leq m \leq k} |U_{t_m} - Y_m|^2. \end{aligned} \tag{2.12}$$

Proof. For $\alpha = (0)$, we have

$$\begin{aligned}
&\mathbb{E} \sup_{0\le m\le n} \left| \sum_{k=0}^{m-1} (f(t_k, U_{t_k}) - f(t_k, Y_k))\Delta t \right|^2 \\
&\quad \le \Delta t^2 \mathbb{E} \sup_{0\le m\le n} m \sum_{k=0}^{m-1} |f(t_k, U_{t_k}) - f(t_k, Y_k)|^2 \\
&\quad \le n\Delta t^2 \mathbb{E} \sum_{k=0}^{n-1} K^2 |U_{t_k} - Y_k|^2 \\
&\quad \le K^2 T \Delta t \sum_{k=0}^{n-1} \mathbb{E} \sup_{0\le m\le k} |U_{t_m} - Y_m|^2.
\end{aligned} \tag{2.13}$$

For $\alpha = (j)$,

$$\begin{aligned}
&\mathbb{E} \sup_{0\le m\le n} \left| \sum_{k=0}^{m-1} (f(t_k, U_{t_k}) - f(t_k, Y_k))\Delta W_k^j \right|^2 \\
&\quad \le 4\mathbb{E} \left| \sum_{k=0}^{n-1} (f(t_k, U_{t_k}) - f(t_k, Y_k))\Delta W_k^j \right|^2 \\
&\quad \le 4 \sum_{k=0}^{n-1} \mathbb{E} |f(t_k, U_{t_k}) - f(t_k, Y_k)|^2 \Delta t \\
&\quad \le 4K^2 \Delta t \sum_{k=0}^{n-1} \mathbb{E} \sup_{0\le m\le k} |U_{t_k} - Y_k|^2.
\end{aligned} \tag{2.14}$$

For longer α's, we just repeat the above two arguments as necessary. ∎

LEMMA 2.2. *Suppose the function f satisfies $|f(t,x)|^2 \le K^2(1+|x|^2)$ for all $t \in [0,T]$ and $x \in \mathbb{R}$, and that α is a multi-index with $\ell(\alpha) \ge 1$. Then*

$$\mathbb{E} \sup_{0\le m\le n} \left| \sum_{k=0}^{m-1} I_\alpha[f(\cdot, U_\cdot)]_{t_k, t_{k+1}} \right|^2 \le \begin{cases} C\Delta t^{2(\ell(\alpha)-1)}(1 + \mathbb{E}|U_0|^2) & \text{if } \ell(\alpha) = n(\alpha), \\ C\Delta t^{\ell(\alpha)+n(\alpha)-1}(1 + \mathbb{E}|U_0|^2) & \text{if } \ell(\alpha) \ne n(\alpha). \end{cases} \tag{2.15}$$

Proof. The ideas of this proof are the same as those in the proof of Lemma 2.1, along with the following bound on the solution U_t (see equation (4.5.16) of [6]):

$$\mathbb{E} \sup_{t_0\le s\le T} |U_s|^2 \le C(1 + \mathbb{E}|U_{t_0}|^2). \tag{2.16}$$

If we apply Lemma 10.8.1 of [6] with $g(s) = f(s, U_s)$, we have

$$\begin{aligned}
&\mathbb{E} \sup_{0\le m\le n} \left| \sum_{k=0}^{m-1} I_\alpha[f(\cdot, U_\cdot)]_{t_k, t_{k+1}} \right|^2 \\
&\quad \le \begin{cases} C\Delta t^{2(\ell(\alpha)-1)} \displaystyle\int_{t_0}^{t_n} \mathbb{E} \sup_{t_0\le s\le t} |f(s, U_s)|^2 \, dt & \text{if } \ell(\alpha) = n(\alpha), \\ C\Delta t^{\ell(\alpha)+n(\alpha)-1} \displaystyle\int_{t_0}^{t_n} \mathbb{E} \sup_{t_0\le s\le t} |f(s, U_s)|^2 \, dt & \text{if } \ell(\alpha) \ne n(\alpha). \end{cases}
\end{aligned} \tag{2.17}$$

Here, the constant C depends only on the length of the time interval $T - t_0$ and on α.

We then apply (2.16) and the growth condition on f, and we have the desired result. ∎

LEMMA 2.3. *Suppose that the sequence of positive numbers Z_n for $n = 0, 1, \ldots, N$ satisfies the inequality*

$$Z_n \le C\left(\Delta t \sum_{k=0}^{n} Z_k + \Delta t^{\gamma}\right) \tag{2.18}$$

for some positive constant C and some $\gamma > 0$. Then $Z_N = O(\Delta t^{\gamma})$ as $\Delta t \to 0$.

Proof. Set $\xi_n = \Delta t \sum_{k=0}^{n} Z_k$, so $Z_n = \frac{1}{\Delta t}(\xi_n - \xi_{n-1})$, and we have

$$\frac{1}{\Delta t}(\xi_n - \xi_{n-1}) \le C\xi_n + C\Delta t^{\gamma}. \tag{2.19}$$

That is,

$$(1 - C\Delta t)\xi_n \le \xi_{n-1} + C\Delta t^{\gamma+1}. \tag{2.20}$$

Therefore,

$$\begin{aligned}
(1 - C\Delta t)^n \xi_n &\le (1 - C\Delta t)^{n-1}\xi_{n-1} + (1 - C\Delta t)^{n-1} C\Delta t^{\gamma+1},\\
(1 - C\Delta t)^{n-1}\xi_{n-1} &\le (1 - C\Delta t)^{n-2}\xi_{n-2} + (1 - C\Delta t)^{n-2} C\Delta t^{\gamma+1},\\
&\vdots\\
(1 - C\Delta t)\xi_1 &\le C\Delta t^{\gamma+1},
\end{aligned}$$

and, summing,

$$\begin{aligned}
(1 - C\Delta t)^n \xi_n &\le (1 + (1 - C\Delta t) + \cdots + (1 - C\Delta t)^{n-1}) C\Delta t^{\gamma+1}\\
&\le \frac{1 - (1 - C\Delta t)^n}{1 - (1 - C\Delta t)} C\Delta t^{\gamma+1}\\
&\le (1 - (1 - C\Delta t)^n)\Delta t^{\gamma}.
\end{aligned} \tag{2.21}$$

Since $(1 - C\Delta t)^N \to e^{-CT}$ as $N = T/\Delta t \to \infty$, we see that $\xi_n \le C\Delta t^{\gamma}$ for some (different) C. Thus, by (2.18),

$$Z_n \le C(\xi_n + \Delta t^{\gamma}) \le C\Delta t^{\gamma}. \qquad \blacksquare \tag{2.22}$$

3. A stochastic Adams–Bashforth scheme. The deterministic Adams–Bashforth scheme for the ordinary differential equation $\phi' = F(\phi)$ takes the form

$$\phi_{n+1} = \phi_n + \frac{\Delta t}{2}[3F(\phi_n) - F(\phi_{n-1})]. \tag{3.1}$$

This scheme is order Δt^2. We will derive a stochastic version of this scheme which maintains the same order.

We begin with the stochastic Taylor expansion to order $\gamma = 2.0$:

$$\begin{aligned}
U_{t+\Delta} &= U_t + b^j \Delta W^j + a\Delta + L^{j_1} b^{j_2} I_{(j_1,j_2)} + L^0 b^j I_{(0,j)} + L^j a I_{(j,0)}\\
&\quad + L^{j_1} L^{j_2} b^{j_3} I_{(j_1,j_2,j_3)} + \frac{1}{2} L^0 a \Delta^2 + L^0 L^{j_1} b^{j_2} I_{(0,j_1,j_2)} + L^{j_1} L^0 b^{j_2} I_{(j_1,0,j_2)}\\
&\quad + L^{j_1} L^{j_2} a I_{(j_1,j_2,0)} + L^{j_1} L^{j_2} L^{j_3} b^{j_4} I_{(j_1,j_2,j_3,j_4)} + \tilde{R}^{\Delta}_{2.0}(t)\\
&= U_t + a\Delta + \frac{1}{2} L^0 a \Delta^2 + M^{\Delta}(t),
\end{aligned} \tag{3.2}$$

where each coefficient is at the point (t, U_t), and each stochastic integral is from t to $t+\Delta$, $\Delta = \Delta t$. We have also used the Einstein summation convention.

Similarly, for $\gamma = 1.5$, we have

$$a(t+\Delta, U_{t+\Delta}) = a + L^0 a\Delta + N^\Delta(t), \tag{3.3}$$

where $N^\Delta(t) = \frac{1}{2}L^0L^0a\Delta^2 + L^j a\Delta W^j + L^0L^j aI_{(0,j)} + L^jL^0aI_{(j,0)} + L^{j_1}L^{j_2}aI_{(j_1,j_2)} + L^{j_1}L^{j_2}L^{j_3}aI_{(j_1,j_2,j_3)} + \tilde{R}^\Delta_{1.5}(t)$, and, for $\gamma = 1.0$,

$$L^0a(t+\Delta, U_{t+\Delta}) = L^0a + P^\Delta(t), \tag{3.4}$$

where $P^\Delta(t) = L^0L^0a\Delta + L^jL^0a\Delta W^j + L^{j_1}L^{j_2}L^0aI_{(j_1,j_2)} + \tilde{R}^\Delta_{1.0}(t)$.

Combining these results, we get

$$\begin{aligned} U_{t+\Delta} = U_t &+ [\alpha a(t+\Delta, U_{t+\Delta}) + (1-\alpha)a]\Delta \\ &+ \left(\frac{1}{2} - \alpha\right)[\beta L^0a(t+\Delta, U_{t+\Delta}) + (1-\beta)L^0a]\Delta^2 \\ &- \alpha\Delta N^\Delta(t) - \left(\frac{1}{2} - \alpha\right)\beta\Delta^2 P^\Delta(t) + M^\Delta(t). \end{aligned} \tag{3.5}$$

In particular, if $t = t_n, \Delta = 2\Delta t, \alpha = 0, \beta = 0$, and writing $U_n = U_{t_n}$,

$$U_{n+2} = U_n + 2a(t_n, U_n)\Delta t + 2L^0a(t_n, U_n)\Delta t^2 + M^{2\Delta t}(t_n), \tag{3.6}$$

and if $t = t_n, \Delta = \Delta t, \alpha = -\frac{3}{2}, \beta = 0$,

$$\begin{aligned} U_{n+1} = U_n &- \frac{3}{2}a(t_{n+1}, U_{n+1})\Delta t + \frac{5}{2}a(t_n, U_n)\Delta t \\ &+ 2L^0a(t_n, U_n)\Delta t^2 + \frac{3}{2}N^{\Delta t}(t_n)\Delta t + M^{\Delta t}(t_n). \end{aligned} \tag{3.7}$$

Therefore,

$$\begin{aligned} U_{n+2} &= U_{n+1} + (U_{n+2} - U_n) - (U_{n+1} - U_n) \\ &= U_{n+1} + \left[\frac{3}{2}a(t_{n+1}, U_{n+1}) - \frac{1}{2}a(t_n, U_n)\right]\Delta t \\ &\quad - \frac{3}{2}\Delta t N^{\Delta t}(t_n) + (M^{2\Delta t}(t_n) - M^{\Delta t}(t_n)). \end{aligned} \tag{3.8}$$

This leads us to consider the following stochastic Adams–Bashforth (SAB) scheme:

$$\begin{aligned} Y_{n+2} = Y_{n+1} &+ \left[\frac{3}{2}a(t_{n+1}, Y_{n+1}) - \frac{1}{2}a(t_n, Y_n)\right]\Delta t \\ &- \frac{3}{2}\Delta t A_n(t_n, Y_n) + B_n(t_n, Y_n), \end{aligned} \tag{3.9}$$

in which

$$A_n(t,x) = L^j a(t,x)\Delta W^j + L^{j_1}L^{j_2}a(t,x)I_{(j_1,j_2)}, \tag{3.10}$$

where the random intervals are from time t_n to t_{n+1}, and

$$\begin{aligned} B_n(t,x) = {} & b^j(t,x)\Delta W^j + L^0 b^j(t,x) I_{(0,j)} + L^j a(t,x) I_{(j,0)} \\ & + L^{j_1} b^{j_2}(t,x) I_{(j_1,j_2)} + L^0 L^{j_1} b^{j_2}(t,x) I_{(0,j_1,j_2)} \\ & + L^{j_1} L^0 b^{j_2}(t,x) I_{(j_1,0,j_2)} + L^{j_1} L^{j_2} a(t,x) I_{(j_1,j_2,0)} \\ & + L^{j_1} L^{j_2} b^{j_3}(t,x) I_{(j_1,j_2,j_3)} + L^{j_1} L^{j_2} L^{j_3} b^{j_4}(t,x) I_{(j_1,j_2,j_3,j_4)}, \end{aligned} \tag{3.11}$$

where the random intervals are those from time t_n to t_{n+2} minus those from time t_n to t_{n+1}.

We then have the following theorem.

THEOREM 3.1. *Suppose that the coefficient functions f_α satisfy*

$$|f_\alpha(t,x) - f_\alpha(t,y)| \le K|x-y| \tag{3.12}$$

for all $\alpha \in \mathcal{A}_{2.0}$, $t \in [0,T]$, and $x, y \in \mathbb{R}^d$;

$$f_{-\alpha} \in C^{1,2} \text{ and } f_\alpha \in \mathcal{H}_\alpha \tag{3.13}$$

for all $\alpha \in \mathcal{A}_{2.0} \cup \mathcal{B}(\mathcal{A}_{2.0})$; and

$$|f_\alpha(t,x)| \le K(1+|x|) \tag{3.14}$$

for all $\alpha \in \mathcal{A}_{2.0} \cup \mathcal{B}(\mathcal{A}_{2.0})$, $t \in [0,T]$, and $x \in \mathbb{R}^d$. Choose $\Delta t \le 1$ and set $N = T/\Delta t$, and define $t_n = n\Delta t$ for $n = 1, \dots, N$. Suppose that Y_0 is some (nonrandom) initial condition and that some appropriate numerical scheme is used to generate Y_1 such that $\mathbb{E}[|U_{t_1} - Y_1|^2 \mid \mathcal{F}_0]^{\frac{1}{2}} \le C\Delta t^2$. Then

$$\mathbb{E}\left[\sup_{0\le n\le N} |U_{t_n} - Y_n|^2 \mid \mathcal{F}_0\right]^{\frac{1}{2}} \le C\Delta t^2. \tag{3.15}$$

Proof. First, we note that

$$\begin{aligned} U_{n+2} = {} & U_{n+1} + \left[\frac{3}{2} a(t_{n+1} U_{n+1}) - \frac{1}{2} a(t_n, U_n)\right] \Delta t \\ & - \frac{3}{2} \Delta t A_n(t_n, U_n) + B_n(t_n, U_n) + R_n, \end{aligned} \tag{3.16}$$

where

$$\begin{aligned} R_n = {} & \frac{3}{2}\Delta t \left[\frac{1}{2} L^0 L^0 a(t_n, U_n)\Delta t^2 + L^0 L^j a(t_n, U_n) I_{(0,j)} + L^j L^0 a(t_n, U_n) I_{(j,0)} \right. \\ & \left. + L^{j_1} L^{j_2} L^{j_3} a(t_n, U_n) I_{(j_1,j_2,j_3)} + \tilde{R}^{\Delta t}_{1.5}(t_n)\right] + \tilde{R}^{2\Delta t}_{2.0}(t_n) - \tilde{R}^{\Delta t}_{2.0}(t_n). \end{aligned}$$

If we iterate (3.9) and (3.16), we arrive at

$$
\begin{cases}
U_n = U_1 + \Delta t \sum_{k=0}^{n-2} \left[\frac{3}{2} a(t_{k+1}, U_{k+1}) - \frac{1}{2} a(t_k, U_k) \right] \\
\qquad - \frac{3}{2} \Delta t \sum_{k=0}^{n-2} A_k(t_k, U_k) + \sum_{k=0}^{n-2} B_k(t_k, U_k) + \sum_{k=0}^{n-2} R_k, \\
Y_n = Y_1 + \Delta t \sum_{k=0}^{n-2} \left[\frac{3}{2} a(t_{k+1}, Y_{k+1}) - \frac{1}{2} a(t_k, Y_k) \right] \\
\qquad - \frac{3}{2} \Delta t \sum_{k=0}^{n-2} A_k(t_k, Y_k) + \sum_{k=0}^{n-2} B_k(t_k, Y_k).
\end{cases} \tag{3.17}
$$

Set $\zeta_n = U_n - Y_n$. Then

$$
\begin{aligned}
\zeta_n = \zeta_1 &+ \Delta t \sum_{k=0}^{n-2} \left[\frac{3}{2} (a(t_{k+1}, U_{k+1}) - a(t_{k+1}, Y_{k+1})) - \frac{1}{2} (a(t_k, U_k) - a(t_k, Y_k)) \right] \\
&- \frac{3}{2} \Delta t \sum_{k=0}^{n-2} (A_k(t_k, U_k) - A_k(t_k, Y_k)) \\
&+ \sum_{i=0}^{n-2} (B_k(t_k, U_k) - B_k(t_k, Y_k)) + \sum_{k=0}^{n-2} R_k.
\end{aligned} \tag{3.18}
$$

Set $Z_n = \mathbb{E}[\sup_{0 \le m \le n} |\zeta_m|^2 \,|\, \mathcal{F}_0]$.
Then we have

$$
\begin{aligned}
Z_n \le C \Bigg(& \mathbb{E}[|\zeta_1|^2 | \mathcal{F}_0] + \Delta t^2 \mathbb{E} \left[\sup_{0 \le m \le n} \left| \sum_{k=0}^{m-2} a(t_{k+1}, U_{k+1}) - a(t_{k+1}, Y_{k+1}) \right|^2 \,\middle|\, \mathcal{F}_0 \right] \\
& + \Delta t^2 \mathbb{E} \left[\sup_{0 \le m \le n} \left| \sum_{k=0}^{m-2} a(t_k, U_k) - a(t_k, Y_k) \right|^2 \,\middle|\, \mathcal{F}_0 \right] \\
& \left. \begin{aligned}
& + \Delta t^2 \mathbb{E} \left[\sup_{0 \le m \le n} \left| \sum_{k=0}^{m-2} (L^j a(t_k, U_k) - L^j a(t_k, Y_k)) \Delta W_k^j \right|^2 \,\middle|\, \mathcal{F}_0 \right] \\
& + \Delta t^2 \mathbb{E} \left[\sup_{0 \le m \le n} \left| \sum_{k=0}^{m-2} (L^{j1} L^{j2} a(t_k, U_k) - L^{j_1} L^{j_2} a(t_k, Y_k)) I_{(j_1, j_2)} \right|^2 \,\middle|\, \mathcal{F}_0 \right]
\end{aligned} \right\} \text{Terms from } A_k \\
& \left. \begin{aligned}
& + \sum_{\alpha \epsilon \mathcal{A}_2^*} \mathbb{E} \left[\sup_{0 \le m \le n} \left| \sum_{k=0}^{m-2} I_\alpha \left[f_\alpha(t_k, U_k) - f_\alpha(t_k, Y_k) \right]_{t_k, t_{k+2}} \right|^2 \,\middle|\, \mathcal{F}_0 \right] \\
& + \sum_{\alpha \epsilon \mathcal{A}_2^*} \mathbb{E} \left[\sup_{0 \le m \le n} \left| \sum_{k=0}^{m-2} I_\alpha \left[f_\alpha(t_k, U_k) - f_\alpha(t_k, Y_k) \right]_{t_k, t_{k+1}} \right|^2 \,\middle|\, \mathcal{F}_0 \right]
\end{aligned} \right\} \text{Terms from } B_k \\
& + \mathbb{E} \left[\sup_{0 \le m \le n} \left| \sum_{k=0}^{m-2} R_k \right|^2 \,\middle|\, \mathcal{F}_0 \right] \Bigg).
\end{aligned}
$$

We go term-by-term:

$$\begin{aligned}
&\Delta t^2\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2} a(t_{k+1},U_{k+1})-a(t_{k+1},Y_{k+1})\right|^2\ \middle|\ \mathcal{F}_0\right]\\
&\quad\le \Delta t^2 n\ \mathbb{E}\left[\sup_{0\le m\le n}\sum_{k=0}^{m-2}\left|a(t_{k+1},U_{k+1})-a(t_{k+1},Y_{k+1})\right|^2\ \middle|\ \mathcal{F}_0\right]\\
&\quad\le T\Delta t\ \mathbb{E}\left[\sum_{k=0}^{n-2}\left|a(t_{k+1},U_{k+1})-a(t_{k+1},Y_{k+1})\right|^2\ \middle|\ \mathcal{F}_0\right] \qquad (3.19)\\
&\quad\le C\Delta t\sum_{k=0}^{n-2}\mathbb{E}\left[|U_{k+1}-Y_{k+1}|^2\mid\mathcal{F}_0\right]\\
&\quad\le C\Delta t\sum_{k=0}^{n-1} Z_k.
\end{aligned}$$

Similarly,

$$\Delta t^2\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2} a(t_k,U_k)-a(t_k,Y_k)\right|^2\ \middle|\ \mathcal{F}_0\right]\le C\Delta t\sum_{k=0}^{n-1} Z_k. \qquad (3.20)$$

Next, we consider the terms from A_k. From Lemma 2.1 with $\alpha=(j)$,

$$\begin{aligned}
&\Delta t^2\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2}(L^j a(t_k,U_k)-L^j a(t_k,Y_k))\Delta W^j_{t_k,t_{k+1}}\right|^2\ \middle|\ \mathcal{F}_0\right] \qquad (3.21)\\
&\quad\le C\Delta t\sum_{k=0}^{n-2} Z_k,
\end{aligned}$$

and from Lemma 2.1 with $\alpha=(j_1,j_2)$,

$$\begin{aligned}
&\Delta t^2\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2}(L^{j1}L^{j2}a(t_k,U_k)-L^{j_1}L^{j_2}a(t_k,Y_k))I_{(j_1,j_2)t_k,t_{k+1}}\right|^2\ \middle|\ \mathcal{F}_0\right] \qquad (3.22)\\
&\quad\le C\Delta t\sum_{k=0}^{n-1} Z_k.
\end{aligned}$$

Now, we consider the terms from B_k. For $\alpha\in\mathcal{A}_2^*$ (i.e., $\alpha\in\mathcal{A}_2,\ell(\alpha)\ne n(\alpha)$),

$$\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2} I_\alpha\left[f_\alpha(t_k,U_k)-f_\alpha(t_k,Y_k)\right]_{t_k,t_{k+1}}\right|^2\ \middle|\ \mathcal{F}_0\right]\le C\Delta t\sum_{k=0}^{n-1} Z_k. \qquad (3.23)$$

The other terms from B_k are similar. This leaves only the terms from R_k. From Lemma 2.2 with $\alpha=(0)$,

$$\begin{aligned}
&\Delta t^2\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2} L^0L^0a(t_n,U_n)\Delta t^2\right|^2\ \middle|\ \mathcal{F}_0\right]\\
&\quad\le \Delta t^4\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{n-2} L^0L^0a(t_n,U_n)\right|^2\ \middle|\ \mathcal{F}_0\right] \qquad (3.24)\\
&\quad\le C\Delta t^4(1+|U_0|^2).
\end{aligned}$$

For $\alpha = (0,j),(j,0),(j_1,j_2,j_3)$, by Lemma 2.2

$$\Delta t^2 \mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2} I_\alpha\left[a_\alpha(t_k,U_k)\right]_{t_k,t_{k+1}}\right|^2 \,\middle|\, \mathcal{F}_0\right] \le C\Delta t^4(1+|U_0|^2), \text{ since } \ell(\alpha)+n(\alpha)=3. \tag{3.25}$$

If $\alpha \in \mathcal{B}(\mathcal{A}\gamma)$ (here $\gamma = 1.5$ or 2.0), we have by Lemma 2.2,

$$\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2} I_\alpha\left[\alpha(\cdot,U_\cdot)\right]_{t_k,t_{k+1}}\right|^2 \,\middle|\, \mathcal{F}_0\right] \le C(1+|U_0|^2)\Delta t^{2\gamma}. \tag{3.26}$$

Therefore, we have

$$\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-2} R_k\right|^2 \,\middle|\, \mathcal{F}_0\right] \le C(1+|U_0|^2)\Delta t^4. \tag{3.27}$$

So, overall, we see that

$$Z_n \le C\left[Z_1 + (1+|U_0|^2)\Delta t^4 + \Delta t\sum_{k=0}^{n-1} Z_k\right]. \tag{3.28}$$

The result then follows from Lemma 2.3. □

Remark 3.1. If we truncate A_n and B_n to

$$A_n(t,x) = L^j a(t,x)\Delta W^j \tag{3.29}$$

and

$$\begin{aligned} B_n(t,x) = {} & b^j(t,x)\Delta W^j + L^0 b^j(t,x) I_{(0,j)} + L^j a(t,x) I_{(j,0)} \\ & + L^{j_1} b^{j_2}(t,x) I_{(j_1,j_2)} + L^{j_1} L^{j_2} b^{j_3}(t,x) I_{(j_1,j_2,j_3)}, \end{aligned} \tag{3.30}$$

the same proof will show that the convergence is now to order $\Delta t^{\frac{3}{2}}$. We note that although the order Δt^2 SAB scheme seems to have no obvious advantages over the standard Δt^2 strong one-step explicit scheme (as in [6]), the order $\Delta t^{\frac{3}{2}}$ SAB scheme does have an advantage over the order $\Delta t^{\frac{3}{2}}$ strong one-step explicit scheme in that the former lacks the terms involving the second derivative of a which are present in the latter.

Remark 3.2. It can be shown that the scheme

$$\begin{aligned} Y_{n+2} = {} & Y_{n+1} + \left[\frac{3}{2}a(t_{n+1},Y_{n+1}) - \frac{1}{2}a(t_n,Y_n)\right]\Delta t \\ & - \frac{3}{2}\Delta t A_n(t_n,Y_n) + B_n(t_n,Y_n), \end{aligned} \tag{3.31}$$

in which

$$A_n(t,x) = L^j a(t,x)\Delta W^j, \tag{3.32}$$

where the random intervals are from time t_n to t_{n+1}, and

$$\begin{aligned} B_n(t,x) = {} & b^j(t,x)\Delta W^j + L^0 b^j(t,x) I_{(0,j)} + L^j a(t,x) I_{(j,0)} \\ & + L^{j_1} b^{j_2}(t,x) I_{(j_1,j_2)}, \end{aligned} \tag{3.33}$$

where the random intervals are those from time t_n to t_{n+2} minus those from time t_n to t_{n+1}, converges to the Itô solution in the weak sense to order 2. As can be seen, this scheme is considerably simpler than the strong scheme, and it avoids the difficulties with generating the higher-order moments that the strong scheme has.

4. A stochastic "implicit leapfrog" scheme. The barotropic vorticity model supplied to us by Cecile Penland and Prashant Sardeshmukh of the National Oceanic and Atmospheric Administration in Boulder, Colorado (see [11]), takes the form

$$\frac{\partial \zeta}{\partial t} = -\nabla \cdot (v\zeta) + S - r\xi - \kappa \nabla^4 \xi, \tag{4.1}$$

where $\zeta = \nabla^2 \psi + f = \xi + f$ and $v = \hat{k} \times \nabla \psi$. Here, ζ is the total vorticity, v is the velocity vector, f is the Coriolis term, S is a (deterministic) forcing, r and κ are constants, and ξ is the local vorticity.

The numerical scheme they provided for this uses spherical harmonics, and, writing F for $-\nabla \cdot (v\zeta)$, the equation becomes

$$\frac{d}{dt}\zeta_n^m = F_n^m + S_n^m - r\xi_n^m - \kappa \left[\frac{n(n+1)}{a^2}\right]^2 \zeta_n^m. \tag{4.2}$$

Then the scheme has two steps. First, a leapfrog step,

$$\tilde{\zeta}_n^m(t + \Delta t) = \zeta_n^m(t - \Delta t) + 2\Delta t[F_n^m(t) + S_n^m(t)], \tag{4.3}$$

followed by an implicit step,

$$\zeta_n^m(t + \Delta t) = \frac{\tilde{\zeta}_n^m(t + \Delta t)}{1 + 2\Delta t \left[r + \kappa \left[\frac{n(n+1)}{a^2}\right]^2\right]}. \tag{4.4}$$

If we simplify notation and write a_1 for $F + S$ and a_2 for $-r\xi - \kappa\nabla^4\xi$, we see that this is just an "implicit leapfrog" scheme

$$\begin{cases} \tilde{Y}(t + \Delta t) = Y(t - \Delta t) + 2\Delta t a_1(t, Y(t)), \\ Y(t + \Delta t) = \tilde{Y}(t + \Delta t) + 2\Delta t a_2(t + \Delta t, Y(t + \Delta t)) \end{cases} \tag{4.5}$$

for the equation

$$dU(t) = [a_1(t, U(t)) + a_2(t, U(t))]\, dt. \tag{4.6}$$

Therefore, we consider a stochastic differential equation of the form

$$dU_t = (a_1(t, U_t) + a_2(t, U_t))\, dt + b(t, U_t)\, dW_t. \tag{4.7}$$

Note that we have simply added a general diffusion term to the deterministic differential equation (4.6).

We will consider the scheme

$$\begin{cases} \tilde{Y}_{n+2} = Y_n + 2a_1(t_{n+1}, Y_{n+1})\Delta t + M_n(Y_n) + M_{n+1}(Y_{n+1}), \\ Y_{n+2} = \tilde{Y}_{n+2} + 2a_2(t_{n+2}, Y_{n+2})\Delta t, \end{cases} \tag{4.8}$$

where

$$M_n(y) = b(t_n, y)\Delta W_n + bb'(t_n, y) I_{(1,1),n}. \tag{4.9}$$

THEOREM 4.1. *Suppose that the coefficient functions f_α satisfy*

$$|f_\alpha(t,x) - f_\alpha(t,y)| \le K|x-y| \tag{4.10}$$

for all $\alpha \in \mathcal{A}_{1.0}$, $t \in [0,T]$, and $x, y \in \mathbb{R}^d$;

$$f_{-\alpha} \in C^{1,2} \quad \text{and} \quad f_\alpha \in \mathcal{H}_\alpha \tag{4.11}$$

for all $\alpha \in \mathcal{A}_{1.0} \cup \mathcal{B}(\mathcal{A}_{1.0})$; and

$$|f_\alpha(t,x)| \le K(1+|x|) \tag{4.12}$$

for all $\alpha \in \mathcal{A}_{1.0} \cup \mathcal{B}(\mathcal{A}_{1.0})$, $t \in [0,T]$, and $x \in \mathbb{R}^d$. Choose $\Delta t \le 1$ and set $N = T/\Delta t$, and define $t_n = n\Delta t$ for $n = 1, \dots, N$. Suppose that some appropriate numerical scheme is used to generate Y_1 such that $\mathbb{E}[|U_{t_1} - Y_1|^2 \mid \mathcal{F}_0]^{\frac{1}{2}} \le C\Delta t$. Then

$$\mathbb{E}\left[\sup_{0\le n\le N} |U_{t_n} - Y_n|^2 \mid \mathcal{F}_0\right]^{\frac{1}{2}} \le C\Delta t. \tag{4.13}$$

Proof. We note first that, by Itô's formula (i.e., the Taylor expansion with $\gamma = 0.0$), the solution U to (4.7) satisfies the following equations (where for notational simplicity we have written U_n for U_{t_n}):

$$\begin{aligned} a_1(t_{n+1}, U_{n+1}) &= a_1(t_n, U_n) + R_{0.0}^{\Delta t, a_1}(t_n); \\ a_2(t_{n+2}, U_{n+2}) &= a_2(t_n, U_n) + R_{0.0}^{2\Delta t, a_2}(t_n); \\ a_2(t_{n+2}, U_{n+2}) &= a_2(t_{n+1}, U_{n+1}) + R_{0.0}^{\Delta t, a_2}(t_{n+1}). \end{aligned} \tag{4.14}$$

Therefore we have

$$\begin{aligned} U_{n+2} &= U_n + (U_{n+2} - U_{n+1}) + (U_{n+1} - U_n) \\ &= U_n + [b(t_{n+1}, U_{n+1})\Delta W_{n+1} + a_1(t_{n+1}, U_{n+1})\Delta t \\ &\quad + a_2(t_{n+1}, U_{n+1})\Delta t + bb'(t_{n+1}, U_{n+1})I_{(1,1),n+1} \\ &\quad + R_{1.0}^{\Delta t}(t_{n+1})] + [b(t_n, U_n)\Delta W_n + a_1(t_n, U_n)\Delta t \\ &\quad + a_2(t_n, U_n)\Delta t + bb'(t_n, U_n)I_{(1,1),n} + R_{1.0}^{\Delta t}(t_n)]. \end{aligned} \tag{4.15}$$

After substituting (4.14) into this, we see that

$$\begin{aligned} U_{n+2} = [U_n + 2a_1(t_{n+1}, U_{n+1})\Delta t + M_n(U_n) + M_{n+1}(U_{n+1})] \\ + 2a_2(t_{n+2}, U_{n+2})\Delta t + R_n; \end{aligned} \tag{4.16}$$

here

$$\begin{aligned} R_n = R_{1.0}^{\Delta t}(t_n) + R_{1.0}^{\Delta t}(t_{n+1}) - \Delta t[R_{0.0}^{\Delta t, a_1}(t_n) \\ + R_{0.0}^{2\Delta t, a_2}(t_n) + R_{0.0}^{\Delta t, a_2}(t_{n+1})]. \end{aligned} \tag{4.17}$$

If we iterate (4.16), we arrive at

$$\begin{aligned} U_n = U_{n^*} + 2\Delta t\left[\sum_{k=1}^{[n/2]} a_1(t_{2k-1+n^*}, U_{2k-1+n^*}) + \sum_{k=1}^{[n/2]} a_2(t_{2k+n^*}, U_{2k+n^*})\right] \\ + \sum_{k=n^*}^{n-1} M_n(U_n) + \sum_{k=0}^{[n/2]} R_{2k-2+n^*}; \end{aligned} \tag{4.18}$$

here n^* is 0 if n is even and 1 if n is odd.

Similarly, we have for Y

$$Y_n = Y_{n^*} + 2\Delta t \left[\sum_{k=1}^{[n/2]} a_1(t_{2k-1+n^*}, Y_{2k-1+n^*}) + \sum_{k=1}^{[n/2]} a_2(t_{2k+n^*}, Y_{2k+n^*}) \right] + \sum_{k=n^*}^{n-1} M_n(Y_n). \tag{4.19}$$

Let us set $Z_n = \mathbb{E}[\sup_{0\le m\le n} |U_m - Y_m|^2 \mid \mathcal{F}_0]$. Then, by subtracting (4.19) from (4.18) and then squaring and taking expectations, we find

$$Z_n \le C\mathbb{E}\left[\sup_{0\le m\le n} Z_{m^*} + \Delta t^2 A_{1,m}^2 + \Delta t^2 A_{2,m}^2 + \left(\sum_{k=m^*}^{m-1} M_k(U_k) - M_k(Y_k) \right)^2 + \left(\sum_{k=0}^{[m/2]} R_{2k-2+m^*} \right)^2 \Bigg| \mathcal{F}_0 \right]. \tag{4.20}$$

In (4.20),

$$A_{1,n} = \sum_{k=1}^{[n/2]} a_1(t_{2k-1+n^*}, U_{2k-1+n^*}) - a_1(t_{2k-1+n^*}, Y_{2k-1+n^*}) \tag{4.21}$$

and

$$A_{2,n} = \sum_{k=1}^{[n/2]} a_2(t_{2k+n^*}, U_{2k+n^*}) - a_2(t_{2k+n^*}, Y_{2k+n^*}). \tag{4.22}$$

We then have the following estimates (where we omit the dependence on t when it is clear):

$$\begin{aligned} \Delta t^2 \mathbb{E}\left[\sup_{0\le m\le n} A_{1,m}^2 \right] &\le \Delta t^2 \mathbb{E}\left[\sup_{0\le m\le n} \left[\frac{m}{2}\right] \sum_{k=1}^{[m/2]} [a_1(U_{2k-1+n^*}) - a_1(Y_{2k-1+n^*})]^2 \right] \\ &\le K\Delta t^2 n \mathbb{E}\left[\sum_{k=1}^{n} (U_k - Y_k)^2 \right] \le KT\Delta t \sum_{k=1}^{n} Z_k. \end{aligned} \tag{4.23}$$

Similarly,

$$\Delta t^2 \mathbb{E}\left[\sup_{0\le m\le n} A_{2,n}^2 \right] \le KT\Delta t \sum_{k=1}^{n} Z_k. \tag{4.24}$$

There are two terms in M_k. For the first one, from Lemma 2.1, with $\alpha = (1)$, we obtain

$$\mathbb{E}\left[\sup_{0\le m\le n} \left| \sum_{k=m^*}^{m-1} [b(U_k) - b(Y_k)]\Delta W_k \right|^2 \right] \le C\Delta t \sum_{k=0}^{n} Z_k. \tag{4.25}$$

The second term is similar, with $\alpha = (1,1)$.

Finally, we show a representative term from the remainder R. From Lemma 2.2, with $\alpha = (1, 0)$,

$$\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}\int_{t_k}^{s} L^0 b(U_r)\,dr\,dW_s\right|^2\right] \le C\Delta t^2(1+U_0^2). \tag{4.26}$$

The remaining terms from R give similar bounds.

Taking all of these estimates into account, we have the inequality

$$Z_n \le K\left[Z_1 + \Delta t\sum_{k=0}^{n} Z_k + \Delta t^2\right]. \tag{4.27}$$

Therefore, from Lemma 2.3, we see that

$$\mathbb{E}\left[\sup_{o\le n\le N}|U_n - Y_n|^2 \,\middle|\, \mathcal{F}_0\right]^{\frac{1}{2}} = O(\Delta t). \qquad \square \tag{4.28}$$

Remark 4.1. It is possible to show that the scheme

$$\begin{cases} \tilde{Y}_{n+2} = Y_n + 2a_1(t_{n+1}, Y_{n+1})\Delta t + b(t_n, Y_n)\Delta W_n \\ \qquad\qquad + b(t_{n+1}, Y_{n+1})\Delta W_{n+1}, \\ Y_{n+2} = \tilde{Y}_{n+2} + 2a_2(t_{n+2}, Y_{n+2})\Delta t \end{cases} \tag{4.29}$$

converges to the Itô solution in the weak sense to order 1. Again, it can be seen that the weak scheme is simpler than the strong scheme to the same order. However, we also note that the weak order 1 Stratonovich scheme

$$\begin{cases} \tilde{Y}_{n+2} = Y_n + 2a_1(t_{n+1}, Y_{n+1})\Delta t + bb'(t_{n+1}, Y_{n+1})(\Delta W_{n+1})^2 \\ \qquad\qquad + b(t_n, Y_n)\Delta W_n + b(t_{n+1}, Y_{n+1})\Delta W_{n+1}, \\ Y_{n+2} = \tilde{Y}_{n+2} + 2a_2(t_{n+2}, Y_{n+2})\Delta t \end{cases} \tag{4.30}$$

is not appreciably simpler than the strong order 1 Stratonovich scheme

$$\begin{cases} \tilde{Y}_{n+2} = Y_n + 2a_1(t_{n+1}, Y_{n+1})\Delta t + M_n(Y_n) + M_{n+1}(Y_{n+1}), \\ Y_{n+2} = \tilde{Y}_{n+2} + 2a_2(t_{n+2}, Y_{n+2})\Delta t, \end{cases} \tag{4.31}$$

where

$$M_n(y) = b(t_n, y)\Delta W_n + \frac{1}{2}bb'(t_n, y)(\Delta W_n)^2. \tag{4.32}$$

5. Discretization of spatial derivatives by finite differences. It sometimes happens in applications that the functions a and b may only be known empirically (i.e., in tables) rather than analytically. In such cases, analytic derivatives of these functions can be difficult to obtain. It is therefore useful to replace these derivatives by discrete approximations. As a first example, consider this modification of Mil'shtein's scheme:

$$\begin{aligned} \hat{Y}_{n+1}^k &= \hat{Y}_n^k + \sum_{j=1}^{d} b^{k,j}(\hat{Y}_n)\Delta W_n^j + a^k(\hat{Y}_n)\Delta t \\ &\quad + \sum_{j_1,j_2,\ell=1}^{d} \frac{1}{\Delta x} b^{\ell,j_1}(\hat{Y}_n)(b^{k,j_2}(\hat{Y}_n + \Delta x e^\ell) - b^{k,j_2}(\hat{Y}_n))I_{(j_1,j_2),n}, \end{aligned} \tag{5.1}$$

where e^ℓ is the vector $(0,\dots,0,1,0,\dots,0)$, with 1 in the ℓth position, and we have chosen $\Delta x > 0$. We have also suppressed the dependence of a and b on time to simplify notation.

We then have the following theorem.

THEOREM 5.1. *Suppose that a and b have the regularity required for Mil'shtein's scheme to converge to the solution U to order Δt. Then*

$$\mathbb{E}\left[\sup_{0\le n\le N}|U_n-\hat{Y}_n|^2\right]^{\frac{1}{2}}=O(\max\{\Delta t,\Delta x\Delta t^{\frac{1}{2}}\}). \tag{5.2}$$

Note that if we want to maintain the order of convergence of Mil'shtein's scheme, we need that $\Delta x = O(\Delta t^{\frac{1}{2}})$.

Proof. We denote Mil'shtein's scheme by Y and recall that it satisfies

$$\mathbb{E}\left[\sup_{0\le n\le N}|Y_n-U_n|^2\right]^{\frac{1}{2}}=O(\Delta t). \tag{5.3}$$

First, we see that (using Einstein's summation convention on repeated indices)

$$\begin{aligned}\hat{Y}^k_{n+1}-Y^k_{n+1}&=\hat{Y}^k_n-Y^k_n+(b^{k,j}(\hat{Y}_n)-b^{k,j}(Y_n))\Delta W^j_n+(a^k(\hat{Y}_n)-a^k(Y_n))\Delta t\\&+\left[b^{\ell,j_1}(\hat{Y}_n)\frac{1}{\Delta x}(b^{k,j_2}(\hat{Y}_n+\Delta x e^\ell)-b^{k,j_2}(\hat{Y}_n))\right.\\&\left.-\ b^{\ell,j_1}(Y_n)\frac{\partial b^{k,j_2}}{\partial x^\ell}(Y_n)\right]I_{(j_1,j_2),n}.\end{aligned} \tag{5.4}$$

Iterating this, we have

$$\begin{aligned}\hat{Y}^k_n-Y^k_n&=\hat{Y}^k_0-Y^k_0+\sum_{i=0}^{n-1}(b^{k,j}(\hat{Y}_i)-b^{k,j}(Y_i))\Delta W^j_i+\sum_{i=0}^{n-1}(a^k(\hat{Y}_i)-a^k(Y_i))\Delta t\\&+\sum_{i=0}^{n-1}\left[b^{\ell,j_1}(\hat{Y}_i)\frac{1}{\Delta x}(b^{k,j_2}(\hat{Y}_i+\Delta x e^\ell)-b^{k,j_2}(\hat{Y}_i))\right.\\&\left.-\ b^{\ell,j_1}(Y_i)\frac{\partial b^{k,j_2}}{\partial x^\ell}(Y_i)\right]I_{(j_1,j_2),i}.\end{aligned} \tag{5.5}$$

Set

$$Z_n=\mathbb{E}\left[\sup_{0\le m\le n}|\hat{Y}_n-Y_n|^2\right]. \tag{5.6}$$

We then have the estimates

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{i=0}^{m-1}(b^{k,j}(\hat{Y}_i)-b^{k,j}(Y_i))\Delta W^j_i\right|^2\right]\\&\le 4\mathbb{E}\left[\left|\sum_{i=0}^{n-1}(b^{k,j}(\hat{Y}_i)-b^{k,j}(Y_i))\Delta W^j_i\right|^2\right]\\&\le 4\mathbb{E}\left[\sum_{i=0}^{n-1}|b^{k,j}(\hat{Y}_i)-b^{k,j}(Y_i)|^2\Delta t\right]\\&\le C\Delta t\sum_{i=0}^{n-1}Z_i,\end{aligned}$$

and

$$\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{i=0}^{m-1}(a^k(\hat{Y}_i)-a^k(Y_i))\Delta t\right|^2\right]\le \Delta t^2 n\mathbb{E}\left[\sum_{i=0}^{n-1}|a^k(\hat{Y}_i)-a^k(Y_i)|^2\right]$$
$$\le C\Delta t\sum_{i=0}^{n-1}Z_i,$$

and, finally,

$$\mathbb{E}\left[\sup_{0\le m\le n}\left|\sum_{i=0}^{m-1}\left[b^{\ell,j_1}(\hat{Y}_i)\frac{1}{\Delta x}[b^{k,j_2}(\hat{Y}_i+\Delta x e^\ell)-b^{k,j_2}(\hat{Y}_i)]\right.\right.\right.$$
$$\left.\left.\left.-\ b^{\ell,j_1}(Y_i)\frac{\partial b^{k,j_2}}{\partial x^\ell}(Y_i)\right]I_{(j_1,j_2),i}\right|^2\right]$$
$$\le C\Delta t^2\sum_{i=0}^{n-1}\mathbb{E}\left[b^{\ell,j_1}(\hat{Y}_i)\frac{1}{\Delta x}[b^{k,j_2}(\hat{Y}_i+\Delta x e^\ell)-b^{k,j_2}(\hat{Y}_i)]-b^{\ell,j_1}(Y_i)\frac{\partial b^{k,j_2}}{\partial x^\ell}(Y_i)\right]^2$$
$$\le C\Delta t^2\sum_{i=0}^{n-1}\mathbb{E}\left[b^{\ell,j_1}(\hat{Y}_i)\frac{\partial b^{k,j_2}}{\partial x^\ell}(\hat{Y}_i)-b^{\ell,j_1}(Y_i)\frac{\partial b^{k,j_2}}{\partial x^\ell}(Y_i)\right]^2$$
$$+C\Delta t^2\sum_{i=0}^{n-1}b^{\ell,j_1}(\hat{Y}_i)^2\left[\frac{1}{\Delta x}(b^{k,j_2}(\hat{Y}_i+\Delta x e^\ell)-b^{k,j_2}(\hat{Y}_i))-\frac{\partial b^{k,j_2}}{\partial x^\ell}(\hat{Y}_i)\right]^2$$
$$\le C\Delta t^2\sum_{i=0}^{n-1}Z_i+C\Delta t\Delta x^2.$$

Therefore, altogether we have

$$Z_n\le C\Delta t\sum_{k=0}^{n-1}Z_k+C\Delta t\Delta x^2, \tag{5.7}$$

and an application of Lemma 2.3 implies that

$$\mathbb{E}\left[\sup_{0\le n\le N}|Y_n-\hat{Y}_n|^2\right]^{\frac{1}{2}}=O(\Delta x\Delta t^{\frac{1}{2}}). \tag{5.8}$$ □

We can apply a similar idea to the SAB scheme. That is, if we replace A_n and B_n in the order $\Delta t^{\frac{3}{2}}$ scheme (see Remark 3.1) by, for instance (with, again, Einstein's summation convention in effect),

$$A_n(t,x)=b^{kj}(t,x)\frac{1}{\Delta x}[a(t,x+\Delta x e^k)-a(t,x)]\Delta W^j, \tag{5.9}$$

where $e^k=(0,\dots,0,1,0,\dots,0)$, with the 1 in the kth position, and

$$
\begin{aligned}
B_n(t,x) = {} & b^j(t,x)\Delta W^j + \frac{1}{\Delta t}[b^j(t+\Delta t,x) - b^j(t,x)]I_{(0,j),n} \\
& + a^k(t,x)\frac{1}{\Delta x}[b^j(t,x+\Delta x e^k) - b^j(t,x)]I_{(0,j),n} \\
& + b^{ki}(t,x)b^{\ell i}(t,x)\frac{1}{8\Delta x^2}[b^j(t,x+\Delta x(e^i+e^\ell)) - b^j(t,x+\Delta x(e^i-e^\ell)) \\
& \quad - b^j(t,x+\Delta x(e^\ell-e^i)) + b^j(t,x-\Delta x(e^i+e^\ell))]I_{(0,j),n} \\
& + b^{kj_1}(t,x)\frac{1}{2\Delta x}[b^{j_2}(t,x+\Delta x e^k) - b^{j_2}(t,x-\Delta x e^k)]I_{(j_1,j_2),n} \\
& + b^{kj}(t,x)\frac{1}{\Delta x}[a(t,x+\Delta x e^k) - a(t,x)]I_{(j,0),n} \\
& + b^{k_1j_1}(t,x)\frac{1}{4\Delta x^2}[b^{k_2j_2}(t,x+\Delta x e^{k_1}) - b^{k_2j_2}(t-x+\Delta x e^{k_1})] \\
& \quad [b^{j_3}(t,x+\Delta x e^{k_2}) - b^{j_3}(t,x-\Delta x e^{k_1})]I_{(j_1,j_2,j_3),n} \\
& + b^{k_1j_1}(t,x)b^{k_2,j_2}(t,x)\frac{1}{4\Delta x^2} \\
& \quad [b^{j_3}(t,x+\Delta x(e^{k_1}+e^{k_2})) - b^{j_3}(t,x+\Delta x(e^{k_1}-e^{k_2})) \\
& \quad - b^{j_3}(t,x+\Delta x(e^{k_2}-e^{k_1})) + b^{j_3}(t,x-\Delta x(e^{k_1}+e^{k_2}))]I_{(j_1,j_2,j_3),n},
\end{aligned}
$$

we could then prove that this scheme converges to order $\max\{\Delta t^{\frac{3}{2}}, \Delta t^{\frac{1}{2}}\Delta x\}$ in a similar fashion.

6. Numerical simulation. The object of this section is to test numerically the accuracy of the scheme of section 3 and compare it to the theoretical result above (i.e., $O(\Delta t^2)$ accuracy) and to the accuracy of the Euler and Mil'shtein schemes (respectively, $O(\Delta t^{\frac{1}{2}})$ and $O(\Delta t)$). All the numerical results below are consistent with the theoretical ones.

We consider the following equation:

$$
dX_t = \beta^2 \sinh X_t \cosh^2 X_t \, dt + \beta \cosh^2 X_t \, dW_t, \tag{6.1}
$$

with $\beta = \frac{1}{10}$ and $X_0 = \frac{1}{2}$. This has the exact solution

$$
X_t = \operatorname{arctanh}(\beta W_t + \tanh X_0), \tag{6.2}
$$

respectively. This can be easily verified using Itô's formula and is just one of many possible examples listed in [6].

We computed approximate solutions Y_n using the Euler and Mil'shtein schemes and the SAB scheme from section 3. Then we computed the following error:

$$
e = \sqrt{\mathbb{E}\left(\sup_{0\le n\le N} |X_n - Y_n|^2\right)}. \tag{6.3}
$$

To estimate the mean value needed, we used 500 sample trajectories.

In Figure 6.1, the order of each scheme is given by the slope of the corresponding line. So we can see that the orders are $\frac{1}{2}$ for Euler, 1 for Mil'shtein, and 2 for the SAB of section 3.

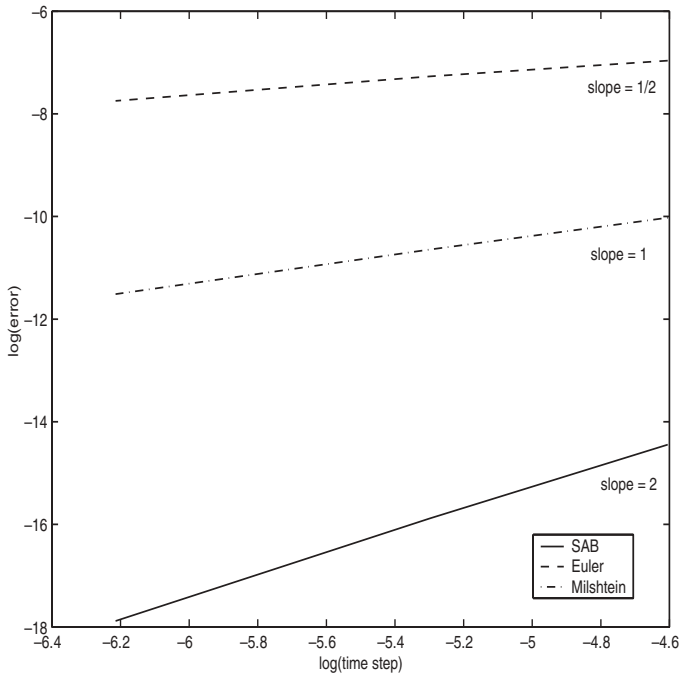


FIG. 6.1. *Results obtained with the stochastic equation* (6.1).

Note that for the SAB scheme, the stochastic integral $I_{(0,1,1)}$ (which is difficult to generate) was approximated by a normal law. The results tend to show that this does not affect the accuracy (at least in these two cases). We will try to improve this point, which seems to raise interesting probabilistic questions, as already mentioned in the introduction.

Acknowledgments. The authors are very grateful to Cecile Penland for bringing these issues to their attention, and they acknowledge very useful discussions with her and with Prashant Sardeshmukh. They are also indebted to Arnaud Debussche for several improvements on an earlier draft, and to Sylvain Faure, who provided the numerical simulations of section 6.

REFERENCES

[1] V. BALLY AND D. TALAY, *The law of the Euler scheme for stochastic differential equations,* I. *Convergence rate of the distribution function*, Probab. Theory Related Fields, 104 (1996), pp. 43–60.

[2] V. BALLY AND D. TALAY, *The law of the Euler scheme for stochastic differential equations,* II. *Convergence rate of the density*, Monte Carlo Methods Appl., 2 (1996), pp. 93–128.

[3] B. D. EWALD, *Numerical Methods for Stochastic Differential Equations in the Geosciences*, Ph.D. dissertation, Indiana University, Bloomington, 2002.

[4] B. D. EWALD, C. PENLAND, AND R. TEMAM, *Accurate integration of stochastic climate models with application to El Niño*, Monthly Weather Rev., 132 (2004), pp. 154–164.

[5] J. G. GAINES AND T. J. LYONS, *Random generation of stochastic area integrals*, SIAM J. Appl. Math., 54 (1994), pp. 1132–1146.

[6] P. E. KLOEDEN AND E. PLATEN, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. 23, Springer-Verlag, Berlin, 1992.

[7] P. E. Kloeden, E. Platen, and H. Schurz, *Numerical Solution of SDE through Computer Experiments*, Universitext, Springer-Verlag, Berlin, 1994.
[8] G. N. Mil'shtein, *Approximate integration of stochastic differential equations*, Theory Probab. Appl., 19 (1974), pp. 557–562.
[9] C. Penland and P. D. Sardeshmukh, *The optimal growth of tropical sea surface temperatures*, J. Climate, 8 (1995), pp. 1999–2024.
[10] W. Rümelin, *Numerical treatment of stochastic differential equations*, SIAM J. Numer. Anal., 19 (1982), pp. 604–613.
[11] P. D. Sardeshmukh and B. J. Hoskins, *The generation of global rotational flow by steady idealized tropical divergence*, J. Atmospheric Sci., 45 (1988), pp. 1228–1251.
[12] D. Talay, *Simulation of stochastic differential equations*, in Probabilistic Methods in Applied Physics, P. Krée and W. Wedig, eds., Springer-Verlag, Berlin, 1995, pp. 54–96.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2277–2297

MULTIPLE SHOOTING FOR UNSTRUCTURED NONLINEAR DIFFERENTIAL-ALGEBRAIC EQUATIONS OF ARBITRARY INDEX*

PETER KUNKEL†, VOLKER MEHRMANN‡, AND RONALD STÖVER§

Abstract. We study multiple shooting methods for the numerical solution of nonlinear boundary value problems for unstructured nonlinear systems of differential-algebraic equations with arbitrary index. We give a convergence analysis and demonstrate the results with some numerical examples.

Key words. nonlinear boundary value problem, differential-algebraic equations, multiple shooting

AMS subject classification. 65L10

DOI. 10.1137/S0036142902418904

1. Introduction. In this paper we consider the numerical solution of nonlinear boundary value problems for systems of differential-algebraic equations of arbitrary index by means of multiple shooting techniques. Multiple shooting is well studied and widely used for ordinary differential equations (see [1]) and also for special classes of systems of differential-algebraic equations (DAEs); see [15, 18].

In this paper we study general nonlinear DAE boundary value problems, i.e., problems of the form

$$\begin{array}{ll} \text{(a)} & F(t,x,\dot{x})=0, \\ \text{(b)} & r(x(\underline{t}),x(\bar{t}))=0, \end{array} \tag{1.1}$$

where $F:[\underline{t},\bar{t}]\times\mathbb{D}_x\times\mathbb{D}_{\dot{x}}\to\mathbb{R}^n$, $r:\mathbb{D}_x\times\mathbb{D}_x\to\mathbb{R}^d$ with $\mathbb{D}_x,\mathbb{D}_{\dot{x}}\subseteq\mathbb{R}^n$ open. The integer d denotes the number of differential components of x. A precise definition will follow in the next section.

The typical feature of shooting methods is that the solution of (1.1) is achieved through the solution of initial value problems, where it is implicitly assumed that they are well conditioned and can be solved sufficiently accurately. The boundary condition, together with the continuity conditions in the case of multiple shooting, then form a system of nonlinear equations for the initial values. In contrast to the application of shooting methods for the solution of ordinary differential equations, however, a problem arises for DAEs due to the fact that initial values have to be consistent with all explicit and hidden algebraic constraints. But even starting with a consistent initial guess, the iterative solver for the nonlinear equation will, in general, produce corrections that lead to inconsistent intermediate iterates. For this reason, shooting methods for nonlinear DAE boundary value problems were considered only

*Received by the editors November 29, 2002; accepted for publication (in revised form) January 6, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/sinum/42-6/41890.html

†Mathematisches Institut, Universität Leipzig, Augustusplatz 10–11, D-04109 Leipzig, Germany (kunkel@math.uni-leipzig.de). The research of this author was supported by DFG research grant Ku964/4.

‡Institut für Mathematik, MA 4-5, Technische Universität Berlin, Straße des 17. Juni 136, D-10623 Berlin, Germany (mehrmann@math.tu-berlin.de). The research of this author was supported by DFG research grant Me790/11.

§Zentrum für Technomathematik, Fachbereich 3, Universität Bremen, Postfach 330 440, D-28334 Bremen, Germany (stoever@math.uni-bremen.de).

in very special cases, where the algebraic constraints are known explicitly [18], or where they can be accessed due to the special structure of the equation [5, 15].

For general linear problems with variable coefficients, the set of consistent initial values at a given point forms an affine space which is numerically accessible; see [9, 12]. Based on this knowledge, shooting methods were developed for this case in [20]. Generalizing this approach to the general nonlinear case is the subject of the present paper. The new method that we present is able to treat boundary value problems for general DAEs of a given arbitrary index, i.e., there are no assumptions on the structure of the equations besides the requirement that the DAE can be assigned a certain kind of index. Note that such an assumption is indispensable because we need existence and uniqueness results on which we can base our method.

The paper is organized as follows. In section 2, we state some preliminaries on the theory of DAEs. In particular, we give the basic index definition and some further results that we need for the construction and investigation of the presented approach. In section 3 we discuss the local uniqueness of solutions of (1.1) via single shooting. A multiple shooting approach is then presented in section 4 together with a special Gauß–Newton-like method. In particular, we show that the arising linear equations can be reduced to shooting systems as they are obtained by multiple shooting for a system of d ordinary differential equations. In section 5, we then discuss the results of a number of numerical experiments. Finally, we give some conclusions in section 6.

2. Preliminaries. In general, the solution of a DAE may depend on derivatives of (1.1a). In particular, we must perform so many differentiations such that we can deduce all algebraic constraints that (1.1a) imposes on possible values for $x(t)$. Assuming in the following that all occurring functions are sufficiently smooth, we first introduce the so-called derivative array functions (see [2, 3])

$$F_\ell(t,x,\dot{x},\dots,x^{(\ell+1)}) = \begin{bmatrix} F(t,x,\dot{x}) \\ \frac{d}{dt}F(t,x,\dot{x}) \\ \vdots \\ (\frac{d}{dt})^\ell F(t,x,\dot{x}) \end{bmatrix} \tag{2.1}$$

that are obtained from (1.1a) by successive differentiation with respect to t. Note that we treat $(t,x,\dot{x},\dots,x^{(\ell+1)})$ here as independent variables such that F_ℓ is a function from some subset of $\mathbb{R}^{(\ell+2)n+1}$ into $\mathbb{R}^{(\ell+1)n}$. Partial derivatives will be denoted by corresponding subscripts as, e.g., in

$$F_{\ell;x} = \tfrac{\partial}{\partial x}F_\ell, \quad F_{\ell;\dot{x},\dots,x^{(\ell+1)}} = \begin{bmatrix} \frac{\partial}{\partial \dot{x}}F_\ell & \cdots & \frac{\partial}{\partial x^{(\ell+1)}}F_\ell \end{bmatrix}.$$

The following hypothesis states the central requirements on the DAE (1.1a); see [10].

HYPOTHESIS 2.1. *There exist (nonnegative) integers μ, a, and d such that for all values $(t,x,\dot{x},\dots,x^{(\mu+1)}) \in \mathbb{L}_\mu$ with*

$$\mathbb{L}_\mu = \{(t,x,\dot{x},\dots,x^{(\mu+1)}) \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(t,x,\dot{x},\dots,x^{(\mu+1)}) = 0\} \neq \emptyset \tag{2.2}$$

associated with F the following properties hold:

1. *We have*

$$\operatorname{rank} F_{\mu;\dot{x},\dots,x^{(\mu+1)}}(t,x,\dot{x},\dots,x^{(\mu+1)}) = (\mu+1)n - a$$

such that there exists a smooth matrix function $\hat{Z}_2$ *on* $\mathbb{L}_\mu$ *with orthonormal columns and size* $((\mu+1)n, a)$ *satisfying*

$$\hat{Z}_2^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}} = 0 \quad \text{on } \mathbb{L}_\mu.$$

2. *We have*

$$\operatorname{rank} \hat{Z}_2^T F_{\mu;x}(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = a$$

such that there exists a smooth matrix function $\hat{T}_2$ *on* $\mathbb{L}_\mu$ *with orthonormal columns and size* (n, d)*, where* $d = n - a$*, satisfying*

$$\hat{Z}_2^T F_{\mu;x} \hat{T}_2 = 0 \quad \text{on } \mathbb{L}_\mu.$$

3. *We have*

$$\operatorname{rank} F_{\dot{x}} \hat{T}_2(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = d$$

such that there exists a smooth matrix function $\hat{Z}_1$ *on* $\mathbb{L}_\mu$ *with orthonormal columns and size* (n, d) *satisfying*

$$\operatorname{rank} \hat{Z}_1^T F_{\dot{x}} \hat{T}_2 = d \quad \text{on } \mathbb{L}_\mu.$$

The minimal number μ (if it exists), such that Hypothesis 2.1 is fulfilled, is called the *strangeness index* of F. The numbers a and d denote the *size of the algebraic and differential part* of (1.1a). In particular, the choice of initial values is restricted by a algebraic constraints. More specific, for an initial value problem consisting of (1.1a) together with $x(t_0) = x_0$ to be solvable, the initial value x_0 must be extendable to a point $(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)})$ in $\mathbb{L}_\mu$. This requirement can be reduced to a conditions on x_0 itself; see [10] for more details.

A typical ingredient in the investigation of numerical methods for boundary value problems is the assumption that a solution of the given problem does exist. We therefore assume that there exists a sufficiently smooth solution $x^* \in C^1([\underline{t}, \overline{t}], \mathbb{R}^n)$ of (1.1) in the sense that

$$\begin{aligned} &\text{(a)} \quad F(t, x^*(t), \dot{x}^*(t)) = 0 \quad \text{for all } t \in [\underline{t}, \overline{t}], \\ &\text{(b)} \quad F_\mu(t, x^*(t), P(t)) = 0 \quad \text{for all } t \in [\underline{t}, \overline{t}], \\ &\text{(c)} \quad r(x^*(\underline{t}), x^*(\overline{t})) = 0, \end{aligned} \tag{2.3}$$

where $P : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n}$ is some smooth function that coincides with $\dot{x}^*$ in the first n components; see [11, Theorem 3] for sufficient conditions for such a function to exist.

Restricting the projectors $\hat{Z}_1$, $\hat{Z}_2$, $\hat{T}_2$ of Hypothesis 2.1 to the path $(t, x^*(t), P(t))$ which lies in $\mathbb{L}_\mu$ due to (2.3b), we obtain functions

$$Z_1 : [\underline{t}, \overline{t}] \to \mathbb{R}^{n,d}, \quad Z_2 : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n,a}, \quad T_2 : [\underline{t}, \overline{t}] \to \mathbb{R}^{n,d} \tag{2.4}$$

that satisfy

$$\begin{aligned} &\text{(a)} \quad Z_2(t)^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t, x^*(t), P(t)) = 0 \quad \text{for all } t \in [\underline{t}, \overline{t}], \\ &\text{(b)} \quad Z_2(t)^T F_{\mu;x}(t, x^*(t), P(t)) T_2(t) = 0 \quad \text{for all } t \in [\underline{t}, \overline{t}], \\ &\text{(c)} \quad \operatorname{rank} Z_1(t)^T F_{\dot{x}}(t, x^*(t), \dot{x}^*(t)) T_2(t) = d \quad \text{for all } t \in [\underline{t}, \overline{t}]. \end{aligned} \tag{2.5}$$

In addition, there exist smooth functions

$$\begin{aligned} &Z_2' : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n,(\mu+1)n-a}, \qquad && T_1 : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n,a}, \\ &T_2' : [\underline{t}, \overline{t}] \to \mathbb{R}^{n,a}, && T_1' : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n,(\mu+1)n-a}, \end{aligned} \tag{2.6}$$

such that the matrix valued functions $[Z_2', Z_2]$, $[T_1', T_1]$, and $[T_2', T_2]$ are pointwise orthogonal and, furthermore,

$$Z_2'(t)^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t, x^*(t), P(t))T_1(t) = 0 \quad \text{for all } t \in [\underline{t}, \overline{t}]. \tag{2.7}$$

It has been shown in [10] that for every

$$(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$$

the DAE (1.1a), if it satisfies Hypothesis 2.1, locally defines a function x from a neighborhood of t_0 into $\mathbb{R}^n$. In particular, the x so obtained solves a DAE of differentiation index at most one that is extracted from the derivative array equations $F_\mu(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = 0$. This solution can be extended until the boundary of the set where F_μ is defined is reached. Since

$$(t_0, x^*(t_0), P(t_0)) \in \mathbb{L}_\mu, \quad t_0 \in [\underline{t}, \overline{t}],$$

defines a solution on $[\underline{t}, \overline{t}]$, the same holds for every $(t_0, x_0, y_0) \in \mathbb{L}_\mu$ in a neighborhood of $(t_0, x^*(t_0), P(t_0))$.

In this section we have briefly presented some results on the solution and formulation of general nonlinear systems of DAEs. In the next section we use these results to analyze the single shooting method and the local uniqueness of solutions to the resulting nonlinear systems.

3. Single shooting and local uniqueness. In this section, which is of a more theoretical nature, we discuss the single shooting method. If initial value problems are uniquely solvable, then we can see the value of the solution at a certain point as a function of the initial value. This means that the boundary condition of a boundary value problem also becomes a function of the initial value. Therefore, a solution of a boundary value problem is said to be *locally unique* if the corresponding initial value is a locally unique solution of the boundary condition. For DAEs we must, of course, take into account that an initial condition must be consistent in the sense that there is a related point in the set $\mathbb{L}_\mu$.

For this reason we locally define a (nonlinear) projection onto $\mathbb{L}_\mu$ by considering the nonlinear system

$$\begin{aligned} &\text{(a)} \quad F_\mu(\underline{t}, \hat{x}, \hat{y}) = 0, \\ &\text{(b)} \quad T_2(\underline{t})^T(\hat{x} - x) = 0, \\ &\text{(c)} \quad T_1(\underline{t})^T(\hat{y} - y) = 0 \end{aligned} \tag{3.1}$$

in the unknowns $(x, y, \hat{x}, \hat{y})$.

If we write (3.1) as

$$H(x, y, \hat{x}, \hat{y}) = 0, \tag{3.2}$$

then a solution of this system is given by $(x^*(\underline{t}), P(\underline{t}), x^*(\underline{t}), P(\underline{t}))$. Since the Jacobian with respect to $\hat{x}, \hat{y}$ satisfies

$$\begin{aligned}
&\operatorname{rank} H_{\hat{x},\hat{y}}(x^*(\underline{t}), P(\underline{t}), x^*(\underline{t}), P(\underline{t})) \\
&\quad = \operatorname{rank} \begin{bmatrix} F_{\mu;x}(\underline{t}, x^*(\underline{t}), P(\underline{t})) & F_{\mu;y}(\underline{t}, x^*(\underline{t}), P(\underline{t})) \\ T_2(\underline{t})^T & 0 \\ 0 & T_1(\underline{t})^T \end{bmatrix} \\
&\quad = \operatorname{rank} \begin{bmatrix} Z_2'(\underline{t})^T F_{\mu;x}(\underline{t}, x^*(\underline{t}), P(\underline{t})) & Z_2'(\underline{t})^T F_{\mu;y}(\underline{t}, x^*(\underline{t}), P(\underline{t})) \\ Z_2(\underline{t})^T F_{\mu;x}(\underline{t}, x^*(\underline{t}), P(\underline{t})) & 0 \\ T_2(\underline{t})^T & 0 \\ 0 & T_1(\underline{t})^T \end{bmatrix},
\end{aligned}$$

and since by construction the matrices

$$\begin{bmatrix} Z_2'(t)^T F_{\mu;y}(t, x^*(t), P(t)) \\ T_1(t)^T \end{bmatrix}, \quad \begin{bmatrix} Z_2(t)^T F_{\mu;x}(t, x^*(t), P(t)) \\ T_2(t)^T \end{bmatrix}$$

are nonsingular for all $t \in [\underline{t}, \overline{t}]$, it follows that $H_{\hat{x},\hat{y}}(x^*(\underline{t}), P(\underline{t}), x^*(\underline{t}), P(\underline{t}))$ is nonsingular. We can therefore solve locally for $(\hat{x}, \hat{y})$ obtaining a function S according to

$$(\hat{x}, \hat{y}) = S(x, y). \tag{3.3}$$

Since $F_\mu(\underline{t}, S(x, y)) = 0$, we have that $(\underline{t}, S(x, y)) \in \mathbb{L}_\mu$ for every (x, y) in a neighborhood of $(x^*(\underline{t}), P(\underline{t}))$. Observing that the initial value problem for (1.1a) together with $(\underline{t}, S(x^*(\underline{t}), P(\underline{t}))) \in \mathbb{L}_\mu$ is solvable on the whole interval $[\underline{t}, \overline{t}]$, the initial value problem remains solvable on the whole interval $[\underline{t}, \overline{t}]$ with an initial condition given by $(\underline{t}, \underline{x}, \underline{y})$ from a neighborhood $\mathbb{L}_\mu \cap \mathbb{U}$ of $(\underline{t}, x^*(\underline{t}), P(\underline{t}))$. Thus the DAE defines a flow

$$\Phi : \mathbb{V} \to \mathbb{R}^n, \quad \mathbb{V} = \{(\underline{x}, \underline{y}) \mid (\underline{t}, \underline{x}, \underline{y}) \in \mathbb{L}_\mu \cap \mathbb{U}\} \tag{3.4}$$

that maps $(\underline{x}, \underline{y}) \in \mathbb{V}$ on the final value $x(\overline{t})$ of the solution x of the associated initial value problem.

For later use, we will need the derivatives of S at $(x^*(\underline{t}), P(\underline{t}))$. These are given by

$$\begin{aligned}
&H_{\hat{x},\hat{y}}(x^*(\underline{t}), P(\underline{t}), x^*(\underline{t}), P(\underline{t})) S_{x,y}(x^*(\underline{t}), P(\underline{t})) \\
&\quad = -H_{x,y}(x^*(\underline{t}), P(\underline{t}), x^*(\underline{t}), P(\underline{t})),
\end{aligned}$$

i.e.,

$$\begin{bmatrix} F_{\mu,x}(\underline{t}, x^*(\underline{t}), P(\underline{t})) & F_{\mu;y}(\underline{t}, x^*(\underline{t}), P(\underline{t})) \\ T_2(\underline{t})^T & 0 \\ 0 & T_1(\underline{t})^T \end{bmatrix} S_{x,y}(x^*(\underline{t}), P(\underline{t})) = \begin{bmatrix} 0 & 0 \\ T_2(\underline{t})^T & 0 \\ 0 & T_1(\underline{t})^T \end{bmatrix}.$$

Let the columns of W be orthonormal and span kernel $F_{\mu;x,y}(\underline{t}, x^*(\underline{t}), P(\underline{t}))$. Setting

$$\tilde{W} = \begin{bmatrix} T_2(\underline{t}) & 0 \\ 0 & T_1(\underline{t}) \end{bmatrix}, \tag{3.5}$$

we see that $\tilde{W}^T W$ is nonsingular, since $H_{\hat{x},\hat{y}}(x^*(\underline{t}), P(\underline{t}), x^*(\underline{t}), P(\underline{t}))$ is nonsingular, and we have

$$S_{x,y}(x^*(\underline{t}), P(\underline{t})) = W(\tilde{W}^T W)^{-1}\tilde{W}^T. \tag{3.6}$$

We then have the following theorem on the local uniqueness of solutions of boundary value problems for DAEs.

THEOREM 3.1. *The function x^* in* (2.3) *is a locally unique solution of the boundary value problem* (1.1) *in the sense that $(x^*(\underline{t}), P(\underline{t}))$ is a solution of*

$$\begin{aligned} &\text{(a)} \quad F_\mu(\underline{t}, \underline{x}, \underline{y}) = 0, \\ &\text{(b)} \quad T_1(\underline{t})^T(\underline{y} - P(\underline{t})) = 0, \\ &\text{(c)} \quad r(\underline{x}, \Phi(S(\underline{x}, \underline{y}))) = 0, \end{aligned} \tag{3.7}$$

with nonsingular Jacobian if and only if

$$\mathcal{E} = CT_2(\underline{t}) + D\Phi_{x,y}(x^*(\underline{t}), P(\underline{t}))S_x(x^*(\underline{t}), P(\underline{t}))T_2(\underline{t}) \tag{3.8}$$

is nonsingular, where $C = r_{x_a}(x^*(\underline{t}), x^*(\bar{t}))$ and $D = r_{x_b}(x^*(\underline{t}), x^*(\bar{t}))$.

Proof. Obviously, $(\underline{x}, \underline{y}) = (x^*(\underline{t}), P(\underline{t}))$ is a solution of (3.7). Moreover, the Jacobian J of (3.7) is given by

$$J = \begin{bmatrix} F_{\mu;x} & F_{\mu;y} \\ 0 & T_1(\underline{t})^T \\ C + D\Phi_{x,y}S_x & D\Phi_{x,y})S_y \end{bmatrix},$$

where we have omitted the arguments $\underline{t}, x^*(\underline{t}), P(\underline{t})$. So we have that $\operatorname{rank} J$ is equal to

$$\operatorname{rank} \begin{bmatrix} Z_2'^T F_{\mu;x} T_2' & Z_2'^T F_{\mu;x} T_2 & Z_2'^T F_{\mu;y} T_1' & 0 \\ Z_2^T F_{\mu;x} T_2' & 0 & 0 & 0 \\ 0 & 0 & 0 & I \\ (C + D\Phi_{x,y}S_x)T_2' & (C + D\Phi_{x,y}S_x)T_2 & D\Phi_{x,y}S_y T_1' & D\Phi_{x,y}S_y T_1 \end{bmatrix}.$$

Since

$$S_x = W(\tilde{W}^T W)^{-1}\begin{bmatrix} T_2^T \\ 0 \end{bmatrix}, \quad S_y = W(\tilde{W}^T W)^{-1}\begin{bmatrix} 0 \\ T_1^T \end{bmatrix}$$

by (3.5), we have $S_x T_2' = 0$ and $S_y T_1' = 0$ by (2.6). Moreover, $Z_2^T F_{\mu;x} T_2'$ and $Z_2'^T F_{\mu;y} T_1'$ are nonsingular by construction. Thus J has full rank if and only if

$$\mathcal{E} = (C + D\Phi_{x,y}S_x)T_2$$

has full rank. ∎

REMARK 3.2. *In the case of linear boundary value problems, i.e., problems* (1.1) *where F and r are linear, condition* (3.8) *coincides with that given in* [14, 20, 21] *and thus yields global uniqueness of the solution x^**

Proof. Since we do not need this result further in the course of this paper, we give the proof in the appendix. ∎

4. Multiple shooting. It is well known that in single shooting, one is faced with the difficulty that the arising initial value problems may be unstable. This may lead to large solution components or even to the problem that the solution does not extend until $\bar{t}$ due to errors in the initial guess. To overcome these difficulties, in multiple shooting the solution interval is split beforehand into smaller subintervals according to

$$\underline{t} = t_0 < t_1 < \cdots < t_{N-1} < t_N = \bar{t}, \quad N \in \mathbb{N}. \tag{4.1}$$

Given initial guesses

$$(x_i, y_i) \in \mathbb{R}^{(\mu+2)n}, \quad i = 0, \ldots, N, \tag{4.2}$$

at these points, the idea is to project (t_i, x_i, y_i) onto $\mathbb{L}_\mu$ and to solve the associated initial value problems on $[t_i, t_{i+1}]$, requiring that the pieces correspond to a continuous solution on the whole interval and that the boundary condition is satisfied.

In contrast to section 3 which was merely dedicated to a theoretical investigation, in this section we present a method that can actually be implemented. We therefore are not allowed to use functions such as Z_2 or T_2 in the definition of the procedure. Instead, we must look for computationally available quantities.

Given (t_i, x_i, y_i) as an initial guess for a point on $\mathbb{L}_\mu$, we can solve $F_\mu(t_i, x, y) = 0$ by the Gauß–Newton method (see, e.g., [16]) to obtain $(t_i, \tilde{x}_i, \tilde{y}_i) \in \mathbb{L}_\mu$. Of course, we must require that the guess (t_i, x_i, y_i) is good enough to guarantee convergence. Applying Hypothesis 2.1, we can then compute quantities $\tilde{Z}_{2,i}$ and $\tilde{T}_{2,i}$, where the columns form orthonormal bases of corange $F_{\mu;y}(t_i, \tilde{x}_i, \tilde{y}_i)$ and kernel $\tilde{Z}_{2,i}^T F_{\mu;x}(t_i, \tilde{x}_i, \tilde{y}_i)$, respectively. In the same way, we can determine quantities $\tilde{Z}'_{2,i}$ and $\tilde{T}_{1,i}$.

Similar to (3.1), the system

$$\begin{aligned} &\text{(a)} \quad F_\mu(t_i, \hat{x}_i, \hat{y}_i) = 0, \\ &\text{(b)} \quad \tilde{T}_{2,i}^T(\hat{x}_i - x_i) = 0, \\ &\text{(c)} \quad \tilde{T}_{1,i}^T(\hat{y}_i - y_i) = 0 \end{aligned} \tag{4.3}$$

locally defines functions S_i according to

$$(\hat{x}_i, \hat{y}_i) = S_i(x_i, y_i) \tag{4.4}$$

in such a way that $(t_i, S_i(x_i, y_i)) \in \mathbb{L}_\mu$. Defining W_i to have columns that form an orthonormal basis of kernel $F_{\mu;x,y}(t_i, \hat{x}_i, \hat{y}_i)$ with $(t_i, \hat{x}_i, \hat{y}_i) \in \mathbb{L}_\mu$ and setting

$$\tilde{W}_i = \begin{bmatrix} \tilde{T}_{2,i} & 0 \\ 0 & \tilde{T}_{1,i} \end{bmatrix}, \tag{4.5}$$

we obtain

$$S_{i;x,y}(\hat{x}_i, \hat{y}_i) = W_i(\tilde{W}_i^T W_i)^{-1} \tilde{W}_i^T \tag{4.6}$$

similar to (3.6) as long as $\tilde{W}_i^T W_i$ is invertible. As was done with Φ in section 3, we define flows Φ_i that map initial values $(\hat{x}_i, \hat{y}_i)$ with $(t_i, \hat{x}_i, \hat{y}_i) \in \mathbb{L}_\mu$ on the value $x(t_{i+1})$ of the solution x of the corresponding initial value problem.

The multiple shooting system is then given by

$$\begin{aligned} &\text{(a)} \quad F_\mu(t_i, x_i, y_i) = 0, && i = 0, \ldots, N, \\ &\text{(b)} \quad \tilde{T}_{2,i+1}^T(x_{i+1} - \Phi_i(S_i(x_i, y_i))) = 0, && i = 0, \ldots, N-1, \\ &\text{(c)} \quad r(x_0, x_N) = 0. \end{aligned} \tag{4.7}$$

Comparing with the single shooting method of section 3, (3.7a) is now required in (4.7a) at all mesh points t_i with corresponding unknowns (x_i, y_i). Besides the boundary condition (4.7c), we impose continuity conditions for the differential components in (4.7b). Condition (3.7b), which was responsible for local uniqueness of the solution in (3.7), cannot be used here because it involves knowledge of the actual solution. Thus, in the present form, system (4.7) is underdetermined. It is therefore solved by a Gauß–Newton-type iteration method which we present in what follows. In the course of the presentation, we will select a suitable generalized inverse of the Jacobian by additional conditions which will turn out to be the appropriate replacement for (3.7b).

Given approximations (x_i, y_i), the Gauß–Newton-type method is defined by the corrections $(\Delta x_i, \Delta y_i)$ that are added to (x_i, y_i) to get updated approximations. In the (underdetermined) ordinary Gauß–Newton method, these corrections satisfy the linearized equations

$$\begin{aligned}
&\text{(a)} \quad F_{\mu;x}(t_i, x_i, y_i)\Delta x_i + F_{\mu;y}(t_i, x_i, y_i)\Delta y_i = -F_\mu(t_i, x_i, y_i),\\
&\text{(b)} \quad \tilde{T}_{2,i+1}^T(\Delta x_{i+1} - \Phi_{i;x,y}(S_i(x_i, y_i))(S_{i;x}(x_i, y_i)\Delta x_i + S_{i;y}(x_i, y_i)\Delta y_i))\\
&\qquad\qquad = -\tilde{T}_{2,i+1}^T(x_{i+1} - \Phi_i(S_i(x_i, y_i))),\\
&\text{(c)} \quad r_{x_a}(x_0, x_N)\Delta x_0 + r_{x_b}(x_0, x_N)\Delta x_N = -r(x_0, x_N).
\end{aligned} \tag{4.8}$$

For an efficient numerical method, however, the structure and the properties of the Jacobian should be utilized. In the following, we will perturb the coefficient matrix in such a way that the system decouples into smaller systems of reasonable size. In particular, the perturbations that we apply will tend to zero when the (x_i, y_i) converge to a solution of (4.7) resulting in a Gauß–Newton-like process with superlinear convergence rate; cf. [4].

In a solution of (4.7), the matrices $F_{\mu;y}(t_i, x_i, y_i)$ will have rank deficiency a. We therefore perturb $F_{\mu;y}(t_i, x_i, y_i)$ to matrices $\tilde{M}_i$ with rank deficiency a. The only condition we must require is that these perturbations tend to zero when the matrices $F_{\mu;y}(t_i, x_i, y_i)$ tend to matrices with rank deficiency a. One possibility for achieving this is to neglect the a smallest singular values of $F_{\mu;y}(t_i, x_i, y_i)$; see, e.g., [6]. The equations (4.8a) are thus replaced by

$$F_{\mu;x}(t_i, x_i, y_i)\Delta x_i + \tilde{M}_i\Delta y_i = -F_\mu(t_i, x_i, y_i). \tag{4.9}$$

Let the columns of $Z_{2,i}$ form an orthonormal basis of corange $\tilde{M}_i$ and let $[Z'_{2,i}, Z_{2,i}]$ be orthogonal. Relation (4.9) then splits into

$$\begin{aligned}
&\text{(a)} \quad Z_{2,i}'^T F_{\mu;x}(t_i, x_i, y_i)\Delta x_i + Z_{2,i}'^T\tilde{M}_i\Delta y_i = -Z_{2,i}'^T F_\mu(t_i, x_i, y_i),\\
&\text{(b)} \quad Z_{2,i}^T F_{\mu;x}(t_i, x_i, y_i)\Delta x_i = -Z_{2,i}^T F_\mu(t_i, x_i, y_i).
\end{aligned} \tag{4.10}$$

Requiring in addition that

$$\tilde{T}_{1,i}^T\Delta y_i = 0 \tag{4.11}$$

as a substitute for (3.7b) and observing that

$$\begin{bmatrix} Z_{2,i}'^T\tilde{M}_i \\ \tilde{T}_{1,i}^T \end{bmatrix}$$

is nonsingular for sufficiently good initial guesses (x_i, y_i), it follows that we can solve (4.10a) with (4.11) for Δy_i in terms of Δx_i.

Let the columns of the matrix $T_{2,i}$ form an orthonormal basis of the space kernel $Z_{2,i}^T F_{\mu;x}(t_i, x_i, y_i)$. For sufficiently good initial guesses (x_i, y_i), $\tilde{T}_{2,i}^T T_{2,i}$ is also nonsingular. Thus, there exists a matrix $T'_{2,i}$ such that $[T'_{2,i}, T_{2,i}]$ is nonsingular and

$$\tilde{T}_{2,i}^T T'_{2,i} = 0. \tag{4.12}$$

Defining $\Delta v'_i$ and Δv_i by the relation

$$\Delta x_i = T'_{2,i} \Delta v'_i + T_{2,i} \Delta v_i, \tag{4.13}$$

(4.10b) becomes

$$Z_{2,i}^T F_{\mu;x}(t_i, x_i, y_i) T'_{2,i} \Delta v'_i = -Z_{2,i}^T F_{\mu}(t_i, x_i, y_i). \tag{4.14}$$

Since $Z_{2,i}^T F_{\mu;x}(t_i, x_i, y_i) T'_{2,i}$ is nonsingular by construction, (4.14) can be solved for $\Delta v'_i$.

Turning to (4.8b), we know that at a solution of (4.8), the relations

$$\begin{aligned} &\text{(a)}\ \ S_{i;x}(x_i, y_i)\Delta x_i = W_i(\tilde{W}_i^T W_i)^{-1} \begin{bmatrix} \tilde{T}_{2,i}^T \\ 0 \end{bmatrix} T_{2,i} \Delta v_i = S_{i;x}(x_i, y_i) T_{2,i} \Delta v_i, \\ &\text{(b)}\ \ S_{i;y}(x_i, y_i)\Delta y_i = W_i(\tilde{W}_i^T W_i)^{-1} \begin{bmatrix} 0 \\ \tilde{T}_{1,i}^T \end{bmatrix} \Delta y_i = 0 \end{aligned} \tag{4.15}$$

hold because of (4.11) and (4.12). Thus, we replace (4.8b) by

$$\begin{aligned} &\tilde{T}_{2,i+1}^T T_{2,i+1} \Delta v_{i+1} - \tilde{T}_{2,i+1}^T \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x}(x_i, y_i) T_{2,i} \Delta v_i \\ &\quad = -\tilde{T}_{2,i+1}^T (x_{i+1} - \Phi_i(S_i(x_i, y_i))), \end{aligned} \tag{4.16}$$

which is again a perturbation that tends to zero when the iteration converges. The main advantage of (4.16) is that we need only the derivative $\Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x}(x_i, y_i)$ in the direction of the d columns of $T_{2,i}$. In particular, if we use numerical differentiation to approximate this derivative, then we need only solve d initial value problems.

Finally, we write (4.8c) in the form

$$\begin{aligned} &r_{x_a}(x_0, x_N) T_{2,0} \Delta v_0 + r_{x_b}(x_0, x_N) T_{2,N} \Delta v_N \\ &\quad = -r(x_0, x_N) - r_{x_a}(x_0, x_N) T'_{2,0} \Delta v'_0 - r_{x_b}(x_0, x_N) T'_{2,N} \Delta v'_N. \end{aligned} \tag{4.17}$$

Setting

$$\begin{aligned} &\text{(a)} \quad G_i = \tilde{T}_{2,i+1}^T \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x}(x_i, y_i) T_{2,i}, \quad i = 0, \ldots, N-1, \\ &\text{(b)} \quad J_i = \tilde{T}_{2,i}^T T_{2,i}, \quad i = 1, \ldots, N, \\ &\text{(c)} \quad \tilde{C} = r_{x_a}(x_0, x_N) T_{2,0}, \quad \tilde{D} = r_{x_b}(x_0, x_N) T_{2,N}, \end{aligned} \tag{4.18}$$

the linear system that we have to solve for the unknowns Δv_i has the shooting-like coefficient matrix

$$\tilde{\mathcal{E}}_N = \begin{bmatrix} -G_0 & J_1 & & & \\ & -G_1 & J_2 & & \\ & & \ddots & \ddots & \\ & & & -G_{N-1} & J_N \\ \tilde{C} & & & & \tilde{D} \end{bmatrix}. \tag{4.19}$$

This system can be solved by standard methods such as Gaussian elimination with pivoting [6]. Since the blocks J_i are invertible for sufficiently good initial guesses (x_i, y_i), it follows that the matrix $\tilde{\mathcal{E}}_N$ is nonsingular if and only if

$$\mathcal{E}_N = \tilde{C} + \tilde{D}(J_N^{-1}G_{N-1})(J_{N-1}^{-1}G_{N-2})\cdots(J_2^{-1}G_1)(J_1^{-1}G_0) \tag{4.20}$$

is nonsingular. Thus, for the method to work, it suffices to show that this is the case at least at the solution and therefore in some neighborhood of it.

At a solution $(x_i, y_i) = (x^*(t_i), y_i^*)$, the matrix $\mathcal{E}_N$ takes the form

$$\begin{aligned}\mathcal{E}_N = CT_{2,0} + DT_{2,N} \prod_{i=N-1}^{i=0} \Big[&(\tilde{T}_{2,i+1}^T T_{2,i+1})^{-1} \tilde{T}_{2,i+1}^T \\ &\times \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x}(x_i, y_i) T_{2,i} \Big].\end{aligned} \tag{4.21}$$

To take into account that $\Phi_i(S_i(x, y))$ is consistent at t_{i+1} for (x, y) in a neighborhood of (x_i, y_i) as the value of a solution of the DAE on $[t_i, t_{i+1}]$, we consider the system

$$\begin{aligned} &\text{(a)} \quad F_\mu(t_i, x, \hat{y}) + Z_{2,i}\alpha = 0, \\ &\text{(b)} \quad \tilde{T}_{1,i}^T(\hat{y} - y_i) = 0. \end{aligned} \tag{4.22}$$

Writing this as

$$H_i(x, \hat{y}, \alpha) = 0, \tag{4.23}$$

we know that $H_i(x_i, y_i, 0) = 0$. Since

$$\begin{aligned}&\operatorname{rank} H_{i;\hat{y},\alpha}(x_i, y_i, 0) \\ &\quad = \operatorname{rank} \begin{bmatrix} F_{\mu;y}(t_i, x_i, y_i) & Z_{2,i} \\ \tilde{T}_{1,i}^T & 0 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} Z_{2,i}'^T F_{\mu;y}(t_i, x_i, y_i) & 0 \\ 0 & I \\ \tilde{T}_{1,i}^T & 0 \end{bmatrix},\end{aligned}$$

the construction of $Z_{2,i}'$ and $\tilde{T}_{1,i}$ guarantees that the matrix $H_{i;\hat{y},\alpha}(x_i, y_i, 0)$ is nonsingular. Thus, (4.22) locally defines functions K_i and L_i according to

$$\hat{y} = K_i(x), \quad \alpha = L_i(x). \tag{4.24}$$

For all x with $L_i(x) = 0$ we have $F_\mu(t_i, x, K_i(x)) = 0$ and x is consistent at t_i. Furthermore, differentiating

$$F_\mu(t_i, x, K_i(x)) + Z_{2,i}L_i(x) = 0,$$

evaluating at x_i and multiplying by $Z_{2,i}^T$ yields

$$L_{i;x}(x_i) = -Z_{2,i}^T F_{\mu;x}(t_i, x_i, y_i).$$

Hence, $L_{i;x}$ has full row rank in a neighborhood of x_i, and all solutions of $L_i(x) = 0$ form a manifold of dimension $d = n - a$ which is a submanifold of the manifold of consistent values at point t_i. Since the dimension of the latter manifold is also d (see [10]), they must coincide.

Thus, given an x that is consistent at t_i, the function K_i yields a $\hat{y}$ such that $(t_i, x, \hat{y}) \in \mathbb{L}_\mu$ while $L_i(x) = 0$. In particular,

$$F_\mu(t_{i+1}, \Phi_i(S_i(x, y)), K_{i+1}(\Phi_i(S_i(x, y)))) = 0 \tag{4.25}$$

holds in a neighborhood of (x_i, y_i). Differentiating this relation with respect to (x, y) and setting $(x, y) = (x_i, y_i)$, we obtain

$$\begin{aligned} &F_{\mu;x}(t_{i+1}, x_{i+1}, y_{i+1})\Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i) \\ &\quad + F_{\mu;y}(t_{i+1}, x_{i+1}, y_{i+1})K_{i+1;x}(x_{i+1})\Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i) = 0. \end{aligned}$$

Multiplying with $Z_{2,i+1}^T$ from the left finally yields

$$Z_{2,i+1}^T F_{\mu;x}(t_{i+1}, x_{i+1}, y_{i+1})\Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i) = 0. \tag{4.26}$$

Hence, the columns of $\Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i)$ lie in the kernel of the matrix $Z_{2,i+1}^T F_{\mu;x}(t_{i+1}, x_{i+1}, y_{i+1})$, which in turn is spanned by the columns of $T_{2,i+1}$. Since the expression $T_{2,i+1}(\tilde{T}_{2,i+1}^T T_{2,i+1})^{-1}\tilde{T}_{2,i+1}^T$ is a projector on this kernel, we have

$$\begin{aligned} &T_{2,i+1}(\tilde{T}_{2,i+1}^T T_{2,i+1})^{-1}\tilde{T}_{2,i+1}^T \Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i) \\ &\quad = \Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i). \end{aligned} \tag{4.27}$$

Thus, (4.21) reduces to

$$\mathcal{E}_N = CT_{2,0} + D\left[\prod_{i=N-1}^{i=0} \Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x}(x_i, y_i)\right] T_{2,0}. \tag{4.28}$$

Finally, defining

$$\Psi_i(x) = (x, K_i(x)) \tag{4.29}$$

and using $\tilde{T}_{1,i}^T K_{i;x}(x_i) = 0$, which holds due to (4.22b), we find that

$$\begin{aligned} &S_{i;x,y}(x_i, y_i)\Psi_{i;x}(x_i) \\ &\quad = W_i(\tilde{W}_i^T W_i)^{-1}\begin{bmatrix} \tilde{T}_{2,i}^T & 0 \\ 0 & \tilde{T}_{1,i}^T \end{bmatrix}\begin{bmatrix} I \\ K_{i;x}(x_i) \end{bmatrix} \\ &\quad = W_i(\tilde{W}_i^T W_i)^{-1}\begin{bmatrix} \tilde{T}_{2,i}^T \\ 0 \end{bmatrix} = S_{i;x}(x_i, y_i). \end{aligned}$$

Hence, (4.28) becomes

$$\mathcal{E}_N = CT_{2,0} + D\left[\prod_{i=N-1}^{i=0} \Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i)\Psi_{i;x}(x_i)\right] T_{2,0}. \tag{4.30}$$

Comparing with (3.8), the term in brackets in (4.30) is nothing more than the derivative $\Phi_{x,y}\tilde{S}_x$ of $\Phi \circ \tilde{S}$ decomposed according to

$$\Phi \circ \tilde{S} = (\Phi_{N-1} \circ S_{N-1}) \circ (\Psi_{N-1} \circ \Phi_{N-2} \circ S_{N-2}) \circ \cdots \circ (\Psi_1 \circ \Phi_0 \circ S_0), \tag{4.31}$$

where $\tilde{S}$ differs from S by replacing $T_1(\underline{t}), T_2(\underline{t})$ with $\tilde{T}_{1,0}, \tilde{T}_{2,0}$ in (3.1). This means that for sufficiently good initial guesses, the matrix $\mathcal{E}_N$ is nonsingular when $\mathcal{E}$ of (3.8) is nonsingular, i.e., when there is a locally unique solution of the boundary value problem in the sense of Theorem 3.1.

Summarizing the results obtained, we have the following convergence theorem.

THEOREM 4.1. *Suppose that the boundary value problem* (1.1) *satisfies Hypothesis* 2.1 *and that* (1.1) *has a locally unique solution according to Theorem* 3.1. *Then, for sufficiently good initial guesses, the iterates of the Gauß–Newton-like procedure developed in the course of this section converge superlinearly to a solution of* (4.7).

Proof. Writing the Gauß–Newton-like procedure for (4.7) in the form

$$z_{\nu+1} = z_\nu - \mathcal{A}_\nu^- \mathcal{F}(z_\nu),$$

where $\mathcal{A}_\nu$ is the chosen perturbation of $\mathcal{F}_z(z_\nu)$ and $\mathcal{A}_\nu^-$ denotes the chosen pseudoinverse due to (4.11), we have

$$\mathcal{A}_\nu \mathcal{A}_\nu^- = I,$$

since $\mathcal{A}_\nu^-$ yields a solution of the perturbed linear system. Thus, we get

$$\begin{aligned} z_{\nu+1} - z_\nu &= -\mathcal{A}_\nu^- \mathcal{F}(z_\nu) \\ &= -\mathcal{A}_\nu^- \big[\mathcal{F}(z_\nu) - \mathcal{F}(z_{\nu-1}) - \mathcal{A}_{\nu-1}(z_\nu - z_{\nu-1})\big] \\ &= -\mathcal{A}_\nu^- \big[\mathcal{F}(z_{\nu-1} + s(z_\nu - z_{\nu-1}))\big|_{s=0}^{s=1} - \mathcal{F}_z(z_{\nu-1})(z_\nu - z_{\nu-1}) \\ &\qquad -(\mathcal{A}_{\nu-1} - \mathcal{F}_z(z_{\nu-1}))(z_\nu - z_{\nu-1})\big] \\ &= -\mathcal{A}_\nu^- \big[\textstyle\int_0^1 (\mathcal{F}_z(z_{\nu-1}+s(z_\nu - z_{\nu-1})) - \mathcal{F}_z(z_{\nu-1}))(z_\nu - z_{\nu-1})\, ds \\ &\qquad -(\mathcal{A}_{\nu-1} - \mathcal{F}_z(z_{\nu-1}))(z_\nu - z_{\nu-1})\big]. \end{aligned}$$

Introducing constants β, γ, and δ_ν according to

$$\|\mathcal{A}_\nu^-\| \le \beta, \quad \|\mathcal{F}_z(u) - \mathcal{F}_z(v)\| \le \gamma\|u - v\|, \quad \|\mathcal{A}_\nu - \mathcal{F}_z(z_\nu)\| = \delta_\nu$$

for some vector norm and its corresponding matrix norm (recalling that we assume sufficient smoothness for the data), we obtain the estimate

$$\|z_{\nu+1} - z_\nu\| \le \frac{1}{2}\beta\gamma\|z_\nu - z_{\nu-1}\|^2 + \beta\delta_{\nu-1}\|z_\nu - z_{\nu-1}\|.$$

Since $\delta_\nu \to 0$ when z_ν converges to a solution, superlinear convergence follows as in [4]. ∎

REMARK 4.2. *In all steps of the above construction, we were forced to include the equation $F_\mu = 0$ in order to get hold of all algebraic constraints posed by the given DAE. For problems with structure it is often much easier to address these algebraic constraints. Sometimes one can even simply write them down. In both cases, the approach presented above can be simplified to obtain more efficient procedures. In particular, one can replace the equation $F_\mu = 0$ by a simpler one that determines the algebraic constraints, or even replace it by them if they are known explicitly.*

REMARK 4.3. *As already mentioned, the main problem in the construction of shooting methods for DAEs is how to deal with inconsistent intermediate iterates. In the method presented here, we used (locally defined) nonlinear projections on $\mathbb{L}_\mu$ to get consistent initial values. A second possibility would have been to shift the manifold $\mathbb{L}_\mu$ in such a way that the given inconsistent iterate then lies in the shifted manifold. In the case of single shooting, if*

$$(\hat{x}, \hat{y}) = S(x_0, y_0),$$

then we would define

$$\hat{\mathbb{L}}_\mu = \{(t, x, y) \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(t, x - x_0 + \hat{x}, y - y_0 + \hat{y}) = 0\}$$

and solve the arising initial value problem with respect to $\hat{\mathbb{L}}_\mu$. The same idea is used by [18] *in the form of so-called relaxed algebraic constraints when these are explicitly*

available. One can show that using the technique of shifting the manifold would yield a method with the same properties as the new method we have presented. However, the method of shifting the manifold has the disadvantage that it requires modifying F_μ for use in the initial value solver.

5. Numerical experiments. The procedure of section 4 has been implemented in `FORTRAN` in the form of a research code. All rank decisions, computations of kernel and corange matrices and their complements, as well as the solution of linear equations, are performed on the basis of singular value decompositions. In particular, no concern was laid on efficiency questions. Due to that, all multiple shooting codes designed for special structured DAEs should outperform the present implementation. All computations were done on a `Sun Blade 100` workstation with 500 MHz in `IEEE` double precision.

The following examples cover a large variety of problem classes of different structure and differentiation index. In all examples, the Gauß–Newton-like procedure of Theorem 4.1 was terminated as soon as $\|\Delta z_\nu\|_2 \le 10^{-5}$, where $\Delta z_\nu = z_{\nu+1} - z_\nu$. The entries G_i of (4.19) were approximated by numerical differentiation according to

$$G_i e_j = \tfrac{1}{\eta} \tilde{T}_{2,i+1}^T [\Phi_i(S_i(x_i + \eta T_{2,i} e_j, y_i)) - \Phi_i(x_i, y_i)], \qquad j = 1, \dots, d,\ i = 0, \dots, N-1,$$

where e_j is the jth canonical basis vector of $\mathbb{R}^d$. We used the choice $\eta = 10^{-7}$. All initial value problems occurring were solved with `GENDA` of [13] using the tolerance 10^{-5}.

Example 5.1. In [7], the model of a periodically driven electronic amplifier is given. The equations with $n = 5$ for the unknowns $(U_1, \dots, U_5)$ read as

$$\begin{aligned}
&(U_E(t) - U_1)/R_0 + C_1(\dot{U}_2 - \dot{U}_1) = 0,\\
&(U_B - U_2)/R_2 - U_2/R_1 + C_1(\dot{U}_1 - \dot{U}_2) - 0.01 f(U_2 - U_3) = 0,\\
&f(U_2 - U_3) - U_3/R_3 - C_2\dot{U}_3 = 0,\\
&(U_B - U_4)/R_4 + C_3(\dot{U}_5 - \dot{U}_4) - 0.99 f(U_2 - U_3) = 0,\\
&-U_5/R_5 + C_3(\dot{U}_4 - \dot{U}_5) = 0
\end{aligned}$$

with

$$\begin{aligned}
U_E(t) &= 0.4 \sin(200\pi t), \quad U_B = 6,\\
f(U) &= 10^{-6}(\exp(U/0.026) - 1),\\
R_0 &= 1000, \quad R_1 = \cdots = R_5 = 9000,\\
C_1 &= 10^{-6}, \quad C_2 = 2 \cdot 10^{-6}, \quad C_3 = 3 \cdot 10^{-6}.
\end{aligned}$$

The problem is known to satisfy Hypothesis 2.1 with $\mu = 0$, $d = 3$, and $a = 2$. If we ask for the periodic response of the amplifier, we are led to the boundary conditions

$$U_l(0) = U_l(0.01), \quad l = 2, 3, 5;$$

thus $\underline{t} = 0$ and $\bar{t} = 0.01$. We used $N = 1$ and determined the initial guess for the unknowns (x_i, y_i), $i = 0, \dots, N$, by integration starting with

$$(0, V_1, V_1, U_B, 0, 0, 0, V_2, 0, 0) \in \mathbb{L}_\mu,$$

where $V_1 = U_B \frac{R_1}{R_1 + R_2}$ and $V_2 = -\frac{V_1}{R_3 C_2}$.

TABLE 5.1
Results for Example 5.1.

ν	$\|\Delta z_\nu\|_2$
0	0.339D+02
1	0.113D-02
2	0.107D-04
3	0.168D-08

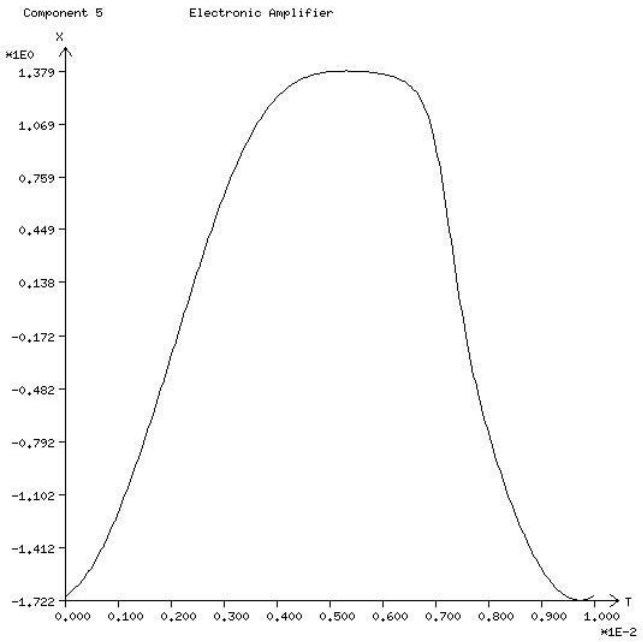


FIG. 5.1. *Periodic response for Example* 5.1.

Then the presented method successfully computed a periodic solution in about 3.1 seconds. The behavior of the Gauß–Newton-like method is given in Table 5.1. One component of the periodic response of the amplifier is shown in Figure 5.1.

Example 5.2. In [17], a multibody system with nonholonomic constraint is presented. The model equations for the unknowns $(\varphi, z_G, z_Z, \lambda)$ are given by

$$\begin{aligned} &I_R\ddot{\varphi} = u, \\ &m_G\ddot{z}_G + d_1(\dot{z}_G - \dot{z}_Z) + c_1(z_G - z_Z) = \lambda, \\ &m_Z\ddot{z}_Z + d_1(\dot{z}_Z - \dot{z}_G) + c_1(z_Z - z_G) = 0, \\ &\dot{z}_G = v_U\varphi, \end{aligned}$$

where

$$\begin{aligned} &I_R = 0.002, \qquad m_G = 3, \qquad m_Z = 10, \\ &v_U = 2.8, \qquad c_1 = 250, \qquad d_1 = 10. \end{aligned}$$

We ask for the time T when, starting from the trivial equilibrium, the maximal allowed $\varphi_{\max} = 0.27$ is reached for an external force $u = 0.001$. The boundary conditions are

TABLE 5.2
Results for Example 5.2.

ν	$\|\Delta z_\nu\|_2$
0	0.727D+00
1	0.405D-01
2	0.345D-04
3	0.107D-09

then given by

$$\varphi(0) = z_G(0) = z_Z(0) = \dot\varphi(0) = \dot z_Z(0) = 0, \quad \varphi(T) = 0.27.$$

Writing the model equations as a first order system and transforming the unknown interval $[0, T]$ to $[0, 1]$, thus introducing T as further unknown, we get a problem with $n = 8$ in the unknowns $(\varphi, z_G, z_Z, \dot\varphi, \dot z_G, \dot z_Z, \lambda, T)$. Hypothesis 2.1 is here satisfied with $\mu = 1$, $d = 6$, and $a = 2$. We set $N = 1$ and took $T = 1$ to get initial guesses for (x_i, y_i) by integration starting with the equilibrium state. Table 5.2 shows the behavior of the Gauß–Newton-like procedure. The computing time was about 4.3 seconds.

Example 5.3. A pendulum in two space dimensions is modeled by

$$\begin{aligned} &\dot p_1 = v_1, \quad \dot v_1 = 2p_1\lambda, \\ &\dot p_2 = v_2, \quad \dot v_2 = 2p_2\lambda - g, \\ &p_1^2 + p_2^2 = 1 \end{aligned}$$

with the gravity constant $g = 9.81$. The unknowns are $(p_1, p_2, v_1, v_2, \lambda)$. In [15] this problem, together with the boundary conditions

$$v_2(0) = 0, \quad p_1(0.55) = 0, \tag{5.1}$$

was used to test an implementation of a multiple shooting method for DAEs with $\mu = 1$. Since for the above formulation we have $\mu = 2$ together with $d = 2$ and $a = 3$, in [15] it was necessary to replace the constraint by its differentiated form

$$2p_1\dot p_1 + 2p_2\dot p_2 = 2p_1v_1 + 2p_2v_2 = 0$$

and to add a further boundary condition due to the additional dynamics introduced. Here we can solve this problem in its original formulation. Instead of (5.1), we also used the boundary conditions

$$v_2(0) = v_2(2.5) = 0, \tag{5.2}$$

thus seeking a periodic orbit. Observe that we must fix the phase of the solution, since the problem is autonomous.

Starting in both cases with the initial guess

$$\begin{aligned} &x_0 = (1, 0.3, 0, 0, 1), \qquad \dot x_0 = (0, 0, 0, -g, 0), \\ &\ddot x_0 = (0, -g, 0, 0, 0), \qquad x_0^{(3)} = (0, 0, 0, 0, 0) \end{aligned}$$

and using $N = 1$, we obtained solutions according to Table 5.3. The computing times were 3.5 and 14.6 seconds, respectively.

TABLE 5.3
Results for Example 5.3.

Boundary condition (5.1)	
ν	$\|\Delta z_\nu\|_2$
0	0.101D+04
1	0.346D+03
2	0.924D+01
3	0.130D+00
4	0.524D-04
5	0.273D-09

Boundary condition (5.2)	
ν	$\|\Delta z_\nu\|_2$
0	0.542D+02
1	0.403D+02
2	0.208D+02
3	0.183D+01
4	0.341D-02
5	0.120D-08

Example 5.4. In [19], the model of a (two-dimensional) truck is given. It has the form of a standard multibody system

$$\begin{aligned} \dot{p} &= v, \\ M\dot{v} &= f(p, v, u, \dot{u}) - g_p(p)^T \lambda, \\ g(p) &= 0, \end{aligned}$$

where p are the (generalized) positions, v the corresponding velocities, and λ the forces introduced by the constraint $g(p) = 0$. In the truck model, p and v have eleven components and λ is scalar. Hypothesis 2.1 is fulfilled with $\mu = 2$, $d = 20$, and $a = 3$. The (scalar) function u models the road profile and is chosen here to be

$$u(t) = \tau \sin(20\pi t).$$

Asking as in [20] for the periodic response of the system for $\tau = 0.05$, we require the boundary conditions

$$\begin{aligned} p_l(0) &= p_l(0.1), \quad l = 1, \ldots, 9, 11, \\ v_l(0) &= v_l(0.1), \quad l = 1, \ldots, 9, 11. \end{aligned}$$

This problem suffers from an extremely bad scaling and high nonlinearity. Therefore, we applied a (fixed) scaling to get reasonable condition numbers and used classical homotopy according to

$$\tau \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$$

to get the desired solution. The homotopy was started with the equilibrium state for $\tau = 0$. The course of the Gauß–Newton-like procedure for $N = 2$ can be found in Table 5.4. The overall computing time was about 84 minutes, with roughly 10 seconds for the solution of a single initial value problem. In particular, more than 95% of the computing time is used for time integration, as is typical for a multiple shooting approach. In this respect, the efficiency of a multiple shooting code mostly depends on the efficiency of the initial value solver.

Example 5.5. The so-called Lotka–Volterra system is the simplest model for a predator/prey interaction and consists (in normalized form) of the two differential equations

$$\dot{x}_1 = x_1(1 - x_2), \quad \dot{x}_2 = -cx_2(1 - x_1)$$

with some constant $c > 0$. It is well known that the quantity

$$H = c(x_1 - \log x_1) + (x_2 - \log x_2)$$

TABLE 5.4
Values $\|\Delta z_\nu\|_2$ for the homotopy of Example 5.4.

ν	$\tau = 0.01$	$\tau = 0.02$	$\tau = 0.03$	$\tau = 0.04$	$\tau = 0.05$
0	0.440D+04	0.481D+04	0.523D+04	0.527D+04	0.464D+04
1	0.639D+03	0.610D+03	0.610D+03	0.762D+03	0.949D+03
2	0.370D+02	0.442D+02	0.354D+02	0.215D+02	0.132D+02
3	0.110D-01	0.359D-01	0.825D-01	0.614D-01	0.288D-01
4	0.984D-05	0.881D-07	0.421D-05	0.119D-04	0.140D-04
5	—	—	—	0.635D-08	0.133D-07

TABLE 5.5
Results for Example 5.5.

ν	$\|\Delta z_\nu\|_2$
0	0.203D+01
1	0.491D+00
2	0.486D-01
3	0.236D-03
4	0.240D-08

stays constant along every componentwise positive solution and that therefore every such solution is periodic. In order to compute a periodic orbit for a given value of H, we can use the above equation for H as the algebraic constraint for the Lotka–Volterra system. But then the system would be overdetermined. We therefore combine the two differential equations such that the resulting relation defines a flow on the manifold defined by the algebraic constraint. Observing that we must fix the phase of the periodic orbit in order to fix a locally unique solution, we obtain the boundary value problem

$$\begin{aligned}
&(1-x_1)\dot{x}_2 - c(1-x_2)\dot{x}_1 + cx_2(1-x_1)^2 + cx_1(1-x_2)^2 = 0,\\
&c(x_1 - \log x_1) + (x_2 - \log x_2) - H = 0,\\
&x_1(0) = x_1(T), \quad x_1(0) = 1.
\end{aligned}$$

Note that the derivatives $\dot{x}_1$ and $\dot{x}_2$ now have solution dependent factors. Transforming the problem finally to the unit interval and using $x_3 = H$ and $x_4 = T$ as further unknowns, the boundary value problem to solve reads as

$$\begin{aligned}
&(1-x_1)\dot{x}_2 - c(1-x_2)\dot{x}_1 + cx_2(1-x_1)^2x_4 + cx_1(1-x_2)^2x_4 = 0,\\
&c(x_1 - \log x_1) + (x_2 - \log x_2) - x_3 = 0, \quad \dot{x}_3 = 0, \quad \dot{x}_4 = 0,\\
&x_1(0) = x_1(1), \quad x_1(0) = 1, \quad x_3(0) = H.
\end{aligned}$$

It has differentiation index one and satisfies Hypothesis 2.1 with $\mu = 0$, $d = 3$, and $a = 1$. Starting with

$$x_0 = (1.0, 0.6, 2.1, 6.0), \quad \dot{x}_0 = (3.2, 0, 0, 0)$$

for the choice $c = 1$, $H = 2.2$ and using $N = 1$, we successfully obtained the solution which has a period $T = 6.4943$. The computation took about 0.8 seconds, and the convergence behavior is reported in Table 5.5.

In summary, we have demonstrated that the multiple shooting method presented is able to solve problems with different values of the index and different structures. The

tables show that the Gauß–Newton-like method developed has very good convergence properties. Indeed, for the examples presented they cannot be distinguished from quadratic convergence.

6. Conclusions. We have presented a multiple shooting approach for the solution of nonlinear boundary value problems for differential-algebraic systems of arbitrary index and without special structure requirements. Using a specific Gauß–Newton-like method for the solution of the nonlinear system of boundary and continuity conditions we have proved superlinear convergence of the method. We have implemented the method on the basis of a new general solver for DAEs of arbitrary index [13] and demonstrated the numerical properties of the method for several examples.

Appendix. Proof of Remark 3.2. In the linear case, the derivative array equations (2.1) have the form

$$M_\ell(t) \begin{bmatrix} \dot{x} \\ \vdots \\ x^{\ell+1} \end{bmatrix} = N_\ell(t)x + g_\ell(t)$$

with

$$M_\ell : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\ell+1)n,(\ell+1)n}, \quad N_\ell : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\ell+1)n,n}, \quad g_\ell : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\ell+1)n}.$$

Defining

$$\begin{aligned} E_1(t) = Z_1(t)^T M_0(t), \quad A_1(t) = Z_1(t)^T N_0(t), \quad f_1(t) = Z_1(t)^T g_0(t), \\ A_2(t) = Z_2(t)^T N_\mu(t), \quad f_2(t) = Z_2(t)^T g_\mu(t), \end{aligned}$$

with Z_1, Z_2 as in the nonlinear case, the given solution x^* satisfies the linear DAE

$$\begin{bmatrix} E_1(t) \\ 0 \end{bmatrix} \dot{x} = \begin{bmatrix} A_1(t) \\ A_2(t) \end{bmatrix} x + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix},$$

which has vanishing strangeness index; see [8, 9]. Following the theory presented there, pointwise nonsingular (smooth) matrix functions $P, Q : [\underline{t}, \overline{t}] \to \mathbb{R}^{n,n}$ exist such that

$$P(t) \begin{bmatrix} E_1(t) \\ 0 \end{bmatrix} Q(t) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix},$$

$$P(t) \begin{bmatrix} A_1(t) \\ A_2(t) \end{bmatrix} Q(t) - P(t) \begin{bmatrix} E_1(t) \\ 0 \end{bmatrix} \dot{Q}(t) = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix},$$

transforming the DAE to

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \dot{\tilde{x}} = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \tilde{x} + \begin{bmatrix} \tilde{f}_1(t) \\ \tilde{f}_2(t) \end{bmatrix}$$

with

$$x(t) = Q(t)\tilde{x}(t) = Q(t) \begin{bmatrix} \tilde{x}_1(t) \\ \tilde{x}_2(t) \end{bmatrix}, \quad \begin{bmatrix} \tilde{f}_1(t) \\ \tilde{f}_2(t) \end{bmatrix} = P(t) \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}.$$

Hence, all solutions have the form

$$x(t) = Q(t) \begin{bmatrix} \tilde{x}_1(\underline{t}) + \int_{\underline{t}}^{t} \tilde{f}_1(s)\, ds \\ \\ -\tilde{f}_2(t) \end{bmatrix},$$

implying that

$$\Phi_{x,y}(x,y) = Q(\bar{t}) \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q(\underline{t})^{-1} & 0 \end{bmatrix}.$$

The equations (3.1) that define S take the form

$$M_\mu(\underline{t})\hat{y} - N_\mu(\underline{t})\hat{x} - g_\mu(\underline{t}) = 0, \quad T_2(\underline{t})^T(\hat{x} - x) = 0, \quad T_1(\underline{t})^T(\hat{y} - y) = 0.$$

Multiplying the first relation with $Z_2(\underline{t})^T$, we obtain

$$Z_2(\underline{t})^T N_\mu(\underline{t})\hat{x} = -Z_2(\underline{t})^T g_\mu(\underline{t})$$

and therefore (recalling the definition of A_2 and f_2)

$$\hat{x} = \begin{bmatrix} A_2(\underline{t}) \\ T_2(\underline{t})^T \end{bmatrix}^{-1} \begin{bmatrix} -f_2(\underline{t}) \\ T_2(\underline{t})^T x \end{bmatrix}.$$

A similar argument yields

$$\hat{y} = \begin{bmatrix} Z_2'(\underline{t})^T M_\mu(\underline{t}) \\ T_1(\underline{t})^T \end{bmatrix}^{-1} \begin{bmatrix} Z_2'(\underline{t})^T N_\mu(\underline{t})\hat{x} + Z_2'(\underline{t})^T g_\mu(\underline{t}) \\ T_1(\underline{t})^T y \end{bmatrix}.$$

Altogether, we have

$$\Phi_{x,y}(x,y) S_x(x,y) = Q(\bar{t}) \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q(\underline{t})^{-1} \begin{bmatrix} A_2(\underline{t}) \\ T_2(\underline{t})^T \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ T_2(\underline{t})^T \end{bmatrix}$$

and therefore

$$\mathcal{E} = C T_2(\underline{t}) + D Q(\bar{t}) \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q(\underline{t})^{-1} \begin{bmatrix} A_2(\underline{t}) \\ T_2(\underline{t})^T \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

Following [20], the transformation P has the block structure

$$P(t) = \begin{bmatrix} P_{11}(t) & P_{12}(t) \\ 0 & P_{22}(t) \end{bmatrix},$$

where $P_{22} : [\underline{t}, \bar{t}] \to \mathbb{R}^{a,a}$ is pointwise nonsingular. Hence,

$$P_{22}(\underline{t}) A_2(\underline{t}) Q(\underline{t}) = [\, 0 \; I \,]$$

and $Q(\underline{t})$ has the form

$$Q(\underline{t}) = [\, T_2(\underline{t}) U \; * \,]$$

with some nonsingular matrix $U \in \mathbb{R}^{d,d}$. In particular,

$$Q(\underline{t}) \begin{bmatrix} U^{-1} \\ 0 \end{bmatrix} = T_2(\underline{t}).$$

Defining $[\,C_{11}\ C_{12}\,] = CQ(\underline{t})$, $[\,D_{11}\ D_{12}\,] = DQ(\bar{t})$, and using

$$\begin{bmatrix} A_2(\underline{t}) \\ T_2(\underline{t})^T \end{bmatrix} T_2(\underline{t}) = \begin{bmatrix} 0 \\ I \end{bmatrix},$$

we find

$$\begin{aligned} \mathcal{E} &= CQ(\underline{t})Q(\underline{t})^{-1}T_2(\underline{t}) + DQ(\bar{t}) \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q(\underline{t})^{-1}T_2(\underline{t}) \\ &= [\,C_{11}\ C_{22}\,] \begin{bmatrix} U^{-1} \\ 0 \end{bmatrix} + [\,D_{11}\ D_{22}\,] \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U^{-1} \\ 0 \end{bmatrix} \\ &= (C_{11} + D_{11})U^{-1}. \end{aligned}$$

Thus, $\mathcal{E}$ is nonsingular if and only if $C_{11} + D_{11}$ is nonsingular.

REFERENCES

[1] U. M. Ascher, R. Mattheij, and R. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, 2nd ed., SIAM, Philadelphia, 1995.

[2] S. L. Campbell, *A general form for solvable linear time varying singular systems of differential equations*, SIAM J. Math. Anal., 18 (1987), pp. 1101–1115.

[3] S. L. Campbell and E. Griepentrog, *Solvability of general differential algebraic equations*, SIAM J. Sci. Comput., 16 (1995), pp. 257–270.

[4] P. Deuflhard and G. Heindl, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal., 16 (1979), pp. 1–10.

[5] R. England, R. Lamour, and J. Lopez-Estrada, *Multiple shooting using a dichotomically stable integrator for solving differential-algebraic equations*, Appl. Numer. Math., 42 (2002), pp. 117–131.

[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, 1996.

[7] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II, Springer-Verlag, Berlin, 1991.

[8] P. Kunkel and V. Mehrmann, *Canonical forms for linear differential-algebraic equations with variable coefficients*, J. Comput. Appl. Math., 56 (1994), pp. 225–251.

[9] P. Kunkel and V. Mehrmann, *A new class of discretization methods for the solution of linear differential-algebraic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1941–1961.

[10] P. Kunkel and V. Mehrmann, *Regular solutions of nonlinear differential-algebraic equations and their numerical determination*, Numer. Math., 79 (1998), pp. 581–600.

[11] P. Kunkel and V. Mehrmann, *Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems*, Math. Control Signals Systems, 14 (2001), pp. 233–256.

[12] P. Kunkel, V. Mehrmann, W. Rath, and J. Weickert, *A new software package for linear differential-algebraic equations*, SIAM J. Sci. Comput., 18 (1997), pp. 115–138.

[13] P. Kunkel, V. Mehrmann, and I. Seufer, *GENDA: A Software Package for the Solution of GEneral Nonlinear Differential-Algebraic Equations*, Technical report 730-02, Institut für Mathematik, TU Berlin, D-10623 Berlin, Germany, 2002.

[14] P. Kunkel and R. Stöver, *Symmetric collocation methods for linear differential-algebraic boundary value problems*, Numer. Math., 91 (2002), pp. 475–501.

[15] R. Lamour, *A shooting method for fully implicit index-2 differential-algebraic equations*, SIAM J. Sci. Comput., 18 (1997), pp. 94–114.

[16] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, London, 1970.

[17] T. Schmidt and M. Hou, *Rollringgetriebe*, Internal Report, Sicherheitstechnische Regelungs- und Meßtechnik, Bergische Universität, GH Wuppertal, Wuppertal, Germany, 1992.

[18] V. H. Schulz, H. G. Bock, and M. C. Steinbach, *Exploiting invariants in the numerical solution of multipoint boundary value problems for DAE*, SIAM J. Sci. Comput., 19 (1998), pp. 440–467.

[19] B. Simeon, F. Grupp, C. Führer, and P. Rentrop, *A nonlinear truck model and its treatment as a multibody system*, J. Comput. Appl. Math., 50 (1994), pp. 523–532.

[20] R. Stöver, *Numerische Lösung von linearen differential-algebraischen Randwertproblemen*, Logos Verlag, Berlin, 1999
[21] R. Stöver, *Collocation methods for solving linear differential-algebraic boundary value problems*, Numer. Math., 88 (2001), pp. 771–795.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2298–2319

HIGHER-ORDER FOURIER APPROXIMATION IN SCATTERING BY TWO-DIMENSIONAL, INHOMOGENEOUS MEDIA*

OSCAR P. BRUNO† AND E. MCKAY HYDE‡

Abstract. This paper provides a theoretical analysis of a higher-order, FFT-based integral equation method introduced recently [*IEEE Trans. Antennas and Propagation*, 48 (2000), pp. 1862–1864] for the evaluation of transverse electric–polarized electromagnetic scattering from a bounded, penetrable inhomogeneity in two-dimensional space. Roughly speaking, this method is based on Fourier smoothing of the integral operator and the refractive index $n(x)$. Here we *prove* that the solution of the resulting integral equation approximates the solution of the exact integral equation with higher-order accuracy, *even when* $n(x)$ *is a discontinuous function*—as suggested by the numerical experiments contained in the paper mentioned above. In detail, we relate the convergence rates of the computed interior and exterior fields to the regularity of the scatterer, and we demonstrate, with a few numerical examples, that the predicted convergence rates are achieved in practice.

Key words. Helmholtz equation, Lippmann–Schwinger integral equation, transverse electric scattering, TM scattering, fast Fourier transform

AMS subject classifications. 35J05, 65N12, 65R20, 78A45

DOI. 10.1137/S0036142903425811

1. Introduction.

Scattering problems find application in a wide range of fields, including communications, materials science, plasma physics, biology, medicine, radar, and remote sensing. The evaluation of useful numerical solutions for scattering problems remains a highly challenging problem, requiring novel mathematical approaches and powerful computational tools. An integral equation method [7, 8] introduced recently for the evaluation of time-harmonic, transverse electric (TE)–polarized, electromagnetic scattering by bounded inhomogeneities in two dimensions has proven highly competitive with currently available approaches. (Note that there is some ambiguity in the naming of the polarization [28, p. R5], with some authors referring to this setting as transverse magnetic (TM)–polarized scattering. To be precise, we consider the case in which the electric field is parallel to the cylindrical axis of the scatterer.) In this paper, we provide a theoretical analysis of the higher-order convergence of this approach. More specifically, we *prove* that the approximating integral equation used in this method, which is based on Fourier approximation of the integral

*Received by the editors April 7, 2003; accepted for publication (in revised form) March 25, 2004; published electronically March 31, 2005. This effort was sponsored in part by the Air Force Office of Scientific Research (AFOSR), Air Force Materials Command, USAF, under the AASERT award F49620-98-1-0368. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

http://www.siam.org/journals/sinum/42-6/42581.html

†Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA 91125 (bruno@acm.caltech.edu). The research of this author was supported by AFOSR grants F49620-96-1-0008, F49620-99-1-0010, and F49620-02-1-0049; the NSF through the NYI award DMS-9596152 and through contracts DMS-9523292, DMS-9816802, and DMS-0104531; and the Powell Research Foundation.

‡Computational and Applied Mathematics - MS 134, Rice University, 6100 Main St., Houston, TX 77005-1892 (hyde@caam.rice.edu). The research of this author was supported by a DOE Computational Science Graduate Fellowship, an Achievement Rewards for College Scientists (ARCS) Fellowship, and an NSF Mathematical Sciences Postdoctoral Research Fellowship.

operator, yields higher-order convergence in the L^∞-norm even when the refractive index $n(x)$ is a discontinuous function. Furthermore, we relate the convergence rates of the computed interior and exterior fields to the regularity of the scatterer, and we demonstrate, with a few numerical examples, that the predicted convergence rates are achieved in practice.

Given an incident field u^i, we denote by u the total electric field—which equals the sum of u^i and the resulting scattered field u^s:

$$u = u^i + u^s. \tag{1.1}$$

Calling λ the wavelength of the incident field and $\kappa = \frac{2\pi}{\lambda}$ the wavenumber, the total field u satisfies [9, p. 2]

$$\Delta u + \kappa^2 n^2(x) u = 0, \quad x \in \mathbb{R}^3, \tag{1.2}$$

where the given incident field u^i is assumed to satisfy

$$\Delta u^i + \kappa^2 u^i = 0, \quad x \in \mathbb{R}^3. \tag{1.3}$$

Finally, to guarantee that the scattered wave is outgoing, u^s is required to satisfy the Sommerfeld radiation condition [9, p. 67]

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial u^s}{\partial r} - i\kappa u^s\right) = 0. \tag{1.4}$$

The algorithms available for computing solutions to this problem fall into two broad classes: (1) finite element and finite difference methods and (2) integral equation methods. Use of finite element and finite difference methods can be advantageous in that, unlike other methods, they lead to sparse linear systems. Their primary disadvantage, on the other hand, lies in the fact that in order to satisfy the Sommerfeld radiation condition (1.4), a relatively large computational domain containing the scatterer must be used, together with appropriate absorbing boundary conditions on the boundary of the computational domain (see, for example, [10, 17, 18, 26, 32]). Thus, these procedures give rise to very large numbers of unknowns and, thus, to very large linear systems.

A second class of algorithms is based on the use of integral equations. An appropriate integral formulation for our two-dimensional TE problem is given by the Lippmann–Schwinger integral equation [9, p. 214], [24],

$$u(x) = u^i(x) - \kappa^2 \int g(x-y) m(y) u(y) dy, \tag{1.5}$$

where $g(x) = \frac{i}{4} H_0^1(\kappa|x|)$ is the fundamental solution of the Helmholtz equation in two dimensions and m is the *compactly supported* function $m = 1 - n^2$. Integral equation approaches are advantageous in a number of ways: they require only discretization of the equation *on the scatterer itself*, and the solutions they produce satisfy the radiation condition at infinity *automatically*. Direct use of integral equation methods is costly, however, since they lead to dense linear systems: a straightforward computation of the required convolution requires $\mathcal{O}(N^2)$ operations per iteration of an iterative linear solver. As mentioned above, however, the *higher-order* integral method that we analyze in this paper, in which the complexity of the convolution evaluation

is reduced to $\mathcal{O}(N \log N)$ operations per iteration, is highly competitive with finite element or finite difference approaches.

Fast solvers for (1.5), based on the fast Fourier transform (FFT), have been available for some time [3, 31, 33]. In these solvers, the convolution with the fundamental solution is computed via Fourier transforms, which can, in turn, be evaluated with low complexity by means of FFTs. These methods do give rise to a reduced complexity for a given discretization but, unfortunately, they are only first-order accurate for discontinuous scatterers. Low-order accuracy results since, for a general nonsmooth and/or nonperiodic function, the FFT provides a poor approximation to the Fourier transform. Our approach also uses FFTs to achieve a reduced complexity but, unlike previous FFT methods, it yields, in addition, *higher-order accuracy*.

Despite the significant advantages exhibited by higher-order methods over their low-order counterparts (see, for example, Appendix B), only limited attempts have been made to develop higher-order methods for the problem under consideration. A higher-order method was proposed in [23] on the basis of a locally corrected Nyström discretization; the complexity of this method, however, is $\mathcal{O}(N^2)$, where N is the total number of unknowns used.

In [27], Vainikko presents two $\mathcal{O}(N \log N)$ methods for solving (1.5). The first applies to m in the Sobolev space $W^{\mu,2}$ and yields $\mathcal{O}(h^\mu)$ L^2-convergence in the near and the far fields. We instead consider piecewise-smooth $m \in C^{k,\alpha}$ (which are arguably the appropriate spaces for scatterers arising in practice). (For the precise definition of the function spaces that we consider, see Definitions 2.4 and 2.5.) In comparing Vainikko's approach with our method, note that a piecewise-smooth $m \in C^{0,\alpha}$ which does not belong to C^1 can, at best, belong to $W^{2,2}$ [22, p. 197], [11, p. 194], for which Vainikko's method predicts $\mathcal{O}(h^2)$ L^∞-convergence in both the near and the far field. Our method, on the other hand, achieves $\mathcal{O}(h^3)$ and $\mathcal{O}(h^5)$ L^∞-convergence in the near and far fields, respectively (see section 3.2 for our convergence results). More generally, a piecewise-smooth $m \in C^{k,\alpha}$ for $k \geq 1$ which does not belong to C^{k+1} can, at best, belong to $W^{k+2,2}$, for which Vainikko's result predicts $\mathcal{O}(h^{k+2})$ L^∞-convergence in both the near and far field, whereas our method achieves $\mathcal{O}(h^{k+3})$ L^∞-convergence in the near field and $\mathcal{O}(h^{k+6})$ L^∞-convergence in the far field.

The second method proposed in [27] applies to piecewise-smooth (possibly discontinuous) m and yields $\mathcal{O}(h^2(1+|\log h|))$ L^∞-convergence in the near and far fields. (This method requires evaluation of the volume fraction of each discretization cell on each side of a discontinuity in $m = 1 - n^2$.) For such inhomogeneities, our method yields $\mathcal{O}(h^2)$ and $\mathcal{O}(h^3)$ L^∞-convergence in the near and far fields, respectively. Thus, our approach, which applies to smooth as well as discontinuous refractive indices, is both fast—it runs in $\mathcal{O}(N \log N)$ operations—and higher-order accurate, substantially exceeding the convergence rates of Vainikko's approach, especially in the far field.

Our method is based on recasting the last term of the integral equation (1.5) by means of the polar coordinate form

$$(Ku)(a,\phi) = -\kappa^2 \int g(a,\phi;r,\theta) m(r,\theta) u(r,\theta) r \, dr \, d\theta. \tag{1.6}$$

An approximate integral equation is obtained from (1.6) by replacing the kernel g by a truncation of its Fourier representation with respect to its angular variables—which, as is known, is given by the addition theorem for the Hankel function; see section 2. As we show in this paper, the solution of this approximate integral equation approximates the solution of the exact integral equation with higher-order accuracy, even for discontinuous functions $n(x)$.

The higher-order convergence of this method relies on the following important fact: although the Fourier representation of the fundamental solution converges slowly, the resulting Fourier representation of the integral converges *rapidly*; clearly, such accuracy improvements for integrated quantities can only occur through a process of error cancellation. In this paper, we prove that this approach does indeed yield higher-order convergence (at least third-order in the exterior field) even in the case of discontinuous inhomogeneities. More precisely, we derive bounds on the convergence rates for the interior and exterior fields as they depend on the regularity of the scatterer (see Theorem 3.5 and Corollaries 3.9 and 3.10).

Our present analysis considers neither a specific numerical discretization for the radial integration nor the method used to solve the resulting linear system. Here we focus instead on the exact solution of the approximate integral equation resulting from the polar Fourier approximation of the fundamental solution, as described briefly above and in detail in section 2; this exact solution of the approximate equation is to be viewed as an approximate solution of the exact equation (1.5). The details of the complete numerical implementation are given in their original form in [7, 8] as well as in the more recent presentations [5, 14], which contain several significant improvements.

As discussed in section 4, our approximate integral formulation allows us to replace the (possibly discontinuous) function n in polar coordinates by its truncated Fourier series of certain orders *without introducing additional errors*. This fact allows us to compute the corresponding angular integrals *exactly* by means of FFTs. (In [14, 15, 16], similar ideas are used in the construction of a fast, higher-order method for the Helmholtz equation in three dimensions.) To conclude this paper we present a number of computational examples that demonstrate that the predicted convergence rates are achieved in practice.

(Note that a direct application of the methods presented in this paper to discontinuous scatterers for either TM or three-dimensional electromagnetic scattering would yield rates of convergence lower than those for the TE case considered here—since in such cases the normal derivatives of the solution are not continuous across surfaces of discontinuity of the refractive index. As shown in [6], however, the convergence rates of our method for all of these problems—TE, TM, and three-dimensional electromagnetic scattering—can be improved significantly by appropriate treatment of thin volumetric regions around surfaces where either discontinuities or reduced regularity of the refractive index occur.)

2. An approximate integral equation. As mentioned in the introduction, our approach produces numerical solutions of (1.5) through consideration of a sequence of approximate integral equations, which result as the fundamental solution is replaced by a truncated Fourier series in an angular variable. In this section we describe our approximate integral equations, and we show that (1) they admit unique solutions and (2) the inverse operators for the approximate problems are uniformly bounded.

To introduce our approximate integral equations we begin by recalling an addition theorem: using polar coordinates $x = ae^{i\phi}$ and $y = re^{i\theta}$, the addition theorem for the Hankel function reads [9, p. 67]

$$H_0^1(\kappa|ae^{i\phi} - re^{i\theta}|) = \sum_{\ell=-\infty}^{\infty} \mathcal{J}_\ell(a, r)e^{i\ell(\phi-\theta)},$$

where, calling J_ℓ and H^1_ℓ the Bessel and Hankel functions of order ℓ, we have denoted

$$\mathcal{J}_\ell(a,r) = H^1_\ell(\kappa \max(a,r))\ J_\ell(\kappa \min(a,r)). \tag{2.1}$$

This identity allows us to obtain another expression for the integral operator K of (1.5),

$$(Ku)(a,\phi) = -\frac{i\kappa^2}{4}\int H^1_0(\kappa|x-y|)m(y)u(y)dy = \sum_{\ell=-\infty}^{\infty}(K_\ell u)(a)e^{i\ell\phi},$$

where, using an annular region $R_0 \le a \le R_1$ containing the support of m, we have set

$$(K_\ell u)(a) = -\frac{i\kappa^2}{4}\int_{R_0}^{R_1}\mathcal{J}_\ell(a,r)\left[\int_0^{2\pi} m(r,\theta)u(r,\theta)e^{-i\ell\theta}d\theta\right] r\,dr. \tag{2.2}$$

Truncating this Fourier series as well as the corresponding Fourier series for the incident field, we obtain the approximate integral equation

$$v(a,\phi) = u^{i,M}(a,\phi) + (K^M v)(a,\phi), \tag{2.3}$$

where

$$u^{i,M}(a,\phi) = \sum_{\ell=-M}^{M} u^i_\ell(a)e^{i\ell\phi}, \tag{2.4}$$

$$(K^M v)(a,\phi) = \sum_{\ell=-M}^{M}(K_\ell v)(a)e^{i\ell\phi}. \tag{2.5}$$

Here and throughout this paper we use a superscript M to denote the truncated Fourier series of order M of a given function.

Decomposing (2.3) into Fourier modes, we observe that a solution of this equation must satisfy

$$v_\ell(a) = \begin{cases} u^i_\ell(a) + (K_\ell v)(a) & \text{for } |\ell| \le M, \\ 0 & \text{for } |\ell| > M. \end{cases} \tag{2.6}$$

Hence,

$$v(a,\phi) = v^M(a,\phi)$$

and solving (2.3) is equivalent to solving the following system of one-dimensional integral equations:

$$v_\ell(a) - (K_\ell v^M)(a) = u^i_\ell(a), \quad \ell = -M,\dots,M. \tag{2.7}$$

To prove existence and uniqueness for this approximate integral equation, we make use of the following technical lemma.

LEMMA 2.1. *There exists a constant $C > 0$ depending only on R_0, R_1, and κ such that*

$$\left\| \int_{R_0}^{R_1} |\mathcal{J}_\ell(a,r)| r\,dr \right\|_\infty \le \frac{C}{\ell^2},$$

where $\mathcal{J}_\ell(a,r)$ is defined in (2.1).

This result, which is proven in Appendix A, allows us to establish the following lemma. (Note: In the bound above and in all similar bounds in this paper, we abuse the notation slightly for $\ell = 0$, in which case the expression on the left-hand side is assumed to be bounded.)

LEMMA 2.2. *For any $m \in L^\infty$,*

$$\|K - K^M\|_\infty \to 0$$

as $M \to \infty$, where the operator norm is the one induced by the L^∞-norm.

Proof. Let $u \in L^\infty$. Then

$$\int_0^{2\pi} |m(r,\theta)u(r,\theta)e^{-i\ell\theta}|d\theta \le 2\pi\|m\|_\infty\|u\|_\infty.$$

Hence, for $M \ge 0$,

$$\begin{aligned}\|(K - K^M)u\|_\infty &\le \frac{\pi\kappa^2}{2}\|m\|_\infty\|u\|_\infty \sum_{|\ell|>M} \left\| \int_{R_0}^{R_1} |\mathcal{J}_\ell(a,r)|r\,dr \right\|_\infty \\ &= \mathcal{O}\left(\sum_{|\ell|>M} \frac{1}{\ell^2} \right) \|u\|_\infty \\ &= \mathcal{O}(M^{-1})\|u\|_\infty.\end{aligned}$$

Therefore, $\|K - K^M\|_\infty = \mathcal{O}(M^{-1}) \to 0$ as $M \to \infty$. ∎

Remark 2.3. The following proof of the existence and uniform boundedness of $(I-K^M)^{-1}$ depends crucially on the existence and boundedness of $(I-K)^{-1}$. By the Riesz–Fredholm theory (see, for example, [20, p. 29]), since K is a compact operator on L^∞, $(I-K)^{-1}$ exists and is bounded if $I-K$ is injective. The injectivity of this operator is equivalent to the uniqueness of solutions of the corresponding Helmholtz equation (1.2). The uniqueness result relevant for our setting follows from corresponding (more general) results for acoustic scattering proved in [30] under assumptions that we state more precisely below with the help of the following definitions.

DEFINITION 2.4. *Given a compact set $D \subset \mathbb{R}^n$, we say that a function f has piecewise continuous derivatives of order k on D, denoted by $f \in C^k_{pw}(D)$, if and only if there exist a finite number of open, disjoint subsets of D, denoted by $D_1, D_2, \dots, D_p$, such that $D = \bigcup_{i=1}^p \overline{D_i}$ and there exist functions $f_i \in C^k(\overline{D_i})$ such that $f|_{D_i} = f_i|_{D_i}$. In an entirely analogous fashion we define spaces of functions with piecewise-Hölder continuous derivatives of order k on D, denoted by $C^{k,\alpha}_{pw}(D)$.*

DEFINITION 2.5. *We say that the scattering inhomogeneity m belongs to $\mathcal{M}$ if and only if* (1) *$m \in C^{0,\alpha}_{pw}(D)$ for some compact set D that properly contains the support of m and* (2) *each of the corresponding subsets $D_1, D_2, \dots, D_p$, as defined in Definition* 2.4*, has a Lipschitz boundary.*

Remark 2.6. With these definitions, we can state the unique solvability result for (1.5), which is based on the uniqueness result of [30], more precisely: $I-K$ admits a bounded inverse on L^∞ for each $m \in \mathcal{M}$. Hence, throughout this paper, we will assume that $m \in \mathcal{M}$. Note that the uniqueness result of [30] makes use of a unique continuation result due to Heinz [13], which assumes C^1 boundary regularity of the subsets D_i defined above. However, more recent unique continuation results make much weaker assumptions (see [19] and the references therein) and hence allow us

to relax the C^1 regularity assumption to Lipschitz regularity (which suffices to allow integration by parts in obtaining the appropriate weak formulation).

We can now establish the following theorem.

THEOREM 2.7. *Given $m \in \mathcal{M}$, for M sufficiently large the operators $(I-K^M)^{-1}$ exist on L^∞ and are uniformly bounded. Thus, given any incident field u^i, (2.3) admits a unique solution $v \in L^\infty$ for all M sufficiently large.*

Proof. Since, by the discussion above, $I-K$ has a bounded inverse, Lemma 2.2 and [20, Theorem 10.1, p. 142] imply that for all sufficiently large M the inverse operators $(I-K^M)^{-1}$ exist and are uniformly bounded. ∎

3. Error bounds. The approximate integral equation (2.3) was obtained by truncating the Fourier series of both the incident field u^i and the integral operator K at each radius; as mentioned above, the exact solution v of this approximate equation is to be viewed as an approximate solution of the exact equation (1.5). As it happens, the function v is a *higher-order approximation* of the exact solution u of (1.5). Roughly speaking, this result follows from the fact that the integral operator Ku and the incident field u^i are smooth and periodic functions of the angular variable, which are thus approximated to higher-order by their truncated Fourier series.

In this section we derive bounds on the error implicit in the approximation of u by v. Of course the full numerical implementation of the method introduces additional errors (e.g., errors arising from radial numerical quadratures), but here we study the accuracy with which v approximates the exact solution u only. Higher-order methods for computing the required radial integrals are discussed in [5, 7, 8, 14].

3.1. Error in approximated Fourier modes. The error in the solution v^M of the approximate integral equation (2.3) at a point $x=(a,\phi)\in\mathbb{R}^2$ is given by

$$|u(x)-v^M(x)| \le |(u-u^M)(x)| + |u^M(x)-v^M(x)|, \tag{3.1}$$

where $(u-u^M)$ is the "tail" of the Fourier series of u,

$$(u-u^M)(a,\phi) = \sum_{|\ell|>M} u_\ell(a)e^{i\ell\phi}.$$

In this section, we derive a bound on the second term on the right-hand side of (3.1).

Subtracting the identities (see (2.6))

$$\begin{aligned} u^M &= u^{i,M} + K^M u, \\ v^M &= u^{i,M} + K^M v^M, \end{aligned}$$

we obtain

$$\begin{aligned} u^M - v^M &= K^M(u-v^M) \\ &= K^M(u^M-v^M) + K^M(u-u^M). \end{aligned}$$

In view of Theorem 2.7 and calling

$$\varepsilon_M = \|u^M - v^M\|_\infty, \tag{3.2}$$

we obtain

$$\varepsilon_M \le B\|K^M(u-u^M)\|_\infty$$

for sufficiently large M, where B is a uniform bound on $\|(I - K^M)^{-1}\|$.

To bound $K^M(u - u^M)$, we note that

$$\|K^M(u - u^M)\|_\infty \leq \sum_{\ell=-M}^{M} \|K_\ell(u - u^M)\|_\infty$$

and

$$\int_0^{2\pi} m(r,\theta)(u - u^M)(r,\theta)e^{-i\ell\theta}d\theta = 2\pi \sum_{|j|>M} m_{\ell-j}(r)u_j(r).$$

Therefore, by Lemma 2.1,

$$\|K_\ell(u - u^M)\|_\infty \leq \frac{C}{\ell^2} \sum_{|j|>M} \|m_{\ell-j}\|_\infty \|u_j\|_\infty. \tag{3.3}$$

We will bound this expression through consideration of bounds on the Fourier coefficients of m and u. To this end, we make use of the following lemma, which is a slight variation of a classical result [34, pp. 48, 71] and can be proved by multiple integrations by parts.

LEMMA 3.1. *If g is a 2π-periodic function such that $g \in C^k([0, 2\pi])$ with $g^{(n)}(0) = g^{(n)}(2\pi)$ for $n = 0, \ldots, k$, and $g^{(k+1)}$ is of bounded variation, then the Fourier coefficients c_ℓ of g satisfy $|c_\ell| \leq C|\ell|^{-(k+2)}$ for some constant C. If $g^{(1)}$ is of bounded variation on $[0, 2\pi]$, then $|c_\ell| \leq C|\ell|^{-1}$.*

The following useful theorem describes the dependence of the regularity of u on the regularity of m. Variations on the results for the Newtonian potential (see [2, p. 223], [11, pp. 78–80], and [12, pp. 53, 56]) give us the following result.

THEOREM 3.2. *Let D be an open set which properly contains the compact support of $m \in \mathcal{M}$, and let u be the solution of (1.5) on D for a given incident field u^i. Then $u \in C^{1,\alpha}(D)$. Furthermore, if Ω is an open subset of D and $m \in C^{k,\alpha}(\Omega)$, then $u \in C^{k+2,\alpha}(\Omega)$.*

Remark 3.3. Since Ω is an arbitrary bounded, open set, this theorem relates the *local* regularity of u to the *local* regularity of m.

To bound the discrete convolution in (3.3) we also need results on the decay rates of the Fourier coefficients of m and u.

LEMMA 3.4. *Let $m \in \mathcal{M}$. Define the annular region $A = \{(a, \phi) : 0 \leq R_0 \leq a \leq R_1\}$ such that A properly contains the support of m. If $m \in C^{k,\alpha}(A) \cap C^{k+2}_{pw}(A)$ for $k \geq 0$, then the Fourier coefficients of the total field u satisfy*

$$\|u_\ell\|_\infty \leq \frac{C}{|\ell|^{k+4}}.$$

If $m \in C^1_{pw}(A)$, then the Fourier coefficients of the total field u satisfy

$$\|u_\ell\|_\infty \leq \frac{C}{|\ell|^3}.$$

Proof. From (2.2), we see that the coefficients in the Fourier series representation of (1.5) are given by

$$u_\ell(a) = u^i_\ell(a) - \frac{i\pi\kappa^2}{2} \int_{R_0}^{R_1} \mathcal{J}_\ell(a, r)(mu)_\ell(r)r\, dr.$$

Since $m \in C^{k,\alpha}(A) \cap C^{k+2}_{pw}(A)$, Theorem 3.2 implies that $u \in C^{k+2,\alpha}(A)$ and hence, $mu \in C^{k,\alpha}(A) \cap C^{k+2}_{pw}(A)$. Therefore, by Lemma 3.1, the Fourier coefficients of mu satisfy

$$\|(mu)_\ell\|_\infty \leq \frac{C_1}{|\ell|^{k+2}}.$$

Again by Theorem 3.2, since u^i solves the homogeneous Helmholtz equation (1.3), $u^i \in C^\infty(\mathbb{R}^2)$. Thus, the Fourier coefficients u^i_ℓ decay faster than $|\ell|^{-p}$ for any positive integer p as $\ell \to \infty$. Therefore, by Lemma 2.1, we obtain

$$\begin{aligned}\|u_\ell\|_\infty &\leq \frac{C_2}{\ell^2}\frac{C_1}{|\ell|^{k+2}} \\ &\leq \frac{C}{|\ell|^{k+4}}.\end{aligned}$$

The proof for $m \in C^1_{pw}$ is similar. ∎

We can now establish the main result of this paper.

THEOREM 3.5. *Let $m \in \mathcal{M}$. Define the annular region $A = \{(a,\phi) : 0 \leq R_0 \leq a \leq R_1\}$ such that A properly contains the support of m.*

If $m \in C^1_{pw}(A)$, then as $M \to \infty$

$$\varepsilon_M = \|u^M - v^M\| \leq B\|K^M(u - u^M)\| = \mathcal{O}\left(\frac{1}{M^3}\right).$$

If $m \in C^{0,\alpha}(A) \cap C^2_{pw}(A)$, then as $M \to \infty$

$$\varepsilon_M = \mathcal{O}\left(\frac{1}{M^5}\right).$$

If $m \in C^{k,\alpha}(A) \cap C^{k+2}_{pw}(A)$ for $k \geq 1$, then as $M \to \infty$

$$\varepsilon_M = \mathcal{O}\left(\frac{1}{M^{k+6}}\right).$$

Proof. We seek a bound on $\|K^M(u - u^M)\|_\infty \leq \sum_{\ell=-M}^{M} \|K_\ell(u - u^M)\|_\infty$. By Lemma 3.4, we obtain

$$\begin{aligned}\|K_\ell(u - u^M)\|_\infty &\leq \frac{C_1}{\ell^2} \sum_{|j|>M} \frac{1}{|\ell - j|^{k+2}} \frac{1}{|j|^{k+4}} \\ &= \frac{C_1}{\ell^2} \sum_{j>M} \frac{1}{j^{k+4}} \left(\frac{1}{(j-\ell)^{k+2}} + \frac{1}{(j+\ell)^{k+2}}\right) \\ &\leq \frac{2C_1}{\ell^2} \sum_{j>M} \frac{1}{j^{k+4}} \frac{1}{(j - |\ell|)^{k+2}}\end{aligned}$$

for $\ell = -M, \ldots, M$. This expression also holds for $m \in C^{1,\alpha}_{pw}(A)$ with $k = -1$. Clearly, it suffices to bound $\|K_\ell(u - u^M)\|_\infty$ for $\ell = 0, \ldots, M$.

Thus, for $k \geq 0$, we obtain

$$\sum_{j>M} \frac{1}{j^{k+4}} \frac{1}{(j-\ell)^{k+2}} \leq \frac{1}{M^{k+4}} \frac{C_2}{(M+1-\ell)^{k+1}}.$$

For $k = -1 (m \in C^{1,\alpha}_{pw}(A))$, on the other hand, we find that

$$\begin{aligned}\sum_{j>M} \frac{1}{j^3}\frac{1}{j-\ell} &\le \frac{1}{M^2}\frac{1}{\ell}\sum_{j>M}\left(\frac{1}{j-\ell}-\frac{1}{j}\right)\\ &\le \frac{1}{M^2}\frac{C_3}{\ell}\log\left(\frac{M+1}{M+1-\ell}\right)\\ &\le \frac{1}{M^2}\frac{C_3}{M+1-\ell},\end{aligned}$$

since $\log x \le x-1$ for $x>0$.

To obtain the final result, it suffices to consider sums of the following form:

$$\sum_{\ell=1}^{M}\frac{1}{\ell^2}\frac{1}{(M+1-\ell)^p}$$

for $p = 1, 2, \ldots$. First, for $p \ge 2$,

$$\begin{aligned}\sum_{\ell=1}^{M}\frac{1}{\ell^2}\frac{1}{(M+1-\ell)^p} &\le \sum_{\ell=1}^{M}\frac{1}{\ell^2}\frac{1}{(M+1-\ell)^2}\\ &\le 2\sum_{\ell=1}^{\lceil \frac{M}{2}\rceil}\frac{1}{\ell^2}\frac{1}{(M+1-\ell)^2}\\ &= \mathcal{O}\left(\frac{1}{M^2}\right)\end{aligned}$$

as $M \to \infty$. Finally, for $p = 1$, we obtain

$$\begin{aligned}\sum_{\ell=1}^{M}\frac{1}{\ell^2}\frac{1}{M+1-\ell} &= \sum_{\ell=1}^{\lceil \frac{M}{2}\rceil}\frac{1}{\ell^2}\frac{1}{M+1-\ell} + \sum_{\ell=\lceil \frac{M}{2}\rceil+1}^{M}\frac{1}{\ell^2}\frac{1}{M+1-\ell}\\ &= \mathcal{O}\left(\frac{1}{M}\right)+\mathcal{O}\left(\frac{\log M}{M^2}\right)\\ &= \mathcal{O}\left(\frac{1}{M}\right)\end{aligned}$$

as $M \to \infty$. Combining these results, the theorem follows. ∎

Remark 3.6. Of course, there are many other conditions on m for which the corresponding convergence rates could be determined; for instance, one might remove the requirement of Hölder continuity. In every case, the convergence rates are directly determined by the rate of decay of the Fourier coefficients of m and u. We do not attempt to provide a comprehensive listing of all possible regularity conditions and their corresponding convergence rates.

Remark 3.7. Numerical experiments indicate that the bounds of Theorem 3.5 are tight. The resulting convergence rates depend on k in a particularly interesting way. As we have shown, the method exhibits *third-order* convergence for $m \in C^1_{pw}(A)$, *fifth-order* convergence for $m \in C^{0,\alpha}(A) \cap C^2_{pw}(A)$, and *seventh-order* convergence for $m \in C^{1,\alpha}(A) \cap C^3_{pw}(A)$. This rather interesting and unexpected k-dependence of the convergence rates is observed in the numerical examples of section 5.

3.2. Total error in the interior and exterior fields. Up to this point, we have computed only convergence rates for the approximated modes, i.e., the modes of order ℓ with $|\ell| \leq M$. Given these convergence rates, we can now easily estimate the total error. We make a distinction here between two types of error: the *interior field error* (the error on the domain of integration $A = \{(a, \phi) : 0 \leq R_0 \leq a \leq R_1\}$) and the *exterior field error* (the error outside of A). The interior field error is simply the difference between the true solution $u(x)$ and the solution $v^M(x)$ of (2.3) on A. Clearly, on A we have

$$\begin{aligned}\|u - v^M\|_\infty &\leq \|u^M - v^M\|_\infty + \|u - u^M\|_\infty \\ &\leq \varepsilon_M + \tau_M,\end{aligned}$$

where ε_M is defined in (3.2) and $\tau_M = \|u - u^M\|_\infty$.

Remark 3.8. Note that the decay rate of $(u - u^M)(x_0)$ for a particular point $x_0 \in A$, as opposed to the maximum error in all of A, depends on the regularity of m in a neighborhood of the circle with radius $r_0 = |x_0|$ centered at the origin. Hence, in general, the convergence rate of $v^M(x_0)$ to $u(x_0)$ may vary with the choice of $x_0 \in A$. In particular, the regularity of m in a neighborhood of the circle with radius $r_0 = |x_0|$ centered at the origin determines the regularity of u in that neighborhood and hence also determines the decay rate of the Fourier coefficients $u_\ell(r_0)$. This decay rate in turn determines whether ε_M or $(u - u^M)(x_0)$ dominates the convergence rate. The pointwise convergence rate is of limited usefulness, however; the following corollary to Theorem 3.5 provides a bound on the maximum error in the computed interior field.

COROLLARY 3.9 (interior field error). *If $m \in C^{k,\alpha}(A) \cap C^{k+2}_{pw}(A)$, then the interior field error is given by*

$$\|u - v^M\|_\infty = \mathcal{O}\left(\frac{1}{M^{k+3}}\right).$$

This result holds with $k = -1$ for $m \in C^1_{pw}(A)$.

Proof. By Lemma 3.4,

$$\tau_M = \sum_{\ell > M} \frac{C}{\ell^{k+4}} = \mathcal{O}\left(\frac{1}{M^{k+3}}\right)$$

as $M \to \infty$. Clearly, by Theorem 3.5, τ_M dominates ε_M for every k. The proof for $m \in C^1_{pw}(A)$ is similar. ∎

Before discussing convergence rates in the exterior field, we describe how to extend the approximate solution v^M, which we have computed only on the interior of A, to the exterior field. Since the integration in (1.5) is performed only over the support of m, one can easily see that, given the solution u on the boundary of A, the solution in the rest of $\mathbb{R}^2$ can be computed simply by an appropriate scaling of the Fourier modes of u^s on the (circular) inner and outer boundaries of A at radii R_0 and R_1, respectively. More precisely, we find that

$$u^s_\ell(a) = \begin{cases} \dfrac{J_\ell(\kappa a)}{J_\ell(\kappa R_0)} u^s_\ell(R_0) & \text{if } 0 \leq a < R_0, \\ \dfrac{H^1_\ell(\kappa a)}{H^1_\ell(\kappa R_1)} u^s_\ell(R_1) & \text{if } a > R_1. \end{cases} \tag{3.4}$$

Our approximate solution v^M is extended to the exterior of A by the same procedure.

COROLLARY 3.10 (exterior field error). *Let $m \in \mathcal{M}$. Given $x_0 \notin A$, extend the approximate solution v^M to the exterior of A by means of* (3.4) *above. More precisely, for $\ell = -M, \ldots, M$, let $r_0 = |x_0|$ and define*

$$v_\ell(r_0) = u_\ell^i(r_0) + \begin{cases} \dfrac{J_\ell(\kappa r_0)}{J_\ell(\kappa R_0)}[v_\ell(R_0) - u_\ell^i(R_0)] & \text{if } 0 \le r_0 < R_0, \\ \dfrac{H_\ell^1(\kappa r_0)}{H_\ell^1(\kappa R_1)}[v_\ell(R_1) - u_\ell^i(R_1)] & \text{if } r_0 > R_1. \end{cases}$$

(Note: If $R_0 = 0$, then the integration domain is a disc and, hence, only the part of the equation above corresponding to $r_0 > R_1$ applies.) Then, the exterior field error at $x_0 \notin A$ is given by

$$|u(x_0) - v^M(x_0)| = \mathcal{O}(\varepsilon_M)$$

as $M \to \infty$, where ε_M, defined in (3.2)*, has bounds given by Theorem* 3.5.

Proof. Assume that $r_0 > R_1$; the proof for $0 \le r_0 < R_0$ is similar. Defining the scaling factors $\beta_\ell(r_0)$ at radius r_0 by

$$\beta_\ell(r_0) = \frac{H_\ell^1(\kappa r_0)}{H_\ell^1(\kappa R_1)}, \tag{3.5}$$

we have

$$\begin{aligned} |u(x_0) - v^M(x_0)| &\le \sum_{\ell=-M}^{M} |\beta_\ell(r_0)|\,|u_\ell(R_1) - v_\ell(R_1)| + |(u - u^M)(x_0)| \\ &\le \varepsilon_M \sum_{\ell=-M}^{M} |\beta_\ell(r_0)| + |(u - u^M)(x_0)|. \end{aligned}$$

As before, let S denote the circle of radius r_0 about the origin. Since $r_0 = |x_0| > R_1$, there exists a neighborhood $N(S)$ of S such that $m|_{N(S)} = 0$. Therefore, $u \in C^\infty(N(S))$ and $|(u - u^M)(x_0)| \le \frac{C}{M^p}$ for *any* integer $p > 0$. This implies that $|(u - u^M)(x_0)|$ is always dominated by ε_M.

Since the Hankel function $H_\ell^1(z) = J_\ell(z) + i\,Y_\ell(z)$, where $Y_\ell(z)$ is the Neumann function of order ℓ, we complete the proof by using the asymptotic expressions for J_ℓ and Y_ℓ [1, p. 365] for fixed z and as $\ell \to \infty$ through positive real values,

$$\begin{aligned} J_\ell(z) &\sim \frac{1}{\sqrt{2\pi\ell}} \left(\frac{ez}{2\ell}\right)^\ell, \\ Y_\ell(z) &\sim -\sqrt{\frac{2}{\pi\ell}} \left(\frac{ez}{2\ell}\right)^{-\ell}. \end{aligned}$$

Therefore, from these asymptotic expressions and from (3.5), we obtain

$$\begin{aligned} |\beta_\ell(r_0)|^2 &= \left|\frac{Y_\ell(\kappa r_0)}{Y_\ell(\kappa R_1)}\right|^2 \frac{1 + \left|\frac{J_\ell(\kappa r_0)}{Y_\ell(\kappa r_0)}\right|^2}{1 + \left|\frac{J_\ell(\kappa R_1)}{Y_\ell(\kappa R_1)}\right|^2} \\ &\sim \left(\frac{R_1}{r_0}\right)^{2\ell} \end{aligned}$$

as $\ell \to \infty$. This implies that $|\beta_\ell(r_0)|$ is summable. We conclude that as $M \to \infty$

$$|u(x_0) - v^M(x_0)| = \mathcal{O}(\varepsilon_M). \qquad \square$$

Note that while $u \in C^\infty$ on the exterior of A, this function may be much less regular on the interior of A (in general, $u \in C^{1,\alpha}$ for $m \in \mathcal{M}$). Hence, the decay of $u - u^M$ on the exterior of A is superalgebraic, whereas $u - u^M$ may decay as slowly as $\mathcal{O}(M^{-2})$ on the interior of A. This fact is responsible for the interesting result that the method converges more rapidly on the exterior of A than on the interior (where $u - u^M$ may dominate ε_M).

These remarks are particularly relevant in the evaluation of radar cross sections, an important measure in many applications. The evaluation of radar cross sections requires the computation of the *far field*. Although Corollary 3.10 does not directly address the error in the far field, we obtain an approximate far field by a scaling of the Fourier modes of v^M just as in the computation of the exterior field. As in [4, p. 6], we define the far field, u_∞, by the asymptotic representation of the scattered field as $r \to \infty$, i.e.,

$$u^s(r,\phi) = e^{i(\kappa r - \frac{\pi}{4})}\sqrt{\frac{2}{\pi\kappa r}}[u_\infty(\phi) + \mathcal{O}(r^{-1})].$$

From (3.4) and the asymptotic expression for $H^1_\ell(z)$ for fixed ℓ as $z \to \infty$ [1, p. 364], we obtain the Fourier modes of u_∞ by a simple scaling of the Fourier modes of u^s:

$$(u_\infty)_\ell = \frac{u^s_\ell(R_1)}{i^\ell H^1_\ell(\kappa R_1)}.$$

If we define the approximate far field v_∞ in the same way, we can prove that

$$\|u_\infty - v_\infty\| = \mathcal{O}(\varepsilon_M)$$

as $M \to \infty$. The proof of this fact is nearly identical to that of Corollary 3.10.

The predicted convergence rates in both the interior field and the far field are verified through several computational examples in section 5.

4. Computation of the angular integral. We have proven that the solution to the approximate integral equation (2.3) provides a higher-order approximation to the solution of the exact integral equation (1.5) for the scattering problem. However, to this point, we have not discussed any methods for computing the required angular and radial integrals. This paper primarily addresses the theoretical aspects of the method; for a discussion of a particular efficient, higher-order radial integrator, we refer to [5, 7, 8, 14]. On the other hand, with regards to the angular integrals, we show below that the Fourier coefficients of $m(r,\theta)v^M(r,\theta)$ can be computed efficiently and *exactly* (except for roundoff) by means of FFTs.

The required angular integrals are given by

$$I_\ell(r) = \int_0^{2\pi} m(r,\theta)v^M(r,\theta)e^{-i\ell\theta}d\theta, \tag{4.1}$$

where v^M solves the approximate integral equation (2.3). We can express this integral

in terms of the Fourier coefficients of m and v, i.e.,

$$\begin{aligned} I_\ell(r) &= \int_0^{2\pi} \left(\sum_{j=-\infty}^{\infty} m_j(r) e^{ij\theta} \right) \left(\sum_{k=-M}^{M} v_k(r) e^{ik\theta} \right) e^{-i\ell\theta} d\theta \\ &= 2\pi \sum_{k=-M}^{M} m_{\ell-k}(r) v_k(r), \end{aligned} \tag{4.2}$$

where $\ell = -M, \ldots, M$. Hence, we obtain a finite discrete convolution of Fourier coefficients of m and v at each radius; since $|\ell| \le M$ and $|k| \le M$, we have $|\ell - k| \le 2M$. Thus, as stated above, given the Fourier coefficients $m_\ell(r)$ for $|\ell| \le 2M$, we can compute the required angular integrals *exactly*. Furthermore, as is well known, such discrete convolutions may be evaluated (with no discretization error) with the help of FFTs [25, pp. 531–537] so that the computational cost at each radius is of the order of $M \log M$.

This method of computing the angular integrals has an interesting implication concerning the dependence of the solution u on the inhomogeneity m. Indeed, since the computation involves only modes m_ℓ, $|\ell| \le 2M$, replacing m with m^{2M} in the integral equation yields no additional error, i.e.,

$$I_\ell(r) = \int_0^{2\pi} m^{2M}(r,\theta) v^M(r,\theta) e^{-i\ell\theta} d\theta. \tag{4.3}$$

Hence, in a sense, the truncation of the Fourier series of the integral operator *implies* an associated truncation of the Fourier series of the refractive index—as a result of the band-limited nature of the solution v^M. Thus, surprisingly, the low-order approximation of a discontinuous refractive index at each radius by its truncated Fourier series yields *no additional error* beyond that of our original, higher-order truncation of the Fourier series of K. This points to the interesting cancellation of errors phenomenon mentioned briefly in the introduction: the large errors in the Fourier approximation of the refractive index *cancel* in the discrete integration process yielding small errors—high-order accurate approximations—in the evaluation of $I_\ell(r)$.

Note that the discrete-convolution approach to the evaluation of $I_\ell(r)$ ($\ell = -M, \ldots, M$) is equivalent to trapezoidal rule integration of (4.3) with a sufficiently large number of integration points N_θ. This follows from the fact that the trapezoidal rule with N_θ points on the interval $[0, 2\pi]$ integrates the Fourier modes $e^{ik\theta}$ for $|k| < N_\theta$ exactly: using N_θ points in the trapezoidal rule to approximate $\int_0^{2\pi} e^{ik\theta} d\theta$, we obtain

$$\frac{2\pi}{N_\theta} \sum_{j=0}^{N_\theta - 1} e^{2\pi i j k / N_\theta} = \begin{cases} 2\pi & \text{if } k = pN_\theta \text{ for } p \in \mathbb{Z}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, since the largest mode in the integrand of (4.3) is $2M + M + M = 4M$, if we choose $N_\theta = cM$, where $c > 4$, the trapezoidal rule computes (4.3) exactly (except for roundoff) and the use of FFTs yields a complexity of $\mathcal{O}(M \log M)$. Algorithmically, this is entirely equivalent to computing the discrete convolution (4.2) using FFTs.

5. Computational examples. In this section, we illustrate the performance of the two-dimensional algorithm for a variety of scattering configurations. We first study the convergence of the method for two scatterers for which analytical solutions

are known. We then verify that the algorithm achieves the predicted convergence rates for three scatterers of varying degrees of regularity.

In each case, we compute the near and far fields produced under plane wave incidence, $u^i(x,y) = e^{i\kappa x}$. To compute the maximum error in the near field, we evaluate the solution computed by our method on an evenly spaced polar grid. On this grid, we evaluate the maximum absolute error as compared with either the analytical solution (when it is available) or the solution computed with a finer discretization. The maximum error in the far field is computed similarly by interpolating to an evenly spaced angular grid.

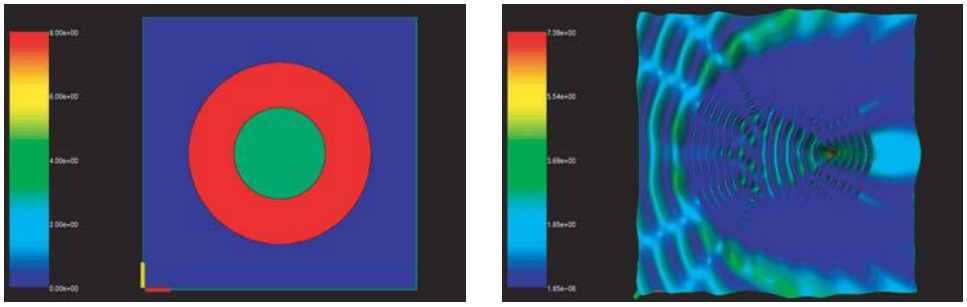
The results for each example are given in the accompanying figures and tables. The figures include visualizations of $-m(x) = n^2(x) - 1$ and the computed near field intensity, $|v^M|^2$. The tables provide values for the number of modes M in the approximate solution v^M, the wall-clock time required, and the maximum absolute errors in the near and far field denoted by ϵ_u^{nf} and ϵ_u^{ff}, respectively. Additionally, the ratios of the errors at successive levels of discretization are listed to illustrate the convergence rates. For some discretizations, the accuracy in the computed solution has reached either machine-precision accuracy (actually just less than machine precision due to round-off errors), the accuracy of the radial integration, or the tolerance of the linear solver. In such a case, we observe no improvement in the error of the solution as we refine the discretization and hence, to indicate a converged solution, we write "Conv." in the ratio column.

Our main goal in this section is to verify the convergence rates established in Theorem 3.5 and Corollaries 3.9 and 3.10. Hence, in this section, we are primarily concerned with the convergence in the number of Fourier modes M, rather than the convergence in the number of radial points. We also seek to demonstrate the $\mathcal{O}(M \log M)$ complexity of the angular integration method. We therefore fix the number of radial points at a sufficiently large value and we hold the number of iterations of the linear solver (GMRES) fixed at a value that produces a sufficiently accurate solution of the linear system. This isolates the dependence of the times and errors on M and allows us to confirm the computational complexity and the predicted convergence rates. All of these results were computed using a 700 MHz Pentium III Xeon workstation.

We first compute the scattering by two obstacles for which an analytical solution is known: (1) a cylindrically symmetric scatterer centered at the origin with piecewise-constant refractive index and (2) a disc centered at $(1\lambda, 0)$ with constant refractive index.

The results for the first example are presented in Figure 5.1 and Table 5.1. Here the inner disc has a radius equal to 1λ and a refractive index $n = 2$; the outer annulus has an outer radius of 2λ and a refractive index $n = 3$. Thus, this scatterer has a diameter of 10 interior wavelengths. (Perhaps the best indication of the difficulty of a scattering problem is given by the size of the scatterer in terms of interior wavelengths, since the numerical method must resolve these wavelengths to provide any accuracy.) One may also observe that the method obtains an exponential convergence rate. This occurs despite the discontinuity in the refractive index because, at each radius, the refractive index is a C^∞ function of the angular variable. Finally, we observe that the time required is consistent with an $\mathcal{O}(M \log M)$ complexity.

The results for the second example are presented in Figure 5.2 and Table 5.2. Here the disc is centered at $(1\lambda, 0)$ and has a diameter of 1λ and a refractive index $n = \sqrt{2}$. Thus, it has a diameter of $\sqrt{2}$ interior wavelengths. As opposed to the previous

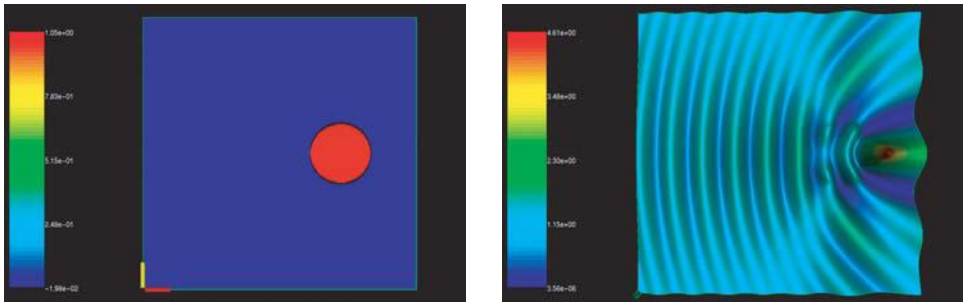


(a) Scatterer ($-m = n^2 - 1$) (b) Near Field Intensity ($|u|^2$)

FIG. 5.1. *Visualizations for a radially layered scatterer. Diameter* = 10 *interior wavelengths.*

TABLE 5.1
Convergence rate for a radially layered scatterer. Diameter = 10 *interior wavelengths.*

M	Time	ϵ_u^{nf}	Ratio	ϵ_u^{ff}	Ratio
15	3.05s	8.50e-2		4.28e-2	
30	3.83s	1.13e-9	7.52e+7	5.46e-13	7.83e+10
60	5.46s	1.68e-12	6.73e+2	4.97e-13	Conv.



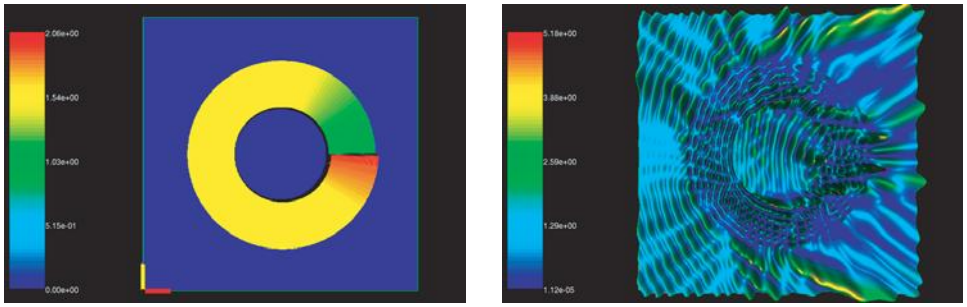
(a) Scatterer ($-m = n^2 - 1$) (b) Near Field Intensity ($|u|^2$)

FIG. 5.2. *Visualizations for an off-center disc. Diameter* = $\sqrt{2}$ *interior wavelengths.*

TABLE 5.2
Convergence for an off-center disc. Diameter = $\sqrt{2}$ *interior wavelengths.*

M	Time	ϵ_u^{nf}	Ratio	ϵ_u^{ff}	Ratio
15	7s	6.22e-2			
30	13s	5.95e-3	10.45	1.58e-3	18.80
60	25s	1.13e-3	5.27	1.83e-4	8.63
120	49s	2.83e-4	3.99	2.27e-5	8.06
240	99s	5.99e-5	4.72	2.84e-6	7.99
480	194s	6.65e-6	9.01	3.56e-7	7.98
960	386s	1.99e-6	3.34	4.42e-8	8.05
1920	808s	2.75e-7	7.24	4.21e-9	10.50

example, however, we do not observe an exponential rate of convergence despite the fact that the disc has a constant refractive index. Since the disc is not centered at the origin, the refractive index at each radius is actually a *discontinuous* function of



(a) Scatterer ($-m = n^2 - 1$) (b) Near Field Intensity ($|u|^2$)

FIG. 5.3. *Visualizations for a discontinuous scatterer. Annulus thickness* ≈ 4.33 *interior wavelengths.*

TABLE 5.3
Convergence rate for a discontinuous scatterer. Annulus thickness ≈ 4.33 *interior wavelengths.*

M	Time	ϵ_u^{nf}	Ratio	ϵ_u^{ff}	Ratio
60	27s	3.24e-2		2.07e-2	
120	52s	4.69e-3	6.91	1.95e-3	10.62
240	109s	6.23e-4	7.53	2.32e-4	8.41
480	228s	9.71e-5	6.42	2.87e-5	8.08
960	458s	1.04e-5	9.34	3.53e-6	8.13
1920	898s	1.45e-6	7.17	3.83e-7	9.22

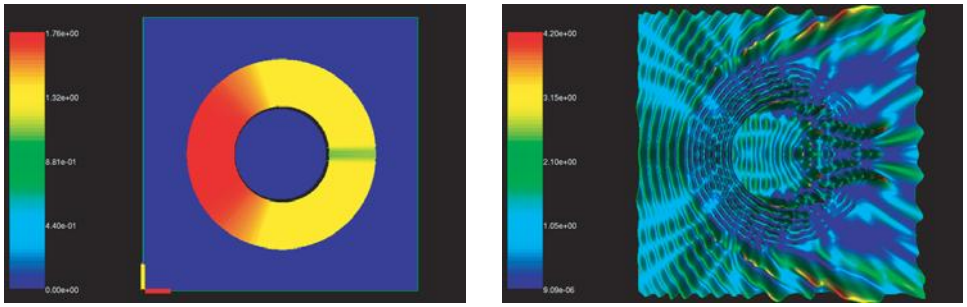
the angular variable. Since the analytical solution in this case is known, the off-center disc provides direct verification of the predicted convergence rates for a *discontinuous* refractive index. The table shows excellent agreement with the predicted third-order convergence in the far field. The convergence in the near field is less steady, but is consistent with the predicted second-order convergence in the near field. As in the previous example, we observe that the computing time scales appropriately with M.

We now illustrate the convergence of the method for a series of three simple scatterers of increasing degrees of regularity. In each case, $m(x) = 1 - n^2(x)$ is given in the following form:

$$m(r,\theta) = \begin{cases} -\dfrac{3}{2} - \dfrac{1}{2\pi} \displaystyle\sum_{|\ell|\geq 1} \left(\frac{i}{\ell}\right)^{k+2} e^{i\ell\theta} & \text{for } \dfrac{5}{2}\lambda \leq r \leq 5\lambda, \\ 0 & \text{otherwise.} \end{cases}$$

Note that for each integer k, this series becomes either a sine or cosine series with real coefficients. If $k = -1$, m is discontinuous and piecewise smooth as a function of θ. Further, for any integer $k \geq 0$, $m \in C^{k,\alpha} \cap C^{\infty}_{pw}$ as a function of θ. The three examples that follow illustrate the convergence of the method for $k = -1, 0, 1$. Because these scatterers are fully inhomogeneous, their size in terms of interior wavelengths is not easily defined. Note, however, that each annular scatterer has a radial thickness of 2.5λ in terms of incident wavelengths; if the refractive index were constant within the annulus and equal to the maximum, then the radial thickness of the annulus would be approximately 4.33, 4.15, and 4.54 interior wavelengths for $k = -1, 0, 1$, respectively.

The results for $k = -1$ are found in Figure 5.3 and Table 5.3. The predicted second-order convergence in the near field is exceeded and the third-order convergence

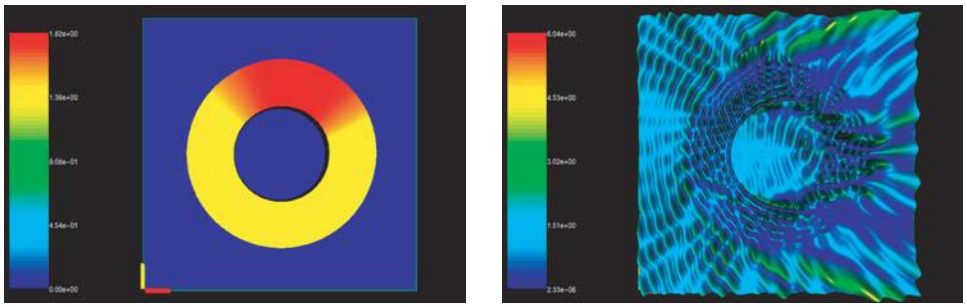


(a) Scatterer ($-m = n^2 - 1$) (b) Near Field Intensity ($|u|^2$)

FIG. 5.4. *Visualizations for a $C^{0,\alpha}$ scatterer. Annulus thickness ≈ 4.15 interior wavelengths.*

TABLE 5.4
Convergence rate for a $C^{0,\alpha}$ scatterer. Annulus thickness ≈ 4.15 interior wavelengths.

M	Time	ϵ_u^{nf}	Ratio	ϵ_u^{ff}	Ratio
60	23s	9.33e-4		7.06e-6	
120	50s	8.91e-5	10.47	1.30e-7	54.31
240	105s	1.15e-5	7.75	3.86e-9	33.68
480	212s	1.46e-6	7.88	1.17e-10	32.99
960	565s	1.83e-7	7.97	1.73e-11	Conv.
1920	1136s	1.98e-8	9.24	1.85e-11	Conv.



(a) Scatterer ($-m = n^2 - 1$) (b) Near Field Intensity ($|u|^2$)

FIG. 5.5. *Visualizations for a $C^{1,\alpha}$ scatterer. Annulus thickness ≈ 4.55 interior wavelengths.*

TABLE 5.5
Convergence rate for a $C^{1,\alpha}$ scatterer. Annulus thickness ≈ 4.55 interior wavelengths.

M	Time	ϵ_u^{nf}	Ratio	ϵ_u^{ff}	Ratio
60	36s	2.16e-5		7.33e-9	
120	72s	4.81e-7	44.91	1.06e-11	691.51
240	160s	1.05e-8	45.81	4.50e-12	Conv.
480	331s	4.76e-10	22.06	4.52e-12	Conv.
960	561s	1.36e-11	35.0	4.61e-12	Conv.
1920	1172s	1.94e-12	Conv.	4.72e-12	Conv.

in the far field is readily observed. The results of $k = 0$ are found in Figure 5.4 and Table 5.4. In this case, the predicted third-order convergence in the near field and

fifth-order convergence in the far field are both matched quite precisely. This example clearly illustrates the interesting jump in the far field convergence rate from third-order for a discontinuous refractive index to *fifth-order* for a $C^{0,\alpha}$ refractive index. Finally, the results for $k = 1$ are found in Figure 5.5 and Table 5.5. In this case, the predicted fourth- and seventh-order convergence rates in the near and far fields, respectively, are clearly exceeded. However, because convergence is so rapid, it is difficult to observe a definite pattern, especially in the far field convergence. In each of these cases, we note that the computing time scales appropriately with M. Finally, we mention that even the largest of these examples required less than 20 minutes and less than 700 MB of memory.

Appendix A. Bound on Fourier coefficients of the fundamental solution. To prove that solutions to the approximate integral equation (2.3) exist and to bound the convergence rate of the method, we need a bound on the decay rate of the Fourier coefficients of the fundamental solution $\mathcal{J}_\ell(a, r)$ defined in (2.1). This decay rate is given in Lemma 2.1.

According to [1, p. 362], for all integers $\ell \geq 0$ and for any real, nonnegative z,

$$|J_\ell(z)| \leq \frac{1}{\ell!}\left(\frac{z}{2}\right)^\ell \leq \frac{z^\ell}{\ell!}. \tag{A.1}$$

The following lemma provides a similar bound for $|Y_\ell(z)|$.

LEMMA A.1. *Let $z \in \mathbb{R}$ with $0 \leq z \leq R$. For all integers $\ell \geq 1$,*

$$|Y_\ell(z)| \leq C\frac{(\ell-1)!}{z^\ell},$$

and for $\ell = 0$,

$$|Y_\ell(z)| \leq C|\log(z)|,$$

where $C > 0$ depends only on R.

Proof. By [4, p. 51], $Y_\ell(z)$ is given for any nonnegative integer ℓ by

$$\begin{aligned} Y_\ell(z) = {} & \frac{2}{\pi}J_\ell(z)\log\left(\frac{z}{2}\right) - \frac{1}{\pi}\sum_{k=0}^{\ell-1}\frac{(\ell-k-1)!}{k!}\left(\frac{z}{2}\right)^{2k-\ell} \\ & -\frac{1}{\pi}\sum_{k=0}^{\infty}\frac{\psi(\ell+k)+\psi(k)}{(-1)^k k!(k+\ell)!}\left(\frac{z}{2}\right)^{2k+\ell}, \end{aligned} \tag{A.2}$$

where $\psi(0) = -\gamma \approx -0.5772$ and $\psi(k) = -\gamma + \sum_{j=1}^{k}\frac{1}{j}$ for $k \geq 1$.

To bound the second term in (A.2), we find that

$$\begin{aligned} \sum_{k=0}^{\ell-1}\frac{(\ell-k-1)!}{k!}\left(\frac{z}{2}\right)^{2k} &\leq (\ell-1)!\sum_{k=0}^{\infty}\frac{1}{k!}\left[\left(\frac{z}{2}\right)^2\right]^k \\ &\leq (\ell-1)!\, e^{\left(\frac{R}{2}\right)^2} \leq C_1(R)(\ell-1)!. \end{aligned}$$

Now note that for $k \geq 1$, $|\psi(0)| \leq 1$ and $0 \leq \psi(k) \leq -\gamma + k \leq k$.

Hence, for a bound on the third term in (A.2), we obtain

$$\sum_{k=0}^{\infty}\frac{|\psi(\ell+k)+\psi(k)|}{k!(k+\ell)!}\left(\frac{z}{2}\right)^{2k+\ell} \leq 2\sum_{k=0}^{\infty}\frac{1}{k!}\frac{[(R/2)^2]^{\frac{\ell}{2}+k}}{(\ell+k-1)!} \leq C_2(R),$$

since

$$\frac{[(R/2)^2]^{\frac{\ell}{2}+k}}{(\ell+k-1)!} \le C_3(R).$$

These bounds together with (A.1) yield the desired result. ∎

We now turn to the proof of the main lemma.

Proof of Lemma 2.1. First note that

$$\begin{aligned}\int_{R_0}^{R_1} |\mathcal{J}_\ell(a,r)|r\,dr &= |H_\ell^1(\kappa a)| \int_{R_0}^{a} |J_\ell(\kappa r)|r\,dr + |J_\ell(\kappa a)| \int_a^{R_1} |H_\ell^1(\kappa r)|r\,dr \\ &\le |J_\ell(\kappa a)| \int_{R_0}^{R_1} |J_\ell(\kappa r)|r\,dr + |J_\ell(\kappa a)| \int_a^{R_1} |Y_\ell(\kappa r)|r\,dr \\ &\quad + |Y_\ell(\kappa a)| \int_{R_0}^{a} |J_\ell(\kappa r)|r\,dr \\ &\le I_{J,J} + I_{J,Y} + I_{Y,J},\end{aligned}$$

where

$$\begin{aligned}I_{J,J} &= |J_\ell(\kappa a)| \int_0^{R_1} |J_\ell(\kappa r)|r\,dr, \\ I_{J,Y} &= |J_\ell(\kappa a)| \int_a^{R_1} |Y_\ell(\kappa r)|r\,dr, \\ I_{Y,J} &= |Y_\ell(\kappa a)| \int_0^{a} |J_\ell(\kappa r)|r\,dr.\end{aligned}$$

Note that $|J_{-\ell}(z)| = |(-1)^\ell J_\ell(z)| = |J_\ell(z)|$ and similarly $|Y_{-\ell}(z)| = |Y_\ell(z)|$. Hence, it suffices to bound these integrals for $\ell \ge 0$.

Thus, for $\ell \ge 0$, by (A.1),

$$I_{J,J} \le \frac{1}{(\ell!)^2} R_1^2 (\kappa R_1)^{2l} \le \frac{C_{J,J}}{\ell^2},$$

where $C_{J,J} > 0$ depends only on κ and R_1. By (A.1) and Lemma A.1, we find that for $\ell > 2$,

$$\begin{aligned}I_{J,Y} &\le C \frac{(\kappa a)^\ell}{\ell!} \int_a^{R_1} \frac{(\ell-1)!}{(\kappa r)^\ell} r\,dr \\ &= C \frac{{R_1}^2}{\ell(\ell-2)} \left[\left(\frac{a}{R_1}\right)^2 - \left(\frac{a}{R_1}\right)^\ell \right] \le \frac{C_{J,Y}}{\ell^2}.\end{aligned}$$

A similar argument shows that $I_{J,Y}$ is also bounded for $\ell = 0, 1, 2$. Finally, for $\ell \ge 1$, we find that

$$\begin{aligned}I_{Y,J} &\le C \frac{(\ell-1)!}{(\kappa a)^\ell} \int_0^{a} \frac{(\kappa r)^\ell}{\ell!} r\,dr \\ &= \frac{a^2}{\ell(\ell+2)} \le \frac{C_{Y,J}}{\ell^2}.\end{aligned}$$

It is not difficult to show that this same bound holds for $\ell = 0$. ∎

TABLE B.1
Relative errors in trapezoidal rule integration.

N	Error	Ratio
1	2.5e-1	
2	9.5e-2	2.6
4	3.5e-2	2.7
8	1.3e-2	2.7
8192	4.2e-7	

(a) $\int_0^{1/2} \sqrt{x}dx \approx 0.2357$

N	Error	Ratio
1	4.8e-2	
2	1.2e-2	4.0
4	2.9e-3	4.1
8	7.4e-4	3.9
8192	7.0e-10	

(b) $\int_0^{\pi/4} e^{\cos^2 x}dx \approx 1.8009$

N	Error	Ratio
1	5.5e-1	
2	6.0e-2	9.2
4	3.1e-4	1.9e+2
8	7.2e-10	4.3e+5
16	2.1e-23	3.4e+13

(c) $\int_0^{\pi} e^{\cos^2 x}dx \approx 5.5084$

Appendix B. Higher-order integration via the trapezoidal rule. When used to integrate a smooth and periodic function over its period, the trapezoidal rule obtains a truly extraordinary convergence rate (see [21, section 9.4] and [29]). As with our numerical method, this convergence behavior is due to the rapid decay of the function's Fourier coefficients (see Lemma 3.1). Since this fact may yet be unfamiliar to some readers, we illustrate trapezoidal rule convergence through three simple, one-dimensional integrals.

In Table B.1, we give the relative errors obtained when computing the integrals of the functions $\sqrt{x}$ and $e^{\cos^2 x}$ by means of the trapezoidal rule with N points. In Table B.1(a), we observe less than second-order convergence when computing $\int_0^{1/2} \sqrt{x}dx$, which is a result of the singularity in its first derivative at the origin. Table B.1(b) shows second-order convergence when computing $\int_0^{\pi/4} e^{\cos^2 x}dx$, which agrees with the well-known convergence rate predicted for the trapezoidal rule when integrating C^2 functions. Finally, in Table B.1(c), we observe an *exponential* convergence rate when computing $\int_0^{\pi} e^{\cos^2 x}$, the same function integrated in Table B.1(b). Note that in this example a relative error of 7×10^{-10} is obtained with 8 points, whereas in Table B.1(b), 8192 points are required for similar accuracy. This extraordinary convergence rate results because we are integrating a *smooth and periodic function over its period*.

Acknowledgments. Color visualizations were generated with the VTK-based visualization tool Vizamrai, developed by Steven Smith at the Center for Applied Scientific Computing (CASC) at Lawrence Livermore National Laboratory. The authors also gratefully acknowledge the constructive suggestions of an anonymous referee.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1965.
[2] L. BERS, F. JOHN, AND M. SCHECHTER, *Partial Differential Equations*, John Wiley and Sons, New York, 1964.
[3] N. N. BOJARSKI, *The k-space formulation of the scattering problem in the time domain*, J. Opt. Soc. Amer., 72 (1982), pp. 570–584.
[4] J. J. BOWMAN, T. B. A. SENIOR, AND P. L. E. USLENGHI, EDS., *Electromagnetic and Acoustic Scattering by Simple Shapes*, North-Holland, Amsterdam, 1969.
[5] O. P. BRUNO AND E. M. HYDE, *An efficient, preconditioned, high-order solver for scattering by two-dimensional, inhomogeneous media*, J. Comput. Phys., 200 (2004), pp. 670–694.
[6] O. P. BRUNO, E. M. HYDE, AND F. REITICH, *An accelerated solver for penetrable obstacle scattering with tunable order of convergence*, in preparation.
[7] O. P. BRUNO AND A. SEI, *A fast high-order solver for EM scattering from complex penetrable bodies: TE case*, IEEE Trans. Antennas and Propagation, 48 (2000), pp. 1862–1864.

[8] O. P. Bruno and A. Sei, *A fast high-order solver for problems of scattering by heterogeneous bodies*, IEEE Trans. Antennas and Propagation, 51 (2003), pp. 3142–3154.
[9] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, New York, 1998.
[10] A. Ditkowski, K. Dridi, and J. S. Hesthaven, *Convergent Cartesian grid methods for Maxwell's equations in complex geometries*, J. Comput. Phys., 170 (2001), pp. 39–80.
[11] G. B. Folland, *Introduction to Partial Differential Equations*, second ed., Princeton University Press, Princeton, NJ, 1995.
[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Heidelberg, New York, 1977.
[13] E. Heinz, *Über die Eindeutigkeit beim Cauchyschen Anfangswertproblem einer elliptischen Differentialgleichung zweiter Ordnung*, Nachr. Akad. Wiss. Göttingen. IIa., 1 (1955), pp. 1–12.
[14] E. M. Hyde, *Fast, High-Order Methods for Scattering by Inhomogeneous Media*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2003.
[15] E. M. Hyde and O. P. Bruno, *A fast, high-order method for scattering by penetrable bodies in three dimensions*, J. Comput. Phys., 202 (2005), 236–261.
[16] E. M. Hyde and O. P. Bruno, *A fast, high-order method for scattering by inhomogeneous media in three dimensions*, Phys. B, 338 (2003), pp. 82–86.
[17] A. Kirsch and P. Monk, *Convergence analysis of a coupled finite element and spectral method in acoustic scattering*, IMA J. Numer. Anal., 10 (1990), pp. 425–447.
[18] A. Kirsch and P. Monk, *An analysis of the coupling of finite-element and Nyström methods in acoustic scattering*, IMA J. Numer. Anal., 14 (1994), pp. 523–544.
[19] H. Koch and D. Tartaru, *Carleman estimates and unique continuation for second-order elliptic equations with nonsmooth coefficients*, Comm. Pure Appl. Math., 54 (2001), pp. 339–360.
[20] R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, 1989.
[21] R. Kress, *Numerical Analysis*, Grad. Texts in Math. 181, Springer-Verlag, New York, 1998.
[22] E. H. Lieb and M. Loss, *Analysis*, Grad. Stud. Math. 14, AMS, Providence, RI, 1997.
[23] G. Liu and S. Gedney, *High-order Nyström solution of the volume EFIE for TM-wave scattering*, Microwave and Optical Technology Letters, 25 (2000), pp. 8–11.
[24] P. A. Martin, *Acoustic scattering by inhomogeneous obstacles*, SIAM J. Appl. Math., 64 (2003), pp. 297–308.
[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran* 77*: The Art of Scientific Computing*, Vol. 1, 2nd ed., Cambridge University Press, Cambridge, 1992.
[26] W. Rachowicz and L. Demkowicz, *An hp-adaptive finite element method for electromagnetics. Part* 1*: Data structure and constrained approximation*, Comput. Methods Appl. Mech. Engrg., 187 (2000), pp. 307–335.
[27] G. M. Vainikko, *Fast solvers of the Lippmann-Schwinger equation*, in Direct and Inverse Problems of Mathematical Physics (Newark, DE, 1997), R. P. Gilbert, J. Kajiwara, and Y. S. Xu, eds., Int. Soc. Anal. Appl. Comput. 5, Kluwer Acad. Publ., Dordrecht, The Netherlands, 2000, pp. 423–440.
[28] K. F. Warnick and W. C. Chew, *Numerical simulation methods for rough surface scattering*, Waves Random Media, 11 (2001), pp. R1–R30.
[29] J. A. C. Weideman, *Numerical integration of periodic functions: A few examples*, Amer. Math. Monthly, 109 (2002), pp. 21–36.
[30] P. Werner, *Zur mathematischen Theorie akustischer Wellenfelder*, Arch. Ration. Mech. Anal., 6 (1960), pp. 231–260.
[31] X. M. Xu and Q. H. Liu, *Fast spectral-domain method for acoustic scattering problems*, IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control, 48 (2001), pp. 522–529.
[32] B. Yang, D. Gottlieb, and J. S. Hesthaven, *Spectral simulations of electromagnetic wave scattering*, J. Comput. Phys., 134 (1997), pp. 216–230.
[33] P. Zwamborn and P. V. den Berg, *Three dimensional weak form of the conjugate gradient FFT method for solving scattering problems*, IEEE Trans. Microwave Theory and Techniques, 40 (1992), pp. 1757–1766.
[34] A. Zygmund, *Trigonometric Series*, 2nd ed., Cambridge University Press, London, 1968.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2320–2341

ROBUST A POSTERIORI ERROR ESTIMATION FOR NONCONFORMING FINITE ELEMENT APPROXIMATION*

MARK AINSWORTH†

Abstract. The equilibrated residual method for a posteriori error estimation is extended to nonconforming finite element schemes for the approximation of linear second order elliptic equations where the permeability coefficient is allowed to undergo large jumps in value across interfaces between differing media. The estimator is shown to provide a computable upper bound on the error and, up to a constant depending only on the geometry, provides two-sided bounds on the error. The robustness of the estimator is also studied and the dependence of the constant on the jumps in permeability is given explicitly.

Key words. robust a posteriori error estimation, nonconforming finite element, Crouzeix–Raviart element, saturation assumption

AMS subject classifications. Primary, 65N30; Secondary, 65N15, 65N50, 76S05

DOI. 10.1137/S0036142903425112

1. Introduction.

A posteriori error estimation for *conforming* finite element schemes has been the subject of extensive investigation, and such methods are now routinely incorporated in adaptive finite element procedures by the engineering and scientific computing communities. In contrast, the treatment of *nonconforming* methods [8] has been subject to sporadic yet sustained attention over the past decade.

The early work of Agouzal [1] was concerned with a posteriori error estimation for nonconforming finite element approximation of Poisson-type problems. The important contribution of Dari et al. [10] presented an explicit a posteriori error estimator based on evaluation of norms of residuals supplemented by jumps in fluxes across interelement edges and showed that the estimator provides two-sided bounds on the error up to generic, unknown constants that are independent of the mesh size. This was subsequently extended to nonconforming mixed finite element approximation of Stokes flow [9] and non-Newtonian flow [3]. The application of hierarchic basis estimators to nonconforming finite element approximation was considered by Hoppe and Wohlmuth [16], where the usual hierarchic basis estimator is augmented with an additional term comparing the nonconforming approximation with a smoothed approximation. Two-sided bounds were obtained [16] under the assumption that a saturation condition is valid. In a related approach, Schieweck [18] proposed a residual-based estimator supplemented with the same additional term as in [16]. However, the analysis of efficiency in [18] was based on additional rather strong assumptions on the regularity of the mesh and the true solution. Carstensen, Bartels, and Jansche [6] derived estimators based on *gradient* averaging (or smoothing) techniques and obtained two-sided bounds. All of the above estimators involve generic, unknown constants and as such provide refinement indicators rather than actual numerical bounds on the error.

*Received by the editors March 25, 2003; accepted for publication (in revised form) July 20, 2004; published electronically March 31, 2005. This work was supported by the Leverhulme Trust through a Leverhulme Trust Fellowship and was completed while the author was visiting the Newton Institute for Mathematical Sciences, Cambridge, UK.

http://www.siam.org/journals/sinum/42-6/42511.html

†Mathematics Department, Strathclyde University, 26 Richmond Street, Glasgow G1 1XH, Scotland (M.Ainsworth@strath.ac.uk).

Destuynder and Métivet [12] derived a posteriori bounds for the error in a conforming approximation obtained by smoothing the nonconforming approximation. Explicit, computable upper bounds on the error measured in the energy norm were obtained for approximation of Poisson's equation. To show that the bounds are efficient, the authors made additional regularity assumptions on the mesh and the true solution and showed that the estimator decays at the same rate as the true error. Unfortunately, these regularity assumptions on the mesh and the true solution generally fail to hold in the context of the solution of practical problems on adaptively refined meshes.

A technique that has proved particularly effective and robust for a posteriori estimation of the error in *conforming* finite element schemes is the *equilibrated residual method*, as described, for example, in [2]. One goal of the present work is to extend the equilibrated residual method to nonconforming finite element schemes for the approximation of a linear second order elliptic problem with variable permeability. The permeability is assumed to be piecewise constant on subdomains corresponding to different media but is allowed to undergo large jumps across interfaces. Particular attention is paid to the robustness of the a posteriori error estimator with respect to the size of the jumps. This issue has also been studied in the setting of conforming finite element approximation by Bernardi and Verfürth [4].

The approach is based on the idea of Dari et al. [10] involving an orthogonal (Helmholtz) decomposition of the error into a conforming part and a nonconforming part. The conforming part is treated using a modification of the standard equilibrated residual method where the weakened continuity requirements for a nonconforming element are fully exploited. Indeed, the usual procedure for conforming approximation can be dramatically simplified to the extent that in its final form the estimator resembles an explicit estimator, but with the advantage that there are no unknown constants. The remaining nonconforming part of the error is estimated using the difference between the nonconforming approximation and a smoothed nonconforming approximation, similarly to [16], and this is shown to give an upper bound without recourse to unknown constants. The final form of the estimator resembles those derived in [10, 12, 16, 18]. However, one byproduct of the method of derivation is that the estimator can be shown to provide computable upper bounds on the error—a feature characteristic of the equilibrated residual method. Furthermore, the bounds are shown to be efficient in the sense that the estimator is bounded above by the true error up to a constant depending only on the shape of the elements. This result is proved without additional assumptions on the regularity of the true solution and the mesh is allowed to be locally refined, as would be the case of adaptive refinements. However, in order to circumvent the saturation assumption [16], we shall assume that the oscillation of the data is sufficiently small. This extends the ideas of Dörfler and Nochetto [13] to nonconforming finite element approximation. Moreover, the analysis takes full account of the large jumps in the permeability across material interfaces and shows that the estimator is robust with respect to the jumps in certain circumstances (such as if the hypothesis assumed by Bernardi and Verfürth [4] is satisfied).

Gradient smoothing procedures are frequently adopted in the setting of conforming finite element approximation. However, for the nonconforming schemes considered here, smoothing is applied directly to the (discontinuous) finite element approximation, as opposed to its gradient. Interestingly, in the case of Laplace's equation, the exact solution of the local residual problem vanishes identically. This means that the estimator reduces to a recovery-based estimator, and in view of the upper bound

property, we arrive at the somewhat surprising observation[1] that the recovery-based estimator provides a guaranteed upper bound on the error (even though the true solution may have singularities and the mesh may be highly unstructured).

The remainder of this paper is organized as follows. After describing the details of the finite element scheme and the conditions on the mesh, the decomposition of the error into a conforming and nonconforming component is presented. The main results of the paper are then outlined, and illustrative numerical examples are presented. Subsequent sections are concerned with the derivation of the upper and lower bounds for each source of error.

2. Preliminaries.

2.1. Model problem.

Consider the model problem

$$-\mathbf{div}(A\,\mathbf{grad}\,u) = f \text{ in } \Omega \tag{2.1}$$

subject to $u = q$ on Γ_D and $\boldsymbol{n} \cdot A\,\mathbf{grad}\,u = g$ on Γ_N, where Ω is a plane polygonal domain, and the disjoint sets Γ_D and Γ_N form a partitioning of the boundary $\Gamma = \partial\Omega$ of the domain. The data satisfy $f \in L_2(\Omega)$, $g \in L_2(\Gamma_N)$, $q \in H^1(\Gamma_D)$ and $A \in L_\infty(\Omega; \mathbb{R}^{2\times 2})$ is positive definite. For simplicity, it will be assumed that the permeability matrix A is piecewise constant on subdomains of Ω. However, the value of A across a subdomain boundary may undergo jumps of many orders of magnitude, corresponding to transition between regions of widely differing permeability.

The variational form of the problem consists of finding $u \in H^1(\Omega)$ such that $u = q$ on Γ_D and

$$(A\,\mathbf{grad}\,u, \mathbf{grad}\,v) = (f, v) + \int_{\Gamma_N} gv\,\mathrm{d}s \quad \forall v \in H^1_E(\Omega), \tag{2.2}$$

where $H^1_E(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$. In general, we shall use the notation $(\cdot,\cdot)_\omega$ to denote the integral inner product over a region ω, and omit the subscript in the case where ω is the physical domain Ω.

Consider a family of partitions $\{\mathcal{P}\}$ of the domain Ω into the union of nonoverlapping, shape regular triangular elements such that the nonempty intersection of a distinct pair of elements is a single common node or single common edge. The family of partitions is assumed to be locally quasi-uniform in the sense that the ratio of the diameters of any pair of neighboring elements is uniformly bounded above and below over the whole family.

In addition, whenever possible, the partitioning is chosen to reflect the structure of the permeability matrix in the sense that individual elements do not straddle a subdomain boundary where the value of A undergoes a large jump. This requirement is reflected in the assumption that, for every element $K \in \mathcal{P}$, there exist positive constants λ_K and Λ_K satisfying

$$\lambda_K \|\boldsymbol{p}\|^2_{L_2(K)} \le (A\boldsymbol{p}, \boldsymbol{p})_K \le \Lambda_K \|\boldsymbol{p}\|^2_{L_2(K)}, \quad \boldsymbol{p} \in L_2(K)^2, \tag{2.3}$$

such that the ratio $\Upsilon_K = \Lambda_K/\lambda_K$ is uniformly bounded over the whole family of partitions. It will be important to develop a posteriori error estimators whose reliability and efficiency is insensitive to the magnitude of the jumps in permeability between differing regions but which are allowed to depend on the variation of A within a region.

[1] We are grateful to an anonymous reviewer for drawing our attention to the fact that this observation was first reported in [14].

2.2. Nonconforming finite element approximation. Let $\mathcal{N}$ index the set of element vertices, let $\partial\mathcal{P}$ denote the set of element edges, let $\mathcal{M} = \{\boldsymbol{m}_\gamma : \gamma \in \partial\mathcal{P}\}$ denote the set of points located at midpoints of edges, and let $\mathbb{P}_1$ denote the space of polynomials of total degree at most one. The *Crouzeix–Raviart* finite element space [8] is defined by

$$X^{\mathsf{nc}} = \left\{ v : \Omega \to \mathbb{R} : v_{|K} \in \mathbb{P}_1 \quad \forall K \in \mathcal{P}, \quad v \text{ is continuous at } \boldsymbol{m}_\gamma \in \mathcal{M}\backslash\Gamma \right\},$$

with the subspace X_E^{nc} defined by

$$X_E^{\mathsf{nc}} = \{ v \in X^{\mathsf{nc}} : v(\boldsymbol{m}_\gamma) = 0 \text{ for } \gamma \subset \Gamma_D \}.$$

Functions belonging to the space X^{nc} and X_E^{nc} may have discontinuities across element interfaces, meaning that X^{nc} is not a subspace of $H^1(\Omega)$ and therefore constitutes a *nonconforming* approximation space [5, 7]. The nonconforming finite element approximation of problem (2.2) consists of finding $u_{\mathsf{nc}} \in X^{\mathsf{nc}}$ such that

$$\begin{aligned} (A\,\mathbf{grad}_{\mathsf{nc}}\, u_{\mathsf{nc}}, \mathbf{grad}_{\mathsf{nc}}\, v) &= (f, v) + \textstyle\int_{\Gamma_N} g v \,\mathrm{d}s \quad \forall v \in X_E^{\mathsf{nc}}, \\ u_{\mathsf{nc}}(\boldsymbol{m}_\gamma) &= q(\boldsymbol{m}_\gamma) \quad \forall \gamma \subset \Gamma_D, \end{aligned} \tag{2.4}$$

where $\mathbf{grad}_{\mathsf{nc}}$ denotes the operator defined by

$$(\mathbf{grad}_{\mathsf{nc}}\, v)_{|K} = \mathbf{grad}(v_{|K}), \quad K \in \mathcal{P}.$$

A Lagrange-type basis $\{\theta_\gamma\}$ for the space X^{nc} may be constructed by choosing $\theta_\gamma \in X^{\mathsf{nc}}$ to be the function uniquely defined by the conditions

$$\theta_\gamma(\boldsymbol{m}_{\gamma'}) = \delta_{\gamma\gamma'}, \quad \gamma' \in \partial\mathcal{P}. \tag{2.5}$$

A nonconforming interpolation operator $\Pi_{\mathsf{nc}} : H^1(\Omega) \to X^{\mathsf{nc}}$ is defined by the conditions

$$\int_\gamma \Pi_{\mathsf{nc}}\, v \,\mathrm{d}s = \int_\gamma v \,\mathrm{d}s \quad \forall \gamma \in \partial\mathcal{P}. \tag{2.6}$$

The representation of the operator relative to the basis (2.5) is given by

$$\Pi_{\mathsf{nc}}\, v = \sum_{\gamma \subset \partial\mathcal{P}} \overline{v}_\gamma \theta_\gamma, \tag{2.7}$$

where $\overline{v}_\gamma$ denotes the average value of v on an edge γ. Observe that the restriction of $\Pi_{\mathsf{nc}}\, v$ to a particular element K is defined entirely in terms of the averages of the function v on the edges of the element and, moreover, Π_{nc} locally preserves constants. These properties (in conjunction with standard scaling arguments) may be used to deduce that there exists a positive constant C depending only on the shape of the element such that the following local elementwise approximation property holds:

$$\|v - \Pi_{\mathsf{nc}}\, v\|_{L_2(K)} + h_K^{1/2} \|v - \Pi_{\mathsf{nc}}\, v\|_{L_2(\partial K)} \le C h_K \|\mathbf{grad}\, v\|_{L_2(K)}. \tag{2.8}$$

2.3. Data oscillation. It will be necessary to impose some notion of regularity on the underlying problem. Often, a *saturation condition* [16] is assumed, but this assumption makes reference to the (unknown) true solution u, meaning that it is difficult to verify a priori. Dörfler and Nochetto [13], working in the context of conforming piecewise affine approximation, showed that the saturation assumption can be removed in favor of an assumption on the magnitude of the *data oscillation*. This condition has the advantage of being formulated directly in terms of the known data for the problem and can therefore be verified a priori. We shall show that a similar conclusion holds in the nonconforming setting considered here.

The oscillation of the data $f \in L_2(\Omega)$ over the finite element partition $\mathcal{P} = \{K\}$ is defined by

$$\operatorname{osc}(f, \{K : K \in \mathcal{P}\})^2 = \sum_{K \in \mathcal{P}} \operatorname{meas}(K) \left\| f - \overline{f}_K \right\|^2_{L_2(K)}, \tag{2.9}$$

where $\overline{f}_K$ is the average value of f over element K. The data oscillation quantifies the variation in the data f with respect to the partition $\mathcal{P}$. Likewise, the oscillation of the Neumann data g is defined by

$$\operatorname{osc}(g, \{\gamma : \gamma \subset \Gamma_N\})^2 = \sum_{\gamma \subset \Gamma_N} \operatorname{meas}(\gamma) \left\| g - \overline{g}_\gamma \right\|^2_{L_2(\gamma)}. \tag{2.10}$$

The appropriate quantity for the Dirichlet data turns out to be

$$\operatorname{osc}\left(\partial q/\partial s, \{\gamma : \gamma \subset \Gamma_D\}\right)^2 = \sum_{\gamma \subset \Gamma_D} \operatorname{meas}(\gamma) \left\| \frac{\partial q}{\partial s} - \mu_\gamma \right\|^2_{L_2(\gamma)}.$$

Here $\overline{g}_\gamma$ and μ_γ denote the average value of g and $\partial q/\partial s$ on edge γ.

2.4. Path permeability. Let Ω_n denote the patch composed from those elements with a vertex located at $\boldsymbol{x}_n$, and let K, $K' \subset \Omega_n$ be distinct elements. It is useful to introduce the notion of permeability $\lambda_{KK'}$ between a pair of elements. Roughly speaking, this measures the permeability of the "most permeable" route between the elements. The precise definition is based on the observation that there is always at least one connected path $\wp(K, K') \subset \mathcal{P}$ passing from K to K' through adjacent elements belonging to the patch Ω_n. The smallest permeability of all the elements in the path $\wp(K, K')$ is given by $\min\{\lambda_M : M \in \wp(K, K')\}$. If $\boldsymbol{x}_n$ is an interior vertex, then there are two such paths, and in this case we take $\wp(K, K')$ to be the path $\wp^*(K, K')$, which maximizes the value of this quantity, and define

$$\lambda_{KK'} = \min\{\lambda_M : M \in \wp^*(K, K')\}. \tag{2.11}$$

If a vertex $\boldsymbol{x}_n$ of element K lies on the Dirichlet boundary Γ_D, then the element may be linked to Γ_D by a connected path $\wp(K, \Gamma_D)$ passing through adjacent elements as before. The permeability $\lambda_{K\Gamma_D}$ between element K and the Dirichlet boundary is then defined using (2.11) with $\wp^*(K, \Gamma_D)$ in place of $\wp^*(K, K')$. The ratio

$$\Upsilon_{KK'} = \frac{\min(\Lambda_K, \Lambda_{K'})}{\lambda_{KK'}} \tag{2.12}$$

measures path permeability relative to the least permeable of the two elements K and K' at the endpoints of the path.

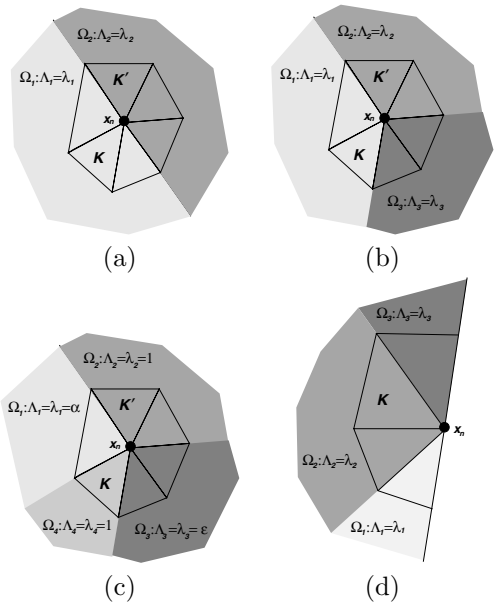


FIG. 2.1. *Value of relative path permeability* (2.12) *for some typical configurations.* (a) *Node on interface:* $\Upsilon_{KK'} = 1$. (b) *Node at crosspoint of three subdomains.* $\Upsilon_{KK'} = 1$. (c) *Node at crosspoint of four subdomains* ($\epsilon \ll 1$). $\Upsilon_{KK'} = 1/\max(\min(1,\alpha),\epsilon)$. (d) *Node on Dirichlet boundary.* $\Upsilon_{K\Gamma_D} = \Lambda_2/\max(\min(\lambda_1,\lambda_2),\min(\lambda_2,\lambda_3))$.

If Ω_n is contained either within a subdomain, on the interface between two subdomains, or at the crosspoint of three subdomains, then the relative path permeability is always unity; see Figure 2.1(a)–(b). However, the relative path permeability may be arbitrarily large at crosspoints where four or more subdomains meet, as would be the case in Figure 2.1(c) if $\alpha \ll 1$. This means that the relative path permeability remains bounded only under additional hypotheses. For example, if, as in Bernardi and Verfürth [4, Hypothesis 2.7], it is assumed that there is always a path between any two elements on which the permeability increases monotonically, then the relative path permeability is always unity, even at boundary nodes $\boldsymbol{x}_n \in \Gamma_D$.

3. A posteriori error estimator.

3.1. Decomposition of the error. The purpose of the present work is to develop methods for obtaining computable estimators for the error $e = u - u_{\mathsf{nc}}$ in the nonconforming approximation measured in the energy norm denoted by $(A\,\mathbf{grad}_{\mathsf{nc}}\, e, \mathbf{grad}_{\mathsf{nc}}\, e)^{1/2}$. The following Helmholtz-type decomposition is essentially taken from Dari et al. [10].

LEMMA 3.1. *Let*

$$\mathcal{H} = \left\{ w \in H^1(\Omega) : \int_\Omega w \,\mathrm{d}\boldsymbol{x} = 0 \text{ and } \frac{\partial w}{\partial s} = 0 \text{ on } \Gamma_N \right\}.$$

The error e may be decomposed into the form

$$A\,\mathbf{grad}_{\mathsf{nc}}\, e = A\,\mathbf{grad}\,\phi + \mathbf{curl}\,\psi, \tag{3.1}$$

where $\phi \in H^1_E(\Omega)$ satisfies

$$(A\,\mathbf{grad}\,\phi, \mathbf{grad}\, v) = (A\,\mathbf{grad}_{\mathsf{nc}}\, e, \mathbf{grad}\, v) \quad \forall v \in H^1_E(\Omega) \tag{3.2}$$

and $\psi \in \mathcal{H}$ satisfies

$$(A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,w) = (\mathbf{grad}_{\mathsf{nc}}\,e, \mathbf{curl}\,w) \quad \forall w \in \mathcal{H}. \tag{3.3}$$

Moreover,

$$(A\,\mathbf{grad}_{\mathsf{nc}}\,e, \mathbf{grad}_{\mathsf{nc}}\,e) = (A\,\mathbf{grad}\,\phi, \mathbf{grad}\,\phi) + (A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi). \tag{3.4}$$

Proof. An application of the Lax–Milgram lemma shows that ϕ exists and is unique. Denote $\boldsymbol{w} = A(\mathbf{grad}_{\mathsf{nc}}\,e - \mathbf{grad}\,\phi) \in L_2(\Omega)^2$. With the aid of Green's formula and (3.2), we deduce that

$$0 = \int_\Omega v\,\mathbf{div}\boldsymbol{w}\,\mathrm{d}\boldsymbol{x} + \int_\Gamma v\,\boldsymbol{n}\cdot\boldsymbol{w}\,\mathrm{d}s \quad \forall v \in H^1_E(\Omega).$$

Consequently, $\boldsymbol{w}$ is divergence free in Ω and $\boldsymbol{n}\cdot\boldsymbol{w} = 0$ on Γ_N. Applying Theorem 3.1 in [15] shows that there exists $\psi \in H^1(\Omega)/\mathbb{R}$ such that

$$A(\mathbf{grad}_{\mathsf{nc}}\,e - \mathbf{grad}\,\phi) = \boldsymbol{w} = \mathbf{curl}\,\psi.$$

Furthermore, $\boldsymbol{n}\cdot\mathbf{curl}\,\psi = \boldsymbol{n}\cdot\boldsymbol{w} = 0$ on Γ_N, and we conclude that $\psi \in \mathcal{H}$. The characterization (3.3) and the orthogonality property (3.4) hold, provided that

$$\int_\Omega \mathbf{grad}\,\phi\cdot\mathbf{curl}\,w\,\mathrm{d}\boldsymbol{x} = 0 \quad \forall w \in \mathcal{H}.$$

This follows directly from an integration by parts on recalling that ϕ vanishes on Γ_D while $\boldsymbol{n}\cdot\mathbf{curl}\,w$ vanishes on Γ_N. ∎

Lemma 3.1 means that the error e in the nonconforming approximation may be split into two parts, as in (3.1). The nature of the contributions ψ and ϕ defined in Lemma 3.1 may be identified as follows. First, suppose that the nonconforming approximation u_{nc} happens to be conforming, by which we mean $u - u_{\mathsf{nc}} \in H^1_E(\Omega)$. The right-hand side of (3.3) then simplifies to

$$(\mathbf{grad}_{\mathsf{nc}}\,e, \mathbf{curl}\,w) = (\mathbf{grad}\,e, \mathbf{curl}\,w) = \int_\Gamma e\,\boldsymbol{n}\cdot\mathbf{curl}\,w\,\mathrm{d}s = 0,$$

since e vanishes on Γ_D and $\boldsymbol{n}\cdot\mathbf{curl}\,w = 0$ on Γ_N for $w \in \mathcal{H}$. Hence, if the approximation u_{nc} is conforming, then the contribution ψ vanishes. For this reason, we shall refer to ψ as the *nonconforming error*. The remaining contribution, ϕ, is referred to as the *conforming error*.

The splitting of the error into nonconforming and conforming components defines an orthogonal decomposition in the sense that the Pythagorean identity (3.4) holds. This means that the problem of obtaining a posteriori error estimators for the total error reduces to the derivation of estimators for the conforming and nonconforming errors *independently*. An estimator for the total error is then given by summing the estimators for the independent contributions.

3.2. Statement of main result. We summarize the results to be proved in sections 5 and 6. For $K \in \mathcal{P}$, let $\boldsymbol{\sigma}_K$ denote the function $\boldsymbol{\sigma}_K = -\frac{1}{2}\overline{f}_K(\boldsymbol{x} - \boldsymbol{x}_K)$, where $\boldsymbol{x}_K$ denotes the element centroid. The estimator for the conforming error is defined in terms of $\boldsymbol{\sigma}_K$ as follows:

$$\eta^2_{\mathsf{cf},K} = \big(A^{-1}\boldsymbol{\sigma}_K, \boldsymbol{\sigma}_K\big)_K = \frac{1}{48}\,\mathrm{meas}(K)\overline{f}^2_K \sum_{\ell=1}^{3} \boldsymbol{w}_\ell^\top A^{-1}\boldsymbol{w}_\ell, \tag{3.5}$$

where $\boldsymbol{w}_\ell$ is the position vector of vertex ℓ of the element relative to the centroid.

The estimator for the nonconforming error is defined in terms of a piecewise affine function $S(u_{\mathsf{nc}})$ on $\mathcal{P}$ obtained by smoothing the nonconforming approximation with values at vertices given by

$$S(u_{\mathsf{nc}})(\boldsymbol{x}_n) = \begin{cases} q(\boldsymbol{x}_n) & \text{if } \boldsymbol{x} \in \Gamma_D, \\ \displaystyle\sum_{K\in\Omega_n} \omega_{K,n}\, u_{\mathsf{nc}|K}(\boldsymbol{x}_n) & \text{otherwise,} \end{cases} \tag{3.6}$$

where the weights $\omega_{K,n}$ are defined by

$$\omega_{K,n} = \frac{\Lambda_K^{1/2}}{\sum_{K'\subset\Omega_n} \Lambda_{K'}^{1/2}}. \tag{3.7}$$

The estimator on element K is then given by

$$\eta_{\mathsf{nc},K}^2 = \left(A\,\mathbf{grad}_{\mathsf{nc}}(u_{\mathsf{nc}} - S(u_{\mathsf{nc}})), \mathbf{grad}_{\mathsf{nc}}(u_{\mathsf{nc}} - S(u_{\mathsf{nc}}))\right)_K. \tag{3.8}$$

The main result may now be stated. For ease of exposition, we suppose the Dirichlet data is homogeneous, although this assumption is subsequently relaxed.

THEOREM 3.2. *Let Δ_K denote the local data oscillation on element K given by*

$$\Delta_K = \operatorname{osc}(f,K) + \operatorname{osc}(g, \{\gamma \subset \Gamma_N \cap \partial K\}), \tag{3.9}$$

and let $\Upsilon_{KK'}$ be the relative path permeability defined in (2.12)*. Then the conforming error may be estimated as*

$$(A\,\mathbf{grad}\,\phi, \mathbf{grad}\,\phi) \le \sum_{K\in\mathcal{P}} (\eta_{\mathsf{cf},K} + C\lambda_K^{-1/2}\Delta_K)^2 \tag{3.10}$$

and

$$c\eta_{\mathsf{cf},K} \le \Upsilon_K^{1/2} (A\,\mathbf{grad}\,\phi, \mathbf{grad}\,\phi)_K^{1/2} + \lambda_K^{-1/2} \operatorname{osc}(f,K). \tag{3.11}$$

Furthermore, the nonconforming error may be estimated as

$$(A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi) \le \sum_{K\in\mathcal{P}} \eta_{\mathsf{nc},K}^2 \tag{3.12}$$

and

$$c\eta_{\mathsf{nc},K}^2 \le (A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi)_{\widetilde{K}} \sum_{K'\subset\widetilde{K}} \Upsilon_{KK'}, \tag{3.13}$$

where $\widetilde{K}$ denotes the patch formed from those elements sharing a common vertex with element K, and c and C are positive constants that depend only on the element geometry. Consequently, the total error may be estimated as

$$(A\,\mathbf{grad}_{\mathsf{nc}}\, e, \mathbf{grad}_{\mathsf{nc}}\, e) \le \sum_{K\in\mathcal{P}} (\eta_{\mathsf{cf},K} + C\lambda_K^{-1/2}\Delta_K)^2 + \eta_{\mathsf{nc},K}^2 \tag{3.14}$$

along with a corresponding lower bound on the total error.

Proof. The estimates for the conforming error are proved in Lemmas 5.2 and 5.3, while the estimates for the nonconforming error are proved in Lemma 6.2 and Theorem 6.4. The upper and lower bounds on the total error follow at once from (3.4). ∎

The resulting estimator is reminiscent of the estimators found in [10, 12, 16, 18]. Here, it is shown that the estimator provides a numerical upper bound for the total global error which does not involve unknown constants. Moreover, the estimator is shown to be efficient and robust (provided the relative path permeability remains bounded) without additional assumptions on the regularity of the true solution or on the mesh.

4. Numerical examples. The behavior of the estimator (3.14) is illustrated for some simple representative problems in this section.

4.1. Laplacian on L-shaped domain. Figure 4.1 shows the sequence of adaptively refined meshes for the solution of Laplace's equation on an L-shaped domain with pure Dirichlet boundary conditions chosen so that the solution is given by $u(r,\theta) = r^{2/3}\sin(2\theta/3)$. The conforming error vanishes in this case and the local error estimator on element K reduces to $\eta_{\mathsf{nc},K}$.

The effectivity index is found to vary in the range 1.5–1.6 in this example, as shown in Table 4.1. The sequence of meshes was constructed adaptively by selecting for refinement all elements where the local error indicator exceeds 30% of the value of the largest local indicator.

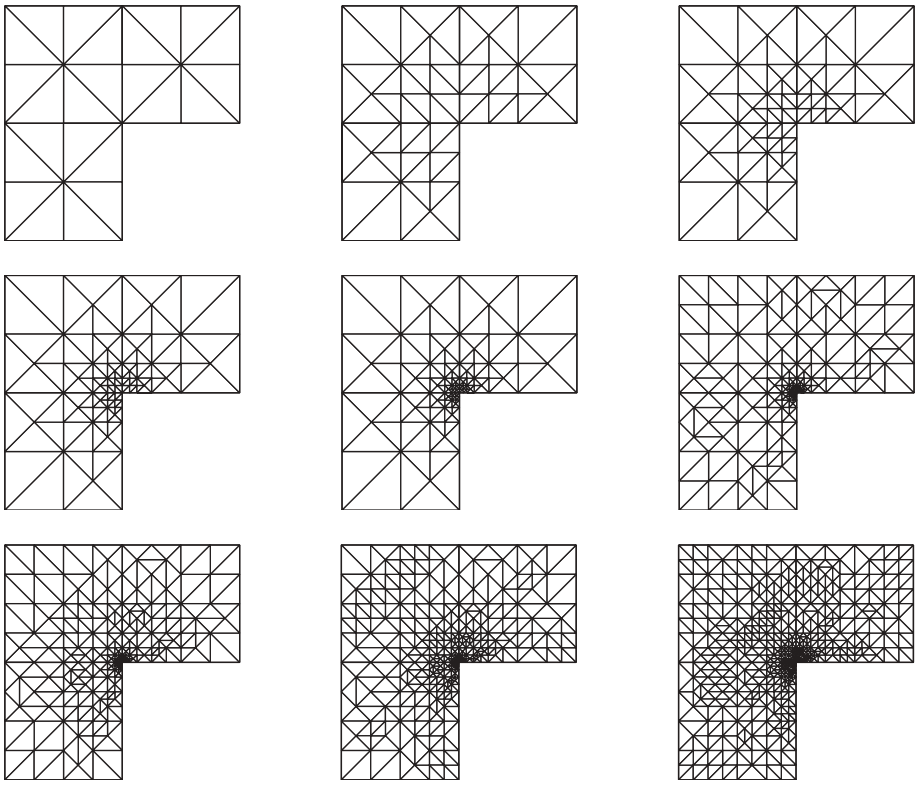


FIG. 4.1. *Sequence of adaptively refined meshes for Laplace example.*

TABLE 4.1
Comparison of estimated and true error for L-shaped domain.

Ndofs	True	Estimated	Effectivity
44	8.87(-2)	2.26(-1)	1.60
99	4.02(-2)	9.97(-2)	1.57
154	2.15(-2)	5.15(-2)	1.55
209	1.41(-2)	3.25(-2)	1.52
264	1.11(-2)	2.50(-2)	1.50
446	5.86(-3)	1.35(-2)	1.52
640	3.86(-3)	8.84(-3)	1.51
933	2.60(-3)	5.96(-3)	1.51
1487	1.57(-3)	3.57(-3)	1.51

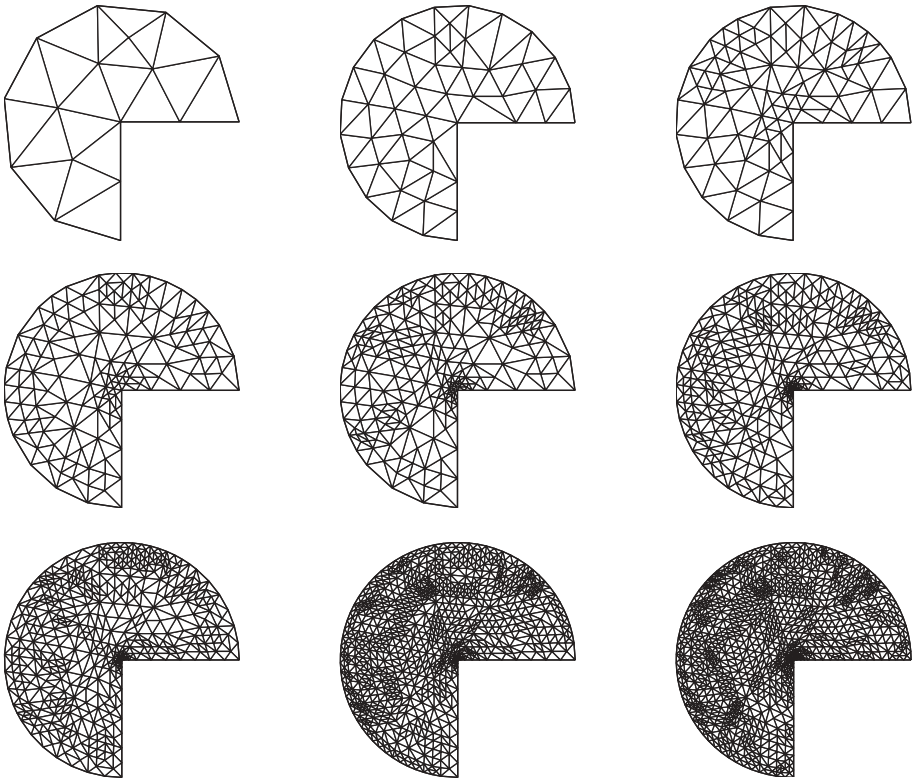


FIG. 4.2. *Sequence of adaptively refined meshes for Poisson problem.*

4.2. Nonzero source term. Figure 4.2 shows the sequence of adaptively refined meshes for the solution of Poisson's equation with homogeneous Dirichlet boundary conditions and source term chosen so that the solution is given by $u(r,\theta) = (r^{2/3} - r^2)\sin(2\theta/3)$. In this example, the conforming and nonconforming errors are both nonzero. The performance of the error estimator is shown in Table 4.2 along with the contributions from the conforming and nonconforming components of the error.

4.3. Variable permeability. The performance of the estimator in the case of variable permeability will be illustrated by considering the simple problem on the domain shown in Figure 4.3 with scalar permeability $A = a_\ell I$ on Ω_ℓ and prescribed flux $g = \boldsymbol{n} \cdot \mathbf{grad}(x^2 - y^2)$ on $\partial\Omega$. The true solution on subdomain Ω_ℓ is given by

TABLE 4.2
Comparison of estimated and true error for Poisson.

Ndofs	True	Estimated	Conforming	Nonconforming	Effectivity
36	4.75(-1)	8.86(-1)	2.13(-1)	6.73(-1)	1.37
119	1.92(-1)	3.58(-1)	7.37(-2)	2.85(-1)	1.36
219	1.11(-1)	2.15(-1)	4.43(-2)	1.70(-1)	1.39
384	6.62(-2)	1.28(-1)	2.73(-2)	1.00(-1)	1.39
647	3.99(-2)	7.61(-2)	1.54(-2)	6.07(-2)	1.38
923	2.61(-2)	5.19(-2)	1.08(-2)	4.11(-2)	1.41
1377	1.84(-2)	3.62(-2)	7.94(-3)	2.82(-2)	1.40
2633	9.05(-3)	1.71(-2)	3.59(-3)	1.35(-2)	1.38
3302	7.10(-3)	1.36(-2)	3.06(-3)	1.05(-2)	1.38

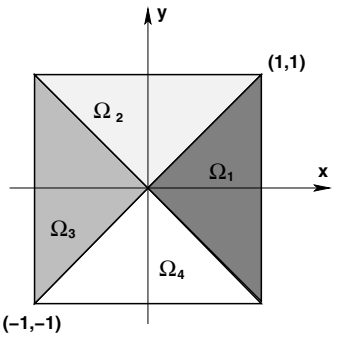


FIG. 4.3. *Geometry and subdomains for problem with variable permeability.*

$u(x,y) = (x^2 - y^2)/a_\ell$. Consider the situation where the local permeability is given by $a_1 = 1$, $a_2 = \alpha^2$, $a_3 = 1$, and $a_4 = \alpha^4$ with an initial mesh consisting of four elements coinciding with the subdomains shown in Figure 4.3.

The smoothing operator S may in principle be constructed using any choice of weights satisfying the condition

$$\sum_{K \subset \Omega_n} \omega_{K,n} = 1. \tag{4.1}$$

This requirement leaves considerable latitude in the selection of the weights. For instance, one might even choose all but one of weights to vanish, as suggested by Schieweck [18]. The obvious choice whereby the weights are chosen to be equal has been utilized by Oswald [17] in the context of multigrid methods. An alternative choice, advocated by Destuynder and Métivet [12, eqn. (17)], is to take $\omega_{K,n}$ proportional to the area of element K. If the mesh is locally quasi-uniform, then the latter two choices are not significantly different. Here, we are specifically concerned with the approximation of problems with highly varying local permeability, and the choice of weights given in (3.7) reflects this by depending on the value of the local permeability. The weights (3.7) are equal if the node $\boldsymbol{x}_n$ is located inside a subdomain where the local permeability is constant but differ markedly on interfaces and crosspoints between subdomains.

The ratio of the estimated error to the true error, $\mathrm{Eff}(\alpha)$, of the estimators obtained using the standard smoothing operator S with equal weighting and the estimator obtained using the weighted smoothing operator with weights depending on the permeability as in (3.7) can be computed explicitly on the initial mesh. For the

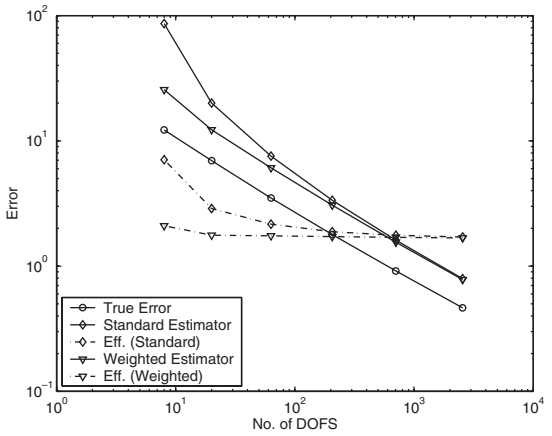


FIG. 4.4. *Comparison of standard estimator with weighted scheme for model problem with variable permeability.*

standard scheme, we obtain

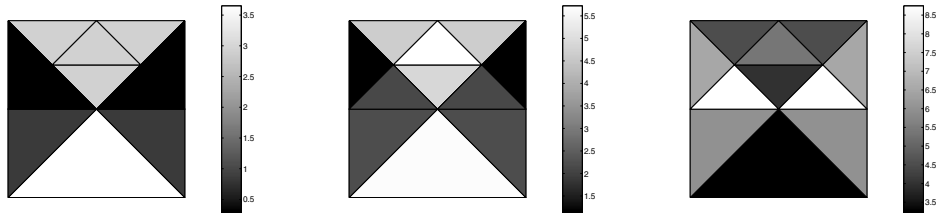
$$\mathrm{Eff}(\alpha)^2 = \begin{cases} \dfrac{3}{8}\alpha^4 + \mathcal{O}(1) & \text{as } \alpha \to \infty, \\ \dfrac{15}{32}\alpha^{-4} + \mathcal{O}(\alpha^{-2}) & \text{as } \alpha \to 0, \end{cases}$$

while for the weighted estimator,

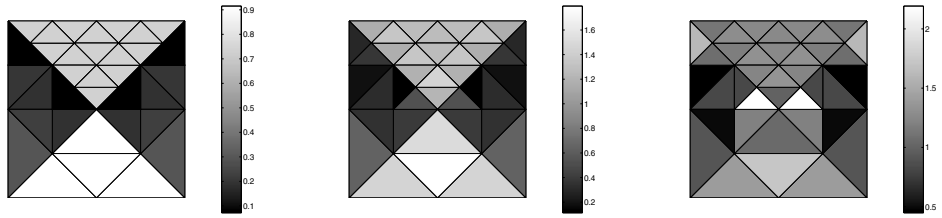
$$\mathrm{Eff}(\alpha)^2 = \begin{cases} \dfrac{27}{4} + \mathcal{O}(\alpha^{-1}) & \text{as } \alpha \to \infty, \\ \dfrac{39}{8} + \mathcal{O}(\alpha) & \text{as } \alpha \to 0. \end{cases}$$

The performance of the standard estimator clearly degenerates rapidly as the local permeability α is varied. The performance of the estimators when α^2 is taken to be 0.1 is presented in Figure 4.4. As would be expected, as the mesh is refined, both estimators tend to give the same value. However, even with this relatively modest value of α, the standard scheme provides very poor estimates for the error on the coarser meshes. The distribution of the local error estimates is compared with the actual local errors, as shown in Figure 4.5. Obviously, the estimators coincide away from interfaces. However, on interfaces and at the crosspoint, it is observed that the weighted estimator gives a more accurate picture of the distribution of the true error. Here, the true error was used to construct the sequence of adaptively refined meshes. Essentially the same sequence of meshes would be obtained if the weighted estimator were used, while if the standard estimator were to be used, then one would obtain a completely different sequence of meshes.

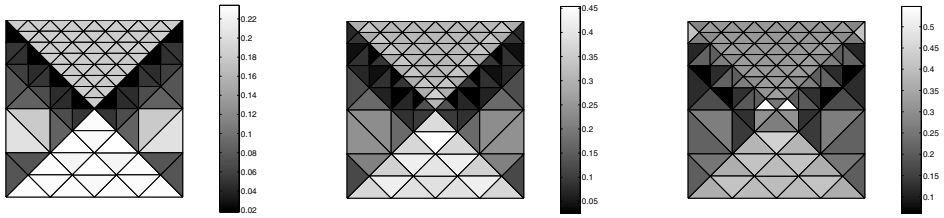
4.4. Neither term can be dropped from the estimator. The fact that the estimator provides an upper bound for the solution of the Laplace equation shows that the nonconforming term must be present in the estimator. It is less clear that the interior residual term must also be present. Consider the Poisson equation on a unit square with pure Neumann data chosen so that the true solution is given by $3x^2 - 2xy + 3y^2$. Observe that the oscillation of f vanishes. Suppose that the



(a) Mesh 2



(b) Mesh 3



(c) Mesh 4

FIG. 4.5. *Distribution of true local error (left) and estimated local error for weighted (center) and standard (right) estimators applied to model problem with variable permeability.*

solution is approximated using the mesh obtained by subdividing the square into a uniform mesh of right-angled triangles with hypotenuse in the direction $(1,1)$. In this scenario, it is found that the nonconforming finite element approximation u_{nc} is actually *conforming*, which means that the nonconforming term vanishes, yet the true error is obviously nonzero. This example shows that the interior residual term is essential for the upper bound and therefore cannot be removed in general.

5. Estimation of the conforming error.

5.1. Upper bound. Our objective is to derive a representation formula for the conforming component ϕ of the error in terms of quantities that can be evaluated explicitly or estimated in terms of the data oscillation. The following preparatory result will be useful in this direction. Let $\overline{\boldsymbol{x}}_K$ denote the centroid of element K, and define

$$\boldsymbol{\sigma}_K = -\frac{1}{2}\overline{f}_K(\boldsymbol{x} - \overline{\boldsymbol{x}}_K), \tag{5.1}$$

where $\overline{f}_K$ denotes the (constant) average value of the data f on element K. The function $\boldsymbol{\sigma}_K$ has the following property.

LEMMA 5.1. *Let $\boldsymbol{\sigma}_K$ be defined as in* (5.1). *Then*

$$(\boldsymbol{\sigma}_K, \mathbf{grad}\, v)_K = (\overline{f}_K, v - \Pi_{\mathsf{nc}}\, v)_K \quad \forall v \in H^1_E(K), \tag{5.2}$$

where Π_{nc} *is the nonconforming interpolation operator defined in* (2.6).

Proof. Let $v \in H^1_E(K)$. Then integration by parts gives

$$(\boldsymbol{\sigma}_K, \mathbf{grad}\, v)_K = -\int_K v\, \mathbf{div}\, \boldsymbol{\sigma}_K \,\mathrm{d}\boldsymbol{x} + \int_{\partial K} v \boldsymbol{n}_K \cdot \boldsymbol{\sigma}_K \,\mathrm{d}s,$$

where $\boldsymbol{n}_K$ denotes the unit outward normal on ∂K. Inserting the expression for $\boldsymbol{\sigma}_K$ into the first term on the right-hand side gives

$$-\int_K v\, \mathbf{div}\, \boldsymbol{\sigma}_K \,\mathrm{d}\boldsymbol{x} = (\overline{f}_K, v)_K$$

since $\overline{f}_K$ is constant. It therefore suffices to show that the second term satisfies

$$\int_{\partial K} v \boldsymbol{n}_K \cdot \boldsymbol{\sigma}_K \,\mathrm{d}s = -(\overline{f}_K, \Pi_{\mathsf{nc}}\, v)_K. \tag{5.3}$$

Substituting the expression for $\boldsymbol{\sigma}_K$ gives

$$\int_{\partial K} v \boldsymbol{n}_K \cdot \boldsymbol{\sigma}_K \,\mathrm{d}s = -\frac{1}{2}\overline{f}_K \sum_{\gamma \subset \partial K} \int_\gamma v \boldsymbol{n}_K \cdot (\boldsymbol{x} - \overline{\boldsymbol{x}}_K) \,\mathrm{d}s.$$

Elementary geometry reveals that $\boldsymbol{n}_K \cdot (\boldsymbol{x} - \overline{\boldsymbol{x}}_K)|_\gamma = 2\,\mathrm{meas}(K)/3\,\mathrm{meas}(\gamma)$, and hence

$$\int_\gamma v \boldsymbol{n}_K \cdot (\boldsymbol{x} - \overline{\boldsymbol{x}}_K) \,\mathrm{d}s = \frac{2}{3}\,\mathrm{meas}(K)\overline{v}_\gamma,$$

where $\overline{v}_\gamma$ denotes the average value of v on the edge γ, and the left-hand side of (5.3) may therefore be written in the form

$$\int_{\partial K} v \boldsymbol{n}_K \cdot \boldsymbol{\sigma}_K \,\mathrm{d}s = -\frac{1}{3}\,\mathrm{meas}(K)\overline{f}_K \sum_{\gamma \subset \partial K} \overline{v}_\gamma. \tag{5.4}$$

An elementary computation using (2.6) reveals that

$$\int_K \Pi_{\mathsf{nc}}\, v \,\mathrm{d}\boldsymbol{x} = \frac{1}{3}\,\mathrm{meas}(K) \sum_{\gamma \subset \partial K} \overline{v}_\gamma,$$

and the identity (5.3) now follows by combining this result with (5.4). ∎

We now turn to the error representation formula. Let $v \in H^1_E(\Omega)$ be given; then, using (3.2) and (2.2), we have

$$\begin{aligned}(A\,\mathbf{grad}\,\phi, \mathbf{grad}\, v) &= (A\,\mathbf{grad}_{\mathsf{nc}}\, e, \mathbf{grad}\, v)\\ &= (f, v) + \textstyle\int_{\Gamma_N} gv \,\mathrm{d}s - (A\,\mathbf{grad}_{\mathsf{nc}}\, u_{\mathsf{nc}}, \mathbf{grad}\, v).\end{aligned} \tag{5.5}$$

By integrating by parts and observing that $\mathbf{div}(A\,\mathbf{grad}_{\mathsf{nc}}\, u_{\mathsf{nc}})$ vanishes, we obtain

$$(A\,\mathbf{grad}_{\mathsf{nc}}\, u_{\mathsf{nc}}, \mathbf{grad}(v - \Pi_{\mathsf{nc}}\, v))_K = \int_{\partial K} \boldsymbol{n}_K \cdot A\,\mathbf{grad}_{\mathsf{nc}}\, u_{\mathsf{nc}}|_K\, (v - \Pi_{\mathsf{nc}}\, v) \,\mathrm{d}s = 0, \tag{5.6}$$

where the final step follows from the definition of Π_{nc} and the fact that the normal component of $A\,\mathbf{grad}_{\mathsf{nc}}\,u_{\mathsf{nc}}$ is piecewise constant on the element edges. Accumulating contributions over all elements shows that, in view of (2.4),

$$\begin{aligned}(A\,\mathbf{grad}_{\mathsf{nc}}\,u_{\mathsf{nc}},\mathbf{grad}\,v) &= (A\,\mathbf{grad}_{\mathsf{nc}}\,u_{\mathsf{nc}},\mathbf{grad}\,\Pi_{\mathsf{nc}}\,v)\\ &= (f,\Pi_{\mathsf{nc}}\,v)+\textstyle\int_{\Gamma_N} g\,\Pi_{\mathsf{nc}}\,v\,\mathrm{d}s.\end{aligned}\tag{5.7}$$

Equation (5.5) may therefore be rewritten as

$$(A\,\mathbf{grad}\,\phi,\mathbf{grad}\,v)=(f,v-\Pi_{\mathsf{nc}}\,v)+\int_{\Gamma_N} g(v-\Pi_{\mathsf{nc}}\,v)\,\mathrm{d}s.\tag{5.8}$$

Thanks to Lemma 5.1,

$$(f,v-\Pi_{\mathsf{nc}}\,v)_K=(\boldsymbol{\sigma}_K,\mathbf{grad}\,v)_K+(f-\overline{f}_K,v-\Pi_{\mathsf{nc}}\,v)_K,$$

while properties of the nonconforming interpolation operator mean that

$$\int_\gamma \overline{g}_\gamma(v-\Pi_{\mathsf{nc}}\,v)\,\mathrm{d}s=0,$$

where $\overline{g}_\gamma$ is the (constant) average value of g on an edge $\gamma\subset\Gamma_N$. With the aid of these results, we arrive at the following representation formula for the conforming component of the error:

$$(A\,\mathbf{grad}\,\phi,\mathbf{grad}\,v)=\sum_{K\in\mathcal{P}} R_K(v)\tag{5.9}$$

for all $v\in H^1_E(\Omega)$, where

$$R_K(v)=(\boldsymbol{\sigma}_K,\mathbf{grad}\,v)_K+(f-\overline{f}_K,v-\Pi_{\mathsf{nc}}\,v)_K+\sum_{\gamma\subset\Gamma_N\cap\partial K}\int_\gamma(g-\overline{g}_\gamma)(v-\Pi_{\mathsf{nc}}\,v)\,\mathrm{d}s.$$

The first term is estimated using a Cauchy–Schwarz inequality to obtain

$$(\boldsymbol{\sigma}_K,\mathbf{grad}\,v)_K\le(A^{-1}\boldsymbol{\sigma}_K,\boldsymbol{\sigma}_K)^{1/2}(A\,\mathbf{grad}\,v,\mathbf{grad}\,v)_K^{1/2},$$

while the remaining terms are estimated in terms of the data oscillation (2.9) and (2.10) using (2.3) and (2.8) to derive

$$(f-\overline{f}_K,v-\Pi_{\mathsf{nc}}\,v)_K\le C\lambda_K^{-1/2}\,\mathrm{osc}(f,K)(A\,\mathbf{grad}\,v,\mathbf{grad}\,v)_K^{1/2}$$

and

$$\begin{aligned}&\sum_{\gamma\subset\Gamma_N\cap\partial K}\int_\gamma(g-\overline{g}_\gamma)(v-\Pi_{\mathsf{nc}}\,v)\,\mathrm{d}s\\ &\le C\lambda_K^{-1/2}\,\mathrm{osc}(g,\{\gamma\subset\Gamma_N\cap\partial K\})(A\,\mathbf{grad}\,v,\mathbf{grad}\,v)_K^{1/2}.\end{aligned}$$

By inserting these bounds into (5.9), choosing $v=\phi$, and using a discrete Cauchy–Schwarz inequality, we arrive at the following estimate for the conforming component of the error.

LEMMA 5.2. *Let $\boldsymbol{\sigma}_K$ be defined as in* (5.1)*; then there exists a positive constant C depending only on the element geometry, such that*

$$(A\,\mathbf{grad}\,\phi, \mathbf{grad}\,\phi) \le \sum_{K\in\mathcal{P}} \left\{ (A^{-1}\boldsymbol{\sigma}_K, \boldsymbol{\sigma}_K)^{1/2} + C\lambda_K^{-1/2}\Delta_K \right\}^2, \tag{5.10}$$

where

$$\Delta_K = \operatorname{osc}(f, K) + \operatorname{osc}(g, \{\gamma \subset \Gamma_N \cap \partial K\}) \tag{5.11}$$

measures the local data oscillation over element K.

Remark 1. The estimator for the conforming error derived above may be regarded as the generalization of the equilibrated residual method [2] to nonconforming finite element approximation. Specifically, we define piecewise constant flux functions $g_K \in L_2(\partial K)$ by the rule

$$g_K = \frac{1}{h_\gamma}\left\{(\mathbf{grad}_{\mathsf{nc}}\, u_{\mathsf{nc}}, \mathbf{grad}_{\mathsf{nc}}\,\theta_\gamma)_K - (f, \theta_\gamma)_K\right\} \text{ on } \gamma \subset \partial K \backslash \Gamma_N \tag{5.12}$$

with $g_K = g$ on $\gamma \subset \partial K \cap \Gamma_N$. Choosing $v = \theta_\gamma$ in (2.4) shows that $g_K + g_{K'} = 0$ on $\partial K \cap \partial K'$. An elementary computation using properties (2.6) and (5.6) reveals that

$$\begin{aligned}&(f, v - \Pi_{\mathsf{nc}}\, v)_K + \int_{\partial K\cap\Gamma_N} g(v - \Pi_{\mathsf{nc}}\, v)\,\mathrm{d}s \\ &= (f, v)_K + \int_{\partial K} g_K v\,\mathrm{d}s - (A\,\mathbf{grad}_{\mathsf{nc}}\, u_{\mathsf{nc}}, \mathbf{grad}\, v)_K \quad \forall v \in H^1_E(K),\end{aligned}$$

where $H^1_E(K) = \{v|_K : v \in H^1_E(\Omega)\}$, and it is trivially seen that both sides vanish whenever v is an affine function. It follows that $\{g_K\}$ is a set of equilibrated fluxes in the sense of [2] and, thanks to (5.8), the conforming error may be decomposed into local contributions,

$$(A\,\mathbf{grad}\,\phi, \mathbf{grad}\, v) = \sum_{K\in\mathcal{P}} \left\{ (f, v)_K + \int_{\partial K} g_K v\,\mathrm{d}s - (A\,\mathbf{grad}_{\mathsf{nc}}\, u_{\mathsf{nc}}, \mathbf{grad}\, v)_K \right\}$$

for all $v \in H^1_E(\Omega)$, which is the starting point for the analysis of the equilibrated residual method.

5.2. Local lower bounds. The next result shows that the estimator suggested by Lemma 5.2 provides a lower bound for the conforming component of the error up to local data oscillation.

LEMMA 5.3. *There exists positive constant c depending only on the element geometry, such that*

$$c(A^{-1}\boldsymbol{\sigma}_K, \boldsymbol{\sigma}_K)_K^{1/2} \le \Upsilon_K^{1/2}(A\,\mathbf{grad}\,\phi, \mathbf{grad}\,\phi)_K^{1/2} + \lambda_K^{-1/2}\operatorname{osc}(f, K). \tag{5.13}$$

Proof. Let $\chi \in H^1_0(K)$ be the cubic (bubble) function whose value is unity at the element centroid. Then, up to constants independent of the element size h_K,

$$\int_K \chi\,\mathrm{d}\boldsymbol{x} \approx h_K^2; \quad \|\chi\|_{L_2(K)} \approx h_K; \quad \|\mathbf{grad}\,\chi\|_{L_2(K)} \approx 1.$$

Thanks to (2.3), the latter estimate implies that $(A\,\mathbf{grad}\,\chi,\mathbf{grad}\,\chi)_K \le C\Lambda_K$. Choosing $v=\chi$ in (5.8) gives

$$(A\,\mathbf{grad}\,\phi,\mathbf{grad}\,\chi)=(f,\chi)_K$$

since $\Pi_{\mathsf{nc}}\,\chi=0$. Equally well, for constant $\overline{f}_K$,

$$\overline{f}_K(1,\chi)_K=(A\,\mathbf{grad}\,\phi,\mathbf{grad}\,\chi)_K-(f-\overline{f}_K,\chi)_K,$$

and so, applying Cauchy–Schwarz inequalities and the properties of χ recorded above, we deduce that

$$\left|\overline{f}_K(1,\chi)_K\right|\le C\left\{\Lambda_K^{1/2}(A\,\mathbf{grad}\,\phi,\mathbf{grad}\,\phi)_K^{1/2}+h_K\left\|f-\overline{f}_K\right\|_{L_2(K)}\right\}.$$

Again exploiting properties of χ, we deduce that, up to constants depending only on the element geometry, $h_K\left\|\overline{f}_K\right\|_{L_2(K)}\approx\left|\overline{f}_K(1,\chi)_K\right|$ and hence

$$h_K\left\|\overline{f}_K\right\|_{L_2(K)}\le C\left\{\Lambda_K^{1/2}(A\,\mathbf{grad}\,\phi,\mathbf{grad}\,\phi)_K^{1/2}+\operatorname{osc}(f,K)\right\}. \tag{5.14}$$

Using (2.3) and (3.5), we deduce that

$$(A^{-1}\boldsymbol{\sigma}_K,\boldsymbol{\sigma}_K)_K\le C\lambda_K^{-1}h_K^2\left\|\overline{f}_K\right\|^2_{L_2(K)},$$

and then, thanks to (5.14), we obtain

$$(A^{-1}\boldsymbol{\sigma}_K,\boldsymbol{\sigma}_K)_K^{1/2}\le C\left\{\Upsilon_K^{1/2}(A\,\mathbf{grad}\,\phi,\mathbf{grad}\,\phi)_K^{1/2}+\lambda_K^{-1/2}\operatorname{osc}(f,K)\right\}$$

as claimed. ∎

6. Estimation of the nonconforming error. We turn to the problem of estimation of the nonconforming component of the total error defined by (3.3). The following result forms the basis for developing upper bounds.

LEMMA 6.1. *Let ψ be defined in* (3.3). *Then*

$$(A^{-1}\,\mathbf{curl}\,\psi,\mathbf{curl}\,\psi)=\min_{\substack{u^*\in H^1(\Omega):\\ u^*=q\text{ on }\Gamma_D}}(A\,\mathbf{grad}_{\mathsf{nc}}(u^*-u_{\mathsf{nc}}),\mathbf{grad}_{\mathsf{nc}}(u^*-u_{\mathsf{nc}})). \tag{6.1}$$

Proof. Let $u^*\in H^1(\Omega)$ satisfy $u^*=q$ on Γ_D. In particular, $u-u^*\in H^1_E(\Omega)$, and hence, applying Green's formula gives for each $w\in\mathcal{H}$,

$$(\mathbf{grad}_{\mathsf{nc}}(u-u^*),\mathbf{curl}\,w)=\int_{\partial\Omega}(u-u^*)\frac{\partial w}{\partial s}\,\mathrm{d}s=0$$

since the first term vanishes on Γ_D, while the second vanishes on Γ_N. Consequently,

$$(A^{-1}\,\mathbf{curl}\,\psi,\mathbf{curl}\,w)=(\mathbf{grad}_{\mathsf{nc}}\,e,\mathbf{curl}\,w)=(\mathbf{grad}_{\mathsf{nc}}(u^*-u_{\mathsf{nc}}),\mathbf{curl}\,w).$$

Therefore, choosing $w=\psi$ and applying a Cauchy–Schwarz inequality reveals that

$$(A^{-1}\,\mathbf{curl}\,\psi,\mathbf{curl}\,\psi)\le(A\,\mathbf{grad}_{\mathsf{nc}}(u^*-u_{\mathsf{nc}}),\mathbf{grad}_{\mathsf{nc}}(u^*-u_{\mathsf{nc}})).$$

It remains to show that the lower bound is attained. Let $\phi\in H^1_E(\Omega)$ be defined as in (3.2), and choose u^* to be the function $u-\phi$. Identity (3.1) reveals that

$$A\,\mathbf{grad}_{\mathsf{nc}}(u^*-u_{\mathsf{nc}})=A\,\mathbf{grad}_{\mathsf{nc}}(e-\phi)=\mathbf{curl}\,\psi,$$

which shows the lower bound is attained. ∎

The significance of Lemma 6.1 is, given *any* admissible function u^* in (6.1), we immediately obtain an upper bound on the nonconforming error. The accuracy of the bound will of course depend on the particular choice of function u^*. Equally well, the ease with which the bound may be evaluated will depend on the actual form of the function. Simple choices of u^* are ruled out by the condition on the Dirichlet boundary (except those where the Dirichlet data q is trivial), and it will be worthwhile to relax this restriction (at the expense of introducing an oscillation term for the Dirichlet data). We shall base our choice of u^* on a continuous, piecewise affine function $\mathcal{S}(u_{\mathsf{nc}})$ obtained by postprocessing the nonconforming approximation. The restriction of the function $\mathcal{S}(u_{\mathsf{nc}})$ to the Dirichlet boundary is chosen to be the continuous piecewise linear interpolant q_I of the Dirichlet data q at element vertices on Γ_D.

LEMMA 6.2. *Let $\mathcal{S}(u_{\mathsf{nc}})$ be any piecewise affine function whose restriction to the Dirichlet boundary Γ_D coincides with q_I. Then there exists a constant C depending only on the shape of the elements, such that*

$$\begin{aligned}(A^{-1}\,\mathbf{curl}\,\psi,\mathbf{curl}\,\psi)^{1/2} &\le (A\,\mathbf{grad}_{\mathsf{nc}}(u_{\mathsf{nc}}-\mathcal{S}(u_{\mathsf{nc}})),\mathbf{grad}_{\mathsf{nc}}(u_{\mathsf{nc}}-\mathcal{S}(u_{\mathsf{nc}})))^{1/2}\\ &\quad + C\Lambda_{\Gamma_D}^{1/2}\operatorname{osc}\left(\partial q/\partial s,\{\gamma:\gamma\subset\Gamma_D\}\right),\end{aligned}\tag{6.2}$$

where $\Lambda_{\Gamma_D}=\max\{\Lambda_K : K$ has an edge on $\Gamma_D\}$.

Proof. First, let $\gamma\subset\partial\mathcal{P}\cap\Gamma_D$ and observe that $q-q_I\in H_{00}^{1/2}(\gamma)$. Moreover, the convexity of the $H_{00}^{1/2}$-norm and standard (one-dimensional) approximation properties of the interpolant reveal that

$$\|q-q_I\|^2_{H_{00}^{1/2}(\gamma)}\le Ch_\gamma\,\|\partial q/\partial s\|^2_{L_2(\gamma)}\,.$$

The same argument applies if we replace q by $q-\alpha s$, where $\alpha\in\mathbb{R}$ is arbitrary and s denotes the arc-length, yielding the estimate

$$\|q-q_I\|_{H_{00}^{1/2}(\gamma)}\le C\operatorname{meas}(\gamma)\inf_{\alpha\in\mathbb{R}}\|\partial q/\partial s-\alpha\|^2_{L_2(\gamma)}=C\operatorname{osc}(\partial q/\partial s,\gamma)^2.\tag{6.3}$$

The function u^* is then chosen to be $u^*=\mathcal{S}(u_{\mathsf{nc}})+\xi$. The function $\xi\in H^1(\Omega)$ is defined elementwise by $\xi_{|K}=q-q_I$ on $\partial K\cap\Gamma_D$, $\xi_{|K}=0$ on $\partial K\backslash\Gamma_D$ and extended onto the domain interior as a harmonic function so that $\|\xi\|_{H^1(K)}\le C\,\|q-q_I\|_{H_{00}^{1/2}(\partial K\cap\Gamma_D)}$. Hence, thanks to (2.3) and (6.3),

$$(A\,\mathbf{grad}\,\xi,\mathbf{grad}\,\xi)^{1/2}\le C\Lambda_K^{1/2}\operatorname{osc}(\partial q/\partial s,\{\gamma:\gamma\subset\Gamma_D\}).$$

Inserting this choice of u^* in Lemma 6.1 and applying the triangle inequality gives the result claimed. ∎

6.1. Local smoothing operator. Given a particular choice of the function $S(u_{\mathsf{nc}})$, Lemma 6.2 shows how to obtain a computable upper bound. The tightness of the bound and the efficiency of the resulting estimator will hinge on the particular construction chosen for S.

The affine function $\mathcal{S}(u_{\mathsf{nc}})$ is uniquely defined by the values at the nodes of the partition given in (3.6). It is clear that S is a linear operator. The next result shows that S is continuous and can be bounded in terms of the choice of weights and the path permeability between elements discussed in section 2.4.

LEMMA 6.3. *Let $n \in \mathcal{N}$ and $K \in \Omega_n$. Then*

$$\begin{aligned}&\left|u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - \mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{x}_n)\right| \\ &\leq C \begin{cases} \left(A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi\right)_{\Omega_n}^{1/2} \sum_{K' \subset \Omega_n} \omega_{K',n} \lambda_{KK'}^{-1/2} & \text{if } \boldsymbol{x}_n \notin \Gamma_D, \\ \lambda_{K\Gamma_D}^{-1/2} \left(A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi\right)_{\Omega_n}^{1/2} + \operatorname{osc}(\partial q/\partial s, \{\gamma \in \mathcal{E}_n \cap \Gamma_D\}) & \\ \qquad\qquad \text{if } \boldsymbol{x}_n \in \Gamma_D. \end{cases}\end{aligned}$$

Proof. Case (i): $\boldsymbol{x}_n \notin \Gamma_D$. Inserting definition (3.6) and using property (4.1), we obtain

$$u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - \mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{x}_n) = \sum_{K' \subset \Omega_n} \omega_{K',n} \left(u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - u_{\mathsf{nc}|K'}(\boldsymbol{x}_n)\right). \tag{6.4}$$

To begin with, we consider the contribution to this quantity arising when an element K' shares a common edge γ with element K.

An elementary computation reveals that

$$u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - u_{\mathsf{nc}|K'}(\boldsymbol{x}_n) = \frac{h_\gamma}{2} \left[\frac{\partial\, u_{\mathsf{nc}}}{\partial s}\right]_\gamma, \quad \gamma = \partial K \cap \partial K'. \tag{6.5}$$

Let β_γ denote the continuous piecewise quadratic function that takes the value $3/4$ at the midpoint of edge γ and vanishes at all remaining nodes and midpoints. The function β_γ is supported on the patch $K \cup K'$. In (3.3), we choose $w \in \mathcal{H}$ to be the difference between β_γ and its (constant) average value over the domain Ω to obtain

$$(A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\beta_\gamma)_{K\cup K'} = (\mathbf{grad}_{\mathsf{nc}}\, e, \mathbf{curl}\,\beta_\gamma).$$

Integration by parts allows the right-hand side to be rewritten in the form

$$(A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\beta_\gamma)_{K\cup K'} = \int_\gamma \left[\frac{\partial\, u_{\mathsf{nc}}}{\partial s}\right]_\gamma \beta_\gamma \,\mathrm{d}s = \frac{h_\gamma}{2} \left[\frac{\partial\, u_{\mathsf{nc}}}{\partial s}\right]_\gamma,$$

where the fact that the jump is constant on an interior edge has been used. Together with (6.5), this identity implies

$$u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - u_{\mathsf{nc}|K'}(\boldsymbol{x}_n) = (A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\beta_\gamma)_{K\cup K'}. \tag{6.6}$$

This relation is valid for pairs of elements K and K' sharing a common edge γ. More generally, suppose elements K and K' share only a common node $\boldsymbol{x}_n$. The path $\wp^*(K, K')$ appearing in (2.11) links the elements K and K' by a set of elements having a common endpoint at $\boldsymbol{x}_n$. The set of edges shared by these elements is denoted by $\partial\wp^*(K, K')$. Relation (6.6) holds on each edge along the path, and so, by summing (6.6) over edges, we obtain a telescoping sum of differences of u_{nc} across neighboring edges, which simplifies to give

$$u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - u_{\mathsf{nc}|K'}(\boldsymbol{x}_n) = \left(A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\beta_{KK'}\right),$$

where $\beta_{KK'} = \sum_{\gamma \in \partial\wp^*(K,K')} \beta_\gamma$. Applying the Cauchy–Schwarz inequality gives the upper bound

$$\begin{aligned}&\left|u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - u_{\mathsf{nc}|K'}(\boldsymbol{x}_n)\right| \\ &\leq \left(A^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi\right)_{\Omega_n}^{1/2} \left(A^{-1}\,\mathbf{curl}\,\beta_{KK'}, \mathbf{curl}\,\beta_{KK'}\right)^{1/2}.\end{aligned} \tag{6.7}$$

The permeability on each element in the path $\wp^*(K, K')$ is bounded below by $\lambda_{KK'}$, implying that

$$\left(A^{-1}\,\mathbf{curl}\;\beta_{KK'}, \mathbf{curl}\;\beta_{KK'}\right) \leq \lambda_{KK'}^{-1}\,\|\mathbf{curl}\;\beta_{KK'}\|^2 \leq C\lambda_{KK'}^{-1},$$

where C depends only on the shape of the elements. Inserting this estimate into (6.7) and recalling (6.4) completes the proof in the first case.

Case (ii). If $\boldsymbol{x}_n \in \Gamma_D$, then $\mathcal{S}(u_{\mathsf{nc}})$ interpolates the Dirichlet data q at the node, so

$$u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - \mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{x}_n) = u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - q(\boldsymbol{x}_n).$$

First consider the case when element K abuts the Dirichlet boundary Γ_D. It is not difficult to show that

$$u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - q(\boldsymbol{x}_n) = \frac{h_\gamma}{2}\left.\frac{\partial\, u_{\mathsf{nc}}}{\partial s}\right|_\gamma - \left(q(\boldsymbol{x}_n) - q(\boldsymbol{m}_\gamma)\right). \tag{6.8}$$

Let μ_γ denote the average value of $\partial q/\partial s$ on edge γ,

$$\left(A^{-1}\,\mathbf{curl}\;\psi, \mathbf{curl}\;\beta_\gamma\right) = \frac{h_\gamma}{2}\left.\frac{\partial\, u_{\mathsf{nc}}}{\partial s}\right|_\gamma - \int_\gamma \left(\frac{\partial q}{\partial s} - \mu_\gamma\right)\beta_\gamma\,\mathrm{d}s - \frac{1}{2}\left(q(\boldsymbol{x}_n) - q(\boldsymbol{x}_m)\right).$$

Subtracting this from (6.8) gives

$$\begin{aligned} &u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - \mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{x}_n) \\ &= \left(A^{-1}\,\mathbf{curl}\;\psi, \mathbf{curl}\;\beta_\gamma\right) + \int_\gamma \left(\frac{\partial q}{\partial s} - \mu_\gamma\right)\beta_\gamma\,\mathrm{d}s + q(\boldsymbol{m}_\gamma) - \frac{1}{2}\left(q(\boldsymbol{x}_n) + q(\boldsymbol{x}_m)\right). \end{aligned}$$

Applying the Peano kernel theorem [11], we write

$$q(\boldsymbol{m}_\gamma) - \frac{1}{2}\left(q(\boldsymbol{x}_n) + q(\boldsymbol{x}_m)\right) = \int_\gamma \frac{\partial q}{\partial s} w\,\mathrm{d}s = \int_\gamma \left(\frac{\partial q}{\partial s} - \mu_\gamma\right) w\,\mathrm{d}s,$$

where $w = 1/2$ on $(\boldsymbol{x}_m, \boldsymbol{m}_\gamma)$ and $w = -1/2$ on $(\boldsymbol{m}_\gamma, \boldsymbol{x}_n)$. As a consequence, we obtain

$$u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - \mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{x}_n) = \left(A^{-1}\,\mathbf{curl}\;\psi, \mathbf{curl}\;\beta_\gamma\right) + \int_\gamma \left(\frac{\partial q}{\partial s} - \mu_\gamma\right)(\beta_\gamma - w)\;\mathrm{d}s.$$

Bounding the second term above by the oscillation of $\partial q/\partial s$ on γ leads to the estimate

$$\begin{aligned} &\left|u_{\mathsf{nc}|K}(\boldsymbol{x}_n) - \mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{x}_n)\right| \\ &\leq C\lambda_{K\Gamma_D}^{-1/2}\left(A^{-1}\,\mathbf{curl}\;\psi, \mathbf{curl}\;\psi\right)_{\Omega_n}^{1/2} + C\,\mathrm{osc}(\partial q/\partial s, \gamma). \end{aligned}$$

This proves the result in the case of an element $K \in \Omega_n$ which has an edge γ on Γ_D. The result may be extended to cover a general element $K' \subset \Omega_n$ by connecting to the exterior boundary along the path $\wp^*(K, \Gamma_D)$ and arguing as before. ∎

6.2. Efficiency of the estimator. The next result concerns the efficiency of the estimator when the weights are chosen as in (3.7).

THEOREM 6.4. *Let $\mathcal{S}(u_{\mathsf{nc}})$ denote the postprocessed approximation defined in (3.6). If $K \in \mathcal{P}$ has no vertices belonging to Γ_D, then there exists a positive constant c, independent of any mesh size, such that*

$$\begin{aligned} &c\left(A\,\mathbf{grad}_{\mathsf{nc}}(u_{\mathsf{nc}}-\mathcal{S}(u_{\mathsf{nc}})),\mathbf{grad}_{\mathsf{nc}}(u_{\mathsf{nc}}-\mathcal{S}(u_{\mathsf{nc}}))\right)_K \\ &\leq (A^{-1}\,\mathbf{curl}\,\psi,\mathbf{curl}\,\psi)_{\widetilde{K}}\sum_{K'\subset\widetilde{K}}\Upsilon_{KK'}, \end{aligned} \tag{6.9}$$

where $\widetilde{K}$ denotes the patch formed from those elements sharing a common vertex with element K. In the case where K has a vertex $\boldsymbol{x}_n \in \Gamma_D$, then the same estimate holds if the right-hand side is supplemented with the term

$$\Upsilon_{K\Gamma_D}(A^{-1}\,\mathbf{curl}\,\psi,\mathbf{curl}\,\psi)_{\widetilde{K}}+\operatorname{osc}(\partial q/\partial s,\{\gamma\in\mathcal{E}_n\cap\Gamma_D\})^2, \tag{6.10}$$

where $\mathcal{E}_n$ denotes the set of edges having an endpoint at $\boldsymbol{x}_n$.

Proof. Applying (2.3) and an inverse estimate shows that the left-hand side of (6.9) is bounded above by

$$C\Lambda_K h_K^{-2}\left\|u_{\mathsf{nc}}-\mathcal{S}(u_{\mathsf{nc}})\right\|^2_{L_2(K)},$$

and then evaluating this integral (using, for instance, the quadrature rule based on edge midpoints which is exact for quadratic functions) gives

$$C\Lambda_K\sum_{\gamma\subset\partial K}\left|u_{\mathsf{nc}}(\boldsymbol{m}_\gamma)-\mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{m}_\gamma)\right|^2.$$

The restriction of $u_{\mathsf{nc}|K}-\mathcal{S}(u_{\mathsf{nc}})$ to an edge $\gamma\subset\partial K$ is a linear function of arc-length, which means that the value at the midpoint $\boldsymbol{m}_\gamma$ is the average of the values at the endpoints of the edge, and therefore

$$\left|u_{\mathsf{nc}}(\boldsymbol{m}_\gamma)-\mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{m}_\gamma)\right|\leq\frac{1}{2}\sum_{\boldsymbol{x}_n\in\gamma}\left|u_{\mathsf{nc}|K}(\boldsymbol{x}_n)-\mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{x}_n)\right|.$$

Hence,

$$\begin{aligned} &\left(A\,\mathbf{grad}_{\mathsf{nc}}(u_{\mathsf{nc}}-\mathcal{S}(u_{\mathsf{nc}})),\mathbf{grad}_{\mathsf{nc}}(u_{\mathsf{nc}}-\mathcal{S}(u_{\mathsf{nc}}))\right)_K \\ &\leq C\Lambda_K\sum_{\boldsymbol{x}_n\in K}\left|u_{\mathsf{nc}|K}(\boldsymbol{x}_n)-\mathcal{S}(u_{\mathsf{nc}})(\boldsymbol{x}_n)\right|^2. \end{aligned} \tag{6.11}$$

Suppose that no vertex of K belongs to Γ_D; then the first estimate in Lemma 6.3 gives the following upper bound for (6.11)

$$\sum_{\boldsymbol{x}_n\in K}\left(A^{-1}\,\mathbf{curl}\,\psi,\mathbf{curl}\,\psi\right)_{\Omega_n}\leq\sum_{K'\subset\Omega_n}\frac{\min(\Lambda_K,\Lambda_{K'})}{\lambda_{KK'}},$$

which in turn may be bounded above by the right-hand side of (6.11). If K has a vertex $\boldsymbol{x}_n\in\Gamma_D$, then the second estimate in Lemma 6.3 must be used, which gives rise to the additional term (6.10). ∎

Acknowledgment. It is a pleasure to thank Prof. Dr. Willy Dörfler for his comments on an earlier version of the manuscript.

REFERENCES

[1] A. Agouzal, *A posteriori error estimator for nonconforming finite element methods*, Appl. Math. Lett., 7 (1994), pp. 1017–1033.
[2] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Pure Appl. Math., Wiley-Interscience, New York, 2000.
[3] W. Z. Bao and J. W. Barrett, *A priori and a posteriori error bounds for a nonconforming linear finite element approximation of a non-Newtonian flow*, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 843–858.
[4] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Numer. Math., 85 (2000), pp. 579–608.
[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, New York, 1994.
[6] C. Carstensen, S. Bartels, and S. Jansche, *A posteriori error estimates for nonconforming finite element methods*, Numer. Math., 92 (2002), pp. 233–256.
[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Elsevier, North-Holland, 1978; reprinted as Classics Appl. Math. 40, SIAM, Philadelphia, 2002.
[8] M. Crouzeix and P. A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO Modél. Math. Anal. Numér., 3 (1973), pp. 33–75.
[9] E. Dari, R. Duran, and C. Padra, *Error estimators for nonconforming finite-element approximations of the Stokes problem*, Math. Comp., 64 (1995), pp. 1017–1033.
[10] E. Dari, R. Duran, C. Padra, and V. Vampa, *A posteriori error estimators for nonconforming finite element methods*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 385–400.
[11] P. J. Davis, *Interpolation and Approximation*, Dover, New York, 1976.
[12] P. Destuynder and B. Métivet, *Explicit error bounds for a nonconforming finite element method*, SIAM J. Numer. Anal., 35 (1998), pp. 2099–2115.
[13] W. Dörfler and R. H. Nochetto, *Small data oscillation implies the saturation assumption*, Numer. Math., 91 (2002), pp. 1–12.
[14] R. Durán and C. Padra, *An error estimator for nonconforming approximations of a nonlinear problem*, in Finite Element Methods: Fifty Years of the Courant Element, M. Krizek, P. Neittaanmaki, and R. Stenberg, eds., Marcel Dekker, New York, 1994, pp. 201–205.
[15] V. Girault and P. A. Raviart, *Finite Element Methods for Navier Stokes Equations*, Springer Ser. Comput. Math. 5, Springer-Verlag, New York, 1986.
[16] R. H. W. Hoppe and B. Wohlmuth, *Element-oriented and edge-oriented local error estimators for nonconforming finite element methods*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 237–263.
[17] P. Oswald, *Intergrid transfer operators and multilevel preconditioners for nonconforming discretizations*, Appl. Numer. Math., 23 (1997), pp. 139–158.
[18] F. Schieweck, *A posteriori error estimates with post-processing for nonconforming finite elements*, ESAIM Math. Mod. Numer. Anal., 36 (2002), pp. 489–503.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2342–2365

PERTURBATIONS OF FORMS AND ERROR ESTIMATES FOR THE FINITE ELEMENT METHOD AT A POINT, WITH AN APPLICATION TO IMPROVED SUPERCONVERGENCE ERROR ESTIMATES FOR SUBSPACES THAT ARE SYMMETRIC WITH RESPECT TO A POINT*

ALFRED H. SCHATZ†

Abstract. We first derive a variety of local error estimates for $u - u_h$ at a point x_0, where u_h belongs to a finite element space S_r^h and is an approximation to u satisfying the local equations $A(u - u_h, \varphi) = F(\varphi)$ for all φ in S_r^h with compact support in a neighborhood of x_0. Here the $A(\cdot, \cdot)$ are bilinear forms associated with second order elliptic equations and the F are linear functionals. In the case that $F \equiv 0$ our results coincide with those of Schatz [*SIAM J. Numer. Anal.*, 38 (2000), pp. 1269–1293] but are improvements when $F \neq 0$. We apply these results to improve the superconvergence error estimates obtained by Schatz, Sloan, and Wahlbin [*SIAM J. Numer. Anal.*, 33 (1996), pp. 505–521] at points x_0 where the subspaces are symmetric with respect to x_0.

Key words. finite elements, weighted pointwise estimates, superconvergence, symmetry

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/S0036142902408131

1. Introduction, preliminaries, and statement of results. Our aim here is twofold. On the one hand we shall extend part of the results in Schatz [11] on finite element error estimates at a point, in particular those that deal with the perturbations of forms. These will then be applied to improve the results by Schatz, Sloan, and Wahlbin [12] on superconvergence at so-called symmetry points of subspaces.

The paper [11] dealt with two types of interior error estimates at a point for finite element approximations of solutions of second order elliptic problems. Roughly speaking, let $B_d \subset\subset \Omega \subset\subset \mathbb{R}^N$, $N \geq 2$, and let $A(\cdot, \cdot)$ be the usual type of locally defined bilinear form associated with the weak formulation of a second order elliptic problem on B_d (see below for a precise definition). Suppose further that u_h is a finite element approximation to a solution u. The first type of estimates derived in [11] were for $u - u_h$ satisfying, for all φ in an appropriate finite element space on B_d,

$$A(u - u_h, \varphi) = 0. \tag{1.1}$$

We shall be concerned with the second type of estimate, extensions of the first, that deal with $u - u_h$ satisfying a more general equation of the form

$$A(u - u_h, \varphi) = F(\varphi). \tag{1.2}$$

Here $F(\varphi)$ is a linear functional with certain technical properties which will be described in detail below. We remark that $F(\phi)$ may depend on $u - u_h$.

*Received by the editors May 22, 2002; accepted for publication (in revised form) August 28, 2003; published electronically March 31, 2005. This work was supported by the National Science Foundation.

http://www.siam.org/journals/sinum/42-6/40813.html

†Department of Mathematics, Cornell University, White Hall, Ithaca, NY 14853 (schatz@math.cornell.edu).

Typically, nonvanishing F arise in a variety of different problems. They often may be identified as perturbation terms of the bilinear form A. This includes some problems that, in the literature, are called "variational crimes" (see, e.g., Brenner and Scott [2]). As an example related to superconvergence, we note that in Nitsche and Schatz [8], they arose in proving superconvergence estimates for difference quotients of $u - u_h$ on translation invariant meshes. In general, for $u - u_h$ satisfying (1.1), difference quotients of $u - u_h$ do not satisfy (1.1) but rather an equation of the form (1.2).

In this paper the application we give is to prove superconvergence at a point with respect to which the mesh is symmetric. As in [12], we shall use the odd and even parts of the error to prove our results. In general the odd and even parts do not satisfy an equation of the form (1.1) but rather an equation of the form (1.2). For additional relevant problems see [2] and [18].

There are a great many papers in the literature devoted to maximum norm estimates on irregular grids. We shall mention a few where additional references may be found. In addition to the papers referred to above, we cite Natterer [6], Scott [16], Nitsche [7], Rannacher and Scott [9], Schatz and Wahlbin [14], and Schatz [10]. The paper most relevant to our work here is [11].

There is an enormous literature on superconvergence. Besides the main reference [12] given above, we mention two monographs devoted to superconvergence, namely, [5] and [18].

In section 1.1 we give some preliminaries for interior estimates, and section 1.2 contains statements of the main results on error estimates at a point with perturbed forms. Theorem 1 is concerned with estimates for the error $u - u_h$, and Theorem 2 deals with first derivatives of the error. Corollaries 1 and 2 are so-called asymptotic expansion inequalities that are useful for our application to superconvergence. These expansions are simple consequences of Theorems 1 and 2. Section 1.3 contains preliminaries for our study of the symmetry theory of superconvergence, and section 1.4 contains statements of the main results. In particular, Theorem 3 is concerned with superconvergence of $u - u_h$, and Theorem 4 deals with superconvergence for first derivatives of $u - u_h$. In section 1.5 we give applications of the estimates to specific boundary value problems.

Section 2 contains a proof of Theorem 1, and a discussion of the proof of Theorem 2 is given in section 2.1. Theorem 3 is proved in section 3, and Theorem 4 is proved in section 4. Finally, the appendix contains the main assumptions on the subspaces.

1.1. Notation and preliminaries. Let Ω be a bounded domain in $\mathbb{R}^N$, $N \geq 1$. For $m \geq 0$ an integer, and $1 \leq p \leq \infty$, $W_p^m(\Omega)$ will denote the usual Sobolev space of functions having distributional derivatives up to order m in L_p. The norm is given for $1 \leq p < \infty$ by

$$\|u\|_{W_p^m(\Omega)} = \left(\sum_{|\alpha| \leq m} \|D^\alpha u\|^p_{L_p(\Omega)} \right)^{1/p}$$

with the usual modification for $p = \infty$. $\mathring{W}_p^m(\Omega)$ is the closure of $C_0^\infty(\Omega)$ with respect to the norm of $W_p^m(\Omega)$. Furthermore, $|u|_{W_p^m(\Omega)}$ will denote the seminorm

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

For $m > 0$, $W_p^{-m}(\Omega)$ is the dual of $\mathring{W}_q^m(\Omega)$ with the norm

$$\|u\|_{W_p^{-m}(\Omega)} = \sup_{\substack{v \in \mathring{W}_q^m(\Omega) \\ \|v\|_{W_q^m(\Omega)}=1}} \int_\Omega uv dx, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

For given $x_0 \in \mathbb{R}^N$ and $d > 0$, let

$$B_d = B_d(x_0) = \{x \in \mathbb{R}^N : |x - x_0| < d\},$$

the ball of radius d centered at x_0. Suppose now that u satisfies the local equation in $B_d(x_0) \subset\subset \Omega$,

$$Lu = -\sum_{i,j=1}^N \frac{\partial}{\partial x_j}\Big(a_{ij}(x)\frac{\partial u}{\partial x_i}\Big) + \sum_{i=1}^N b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f \text{ in } B_d(x_0). \tag{1.3}$$

It will be assumed that the coefficients of L are smooth and that L is uniformly elliptic in $\overline{B}_d(x_0)$. That is, there exists a constant $C_{\text{ell}} > 0$, independent of $x \in \overline{B_d(x_0)}$ such that for all $\zeta \in \mathbb{R}^N$

$$C_{\text{ell}}|\zeta|^2 \le \sum_{i,j=1}^N a_{ij}\zeta_i\zeta_j.$$

In weak form, $u \in W_2^1(B_d(x_0))$ satisfies the local equations

$$\begin{aligned} A(u,v) &\equiv \int_{B_d}\left(\sum_{i,j=1}^N a_{ij}\frac{\partial u}{\partial x_i}\,\frac{\partial v}{\partial x_j} + \sum_{i=1}^N b_i\frac{\partial u}{\partial x_i}v + cuv\right)dx \\ &= \int_{B_d} fv dx \equiv (f,v) \text{ for all } v \in \mathring{W}_2^1(B_d(x_0)). \end{aligned} \tag{1.4}$$

Consider now a finite element approximation u_h of u. To this end, for each $0 < h < \frac{1}{2}$ and integer $r \ge 2$, let $S_r^h(\Omega) \subset W_\infty^1(\Omega)$ be a one-parameter family of finite element spaces defined on a disjoint partition $\{\tau_h\}$ of Ω that covers $B_d(x_0)$. It will be assumed that the partition is quasi-uniform of size h. Furthermore, it will be assumed that Assumptions A.1–A.4 of the appendix are satisfied. These are essentially the same assumptions as in [11]. They are satisfied by large classes of finite element spaces, e.g., by conforming piecewise polynomials defined on quasi-uniform partitions of Ω of size h, whose restriction to each disjoint set τ_h of the partition contains all polynomials of degree $\le r-1$. The spaces we use have the standard approximation property that they can approximate functions to order h^r and derivatives to order h^{r-1} in L_∞.

We shall consider approximations $u_h \in S_r^h(B_d)$ satisfying the local equations

$$A(u - u_h, \varphi) = F(\varphi) \text{ for all } \varphi \in \mathring{S}_r^h(B_d), \tag{1.5}$$

where $F(\varphi)$ is a linear functional on $\mathring{S}_r^h(B_d)$. Here $\mathring{S}_r^h(B_d)$ is the subspace of functions $\varphi \in S_r^h(B_d)$, whose support is contained in B_d.

Before stating our first results, we shall need to introduce various weighted Sobolev norms. Let $x_0 \in \mathbb{R}^N$ and $s \in \mathbb{R}$ be fixed and let $y \in \mathbb{R}^N$ be arbitrary. Consider the weight function

$$\sigma_{x_0,h}^s(y) = \Big(\frac{h}{|x_0 - y| + h}\Big)^s \tag{1.6}$$

and, for $p = 1, \infty$ the weighted seminorms

$$|u|_{W_p^j(B_d),x_0,s} = \sum_{|\alpha|=j} \|\sigma_{x_0,h}^s(y) D^\alpha u(y)\|_{L_p(B_d)} \tag{1.7}$$

and the norms

$$\|u\|_{W_p^m(B_d),x_0,s} = \sum_{j=0}^m |u|_{W_p^j(B_d),x_0,s}. \tag{1.8}$$

We shall need some norms for the linear functional F that will allow us to obtain estimates for $(u - u_h)(x_0)$ and $\frac{\partial}{\partial x_i}(u - u_h)(x_0)$. We shall first introduce these norms and then try to motivate their choice, after stating our main results in section 1.2.

We begin with the simplest. For $u \in \mathring{W}_1^1(B_d)$ and $s \in \mathbb{R}$ let

$$\|u\|_{W_1^{0,1}(B_d),x_0,s} = h^{-1}\|u\|_{L_1(B_d),x_0,s+1} + \|\nabla u\|_{L_1(B_d),x_0,s}. \tag{1.9}$$

Then we define the norm of a linear functional bounded with respect to this norm,

$$|||F|||_{W_\infty^{0,-1}(B_d),x_0,s} = \sup_{\substack{\varphi \in \varphi \in \mathring{W}_1^1(B_d) \\ \|\varphi\|_{W_1^{0,1}(B_d),x_0,-s} = 1}} F(\varphi). \tag{1.10}$$

Furthermore, let $\ell > k$ be an arbitrary but fixed integer, where $k = 1, 2$.

For $u \in \mathring{W}_1^\ell(B_d)$, we define the norm

$$\|u\|_{W_j^{0,\ell,k}(B_d),x_0} = \sum_{j=0}^{\ell} \sum_{|\alpha|=j} \Big(\ln \frac{1}{h}\Big)^{\overline{j}} \int_{B_d} (|x - x_0| + h)^{j-k} |D^\alpha u| dx, \tag{1.11}$$

where $\overline{j} = 1$ if $N = 2$, $k = 2$, and $j = 0$ and $\overline{j} = 0$ otherwise. Then we define the norm of a linear functional bounded with respect to this norm,

$$|||F|||_{W_\infty^{0,-\ell,k}(B_d),x_0} = \sup_{\substack{\varphi \in \mathring{W}_1^\ell(B_d) \\ \|\varphi\|_{W_1^{0,\ell,k}(B_d),x_0} = 1}} F(\varphi). \tag{1.12}$$

We are now in a position to state our main results for interior estimates.

1.2. A statement of results for interior error estimates. We begin with error estimates for $(u - u_h)$.

THEOREM 1. *Suppose that the assumptions on the finite element spaces $S_r^h(B_d)$ given in* A.1–A.4 *are satisfied on $B_d(x_0)$. Let $1 \le p \le \infty$, t be a nonnegative integer and $0 \le s \le r - 2$ be given. There exist positive constants C and k, which depend at most on p, t, N, s, C_{ell}, the maximum norms of the coefficients of A and a sufficient number of their derivatives in $\overline{B}_d(x_0)$, and the constants in Assumptions* A.1–A.4, *such that if $u \in W_\infty^1(B_d(x_0))$ and $u_h \in S_r^h(B_d)$ satisfy* (1.5), *where $d \ge kh$, then*

$$\begin{aligned} \|u - u_h\|_{L_\infty(B_h(x_0))} \le{}& Ch\Big(\ln \frac{d}{h}\Big)^{\overline{s}} \min_{\chi \in S_r^h(B_d(x_0))} \|u - \chi\|_{W_\infty^1(B_d(x_0)),x_0,s} \\ &+ Cd^{-t-N/p}\|u - u_h\|_{W_p^{-t}(B_d(x_0))} \\ &+ C\Big(h\Big(\ln \frac{d}{h}\Big)^{\overline{s}} |||F|||_{W_\infty^{0,-1}(B_d(x_0)),x_0,s} \\ &\quad + \Big(\ln \frac{d}{h}\Big)|||F|||_{W_\infty^{0,-r,2}(B_d(x_0)),x_0}\Big). \end{aligned} \tag{1.13}$$

Here $\overline{s} = 0$ if $0 \le s < r - 2$, and $\overline{s} = 1$ if $s = r - 2$.

We again remark that in the case that $F \equiv 0$, Theorem 1 coincides with Theorem 1.1 of [11]. However, in the case that $F \not\equiv 0$, then Theorem 1 is an improvement in that the norms of F are weaker.

The next result is the analogue of Theorem 1 for first derivatives of $u - u_h$.

THEOREM 2. *Suppose that* A.1–A.4, *the assumptions on the finite element spaces* $S_1^h(\infty)$ *given in the appendix, are satisfied on* $B_d(x_0)$. *Let* $1 \le p \le \infty$, t *be a nonnegative integer and* $0 \le s \le r-1$ *be given. There exist positive constants* C *and* k, *which depend at most on* p, t, N, s, C_{ell}, *the maximum norms of the coefficients of* A *and a sufficient number of their derivatives in* $B_d(x_0)$, *and the constants in* A.1–A.4, *such that if* $u \in W_\infty^1(B_d(x_0))$ *and* $u_h \in S_r^h(B_d(x_0))$ *satisfy* (1.5) *with* $d \ge kh$, *then*

$$
\begin{aligned}
&\|u-u_h\|_{W_\infty^1(B_h(x_0))} \\
&\quad \le C\Big(\ln\frac{d}{h}\Big)^{\overline{\overline{s}}} \min_{\chi\in S_r^h(B_d(x_0))} \Big(|u-\chi|_{W_\infty^1(B_d(x_0)),x_0,s} + d^{-1}|u-\chi|_{L_\infty(B_d(x_0)),x_0,s}\Big) \\
&\quad + Cd^{-t-N/p-1}\|u-u_h\|_{W_p^{-t}(B_d(x_0))} \qquad (1.14)\\
&\quad + C\Big(\ln\frac{d}{h}\Big)^{\overline{\overline{s}}} |||F|||_{W_\infty^{0,-1}(B_d(x_0)),x_0,s} \\
&\quad + C\Big(\ln\frac{d}{h}\Big) |||F|||_{W_\infty^{0,-r,1}(B_d(x_0)),x_0}.
\end{aligned}
$$

Here $\overline{\overline{s}} = 0$ *if* $0 \le s < r-1$, *and* $\overline{\overline{s}} = 1$ *if* $s = r-1$.

Analogous to the remark made for Theorem 1, we remark that in the case that $F \equiv 0$, Theorem 2 coincides with Theorem 1.2 of [11], and in the case $F \neq 0$, Theorem 2 is an improvement in that the norms of F are weaker.

Let us digress to motivate the choice of norms for the functional F. We limit ourselves to a discussion of Theorem 1. The choice in Theorem 2 follows in an analogous fashion. In Theorem 1, the contribution to the error $|(u-u_h)(x_0)|$ made by the functional F may be seen in (2.14), in the case $d=1$, to be a term of the form

$$|F(wg_h^{x_0}(x))|.$$

Here $w(x) \in C_0^\infty(B_{1/2}(x_0))$ is a cutoff function, and $g_h^{x_0}(x)$ is the Galerkin approximation to $g^{x_0}(x)$ on $B_1(x_0)$, where $g^{x_0}(x)$ is a smoothed Green's function for a certain auxilliary problem. Using the triangle inequality we have

$$|F(wg_h^{x_0}(x))| \le |F(wg^{x_0}(x))| + |F(w(g^{x_0}(x) - g_h^{x_0}(x))|.$$

The idea now is to treat these two terms separately, requiring F to have different properties in each case.

For the first term on the right we can require that F be bounded when acting on any space in which $g^{x_0}(x)$ is bounded. We shall prove in Lemma 2 that

$$\|g^{x_0}(\cdot)\|_{W_1^{0,r,2}(B_1(x_0)),x_0} \le C\ln\frac{1}{h}.$$

Note that this norm was designed to capture the behavior of the smoothed Green's function and it is natural to bound $F(wg^{x_0}(x))$ by

$$
\begin{aligned}
|F(wg^{x_0}(x))| &\le |||F|||_{W_\infty^{0,-r,2}(B_1(x_0)),x_0} \|wg\|_{W_1^{0,r,2}(B_1(x_0)),x_0} \\
&\le C\ln\frac{1}{h} |||F|||_{W_\infty^{0,-r,2}(B_1(x_0)),x_0}.
\end{aligned}
$$

Similarly for the second term $F(w(g^{x_0}-g_h^{x_0}))$, we seek a norm with respect to which $w(g^{x_0}(x)-g_h^{x_0}(x))$ is bounded. From (2.11) we have that for $0\le s\le r-2$, $g_h^{x_0}(x)$ approximates $g^{x_0}(x)$ according to

$$\|g^{x_0}(x)-g_h^{x_0}(x)\|_{W_1^{0,1}(B_d(x_0)),x_0,-s}\le Ch\Big(\ln\frac{1}{h}\Big)^{\overline{s}}.$$

Using this we easily obtain

$$\begin{aligned}|F(w(g^{x_0}(x)-g_h^{x_0}(x)))|&\le |||F|||_{W_\infty^{0,-1}(B_d(x_0)),x_0,s}\|w(g^{x_0}(x)-g_h^{x_0}(x))\|_{W_1^{0,1}(B_1(x_0)),x_0,-s}\\&\le Ch\Big(\ln\frac{1}{h}\Big)^{\overline{s}}|||F|||_{W_\infty^{0,-1}(B_d(x)),x_0,s}.\end{aligned}$$

We shall now state some simple consequences of Theorems 1 and 2, so-called error expansion inequalities, that will be the basic estimates used in our superconvergence study to be given in section 1.4. We shall need a strengthened form of A.1 that we state here as a separate assumption.

Assumption A.5. Assume that for $p=\infty$ and for $D_1\subset\subset D\subset B_d(x_0)$, with $\operatorname{dist}(D_1,\partial D)\ge k_0h$, the function $\chi\in S_r^h(D)$ in A.1 satisfies

$$\|u-\chi\|_{W_\infty^1(D_1)}\le Ch^{r-1}|u|_{W_\infty^r(D)},\tag{1.15}$$

where $|\cdot|_{W_\infty^r(D)}$ is the seminorm defined in (1.7).

COROLLARY 1. *Suppose that the conditions of Theorem* 1 *hold and in addition Assumption* A.5 *holds. Let $u\in W^\gamma(B_d(x_0))$, where γ is an integer $r+1\le\gamma\le 2r-2$. Let $\widehat{C}>0$ be a fixed but arbitrary constant with $\widehat{C}h\le d$, and let $\widehat{x}$ satisfy $|x_0-\widehat{x}|\le\widehat{C}h$. Then* (1.13) *holds with*

$$Ch\Big(\ln\frac{d}{h}\Big)^{\overline{s}}\min_{\chi\in S_r^h(B_d(x_0))}\|u-\chi\|_{W_\infty^1(B_d(x_0)),x_0,s}\tag{1.16}$$

in (1.14) *replaced by*

$$C\Big(\ln\frac{d}{h}\Big)^{\overline{\gamma}}\bigg(h^r\sum_{|\alpha|=r}|D^\alpha u(\widehat{x})|+\cdots+h^{\gamma-1}\sum_{|\alpha|=\gamma-1}|D^\alpha u(\widehat{x})|+h^\gamma|u|_{W_\infty^\gamma(B_d(x_0))}\bigg).\tag{1.17}$$

Here $\overline{\gamma}=0$ if $r+1\le\gamma<2r-2$ and $\overline{\gamma}=1$ if $\gamma=2r-2$. C is the same as in Theorem 1 *except that it also depends on $\widehat{C}$.*

The corresponding error expansion inequality for first derivatives of the error is as follows.

COROLLARY 2. *Suppose that the conditions of Theorem* 2 *hold and in addition Assumption* A.5 *holds. Let $u\in W_\infty^{\gamma+1}(B_d(x_0))$, where γ is an integer $r\le\gamma\le 2r-2$. Let $\widehat{C}$ be an arbitrary but fixed positive constant with $\widehat{C}h<d$, and let $\widehat{x}$ satisfy $|x_0-\widehat{x}|\le\widehat{C}h$. Then* (1.14) *holds with*

$$C\Big(\ln\frac{d}{h}\Big)^{\overline{\overline{s}}}\min_{\chi\in S_r^h(B_d(x_0))}\big(|u-\chi|_{W_\infty^1(B_d(x_0)),x_0,s}+d^{-1}|u-\chi|_{L_\infty(B_d(x_0)),x_0,s}\big)\tag{1.18}$$

replaced by

$$C\Big(\ln\frac{d}{h}\Big)^{\overline{\overline{\gamma}}}\bigg(h^{r-1}\sum_{|\alpha|=r}|D^\alpha u(\widehat{x})|+\cdots+h^{\gamma-1}\sum_{|\alpha|=\gamma}|D^\alpha u(\widehat{x})|+h^\gamma|u|_{W_\infty^{\gamma+1}(B_d(x_0))}\bigg).\tag{1.19}$$

Here $\overline{\overline{\gamma}} = 0$ *if* $r \le \gamma < 2r-2$ *and* $\overline{\overline{\gamma}} = 1$ *if* $\gamma = 2r-2$. C *is the same as in Theorem* 2 *except that it also depends on* $\widehat{C}$.

1.3. Superconvergence and subspaces symmetric with respect to a point—preliminaries. In [12], a new theory was presented for obtaining superconvergence. Assuming now that $u - u_h$ satisfies

$$A(u-u_h,\varphi) = 0 \quad \text{for all} \quad \varphi \in \mathring{S}_r^h(B_d(x_0)), \tag{1.20}$$

superconvergence at a point x_0 was achieved by requiring the subspaces to be symmetric with respect to x_0 in $B_d(x_0)$, which is defined as follows.

Assumption A.6. Let x_0 and $d > 0$ be given with $B_d(x_0) \subset\subset \Omega$. Assume that whenever $\varphi(x) \in S_r^h(B_d(x_0))$, then $\overline{\varphi}(x) = \varphi(x_0 - (x - x_0))$ also belongs to $S_r^h(B_d(x_0))$. In this case we shall call x_0 a symmetry point of $S_r^h(B_d(x_0))$.

Let us now turn to some simple examples of finite element spaces that satisfy Assumptions A.1–A.6. Let $\{\tau_h\}$ be a quasi-uniform partition of size h of Ω (which obviously covers $B_d(x_0)$). For our finite element spaces $S_r^h(B_d(x_0))$, we shall start with the set of functions that are in $C^\ell(B_d(x_0))$ for some integer $0 \le \ell \le r-2$ and that, on each element τ_h, are polynomials of the form

$$\varphi(x) = \sum_{\alpha \in I} C_\alpha x^\alpha, \ C_\alpha \text{ arbitrary constants.} \tag{1.21}$$

Here, I is a fixed set of multi-indices. Many well-known finite element spaces fitting this description satisfy Assumptions A.1–A.5. It is easy to see that A.6 is satisfied if and only if

(i) the partition $\{\tau_h\}$ is invariant under the antipodal mapping $x \to x_0 - (x - x_0)$, and

(ii) functions of the form (1.21) are invariant in form under the same antipodal mapping $x \to x_0 - (x - x_0)$.

We refer the reader to the examples of partitions $\{\tau_h\}$ satisfying (i) that are given in [12], and granting these we only have to check (ii). If $\alpha \in I$, $\alpha = (\alpha_1, \dots, \alpha_N)$, a multi-index, then x^α transforms to $(2x_0 - x)^\alpha$, and it is easy to see via a binomial expansion that $\alpha \in I$, if and only if $\beta \in I$ for every multi-index $\beta \le \alpha$. This is a necessary and sufficient condition that functions of the form (1.21), defined on partitions satisfying (i), satisfy A.6. A simple example is $P_{r-1}(\tau_h)$, the set of polynomials of degree $\le r-1$. In this case $\alpha \in I$, if and only if $|\alpha| \le r-1$. Another simple example is $Q_{r-1}(\tau_h)$, the tensor products of one-dimensional polynomials of degree $r-1$. In this case $\alpha \in I$ if and only if $\alpha_i \le r-1$ for $i = 1, \dots, N$.

1.4. Statement of results for superconvergence at symmetry points. We are now in a position to state our main superconvergence results. Applications to specific boundary value problems are given in section 1.5. We begin with estimates for the error $(u - u_h)(x_0)$ at a symmetry point x_0.

THEOREM 3. *Suppose that the conditions of Corollary* 1 *hold and in addition* A.6 *is satisfied. Then if* $r \ge 3$ *is odd,*

$$|(u-u_h)(x_0)| \le C\Big[h^{r+1}\Big(\ln\frac{d}{h}\Big)^2 \|u\|_{W_\infty^{r+1}(B_d(x_0))} + \Big(\ln\frac{d}{h}\Big) d^{-t-N/p} \|e\|_{W_p^{-t}(B_d(x_0))}\Big]. \tag{1.22}$$

Remark. We point out that the first term on the right side of (1.22) is superconvergent of order $h^{r+1}(\ln \frac{d}{h})^2$ and hence superconvergence of the same order will occur

at x_0 provided a similar bound holds for the second term on the right which measures the effects of the discretization error from outside of $B_d(x_0)$. As is customary, this can be determined separately for each particular global problem. Examples are given in the next section.

Our next result involves superconvergent approximations to first derivatives at symmetry points x_0. As noted in [12], u_h may not be differentiable at x_0, in which case our result will give many possible approximations to $\frac{\partial u}{\partial x_i}(x_0)$ given by $\frac{\partial \widehat{u}_h(x_0,\beta)}{\partial x_i}$, where β is any unit vector such that for $0 \le s \le s_0$ sufficiently small, $\frac{\partial u_h}{\partial x_i}$ has both left- and right-hand limits as $s \to 0$ along the line $x = x_0 + s\beta$. We then define

$$\frac{\partial \widehat{u}_h}{\partial x_i}(x_0,\beta) = \frac{1}{2}\lim_{s\to 0}\Big(\frac{\partial u_h(x_0+s\beta)}{\partial x_i} + \frac{\partial u_h(x-s\beta)}{\partial x_i}\Big).$$

Our result is then as follows.

THEOREM 4. *Suppose that the conditions of Corollary* 2 *hold and in addition* A.6 *is satisfied. Then if* $r \ge 2$ *is even and* $i = 1, \dots, N$,

$$\begin{aligned}&\Big|\frac{\partial u}{\partial x_i}(x_0) - \frac{\partial \widehat{u}(x_0,\beta)}{\partial x_i}\Big| \\ &\quad \le C\Big(h^r\Big(\ln\frac{d}{h}\Big)^2\|u\|_{W_\infty^{r+1}(B_{d(x_0)})} + \ln\Big(\frac{d}{h}\Big)d^{-1-t-N/p}\|e\|_{W_p^{-t}(B_{d(x_0)})}\Big).\end{aligned} \tag{1.23}$$

Remark. The first term on the right side of (1.23) is superconvergent of order $h^r(\ln\frac{d}{h})^2$, and we now give specific examples in which the second term on the right also exhibits superconvergent behavior.

1.5. Examples: Applications to specific boundary value problems. In this section we apply our results to the finite element method for approximating the solutions of various boundary value problems. For simplicity we restrict ourselves to the examples treated in section 3 of [12]. The superconvergence results given below are all significant improvements of those given there. Our boundary value problems will be of the form

$$Lu = -\sum_{i,j=1}^{N}\frac{\partial}{\partial x_j}\Big(a_{ij}(x)\frac{\partial u}{\partial x_i}\Big) + \sum_{i=1}^{N} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f \text{ in } \Omega \tag{1.24}$$

with some boundary condition on $\partial\Omega$.

It will be assumed that the coefficients of L are smooth and satisfy a uniform ellipticity conditon on Ω.

Example 1. Let the form (1.4) be defined on all of Ω and coercive on $W_2^1(\Omega)$ so that it corresponds to a problem with homogeneous conormal derivative boundary conditions. Suppose that $\partial\Omega$ is smooth and that the triangulations fit the boundary exactly. Furthermore, suppose that the subspaces have the property that for any $1 \le q \le \infty$,

$$\min_{\chi\in S_r^h(\Omega)}\|u-\chi\|_{W_q^1(\Omega)} \le Ch^{r-1}\|u\|_{W_q^r(\Omega)}. \tag{1.25}$$

We investigate the standard finite element method

$$A(u-u_h,\varphi) = 0 \text{ for all } \varphi \in S_r^h(\Omega). \tag{1.26}$$

It is well known (see, for example, [12, Example 3.1]) that

$$\|e\|_{W_\infty^{2-r}(\Omega)} \le Ch^{2r-2}\Big(\ln\frac{1}{h}\Big)\|u\|_{W_\infty^r(\Omega)}. \tag{1.27}$$

Applying (1.29) to Theorem 3, we obtain at symmetry points x_0 that if $r \ge 3$ is odd and u is sufficiently smooth, then

$$\begin{aligned} |e(x_0)| &\le C(u)\Big(\ln\frac{1}{h}\Big)^2(h^{r+1} + d^{2-r}h^{2r-2}) \\ &\le C(u)\Big(\ln\frac{1}{h}\Big)^2 h^{r+1} \text{ for } d \ge h^{1-\frac{1}{r-2}} \quad (r \ge 3 \text{ odd}). \end{aligned} \tag{1.28}$$

Thus if we disregard logarithmic factors we have superconvergence of one order higher than the global optimal rate, provided the subspace satisfies the symmetry condition in an $0\big(h^{1-\frac{1}{r-2}}\big)$ neighborhood of x_0.

The corresponding result for first derivatives is for $r \ge 2$ and even. In this case it follows from (1.27) and (1.23) that

$$\begin{aligned} &\Big|\frac{\partial u}{\partial x_i}(x_0) - \frac{\partial \widehat{u}_h(x_0,\beta)}{\partial x_i}\Big| \\ &\qquad \le C(u)\Big(\ln\frac{1}{h}\Big)^2(h^r + d^{1-r}h^{2r-2}) \\ &\qquad \le C(u)\Big(\ln\frac{1}{h}\Big)^2(h^r) \quad \text{for} \quad d \ge h^{1-\frac{1}{r-1}} \quad (r \ge 2, \text{ even}). \end{aligned} \tag{1.29}$$

We also note that a superconvergence rate of order $h^{r+\varepsilon}$ for $e(x_0)$ can be achieved with $d \ge h^{1-\frac{\varepsilon}{r-2}}$, and a superconvergence rate of order $h^{r-1+\varepsilon}$ for first derivatives can be achieved with $d \ge h^{1-\frac{\varepsilon}{r-1}}$. This indicates that we can obtain superconvergence fairly close to boundaries.

Example 2 (Dirichlet's problem on smooth domains in $\mathbb{R}^N$). Here we consider homogeneous Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. For our finite element spaces we take isoparametric elements defined on a mesh that approximates the boundary to order h^r and that vanish on the approximate boundary. For u sufficiently smooth it was proved in Schatz and Wahlbin [15] that

$$\|e\|_{L_\infty(\Omega)} \le Ch^r\Big(\ln\frac{1}{h}\Big)^{\overline{r}}. \tag{1.30}$$

Applying this to (1.23) we obtain for the derivatives

$$\begin{aligned} \Big|\frac{\partial u(x_0)}{\partial x_i} - \frac{\partial \widehat{u}_h(x_0,\beta)}{\partial x_i}\Big| &\le C\Big(\ln\frac{1}{h}\Big)^2\big(h^r\|u\|_{W_\infty^{r+1}(B_{d(x_0)})} + h^r d^{-1}\big) \\ &\le C\Big(\ln\frac{1}{h}\Big)^2 h^r \end{aligned} \tag{1.31}$$

for $r \ge 2$ even and $d = 0(1)$.

Results analogous to (1.28) and (1.29) would hold if (1.27) were satisfied. This can be accomplished, for example, by using superparametric elements that approximate the boundary to order h^{2r-2}.

Example 3 (Dirichlet problems on smooth plane domains). Scott [16], [17] treated Dirichlet problems on plane smooth domains in a special way so that

$$\|e\|_{W_2^{2-r}(\Omega)} \le Ch^{2r-2}. \tag{1.32}$$

In this case it is easily seen that the superconvergence estimate (1.28) holds this time with $d \ge Ch^{\frac{r-3}{r-1}}$. Furthermore, (1.29) holds with $d \ge Ch^{\frac{r-2}{r}}$.

Example 4 (Dirichlet problems on plane polygonal domains). Consider the problem

$$-\Delta u = f \text{ in } \Omega,\ u = 0 \text{ on } \partial\Omega,$$

where Ω is a plane polygonal domain. The behavior of solutions for f sufficiently smooth is well known (see, for example, Grisvard [3, Theorem 5.1.3.5]), and using suitable mesh refinements if necessary [1], we can assume that

$$\min_{\chi \in S_r^h(\Omega)} \|u - \chi\|_{W_2^1(\Omega)} \le Ch^{r-1}\|f\|_{W_2^{r-2}(\Omega)}.$$

A standard duality argument gives that (1.32) holds.

Then for $B_d(x) \subset\subset \Omega$ on which the meshes are not refined and for which Corollaries 1 and 2 hold, the same results as for Example 3 hold.

2. A proof of Theorem 1. The proof of Theorem 1 is quite lengthy. Parts of it have a great deal in common with the proof of Theorem 1.1 of [11]. We give an outline of the common parts and details on what is new. We shall simplify by considering only the case $d = 1$. The case $d < 1$ follows from this by a scaling argument. Without loss of generality we may assume that A is coercive on $W_2^1(B_1)$. The case of noncoercive A can be easily reduced to this case (see, e.g., [11]) Furthermore, the proof may be reduced in a now-standard way to estimates for an auxilliary Neumann problem with homogeneous conormal derivative boundary conditions as follows. Let $w(x) \in C_0^\infty(\mathbb{R}^N)$ be a cutoff function with $0 \le w(x) \le 1$, $w(x) \equiv 1$ for $|x - x_0| \le .25$, and $w(x) \equiv 0$ for $|x - x_0| \ge .5$. Set $\widehat{u} = wu$ and let $(\widehat{u})_h$ be the finite element approximation of $\widehat{u}$ defined by

$$A(\widehat{u} - (\widehat{u})_h, \varphi) = F(w\varphi) \quad \text{for all} \quad \varphi \in S_r^h(B_1). \tag{2.1}$$

Using the triangle inequality,

$$\|u - u_h\|_{L_\infty(B_h)} \le \|\widehat{u} - (\widehat{u})_h\|_{L_\infty(B_h)} + \|(\widehat{u})_h - u_h\|_{L_\infty(B_h)}. \tag{2.2}$$

The remainder of this section is devoted to proving the following lemma, from which (1.13) follows.

LEMMA 1. *With u, u_h, $\widehat{u}$, and $(\widehat{u})$ defined above,*

$$\begin{aligned}\|\widehat{u} - (\widehat{u})_h\|_{L_\infty(B_h)} \le C\Big(&h\Big(\ln\frac{1}{h}\Big)^{\overline{s}}\|\widehat{u}\|_{W_\infty^1(B_1),x_0,s} \\ &+ h\Big(\ln\frac{1}{h}\Big)^{\overline{s}}|||F|||_{W_\infty^{0,-1}(B_1),x_0,s} \\ &+ \Big(\ln\frac{1}{h}\Big)|||F|||_{W_\infty^{0,-r,2}(B_1),x_0}\Big)\end{aligned} \tag{2.3}$$

and

$$\|(\widehat{u})_h - u_h\|_{L_\infty(B_h)} \le C\Big(h\|\widehat{u}\|_{W^1_\infty(B_1),x_0,s} + \|u-u_h\|_{W_p^{-t}(B_1)} + h|||F|||_{W^{0,-1}_\infty(B_1),x_0,s} + |||F|||_{W^{0,-r,2}_\infty(B_1),x_0}\Big). \tag{2.4}$$

Before proceeding with a proof of this lemma, let us show how Theorem 1 follows. Since

$$\|\widehat{u}\|_{W^1_\infty(B_1),x_0,s} \le C\|u\|_{W^1_\infty(B_1)x_0,s},$$

from (2.3), (2.4), and (2.2), we have that

$$\|u-u_h\|_{L_\infty(B_h)} \le C\Big(h\Big(\ln\frac{1}{h}\Big)^{\overline{s}}\|u\|_{W^1_\infty(B_1),x_0,s} + \|u-u_h\|_{W_p^{-t}(B_1)} + h\Big(\ln\frac{1}{h}\Big)^{\overline{s}}|||F|||_{W^{0,-1}_\infty(B_1),x_0,s} + \Big(\ln\frac{1}{h}\Big)|||F|||_{W^{0,-r,2}_\infty(B_1),x_0}\Big). \tag{2.5}$$

The inequality (1.14) follows easily on applying (2.5) to $u-u_h \equiv u-\chi-(u_h-\chi)$ for any $\chi \in S^h_r(B_1)$.

We first prove the estimate (2.3). For this purpose let k be a fixed positive integer, and let $\psi \in \mathring{W}_2^{r-2}(B_{kh})$ with $\|\psi\|_{W_2^{r-2}(B_{kh})} = 1$. Define $g^{x_0}(y)$ as the unique solution of

$$A(v, g^{x_0}) = (v, h^{-N/2+2-r}\psi) \quad \text{for all} \quad v \in W^1_2(B_1(x_0)). \tag{2.6}$$

Note that $g^{x_0}(y) \in W^r_2(B_1)$ and $h^{-N/2+2-r}\|\psi\|_{L_1(B_{kh})} \le C$, so that g^{x_0} may be thought of as a smoothed and renormalized Green's function, in general smoother than that defined in [11]. Furthermore, let $g^{x_0}_h \in S^h_r(B_1)$ be its finite element approximation defined by

$$A(\varphi, g^{x_0} - g^{x_0}_h) = 0 \quad \text{for all} \quad \varphi \in S^h_r(B_1). \tag{2.7}$$

The following estimates play a critical role in our new results.

LEMMA 2. *For g^{x_0} and $g^{x_0}_h$ as defined above,*

$$\|g^{x_0}\|_{W^j_2(B_1)} \le Ch^{-N/2-j+2} \quad \textit{for} \quad j = 2, \dots, r \tag{2.8}$$

and

$$\|g^{x_0}\|_{W^{0,r,2}_1(B_1),x_0} \le C\ln\frac{1}{h}. \tag{2.9}$$

Furthermore,

$$\|g^{x_0} - g^{x_0}_h\|_{W^1_1(B_1),x_0,-s} \le Ch\Big(\ln\frac{1}{h}\Big)^{\overline{s}} \tag{2.10}$$

and

$$\|g^{x_0} - g^{x_0}_h\|_{W^{0,1}_1(B_{.5}),x_0,-s} \le Ch\Big(\ln\frac{1}{h}\Big)^{\overline{s}}, \tag{2.11}$$

where the constants C are independent of ψ.

Let us first show how, using this, one may estimate $\|\widehat{u} - (\widehat{u})_h\|_{L_\infty(B_h)}$ in (2.2). Letting $\widehat{e} = \widehat{u} - (\widehat{u})_h$ and using standard inverse properties of $S_r^h(B_1)$, we have for any $\chi \in S_r^h(B_1)$ that there is a $k > 0$, sufficiently large but fixed, such that

$$\|\widehat{e}\|_{L_\infty(B_h)} \le C\|\widehat{u} - \chi\|_{L_\infty(B_{kh})} + Ch^{-N/2+2-r}\|\widehat{e}\|_{W_2^{2-r}(B_{kh})}. \tag{2.12}$$

Then by duality,

$$h^{-N/2+2-r}\|\widehat{e}\|_{W_2^{2-r}(B_{kh})} = \sup_{\substack{\psi \in \mathring{W}_2^{r-2}(B_{kh}) \\ \|\psi\|_{W_2^{r-2}(B_{kh})}=1}} (\widehat{e}, h^{-N/2+2-r}\psi). \tag{2.13}$$

Now for each such ψ, we have in view of (2.1) and (2.7) that

$$\begin{aligned}(\widehat{e}, h^{-N/2+2-r}\psi) &= A(\widehat{e}, g^{x_0}) = A(\widehat{e}, g^{x_0} - g_h^{x_0}) + F(wg_h^{x_0}) \\ &= A(\widehat{u} - \chi, g^{x_0} - g_h^{x_0}) + F(w(g_h^{x_0} - g^{x_0})) + F(wg^{x_0}).\end{aligned} \tag{2.14}$$

Combining (2.12), (2.13), and (2.14) with Lemma 2 we arrive at

$$\begin{aligned}\|\widehat{e}\|_{L_\infty(B_h)} &\le C\Big(\|\widehat{u} - \chi\|_{L_\infty(B_{kh})} + \|\widehat{u} - \chi\|_{W_\infty^1(B_1),x_0,s}\|g^{x_0} - g_h^{x_0}\|_{W_1^1(B_1),x_0,-s} \\ &\quad + |||F|||_{W_\infty^{0,-1}(B_{.5}),x_0,s}\|g^{x_0} - g_h^{x_0}\|_{W_1^{0,1}(B_{.5}),x_0,-s} \\ &\quad + |||F|||_{W_\infty^{0,-r,2}(B_1),x_0}\|g^{x_0}\|_{W_1^{0,r,2}(B_1),x_0}\Big) \\ &\le C\Big(h\Big(\ln\frac{1}{h}\Big)^{\overline{s}}\|\widehat{u}\|_{W_\infty^1(B_1),x_0,s} + h\Big(\ln\frac{1}{h}\Big)^{\overline{s}}|||F|||_{W_\infty^{0,-1}(B_1),x_0,s} \\ &\quad + \Big(\ln\frac{1}{h}\Big)|||F|||_{W_\infty^{0,-r,2}(B_1),x_0}\Big),\end{aligned}$$

which completes the proof of (2.3), once we have proved Lemma 2.

Proof of Lemma 2. The inequality (2.8) follows from a standard a priori estimate and Poincaré's inequality. In fact, for $j = 2, \dots, r$,

$$\|g^{x_0}\|_{W_2^j(B_1)} \le Ch^{-N/2+2-r}\|\psi\|_{W_2^{j-2}(B_{kh})} \le Ch^{-N/2-j}\|\psi\|_{W_2^{r-2}(B_{kh})}.$$

We prove (2.9) in three steps. Set

$$M_\alpha = \int_{B_1} (|x - x_0| + h)^{j-2}|D^\alpha g^{x_0}|dx, \quad |\alpha| = j.$$

Step 1. We begin by estimating M_α for $j = 2, \dots, r$:

$$\begin{aligned}M_\alpha &\le \int_{B_{2kh}} (|x - x_0| + h)^{j-2}|D^\alpha g^{x_0}|dx + \sum_{j=0}^{J}\int_{\Omega_\ell} (|x - x_0| + h)^{j-2}|D^\alpha g^{x_0}|dx \\ &= I_{in} + \sum_{\ell=0}^{J} I_\ell.\end{aligned} \tag{2.15}$$

Here for $\ell = 0, \dots, J = [\ln_2 \frac{1}{kh}] + 1$, Ω_ℓ is the annulus

$$\Omega_\ell = \{x : 2^{\ell-1} \le |x - x_0| \le 2^{-\ell}\}.$$

Now using (2.8),

$$|I_{in}| \le Ch^{j-2+N/2} \|g^{x_0}\|_{W_2^j(B_1)} \le C. \tag{2.16}$$

Furthermore, for any $\ell = 0, \dots, J$ and any $j = 2, \dots, r$, and $d_\ell = 2^{-\ell}$,

$$|I_\ell| \le Cd_\ell^{j+N} \|D_x^\alpha g^{x_0}\|_{L_\infty(\Omega_\ell)}. \tag{2.17}$$

Now let $G^x(y)$ be the Green's function for the auxilliary Neumann problem. It follows that since $x \in \Omega_\ell$ and $\Omega_\ell \cap B_{kh} = \phi$, then

$$D_x^\alpha g^{x_0}(x) = h^{-N/2+2-r} \int_{B_{kh}} \psi(y) D_x^\alpha G^x(y) dy.$$

By an estimate of Krasovskii [4] we have for $N \ge 2$,

$$|D_x^\alpha D_y^\beta G^x(y)| \le C \begin{cases} \left(\ln \dfrac{1}{|x-y|}\right) + 1 & \text{if } N = 2 \text{ and } |\alpha| + |\beta| = 0, \\ \dfrac{1}{|x-y|^{N-2+|\alpha|+|\beta|}} & \text{otherwise.} \end{cases} \tag{2.18}$$

Thus for $|\alpha| = j \ge 2$

$$\|D_x^\alpha G^x(y)\|_{L_\infty(\Omega_\ell)} \le \frac{C}{d_\ell^{N-2+j}}, \ x \in \Omega_\ell, \quad y \in B_{kh}.$$

Hence it follows from this that

$$\|D_x^\alpha g^{x_0}\|_{L_\infty(\Omega_\ell)} \le h^{-N/2+2-r} \|\psi\|_{L_1(B_{kh})} d_\ell^{-N-j+2} \le Cd_\ell^{-N-j+2},$$

where we used the Cauchy–Schwarz and Poincaré inequalities. Combining this last inequality with (2.17) we arrive at

$$|I_\ell| \le C. \tag{2.19}$$

Summing (2.19) over ℓ and adding the result to (2.16) we finally obtain from (2.15) that for any $|\alpha| = j$ and $j = 2, \dots, r$,

$$M_\alpha \le C \ln \frac{1}{h}, \tag{2.20}$$

which completes Step 1.

Step 2. We estimate M_α for $j = 1$ and $N \ge 2$, or $j = 0$ and $N \ge 3$. In this case

$$\begin{aligned} M_\alpha &\le h^{-N/2+2-r} \int_{B_1} (|x - x_0| + h)^{j-2} \left(\int_{B_{kh}} |\psi(y)| \ |D^\alpha G^x(y)| dy \right) dx \\ &\le h^{-N/2+2-r} \int_{B_{kh}} |\psi(y)| \left(\int_{B_1} (|x - x_0| + h)^{j-2} |D^\alpha G^x(y)| dx \right) dy, \end{aligned}$$

where we interchanged orders of integration. Setting $\rho = |x-y|$ and $\sigma = |x-x_0|$ it follows that

$$M_\alpha \le Ch^{-N/2+2-r}\int_{B_{kh}} |\psi(y)|\Big(\int_0^{3/2}(\sigma+h)^{j-2}\rho^{1-j}d\rho\Big)dy,$$

where we used (2.18). Now let $j=1$; then

$$\begin{aligned} M_\alpha &\le Ch^{-N/2+2-r}\int_{B_{kh}} |\psi(y)|\Big(\int_0^{2}(\sigma+h)^{-1}d\sigma\Big) \\ &\le Ch^{-N/2+2-r}\ln\frac{1}{h}\|\psi\|_{L_1(B_{kh})} \le C\ln\frac{1}{h}. \end{aligned} \tag{2.21}$$

In the case that $j=0$ and $N\ge 3$, it follows from (2.18) that

$$\begin{aligned} M_\alpha &\le Ch^{-N/2+2-r}\int_{B_{kh}} |\psi(y)|\Big(\int_0^{3/2}\frac{\rho}{(\sigma+h)^2}d\rho\Big)dy \\ &\le Ch^{-N/2+2-r}\int_{B_{kh}} |\psi(y)|\Big(\int_0^{2}(\sigma+h)^{-1}d\sigma\Big)dy \\ &\le C\ln\frac{1}{h}, \end{aligned} \tag{2.22}$$

where we used $\rho \le \sigma + 2kh \le 2k(\sigma+h)$. This completes Step 2.

Step 3. $j=0$ and $N=2$. Again using (2.18) we have

(2.23)

$$\begin{aligned} M_\alpha &\le Ch^{-N/2+2-r}\int_{B_{kh}} |\psi(y)|\Big(\int_{B_1}\frac{\big(\ln\frac{1}{|x-y|}+1\big)}{(|x-x_0|+h)^2}\,dx\Big)dy \\ &\le Ch^{-N/2+2-r}\int_{B_{kh}} |\psi(y)|\Big(\int_0^{h}\frac{\rho\big(\ln\frac{1}{\rho}+1\big)}{(\sigma+h)^2}d\rho + \int_h^{1.5}\frac{\rho\big(\ln\frac{1}{\rho}+1\big)}{(\sigma+h)^2}d\rho\Big)dy. \end{aligned}$$

Furthermore,

$$\int_0^h \frac{\rho\big(\ln\frac{1}{\rho}+1\big)}{(\sigma+h)^2}d\rho \le \frac{1}{h^2}\int_0^h \rho\Big(\ln\frac{1}{\rho}+1\Big)d\rho \le C\ln\frac{1}{h}, \tag{2.24}$$

which is obtained by integrating the last integral by parts. Again, since $\rho \le 2k(\sigma+h)$,

$$\int_h^{1.5}\frac{\rho\big(\ln\frac{1}{\rho}+1\big)}{(\sigma+h)^2}d\rho \le C\Big(\ln\frac{1}{h}\Big)\int_0^2(\sigma+h)^{-1}d\sigma \le C\Big(\ln\frac{1}{h}\Big)^2. \tag{2.25}$$

Substituting (2.24) and (2.25) into (2.23) we obtain for $j=0$ and $N=2$,

$$M_\alpha \le C\Big(\ln\frac{1}{h}\Big)^2 \text{ for } |\alpha| = j = 0 \text{ and } N=2, \tag{2.26}$$

which completes Step 3. The inequality (2.9) now easily follows from (2.20), (2.21), (2.22), and (2.26).

We next note that the inequality (2.10) was proved in Lemma 2.7 of [11] for a less smoothed Green's function than that treated here. The proof given there yields the result (2.10). This completes the proof of Lemma 2.

Remark. The extra smoothness of the smoothed Green's function constructed here was used in the inequalities (2.8) and (2.9) and could be used to estimate higher derivatives of the error in (2.11).

We now turn to the proof of (2.4). To begin with, notice that $(\widehat{u})_h - u_h$, where $\widehat{u} = wu$, is discrete "A harmonic" on $B_{.25}$; i.e., it satisfies

$$A((\widehat{u})_h - u_h, \varphi) = 0 \quad \text{for all} \quad \varphi \in \mathring{S}_r^h(B_{.25}). \tag{2.27}$$

It was proved in [13] (see also [11]) that

$$\begin{aligned} \|(\widehat{u})_h - u_h\|_{L_\infty(B_h)} &\le C\|(\widehat{u})_h - u_h\|_{W_{p'}^{-t'}(B_{.25})} \\ &\le C\Big(\|u - u_h\|_{W_{p'}^{-t'}(B_1)} + \|\widehat{u} - (\widehat{u})_h\|_{W_{p'}^{-t'}(B_1)}\Big), \end{aligned} \tag{2.28}$$

where $t' \ge 0$ and $1 \le p' \le \infty$ are arbitrary. Now t' may be chosen so large and $p' = 1$, so that for arbitrary $t \ge 0$ and $1 \le p \le \infty$ (2.28) becomes

$$\|(\widehat{u})_h - u_h\|_{L_\infty(B_h)} \le C\Big(\|u - u_h\|_{W_p^{-t}(B_1)} + \|\widehat{u} - (\widehat{u})_h\|_{W_2^{-\lambda}(B_1)}\Big), \tag{2.29}$$

where $\lambda = \max(r-2, [N/2])$ and $[N/2]$ denotes the integer part of $N/2$. The proof of (2.4), and hence of Theorem 1, is complete once we have shown that

$$\begin{aligned} &\|\,\widehat{u} - (\widehat{u})_h\|_{W_2^{-\lambda}(B_1)} \\ &\quad \le C\Big(h\|\widehat{u}\|_{W_\infty^1(B_1),x_0,s} + h|||F|||_{W_\infty^{0,-1}(B_1),x_0,s} + \ln\frac{1}{h}|||F|||_{W_\infty^{0,-r,2}(B_1),x_0}\Big). \end{aligned} \tag{2.30}$$

The proof of this follows by a standard duality argument that we shall just outline.

$$\|\widehat{e}\|_{W_2^{-\lambda}(B_1)} = \sup_{\substack{\eta \in \mathring{W}_2^{\lambda}(B_1) \\ \|\eta\|_{W_2^{\lambda}(B_1)} = 1}} (\widehat{e}, \eta). \tag{2.31}$$

For each such η let $z \in W_2^r(B_1)$ satisfy

$$A(v, z) = (v, \eta) \quad \text{for all} \quad v \in W_2^1(B_1) \tag{2.32}$$

and let $z_h \in S_r^h(B_1)$ satisfy

$$A(\varphi, z - z_h) = 0 \quad \text{for all} \quad \varphi \in S_r^h(B_1). \tag{2.33}$$

Then

$$\begin{aligned} (\widehat{e}, \eta) = A(\widehat{e}, z) &= A(\widehat{e}, z - z_h) + F(wz_h) \\ &= A(\widehat{u}, z - z_h) + F(w(z_h - z)) + F(wz). \end{aligned}$$

This together with (2.31) yields

$$\begin{aligned} \|\widehat{e}\|_{W_2^{-\lambda}(B_1)} \le C\ \Big(&\|\widehat{u}\|_{W_\infty^1(B_1),x_0,s}\|z - z_h\|_{W_1^1(B_1),x_0,-s} \\ &+ |||F|||_{W_\infty^{0,-1}(B_{.5}),x_0,s}\|z - z_h\|_{W_1^{0,1}(B_{.5}),x_0,s} \\ &+ |||F|||_{W_\infty^{0,-r,2}(B_{.5}),x_0}\|z\|_{W_1^{0,r,2}(B_{.5}),x_0}\Big). \end{aligned} \tag{2.34}$$

The inequality (2.30) now easily follows from (2.34) and the following lemma.

LEMMA 3. *With z and z_h defined as above,*

$$\|z - z_h\|_{W_1^1(B_1),x_0,-s} \le Ch, \tag{2.35}$$

$$\|z - z_h\|_{W_1^{0,1}(B_{.5}),x_0,-s} \le Ch, \tag{2.36}$$

$$\|z\|_{W_1^{0,r,2}(B_1),x_0} \le C\Big(\ln\frac{1}{h}\Big). \tag{2.37}$$

Proof. We begin by noting that

$$\|z\|_{W_2^{\lambda+2}(B_1)} \le C\|\eta\|_{W_2^{\lambda}(B_1)} \le C. \tag{2.38}$$

Then a crude estimate yields

$$\|z - z_h\|_{W_1^1(B_1),x_0,-s} \le Ch^{-s}\|z - z_h\|_{W_2^1(B_1)} \le Ch^{r-1-s}\|z\|_{W_2^r(B_1)} \le Ch, \tag{2.39}$$

where we used (2.38) and the fact that $0 \le s \le r-2$. This proves (2.35). We leave the proof of (2.36) to the reader.

To prove (2.37) we let

$$P_\alpha = \int_{B_1} (|x - x_0| + h)^{j-2}|D^\alpha(\omega z)|dx,\ |\alpha| = j,\ j = 0, \dots, r.$$

When $j = 2, \dots, r$,

$$P_\alpha \le C\|D^\alpha(\omega z)\|_{L_2(B_1)} \le C\|\eta\|_{W_2^{r-2}(B_1)} \le C. \tag{2.40}$$

In the case that $j = 0, 1$ we have

$$P_\alpha = \|(|x - x_0| + h)^{j-2}\|_{L_1(B_1)}\|D^\alpha(\omega z)\|_{L_\infty(B_1)}.$$

Now

$$\|(|x - x_0| + h)^{j-2}\|_{L_1(B_1)} \le C\begin{cases}\ln\dfrac{1}{h} & \text{when } j = 0 \text{ and } N = 2,\\ 1 & \text{when } j = 0 \text{ and } N \ge 3 \text{ or } j = 1 \text{ and } N \ge 2.\end{cases}$$

Thus in this case we certainly have

$$P_\alpha \le C\ln\frac{1}{h},\ j = 0, 1,\ |\alpha| = j. \tag{2.41}$$

The inequality (2.37) now easily follows from (2.40) and (2.41).

This completes the proof of Lemma 3 and Theorem 1. ∎

2.1. A sketch of the proof of Theorem 2. In general outline, the proof of Theorem 2 follows that of Theorem 1. We give only a sketch of the proof and indicate the differences. We again restrict ourselves to the case $d = 1$ and this time, using $\widehat{u} - (\widehat{u})_h$, the solution of (2.1), we have instead of (2.2), that for any $i = 1, \dots, N$,

$$\Big\| \frac{\partial}{\partial x_i}(u - u_h) \Big\|_{L_\infty(B_h)} \le \Big\| \frac{\partial}{\partial x_i}(\widehat{u} - (\widehat{u})_h) \Big\|_{L_\infty(B_h)} + \Big\| \frac{\partial}{\partial x_i}((\widehat{u})_h - u_h) \Big\|_{L_\infty(B_h)}. \tag{2.42}$$

Then instead of Lemma 1 we have the following lemma.

LEMMA 4. *With u, u_h, $\widehat{u}$, and $\widehat{u}_h$ as in Lemma 1,*

$$\begin{aligned} &\Big\| \frac{\partial}{\partial x_i}(\widehat{u} - (\widehat{u})_h) \Big\|_{L_\infty(B_h)} \\ &\quad \le C\Big(\Big(\ln \frac{1}{h} \Big)^{\overline{\overline{s}}} \|\widehat{u}\|_{W^1_\infty(B_1),x_0,s} + \Big(\ln \frac{1}{h} \Big)^{\overline{\overline{s}}} |||F|||_{W^{0,-1}_\infty(B_1),x_0,s} \\ &\qquad + \Big(\ln \frac{1}{h} \Big) |||F|||_{W^{0,-r,1}_\infty(B_1),x_0} \Big) \end{aligned} \tag{2.43}$$

and

$$\begin{aligned} &\Big\| \frac{\partial}{\partial x_i}((\widehat{u})_h - u_h) \Big\|_{L_\infty(B_n)} \\ &\quad \le C\Big(\|u\|_{W^1_\infty(B_1),x_0,s} + \|u - u_h\|_{W_p^{-t}(B_1)} \\ &\qquad + |||F|||_{W^{0,-1}_\infty(B_1),x_0,s} + |||F|||_{W^{0,-r,1}_\infty(B_1),x_0} \Big). \end{aligned} \tag{2.44}$$

The proof of Theorem 2 follows easily from (2.40), (2.43), and (2.44). Therefore we need only prove (2.43) and (2.44).

To prove (2.43), let $\psi \in \mathring{W}_2^{r-1}(B_{kh})$ with $\|\psi\|_{W_2^{r-1}(B_{kh})} = 1$, and define $\widetilde{g}^{x_0}$ as the solution of

$$A(v, \widetilde{g}^{x_0}) = \Big(v, -h^{-N/2-r+1} \frac{\partial \psi}{\partial x_i} \Big). \tag{2.45}$$

Furthermore, let $\widetilde{g}_h^{x_0}$ be the finite element approximation defined by

$$A(v, \widetilde{g}^{x_0} - \widetilde{g}_h^{x_0}) = 0 \quad \text{for all} \quad v \in W_2^1(B_1). \tag{2.46}$$

The analogue of Lemma 2 is as follows.

LEMMA 5. *Let $\widetilde{g}^{x_0}$ and $\widetilde{g}_h^{x_0}$ as defined above. Then*

$$\|\widetilde{g}^{x_0}\|_{W_2^j(B_1)} \le C h^{-N/2-j+1} \quad \textit{for} \quad j = 1, \dots, r, \tag{2.47}$$

$$\|\widetilde{g}^{x_0}\|_{W_1^{0,r,1}(B_1),x_0} \le C \ln \frac{1}{h}, \tag{2.48}$$

$$\|\widetilde{g}^{x_0} - \widetilde{g}_h^{x_0}\|_{W_1^1(B_1),x_0,-s} \le C\Big(\ln\frac{1}{h}\Big)^{\overline{\overline{s}}}, \tag{2.49}$$

$$\|\widetilde{g}^{x_0} - \widetilde{g}_h^{x_0}\|_{W_1^{0,1}(B_{1/2}),x_0,-s} \le C\Big(\ln\frac{1}{h}\Big). \tag{2.50}$$

The proof of Lemma 5 follows along the lines of Lemma 2 and is left for the reader.

To prove (2.43) we first find, with $\widehat{e} = \widehat{u} - (\widehat{u})_h$, that for any $\chi \in S_r^h(B_1)$,

$$\Big\|\frac{\partial\widehat{e}}{\partial x_i}\Big\|_{L_\infty(B_h)} \le \Big\|\frac{\partial(\widehat{u}-\chi)}{\partial x_i}\Big\|_{L_\infty(B_{kh})} + h^{-N/2+1-r}\Big\|\frac{\partial\widehat{e}}{\partial x_i}\Big\|_{W_2^{1-r}(B_{kh})}. \tag{2.51}$$

By duality and integration by parts

$$h^{-N/2+1-r}\Big\|\frac{\partial\widehat{e}}{\partial x_i}\Big\|_{W_2^{1-r}(B_{kh})} = \sup_{\substack{\psi\in\mathring{W}_2^{r-1}(B_{kh})\\ \|\psi\|_{W_2^{r-1}(B_{kh})}=1}} \Big(\widehat{e}, -h^{-N/2+1-r}\frac{\partial\psi}{\partial x_i}\Big). \tag{2.52}$$

For each such ψ

$$\Big(\widehat{e}, -h^{N/2+1-r}\frac{\partial\psi}{\partial x_i}\Big) = A(\widehat{e}, \widetilde{g}^{x_0} - \widetilde{g}_h^{x_0}) + F(\widetilde{w(g_h^{x_0} - \widetilde{g}^{x_0})}) + F(w\widetilde{g}^{x_0}).$$

It follows that

$$\begin{aligned}
&\Big|\Big(\widehat{e}, -h^{N/2+1-r}\frac{\partial\psi}{\partial x_i}\Big)\Big| \\
&\quad\le C\Big(\|\widehat{u}-\chi\|_{W_\infty^1(B_1),x_0,s}\|\widetilde{g}^{x_0} - \widetilde{g}_h^{x_0}\|_{W_1^1(B_1),x_0,-s} \\
&\qquad + |||F|||_{W_\infty^{0,-1}(B_1),x_0,s}\|\widetilde{g}^{x_0} - \widetilde{g}_h^{x_0}\|_{W_1^{0,1}(B_{1/2}),x_0,-s} \\
&\qquad + |||F|||_{W_\infty^{0,-r,1}(B_1),x_0}\|\widetilde{g}^{x_0}\|_{W_1^{0,r,1}(B_1),x_0}\Big).
\end{aligned} \tag{2.53}$$

The inequality (2.43) now follows by combining (2.48), (2.52), and (2.53) with Lemma 5.

The proof of (2.44) follows the proof of (2.4) very closely except that instead of (2.21), we now use

$$\Big\|\frac{\partial}{\partial x_i}((\widehat{u})_h - u_h)\Big\|_{L_\infty(B_h)} \le C\|(\widehat{u})_h - u_h\|_{W_p^{-t}(B_{.25})},$$

which was proved in [14] (see also [11]). We leave the details to the reader.

3. A proof of Theorem 3. Consider $\overline{e}(x) = e(x_0 - (x - x_0))$, where $e(x) = u(x) - u_h(x)$. For any $\varphi \in \mathring{S}_r^h(B_d(x_0))$, $\overline{e}(x)$ satisfies (after a change of variables

$y = x_0 - (x - x_0))$

$$
\begin{aligned}
A(\overline{e},\varphi) &= \int_{B_d(x_0)} \left(\sum_{i,j=1}^{N} a_{ij} \frac{\partial \overline{e}}{\partial x_i} \frac{\partial \varphi}{\partial x_j} + \sum_{i=1}^{N} b_i \frac{\partial \overline{e}}{\partial x_i} \varphi + c\overline{e}\varphi \right) dx \\
&= (-1)^N \int_{B_d(x_0)} \left(\sum_{i,j=1}^{N} \overline{a}_{ij} \frac{\partial e}{\partial y_i} \frac{\partial \overline{\varphi}}{\partial y_j} - \sum_{i=1}^{N} \overline{b}_i \frac{\partial e}{\partial y} \overline{\varphi} + \overline{C} e \overline{\varphi} \right) dy \\
&= (-1)^N A(e,\overline{\varphi}) \\
&\quad + (-1)^N \int_{B_d(x_0)} \left(\sum_{i,j=1}^{N} (\overline{a}_{ij} - a_{ij}) \frac{\partial e}{\partial y_i} \frac{\partial \overline{\varphi}}{\partial y_j} - \sum_{i=1}^{N} (\overline{b}_i + b_i) \frac{\partial e}{\partial y_i} \overline{\varphi} \right) dy \\
&\quad + (-1)^N \textstyle\int_{B_d(x_0)} (\overline{C} - C) e \overline{\varphi} dy.
\end{aligned} \tag{3.1}
$$

Since $\overline{\varphi} \in \mathring{S}_r^h(B_d(x_0)), A(e,\overline{\varphi}) = 0$, and it follows that (3.1) may be rewritten as

$$
\begin{aligned}
A(\overline{e},\varphi) &= 2(-1)^N \int_{B_d(x_0)} \left(\sum_{i,j=1}^{N} a_{ij,\mathrm{odd}} \frac{\partial e}{\partial y_i} \frac{\partial \overline{\varphi}}{\partial y_j} - \sum_{i=1}^{N} b_{i,\mathrm{even}} \frac{\partial e}{\partial y_i} \overline{\varphi} + C_{\mathrm{odd}} e \overline{\varphi} \right) dy \\
&= 2(-1)^n [F_1(\varphi) + F_2(\varphi) + F_3(\varphi)] \equiv F(\varphi).
\end{aligned} \tag{3.2}
$$

Here, for any function $w(x)$

$$
w_{\mathrm{even}} = \frac{w(x) + \overline{w}(x)}{2}, \qquad w_{\mathrm{odd}} = \frac{w(x) - \overline{w}(x)}{2}.
$$

Note that since $A(e,\varphi) = 0$ for all $\varphi \in \mathring{S}_r^h(B_d(x_0))$, then

$$
A(e_{\mathrm{even}},\varphi) = F(\varphi) \quad \text{for all} \quad \varphi \in \mathring{S}_r^h(B_d(x_0)). \tag{3.3}
$$

Applying Corollary 1 to $e_{\mathrm{even}} = u_{\mathrm{even}} - u_{h,\mathrm{even}}$ with the choice $\gamma = r+1$, we obtain that at the symmetry point x_0 where $e_{\mathrm{even}}(x_0) = e(x_0)$,

$$
\begin{aligned}
|e(x_0)| \le C &\left[\left(\ln \frac{d}{h} \right)^{\overline{r}} \left(h^r \sum_{|\alpha|=r} |D^\alpha u_{\mathrm{even}}(x_0)| + h^{r+1} \|u_{\mathrm{even}}\|_{W_\infty^{r+1}(B_d(x_0))} \right) \right. \\
&\left. + d^{-t-N/p} \|e_{\mathrm{even}}\|_{W_p^{-t}(B_d(x_0))} \right] \\
&+ C \ln \left(\frac{d}{h} \right) \left[h |||F|||_{W_\infty^{0,-1}(B_d(x_0)),x_0,r-2} + |||F|||_{W_\infty^{0,-r,1}(B_d(x_0)),x_0} \right].
\end{aligned}
$$

Here $\overline{r} = 1$ if $r = 3$, and $\overline{r} = 0$ otherwise. Since r is odd, $D^\alpha u_{\mathrm{even}}(x_0) = 0$ for all $|\alpha| = r$ and it easily follows that

$$
\begin{aligned}
|e(x_0)| \le C &\left[\left(\ln \frac{d}{h} \right)^{\overline{r}} h^{r+1} \|u\|_{W_\infty^{r+1}(B_d(x_0))} + d^{-t-N/p} \|e\|_{W_p^{-t}(B_d(x_0))} \right] \\
&+ C \left(\ln \frac{d}{h} \right) \left[h |||F|||_{W_\infty^{0,-1}(B_d(x_0),x_0,r-2} + |||F|||_{W_\infty^{0,-r,1}(B_d(x_0)),x_0} \right].
\end{aligned} \tag{3.4}
$$

We now estimate the last two terms on the right side of (3.4).

We begin by estimating $F_3(\varphi)$ defined in (3.2). Setting $B_d \equiv B_d(x_0)$,

$$
\begin{aligned}
|F_3(\varphi)| &= \Big| \int_{B_d} C_{\text{odd}} e\varphi dy \Big| \le C\|e\|_{W_\infty^{-1}(B_d)} \|\varphi\|_{W_1^1(B_d)} \\
&\le C\|e\|_{W_\infty^{-1}(B_d)} \|\varphi\|_{W_1^{0,-1}(B_d),x_0,r-2} \quad \text{for all} \quad \varphi \in \mathring{W}_1^r(B_d).
\end{aligned} \tag{3.5}
$$

On the other hand,

$$
\begin{aligned}
|F_3(\varphi)| &\le C\|e\|_{W_\infty^{-2}(B_d)} \|\varphi\|_{W_1^2(B_d)} \\
&\le C\|e\|_{W_\infty^{-2}(B_d)} \|\varphi\|_{W_1^{0,r,2}(B_d),x_0} \quad \text{for all} \quad \varphi \in \mathring{W}_1^r(B_d).
\end{aligned} \tag{3.6}
$$

Hence taken together, (3.5) and (3.6) imply

$$
\begin{aligned}
&\Big(\ln \frac{d}{h} \Big) \Big[h|||F_3|||_{W_\infty^{0,-1}(B_d),x_0,r-2} + |||F_3|||_{W_\infty^{0,-r}(B_d),x_0} \Big] \\
&\qquad \le C\Big(\ln \frac{d}{h} \Big) \big[h\|e\|_{W_\infty^{-1}(B_d)} + \|e\|_{W_\infty^{-2}(B_d)} \big].
\end{aligned} \tag{3.7}
$$

To estimate $F_2(\varphi)$ we use the triangle inequality and integration by parts to obtain

$$
\begin{aligned}
|F_2(\varphi)| &\le \sum_{i=1}^N \Big| \int_{B_d} b_{i,\text{even}} \frac{\partial e}{\partial y_i} \varphi dy \Big| \\
&\le \sum_{i=1}^N \Big(\Big| \int_{B_d} e\Big(b_{i,\text{even}} \frac{\partial \varphi}{\partial y_i} \Big) dy \Big| + \Big| \int_{B_d} e \frac{\partial b_{i,\text{even}}}{\partial x_i} \varphi dy \Big| \Big) \\
&\le C\|e\|_{L_\infty(B_d)} \|\varphi\|_{W_1^1(B_d)} \le C\|e\|_{L_\infty(B_d)} \|\varphi\|_{W_1^{0,1}(B_d),x_0,r-2}.
\end{aligned} \tag{3.8}
$$

Furthermore,

$$
|F_2(\varphi)| \le C\|e\|_{W_\infty^{-1}(B_d)} \|\varphi\|_{W_1^2(B_d)} \le C\|e\|_{W_\infty^{-1}(B_d)} \|\varphi\|_{W_1^{0,r,2}(B_d),x_0}. \tag{3.9}
$$

Taken together, (3.8) and (3.9) imply

$$
\begin{aligned}
&\Big(\ln \frac{d}{h} \Big) \Big[h|||F_2|||_{W_\infty^{0,-1}(B_d)x_0,r-2} + |||F_2|||_{W_\infty^{0,-r,2}(B_d),x_0} \Big] \\
&\qquad \le C\Big(\ln \frac{d}{h} \Big) \Big[h\|e\|_{L_\infty(B_d)} + \|e\|_{W_\infty^{-1}(B_d)} \Big].
\end{aligned} \tag{3.10}
$$

Last, we estimate $F_1(\varphi)$. On the one hand we have

$$
\begin{aligned}
|F_1(\varphi)| &\le \sum_{i,j=1}^N \Big| \int_{B_d} a_{ij,\text{odd}} \frac{\partial e}{\partial y_i} \frac{\partial \varphi}{\partial y_j} dy \Big| \\
&\le \sum_{i,j=1}^N \Big\| \frac{\partial e}{\partial y_i} \Big\|_{L_\infty(B_d)} \Big\| a_{ij,\text{odd}} \frac{\partial \varphi}{\partial y_j} \Big\|_{L_1(B_d)} \\
&\le Ch\|e\|_{W_\infty^1(B_d)} \|\varphi\|_{W_1^1(B_d),x_0,-1},
\end{aligned} \tag{3.11}
$$

where in the last step we used

$$
|a_{ij,\text{odd}}| \le C|x - x_0|. \tag{3.12}
$$

Furthermore, using integration by parts and the triangle inequality,

$$|F_1(\varphi)| \le \sum_{i,j=1}^{N} \left(\left| \int_{B_d} e \frac{\partial a_{ij,\text{odd}}}{\partial y_i} \frac{\partial \varphi}{\partial y_j} dy \right| + \left| \int_{B_d} e a_{ij,\text{odd}} \frac{\partial^2 \varphi}{\partial y_i \partial y_j} dy \right| \right)$$
$$\le C \|e\|_{W_\infty^{-1}(B_d)} \left(\|\varphi\|_{W_1^2(B_d)} + \sum_{i,j=1}^{N} \left\| a_{ij,\text{odd}} \frac{\partial^2 \varphi}{\partial y_i \partial y_j} \right\|_{W_1^1(B_d)} \right).$$

A simple calculation shows that

$$\sum_{i,j=1}^{N} \left\| a_{ij,\text{odd}} \frac{\partial^2 \varphi}{\partial y_i \partial y_j} \right\|_{W_1^1(B_d)} \le C \|\varphi\|_{W_1^{0,r,2}(B_d),x_0},$$

where we again used (3.12). This leads to

$$|F_1(\varphi)| \le C \|e\|_{W_\infty^{-1}(B_d)} \|\varphi\|_{W_1^{0,r,2}(B_d),x_0}. \tag{3.13}$$

Taken together, (3.11) and (3.13) imply that

$$\left(\ln \frac{d}{h} \right) \left[h |||F_1|||_{W_\infty^{0,-1}(B_d),x_0,r-2} + |||F_1|||_{W_\infty^{0,-r,2}(B_d),x_0} \right]$$
$$\le C \left(\ln \frac{d}{h} \right) \left[h^2 \|e\|_{W_\infty^1(B_d)} + \|e\|_{W_\infty^{-1}(B_d)} \right]. \tag{3.14}$$

Combining (3.14), (3.10), and (3.7) we arrive at

$$\left(\ln \frac{d}{h} \right) \left[h |||F|||_{W_\infty^{0,-1}(B_d),x_0,r-2} + |||F|||_{W_\infty^{0,-r,2}(B_d),x_0} \right]$$
$$\le C \left(\ln \frac{d}{h} \right) \left[h^2 \|e\|_{W_\infty^1(B_d)} + h\|e\|_{L_\infty(B_d)} + \|e\|_{W_\infty^{-1}(B_d)} \right]. \tag{3.15}$$

Now with an inconsequential change in domains, it follows from (3.4) and (3.15) that

$$|e(x_0)| \le C \left(h^{r+1} \left(\ln \frac{d}{h} \right)^{\overline{r}} \|u\|_{W_\infty^{r+1}(B_{d/2})} + d^{-t-N/p} \|e\|_{W_p^{-t}(B_{d/2})} \right)$$
$$+ \left(\ln \frac{d}{h} \right) \left(h^2 \|e\|_{W_\infty^1(B_{d/2})} + h\|e\|_{L_\infty(B_{d/2})} + \|e\|_{W_\infty^{-1}(B_{d/2})} \right). \tag{3.16}$$

Now it was proved in [14], and may be seen from (1.15), that for $r \ge 3$ and $d \ge kh$

$$h^2 \|e\|_{W_\infty^1(B_{d/2})} \le C \left(h^{r+1} \|u\|_{W_\infty^r(B_d)} + h^2 d^{-1-N/p-t} \|e\|_{W_p^{-t}(B_d)} \right), \tag{3.17}$$

and from (1.13),

$$h\|e\|_{L_\infty(B_{d/2})} \le C \left(h^{r+1} \|u\|_{W_\infty^r(B_d)} + h d^{-N/p-t} \|e\|_{W_p^{-t}(B_d)} \right). \tag{3.18}$$

Furthermore, it is not difficult to show by a local duality argument that for $t \ge 1$ and $r \ge 3$

$$\|e\|_{W_\infty^{-1}(B_{d/2})} \le C \left(h^{r+1} \left(\ln \frac{d}{h} \right) \|u\|_{W_\infty^r(B_d)} + d^{1-N/p-t} \|e\|_{W_p^{-t}(B_d)} \right). \tag{3.19}$$

Substituting (3.17), (3.18), and (3.19) into (3.16), we easily arrive at

$$|e(x_0)| \le C \left(\left(\ln \frac{d}{h} \right)^2 h^{r+1} \|u\|_{W_\infty^{r+1}(B_d)} + \left(\ln \frac{d}{h} \right) d^{-N/p-t} \|e\|_{W_p^{-t}(B_d)} \right). \tag{3.20}$$

This completes the proof of Theorem 3.

4. A proof of Theorem 4. In view of (3.2) and (1.20), we have that

$$A(e_{\rm odd},\varphi) = -F(\varphi) \quad \text{for all} \quad \varphi \in \mathring{S}_r^h(B_d(x_0)). \tag{4.1}$$

Then using (1.20) of Corollary 2 and the fact that since $r \geq 2$ is even, $D^\alpha u_{\rm odd}(x_0) = 0$ for all $|\alpha| = r$, we obtain

$$\begin{aligned} &\| e_{\rm odd}\|_{W_\infty^1(B_h(x_0))} \\ &\quad \leq C\Big[h^r\Big(\ln\frac{d}{h}\Big)^{\overline{\overline{r}}}\|u\|_{W_\infty^{r+1}(B_d(x_0))} + d^{-1-t-N/p}\|e\|_{W_p^{-t}(B_d(x_0))}\Big] \\ &\qquad + \Big(\ln\frac{d}{h}\Big)\Big[|||F|||_{W_\infty^{0,-1}(B_d(x_0)),x_0,r-1} + |||F|||_{W_\infty^{0,-r,1}(B_d(x_0)),x_0}\Big]. \end{aligned} \tag{4.2}$$

Proceeding as in the proof of (3.7) we arrive at

$$\begin{aligned} &\Big(\ln\frac{d}{h}\Big)\Big(|||F|||_{W_\infty^{0,-1}(B_d),x_0,r-1} + |||F|||_{W_\infty^{0,-r,1}(B_d),x_0}\Big) \\ &\qquad \leq \Big(\ln\frac{d}{h}\Big)\Big(\|e\|_{L_\infty(B_d)} + h\|e\|_{W_\infty^1(B_d)}\Big). \end{aligned} \tag{4.3}$$

Using (3.17) and (3.18) in (4.3), again after an inconsequential change in domains, and substituting the result into (4.2) we obtain

$$\begin{aligned} &\| e_{\rm odd}\|_{W_\infty^1(B_h(x_0))} \\ &\quad \leq C\Big(h^r\Big(\ln\frac{d}{h}\Big)^2\|u\|_{W_\infty^{r+1}(B_d)} + \Big(\ln\frac{d}{h}\Big)d^{-1-N/p-t}\|e\|_{W_p^{-t}(B_d)}\Big). \end{aligned} \tag{4.4}$$

By definition,

$$\frac{\partial}{\partial x_i}e_{\rm odd}(x) = \left(\frac{\frac{\partial e(x)}{\partial x_i} + \frac{\partial e}{\partial x_i}(x_0 - (x - x_0))}{2}\right),$$

wherever these derivatives exist. The inequality (1.23) now follows on setting $x - x_0 = s\beta$ for an appropriate direction β and scalar $0 \leq s \leq s_0$ and taking the limit as $s \to 0$. This completes the proof of Theorem 4.

Appendix. Assumed properties of finite element spaces. We now state our assumptions on the finite element spaces used in this paper. They are the same as those used in [11], which in turn are essentially the same, with some simplifications, as those used in [13] and [14].

Let $\mathcal{D} \subset\subset \Omega \subset \mathbb{R}^N$ and for $r \geq 2$ an integer and $0 < h < 1$ a parameter, $S_r^h(\mathcal{D})$ will denote a family of finite dimensional subspaces of $W_\infty^1(\mathcal{D})$. For $G \subseteq \mathcal{D}$, $S_r^h(G)$ is the restriction of $S_r^h(\mathcal{D})$ to G and

$$\mathring{S}_r^h(G) = \varphi : \varphi \in S_r^h(G), \operatorname{supp}(\varphi) \subset G.$$

We assume that there exist positive constants k_0, δ, c_1, c_2, c_3 so that the following Assumptions A.1–A.4 are satisfied for any $\mathcal{D}_1 \subset\subset \mathcal{D}_2 \subset\subset \mathcal{D}_3 \subset\subset \mathcal{D}_4$ with $\operatorname{dist}(\mathcal{D}_i, \partial\mathcal{D}_{i+1}) \geq k_0 h$, $i = 1, 2, 3$.

Assumption A.1 (approximation). For each $v \in W_p^\ell(\mathcal{D}_2)$ there exists a $\chi \in S_r^h(\mathcal{D}_2)$ such that if $t = 0, 1$, $1 \leq \ell \leq r$, and $1 \leq p \leq \infty$,

$$\|w - \chi\|_{W_p^t(\mathcal{D}_1)} \leq c_1 h^{\ell - t}\|v\|_{W_p^\ell(\mathcal{D}_2)}. \tag{A.1}$$

In addition there exists a χ such that

$$\|v-\chi\|_{W^1_\infty(\mathcal{D}_1)} \le c_1 h^{r-1-N/p}\|v\|_{W^r_p(\mathcal{D}_2)}. \tag{A.2}$$

Furthermore, if $\operatorname{supp}(v) \subseteq \mathcal{D}_1$, then $\chi \in \mathring{S}^h(\mathcal{D}_2)$.

Assumption A.2 (inverse properties). For $\ell = 0, 1$, $t \ge 0$ an integer, and $1 \le q \le p \le \infty$,

$$\|\chi\|_{W^\ell_p(\mathcal{D}_1)} \le c_2 h^{-[\frac{N}{q}-\frac{N}{p}]-\ell-t}\|\chi\|_{W^{-t}_q(\mathcal{D}_2)} \text{ for all } \chi \in S^h_r(\mathcal{D}_2). \tag{A.3}$$

Furthermore, for any $\tau_h \subset B_d$ and integer $j = 1, \dots, r-2$,

$$\|\chi\|_{W^{1+j}_1(\tau_h)} \le Ch^{-j}\|\chi\|_{W^1_1(\tau_h)}. \tag{A.4}$$

Assumption A.3 (superapproximation). Let $\omega \in C^\infty_0(\mathcal{D}_3)$; then for each $\chi \in S^h_r(\mathcal{D}_4)$ there exists an $\eta \in \mathring{S}^h_r(\mathcal{D}_4)$ such that for $\ell = 0, 1$

$$\|\omega\chi - \eta\|_{W^\ell_2(\mathcal{D}_4)} \le c_3 h\|\omega\|_{W^\delta_\infty(\mathcal{D}_3)}\|\chi\|_{W^\ell_2(\mathcal{D}_4)}. \tag{A.5}$$

Furthermore, if $\omega \equiv 1$ on $\mathcal{D}_2$, then $\eta = \chi$ in $\mathcal{D}_1$ and

$$\|\omega\chi - \eta\|_{W^\ell_2(\mathcal{D}_4)} \le c_3 h\|\omega\|_{W^\delta_\infty(\mathcal{D}_3)}\|\chi\|_{W^\ell_2(\mathcal{D}_4\setminus\mathcal{D}_1)}. \tag{A.6}$$

Assumption A.4 (scaling). Let $x \in \mathcal{D}$ and $d \ge k_0 h$ be such that $B_d(x) \subseteq \mathcal{D}$. The linear transformation $z = (y-x)/d$ takes $B_d(x) = \{y : |y-x| < d\}$ into a new domain $\widehat{B}_1(x)$ and $S^h_r(B_d(x))$ into a new function space $\widehat{S}^{h/d}_r(\widehat{B}_1(x))$. Then $\widehat{S}^{h/d}_r(\widehat{B}_1(x))$ satisfies A.1, A.2, and A.3 with h replaced by h/d. The constants occurring in A.1, A.2, and A.3 remain the same and are independent of d.

For a discussion of these properties see [13].

REFERENCES

[1] I. Babuska, *Finite element method for domains with corners*, Computing, 6 (1970), pp. 264–273.

[2] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

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

[4] J. P. Krasovskii, *Isolation of singularities of the Green's function*, Mat. USSR-Izv., 1 (1967), pp. 935–966.

[5] P. Krizek and P. Neittaanmaki, *On convergence techniques*, Acta Appl. Math., 9 (1987), pp. 175–198.

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

[7] J. A. Nitsche, *L_∞ convergence of finite element convergence*, in Proceedings of the Second Conference on Finite Elements, Reines, France, 1975.

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

[9] R. Rannacher and R. Scott, *Some optimal error estimates for piecewise linear finite element approximations*, Math. Comp., 38 (1982), pp. 437–445.

[10] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part* I. *Global estimates*, Math. Comp., 67 (1998), pp. 877–899.

[11] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part* II. *Interior estimates*, SIAM J. Numer. Anal., 38 (2000), pp. 1269–1293.

[12] A. H. Schatz, I. Sloan, and L. B. Wahlbin, *Superconvergence in finite element methods and meshes which are locally symmetric with respect to a point*, SIAM J. Numer. Anal., 33 (1996), pp. 505–521.

[13] A. H. Schatz and L. B. Wahlbin, *Interior maximum norm estimates for finite element methods*, Math. Comp., 31 (1977), pp. 414–442.

[14] A. H. Schatz and L. B. Wahlbin, *Interior maximum norm estimates for finite element methods, Part* II, Math. Comp., 64 (1995), pp. 907–928.

[15] A. H. Schatz and L. B. Wahlbin, *On the quasi-optimality in L_∞ of the H_0^1 projection into finite element spaces*, Math. Comp., 38 (1982), pp. 1–21.

[16] R. Scott, *Interpolated boundary conditions in the finite element method*, SIAM J. Numer. Anal., 12 (1975), pp. 404–427.

[17] R. Scott, *Optimal L_∞ estimates for the finite element method on irregular grids*, Math. Comp., 30 (1976), pp. 687–697.

[18] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Springer-Verlag, New York, 1995.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2366–2386

ANALYSIS OF A NEW ERROR ESTIMATE FOR COLLOCATION METHODS APPLIED TO SINGULAR BOUNDARY VALUE PROBLEMS*

WINFRIED AUZINGER†, OTHMAR KOCH†, AND EWA WEINMÜLLER†

Abstract. We discuss an a posteriori error estimate for the numerical solution of boundary value problems for nonlinear systems of ordinary differential equations with a singularity of the first kind. The estimate for the global error of an approximation obtained by collocation with piecewise polynomial functions is based on the defect correction principle. We prove that for collocation methods which are not superconvergent, the error estimate is asymptotically correct. As an essential prerequisite we derive convergence results for collocation methods applied to nonlinear singular problems.

Key words. boundary value problems, singularity of the first kind, collocation methods, error estimate, defect correction, asymptotical correctness

AMS subject classification. 65L05

DOI. 10.1137/S0036142902418928

1. Introduction.

In this paper, we discuss the numerical solution of singular boundary value problems of the form

$$z'(t) = \frac{M(t)}{t} z(t) + f(t, z(t)), \quad t \in (0, 1], \tag{1.1a}$$

$$B_a z(0) + B_b z(1) = \beta, \tag{1.1b}$$

$$z \in C[0, 1], \tag{1.1c}$$

where z is an n-dimensional real function, M is a smooth $n \times n$ matrix, and f is an n-dimensional smooth function on a suitable domain. B_a and B_b are constant $r \times n$ matrices, with $r < n$. In section 3 we will demonstrate that condition (1.1c) is equivalent to a set of $n - r$ linearly independent conditions $z(0)$ must satisfy. These boundary conditions are augmented by (1.1b) to yield an isolated solution z. In this paper, we restrict our attention to the class of singular boundary value problems which are equivalent to a well-posed singular initial value problem, where all boundary conditions are posed at $t = 0$. In this case, a shooting argument can be used to derive a representation of the solution convenient for our analysis. This implies certain restrictions on the spectrum of the matrix $M(0)$, which will be discussed in section 3.

The search for an efficient numerical method to solve problems (1.1) is strongly motivated by numerous applications from physics, chemistry, mechanics, or ecology; see, for example, [15], [28]. Also, research activities in related fields, like the computation of connecting orbits in dynamical systems [21] or singular Sturm–Liouville problems [6], may benefit from techniques developed for problems of the form (1.1). The problem class discussed in this paper, where $M(0)$ has no eigenvalues with positive real parts, arises in applications from mechanics (buckling of spherical shells

*Received by the editors December 2, 2002; accepted for publication (in revised form) April 13, 2004; published electronically March 31, 2005. This project was supported by the Austrian Science Fund (FWF) under grant P-15072-MAT.

http://www.siam.org/journals/sinum/42-6/41892.html

†Institute for Analysis and Scientific Computing, Vienna University of Technology, A-1040 Vienna, Austria (w.auzinger@tuwien.ac.at, othmar@othmar-koch.org, e.weinmueller@tuwien.ac.at).

[22], [25]), chemical reactor theory (cf. [14]), and avalanche dynamics (see [18] and [19]). Moreover, Dirichlet problems for certain nonlinear elliptic equations lead to this problem class when certain symmetries are present; see [23]. The computation of self-similar solution profiles for the nonlinear Schrödinger equation is also essentially reduced to this problem type; see [7]. However, our restriction on the spectrum of $M(0)$ excludes problems of the type

$$y''(t) + \frac{1}{t}y'(t) - \frac{1}{t^2}y(t) = f(t)$$

(see [20], [22]) from the treatment. The first order system resulting from the Euler transformation $z(t) = (y(t), ty'(t))$ does not belong to the class considered here.

To compute the numerical solution of (1.1), we use polynomial collocation at collocation points placed in the interior of every collocation interval. Collocation has been used in one of the best established standard codes for (regular) boundary value problems, COLSYS (COLNEW); see [1] and [2]. In COLSYS, (superconvergent) collocation at Gaussian points is used; cf. [8]. Our decision to use collocation was motivated by its advantageous convergence properties for (1.1), while in the presence of a singularity other high order methods show order reductions and become inefficient (see, for example, [11]). For linear problems (1.1) which can equivalently be posed as initial value problems, it was shown in [10] that the convergence order of collocation methods is at least equal to the *stage order* of the method. We will discuss the restrictions implied by the latter requirement in section 3. For the general class (1.1), numerical evidence suggests that the convergence order is at least equal to the stage order for both the linear and the nonlinear case;[1] cf. [5]. However, we cannot expect to observe superconvergence (cf. [8]) when collocation is applied to (1.1) in general. At most, a convergence order of $O(|\ln(h)|^{n_0-1}h^{m+1})$, for some positive integer n_0, holds for a method of stage order m; see [10]. Consequently, a restriction to collocation at an even number of equidistant points, which implies that the convergence order is at most $O(h^m)$, does not limit the method's accuracy significantly. We use these collocation nodes in practice, since it turns out that the error of the error estimate we propose in this paper is $O(|\ln(h)|^{n_0-1}h^{m+1})$. This means that the estimate is asymptotically correct when the order of the collocation method is not higher than the stage order.

Our main aim was to construct an efficient asymptotically correct error estimate for the global error of the numerical solution obtained by collocation. This estimate, introduced in [5], is based on the defect correction principle, which was first considered in [29] for the estimation of the global error of Runge–Kutta methods. In [29], the estimate for the error at the mesh points is obtained by applying the (high order) basic numerical scheme twice, to the original and to a suitably defined "neighboring problem." An extension of this idea proposed in [12], [24] avoids the second application of the high order scheme, using a cheap low order method instead. Again, this estimate is asymptotically correct at the mesh points only. A further modification proposed by the authors provides an error estimate which is asymptotically correct at both the mesh and the collocation points. The analysis of this estimate in the context of nonlinear regular problems was given in [5]. It could be shown that for a collocation method of stage order $O(h^m)$, the error of the estimate (the difference between the global error and its estimate) is of order $O(h^{m+1})$. Numerical evidence suggests that

[1] The analysis given in [26] for second order problems might provide tools to prove this assertion.

this is also true for singular problems. In this paper, we will prove this assertion for the class of singular problems (1.1).

The collocation method and error estimate described in this paper were also implemented in the MATLAB code `sbvp` designed especially to solve singular boundary value problems. The error estimate yields a reliable basis for a mesh selection procedure which enables an efficient computation of the numerical solution. A description of the code and experimental evidence of its advantageous properties are given in [4].

The paper is organized as follows: The analytical properties of (1.1) which were discussed in detail in [9] are briefly recapitulated in section 3. In section 4.1, the results for collocation methods according to [10] are given. Using these results, we derive new, refined bounds for the errors of the numerical solution and its derivative, and we extend these results to the nonlinear case. This analysis is carried out in section 4.2. In section 5 we use these estimates for the collocation solution in order to prove that our version of the error estimate is asymptotically correct for problem (1.1). Finally, in section 6 we give a numerical example which illustrates the theory.

2. Preliminaries. Throughout the paper, the following notation is used. We denote by $\mathbb{R}^n$ the space of real vectors of dimension n and use $|\cdot|$,

$$|x| = |(x_1, x_2, \ldots, x_n)^T| := \max_{1\le i\le n} |x_i|,$$

to denote the maximum norm in $\mathbb{R}^n$. $C_n^p[0,1]$ is the space of real vector-valued functions which are p times continuously differentiable on $[0,1]$. For functions $y \in C_n^0[0,1]$ we define the maximum norm,

$$\|y\|_{[0,1]} := \max_{0\le t\le 1} |y(t)|,$$

or more generally for an interval $J \subseteq [0,1]$,

$$\|y\|_J := \max_{t\in J} |y(t)|.$$

$C_{n\times n}^p[0,1]$ is the space of real $n \times n$ matrices with columns in $C_n^p[0,1]$. For a matrix $A = (a_{ij})_{i,j=1}^n$, $A \in C_{n\times n}^0[0,1]$, $\|A\|_{[0,1]}$ is the induced norm,

$$\|A\|_{[0,1]} = \max_{0\le t\le 1} |A(t)| = \max_{0\le t\le 1} \left(\max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}(t)| \right).$$

Where there is no confusion, we will omit the subscripts n and $n \times n$ and denote $C[0,1] = C^0[0,1]$.

For the numerical analysis, we define meshes

$$\Delta := (\tau_0, \tau_1, \ldots, \tau_N),$$

and $h_i := \tau_{i+1} - \tau_i,\ i = 0, \ldots, N-1,\ \tau_0 = 0,\ \tau_N = 1$. On Δ, we define corresponding grid vectors

$$u_\Delta := (u_0, \ldots, u_N) \in \mathbb{R}^{(N+1)n}.$$

The norm on the space of grid vectors is given by

$$\|u_\Delta\|_\Delta := \max_{0\le k\le N} |u_k|.$$

For a continuous function $y \in C[0,1]$, we denote by R_Δ the pointwise projection onto the space of grid vectors,

$$R_\Delta(y) := (y(\tau_0), \dots, y(\tau_N)).$$

For collocation, m points spaced at distances $h_i\delta_j,\ j = 1, \dots, m$, are inserted in each subinterval $J_i := [\tau_i, \tau_{i+1}]$. This yields the (fine) grid[2] (see Figure 2.1)

$$\Delta^m := \left\{ t_{i,j} :\ t_{i,j} = \tau_i + h_i \sum_{k=0}^{j} \delta_k,\ i = 0, \dots, N-1,\ j = 0, \dots, m+1 \right\}. \tag{2.1}$$

We restrict ourselves to grids where $\delta_1 > 0$ to avoid a special treatment of the singular point $t = 0$. For the analysis of collocation methods, we allow $\delta_{m+1} = 0$. In the discussion of the error estimate, we use the further restriction $\delta_{m+1} > 0$. This requirement is satisfied for equidistant collocation points which we use in practice (see section 1), where

$$\delta_j := \frac{1}{m+1}, \quad j = 1, \dots, m+1. \tag{2.2}$$

For a grid Δ^m, u_{Δ^m}, $\|\cdot\|_{\Delta^m}$, and R_{Δ^m} are defined accordingly.

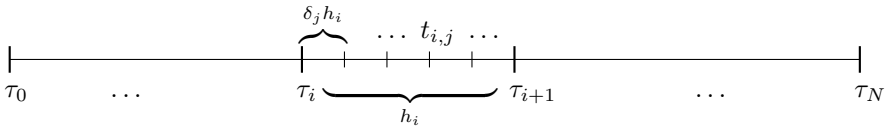


FIG. 2.1. *The computational grid.*

3. Analytical results. In this section we discuss the analytical properties of (1.1); cf. [9]. Here, we assume all eigenvalues of $M(0)$ have nonpositive real parts. Moreover, the only eigenvalue of $M(0)$ on the imaginary axis is zero. These restrictions are necessary to ensure that we can use a shooting argument to derive a representation of the solution convenient for our theory.[3]

First, we treat the linear case,

$$z'(t) = \frac{M(t)}{t} z(t) + f(t), \quad t \in (0,1], \tag{3.1a}$$

$$B_a z(0) + B_b z(1) = \beta, \tag{3.1b}$$

$$z \in C[0,1], \tag{3.1c}$$

where $B_a, B_b \in \mathbb{R}^{r\times n},\ r < n$, are constant matrices, and $\beta \in \mathbb{R}^r$ is a constant vector.

Throughout, we assume $M \in C^1[0,1]$. Consequently, we can rewrite $M(t)$ and obtain

$$M(t) = M(0) + tC(t) \tag{3.2}$$

[2]For convenience, we denote τ_i by $t_{i,0} \equiv t_{i-1,m+1},\ i = 1, \dots, N-1$. Moreover, we define $\delta_0 := 0,\ \delta_{m+1} := (t_{i,m+1} - t_{i,m})/h_i$. Note that we choose the same distribution of collocation points in every subinterval J_i, and that $\sum_{j=0}^{m+1} \delta_j = 1$ holds for $i = 0, \dots, N-1$.

[3]Note, however, that we do not use shooting when we actually compute the numerical solution.

with a continuous matrix $C(t)$.

Let X_0 be the kernel of $M(0)$ and let R be a projection onto X_0, where the rank of R is equal to r. We define

$$S := I_n - R,$$

where we denote by I_n the $n \times n$ identity matrix. The necessary and sufficient condition for z to be continuous on $[0,1]$ is

$$Sz(0) = 0.$$

This yields

$$z(0) = (S+R)z(0) = Rz(0),$$

and due to

$$M(0)z(0) = MRz(0) = 0$$

it follows that (3.1c) is equivalent to $z(0) \in \ker(M(0))$. These conditions are augmented by (3.1b) to yield a unique solution.

We denote by $\tilde{E}$ the $n \times r$ matrix consisting of a maximal set of linearly independent columns of R. Moreover, let $Z(t) = (Z_1(t), \ldots, Z_r(t))$ be the fundamental solution matrix of the initial value problem

$$Z'(t) = \frac{M(t)}{t} Z(t), \quad t \in (0,1], \tag{3.3a}$$

$$Z(0) = \tilde{E}. \tag{3.3b}$$

The necessary and sufficient condition for problem (3.1) to have a unique solution is that the $r \times r$ matrix Q,

$$Q := B_a \tilde{E} + B_b Z(1), \tag{3.4}$$

be nonsingular. In this case, we can represent the solution z of (3.1) by

$$z(t) = \sum_{k=1}^{r} a_k Z_k(t) + \tilde{z}(t), \tag{3.5}$$

where $\tilde{z}$ is the solution of

$$\tilde{z}'(t) = \frac{M(t)}{t} \tilde{z}(t) + f(t), \quad t \in (0,1], \tag{3.6a}$$

$$\tilde{z}(0) = 0. \tag{3.6b}$$

The coefficients $a = (a_1, \ldots, a_r)$ are uniquely determined by $Qa = \beta - B_b \tilde{z}(1)$.

For the solution of the linear problem (3.1), $z \in C^{k+1}[0,1]$ holds if $f \in C^k[0,1]$ and $M \in C^{k+1}[0,1]$.

Now we discuss the nonlinear problem[4]

$$z'(t) = \frac{M(t)}{t} z(t) + f(t, z(t)), \quad t \in (0,1], \tag{3.7a}$$

$$B_a z(0) + B_b z(1) = \beta, \tag{3.7b}$$

$$M(0)z(0) = 0. \tag{3.7c}$$

In order to formulate analogous smoothness properties for z, we make the following assumptions:

[4]Again, we assume that $M(0)$ has only eigenvalues with negative real parts or the eigenvalue 0.

1. $f : D_1 \to \mathbb{R}^n$ is a nonlinear mapping, where $D_1 \subseteq [0,1] \times \mathbb{R}^n$ is a suitable set.
2. Equation (3.7) has a solution $z \in C[0,1] \cap C^1(0,1]$. With this solution and a $\rho > 0$ we associate the closed balls

$$S_\rho(z(t)) := \{x \in \mathbb{R}^n : \ |z(t) - x| \le \rho\}$$

and the tube

$$T_\rho(z) := \{(t,x) : \ t \in [0,1], \ x \in S_\rho(z(t))\}.$$

3. $f(t,z)$ is continuously differentiable with respect to z, and $\frac{\partial f(t,z)}{\partial z}$ is continuous on $T_\rho(z)$.
4. The solution z is isolated. This means that

$$\begin{aligned} &u'(t) = \frac{M(t)}{t}u(t) + A(t)u(t), \quad t \in (0,1], \\ &B_a u(0) + B_b u(1) = 0, \\ &M(0)u(0) = 0, \end{aligned}$$

where

$$A(t) := \frac{\partial f}{\partial z}(t, z(t))$$

has only the trivial solution.

Under these assumptions and for $f \in C^k(T_\rho(z)), \ M \in C^{k+1}[0,1]$, the solution z of (3.7) satisfies $z \in C^{k+1}[0,1]$.

For further details and proofs see [9].

4. Collocation methods. In this section, we derive new, refined error bounds for collocation methods applied to (1.1), relying on earlier results formulated in [10]. Moreover, we extend the convergence analysis to the nonlinear case. For reasons of simplicity, we restrict the discussion to equidistant meshes, $h_i = h, \ i = 0, \dots, N-1$, because the results from [10] are formulated for this situation. However, the results also hold for nonuniform meshes which have a limited variation in the stepsizes; see [10, section 6].

Let us denote by B the Banach space of continuous, piecewise polynomial functions $q \in \mathbb{P}_m$ of degree $\le m, \ m \in \mathbb{N}$ (m is called the *stage order* of the method), equipped with the norm $\|\cdot\|_{[0,1]}$. As an approximation for the exact solution z of (1.1), we define an element of B which satisfies the differential equation (1.1a) at a finite number of points and which is subject to the same boundary conditions. Since we require the numerical solution to satisfy (1.1c), we introduce the space $B_1 \subset B$, such that $M(0)q(0) = 0 \ \forall\, q \in B_1$. Thus, we are seeking a function $p(t) = p_i(t), \ t \in J_i, \ i = 0, \dots, N-1$, in B_1 which satisfies

$$p_i'(t_{i,j}) = \frac{M(t_{i,j})}{t_{i,j}} p_i(t_{i,j}) + f(t_{i,j}, p_i(t_{i,j})), \quad i = 0, \dots, N-1, \ j = 1, \dots, m, \tag{4.1a}$$

$$B_a p(0) + B_b p(1) = \beta. \tag{4.1b}$$

We consider collocation on general grids Δ^m as defined in section 1, subject to the restriction $\delta_1 > 0$.

4.1. Earlier results. In [10], collocation methods for linear problems were studied. For the analysis of the nonlinear case in section 4.2, bounds for the collocation solution $p \in B_1$ need to be specified. Here, the relevant preliminaries from [10] are recapitulated.

Thus, we consider the solution $p \in B_1$ of

$$p'(t_{i,j}) = \frac{M(t_{i,j})}{t_{i,j}} p(t_{i,j}) + f(t_{i,j}), \quad i = 0, \dots, N-1,\ j = 1, \dots, m, \tag{4.2a}$$

$$B_a p(0) + B_b p(1) = \beta. \tag{4.2b}$$

LEMMA 4.1. *For $\mu, \beta \in \{0, 1\}$ and arbitrary constants $c_{i,j}$, there exists a unique $p \in B_1$ which satisfies*

$$p'(t_{i,j}) = \frac{M(0)}{t_{i,j}} p(t_{i,j}) + \frac{M(0)^\mu}{t_{i,j}^\beta} c_{i,j}, \quad i = 0, \dots, N-1,\ j = 1, \dots, m, \tag{4.3a}$$

$$p(0) = 0. \tag{4.3b}$$

Furthermore,

$$\|p\|_{J_i} \le \text{const.}\, \tau_{i+1}^{1-\beta} |\ln(h)|^{(\beta(n_0-\mu))_+} C_i, \quad i = 0, \dots, N-1, \tag{4.4}$$

where n_0 is the dimension of the largest Jordan block of $M(0)$ corresponding to the eigenvalue 0,

$$(x)_+ := \begin{cases} x, & x \ge 0, \\ 0, & x < 0, \end{cases}$$

and

$$C_i := \max_{\substack{l = 0, \dots, i \\ j = 1, \dots, m}} |c_{l,j}|.$$

Proof. See [10, Lemma 4.4]. ∎

The following result is a slightly modified version of [10, Theorem 4.1].

THEOREM 4.2. *For $\mu, \beta \in \{0, 1\}$, consider the problem*

$$p'(t_{i,j}) = \frac{M(t_{i,j})}{t_{i,j}} p(t_{i,j}) + \frac{M(0)^\mu}{t_{i,j}^\beta} c_{i,j}, \quad i = 0, \dots, N-1,\ j = 1, \dots, m, \tag{4.5a}$$

$$p(0) = \delta \in \ker(M(0)). \tag{4.5b}$$

There exists a unique solution of (4.5) when h is sufficiently small, and this solution satisfies

$$\|p\|_{J_i} \le \text{const.}\, (|\delta| + \tau_{i+1}^{1-\beta} |\ln(h)|^{(\beta(n_0-\mu))_+} C_i), \quad i = 0, \dots, i_0, \tag{4.6}$$

for a suitable $i_0 \le N-1$.

Proof. In [10, Theorem 4.1], the estimate following [10, formula (4.15)] can be replaced by

$$\|p\|_{J_i} \le \kappa(\tau_{i+1} \|p\|_{[0,\tau_{i+1}]} + |\delta| + \tau_{i+1}^{1-\beta} |\ln(h)|^{(\beta(n_0-\mu))_+} C_i), \quad i = 0, \dots, i_0, \tag{4.7}$$

if the results of [10, Lemma 4.4] are suitably applied. Substitution of the bound for p derived in [10, Theorem 4.1] into the right-hand side of (4.7) yields the result. ∎

Note that the existence of the solution of (4.5) and the estimate (4.6) are shown only on an interval $[0, b]$, where b is sufficiently small (but independent of h). Thus, we need to use classical theory for regular problems to ensure the existence of the solution on the whole interval. In what follows, we treat the underlying singular problems only on the restricted interval, and we apply classical results for collocation from [3], and the error estimate analysis for regular problems from [5], to complete the proofs.

4.2. New error bounds. First, we use Theorem 4.2 to derive bounds for the solution $p \in B_1$ of the general linear problem (4.2) and for its derivative p'. By the superposition principle, p can be written in the form

$$p(t) = \sum_{k=1}^{r} b_k P_k(t) + \tilde{p}(t), \tag{4.8}$$

analogous to (3.5) for the exact solution. Here, $P(t) = (P_1(1), \ldots, P_r(t))$ is the $n \times r$ matrix solution of

$$P'(t_{i,j}) = \frac{M(t_{i,j})}{t_{i,j}} P(t_{i,j}), \quad i = 0, \ldots, N-1,\ j = 1, \ldots, m, \tag{4.9a}$$

$$P(0) = \tilde{E}, \tag{4.9b}$$

whose columns are in B_1, and $\tilde{p}$ satisfies

$$\tilde{p}'(t_{i,j}) = \frac{M(t_{i,j})}{t_{i,j}} \tilde{p}(t_{i,j}) + f(t_{i,j}), \quad i = 0, \ldots, N-1,\ j = 1, \ldots, m, \tag{4.10a}$$

$$\tilde{p}(0) = 0. \tag{4.10b}$$

It was shown in [10, Theorem 4.4] that the representation (4.8) is well defined. Of course, the coefficients b_k could be computed in principle from the boundary conditions, as in the case of the analytical problem; the representation (4.8) is used only to describe the structure of the solution p, and therefore we refrain from specifying b_k explicitly. Next, we derive convergence results for the quantities appearing in the representation (4.8) using arguments similar to those given in [10, Theorem 4.2].

Consider the solutions z and q of (3.1a) and (4.2a), respectively, subject to the initial conditions $z(0) = q(0) = \delta \in \ker(M(0))$. We define an error function $e \in B_1$ by

$$\begin{aligned} &e'(t_{i,j}) = z'(t_{i,j}) - q'(t_{i,j}), \quad i = 0, \ldots, N-1,\ j = 1, \ldots, m, \\ &e(0) = 0. \end{aligned}$$

From standard results for interpolation (see, for example, [13]), we conclude that

$$e(t) = z(t) - q(t) + tO(h^m)$$

if z is sufficiently smooth, whence

$$e'(t_{i,j}) = \frac{M(t_{i,j})}{t_{i,j}} e(t_{i,j}) + O(h^m), \quad i = 0, \ldots, N-1,\ j = 1, \ldots, m, \tag{4.11a}$$

$$e(0) = 0. \tag{4.11b}$$

Now, Theorem 4.2 yields

$$\|e\|_{J_i} \le \tau_{i+1} O(h^m), \quad i = 0, \ldots, i_0, \tag{4.12}$$

and consequently

$$\|z-q\|_{J_i} \le \tau_{i+1} O(h^m), \quad i = 0, \dots, i_0. \tag{4.13}$$

It follows from (4.11a) and (4.12) that $e'(t_{i,j}) = O(h^m)$, which implies

$$\|z' - q'\|_{[0,1]} = O(h^m). \tag{4.14}$$

Finally, we show that the residual of q with respect to (3.1a) has the same asymptotic quality. Since $q \in C[0,1]$ and q' has only a finite number of jump discontinuities in $[0,1]$, we can use the representations

$$q(t) = \delta + t \int_0^1 q'(st)\, ds, \tag{4.15a}$$

$$z(t) = \delta + t \int_0^1 z'(st)\, ds \tag{4.15b}$$

to conclude that

$$\begin{aligned} q'(t) - \frac{M(t)}{t} q(t) - f(t) &= q'(t) - z'(t) + \frac{M(t)}{t} t \int_0^1 (q'(st) - z'(st))\, ds \\ &= O(h^m), \quad t \in [0,1]. \end{aligned} \tag{4.16}$$

This means that the refined bounds (4.13), (4.14), and (4.16) hold for the fundamental modes P_k and the particular solution $\tilde{p}$ in (4.8). To show that these bounds also hold for the solution p of (4.2), we have to estimate the differences $|a_k - b_k|$ for $k = 1, \dots, r$. We substitute (3.5) and (4.8) into (3.1b) and obtain a system of linear equations for $a_k - b_k$. This system is nonsingular since Q from (3.4) is nonsingular and $P(1) = Z(1) + O(h^m)$. This implies

$$b_k = a_k + O(h^m), \quad k = 1, \dots, r; \tag{4.17}$$

see also [10, Theorem 4.5].

Consequently, the following result holds.

THEOREM 4.3. *Consider the solution $p \in B_1$ of (4.2) as an approximation of the (sufficiently smooth[5]) solution z of (3.1). Then, for a sufficiently small stepsize h and a suitable $i_0 \le N-1$, the following bounds hold:*

$$z(t) - p(t) = \tilde{E} O(h^m) + \tau_{i+1} O(h^m), \quad t \in J_i,\ i = 0, \dots, i_0, \tag{4.18a}$$

$$\|z' - p'\|_{[0,1]} = O(h^m), \tag{4.18b}$$

$$\left| p'(t) - \frac{M(t)}{t} p(t) - f(t) \right| = O(h^m), \quad t \in [0,1]. \tag{4.18c}$$

Proof. The result follows immediately on noting that $P(t)$ can also be written in a form given by (4.15a), and therefore

$$\begin{aligned} z(t) - p(t) &= \sum_{k=1}^{r} a_k (Z_k(t) - P_k(t)) + \sum_{k=1}^{r} (a_k - b_k) P_k(t) + \tilde{z}(t) - \tilde{p}(t) \\ &= \tau_{i+1} O(h^m) + (\tilde{E} + tO(1)) O(h^m) + \tau_{i+1} O(h^m), \quad t \in J_i. \end{aligned}$$

[5] We require that $z \in C^{m+1}[0,1]$, which holds if $f \in C^m[0,1]$ and $M \in C^{m+1}[0,1]$.

The bounds (4.18b) and (4.18c) are direct consequences of this representation. ∎

To prove the analogous convergence results for nonlinear problems, we use techniques developed in [17]. In order to show the existence of the solution and derive the error bounds, we rewrite the problem in an abstract Banach space setting and apply the Banach fixed point theorem. The arguments are similar to those given in the proof of [17, Theorem 3.6], but we cannot use this theorem directly, because some of the assumptions made there are violated and, also, a refined error estimate is required. Therefore, we need to repeat the main steps of the proof.

We write the collocation problem as an operator equation

$$F(p) = 0, \tag{4.19}$$

where $F : B_1 \to B_2$ is defined by

$$F(p) = \begin{pmatrix} p'(t_{i,j}) - \frac{M(t_{i,j})}{t_{i,j}} p(t_{i,j}) - f(t_{i,j}, p(t_{i,j})), \quad i = 0, \dots, N-1,\ j = 1, \dots, m \\ B_a p(0) + B_b p(1) - \beta \end{pmatrix},$$

and B_1 and B_2 are Banach spaces,

$$B_1 = (\{q \in \mathbb{P}_m :\ M(0)q(0) = 0\}, \|\cdot\|_{[0,1]}), \quad B_2 = (\mathbb{R}^{Nmn+r}, |\cdot|).$$

For $p \in B_1$, the Fréchet derivative $DF(p) : B_1 \to B_2$ of F is given by

$$DF(p)q = \begin{pmatrix} q'(t_{i,j}) - \frac{M(t_{i,j})}{t_{i,j}} q(t_{i,j}) - D_2 f(t_{i,j}, p(t_{i,j}))q(t_{i,j}), \quad i = 0, \dots, N-1,\ j = 1, \dots, m \\ B_a q(0) + B_b q(1) \end{pmatrix},$$

where $D_2 f(t, z)$ is the Fréchet derivative of f with respect to z.

If $D_2 f$ is Lipschitz, then DF also satisfies a Lipschitz condition with the same constant,

$$\begin{aligned} |(DF(p_1) - DF(p_2))q| &= \left| \begin{pmatrix} (D_2 f(t_{i,j}, p_1(t_{i,j})) - D_2 f(t_{i,j}, p_2(t_{i,j})))q(t_{i,j}) \ \ \forall\, i, j \\ 0 \end{pmatrix} \right| \\ &\le L \|p_1 - p_2\|_{[0,1]} \|q\|_{[0,1]}. \end{aligned}$$

For the convergence proof, we require all assumptions from section 3 to hold. In particular, this means that an isolated, smooth solution z of (1.1) exists. Using this function, we now construct an auxiliary element $p_{\mathrm{ref}} \in B_1$ for the proof of the existence of a solution p of (4.1). We require that p_{ref} satisfy

$$p'_{\mathrm{ref}}(t_{i,j}) = z'(t_{i,j}), \quad i = 0, \dots, N-1,\ j = 1, \dots, m, \tag{4.20a}$$

$$B_a p_{\mathrm{ref}}(0) + B_b p_{\mathrm{ref}}(1) = \beta. \tag{4.20b}$$

Since p'_{ref} is a piecewise polynomial of degree $\le m-1$, it is uniquely defined by the system (4.20a). Moreover,

$$\|z' - p'_{\mathrm{ref}}\|_{[0,1]} = O(h^m). \tag{4.21}$$

Representing p_{ref} by means of (4.15a), we conclude that

$$z(t) - p_{\mathrm{ref}}(t) = \tilde{E}(r_1 - r_2) + tO(h^m), \quad r_1, r_2 \in \mathbb{R}^r.$$

Substitution into (4.20b) yields

$$(B_a + B_b)\tilde{E}(r_1 - r_2) = O(h^m).$$

For further analysis, we assume that

$$\tilde{Q} := (B_a + B_b)\tilde{E} \quad \text{is nonsingular.} \tag{4.22}$$

This implies $r_1 - r_2 = O(h^m)$, and consequently

$$z(t) - p_{\mathrm{ref}}(t) = \tilde{E}O(h^m) + tO(h^m). \tag{4.23}$$

Remark. Assumption (4.22) is quite natural. If we require that boundary value problems consisting of (1.1a) posed on intervals $(0, b]$, $0 < b \leq 1$, and boundary conditions $M(0)z(0) = 0$ and $B_a z(0) + B_b z(b) = \beta$ have unique, continuous solutions, then (4.22) follows. Moreover, we can interpret (4.20) as the (regular) collocation problem associated with the boundary value problem

$$\begin{aligned} &y'(t) = z'(t), \quad t \in (0, 1], \\ &B_a y(0) + B_b y(1) = \beta, \\ &M(0)y(0) = 0. \end{aligned}$$

Obviously, $y(t) = z(t)$ is a solution of this reconstruction problem, and if we require the solution to be unique, then (4.22) must hold. Note that (4.22) always holds for problems with separated boundary conditions.

We now use (4.23) to derive the following relation:

$$\begin{aligned} F(p_{\mathrm{ref}}) &= \begin{pmatrix} p'_{\mathrm{ref}}(t_{i,j}) - \frac{M(t_{i,j})}{t_{i,j}} p_{\mathrm{ref}}(t_{i,j}) - f(t_{i,j}, p_{\mathrm{ref}}(t_{i,j})) \ \ \forall\, i,j \\ B_a p_{\mathrm{ref}}(0) + B_b p_{\mathrm{ref}}(1) - \beta \end{pmatrix} \\ &= \begin{pmatrix} p'_{\mathrm{ref}}(t_{i,j}) - z'(t_{i,j}) - \frac{M(t_{i,j})}{t_{i,j}} (p_{\mathrm{ref}}(t_{i,j}) - z(t_{i,j})) \\ \qquad\qquad -f(t_{i,j}, p_{\mathrm{ref}}(t_{i,j})) + f(t_{i,j}, z(t_{i,j})) \ \ \forall\, i,j \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{M(0)}{t_{i,j}} (\tilde{E}O(h^m) + t_{i,j}O(h^m)) + O(h^m) \ \ \forall\, i,j \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} O(h^m) \\ 0 \end{pmatrix}. \end{aligned} \tag{4.24}$$

Finally, we give an estimate for $DF^{-1}(p_{\mathrm{ref}})$. Note that

$$q := DF^{-1}(p_{\mathrm{ref}}) \left(\begin{pmatrix} \gamma_{i,j} \ \ \forall\, i,j \\ \tilde{\beta} \end{pmatrix} \right)$$

is the solution of the linear collocation problem

$$q'(t_{i,j}) = \frac{M(t_{i,j})}{t_{i,j}} q(t_{i,j}) + D_2 f(t_{i,j}, p_{\mathrm{ref}}(t_{i,j}))q(t_{i,j}) + \gamma_{i,j} \ \ \forall\, i,j, \tag{4.25a}$$

$$B_a q(0) + B_b q(1) = \tilde{\beta}. \tag{4.25b}$$

Since for sufficiently small h, p_{ref} is in $T_\rho(z)$, this problem is well defined. Finally, from Theorem 4.2, we have

$$\|q\|_{J_i} \leq \text{const.}\, (|\tilde{\beta}| + \tau_{i+1}\gamma_i), \tag{4.26}$$

where

$$\gamma_i = \max_{\substack{l=0,\dots,i \\ j=1,\dots,m}} |\gamma_{l,j}|.$$

With these preliminary results we can prove the next theorem.

THEOREM 4.4. *Let z be an isolated, sufficiently smooth solution of* (1.1). *For sufficiently small h and $\rho > 0$, the nonlinear collocation scheme* (4.1) *has a unique solution p in the tube $T_\rho(z)$ around z. Moreover, the estimates* (4.18) *hold.*

Proof. We proceed in a manner similar to the proof of [17, Theorem 3.6]. Define a mapping $G : B_1 \to B_1$,

$$G(q) := q - DF^{-1}(p_{\mathrm{ref}})F(q). \tag{4.27}$$

Obviously, $F(p) = 0$ is equivalent to the fixed point equation $G(p) = p$. We use the Banach fixed point theorem to show that this equation has a unique solution in a suitably chosen closed ball

$$K := K(p_{\mathrm{ref}}, \rho_0) := \{q \in B_1 : \ \|q - p_{\mathrm{ref}}\|_{[0,1]} \le \rho_0\}.$$

To show that G is a contraction, we write

$$q := G(p_1) - G(p_2) = DF^{-1}(p_{\mathrm{ref}})(DF(p_{\mathrm{ref}}) - \widehat{DF}(p_1, p_2))(p_1 - p_2)$$

for $p_1, p_2 \in K$, where

$$\widehat{DF}(p_1, p_2) := \int_0^1 DF(\tau p_1 + (1-\tau)p_2)\, d\tau.$$

Consequently, q is the solution of the scheme (4.25), where

$$\begin{aligned}\left|\begin{pmatrix} \gamma_{i,j} \ \ \forall\, i,j \\ \tilde{\beta} \end{pmatrix}\right| &= \left|\int_0^1 (DF(p_{\mathrm{ref}}) - DF(\tau p_1 + (1-\tau)p_2))\, d\tau (p_1 - p_2)\right| \\ &\le L\rho_0 \|p_1 - p_2\|_{[0,1]}\end{aligned}$$

due to the Lipschitz condition which DF satisfies. Thus, it follows from (4.26) that G is a contraction with constant $\tilde{L} < 1$ if ρ_0 is sufficiently small. To show that G maps K into itself, we estimate for $q \in K$,

$$\|p_{\mathrm{ref}} - G(q)\|_{[0,1]} \le \|p_{\mathrm{ref}} - G(p_{\mathrm{ref}})\|_{[0,1]} + \|G(p_{\mathrm{ref}}) - G(q)\|_{[0,1]},$$

where $p_{\mathrm{ref}} - G(p_{\mathrm{ref}}) = DF^{-1}(p_{\mathrm{ref}})F(p_{\mathrm{ref}})$ is the solution of (4.25) with $\gamma_{i,j} = O(h^m)$ and $\tilde{\beta} = 0$; cf. (4.24). Thus,

$$\|p_{\mathrm{ref}} - G(q)\|_{[0,1]} \le O(h^m) + \tilde{L}\rho_0 \le \rho_0, \tag{4.28}$$

provided that h is sufficiently small. The Banach fixed point theorem now implies that a solution $p \in B_1$ of (4.1) exists.

We now prove the convergence results (4.18). From

$$\begin{aligned}\|p_{\mathrm{ref}} - p\|_{J_i} = \|p_{\mathrm{ref}} - G(p)\|_{J_i} &\le \|p_{\mathrm{ref}} - G(p_{\mathrm{ref}})\|_{J_i} + \|G(p_{\mathrm{ref}}) - G(p)\|_{J_i} \\ &\le \tau_{i+1} O(h^m) + \tilde{L}\|p_{\mathrm{ref}} - p\|_{J_i}\end{aligned}$$

we have $\|p_{\rm ref} - p\|_{J_i} \le \tau_{i+1} O(h^m)$, which together with (4.23) yields

$$\begin{aligned} z(t) - p(t) &= z(t) - p_{\rm ref}(t) + p_{\rm ref}(t) - p(t) \\ &= \tilde{E} O(h^m) + t O(h^m) + \tau_{i+1} O(h^m), \quad t \in J_i. \end{aligned} \tag{4.29}$$

Consequently, (4.18a) follows. Next, we choose a piecewise polynomial function $e \in B_1$ satisfying $e'(t_{i,j}) = z'(t_{i,j}) - p'(t_{i,j})$. Therefore, $e'(t) = z'(t) - p'(t) + O(h^m)$. Moreover, (4.29) implies

$$\begin{aligned} e'(t_{i,j}) &= z'(t_{i,j}) - p'(t_{i,j}) \\ &= \frac{M(t_{i,j})}{t_{i,j}} (z(t_{i,j}) - p(t_{i,j})) + f(t_{i,j}, z(t_{i,j})) - f(t_{i,j}, p(t_{i,j})) \\ &= O(h^m), \quad i = 0, \dots, N-1,\ j = 1, \dots, m. \end{aligned}$$

Thus $e'(t) = O(h^m) = z'(t) - p'(t) + O(h^m)$ and (4.18b) follows. Finally, (4.18c) is shown by using (4.18b), (4.29), and the Lipschitz condition for f in

$$\begin{aligned} p'(t) - \frac{M(t)}{t} p'(t) - f(t, p(t)) \\ = p'(t) - z'(t) + \frac{M(t)}{t} (p(t) - z(t)) - f(t, p(t)) + f(t, z(t)) \\ = O(h^m), \quad t \in [0, 1]. \qquad \square \end{aligned}$$

Under the previous assumptions we can also show that Newton's method converges quadratically when it is applied to compute the collocation solution p, provided that the starting approximation $p^{[0]}$ is chosen sufficiently close to $p_{\rm ref}$.

THEOREM 4.5. *Let all assumptions of Theorem* 4.4 *hold. Newton's method converges quadratically to the solution* $p \in K(p_{\rm ref}, \rho_0)$ *of* (4.1) *if the starting iterate* $p^{[0]}$ *is chosen in a ball* $K(p_{\rm ref}, \rho_1)$, $\rho_1 \le \rho_0$, *provided that* ρ_0, ρ_1, *and the stepsize* h *are sufficiently small.*

Proof. The proof is analogous to that of [17, Theorem 3.7], taking into account the modifications made earlier in the proof of Theorem 4.4.

We write[6]

$$DF(q) = DF(p_{\rm ref})(I + DF^{-1}(p_{\rm ref})(DF(q) - DF(p_{\rm ref})))$$

for $q \in K(p_{\rm ref}, \rho_0)$ and use the bound for $DF^{-1}(p_{\rm ref})$, the Lipschitz condition for DF, and the Banach lemma to show that $DF^{-1}(q)$ is bounded if ρ_0 is sufficiently small,

$$\|DF^{-1}(q)\|_{[0,1]} \le K_{\rho_0}, \tag{4.30}$$

where K_{ρ_0} is a constant depending on ρ_0. Furthermore, let $p^{[0]} \in K(p_{\rm ref}, \rho_1)$; then

$$\begin{aligned} p^{[1]} - p^{[0]} &= -DF^{-1}(p^{[0]}) F(p^{[0]}) \\ &= -DF^{-1}(p^{[0]}) F(p_{\rm ref}) + DF^{-1}(p^{[0]}) (\widehat{DF}(p_{\rm ref}, p^{[0]})(p_{\rm ref} - p^{[0]})) \end{aligned}$$

[6] I is the identical mapping on the space of operators mapping $B_1 \to B_2$, that is, $I : DF(p_{\rm ref}) \mapsto DF(p_{\rm ref})$.

holds, with $\widehat{DF}(p_1, p_2)$ specified in Theorem 4.4. Using the Lipschitz condition for DF, we obtain

$$\begin{aligned}
&\|DF^{-1}(p^{[0]})\widehat{DF}(p_{\mathrm{ref}}, p^{[0]})(p_{\mathrm{ref}} - p^{[0]})\|_{[0,1]} \\
&\quad = \|p_{\mathrm{ref}} - p^{[0]} + DF^{-1}(p^{[0]})(\widehat{DF}(p_{\mathrm{ref}}, p^{[0]}) - DF(p^{[0]}))(p_{\mathrm{ref}} - p^{[0]})\|_{[0,1]} \\
&\quad \le \left(1 + \frac{L\rho_1}{2} K_{\rho_0}\right) \rho_1 =: C\rho_1.
\end{aligned}$$

Finally, we conclude

$$\|p^{[1]} - p^{[0]}\|_{[0,1]} \le K_{\rho_0} O(h^m) + C\rho_1.$$

Consider a ball $K(p^{[0]}, r)$. For a sufficiently small ρ_1 it is possible to choose the radius $r \le \rho_0$ in such a way that $K(p^{[0]}, r) \subseteq K(p_{\mathrm{ref}}, \rho_0)$. Moreover, let

$$\|DF^{-1}(p^{[0]})(DF(q_1) - DF(q_2))\|_{[0,1]} \le \omega(\|q_1 - q_2\|_{[0,1]}) \quad \forall\, q_1, q_2 \in K(p^{[0]}, r),$$

and choose r such that the condition $\omega(r) = 2K_{\rho_0} L r \le 1/2$ holds. Consequently,

$$\|p^{[1]} - p^{[0]}\|_{[0,1]} \le K_{\rho_0} O(h^m) + C\rho_1 \le (1 - 2\omega(r))r,$$

provided that ρ_1 and h are sufficiently small; cf. [16, formula (6c)]. This implies that the assumptions of [16, Theorem 1] are satisfied and the quadratic convergence of Newton's method in $K(p^{[0]}, r)$ follows. ∎

5. The error estimate. In this section, we analyze an error estimate based on the defect correction principle for the numerical solution p on the collocation grid Δ^m. For reasons explained in section 1, it is sufficient for practical purposes to consider equidistant collocation (cf. (2.2)), where we choose m even. However, the argument is valid on any collocation grid with $t_{i,m} < t_{i,m+1},\ i = 0, \ldots, N-1$.

Our estimate was introduced in [5], where it was shown to be asymptotically correct for regular problems. The numerical solution p obtained by collocation is used to define a "neighboring problem" to (1.1). The original and neighboring problems are solved by the backward Euler method at the points $t_{i,j},\ i = 0, \ldots, N-1,\ j = 1, \ldots, m+1$. This yields the grid vectors[7] $\xi_{i,j}$ and $\pi_{i,j}$ as the solutions of the following schemes, subject to boundary conditions (1.1b) and (1.1c):

$$\frac{\xi_{i,j} - \xi_{i,j-1}}{t_{i,j} - t_{i,j-1}} = \frac{M(t_{i,j})}{t_{i,j}} \xi_{i,j} + f(t_{i,j}, \xi_{i,j}), \quad \text{and} \tag{5.1a}$$

$$\frac{\pi_{i,j} - \pi_{i,j-1}}{t_{i,j} - t_{i,j-1}} = \frac{M(t_{i,j})}{t_{i,j}} \pi_{i,j} + f(t_{i,j}, \pi_{i,j}) + \bar{d}_{i,j}, \tag{5.1b}$$

where $\bar{d}_{i,j}$ is a defect term defined by

$$\bar{d}_{i,j} := \frac{p(t_{i,j}) - p(t_{i,j-1})}{t_{i,j} - t_{i,j-1}} - \sum_{k=1}^{m+1} \alpha_{j,k} \left(\frac{M(t_{i,k})}{t_{i,k}} p(t_{i,k}) + f(t_{i,k}, p(t_{i,k})) \right). \tag{5.2}$$

Here, the coefficients $\alpha_{j,k}$ are chosen in such a way that the quadrature rules given by

$$\frac{1}{t_{i,j} - t_{i,j-1}} \int_{t_{i,j-1}}^{t_{i,j}} \varphi(\tau)\, d\tau \approx \sum_{k=1}^{m+1} \alpha_{j,k} \varphi(t_{i,k})$$

[7] Here and in Theorem 5.1, we assume throughout $i = 0, \ldots, N-1,\ j = 1, \ldots, m+1$.

have precision $m+1$.

In the next theorem, we show that the difference $\xi_{\Delta^m} - \pi_{\Delta^m}$ is an asymptotically correct estimate for the global error of the collocation solution, $R_{\Delta^m}(z) - R_{\Delta^m}(p)$.

THEOREM 5.1. *Assume that the singular boundary value problem* (1.1) *has an isolated (sufficiently smooth*[8]*) solution z. Then, provided that h is sufficiently small, the following estimate holds:*

$$\|(R_{\Delta^m}(z) - R_{\Delta^m}(p)) - (\xi_{\Delta^m} - \pi_{\Delta^m})\|_{\Delta^m} = O(|\ln(h)|^{n_0-1} h^{m+1}), \tag{5.3}$$

with n_0 specified in Lemma 4.1.

Proof. The general idea of the proof is similar to that for regular problems. In particular, the smooth nonlinear part in the right-hand side of (1.1a) can be treated analogously. Therefore, we give a general outline of the proof here and discuss those aspects which are crucial for the singular case. For further technical details we refer the reader to [5].

Let

$$\varepsilon_{\Delta^m} := \xi_{\Delta^m} - R_{\Delta^m}(z), \quad \bar{\varepsilon}_{\Delta^m} := \pi_{\Delta^m} - R_{\Delta^m}(p); \tag{5.4}$$

then the quantity to be estimated is

$$\tilde{\varepsilon}_{\Delta^m} := (R_{\Delta^m}(p) - R_{\Delta^m}(z)) - (\pi_{\Delta^m} - \xi_{\Delta^m}) = \varepsilon_{\Delta^m} - \bar{\varepsilon}_{\Delta^m}. \tag{5.5}$$

Here, ε_{Δ^m}, the error of the backward Euler scheme applied to the original problem, satisfies

$$\begin{aligned} \frac{\varepsilon_{i,j} - \varepsilon_{i,j-1}}{t_{i,j} - t_{i,j-1}} &= \frac{M(t_{i,j})}{t_{i,j}} \xi_{i,j} + f(t_{i,j}, \xi_{i,j}) - \frac{z(t_{i,j}) - z(t_{i,j-1})}{t_{i,j} - t_{i,j-1}} \\ &= \frac{M(t_{i,j})}{t_{i,j}} \xi_{i,j} + f(t_{i,j}, \xi_{i,j}) \\ &\quad - \sum_{k=1}^{m+1} \alpha_{j,k} \left(\frac{M(t_{i,k})}{t_{i,k}} z(t_{i,k}) + f(t_{i,k}, z(t_{i,k})) \right) + O(h^{m+1}), \end{aligned} \tag{5.6}$$

since the $\alpha_{j,k}$ define quadrature rules of precision $O(h^{m+1})$. Moreover, $\bar{\varepsilon}_{\Delta^m}$ satisfies

$$\begin{aligned} \frac{\bar{\varepsilon}_{i,j} - \bar{\varepsilon}_{i,j-1}}{t_{i,j} - t_{i,j-1}} &= \frac{M(t_{i,j})}{t_{i,j}} \pi_{i,j} + f(t_{i,j}, \pi_{i,j}) + \bar{d}_{i,j} - \frac{p(t_{i,j}) - p(t_{i,j-1})}{t_{i,j} - t_{i,j-1}} \\ &= \frac{M(t_{i,j})}{t_{i,j}} \pi_{i,j} + f(t_{i,j}, \pi_{i,j}) \\ &\quad - \sum_{k=1}^{m+1} \alpha_{j,k} \left(\frac{M(t_{i,k})}{t_{i,k}} p(t_{i,k}) + f(t_{i,k}, p(t_{i,k})) \right). \end{aligned} \tag{5.7}$$

Both (5.6) and (5.7) hold for $i = 0, \ldots, N-1$, $j = 1, \ldots, m+1$, and ε_{Δ^m} as well as $\bar{\varepsilon}_{\Delta^m}$ satisfy homogeneous boundary conditions.

In order to proceed, we use Taylor's theorem to conclude that

$$\begin{aligned} f(t_{i,j}, \xi_{i,j}) - f(t_{i,j}, z(t_{i,j})) &= \int_0^1 D_2 f(t_{i,j}, z(t_{i,j}) + \tau(\xi_{i,j} - z(t_{i,j}))) \, d\tau \cdot \varepsilon_{i,j} \\ &=: A(t_{i,j}) \varepsilon_{i,j}, \end{aligned} \tag{5.8}$$

[8] In fact, we require $z \in C^{m+2}[0,1]$.

and analogously

$$f(t_{i,j},\pi_{i,j}) - f(t_{i,j},p(t_{i,j})) =: \bar{A}(t_{i,j})\bar{\varepsilon}_{i,j}. \tag{5.9}$$

Next, we note that due to (4.18c),

$$\begin{aligned}\bar{d}_{i,j} &= \frac{p(t_{i,j}) - p(t_{i,j-1})}{t_{i,j} - t_{i,j-1}} - \sum_{k=1}^{m+1} \alpha_{j,k} \left(\frac{M(t_{i,k})}{t_{i,k}} p(t_{i,k}) + f(t_{i,k}, p(t_{i,k})) \right) \\ &= \frac{1}{t_{i,j} - t_{i,j-1}} \int_{t_{i,j-1}}^{t_{i,j}} p'(\tau)\, d\tau - \sum_{k=1}^{m+1} \alpha_{j,k} p'(t_{i,k}) \\ &\quad + \alpha_{j,m+1} \left(p'(t_{i,m+1}) - \frac{M(t_{i,m+1})}{t_{i,m+1}} p(t_{i,m+1}) - f(t_{i,m+1}, p(t_{i,m+1})) \right) \\ &= O(h^m). \end{aligned} \tag{5.10}$$

From this we conclude that $\xi_{i,j} = \pi_{i,j} + O(h^m)$ using the following argument.

The backward Euler schemes (5.1a) and (5.1b) can be written as collocation methods with $m = 1$ and the collocating condition posed at the right endpoint of each interval $[t_{i,j-1}, t_{i,j}]$. Thus, we discuss the collocation solutions $\xi(t)$, $\pi(t)$ of two singular boundary value problems whose right-hand sides differ by a term $O(h^m)$. This term can be assumed to be smooth if a suitable interpolant g of $\bar{d}_{i,j}$ is used. More precisely, $\xi(t)$ is an approximation to the solution z of (1.1), and $\pi(t)$ is an approximation to the solution of

$$\begin{aligned} & z'_{\mathrm{def}}(t) = \frac{M(t)}{t} z_{\mathrm{def}}(t) + f(t, z_{\mathrm{def}}(t)) + g(t), \quad t \in (0,1], \\ & B_a z_{\mathrm{def}}(0) + B_b z_{\mathrm{def}}(1) = \beta, \\ & M(0) z_{\mathrm{def}}(0) = 0. \end{aligned}$$

For (1.1), we make the assumption that the analytical problem is stable in the sense that

$$\|z - z_{\mathrm{def}}\|_{[0,1]} \le \text{const.}\, \|g\|_{[0,1]} = O(h^m)$$

holds. For results on this type of stability analysis, see [27].

As in section 4 we can prove that (locally) unique solutions $\xi(t)$ and $\pi(t)$ of (5.1a) and (5.1b) exist in a neighborhood of z and z_{def}, respectively.

Subtracting (5.1b) from (5.1a) and using Taylor expansion about $\pi(t_{i,j})$, we can show that $q(t) := \xi(t) - \pi(t)$ satisfies the linear scheme

$$\frac{q(t_{i,j}) - q(t_{i,j-1})}{t_{i,j} - t_{i,j-1}} = \frac{M(t_{i,j}) + t_{i,j} B(t_{i,j})}{t_{i,j}} q(t_{i,j}) + O(h^m),$$

with a suitable, bounded matrix B and homogeneous boundary conditions. Since this is equivalent to a collocation scheme, we may use [10, Theorem 4.4] with $\gamma = \delta = 0$ (or alternatively a combination of stability results from section 4) to conclude that $\|R_{\Delta^m}(q)\|_{\Delta^m} = \|\xi_{\Delta^m} - \pi_{\Delta^m}\|_{\Delta^m} = O(h^m)$.

Since $\varepsilon_{i,j} = O(h)$ and $\bar{\varepsilon}_{i,j} = O(h)$, we may finally write (see [5])

$$\bar{A}(t_{i,j})\bar{\varepsilon}_{i,j} = A(t_{i,j})\bar{\varepsilon}_{i,j} + (\bar{A}(t_{i,j}) - A(t_{i,j}))\bar{\varepsilon}_{i,j} = A(t_{i,j})\bar{\varepsilon}_{i,j} + O(h^{m+1}).$$

Now we use (5.8), (5.9) to rewrite (5.6), (5.7) and obtain

$$\begin{aligned}\frac{\varepsilon_{i,j}-\varepsilon_{i,j-1}}{t_{i,j}-t_{i,j-1}} &= \frac{M(t_{i,j})}{t_{i,j}}\varepsilon_{i,j} + A(t_{i,j})\varepsilon_{i,j} + \frac{M(t_{i,j})}{t_{i,j}}z(t_{i,j}) + f(t_{i,j},z(t_{i,j})) \\ &\quad - \sum_{k=1}^{m+1}\alpha_{j,k}\left(\frac{M(t_{i,k})}{t_{i,k}}z(t_{i,k}) + f(t_{i,k},z(t_{i,k}))\right) + O(h^{m+1}) \qquad (5.11)\end{aligned}$$

and

$$\begin{aligned}\frac{\bar\varepsilon_{i,j}-\bar\varepsilon_{i,j-1}}{t_{i,j}-t_{i,j-1}} &= \frac{M(t_{i,j})}{t_{i,j}}\bar\varepsilon_{i,j} + A(t_{i,j})\bar\varepsilon_{i,j} + \frac{M(t_{i,j})}{t_{i,j}}p(t_{i,j}) + f(t_{i,j},p(t_{i,j})) \\ &\quad - \sum_{k=1}^{m+1}\alpha_{j,k}\left(\frac{M(t_{i,k})}{t_{i,k}}p(t_{i,k}) + f(t_{i,k},p(t_{i,k}))\right) + O(h^{m+1}). \qquad (5.12)\end{aligned}$$

Systems (5.11) and (5.12) are a pair of "parallel" backward Euler schemes, with related inhomogeneous terms. Let us use the shorthand notation $\phi(t) := f(t,p(t)) - f(t,z(t))$. It can be shown that for the difference in the smooth parts of the inhomogeneous terms, the estimate

$$\begin{aligned}|\phi(t_{i,j}) - \sum_{k=1}^{m+1}\alpha_{j,k}\phi(t_{i,k})| &\le \text{const.}\, h_i\|\phi'\|_{J_i} \\ &\le \text{const.}\, h_i(\|z-p\|_{J_i} + \|z'-p'\|_{J_i}) \le O(h^{m+1}) \qquad (5.13)\end{aligned}$$

holds. To see this we use Taylor expansion of $\phi(t_{i,k})$ about $t_{i,j}$ and the fact that $\sum_{k=1}^{m+1}\alpha_{j,k} = 1\ \forall j$; see [5]. The estimate finally follows from Theorem 4.4.

In the next step, we derive a representation for the difference in the singular terms occurring in the inhomogeneous parts of schemes (5.11) and (5.12). With $\epsilon(t) := z(t) - p(t)$ and with $\sigma := t_{i,j} + \tau(t_{i,k} - t_{i,j})$, we rewrite

$$\begin{aligned}&\frac{M(t_{i,j})}{t_{i,j}}\epsilon(t_{i,j}) - \sum_{k=1}^{m+1}\alpha_{j,k}\frac{M(t_{i,k})}{t_{i,k}}\epsilon(t_{i,k}) \qquad (5.14)\\ &\quad= \frac{M(t_{i,j})}{t_{i,j}}\epsilon(t_{i,j}) - \sum_{k=1}^{m+1}\alpha_{j,k}\left(\frac{M(t_{i,j})}{t_{i,j}}\epsilon(t_{i,j})\right.\\ &\qquad\left. + \int_0^1 \frac{d}{d\sigma}\left(\frac{M(\sigma)}{\sigma}\epsilon(\sigma)\right)\,d\tau(t_{i,k}-t_{i,j})\right)\\ &\quad= \sum_{k=1}^{m+1}\alpha_{j,k}(t_{i,j}-t_{i,k})\int_0^1\left(\frac{M(0)}{\sigma}\epsilon'(\sigma)\right.\\ &\qquad\left. - \frac{M(0)}{\sigma^2}\epsilon(\sigma) + C'(\sigma)\epsilon(\sigma) + C(\sigma)\epsilon'(\sigma)\right)d\tau\\ &\quad= \frac{M(0)}{t_{i,j}}O(h^{m+1}) + O(h^{m+1})\end{aligned}$$

on noting that

$$\frac{1}{\sigma} \le \frac{m}{t_{i,j}}, \quad k = 1,\dots,m+1,\ j = 1,\dots,m,\ \tau\in[0,1],$$

and using the results of Theorem 4.4.

Altogether, we have shown that the error of the error estimate $\tilde{\varepsilon}_{\Delta^m}$ (cf. (5.5)) satisfies a linear Euler difference scheme

$$\frac{\tilde{\varepsilon}_{i,j}-\tilde{\varepsilon}_{i,j-1}}{t_{i,j}-t_{i,j-1}} = \frac{M(t_{i,j})}{t_{i,j}}\tilde{\varepsilon}_{i,j} + A(t_{i,j})\tilde{\varepsilon}_{i,j} + \frac{M(0)}{t_{i,j}}O(h^{m+1}) + O(h^{m+1}) \ \forall\, i,j, \tag{5.15a}$$

$$B_a\tilde{\varepsilon}_{0,0} + B_b\tilde{\varepsilon}_{N-1,m+1} = 0, \tag{5.15b}$$

$$M(0)\tilde{\varepsilon}_{0,0} = 0. \tag{5.15c}$$

This scheme can also be interpreted as a collocation scheme with $m = 1$ where the only collocation point is the right endpoint of every collocation interval. To estimate the solution of (5.15) we use a representation according to (4.8) for $\tilde{\varepsilon}_{\Delta^m}$. Then we apply Theorem 4.2 to derive bounds for the quantities occurring in (4.8), and we conclude that altogether the estimate (5.3) holds for the solution of (5.15). ∎

Remark. Obviously, the arguments used to prove the last theorem are valid for any choice of collocation nodes. The only necessary restriction is $t_{i,m+1} > t_{i,m}$. However, if we consider superconvergent schemes, the error estimate is no longer asymptotically correct, because the basic collocation solution has a higher convergence order in that case. Therefore we restrict ourselves to an even number of equidistant collocation points. This restriction is not severe, since in the case of singular problems, the highest convergence order that can generally be expected at the mesh points τ_i is $O(|\ln(h)|^{n_0-1}h^{m+1})$; see [10].

Finally, we would like to mention an alternative variant of our error estimate closely related to the so-called version B of defect correction according to Stetter [24]. If instead of (5.1) we solve

$$\frac{\zeta_{i,j}-\zeta_{i,j-1}}{t_{i,j}-t_{i,j-1}} = \frac{M(t_{i,j})}{t_{i,j}}\zeta_{i,j} + D(t_{i,j})\zeta_{i,j} - \bar{d}_{i,j}, \tag{5.16}$$

where

$$D(t_{i,j}) := D_2 f(t_{i,j}, p(t_{i,j})),$$

then $\zeta_{i,j}$ is an asymptotically correct error estimate. To see this, we note that the difference between this error estimate and the estimate analyzed earlier in this paper,

$$x_{i,j} := (\xi_{i,j} - \pi_{i,j}) - \zeta_{i,j},$$

satisfies

$$\begin{aligned}\frac{x_{i,j}-x_{i,j-1}}{t_{i,j}-t_{i,j-1}} &= \frac{M(t_{i,j})}{t_{i,j}}x_{i,j} + f(t_{i,j},\xi_{i,j}) - f(t_{i,j},\pi_{i,j}) - D(t_{i,j})\zeta_{i,j}\\ &= \frac{M(t_{i,j})}{t_{i,j}}x_{i,j} + O(h^{m+1}).\end{aligned}$$

Consequently, the error of this error estimate has a bound analogous to (5.3). Note that for linear problems, this alternative error estimate coincides with the variant discussed earlier in the paper. For nonlinear problems, the practical usability and numerical stability of the new estimate still has to be carefully assessed.

6. Numerical examples. To illustrate the theory, we first consider the following nonlinear problem:

$$z'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} z(t) + t \begin{pmatrix} 0 \\ -\frac{2(t^2+2)+8}{(t^2+2)^2} z_1^2(t) + \frac{8t^2}{(t^2+2)^2} z_1^3(t) \end{pmatrix}, \tag{6.1a}$$

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} z(0) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} z(1) = \begin{pmatrix} 0 \\ 1/\ln(3) \end{pmatrix}. \tag{6.1b}$$

Its exact solution is

$$z(t) = \left(\frac{1}{\ln(t^2+2)}, -\frac{2t^2}{(t^2+2)\ln^2(t^2+2)} \right)^T.$$

The computations were carried out with the subroutines from our MATLAB code `sbvp` (cf. [4]) on fixed, equidistant grids. For the purpose of determining the empirical convergence orders the mesh adaptation strategy was disabled. The tests were performed in IEEE double precision with EPS $\approx 1.11 \cdot 10^{-16}$. In Table 6.1, we give the exact global errors $\mathrm{err}_{\mathrm{coll}}$ of the collocation solutions for the respective mesh width h and the convergence orders p_{coll} computed from the errors for two consecutive stepsizes. Moreover, the errors of the error estimate with respect to the exact global errors, $\mathrm{err}_{\mathrm{est}}$, are recorded, together with associated empirical convergence orders p_{est}. In accordance with the theoretical results from sections 4 and 5, convergence orders $O(h^4)$ for collocation and $O(h^5)$ for the error estimate are observed. This illustrates the asymptotical correctness of the error estimate analyzed in this paper. Test runs given in [4] demonstrate that this error estimate can be used as a dependable basis for a mesh adaptation algorithm, providing an efficient, high precision numerical solver.

TABLE 6.1
Convergence orders of collocation and error estimate for (6.1).

h	$\mathrm{err}_{\mathrm{coll}}$	p_{coll}	$\mathrm{err}_{\mathrm{est}}$	p_{est}
2^{-2}	1.5763e−04		2.2232e−05	
2^{-3}	9.5865e−06	4.04	6.5978e−07	5.07
2^{-4}	5.9574e−07	4.01	1.7873e−08	5.21
2^{-5}	3.7189e−08	4.00	5.1077e−10	5.13
2^{-6}	2.3237e−09	4.00	1.5205e−11	5.07
2^{-7}	1.4522e−10	4.00	4.6274e−13	5.04
2^{-8}	9.0772e−12	4.00	1.4655e−14	4.98

Finally, we demonstrate the favorable performance of our error estimate for a practically relevant example from applications. The following boundary value problem is a model from the theory of shallow spherical shells; see [22], [25]. The transformation of the original two-dimensional system of second order to the first order form yields

$$z'(t) = \frac{1}{t}\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} z(t) + t \begin{pmatrix} 0 \\ 0 \\ z_2(t)(-\mu^2 + z_1(t)) - 2\gamma \\ z_1(t)(\mu^2 - \frac{1}{2} z_1(t)) \end{pmatrix}, \tag{6.2}$$

where the eigenvalues of $M(0)$ are $\lambda = 0,\ 0,\ -2,\ -2$. The boundary conditions read $z_3(0) = z_4(0) = z_1(1) = 0,\ z_4(1) + 2/3 z_2(1) = 0$, and the parameters are chosen as $\mu = 9,\ \gamma = 6000$. We solve (6.2) using our code `sbvp` [4] equipped with the

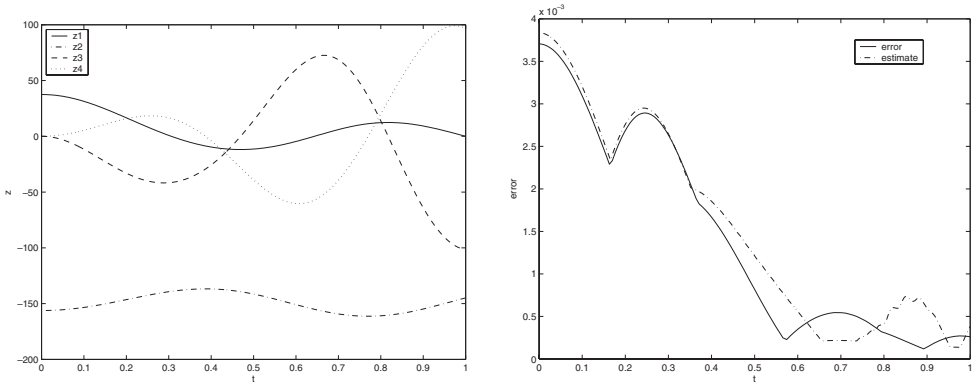


FIG. 6.1. *Solution, global error, and error estimate for* (6.2).

error estimate from section 5 and the adaptive mesh selection routine. The numerical solution satisfies a mixed tolerance requirement with absolute and relative tolerance equal to 10^{-4} at a mesh containing 124 mesh points, where the variation in the mesh width is just below 2. In Figure 6.1 four components of the numerical solution are given, and the estimate of the global error on the final mesh is compared with the error of the collocation solution. In order to calculate the error of the collocation solution we used a reference solution computed with tolerances $5 \cdot 10^{-6}$. The maximum of the error estimate is 0.0038367, and the maximum of the error with respect to the reference solution is 0.003706. For most of the integration interval, the estimate slightly overestimates the "true" error.

Acknowledgment. We wish to thank the referees for their valuable suggestions, in particular for pointing out the alternative variant (5.16).

REFERENCES

[1] U. ASCHER, J. CHRISTIANSEN, AND R. RUSSELL, *A collocation solver for mixed order systems of boundary values problems*, Math. Comp., 33 (1978), pp. 659–679.

[2] U. ASCHER, J. CHRISTIANSEN, AND R. RUSSELL, *Collocation software for boundary value ODEs*, ACM Trans. Math. Software, 7 (1981), pp. 209–222.

[3] U. ASCHER, R. MATTHEIJ, AND R. RUSSELL, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[4] W. AUZINGER, G. KNEISL, O. KOCH, AND E. B. WEINMÜLLER, *A collocation code for boundary value problems in ordinary differential equations*, Numer. Algorithms, 33 (2003), pp. 27–39.

[5] W. AUZINGER, O. KOCH, AND E. B. WEINMÜLLER, *Efficient collocation schemes for singular boundary value problems*, Numer. Algorithms, 31 (2002), pp. 5–25.

[6] P. BAILEY, W. EVERITT, AND A. ZETTL, *Computing eigenvalues of singular Sturm-Liouville problems*, Results Math., 20 (1991), pp. 391–423.

[7] C. J. BUDD, O. KOCH, AND E. B. WEINMÜLLER, *Self-similar Blow-up in Nonlinear PDEs*, Technical report, AURORA TR2004-15.

[8] C. DE BOOR AND B. SWARTZ, *Collocation at Gaussian points*, SIAM J. Numer. Anal., 10 (1973), pp. 582–606.

[9] F. R. DE HOOG AND R. WEISS, *Difference methods for boundary value problems with a singularity of the first kind*, SIAM J. Numer. Anal., 13 (1976), pp. 775–813.

[10] F. R. DE HOOG AND R. WEISS, *Collocation methods for singular boundary value problems*, SIAM J. Numer. Anal., 15 (1978), pp. 198–217.

[11] F. DE HOOG AND R. WEISS, *The application of Runge-Kutta schemes to singular initial value problems*, Math. Comp., 44 (1985), pp. 93–103.

[12] R. Frank and C. Überhuber, *Iterated defect correction for differential equations, part* I: *Theoretical results*, Computing, 20 (1978), pp. 207–228.
[13] F. B. Hildebrand, *Introduction to Numerical Analysis*, 2nd ed., McGraw-Hill, New York, 1974.
[14] V. Hlavacek, M. Marek, and M. Kubicek, *Modelling of chemical reactors*, Chem. Engrg. Sci., 23 (1968), pp. 1083–1097.
[15] T. Kapitula, *Existence and stability of singular heteroclinic orbits for the Ginzburg-Landau equation*, Nonlinearity, 9 (1996), pp. 669–685.
[16] H. Keller, *Newton's method under mild differentiability conditions*, J. Comput. System Sci., 4 (1970), pp. 15–28.
[17] H. Keller, *Approximation methods for nonlinear problems with application to two-point boundary value problems*, Math. Comp., 29 (1975), pp. 464–474.
[18] O. Koch and E. B. Weinmüller, *Analytical and numerical treatment of a singular initial value problem in avalanche modeling*, Appl. Math. Comput., 148 (2003), pp. 561–570.
[19] D. M. McClung and A. I. Mears, *Dry-flowing avalanche run-up and run-out*, J. Glaciol., 41 (1995), pp. 359–369.
[20] H. Meissner and P. Tholfsen, *Cylindrically symmetric solutions of the Ginzburg-Landau equation*, Phys. Rev., 169 (1968), pp. 413–416.
[21] G. Moore, *Geometric methods for computing invariant manifolds*, Appl. Numer. Math., 17 (1995), pp. 319–331.
[22] P. Rentrop, *Eine Taylorreihenmethode zur numerischen Lösung von Zwei-Punkt Randwertproblemen mit Anwendung auf singuläre Probleme der nichtlinearen Schalentheorie*, Technical report, TUM-MATH-7733, Technische Universität München, 1977.
[23] R. D. Russell and L. F. Shampine, *Numerical methods for singular boundary value problems*, SIAM J. Numer. Anal., 12 (1975), pp. 13–36.
[24] H. J. Stetter, *The defect correction principle and discretization methods*, Numer. Math., 29 (1978), pp. 425–443.
[25] H. Weinitschke, *On the stability problem for shallow spherical shells*, J. Math. and Phys., 38 (1959), pp. 209–231.
[26] E. B. Weinmüller, *Collocation for singular boundary value problems of second order*, SIAM J. Numer. Anal., 23 (1986), pp. 1062–1095.
[27] E. B. Weinmüller, *Stability of singular boundary value problems and their discretization by finite differences*, SIAM J. Numer. Anal., 26 (1989), pp. 180–213.
[28] C.-Y. Yeh, A.-B. Chen, D. Nicholson, and W. Butler, *Full-potential Korringa-Kohn-Rostoker band theory applied to the Mathieu potential*, Phys. Rev. B, 42 (1990), pp. 10976–10982.
[29] P. Zadunaisky, *On the estimation of errors propagated in the numerical integration of ODEs*, Numer. Math., 27 (1976), pp. 21–39.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2387–2407

A FIRST-ORDER SYSTEM LEAST SQUARES FINITE ELEMENT METHOD FOR THE SHALLOW WATER EQUATIONS*

GERHARD STARKE†

Abstract. A least squares finite element method for the first-order system of the shallow water equations is proposed and studied. The method combines a characteristic-based time discretization with a least squares finite element approach which approximates the water level and the velocity field in H^1 and $H(\mathrm{div})$, respectively. The linearized least squares functional is shown to be elliptic, uniformly as the time-step size τ approaches zero, with respect to a suitably weighted norm. Moreover, Lipschitz continuity of the Fréchet derivative is shown with respect to this norm. This implies that the local evaluation of the nonlinear least squares functional constitutes an a posteriori error estimator on which an adaptive refinement technique may be based. The efficiency of such an adaptive finite element approach is tested numerically for a test problem involving the surface flow in a widening channel leading to a recirculating velocity field.

Key words. least squares finite element method, first-order system, shallow water equations, method of characteristics

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/S0036142903438124

1. Introduction.

The numerical simulation of surface flow by the system of shallow water equations is of interest for the prediction of, for example, water levels in coastal regions or flood zones. The shallow water equations which are obtained from depth-averaging the Navier–Stokes equations for fluid flow simplify the flow model to a two-dimensional situation. This reduction is permissible under certain assumptions including the fact that the horizontal dimension of the domain is much larger than the vertical one. A lot of effort has been dedicated to the development of accurate and efficient numerical methods for the approximate solution of shallow water equations in recent years. Time discretization based on characteristics has become a standard approach to the numerical simulation of surface flows in the meantime; see, e.g., [18, 9, 14, 17, 2, 12, 8, 10]. The discretization in time based on a coordinate transformation along the characteristics is also used in the analytical study of a numerical approach to the shallow water equations in [7].

Most often, the finite element discretization of the shallow water system uses piecewise quadratic approximations for the velocity components and piecewise linear elements for the elevation (see, e.g., [11]). This combination of finite element spaces constitutes the well-known Taylor–Hood elements and is known to be stable for the mixed variational formulation of the Stokes problem. If viscosity is neglected in the shallow water model, the velocity field need only be $H(\mathrm{div})$-conforming and Raviart–Thomas elements may be used for its approximation. The lowest-order Raviart–Thomas spaces for the velocity field combined with piecewise constants for the elevation are used in [12, 8] for this purpose. The resulting mixed finite element approach for the two-dimensional shallow water equations is contained as a special case in the more general Quasi-three-dimensional multilayer approach treated in [12, 8].

*Received by the editors November 20, 2003; accepted for publication (in revised form) May 24, 2004; published electronically March 31, 2005. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) under grant STA 402/7-1.

http://www.siam.org/journals/sinum/42-6/43812.html

†Institut für Angewandte Mathematik, Universität Hannover, Welfengarten 1, 30167 Hannover, Germany (starke@ifam.uni-hannover.de).

The mixed finite element method by Raviart and Thomas was also employed in [10] where the emphasis was on the incorporation of transparent boundary conditions into the weak formulation.

In this paper, we present a least squares finite element method for the system of shallow water equations. This method does not require the finite element spaces for velocity and elevation to satisfy a compatibility condition. Raviart–Thomas spaces (of arbitrary degree) for the velocity field may be combined with H^1-conforming elements (of arbitrary degree) for the elevation, for example. The elementwise evaluation of the least squares functional constitutes an a posteriori error estimator at no additional cost. This a posteriori error estimator gives rise to adaptive refinement strategies which dramatically increase the accuracy and efficiency of numerical methods in many practical situations (see [3] for a study of such techniques in the least squares finite element context). The use of the local evaluation of nonlinear least squares functionals as an a posteriori error estimator was previously studied in connection to subsurface flow problems in [16, 15]. For surface flow in domains with complicated boundaries, like coastal regions, adaptivity is mandatory to achieve satisfactory results at reasonable computational cost. Adaptive mesh refinement techniques based on error indicators which are conceptually different from the ones derived in this paper are proposed and tested for shallow water flow in [11].

For the above reasons, among others, least squares finite element methods have become increasingly popular in recent years for a number of different application problems; see [5] for an overview. Several least squares formulations for the Navier–Stokes equations have been studied in [4, 6] where the partial derivatives of the velocity field are introduced as additional variables.

In section 2 we introduce the system of shallow water equations and describe the time discretization using a coordinate transformation along the characteristics. The least squares variational formulation in the space of the resulting system is investigated in section 3. In particular, ellipticity of the linearized least squares functional and Lipschitz continuity of the Fréchet derivative is shown with respect to a properly weighted norm. Section 4 treats the discretization by appropriate finite element spaces including adaptive refinement strategies and the solution of the resulting nonlinear least squares problems by Gauss–Newton iterations. Finally, computational results for a test problem modelling recirculating shallow water flow in a channel are presented in section 5.

2. The shallow water equations and method of characteristics. The process variables to be determined are H, the total elevation above the sea bottom, and $\mathbf{u}$, the depth-averaged horizontal velocities. The system of the shallow water equations is then given by

$$\begin{aligned} \partial_t H + \operatorname{div}(H\mathbf{u}) &= 0, \\ \partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + g\nabla(H - H_b) + c_f \frac{\mathbf{u}|\mathbf{u}|}{H} + \mathbf{F} &= \mathbf{0}, \end{aligned} \tag{2.1}$$

to be satisfied in a region $\Omega \subset \mathbb{R}^2$, where H_b denotes the bathymetric depth under the reference plane, g is the acceleration due to gravity, c_f is the so-called Chezy coefficient for the bottom friction term, and $\mathbf{F}$ comprises additional body and surface forces (cf. [7]). Note that we have neglected the fluid viscosity in (2.1), an assumption that is often admissible in practical situations (see, e.g., [8, 12, 10]).

In order to properly take into account the hyperbolic nature of system (2.1), characteristic-based time discretizations are widely used for the shallow water equa-

tions. To this end, the system (2.1) is reformulated with respect to transformed time-space coordinates. The time coordinate is replaced by the local direction of the characteristics while the spatial coordinates remain unchanged. The characteristic curves $\mathbf{X}(t;s,\mathbf{x})$ associated with the system (2.1) are given by

$$\frac{d}{dt}\mathbf{X}(t;s,\mathbf{x}) = \mathbf{u}(t,\mathbf{X}(t;s,\mathbf{x})) \quad \text{for } t>0,$$
$$\mathbf{X}(s;s,\mathbf{x}) = \mathbf{x},$$

where $\mathbf{X}(t;s,\mathbf{x})$ is the characteristic curve, parametrized by t, passing through the time-space point $(s,\mathbf{x})$. For the transformed process variables

$$\hat{H}(t,\mathbf{x}) = H(t,\mathbf{X}(t;s,\mathbf{x})) \quad \text{and} \quad \hat{\mathbf{u}}(t,\mathbf{x}) = \mathbf{u}(t,\mathbf{X}(t;s,\mathbf{x})),$$

we have

$$\partial_t \hat{H} = \partial_t H + \mathbf{u}\cdot\nabla H,$$
$$\partial_t \hat{\mathbf{u}} = \partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u}.$$

With respect to the transformed coordinates, (2.1) therefore turns into

$$\partial_t \hat{H} + \hat{H}\,\mathrm{div}\,\hat{\mathbf{u}} = 0,$$
$$\partial_t \hat{\mathbf{u}} + g\nabla(\hat{H}-H_b) + c_f\frac{\hat{\mathbf{u}}|\hat{\mathbf{u}}|}{\hat{H}} + \mathbf{F} = \mathbf{0}.$$

Using an implicit Euler time discretization we end up with a first-order system of the form

$$\begin{aligned} &\frac{H-\hat{H}^{\mathrm{old}}}{\tau} + H\,\mathrm{div}\,\mathbf{u} = 0 \quad \text{for all } \mathbf{x}\in\Omega,\\ &\frac{\mathbf{u}-\hat{\mathbf{u}}^{\mathrm{old}}}{\tau} + g\nabla(H-H_b) + c_f\frac{\mathbf{u}|\mathbf{u}|}{H} + \mathbf{F} = \mathbf{0} \quad \text{for all } \mathbf{x}\in\Omega \end{aligned} \tag{2.2}$$

for each time step. Here, τ denotes the time-step length, H and $\mathbf{u}$ stand for the unknown solutions at the current time t, while $\hat{H}^{\mathrm{old}}$ and $\hat{\mathbf{u}}^{\mathrm{old}}$ are the solutions at the previous time level $t-\tau$ evaluated at the point $\hat{\mathbf{x}}$ backwards along the characteristic. An implicit Euler scheme is usually sufficient for the approximation of the characteristics. In that case we obtain $\hat{\mathbf{x}} = \mathbf{x} - \tau\mathbf{u}$.

Initial conditions prescribed for H and $\mathbf{u}$ allow the time-stepping procedure to start from time $t=0$. Furthermore, we consider the boundary conditions

$$\begin{aligned} &H = H^D \quad \text{for all } \mathbf{x}\in\Gamma_D,\\ &\mathbf{u}\cdot\mathbf{n} = 0 \quad \text{for all } \mathbf{x}\in\Gamma_N, \end{aligned} \tag{2.3}$$

where Γ_D and Γ_N are boundary segments such that $\Gamma_D\cup\Gamma_N = \partial\Omega$. Physically, this means that at each boundary point either the elevation is prescribed or the normal velocity is set to zero.

3. Least squares formulation of the time-discretized first-order system. Our variational formulation for the solution of the first-order system (2.2) with boundary conditions (2.3) will be based on the spaces

$$H^1_{\Gamma_D}(\Omega) = \{\eta\in H^1(\Omega) : \eta = 0 \text{ on } \Gamma_D\},$$
$$H_{\Gamma_N}(\mathrm{div},\Omega) = \{\mathbf{v}\in H(\mathrm{div},\Omega) : \mathbf{v}\cdot\mathbf{n} = 0 \text{ on } \Gamma_N\}.$$

In addition to the standard norm in $L^2(\Omega)$ which we denote by $\|\cdot\|_{0,\Omega}$, we will also make use of the supremum norm $\|\cdot\|_{\infty,\Omega}$ on $L^\infty(\Omega)$ and the Sobolev space

$$W^{1,\infty}(\Omega) = \{\eta \in L^\infty(\Omega) : \nabla\eta \in L^\infty(\Omega)^2\}.$$

Multiplying the first-order system (2.2) with τ leads to

$$\mathcal{R}(H, \mathbf{u}; \hat{H}^{\mathrm{old}}, \hat{\mathbf{u}}^{\mathrm{old}}) = \begin{pmatrix} H - \hat{H}^{\mathrm{old}} + \tau H \operatorname{div} \mathbf{u} \\ \mathbf{u} - \hat{\mathbf{u}}^{\mathrm{old}} + \tau \left(g\nabla(H - H_b) + c_f \dfrac{\mathbf{u}|\mathbf{u}|}{H} + \mathbf{F} \right) \end{pmatrix} = \mathbf{0}. \tag{3.1}$$

Our aim is to solve (3.1) for $H = H^D + \hat{H}$ with $\hat{H} \in H^1_{\Gamma_D}(\Omega)$ and for $\mathbf{u} \in H_{\Gamma_N}(\operatorname{div}, \Omega)$. To this end, we consider the least squares minimization problem of finding $(\hat{H}, \mathbf{u}) \in H^1_{\Gamma_D}(\Omega) \times H_{\Gamma_N}(\operatorname{div}, \Omega)$ such that

$$\|\mathcal{R}(H^D + \hat{H}, \mathbf{u}; \hat{H}^{\mathrm{old}}, \hat{\mathbf{u}}^{\mathrm{old}})\|^2_{0,\Omega} \le \|\mathcal{R}(H^D + \eta, \mathbf{v}; \hat{H}^{\mathrm{old}}, \hat{\mathbf{u}}^{\mathrm{old}})\|^2_{0,\Omega} \tag{3.2}$$

holds for all $(\eta, \mathbf{v}) \in H^1_{\Gamma_D}(\Omega) \times H_{\Gamma_N}(\operatorname{div}, \Omega)$.

For the linearization, the Fréchet derivative of $\mathcal{R}$ with respect to $(H, \mathbf{u})$ is of interest. It is given by

$$\mathcal{J}(H, \mathbf{u})[\eta, \mathbf{v}] = \begin{pmatrix} \eta + \tau(H \operatorname{div} \mathbf{v} + \eta \operatorname{div} \mathbf{u}) \\ \mathbf{v} + \tau g \nabla\eta + \tau c_f \left(\dfrac{|\mathbf{u}|}{H}\mathbf{v} + \dfrac{\mathbf{u}}{|\mathbf{u}|H}(\mathbf{u} \cdot \mathbf{v}) \right) - \tau c_f \dfrac{\mathbf{u}|\mathbf{u}|}{H^2}\eta \end{pmatrix}. \tag{3.3}$$

For the well-posedness of the least squares minimization problem (3.2) we need to establish an equivalence of the type

$$\alpha |||(\eta, \mathbf{v})|||^2 \le \|\mathcal{J}(H, \mathbf{u})[\eta, \mathbf{v}]\|^2_{0,\Omega} \le \beta |||(\eta, \mathbf{v})|||^2$$

for all $(\eta, \mathbf{v}) \in H^1_{\Gamma_D}(\Omega) \times H_{\Gamma_N}(\operatorname{div}, \Omega)$ with positive constants α, β which are independent of τ and with respect to a suitable norm $|||(\cdot, \cdot)|||$. Such an equivalence will also be needed in the next section for the well-posedness of the linear subproblems in a Gauss–Newton iteration and for deriving the property of the least squares functional as an a posteriori error estimator.

Let us start by considering the special case $H \equiv \bar{H} > 0$, $\mathbf{u} \equiv \mathbf{0}$ which is the unique solution of a stationary problem with constant water level. For this rather simple situation we obtain

$$\mathcal{J}(\bar{H}, \mathbf{0})[\eta, \mathbf{v}] = \begin{pmatrix} \eta + \tau \bar{H} \operatorname{div} \mathbf{v} \\ \mathbf{v} + \tau g \nabla\eta \end{pmatrix},$$

which is obviously uniformly equivalent to

$$\tilde{\mathcal{J}}(\bar{H}, \mathbf{0})[\eta, \mathbf{v}] = \begin{pmatrix} \bar{H}^{-1/2}\eta + \tau \bar{H}^{1/2} \operatorname{div} \mathbf{v} \\ g^{-1/2}\mathbf{v} + \tau g^{1/2} \nabla\eta \end{pmatrix}$$

in the sense that

$$\min\{\bar{H}, g\} \|\tilde{\mathcal{J}}(\bar{H}, \mathbf{0})[\eta, \mathbf{v}]\|^2_{0,\Omega} \le \|\mathcal{J}(\bar{H}, \mathbf{0})[\eta, \mathbf{v}]\|^2_{0,\Omega} \le \max\{\bar{H}, g\} \|\tilde{\mathcal{J}}(\bar{H}, \mathbf{0})[\eta, \mathbf{v}]\|^2_{0,\Omega}$$

holds for all $(\eta, \mathbf{v}) \in H^1_{\Gamma_D}(\Omega) \times H_{\Gamma_N}(\operatorname{div}, \Omega)$. The identity

$$\|\tilde{\mathcal{J}}(\bar{H}, \mathbf{0})[\eta, \mathbf{v}]\|^2_{0,\Omega} = \bar{H}^{-1}\|\eta\|^2_{0,\Omega} + \tau^2\bar{H}\|\operatorname{div}\mathbf{v}\|^2_{0,\Omega} + g^{-1}\|\mathbf{v}\|^2_{0,\Omega} + \tau^2 g\|\nabla\eta\|^2_{0,\Omega}$$

(the cross terms vanish due to integration by parts) suggests the use of the norm

$$|||(\eta, \mathbf{v})||| = (\|\eta\|^2_{0,\Omega} + \tau^2\|\operatorname{div}\mathbf{v}\|^2_{0,\Omega} + \|\mathbf{v}\|^2_{0,\Omega} + \tau^2\|\nabla\eta\|^2_{0,\Omega})^{1/2} \tag{3.4}$$

on $H^1_{\Gamma_D}(\Omega) \times H_{\Gamma_N}(\operatorname{div}, \Omega)$.

For the investigation of the general case, it is convenient to use a transformation of variables. To this end, we write (3.3) as

$$\mathcal{J}(H, \mathbf{u})[\eta, \mathbf{v}] = \begin{pmatrix} (1 + \tau\operatorname{div}\mathbf{u})\eta + \tau H\operatorname{div}\mathbf{v} \\ D(H, \mathbf{u})\mathbf{v} + \tau g\nabla\eta - \tau c_f\dfrac{\mathbf{u}|\mathbf{u}|}{H^2}\eta \end{pmatrix}, \tag{3.5}$$

where the matrix

$$D(H, \mathbf{u}) = \left(1 + \tau c_f\frac{|\mathbf{u}|}{H}\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{\tau c_f}{|\mathbf{u}|H}\begin{pmatrix} u_1^2 & u_1u_2 \\ u_1u_2 & u_2^2 \end{pmatrix}$$

is invertible with

$$D(H, \mathbf{u})^{-1} = \frac{1}{1 + \tau c_f\frac{|\mathbf{u}|}{H}}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{\tau c_f}{|\mathbf{u}|H(1 + \tau c_f\frac{|\mathbf{u}|}{H})(1 + 2\tau c_f\frac{|\mathbf{u}|}{H})}\begin{pmatrix} u_1^2 & u_1u_2 \\ u_1u_2 & u_2^2 \end{pmatrix}.$$

Of course, this formula is valid only for $\mathbf{u} \neq \mathbf{0}$, and for vanishing $\mathbf{u}$ we simply have $D(H, \mathbf{u}) = I$. Note that, in particular,

$$D(H, \mathbf{u})^{-1}\mathbf{u} = \frac{1}{1 + \tau c_f\frac{|\mathbf{u}|}{H}}\left(1 - \frac{\tau c_f\frac{|\mathbf{u}|}{H}}{1 + 2\tau c_f\frac{|\mathbf{u}|}{H}}\right)\mathbf{u} = \frac{1}{1 + 2\tau c_f\frac{|\mathbf{u}|}{H}}\mathbf{u}. \tag{3.6}$$

Obviously, for any $\mathbf{z} \in H_{\Gamma_N}(\operatorname{div}, \Omega)$,

$$\mathbf{z} \cdot D(H, \mathbf{u})\mathbf{z} = \left(1 + \tau c_f\frac{|\mathbf{u}|}{H}\right)|\mathbf{z}|^2 + \tau c_f\frac{1}{|\mathbf{u}|H}(\mathbf{u} \cdot \mathbf{z})^2 \geq \left(1 + \tau c_f\frac{|\mathbf{u}|}{H}\right)|\mathbf{z}|^2. \tag{3.7}$$

Similarly, for $D(H, \mathbf{u})^{-1}$ the estimate

$$\begin{aligned} \mathbf{z} \cdot D(H, \mathbf{u})^{-1}\mathbf{z} &= \frac{1}{1 + \tau c_f\frac{|\mathbf{u}|}{H}}\left(|\mathbf{z}|^2 - \frac{\tau c_f(\mathbf{u} \cdot \mathbf{z})^2}{|\mathbf{u}|H(1 + 2\tau c_f\frac{|\mathbf{u}|}{H})}\right) \\ &\geq \frac{1}{1 + \tau c_f\frac{|\mathbf{u}|}{H}}|\mathbf{z}|^2\left(1 - \tau c_f\frac{|\mathbf{u}|}{H(1 + 2\tau c_f\frac{|\mathbf{u}|}{H})}\right) = \frac{|\mathbf{z}|^2}{1 + 2\tau c_f\frac{|\mathbf{u}|}{H}} \end{aligned} \tag{3.8}$$

holds.

In what follows, we abbreviate our notation by omitting the dependence of D on H and $\mathbf{u}$. Inserting

$$\mathbf{v} = \mathbf{w} + \tau c_f \frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\eta$$

into (3.5), $\mathcal{J}(H,\mathbf{u})[\eta,\mathbf{v}]$ turns into

$$\tilde{\mathcal{J}}(H,\mathbf{u})[\eta,\mathbf{w}] = \begin{pmatrix} (1+\tau \operatorname{div}\mathbf{u})\eta + \tau H \operatorname{div}\mathbf{w} + \tau^2 c_f H \operatorname{div}\left(\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\eta\right) \\ D\mathbf{w} + \tau g \nabla\eta \end{pmatrix}. \tag{3.9}$$

LEMMA 3.1. *If $H \in W^{1,\infty}(\Omega)$, $\mathbf{u} \in W^{1,\infty}(\Omega)^2$, and $H \geq H_*$ with a constant $H_* > 0$, then, for any $\eta \in H^1_{\Gamma_D}(\Omega)$, the mapping*

$$\mathbf{v} \mapsto \mathbf{v} - \tau c_f \frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\eta =: \mathbf{w}$$

is bijective in $H_{\Gamma_N}(\operatorname{div},\Omega)$. Moreover, there is a constant $\gamma > 0$ which is independent of τ such that, under the above transformation,

$$\frac{1}{\gamma}|||(\eta,\mathbf{v})||| \leq |||(\eta,\mathbf{w})||| \leq \gamma|||(\eta,\mathbf{v})||| \tag{3.10}$$

holds, uniformly as $\tau \to 0$, for all $(\eta,\mathbf{v}) \in H^1_{\Gamma_D}(\Omega) \times H_{\Gamma_N}(\operatorname{div},\Omega)$.

Proof. Since $\mathbf{n}\cdot\mathbf{u} = 0$ on Γ_N, (3.6) implies that also $\mathbf{n}\cdot(D^{-1}\mathbf{u}) = 0$ on Γ_N. Therefore, $\mathbf{n}\cdot\mathbf{v} = 0$ on Γ_N if and only if $\mathbf{n}\cdot\mathbf{w} = 0$ on Γ_N. Moreover, from (3.6) we obtain

$$\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u}) = \frac{|\mathbf{u}|}{H(H+2\tau c_f|\mathbf{u}|)}\mathbf{u} \tag{3.11}$$

leading to

$$\left\|\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right\|_{\infty,\Omega} \leq \left\|\frac{\mathbf{u}}{H}\right\|^2_{\infty,\Omega} \leq \left(\frac{\|\mathbf{u}\|_{\infty,\Omega}}{H_*}\right)^2 =: C_1. \tag{3.12}$$

From (3.11) we also obtain

$$\begin{aligned}\operatorname{div}\left(\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right) &= \mathbf{u}\cdot\nabla\left(\frac{|\mathbf{u}|}{H(H+2\tau c_f|\mathbf{u}|)}\right) + \frac{|\mathbf{u}|}{H(H+2\tau c_f|\mathbf{u}|)}\operatorname{div}\mathbf{u} \\ &= \frac{\mathbf{u}\cdot(H^2\nabla(|\mathbf{u}|) - 2|\mathbf{u}|(H+\tau c_f|\mathbf{u}|)\nabla H)}{H^2(H+2\tau c_f|\mathbf{u}|)^2} + \frac{|\mathbf{u}|}{H(H+2\tau c_f|\mathbf{u}|)}\operatorname{div}\mathbf{u}.\end{aligned}$$

Using $|\nabla(|\mathbf{u}|)| \leq |\nabla\mathbf{u}|$ the above identity leads to

$$\begin{aligned}&\left\|\operatorname{div}\left(\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right)\right\|_{\infty,\Omega} \\ &\quad\leq \left\|\frac{|\mathbf{u}|\,|\nabla\mathbf{u}|}{H^2}\right\|_{\infty,\Omega} + 2\left\|\frac{|\mathbf{u}|^2|\nabla H|}{H^3}\right\|_{\infty,\Omega} + \left\|\frac{|\mathbf{u}|\,|\operatorname{div}\mathbf{u}|}{H^2}\right\|_{\infty,\Omega} \\ &\quad\leq 2\left\|\frac{\mathbf{u}}{H}\right\|_{\infty,\Omega}\left(\left\|\frac{\nabla\mathbf{u}}{H}\right\|_{\infty,\Omega} + \left\|\frac{\mathbf{u}}{H}\right\|_{\infty,\Omega}\left\|\frac{\nabla H}{H}\right\|_{\infty,\Omega}\right) \\ &\quad\leq 2\frac{\|\mathbf{u}\|_{\infty,\Omega}}{H_*}\left(\frac{\|\nabla\mathbf{u}\|_{\infty,\Omega}}{H_*} + \frac{\|\mathbf{u}\|_{\infty,\Omega}\|\nabla H\|_{\infty,\Omega}}{H_*^2}\right) =: C_2.\end{aligned} \tag{3.13}$$

Therefore,

$$\begin{aligned}\|\mathbf{w}\|_{0,\Omega} &= \left\|\mathbf{v} - \tau c_f \frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\eta\right\|_{0,\Omega} \le \|\mathbf{v}\|_{0,\Omega} + \tau c_f \left\|\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right\|_{\infty,\Omega} \|\eta\|_{0,\Omega} \\ &\le \|\mathbf{v}\|_{0,\Omega} + \tau c_f C_1 \|\eta\|_{0,\Omega}\end{aligned}$$

and

$$\begin{aligned}\|\operatorname{div}\mathbf{w}\|_{0,\Omega} &= \left\|\operatorname{div}\mathbf{v} - \tau c_f \operatorname{div}\left(\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right)\eta - \tau c_f \frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\cdot\nabla\eta\right\|_{0,\Omega} \\ &\le \|\operatorname{div}\mathbf{v}\|_{0,\Omega} + \tau c_f \left\|\operatorname{div}\left(\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right)\right\|_{\infty,\Omega} \|\eta\|_{0,\Omega} \\ &\quad + \tau c_f \left\|\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right\|_{\infty,\Omega} \|\nabla\eta\|_{0,\Omega} \\ &\le \|\operatorname{div}\mathbf{v}\|_{0,\Omega} + \tau c_f C_2 \|\eta\|_{0,\Omega} + c_f C_1 \tau \|\nabla\eta\|_{0,\Omega},\end{aligned}$$

which leads to

$$\begin{aligned}|||(\eta,\mathbf{w})|||^2 &= \|\eta\|_{0,\Omega}^2 + \tau^2\|\nabla\eta\|_{0,\Omega}^2 + \|\mathbf{w}\|_{0,\Omega}^2 + \tau^2\|\operatorname{div}\mathbf{w}\|_{0,\Omega}^2 \\ &\le (1 + 2\tau^2 c_f^2(C_1^2 + C_2^2))\|\eta\|_{0,\Omega}^2 + (1 + 2c_f^2 C_1^2)\tau^2\|\nabla\eta\|_{0,\Omega}^2 \\ &\quad + 2\|\mathbf{v}\|_{0,\Omega}^2 + 2\tau^2\|\operatorname{div}\mathbf{v}\|_{0,\Omega}^2.\end{aligned}$$

This proves the right inequality in (3.10) with $\gamma = 1 + \max\{1, 2\tau^2 c_f^2(C_1^2 + C_2^2), 2c_f^2 C_1^2\}$. The left inequality follows similarly. ∎

Remark. The linear first-order operator in (3.9) may be reformulated as a second-order operator with respect to η alone. For the resulting second-order differential equation

$$(1 + \tau \operatorname{div}\mathbf{u})\eta + \tau^2 H \operatorname{div}\left(c_f \frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\eta - gD^{-1}\nabla\eta\right) = 0,$$

however, the unique solvability can only be guaranteed under additional conditions on the coefficients of the operator. In the following theorem ellipticity of the associated least squares functional is established under suitable conditions on the coefficients which are formulated as restrictions on the time-step size.

THEOREM 3.2. *Assume as in Lemma* 3.1 *that* $H \in W^{1,\infty}(\Omega)$, $\mathbf{u} \in W^{1,\infty}(\Omega)^2$, *and that there is a positive constant* H_* *such that* $H \ge H_*$ *holds. If, moreover, the time-step size* τ *satisfies the condition*

$$\tau \le \min\left\{\frac{1}{2\|\operatorname{div}\mathbf{u}\|_{\infty,\Omega}}, \frac{1}{(8c_f C_2 H^*)^{1/2}}, \frac{g}{4c_f C_1 (H^*(1 + 2\tau c_f C_1^{1/2}))^{1/2}}\right\}, \tag{3.14}$$

where $H^* = \|H\|_{\infty,\Omega}$ *and* C_1 *and* C_2 *are the constants defined in* (3.12) *and* (3.13), *respectively, then there are positive constants* α *and* β *such that*

$$\alpha|||(\eta,\mathbf{v})|||^2 \le \|\mathcal{J}(H,\mathbf{u})[\eta,\mathbf{v}]\|_{0,\Omega}^2 \le \beta|||(\eta,\mathbf{v})|||^2 \tag{3.15}$$

holds for all $(\eta,\mathbf{v}) \in H^1_{\Gamma_D}(\Omega) \times H_{\Gamma_N}(\operatorname{div},\Omega)$, *uniformly as* $\tau \to 0$.

Proof. In view of Lemma 3.1 we only need to show that

$$\tilde{\alpha}|||(\eta,\mathbf{w})|||^2 \le \|\tilde{\mathcal{J}}(H,\mathbf{u})[\eta,\mathbf{w}]\|_{0,\Omega}^2 \le \tilde{\beta}|||(\eta,\mathbf{w})|||^2$$

holds for all $(\eta,\mathbf{w}) \in H^1_{\Gamma_D}(\Omega)\times H_{\Gamma_N}(\operatorname{div},\Omega)$, uniformly in τ. Moreover, $\tilde{\mathcal{J}}(H,\mathbf{u})[\eta,\mathbf{w}]$ is equivalent to

$$\mathcal{J}^\circ(H,\mathbf{u})[\eta,\mathbf{w}] = \begin{pmatrix} \left(\dfrac{1+\tau\operatorname{div}\mathbf{u}}{H}\right)^{1/2}\eta + \tau\left(\dfrac{H}{1+\tau\operatorname{div}\mathbf{u}}\right)^{1/2}\operatorname{div}\mathbf{w} + \cdots \\ \cdots + \tau^2 c_f\left(\dfrac{H}{1+\tau\operatorname{div}\mathbf{u}}\right)^{1/2}\operatorname{div}\left(\dfrac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\eta\right) \\ \dfrac{1}{(2g)^{1/2}}D^{1/2}\mathbf{w} + \tau\left(\dfrac{g}{2}\right)^{1/2}D^{-1/2}\nabla\eta \end{pmatrix}$$

in the sense that

$$\begin{aligned} &\min\left\{H_*(1-\tau\|\operatorname{div}\mathbf{u}\|_{\infty,\Omega}), 2g\left(1-\tau c_f\frac{\|\mathbf{u}\|_{\infty,\Omega}}{H_*}\right)\right\}\|\mathcal{J}^\circ(H,\mathbf{u})[\eta,\mathbf{w}]\|_{0,\Omega}^2 \\ &\quad\le \|\tilde{\mathcal{J}}(H,\mathbf{u})[\eta,\mathbf{w}]\|_{0,\Omega}^2 \\ &\quad\le \max\left\{H^*(1+\tau\|\operatorname{div}\mathbf{u}\|_{\infty,\Omega}), 2g\left(1+2\tau c_f\frac{\|\mathbf{u}\|_{\infty,\Omega}}{H_*}\right)\right\}\|\mathcal{J}^\circ(H,\mathbf{u})[\eta,\mathbf{w}]\|_{0,\Omega}^2 \end{aligned}$$

holds for all $(\eta,\mathbf{w}) \in H^1_{\Gamma_D}(\Omega)\times H_{\Gamma_N}(\operatorname{div},\Omega)$, uniformly in τ. Note that it is possible to take the square roots above since $1+\tau\operatorname{div}\mathbf{u} > 0$ in Ω due to (3.14). Therefore, all that is left to show is that

$$\alpha^\circ|||(\eta,\mathbf{w})|||^2 \le \|\mathcal{J}^\circ(H,\mathbf{u})[\eta,\mathbf{w}]\|_{0,\Omega}^2 \le \beta^\circ|||(\eta,\mathbf{w})|||^2 \tag{3.16}$$

holds for all $(\eta,\mathbf{w}) \in H^1_{\Gamma_D}(\Omega)\times H_{\Gamma_N}(\operatorname{div},\Omega)$ with constants α° and β° which are independent of τ. We only prove the left inequality in (3.16) which is the hard part.

Using (3.12) and (3.13) again, we obtain

$$\begin{aligned} &\left\|\left(\frac{H}{1+\tau\operatorname{div}\mathbf{u}}\right)^{1/2}\operatorname{div}\left(\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\eta\right)\right\|_{0,\Omega} \\ &\le \left\|\frac{H}{1+\tau\operatorname{div}\mathbf{u}}\right\|_{\infty,\Omega}^{1/2}\left(\left\|\operatorname{div}\left(\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right)\right\|_{\infty,\Omega}\|\eta\|_{0,\Omega} + \left\|\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\right\|_{\infty,\Omega}\|\nabla\eta\|_{0,\Omega}\right) \\ &\le \left(\frac{H^*}{1-\tau\|\operatorname{div}\mathbf{u}\|_{\infty,\Omega}}\right)^{1/2}(C_2\|\eta\|_{0,\Omega} + C_1\|\nabla\eta\|_{0,\Omega}). \end{aligned}$$

Combined with the restriction on the time-step size (3.14), this leads to

$$\left\|\left(\frac{H}{1+\tau\operatorname{div}\mathbf{u}}\right)^{1/2}\operatorname{div}\left(\frac{|\mathbf{u}|}{H^2}(D^{-1}\mathbf{u})\eta\right)\right\|_{0,\Omega} \le (2H^*)^{1/2}(C_2\|\eta\|_{0,\Omega} + C_1\|\nabla\eta\|_{0,\Omega}).$$

This implies

$$
\begin{aligned}
&\|\mathcal{J}^\circ(H,\mathbf{u})[\eta,\mathbf{w}]\|_{0,\Omega}^2 \\
&\quad\geq \frac{1}{2}\left\|\left(\frac{1+\tau\,\mathrm{div}\,\mathbf{u}}{H}\right)^{1/2}\eta+\tau\left(\frac{H}{1+\tau\,\mathrm{div}\,\mathbf{u}}\right)^{1/2}\mathrm{div}\,\mathbf{w}\right\|_{\Omega}^2 \\
&\qquad-2\tau^4c_f^2H^*(C_2\|\eta\|_{0,\Omega}+C_1\|\nabla\eta\|_{0,\Omega})^2+\left\|\frac{1}{(2g)^{1/2}}D^{1/2}\mathbf{w}+\tau\left(\frac{g}{2}\right)^{1/2}D^{-1/2}\nabla\eta\right\|_{0,\Omega}^2 \\
&\quad=\frac{1}{2}\left\|\left(\frac{1+\tau\,\mathrm{div}\,\mathbf{u}}{H}\right)^{1/2}\eta\right\|_{0,\Omega}^2+\frac{\tau^2}{2}\left\|\left(\frac{H}{1+\tau\,\mathrm{div}\,\mathbf{u}}\right)^{1/2}\mathrm{div}\,\mathbf{w}\right\|_{0,\Omega}^2 \\
&\qquad-2\tau^4c_f^2H^*(C_2\|\eta\|_{0,\Omega}+C_1\|\nabla\eta\|_{0,\Omega})^2+\frac{1}{2g}\|D^{1/2}\mathbf{w}\|_{0,\Omega}^2+\frac{g}{2}\tau^2\|D^{-1/2}\nabla\eta\|_{0,\Omega}^2,
\end{aligned}
$$

where we used the identity $(\eta,\mathrm{div}\,\mathbf{w})_{0,\Omega}+(\mathbf{w},\nabla\eta)_{0,\Omega}=0$. Using (3.7) and (3.8), this may be further bounded from below to obtain

$$
\begin{aligned}
&\|\mathcal{J}^\circ(H,\mathbf{u})[\eta,\mathbf{w}]\|_{0,\Omega}^2 \\
&\quad\geq\frac{1-\tau\|\,\mathrm{div}\,\mathbf{u}\|_{\infty,\Omega}}{2H^*}\|\eta\|_{0,\Omega}^2+\frac{\tau^2H_*}{2(1+\tau\|\,\mathrm{div}\,\mathbf{u}\|_{\infty,\Omega})}\|\,\mathrm{div}\,\mathbf{w}\|_{0,\Omega}^2 \\
&\qquad-4\tau^4c_f^2H^*(C_2^2\|\eta\|_{0,\Omega}^2+C_1^2\|\nabla\eta\|_{0,\Omega}^2)+\frac{1}{2g}\|\mathbf{w}\|_{0,\Omega}^2+\frac{g\tau^2}{2(1+2\tau c_fC_1^{1/2})}\|\nabla\eta\|_{0,\Omega}^2 \\
&\quad\geq\left(\frac{1}{4H^*}-4\tau^4c_f^2H^*C_2^2\right)\|\eta\|_{0,\Omega}^2+\left(\frac{g\tau^2}{2(1+2\tau c_fC_1^{1/2})}-4\tau^4c_f^2H^*C_1^2\right)\|\nabla\eta\|_{0,\Omega}^2 \\
&\qquad+\frac{1}{2g}\|\mathbf{w}\|_{0,\Omega}^2+\frac{\tau^2H_*}{2(1+\tau\|\,\mathrm{div}\,\mathbf{u}\|_{\infty,\Omega})}\|\,\mathrm{div}\,\mathbf{w}\|_{0,\Omega}^2.
\end{aligned}
$$

The restriction on the time-step size (3.14) leads finally to

$$
\begin{aligned}
\|\mathcal{J}^\circ(H,\mathbf{u})[\eta,\mathbf{w}]\|_{0,\Omega}^2\geq{}&\frac{1}{8H^*}\|\eta\|_{0,\Omega}^2+\frac{g}{4(1+2\tau c_fC_1^{1/2})}\tau^2\|\nabla\eta\|_{0,\Omega}^2 \\
&+\frac{1}{2g}\|\mathbf{w}\|_{0,\Omega}^2+\frac{H_*}{2(1+\tau\|\,\mathrm{div}\,\mathbf{u}\|_{\infty,\Omega})}\tau^2\|\,\mathrm{div}\,\mathbf{w}\|_{0,\Omega}^2. \qquad \square
\end{aligned}
$$

Remark. Theorem 3.2 shows that the least squares minimization problem (3.2) is well-posed in a neighborhood of the solution $(H,\mathbf{u})$ of the nonlinear elliptic problem (3.1).

We end this section by showing that the Fréchet derivative $\mathcal{J}$ associated with our least squares formulation is Lipschitz continuous. The derivation of a posteriori error estimators in the next section will be done on the basis of this result. It is also useful in the convergence study of Gauss–Newton methods for the iterative solution of the discrete nonlinear variational problems.

LEMMA 3.3. *For* $\mathbf{u},\bar{\mathbf{u}}\in L^2(\Omega)^2$,

$$
\left\|\frac{1}{|\bar{\mathbf{u}}|}\begin{pmatrix}\bar{u}_1^2 & \bar{u}_1\bar{u}_2\\ \bar{u}_1\bar{u}_2 & \bar{u}_2^2\end{pmatrix}-\frac{1}{|\mathbf{u}|}\begin{pmatrix}u_1^2 & u_1u_2\\ u_1u_2 & u_2^2\end{pmatrix}\right\|_{0,\Omega}\leq 3\|\bar{\mathbf{u}}-\mathbf{u}\|_{0,\Omega}. \tag{3.17}
$$

Proof. Using elementary matrix calculations,

$$\left|\frac{1}{|\bar{\mathbf{u}}|}\begin{pmatrix}\bar{u}_1^2 & \bar{u}_1\bar{u}_2\\ \bar{u}_1\bar{u}_2 & \bar{u}_2^2\end{pmatrix}-\frac{1}{|\mathbf{u}|}\begin{pmatrix}u_1^2 & u_1u_2\\ u_1u_2 & u_2^2\end{pmatrix}\right|^2=\left|\frac{\bar{\mathbf{u}}\bar{\mathbf{u}}^T}{|\bar{\mathbf{u}}|}-\frac{\mathbf{u}\mathbf{u}^T}{|\mathbf{u}|}\right|^2$$

$$\begin{aligned}&=|\bar{\mathbf{u}}|^2+|\mathbf{u}|^2-2\frac{(\mathbf{u}^T\bar{\mathbf{u}})^2}{|\mathbf{u}|\,|\bar{\mathbf{u}}|}\\ &=|\mathbf{u}+(\bar{\mathbf{u}}-\mathbf{u})|^2+|\mathbf{u}|^2-2\frac{(\mathbf{u}^T(\mathbf{u}+(\bar{\mathbf{u}}-\mathbf{u})))^2}{|\mathbf{u}|\,|\mathbf{u}+(\bar{\mathbf{u}}-\mathbf{u})|}\\ &=2|\mathbf{u}|^2+2\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})+|\bar{\mathbf{u}}-\mathbf{u}|^2-2\frac{|\mathbf{u}|^4+2|\mathbf{u}|^2\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})+(\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u}))^2}{|\mathbf{u}|^2(1+\frac{2}{|\mathbf{u}|^2}\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})+\frac{|\bar{\mathbf{u}}-\mathbf{u}|^2}{|\mathbf{u}|^2})^{1/2}}\end{aligned}\tag{3.18}$$

is obtained. The two cases $|\bar{\mathbf{u}}-\mathbf{u}|\geq|\mathbf{u}|$ and $|\bar{\mathbf{u}}-\mathbf{u}|\leq|\mathbf{u}|$ are considered separately. For the first case, $|\bar{\mathbf{u}}-\mathbf{u}|\geq|\mathbf{u}|$ implies

$$\left|\frac{\bar{\mathbf{u}}\bar{\mathbf{u}}^T}{|\bar{\mathbf{u}}|}-\frac{\mathbf{u}\mathbf{u}^T}{|\mathbf{u}|}\right|^2\leq 2|\mathbf{u}|^2+2\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})+|\bar{\mathbf{u}}-\mathbf{u}|^2\leq 5|\bar{\mathbf{u}}-\mathbf{u}|^2.\tag{3.19}$$

For the second case, $|\bar{\mathbf{u}}-\mathbf{u}|\leq|\mathbf{u}|$ implies

$$\frac{2}{|\mathbf{u}|^2}\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})+\frac{1}{|\mathbf{u}|^2}|\bar{\mathbf{u}}-\mathbf{u}|^2\leq 3$$

which leads to

$$\left(1+\frac{2}{|\mathbf{u}|^2}\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})+\frac{1}{|\mathbf{u}|^2}|\bar{\mathbf{u}}-\mathbf{u}|^2\right)^{-1/2}\geq 1-\frac{1}{|\mathbf{u}|^2}\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})-\frac{1}{2|\mathbf{u}|^2}|\bar{\mathbf{u}}-\mathbf{u}|^2.$$

Combined with (3.18), we therefore obtain

$$\begin{aligned}&\left|\frac{\bar{\mathbf{u}}\bar{\mathbf{u}}^T}{|\bar{\mathbf{u}}|}-\frac{\mathbf{u}\mathbf{u}^T}{|\mathbf{u}|}\right|^2\\ &\quad\leq 2|\mathbf{u}|^2+2\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})+|\bar{\mathbf{u}}-\mathbf{u}|^2\\ &\quad-2\left(|\mathbf{u}|^2+2\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})+\frac{(\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u}))^2}{|\mathbf{u}|^2}\right)\left(1-\frac{1}{|\mathbf{u}|^2}\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})-\frac{|\bar{\mathbf{u}}-\mathbf{u}|^2}{2|\mathbf{u}|^2}\right)\\ &\quad=2\frac{(\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u}))^2}{|\mathbf{u}|^2}+2\frac{(\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u}))^3}{|\mathbf{u}|^4}+2|\bar{\mathbf{u}}-\mathbf{u}|^2+\frac{\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u})|\bar{\mathbf{u}}-\mathbf{u}|^2}{|\mathbf{u}|^4}\\ &\qquad+2\frac{(\mathbf{u}^T(\bar{\mathbf{u}}-\mathbf{u}))^2|\bar{\mathbf{u}}-\mathbf{u}|^2}{|\mathbf{u}|^2}\leq 4|\bar{\mathbf{u}}-\mathbf{u}|^2+4\frac{|\bar{\mathbf{u}}-\mathbf{u}|^3}{|\mathbf{u}|}+\frac{|\bar{\mathbf{u}}-\mathbf{u}|^4}{|\mathbf{u}|^2}\leq 9|\bar{\mathbf{u}}-\mathbf{u}|^2.\end{aligned}\tag{3.20}$$

The estimate (3.17) follows from integrating (3.19) and (3.20) over Ω. ∎

THEOREM 3.4. *Assume that $\mathbf{u},\bar{\mathbf{u}}\in W^{1,\infty}(\Omega)^2$ with $\mathbf{u}\cdot\mathbf{n}=0$ on Γ_N, $H,\bar{H}\in H^1(\Omega)$ with $H=\bar{H}=H^D$ on Γ_D, and that there is a positive constant H_* such that $H\geq H_*$ and $\bar{H}\geq H_*$ uniformly in Ω. Then,*

$$\|\mathcal{J}(\bar{H},\bar{\mathbf{u}})[\eta,\mathbf{v}]-\mathcal{J}(H,\mathbf{u})[\eta,\mathbf{v}]\|_{0,\Omega}\leq L|||(\bar{H}-H,\bar{\mathbf{u}}-\mathbf{u})|||\,|||(\eta,\mathbf{v})|||,\tag{3.21}$$

uniformly as $\tau\to 0$, for all $(\eta,\mathbf{v})\in H^1_{\Gamma_D}(\Omega)\times H_{\Gamma_N}(\mathrm{div},\Omega)$, where the constant L does not depend on τ.

Proof. Equation (3.5) implies

$$
\begin{aligned}
&\|\mathcal{J}(\bar H,\bar{\mathbf{u}})[\eta,\mathbf{v}]-\mathcal{J}(H,\mathbf{u})[\eta,\mathbf{v}]\|_{0,\Omega}\\
&\quad=\left\|\begin{pmatrix}\tau\operatorname{div}(\bar{\mathbf{u}}-\mathbf{u})\eta+\tau(\bar H-H)\operatorname{div}\mathbf{v}\\ (D(\bar H,\bar{\mathbf{u}})-D(H,\mathbf{u}))\mathbf{v}-\tau c_f\left(\dfrac{\bar{\mathbf{u}}|\bar{\mathbf{u}}|}{\bar H^2}-\dfrac{\mathbf{u}|\mathbf{u}|}{H^2}\right)\eta\end{pmatrix}\right\|_{0,\Omega}\\
&\quad\le\tau\|\operatorname{div}(\bar{\mathbf{u}}-\mathbf{u})\|_{0,\Omega}\|\eta\|_{0,\Omega}+\tau\|\bar H-H\|_{0,\Omega}\|\operatorname{div}\mathbf{v}\|_{0,\Omega}\\
&\qquad+\|D(\bar H,\bar{\mathbf{u}})-D(H,\mathbf{u})\|_{0,\Omega}\|\mathbf{v}\|_{0,\Omega}+\tau c_f\left\|\frac{\bar{\mathbf{u}}|\bar{\mathbf{u}}|}{\bar H^2}-\frac{\mathbf{u}|\mathbf{u}|}{H^2}\right\|_{0,\Omega}\|\eta\|_{0,\Omega}\\
&\quad\le\sqrt{2}\left(\tau^2\|\operatorname{div}(\bar{\mathbf{u}}-\mathbf{u})\|_{0,\Omega}^2+\|\bar H-H\|_{0,\Omega}^2\right.\\
&\qquad\left.+\|D(\bar H,\bar{\mathbf{u}})-D(H,\mathbf{u})\|_{0,\Omega}^2+\tau^2c_f^2\left\|\frac{\bar{\mathbf{u}}|\bar{\mathbf{u}}|}{\bar H^2}-\frac{\mathbf{u}|\mathbf{u}|}{H^2}\right\|_{0,\Omega}^2\right)^{1/2}|||(\eta,\mathbf{v})|||.
\end{aligned}
\tag{3.22}
$$

The definition of $D(H,\mathbf{u})$ (cf. (3.5)) implies

$$
\begin{aligned}
D(\bar H,\bar{\mathbf{u}})-D(H,\mathbf{u})=\tau c_f\left(\left(\frac{|\bar{\mathbf{u}}|}{\bar H}-\frac{|\mathbf{u}|}{H}\right)\begin{pmatrix}1&0\\0&1\end{pmatrix}\right.\\
\left.+\frac{1}{|\bar{\mathbf{u}}|\bar H}\begin{pmatrix}\bar u_1^2&\bar u_1\bar u_2\\ \bar u_1\bar u_2&\bar u_2^2\end{pmatrix}-\frac{1}{|\mathbf{u}|H}\begin{pmatrix}u_1^2&u_1u_2\\ u_1u_2&u_2^2\end{pmatrix}\right)
\end{aligned}
$$

and therefore

$$
\begin{aligned}
&\|D(\bar H,\bar{\mathbf{u}})-D(H,\mathbf{u})\|_{0,\Omega}\\
&\quad\le\tau c_f\left(\sqrt{2}\left\|\frac{|\bar{\mathbf{u}}|}{\bar H}-\frac{|\mathbf{u}|}{H}\right\|_{0,\Omega}+\left\|\left(\frac{1}{\bar H}-\frac{1}{H}\right)\frac{1}{|\mathbf{u}|}\begin{pmatrix}u_1^2&u_1u_2\\ u_1u_2&u_2^2\end{pmatrix}\right\|_{0,\Omega}\right.\\
&\qquad\left.+\left\|\frac{1}{\bar H}\left(\frac{1}{|\bar{\mathbf{u}}|}\begin{pmatrix}\bar u_1^2&\bar u_1\bar u_2\\ \bar u_1\bar u_2&\bar u_2^2\end{pmatrix}-\frac{1}{|\mathbf{u}|}\begin{pmatrix}u_1^2&u_1u_2\\ u_1u_2&u_2^2\end{pmatrix}\right)\right\|_{0,\Omega}\right)\\
&\quad\le\tau c_f\left(\frac{\sqrt{2}}{H_*}\||\bar{\mathbf{u}}|-|\mathbf{u}|\|_{0,\Omega}+(1+\sqrt{2})\|\mathbf{u}\|_{\infty,\Omega}\left\|\frac{1}{\bar H}-\frac{1}{H}\right\|_{0,\Omega}\right.\\
&\qquad\left.+\frac{1}{H_*}\left\|\frac{1}{|\bar{\mathbf{u}}|}\begin{pmatrix}\bar u_1^2&\bar u_1\bar u_2\\ \bar u_1\bar u_2&\bar u_2^2\end{pmatrix}-\frac{1}{|\mathbf{u}|}\begin{pmatrix}u_1^2&u_1u_2\\ u_1u_2&u_2^2\end{pmatrix}\right\|_{0,\Omega}\right)\\
&\quad\le\tau c_f\left(\frac{\sqrt{2}}{H_*}\|\bar{\mathbf{u}}-\mathbf{u}\|_{0,\Omega}+\frac{1+\sqrt{2}}{H_*^2}\|\mathbf{u}\|_{\infty,\Omega}\|\bar H-H\|_{0,\Omega}+\frac{3}{H_*}\|\bar{\mathbf{u}}-\mathbf{u}\|_{0,\Omega}\right),
\end{aligned}
$$

where we used (3.17) for estimating the last term. Moreover,

$$
\begin{aligned}
\left\|\frac{\bar{\mathbf{u}}|\bar{\mathbf{u}}|}{\bar H^2}-\frac{\mathbf{u}|\mathbf{u}|}{H^2}\right\|_{0,\Omega}&=\left\|\frac{|\bar{\mathbf{u}}|}{\bar H^2}(\bar{\mathbf{u}}-\mathbf{u})+\frac{\mathbf{u}}{\bar H^2}(|\bar{\mathbf{u}}|-|\mathbf{u}|)+\mathbf{u}|\mathbf{u}|\left(\frac{1}{\bar H}+\frac{1}{H}\right)\frac{H-\bar H}{\bar H H}\right\|_{0,\Omega}\\
&\le\frac{\|\bar{\mathbf{u}}\|_{\infty,\Omega}+\|\mathbf{u}\|_{\infty,\Omega}}{H_*^2}\|\bar{\mathbf{u}}-\mathbf{u}\|_{0,\Omega}+2\frac{\|\mathbf{u}\|_{\infty,\Omega}^2}{H_*^3}\|\bar H-H\|_{0,\Omega}.
\end{aligned}
$$

Inserting all this into (3.22) implies

$$\begin{aligned}
&\|\mathcal{J}(\bar H,\bar{\mathbf{u}})[\eta,\mathbf{v}]-\mathcal{J}(H,\mathbf{u})[\eta,\mathbf{v}]\|_{0,\Omega}\\
&\quad\le\sqrt{2}\left(\tau^2\|\operatorname{div}(\bar{\mathbf{u}}-\mathbf{u})\|_{0,\Omega}^2+\|\bar H-H\|_{0,\Omega}^2\right.\\
&\qquad+2\tau^2c_f^2\left(\frac{11}{H_*^2}\|\bar{\mathbf{u}}-\mathbf{u}\|_{0,\Omega}^2+\frac{6}{H_*^4}\|\mathbf{u}\|_{\infty,\Omega}^2\|\bar H-H\|_{0,\Omega}^2\right)\\
&\qquad+2\tau^2c_f^2\left(\frac{(\|\bar{\mathbf{u}}\|_{\infty,\Omega}+\|\mathbf{u}\|_{\infty,\Omega})^2}{H_*^4}\|\bar{\mathbf{u}}-\mathbf{u}\|_{0,\Omega}^2\right.\\
&\qquad\left.\left.+4\frac{\|\mathbf{u}\|_{\infty,\Omega}^4}{H_*^6}\|\bar H-H\|_{0,\Omega}^2\right)\right)^{1/2}|||(\eta,\mathbf{v})|||
\end{aligned}$$

which proves (3.21). ∎

We end this section with an investigation of the behavior of the least squares functional near the solution $(H,\mathbf{u})$ of (3.1). Using the fact that $\mathcal{R}(H,\mathbf{u};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}})=0$ and the Lipschitz continuity shown in Theorem 3.4, it can be shown that for $(\bar H,\bar{\mathbf{u}})$, $(\tilde H,\tilde{\mathbf{u}})$ sufficiently close to $(H,\mathbf{u})$, up to quadratic terms in $(\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}})$,

$$\begin{aligned}
&(\mathcal{R}(\bar H,\bar{\mathbf{u}};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}}),\mathcal{J}(\bar H,\bar{\mathbf{u}})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}\\
&\quad-(\mathcal{R}(\tilde H,\tilde{\mathbf{u}};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}}),\mathcal{J}(\tilde H,\tilde{\mathbf{u}})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}\\
&=(\mathcal{R}(\bar H,\bar{\mathbf{u}};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}})-\mathcal{R}(H,\mathbf{u};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}}),\mathcal{J}(\bar H,\bar{\mathbf{u}})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}\\
&\quad-(\mathcal{R}(\tilde H,\tilde{\mathbf{u}};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}})-\mathcal{R}(H,\mathbf{u};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}}),\mathcal{J}(\tilde H,\tilde{\mathbf{u}})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}\\
&\approx(\mathcal{J}(H,\mathbf{u})[\bar H-H,\bar{\mathbf{u}}-\mathbf{u}],\mathcal{J}(H,\mathbf{u})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}\\
&\quad-(\mathcal{J}(H,\mathbf{u})[\tilde H-H,\tilde{\mathbf{u}}-\mathbf{u}],\mathcal{J}(H,\mathbf{u})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}\\
&=(\mathcal{J}(H,\mathbf{u})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}],\mathcal{J}(H,\mathbf{u})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}.
\end{aligned}$$

Combined with (3.15), this implies that

$$\begin{aligned}
&(\mathcal{R}(\bar H,\bar{\mathbf{u}};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}}),\mathcal{J}(\bar H,\bar{\mathbf{u}})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}\\
&\quad-(\mathcal{R}(\tilde H,\tilde{\mathbf{u}};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}}),\mathcal{J}(\tilde H,\tilde{\mathbf{u}})[\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}}])_{0,\Omega}\ge\breve\alpha|||(\bar H-\tilde H,\bar{\mathbf{u}}-\tilde{\mathbf{u}})|||^2
\end{aligned}$$

holds with a constant $\breve\alpha>0$. This condition is equivalent to the statement that the least squares functional $\|\mathcal{R}(\bar H,\bar{\mathbf{u}};\hat H^{\mathrm{old}},\hat{\mathbf{u}}^{\mathrm{old}})\|_{0,\Omega}^2$ is uniformly convex in a neighborhood of $(H,\mathbf{u})$ (cf. [1, section 4.3.4]).

4. Finite element discretization, approximate mass balance, and Gauss–Newton iteration. The least squares finite element method consists in setting the minimization problem (3.2) in finite-dimensional subspaces $Q_h\subset H^1_{\Gamma_D}(\Omega)$ and $\mathbf{V}_h\subset H_{\Gamma_N}(\operatorname{div},\Omega)$. Suitable finite element spaces are standard H^1-conforming piecewise polynomial functions for Q_h and the $H(\operatorname{div})$-conforming Raviart–Thomas elements for $\mathbf{V}_h$. The minimization problem in these finite-dimensional spaces has a unique solution in a neighborhood of $(H,\mathbf{u})$ since the least squares functional is uniformly convex (cf. [1, section 3.2.2]).

The following general relation holds between the residual $\mathcal{R}$ and its derivative $\mathcal{J}$:

$$\begin{aligned}\mathcal{R}(H_h,\mathbf{u}_h;\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}}) &= \mathcal{R}(H_h,\mathbf{u}_h;\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}}) - \mathcal{R}(H,\mathbf{u};\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}})\\ &= \mathcal{J}(H,\mathbf{u})[H_h-H,\mathbf{u}_h-\mathbf{u}]\\ &\quad + \int_0^1 \left(\mathcal{J}(H+s(H_h-H),\mathbf{u}+s(\mathbf{u}_h-\mathbf{u})) - \mathcal{J}(H,\mathbf{u})\right)[H_h-H,\mathbf{u}_h-\mathbf{u}]\,ds.\end{aligned}$$

Using the Lipschitz continuity of $\mathcal{J}$ shown in Theorem 3.4, this leads to

$$\|\mathcal{R}(H_h,\mathbf{u}_h;\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}})\|_{0,\Omega} \le \|\mathcal{J}(H,\mathbf{u})[H_h-H,\mathbf{u}_h-\mathbf{u}]\|_{0,\Omega} + \frac{L}{2}|||(H_h-H,\mathbf{u}_h-\mathbf{u})|||^2$$

and

$$\|\mathcal{R}(H_h,\mathbf{u}_h;\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}})\|_{0,\Omega} \ge \|\mathcal{J}(H,\mathbf{u})[H_h-H,\mathbf{u}_h-\mathbf{u}]\|_{0,\Omega} - \frac{L}{2}|||(H_h-H,\mathbf{u}_h-\mathbf{u})|||^2.$$

Combined with Theorem 3.2, this implies that there exist positive constants $\ddot{\alpha}$ and $\dddot{\beta}$, such that

$$\ddot{\alpha}|||(H_h-H,\mathbf{u}_h-\mathbf{u})|||^2 \le \|\mathcal{R}(H_h,\mathbf{u}_h;\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}})\|_{0,\Omega}^2 \le \dddot{\beta}|||(H_h-H,\mathbf{u}_h-\mathbf{u})|||^2 \tag{4.1}$$

holds if $|||(H_h-H,\mathbf{u}_h-\mathbf{u})|||$ is sufficiently small. The practical implication of (4.1) is that for a given triangulation $\mathcal{T}_h$ of Ω, the local evaluation of the least squares functional,

$$\|\mathcal{R}(H_h,\mathbf{u}_h;\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}})\|_{0,T}^2 \quad \text{for each } T\in\mathcal{T}_h,$$

serves as an a posteriori error estimator. Note that only the constants $\ddot{\alpha}$ and $\dddot{\beta}$ affect the sharpness of this a posteriori error estimator.

For some applications, it is of special importance that the mass balance error which is associated with the first equation in (2.2) is small. More precisely, let $\mathcal{T}$ be a fixed triangulation of Ω, then a measure for the local mass balance error is given by

$$\mathcal{M}(H_h,\mathbf{u}_h) = \sum_{K\in\mathcal{T}} \left| \int_K (H_h - \hat{H}^{\text{old}} + \tau H_h \operatorname{div}\mathbf{u}_h)\,d\mathbf{x} \right|. \tag{4.2}$$

For the Raviart–Thomas mixed finite element method employed in [8, 12], this mass balance error is zero (up to interpolation errors for $\hat{H}^{\text{old}}$ on the current triangulation). In our least squares approach $\mathcal{M}(H_h,\mathbf{u}_h)$ does not vanish, in general, but it is rather small, and decreases with refinement, as illustrated by the numerical results in the next section. Similar results were observed in connection to the study of mass conservation of a least squares mixed finite element method applied to variably saturated subsurface flow (see [16]).

For the solution of the discrete version of (3.2), i.e., to find $(\hat{H}_h,\mathbf{u}_h)\in Q_h\times\mathbf{V}_h$ such that

$$\|\mathcal{R}(H^D+\hat{H}_h,\mathbf{u}_h;\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}})\|_{0,\Omega}^2 \le \|\mathcal{R}(H^D+\eta_h,\mathbf{v}_h;\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}})\|_{0,\Omega}^2 \tag{4.3}$$

holds for all $(\eta_h,\mathbf{v}_h)\in Q_h\times\mathbf{V}_h$, we use a Gauss–Newton iteration with suitable stopping criterion (see [16, 15]). The kth Gauss–Newton step consists of computing $\delta_h^H\in Q_h$ and $\boldsymbol{\delta}_h^u\in\mathbf{V}_h$ such that

$$\left\| \mathcal{R}\left(H_h^{(k)},\mathbf{u}_h^{(k)};\hat{H}^{\text{old}},\hat{\mathbf{u}}^{\text{old}}\right) + \mathcal{J}\left(H_h^{(k)},\mathbf{u}_h^{(k)}\right)[\delta_h^H,\boldsymbol{\delta}_h^u] \right\|_{0,\Omega}^2$$

is minimized, and setting

$$H_h^{(k+1)} = H_h^{(k)} + \delta_h^H, \qquad \mathbf{u}_h^{(k+1)} = \mathbf{u}_h^{(k)} + \boldsymbol{\delta}_h^u.$$

This is equivalent to the variational formulation of finding $\delta_h^H \in Q_h$ and $\boldsymbol{\delta}_h^u \in \mathbf{V}_h$ such that

(4.4)
$$\left(\mathcal{J}\left(H_h^{(k)}, \mathbf{u}_h^{(k)}\right)[\delta_h^H, \boldsymbol{\delta}_h^u] + \mathcal{R}\left(H_h^{(k)}, \mathbf{u}_h^{(k)}; \hat{H}^{\text{old}}, \hat{\mathbf{u}}^{\text{old}}\right), \mathcal{J}\left(H_h^{(k)}, \mathbf{u}_h^{(k)}\right)[\eta_h, \mathbf{v}_h]\right)_{0,\Omega} = 0$$

for all $\eta_h \in Q_h$ and $\mathbf{v}_h \in \mathbf{V}_h$. The size of the algebraic residual

$$\left(\mathcal{R}\left(H_h^{(k)}, \mathbf{u}_h^{(k)}; \hat{H}^{\text{old}}, \hat{\mathbf{u}}^{\text{old}}\right), \mathcal{J}\left(H_h^{(k)}, \mathbf{u}_h^{(k)}\right)[\eta_h, \mathbf{v}_h]\right)_{0,\Omega}$$

may be used as a stopping criterion for the Gauss–Newton iteration (cf. [15]). In general, the Gauss–Newton iteration converges only locally, i.e., if $(H_h^{(0)}, \mathbf{u}_h^{(0)})$ is already sufficiently close to $(H_h, \mathbf{u}_h)$. Therefore, the scheme has to be combined with a line search or trust region strategy to ensure convergence (cf. [13, Chapter 10]). The linear variational problems (4.4) arising in each Gauss–Newton step may most efficiently be solved by multilevel methods (see [15] for such a strategy for a related nonlinear least squares formulation). Embedding such an inexact Gauss–Newton iteration into a full multigrid framework can lead to an overall method with a computational cost which grows only linearly in the dimension of the system. This requires the careful adjustment of the different tolerances for the linear and nonlinear algebraic errors with respect to the discretization error (cf. [15]).

5. Computational results. This section is concerned with the validation of our least squares approach by numerical results obtained for a test example of shallow water flow taken from [8]. The recirculation of the flow behind an abrupt widening of a channel is only captured by sufficiently fine meshes. Adaptive refinement strategies are helpful for this purpose.

The domain Ω is best defined by its boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$ with Γ_D and Γ_N given by

$$\begin{aligned}
\Gamma_D &= \{\mathbf{x} = (x_1, x_2) : x_1 = 100,\ 0 \le x_2 \le 100\} \\
&\quad \cup \{\mathbf{x} = (x_1, x_2) : x_1 = 1100,\ 0 \le x_2 \le 200\}, \\
\Gamma_N &= \{\mathbf{x} = (x_1, x_2) : x_2 = 0,\ 100 \le x_1 \le 1100\} \\
&\quad \cup \{\mathbf{x} = (x_1, x_2) : x_2 = 100,\ 100 \le x_1 \le 200\} \\
&\quad \cup \{\mathbf{x} = (x_1, x_2) : x_1 = 200,\ 100 \le x_2 \le 200\} \\
&\quad \cup \{\mathbf{x} = (x_1, x_2) : x_2 = 200,\ 200 \le x_1 \le 1100\}.
\end{aligned}$$

As boundary conditions on Γ_D we have

$$\begin{aligned}
H(\mathbf{x}) &= 2 + \frac{1}{4}\sin\left(\frac{t\pi}{300}\right) \quad \text{if } x_1 = 100, \\
H(\mathbf{x}) &= 2 \quad \text{if } x_1 = 1100,
\end{aligned}$$

modelling a water level that varies with t at the left end of Ω while keeping the elevation constant at the right boundary. On Γ_N the normal component of the velocity, $\mathbf{u} \cdot \mathbf{n}$, is set to zero. The Chezy coefficient was chosen as $c_f = 0.002725$ which is taken from [8]. The initial conditions at $t = 0$ are $H \equiv 2$ and $\mathbf{u} \equiv \mathbf{0}$.

The time discretization is done with a time-step size τ of 5 seconds. These rather small time steps are chosen in order to resolve the variation of the inflow

boundary conditions with sufficient accuracy. Moreover, our experiments showed that for larger time-step sizes the cost associated with the Gauss–Newton iteration for solving the nonlinear systems in each time step is growing significantly. For the space discretization, piecewise quadratic functions are used for Q_h and combined with quadratic Raviart–Thomas spaces for $\mathbf{V}_h$. Starting from a coarse initial triangulation, four steps of adaptive refinement based on the local evaluation of the least squares functional described in the previous section are performed. With a given tolerance ε for the desired value of the functional minimum, all triangles are refined which satisfy

$$\|\mathcal{R}(H_h, \mathbf{u}_h; \hat{H}^{\text{old}}, \hat{\mathbf{u}}^{\text{old}})\|_{0,T}^2 > \frac{\varepsilon}{\#\text{elements}}$$

for the local contribution to the least squares functional.

We present the results of our computations at various stages of the time evolution. As it can be seen in Figures 5.1 and 5.2, at $t = 100$ a sink begins to form behind the reentrant corner. At $t = 200$ the recirculation phase starts, gaining in strength until $t = 300$. Afterwards, at $t = 400$, the flow starts turning until it is directed backwards at $t = 500$. For $t = 600$ this cycle starts again resembling the results for $t = 100$. Figure 5.3 shows the local distribution of the least squares functional for the same times on a logarithmic scale. For each of the shown time steps lighter zones starting from the reentrant corner with relatively large contributions to the least squares functional can be clearly identified. The use of this error estimator results in the adaptively refined triangulations shown in Figure 5.4. Obviously, a high level of refinement is reached primarily in the recirculation zone in order to resolve the local variation of the water level and velocity field in this area.

Tables 5.1 to 5.6 show the reduction of the least squares functional on the resulting sequence of adaptively refined triangulations. The dimensions of the finite element spaces Q_h and $\mathbf{V}_h$ are denoted by N_Q and N_V, respectively. The optimal convergence behavior achievable with piecewise quadratic finite elements would result in a reduction of the least squares functional proportional to $(N_Q + N_V)^{-2}$. Such a behavior is only reached in our numerical tests for $t = 200$ and 300. For the other times only $(N_Q + N_V)^{-\mu}$ with μ well below 2 is reached in average for the reduction of the least squares functional. This suboptimal convergence behavior indicates that the adaptive refinement strategy based on a fixed tolerance for the local contribution of the functional may not be appropriate at these time steps. On the other hand, this phenomenon may also be due to the nonlinear nature of the problem which is possibly not resolved on all refinement levels shown. Note that the reduction exponent μ moves somewhat closer to 2 on finer levels in most of the time steps shown. Interestingly, the largest deviation from the expected optimal convergence behavior is seen at those time steps where relative large recirculation zones are present.

Also shown is the mass balance error $\mathcal{M}(H_h, \mathbf{u}_h)$ associated with these time steps. The coarsest triangulation (level 0) is used for $\mathcal{T}$ in the definition of $\mathcal{M}(H_h, \mathbf{u}_h)$ in (4.2). Our numerical results indicate that the decrease in the mass balance error is at least as fast as for the least squares functional. Note that the computed mass balance error needs to be related to the typical size of the mass variation

$$\left| \int_\Omega (H - \hat{H}^{\text{old}})\, d\mathbf{x} \right| \approx |\Omega|\, |\Delta H| = 1.7 \cdot 10^5 |\Delta H|.$$

The inflow boundary conditions suggest that ΔH is in the range of $10^{-3}\tau = 5 \cdot 10^{-3}$ which leads to an overall mass variation in the range of 10^3. The relative mass balance error in Tables 5.1 to 5.6 is therefore on the order of 10^{-4} to 10^{-3} on the finest level.

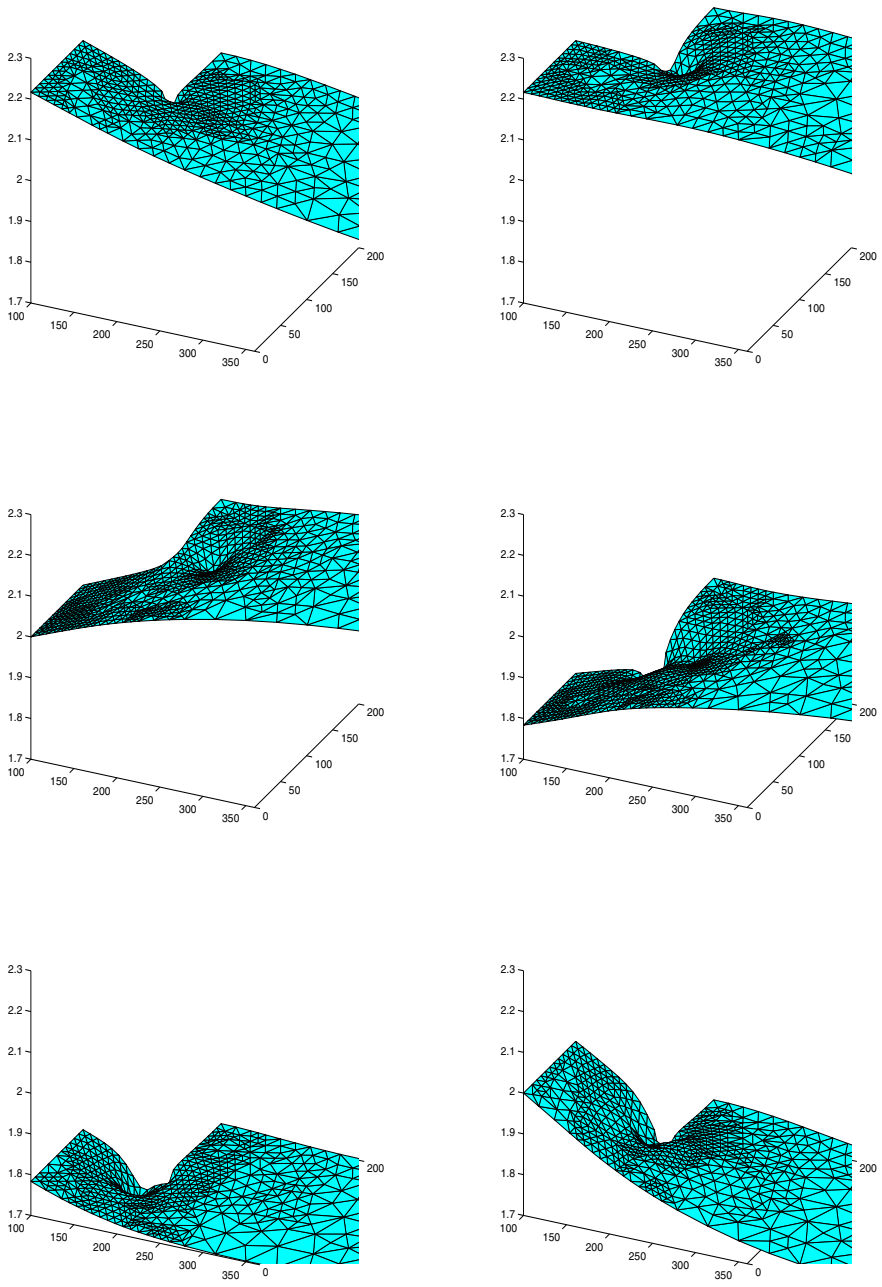


FIG. 5.1. *Elevation for* $t = 100, 200, 300, 400, 500,$ *and* 600.

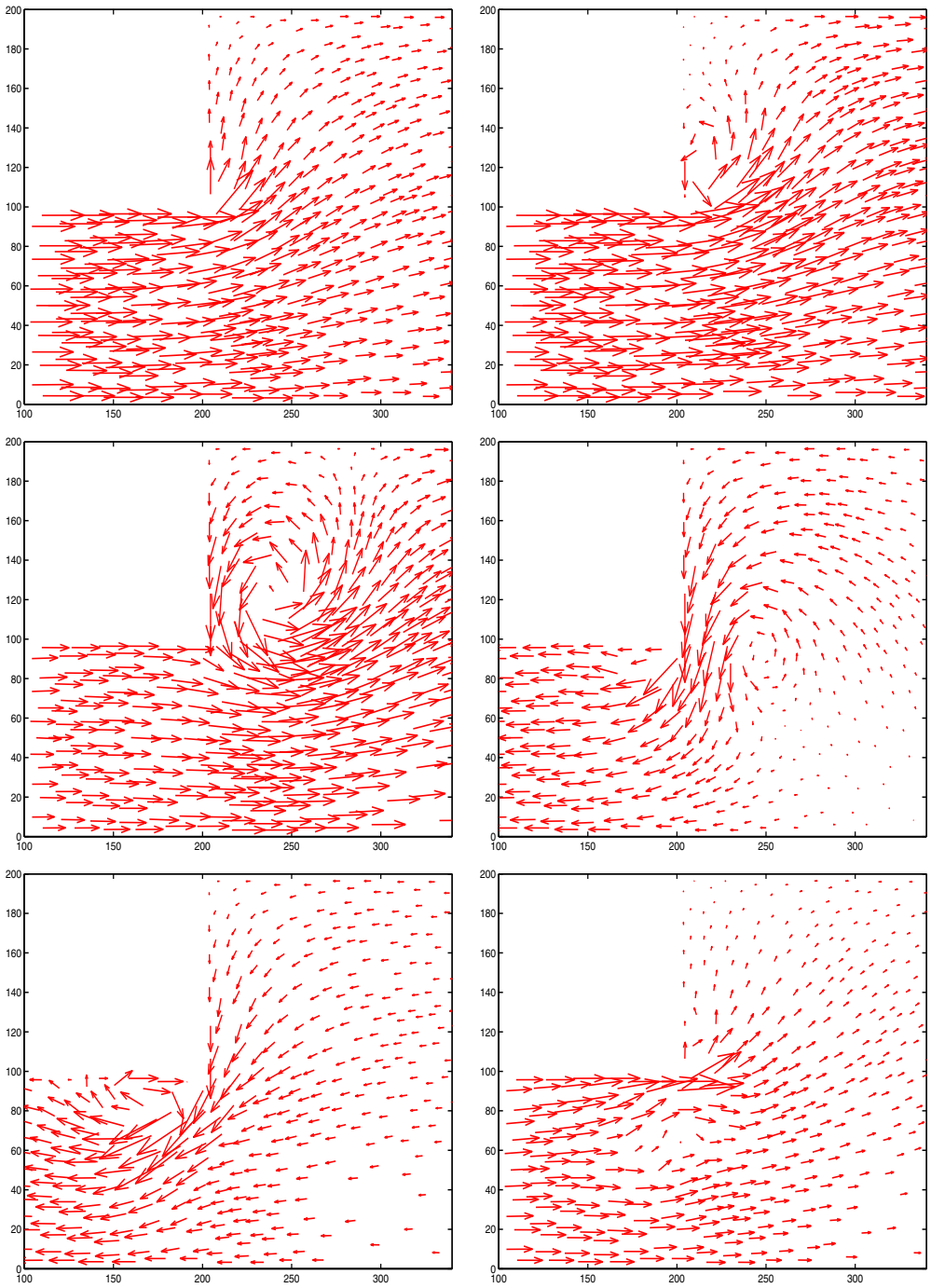


FIG. 5.2. *Velocity field for* $t = 100, 200, 300, 400, 500,$ *and* 600.

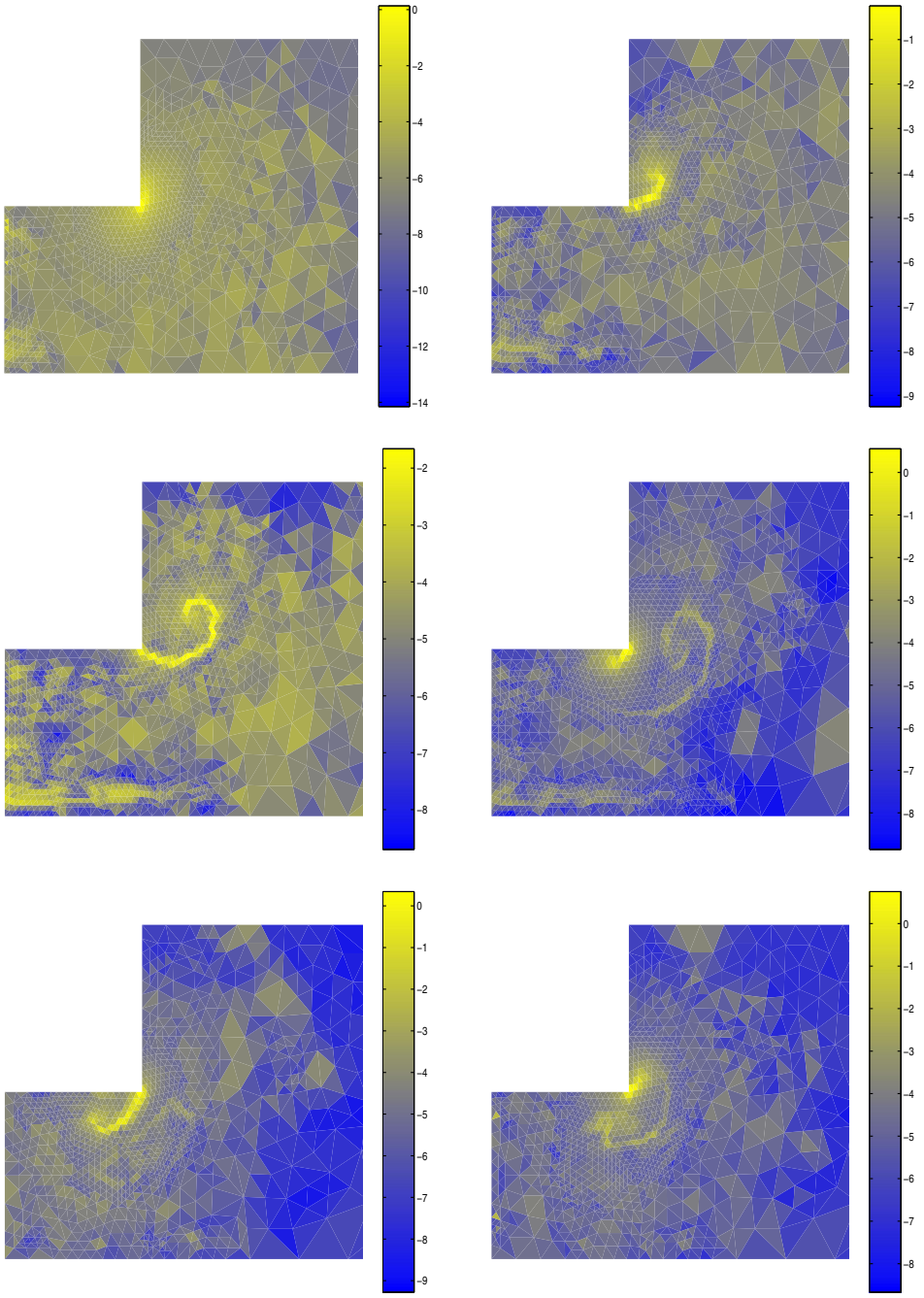


FIG. 5.3. *Least squares functional on level* 3 *for* $t = 100, 200, 300, 400, 500,$ *and* 600.

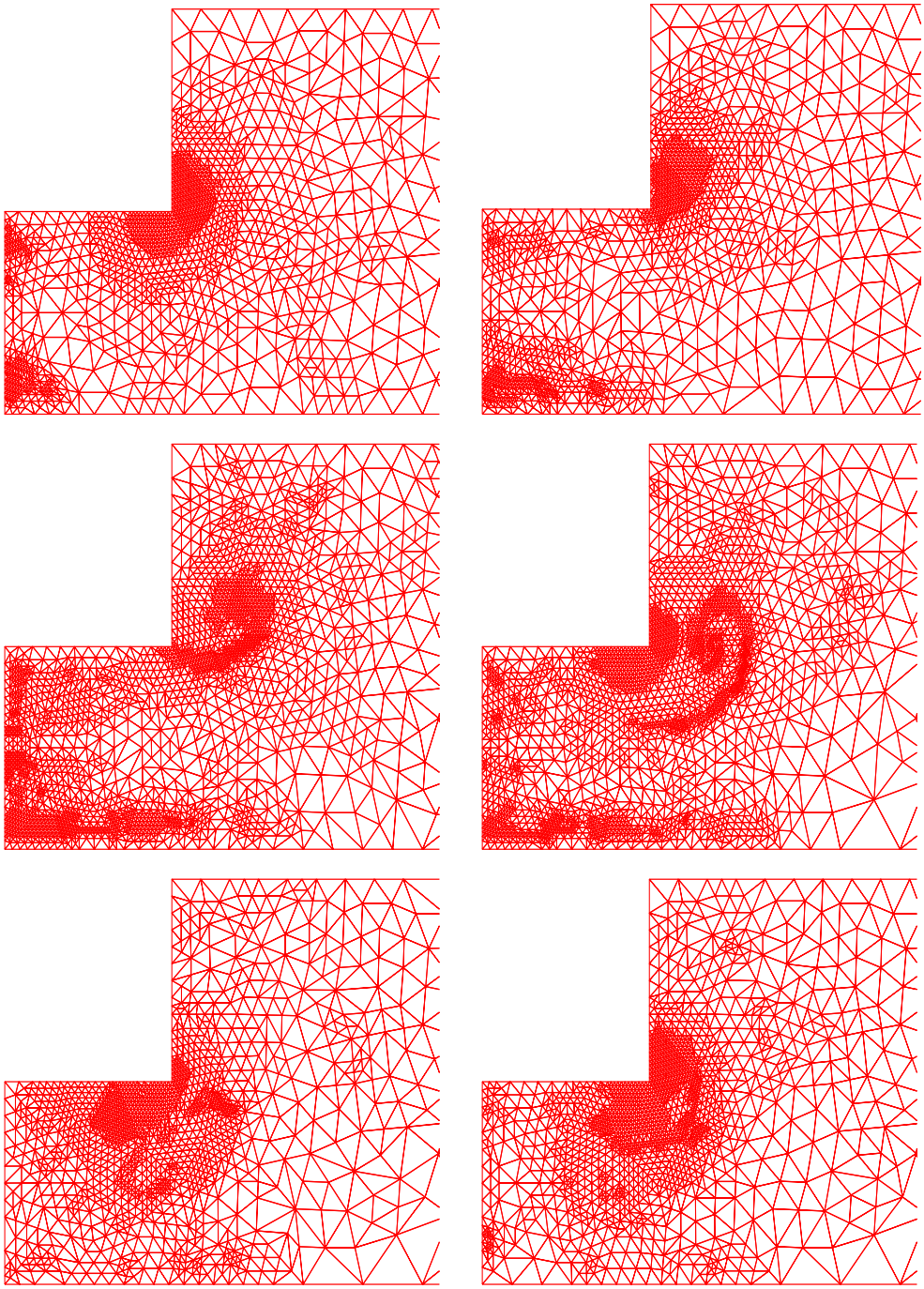


FIG. 5.4. *Adaptively refined triangulations for* $t = 100, 200, 300, 400, 500,$ *and* 600.

TABLE 5.1
Least squares functional for $t = 100$.

l	0	1	2	3	4
# Elements	461	801	1325	2075	3001
N_Q	972	1668	2724	4230	6090
N_V	2254	3938	6550	10294	14916
Functional	31.39	15.66	7.63	2.50	1.21
Reduction exponent μ	—	1.26	1.43	2.49	1.97
Mass balance error	68.55	27.63	13.18	5.01	2.14

TABLE 5.2
Least squares functional for $t = 200$.

l	0	1	2	3	4
# Elements	461	820	1342	2117	2926
N_Q	972	1705	2761	4312	5933
N_V	2254	4034	6632	10506	14548
Functional	92.50	15.41	5.01	2.85	1.49
Reduction exponent μ	—	3.11	2.28	1.24	2.00
Mass balance error	65.36	5.88	2.48	1.53	1.07

TABLE 5.3
Least squares functional for $t = 300$.

l	0	1	2	3	4
# Elements	461	812	1547	2962	4760
N_Q	972	1687	3172	6005	9597
N_V	2254	3996	7656	14728	23722
Functional	30.99	7.08	2.08	0.54	0.13
Reduction exponent μ	—	2.61	1.90	2.06	3.01
Mass balance error	51.55	6.58	1.68	0.83	0.65

TABLE 5.4
Least squares functional for $t = 400$.

l	0	1	2	3	4
# Elements	461	800	1593	3076	4968
N_Q	972	1665	3270	6241	10037
N_V	2254	3934	7880	15290	24738
Functional	80.53	21.67	19.92	9.49	5.50
Reduction exponent μ	—	2.38	0.12	1.13	1.14
Mass balance error	110.87	25.68	23.74	8.51	3.81

TABLE 5.5
Least squares functional for $t = 500$.

l	0	1	2	3	4
# Elements	461	751	1293	2189	3179
N_Q	972	1566	2666	4472	6462
N_V	2254	3690	6384	10850	15790
Functional	52.42	25.14	7.95	6.45	2.41
Reduction exponent μ	—	1.51	2.12	0.40	2.64
Mass balance error	84.87	12.78	6.74	4.20	2.08

TABLE 5.6
Least squares functional for $t = 600$.

l	0	1	2	3	4
# Elements	461	784	1343	2216	3428
N_Q	972	1631	2762	4517	6951
N_V	2254	3856	6638	10994	17044
Functional	115.03	71.01	36.41	13.20	5.66
Reduction exponent μ	—	0.91	1.24	2.03	1.94
Mass balance error	129.10	50.23	23.91	9.84	4.29

Acknowledgment. I am thankful to the anonymous referees for helpful comments and suggestions.

REFERENCES

[1] K. E. ATKINSON AND W. HAN, *Theoretical Numerical Analysis*, Springer, New York, 2001.
[2] J. BEHRENS, *Atmospheric and ocean modeling with an adaptive finite element solver for the shallow-water equations*, Appl. Numer. Math., 26 (1998), pp. 217–226.
[3] M. BERNDT, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Local error estimates and adaptive refinement for first-order system least squares*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43.
[4] P. BOCHEV, Z. CAI, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Analysis of velocity-flux first-order system least-squares principles for the Navier–Stokes equations: Part* I, SIAM J. Numer. Anal., 35 (1998), pp. 990–1009.
[5] P. B. BOCHEV AND M. D. GUNZBURGER, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.
[6] P. BOCHEV, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Analysis of velocity-flux least-squares principles for the Navier–Stokes equations: Part* II, SIAM J. Numer. Anal., 36 (1999), pp. 1125–1144.
[7] C. N. DAWSON AND M. MARTINEZ-CANALES, *A characteristic-Galerkin approximation to a system of shallow water equations*, Numer. Math., 86 (2000), pp. 239–256.
[8] L. FONTANA, E. MIGLIO, A. QUARTERONI, AND F. SALERI, *A finite element method for 3D hydrostatic water flows*, Comput. Vis. Sci., 2 (1999), pp. 85–93.
[9] R. HINKELMANN AND W. ZIELKE, *A parallel 2D Lagrangian-Eulerian model for the shallow water equations*, in Computing in Civil and Building Engineering, Vol. 1, P. J. Pahl and H. Werner, eds., A. A. Balkema Publishers, Rotterdam, The Netherlands, 1995, pp. 537–543.
[10] A. HOLSTAD AND I. LIE, *Transparent boundary conditions for the shallow water equations with a mixed finite element formulation*, Appl. Numer. Math., 44 (2003), pp. 109–138.
[11] M. MARROCU AND D. AMBROSI, *Mesh adaptation strategies for shallow water flow*, Internat. J. Numer. Methods Fluids, 31 (1999), pp. 497–512.
[12] E. MIGLIO, A. QUARTERONI, AND F. SALERI, *Finite element approximation of quasi-3D shallow water equations*, Comput. Methods Appl. Mech. Engrg., 174 (1999), pp. 355–369.
[13] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer, New York, 1999.
[14] J. PETERA AND V. NASSEHI, *A new two-dimensional finite element model for the shallow water equations using a Lagrangian framework constructed along fluid particle trajectories*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 4159–4182.
[15] G. STARKE, *Gauss-Newton multilevel methods for least-squares finite element computations of variably saturated subsurface flow*, Computing, 64 (2000), pp. 323–338.
[16] G. STARKE, *Least-squares mixed finite element solution of variably saturated subsurface flow problems*, SIAM J. Sci. Comput., 21 (2000), pp. 1869–1885.
[17] R. A. WALTERS AND V. CASULLI, *A robust finite element model for hydrostatic surface water flows*, Comm. Numer. Methods Engrg., 14 (1998), pp. 931–940.
[18] O. C. ZIENKIEWICZ AND P. ORTIZ, *A split-characteristic based finite element model for the shallow water equations*, Internat. J. Numer. Methods Fluids, 20 (1995), pp. 1061–1080.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2408–2428

RECONSTRUCTION OF CLOSELY SPACED SMALL INCLUSIONS*

HABIB AMMARI†, HYEONBAE KANG‡, EUNJOO KIM‡, AND MIKYOUNG LIM‡

Abstract. In this paper we establish an explicit asymptotic formula for the steady state voltage perturbations caused by closely spaced small conductivity inhomogeneities. Based on this new formula we design a very effective numerical method to identify the location and some geometric features of these inhomogeneities from a finite number of boundary measurements. The viability of our approach is documented by numerical examples.

Key words. conductivity imaging, closely spaced small inhomogeneities, polarization tensors of multiple inclusions, numerical reconstruction algorithms

AMS subject classifications. 35R30, 35B30

DOI. 10.1137/S0036142903422752

1. Introduction.

The problem of determining interior information about a medium from boundary field measurements is one that is not, in general, well posed. If, however, in advance we have additional structural information about the medium, then we may be able to determine specific features with higher resolution. One particular very promising line of work has been concerned with the reconstruction of small *well-separated* inhomogeneities. While efficient algorithms to determine the location and/or shape of the small inhomogeneities have been developed in that case [3, 5, 6, 7, 8, 10, 17, 18, 23, 28, 4], it remains an interesting open problem to adapt these numerical methods to the close-to-touching case. In this paper, we will concentrate on the reconstruction of conductivity inhomogeneities in a specific setting, namely a bounded domain consisting of a homogeneous conducting background medium in which are embedded *closely spaced* conducting inhomogeneities of small diameter. We believe that our results can be extended, with some modifications, to the case of the Helmholtz equation, the full Maxwell equations, and the Lamé system, as well as to the setting of thin inhomogeneities.

Our objective in this work is threefold. Our first goal is to provide an explicit derivation of the leading order boundary perturbations resulting from the presence of an arbitrary number of closely spaced small conductivity inhomogeneities. This formula expresses the voltage potential in terms of the background potential, a certain Neumann function, and the relevant *polarization tensors of multiple inclusions* and, in a most natural way, generalizes those already derived for a finite set of well-separated small inhomogeneities [1, 10, 16]. Our second goal is to investigate some important properties of these new polarization tensors such as symmetry and positivity. We also estimate their eigenvalues in terms of the total volume of the inclusions and explicitly compute them in the multidisk case. These results do not seem to be available in the literature. In the case of a single inhomogeneity these polarization tensors are reduced

*Received by the editors February 13, 2003; accepted for publication (in revised form) January 19, 2004; published electronically March 31, 2005. The research of the first author was partially supported by ACI Jeunes Chercheurs (0693) from the Ministry of Education and Scientific Research, France and the research of the second author was partially supported by KOSEF 98-0701-03-5 and BK21 at the School of Mathematical Sciences of SNU.

http://www.siam.org/journals/sinum/42-6/42275.html

†Centre de Mathématiques Appliquées, CNRS UMR 7641 and Ecole Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr).

‡School of Mathematical Sciences, Seoul National University, Seoul 151-747, Korea (hkang@math.snu.ac.kr, kej@math.snu.ac.kr, mklim@math.snu.ac.kr).

to the classical Pólya–Szegö polarization tensor which has been extensively studied [10, 16, 21, 25, 26] and higher-order polarization tensors which have been introduced in [1, 2]. It should also be noted that Cheng and Greengard gave in Theorem 2.2 of their interesting paper [11] a solution to the two- and three-disk conductivity problem based on a method of images. Our calculations provide a more general way of solving the multidisk problem.

Our third goal is to apply our explicit asymptotic formula for the purpose of identifying the location and the polarization tensor of closely spaced small inhomogeneities from finitely many current-voltage pairs measured on the boundary.

To fix notation, consider a homogeneous conducting object which occupies a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3. We will assume, for the sake of simplicity, that its conductivity is equal to 1. The background voltage potential, U, is the solution to the boundary value problem

$$\begin{cases} \Delta U = 0 \quad \text{in } \Omega, \\ \dfrac{\partial U}{\partial \nu}\Big|_{\partial\Omega} = g. \end{cases}$$

Here ν denotes the unit outward normal to the domain Ω and g represents the applied boundary current; it belongs to the set $L^2_0(\partial\Omega) = \{g \in L^2(\partial\Omega), \int_{\partial\Omega} g = 0\}$.

Let D denote a set of m closely spaced inhomogeneities inside Ω

$$D = \cup_{l=1}^m D_l := \cup_{l=1}^m (\epsilon B_l + z),$$

where $z \in \Omega$, $\epsilon > 0$ is small, and B_l for $l = 1, \dots, m$ is a bounded Lipschitz domain in $\mathbb{R}^d$. Throughout this paper we suppose that

(H1) the set D is well separated from the boundary $\partial\Omega$ (i.e., $\operatorname{dist}(D, \partial\Omega) > d_0 > 0$);

(H2) there exist positive constants C_1 and C_2 such that

$$C_1 \le \operatorname{diam} B_l \le C_2, \quad \text{and} \quad C_1 \le \operatorname{dist}(B_l, B_{l'}) \le C_2, \quad l \ne l';$$

(H3) and the conductivity of the inhomogeneity D_l for $l = 1, \dots, m$ is equal to some positive constant $k_l \ne 1$.

The conductivity profile of Ω is then

$$\gamma(x) = \begin{cases} k_l, & x \in D_l, \\ 1, & x \in \Omega \setminus \overline{D}. \end{cases} \tag{1.1}$$

The voltage potential in the presence of the set D of conductivity inhomogeneities is denoted u. It is the solution to

$$\begin{cases} \nabla \cdot \left(\chi\left(\Omega \setminus \bigcup_{l=1}^m \overline{D_l} \right) + \sum_{l=1}^m k_l \chi(D_l) \right) \nabla u = 0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu}\Big|_{\partial\Omega} = g. \end{cases}$$

We normalize both the potentials U and u by requiring that $\int_{\partial\Omega} u = \int_{\partial\Omega} U = 0$.

It should be noted that Capdeboscq and Vogelius, in their interesting paper [9], derived a very general representation formula for the boundary voltage perturbations caused by internal conductivity inhomogeneities of small volume fraction. Rather than directly applying this general formula for finding the location and geometric

features of D we derive a more *explicit* one in the particular case of close-to-touching small inhomogeneities from which it is possible to extract *very specific information* about D. The numerical examples presented in this paper show that this explicit formula leads to a *very effective computational algorithm*.

This paper is organized as follows. In section 2, we review some basic facts on the layer potentials of the Laplacian which constitute the basic tools of the present work. In section 3, we give mathematical definitions of the polarization tensors of multiple inclusions. In particular, we rigorously establish some useful properties of these new polarization tensors, carefully study their properties of symmetry and positivity, and estimate their eigenvalues in terms of the total volume of the inclusions. In section 4, we explicitly compute these polarization tensors in the multidisk case. In section 5, we visualize the first-order polarization tensor of multiple inclusions in terms of an equivalent ellipse. In section 6, we rigorously justify the derivation of our asymptotic formula. In section 7, we propose a numerical algorithm based on the asymptotic formula to detect the location and size of closely spaced inclusions. This algorithm makes it possible to find the first-order polarization tensor as well.

2. Layer potentials for the Laplacian. In this section, let us review some well-known properties of the layer potentials for the Laplacian and prove a decomposition formula of the solution u to the following transmission problem:

$$\begin{cases} \nabla \cdot \left(\chi\left(\Omega \setminus \bigcup_{l=1}^{m} \overline{B_l} \right) + \sum_{l=1}^{m} k_l \chi(B_l) \right) \nabla u = 0 \quad \text{in } \mathbb{R}^d, \\ u(x) - H(x) = O(|x|^{1-d}) \quad \text{as } |x| \to \infty, \end{cases} \tag{2.1}$$

where H is a harmonic function in $\mathbb{R}^d$.

The theory of layer potentials has been developed in relation to the boundary value problems. Let D be a bounded domain in $\mathbb{R}^d, d \geq 2$. We assume that ∂D is Lipschitz. Let $\Gamma(x)$ be the fundamental solution of the Laplacian Δ:

$$\Gamma(x) = \begin{cases} \dfrac{1}{2\pi} \ln |x|, & d = 2, \\ \dfrac{1}{(2-d)\omega_d} |x|^{2-d}, & d \geq 3, \end{cases} \tag{2.2}$$

where ω_d is the area of the $(d-1)$-dimensional unit sphere. The single and double layer potentials of the density function ϕ on D are defined by

$$\mathcal{S}_D \phi(x) := \int_{\partial D} \Gamma(x-y) \phi(y) d\sigma(y), \quad x \in \mathbb{R}^d, \tag{2.3}$$

$$\mathcal{D}_D \phi(x) := \int_{\partial D} \frac{\partial}{\partial \nu_y} \Gamma(x-y) \phi(y) d\sigma(y), \quad x \in \mathbb{R}^d \setminus \partial D. \tag{2.4}$$

For a function u defined on $\mathbb{R}^d \setminus \partial D$, we denote

$$\frac{\partial}{\partial \nu^{\pm}} u(x) := \lim_{t \to 0^+} \langle \nabla u(x \pm t\nu_x), \nu_x \rangle, \quad x \in \partial D,$$

if the limit exists. Here ν_x is the outward unit normal to ∂D at x.

The proof of the following trace formula can be found in [14, 15, 24] (for Lipschitz domains, see [27]):

$$\frac{\partial}{\partial \nu^{\pm}} \mathcal{S}_D \phi(x) = \left(\pm \frac{1}{2} I + \mathcal{K}_D^* \right) \phi(x), \quad x \in \partial D, \tag{2.5}$$

$$(\mathcal{D}_D \phi)|_{\pm} = \left(\mp \frac{1}{2} I + \mathcal{K}_D \right) \phi(x), \quad x \in \partial D, \tag{2.6}$$

where

$$\mathcal{K}_D \phi(x) = \frac{1}{\omega_d} \text{p.v.} \int_{\partial D} \frac{\langle y - x, \nu_y \rangle}{|x - y|^d} \phi(y) d\sigma(y)$$

and $\mathcal{K}_D^*$ is the L^2-adjoint of $\mathcal{K}_D$. When ∂D is Lipschitz, $\mathcal{K}_D$ is a singular integral operator and known to be bounded on $L^2(\partial\Omega)$ [12]. Let $L_0^2(\partial D) := \{f \in L^2(\partial D) : \int_{\partial D} f d\sigma = 0\}$. The following results are due to Verchota and Escauriaza, Fabes, and Verchota.

THEOREM 2.1 (see [13], [27]). *$\lambda I - \mathcal{K}_D^*$ is invertible on $L_0^2(\partial D)$ if $|\lambda| \geq \frac{1}{2}$, and for $\lambda \in (-\infty, -\frac{1}{2}] \cup (\frac{1}{2}, \infty)$, $\lambda I - \mathcal{K}_D^*$ is invertible on $L^2(\partial D)$.*

The following theorem will be very useful in the next section.

THEOREM 2.2. *Let H be a harmonic function in $\mathbb{R}^d$ for $d = 2$ or 3. Let u be the solution of the transmission problem (2.1). There are unique functions $\varphi^{(j)} \in L_0^2(\partial B_j)$, $j = 1, \ldots, m$, such that*

$$u(x) = H(x) + \sum_{j=1}^{m} \mathcal{S}_{B_j} \varphi^{(j)}(x). \tag{2.7}$$

The potential $\varphi^{(j)}$, $j = 1, \ldots, m$, satisfy

$$(\lambda_j I - \mathcal{K}_{B_j}^*)\varphi^{(j)} - \sum_{k \neq j} \frac{\partial (\mathcal{S}_{B_k} \varphi^{(k)})}{\partial \nu^{(j)}}\bigg|_{\partial B_j} = \frac{\partial H}{\partial \nu^{(j)}}\bigg|_{\partial B_j} \quad \text{on } \partial B_j, \quad j = 1, \ldots, m, \tag{2.8}$$

where $\nu^{(j)}$ denotes the outward unit normal to ∂B_j and

$$\lambda_j = \frac{k_j + 1}{2(k_j - 1)}, \quad j = 1, \ldots, m.$$

Proof. It is easy to see from (2.5) that u defined by (2.7) and (2.8) is the solution of (2.1). Thus it is enough to show that the integral equation (2.8) has a unique solution.

Let $X := L_0^2(\partial B_1) \times \cdots \times L_0^2(\partial B_m)$. We prove that the operator $T : X \to X$ defined by

$$\begin{aligned} T(\varphi^{(1)}, \ldots, \varphi^{(m)}) &= T_0(\varphi^{(1)}, \ldots, \varphi^{(m)}) + T_1(\varphi^{(1)}, \ldots, \varphi^{(m)}) \\ &:= \left((\lambda_1 I - \mathcal{K}_{B_1}^*)\varphi^{(1)}, \ldots, (\lambda_m I - \mathcal{K}_{B_m}^*)\varphi^{(m)} \right) \\ &\quad - \left(\sum_{k \neq 1} \frac{\partial (\mathcal{S}_{B_k} \varphi^{(k)})}{\partial \nu^{(1)}}\bigg|_{\partial B_1}, \ldots, \sum_{k \neq m} \frac{\partial (\mathcal{S}_{B_k} \varphi^{(k)})}{\partial \nu^{(m)}}\bigg|_{\partial B_m} \right) \end{aligned}$$

is invertible. By Theorem 2.1, T_0 is invertible on X. On the other hand, it is easy to see that T_1 is a compact operator on X. Thus, by the Fredholm alternative, it suffices to show that T is injective on X. If $T(\varphi^{(1)}, \dots, \varphi^{(m)}) = 0$, then $u(x) := \sum_{j=1}^m \mathcal{S}_{B_j}\varphi^{(j)}(x)$, $x \in \mathbb{R}^d$, is the solution of (2.1) with $H = 0$. By the uniqueness of the solution to (2.1), we get $u \equiv 0$. In particular, $\mathcal{S}_{B_j}\varphi^{(j)}$ is smooth across ∂B_j, $j = 1, \dots, m$. Therefore, $\varphi^{(j)} = \frac{\partial(\mathcal{S}_{B_j}\varphi^{(j)})}{\partial\nu^{(j)}}\big|_+ - \frac{\partial(\mathcal{S}_{B_j}\varphi^{(j)})}{\partial\nu^{(j)}}\big|_- = 0$. This completes the proof. ∎

3. Polarization tensors of multiple inclusions. Our aim in this section is to introduce new concepts of polarization tensors of multiple inclusions which generalize the Pólya–Szegö tensor. These concepts are defined in a way analogous to the generalized polarization tensors introduced in [1, 2]. A novel result of this paper is a proof of symmetry and positive-definiteness of these polarization tensors. We also obtain estimations of their eigenvalues in terms of the total volume of the inclusions. These results will turn out to be crucial for our approach to determine the location and some geometric features of closely spaced small conductivity inclusions.

DEFINITION 3.1. *Let $\alpha = (\alpha_1, \dots, \alpha_d), \beta = (\beta_1, \dots, \beta_d) \in \mathbb{N}^d$ be multi-indices. For $l = 1, \dots, m$ let $\varphi_\alpha^{(l)}$ be the solution of*

$$(\lambda_l I - \mathcal{K}_{B_l}^*)\varphi_\alpha^{(l)} - \sum_{k\neq l} \frac{\partial(\mathcal{S}_{B_k}\varphi_\alpha^{(k)})}{\partial\nu^{(l)}}\bigg|_{\partial B_l} = \frac{\partial x^\alpha}{\partial\nu^{(l)}}\bigg|_{\partial B_l} \quad \text{on } \partial B_l. \tag{3.1}$$

Then the polarization tensor $M = (m_{\alpha\beta})$ is defined to be

$$m_{\alpha\beta} = \sum_{l=1}^m \int_{\partial B_l} x^\beta \varphi_\alpha^{(l)}(x) d\sigma. \tag{3.2}$$

If $|\alpha| = |\beta| = 1$, we denote $m_{\alpha\beta}$ by m_{ij}, $i, j = 1, \dots, d$. We call m_{ij} the first-order polarization tensor.

THEOREM 3.2. *The polarization tensor M is symmetric. More precisely, if a_α and b_β are constants such that $\sum_\alpha a_\alpha y^\alpha$ and $\sum_\beta b_\beta y^\beta$ are harmonic polynomials, then*

$$\sum_{\alpha,\beta} a_\alpha b_\beta m_{\alpha\beta} = \sum_{\alpha,\beta} a_\alpha b_\beta m_{\beta\alpha}. \tag{3.3}$$

Proof. Put $f(y) := \sum_\alpha a_\alpha y^\alpha$, $g(y) := \sum_\beta b_\beta y^\beta$, $\varphi^{(l)} := \sum_\alpha a_\alpha \varphi_\alpha^{(l)}$, and $\psi^{(l)} := \sum_\beta b_\beta \varphi_\beta^{(l)}$. Then one can easily see that

$$\sum_{\alpha,\beta} a_\alpha b_\beta m_{\alpha\beta} = \sum_{l=1}^m \int_{\partial B_l} g\varphi^{(l)} d\sigma \quad \text{and} \quad \sum_{\alpha,\beta} a_\alpha b_\beta m_{\beta\alpha} = \sum_{l=1}^m \int_{\partial B_l} f\psi^{(l)} d\sigma.$$

We also put

$$\Phi(x) := \sum_{l=1}^m \mathcal{S}_{B_l}\varphi^{(l)} \quad \text{and} \quad \Psi(x) := \sum_{l=1}^m \mathcal{S}_{B_l}\psi^{(l)}.$$

From the definition of $\varphi_\alpha^{(l)}$, one can readily get

$$k_j \frac{\partial(f+\Phi)}{\partial\nu^{(j)}}\bigg|_- = \frac{\partial(f+\Phi)}{\partial\nu^{(j)}}\bigg|_+ \quad \text{on } \partial B_j, \tag{3.4}$$

and the same relation holds for $g+\Psi$. From (3.1) we obtain

$$\begin{aligned}\frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu^{(l)}}\Big|_+ - k_l\frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu^{(l)}}\Big|_- &= \sum_\alpha a_\alpha\left[\frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)}_\alpha)}{\partial\nu^{(l)}}\Big|_+ - k_l\frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)}_\alpha)}{\partial\nu^{(l)}}\Big|_-\right]\\ &= (k_l-1)\sum_\alpha a_\alpha\frac{\partial}{\partial\nu^{(l)}}\left[x^\alpha+\sum_{j\neq l}\mathcal{S}_{B_j}\varphi^{(j)}_\alpha\right]\\ &= (k_l-1)\frac{\partial}{\partial\nu^{(l)}}\left[f+\sum_{j\neq l}\mathcal{S}_{B_j}\varphi^{(j)}_\alpha\right].\end{aligned}$$

Thus, it follows from (3.4) that

$$\varphi^{(l)} = \frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu^{(l)}}\Big|_+ - \frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu^{(l)}}\Big|_- = (k_l-1)\frac{\partial(f+\Phi)}{\partial\nu^{(l)}}\Big|_- \quad \text{on } \partial B_l. \tag{3.5}$$

Therefore, we get

$$\begin{aligned}\sum_{\alpha,\beta}a_\alpha b_\beta m_{\alpha\beta} &= \sum_{l=1}^m(k_l-1)\int_{\partial B_l} g\frac{\partial(f+\Phi)}{\partial\nu}\Big|_- d\sigma\\ &= \sum_{l=1}^m(k_l-1)\int_{\partial B_l}(g+\Psi)\frac{\partial(f+\Phi)}{\partial\nu}\Big|_- d\sigma - \sum_{l=1}^m(k_l-1)\int_{\partial B_l}\Psi\frac{\partial(f+\Phi)}{\partial\nu}\Big|_- d\sigma\\ &= \sum_{l=1}^m(k_l-1)\int_{\partial B_l}(g+\Psi)\frac{\partial(f+\Phi)}{\partial\nu}\Big|_- d\sigma\\ &\quad - \sum_{l=1}^m\int_{\partial B_l}\Psi\left[\frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu}\Big|_+ - \frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu}\Big|_-\right]d\sigma.\end{aligned} \tag{3.6}$$

Observe now that

$$\begin{aligned}&\sum_{l=1}^m\int_{\partial B_l}\Psi\frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu}\Big|_+ d\sigma = \sum_{j,l}\int_{\partial B_l}\mathcal{S}_{B_j}\psi^{(j)}\frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu}\Big|_+ d\sigma\\ &\quad = -\sum_{l=1}^m\int_{\mathbb{R}^d\setminus\overline{B}_l}\nabla\mathcal{S}_{B_l}\psi^{(l)}\cdot\nabla\mathcal{S}_{B_l}\varphi^{(l)}dx - \frac{1}{2}\sum_{l\neq j}\int_{\mathbb{R}^d\setminus\overline{B_l\cup B_j}}\nabla\mathcal{S}_{B_j}\psi^{(j)}\cdot\nabla\mathcal{S}_{B_l}\varphi^{(l)}dx,\end{aligned}$$

and

$$\begin{aligned}\sum_{l=1}^m\int_{\partial B_l}\Psi\frac{\partial(\mathcal{S}_{B_l}\varphi^{(l)})}{\partial\nu}\Big|_- d\sigma &= \sum_{j,l}\int_{B_l}\nabla\mathcal{S}_{B_j}\psi^{(j)}\cdot\nabla\mathcal{S}_{B_l}\varphi^{(l)}dx\\ &= \sum_{l=1}^m\int_{B_l}\nabla\mathcal{S}_{B_l}\psi^{(l)}\cdot\nabla\mathcal{S}_{B_l}\varphi^{(l)}dx + \frac{1}{2}\sum_{j\neq l}\int_{B_l\cup B_j}\nabla\mathcal{S}_{B_j}\psi^{(j)}\cdot\nabla\mathcal{S}_{B_l}\varphi^{(l)}dx,\end{aligned}$$

to finally arrive at

$$\begin{aligned}\sum_{\alpha,\beta}a_\alpha b_\beta m_{\alpha\beta} &= \sum_{l=1}^m(k_l-1)\langle(g+\Psi),(f+\Phi)\rangle_{B_l} + \sum_{l=1}^m\langle\mathcal{S}_{B_l}\psi^{(l)},\mathcal{S}_{B_l}\varphi^{(l)}\rangle_{\mathbb{R}^d}\\ &\quad + \frac{1}{2}\sum_{j\neq l}\langle\mathcal{S}_{B_j}\psi^{(j)},\mathcal{S}_{B_l}\varphi^{(l)}\rangle_{\mathbb{R}^d}.\end{aligned} \tag{3.7}$$

Here, the notation $\langle u, v\rangle_D := \int_D \nabla u \cdot \nabla v dx$ has been used. The symmetry (3.3) follows immediately from (3.7) and the proof is complete. ∎

THEOREM 3.3. *Suppose that either $k_l - 1 > 0$ or $k_l - 1 < 0$ for all $l = 1, \dots, m$. Let*

$$\kappa := \max_{1 \le l \le m} \left| 1 - \frac{1}{k_l} \right|.$$

For any a_α such that $\sum_\alpha a_\alpha y^\alpha$ is harmonic,

$$\left| \sum_{\alpha,\beta} a_\alpha a_\beta m_{\alpha\beta} \right| \ge \frac{|\kappa - 1|}{m+1} \sum_{l=1}^m |k_l - 1| \int_{B_l} \left| \nabla \left(\sum_\alpha a_\alpha y^\alpha \right) \right|^2 dy. \tag{3.8}$$

In particular, if $k_l - 1 > 0$ (resp., < 0) for all $l = 1, \dots, m$, then (m_{ij}) is positive (resp., negative) definite and if $\sum_{i=1}^d a_i^2 = 1$, then

$$\left| \sum_{i,j} a_i a_j m_{ij} \right| \ge \frac{|\kappa - 1|}{m+1} \sum_{l=1}^m |k_l - 1||B_l|. \tag{3.9}$$

Here $|B|$ is the volume of B.

Proof. Suppose that either $k_l - 1 > 0$ or $k_l - 1 < 0$ for all $l = 1, \dots, m$. Define the quadratic form $Q_D(u)$ by

$$Q_D(u) := \langle u, u\rangle_D.$$

It then follows from (3.7) that

$$\begin{aligned} \sum_{\alpha,\beta} a_\alpha a_\beta m_{\alpha\beta} &= \sum_{l=1}^m (k_l - 1) Q_{B_l}(f + \Phi) + \sum_{l=1}^m Q_{\mathbb{R}^d}(\mathcal{S}_{B_l} \varphi^{(l)}) \\ &\quad + \frac{1}{2} \sum_{j \ne l} \langle \mathcal{S}_{B_j} \varphi^{(j)}, \mathcal{S}_{B_l} \varphi^{(l)} \rangle_{\mathbb{R}^d} \\ &= \sum_{l=1}^m (k_l - 1) Q_{B_l}(f + \Phi) + Q_{\mathbb{R}^d}(\Phi). \end{aligned} \tag{3.10}$$

On the other hand, because of (3.4), we get

$$(k_j - 1) \frac{\partial f}{\partial \nu^{(j)}} = \frac{\partial \Phi}{\partial \nu^{(j)}}\bigg|_+ - k_j \frac{\partial \Phi}{\partial \nu^{(j)}}\bigg|_- \quad \text{on } \partial B_j, \quad j = 1, \dots, d.$$

Thus, it follows from (3.5) that

$$\begin{aligned} \sum_{\alpha,\beta} a_\alpha a_\beta m_{\alpha\beta} &= \sum_{l=1}^m (k_l - 1) \int_{\partial B_l} f \frac{\partial (f + \Phi)}{\partial \nu}\bigg|_- d\sigma \\ &= \sum_{l=1}^m (k_l - 1) Q_{B_l}(f) + \sum_{l=1}^m (k_l - 1) \int_{\partial B_l} \frac{\partial f}{\partial \nu} \Phi d\sigma \\ &= \sum_{l=1}^m (k_l - 1) Q_{B_l}(f) + \sum_{l=1}^m \int_{\partial B_l} \frac{\partial \Phi}{\partial \nu}\bigg|_+ \Phi d\sigma - \sum_{l=1}^m k_l \int_{\partial B_l} \frac{\partial \Phi}{\partial \nu}\bigg|_- \Phi d\sigma \\ &= \sum_{l=1}^m (k_l - 1) Q_{B_l}(f) - \sum_{l=1}^m Q_{\mathbb{R}^d}(\Phi) - \sum_{l=1}^m (k_l - 1) Q_{B_l}(\Phi). \end{aligned} \tag{3.11}$$

By equating (3.10) and (3.11) we have

$$\begin{aligned}&\sum_{l=1}^{m}(k_l-1)Q_{B_l}(f+\Phi)+Q_{\mathbb{R}^d}(\Phi)\\&=\sum_{l=1}^{m}(k_l-1)Q_{B_l}(f)-\sum_{l=1}^{m}Q_{\mathbb{R}^d}(\Phi)-\sum_{l=1}^{m}(k_l-1)Q_{B_l}(\Phi).\end{aligned}\tag{3.12}$$

Since the left-hand side of (3.12) is positive, one can claim that

$$\sum_{l=1}^{m}(k_l-1)Q_{B_l}(f)\geq\sum_{l=1}^{m}k_lQ_{B_l}(\Phi).\tag{3.13}$$

It also follows from (3.12) that

$$Q_{\mathbb{R}^d}(\Phi)=\frac{1}{m+1}\sum_{l=1}^{m}(k_l-1)\left[Q_{B_l}(f)-Q_{B_l}(f+\Phi)-Q_{B_l}(\Phi)\right].\tag{3.14}$$

Substituting (3.14) into (3.10), we obtain

$$\sum_{\alpha,\beta}a_\alpha a_\beta m_{\alpha\beta}=\frac{m}{m+1}\sum_{l=1}^{m}(k_l-1)Q_{B_l}(f+\Phi)+\frac{1}{m+1}\sum_{l=1}^{m}(k_l-1)\left[Q_{B_l}(f)-Q_{B_l}(\Phi)\right],\tag{3.15}$$

and hence

$$\sum_{\alpha,\beta}a_\alpha a_\beta m_{\alpha\beta}\geq\frac{1}{m+1}\sum_{l=1}^{m}(k_l-1)\left[Q_{B_l}(f)-Q_{B_l}(\Phi)\right].\tag{3.16}$$

By (3.13) we get

$$\begin{aligned}\sum_{l=1}^{m}(k_l-1)Q_{B_l}(\Phi)&=\sum_{l=1}^{m}\frac{(k_l-1)}{k_l}k_lQ_{B_l}(\Phi)\\&\leq\kappa\sum_{l=1}^{m}k_lQ_{B_l}(\Phi)\leq\kappa\sum_{l=1}^{m}(k_l-1)Q_{B_l}(f),\end{aligned}$$

and hence (3.8) follows immediately from (3.16). This completes the proof. ∎

4. Explicit formulae for the polarization tensors of multiple disks. In this section, we explicitly compute the solution $\varphi^{(l)}$ of the integral equation (2.8) in the case where all of the domains B_l are two-dimensional disks. Let $B_l=B(z_l,r_l)$ be the disk with center z_l and radius r_l for $l=1,\dots,m$. Let R_l, $l=1,\dots,m$, be the reflection with respect to the disk B_l, i.e.,

$$R_l(x):=\frac{r_l^2(x-z_l)}{|x-z_l|^2}+z_l.$$

We also define the reflection of a function f by

$$(R_lf)(x)=f(R_l(x)),\quad x\in\mathbb{R}^2,\quad l=1,\dots,m.$$

The following lemma will be useful later.

LEMMA 4.1. *For a function u harmonic in $\overline{B_l}$, we have*

$$\mathcal{S}_{B_l}\left(\frac{\partial u}{\partial \nu^{(l)}}\Big|_{\partial B_l}\right)(x) = -\frac{1}{2}R_l u(x) + \frac{1}{2}u(z_l) \quad \text{for } x \in \mathbb{R}^2 \setminus \overline{B_l}. \tag{4.1}$$

Proof. By (2.5), we have

$$\frac{\partial}{\partial \nu}\mathcal{S}_{B_l}\left(\frac{\partial u}{\partial \nu^{(l)}}\Big|_{\P B_l}\right)(x) = \left(\frac{1}{2}I + \mathcal{K}_{B_l}^*\right)\left(\frac{\partial u}{\partial \nu^{(l)}}\Big|_{\partial B_l}\right) + (x).$$

Since B_l is a disk and $\int_{\partial B_l} \frac{\partial u}{\partial \nu} d\sigma = 0$, one can get that $\mathcal{K}_{B_l}^*(\frac{\partial u}{\partial \nu^{(l)}}\big|_{\partial B_l}) = 0$ on ∂B_l. We refer the reader to [19] for a proof of this fact. Therefore, we get

$$\frac{\partial}{\partial \nu}\mathcal{S}_{B_l}\left(\frac{\partial u}{\partial \nu^{(l)}}\Big|_{\partial B_l}\right)(x) = \frac{1}{2}\frac{\partial u}{\partial \nu^{(l)}}\Big|_{\partial B_l} + (x),$$

and thus

$$\mathcal{S}_{B_l}\left(\frac{\partial u}{\partial \nu^{(l)}}\Big|_{\partial B_l}\right)(x) = -\frac{1}{2}R_l u(x) + C$$

for some constant C. Since $\int_{\partial B_l} \frac{\partial u}{\partial \nu^{(l)}} d\sigma = 0$ and hence $\mathcal{S}_{B_l}(\frac{\partial u}{\partial \nu^{(l)}}\big|_{\partial B_l})(x) \to 0$ as $|x| \to \infty$, we have $C = (1/2)u(z_l)$. This completes the proof. ∎

Our main result in this section is the following.

THEOREM 4.2. *For $l = 1, \dots, m$, let*

$$S_l = \{\Theta = (k_1, \dots, k_n) \mid n \in \mathbb{N},\ k_i \in \{1, \dots, m\} \text{ such that } k_1 \neq l \text{ and } k_i \neq k_{i+1}\}.$$

For $\Theta = (k_1, \dots, k_n) \in S_l$, let

$$R_\Theta = R_{k_1} R_{k_2} \cdots R_{k_n} \text{ and } \Lambda_\Theta = \prod_{i=1}^{n}\left(-\frac{1}{2\lambda_{k_i}}\right).$$

Then, for a given harmonic function H, the solution of (2.8) *is given by*

$$\varphi^{(l)} = \frac{1}{\lambda_l}\sum_{\Theta \in S_l} \Lambda_\Theta \frac{\partial}{\partial \nu_l}(R_\Theta H)\big|_{\partial B_l} + \frac{1}{\lambda_l}\frac{\partial H}{\partial \nu_l}\Big|_{\partial B_l}, \quad l = 1, \dots, m, \tag{4.2}$$

provided that

$$\min_{1 \le i \neq j \le m} dist(B_i, B_j) > (\sqrt{m-1} - 1)\max_{1 \le i \le m} r_i. \tag{4.3}$$

The series in (4.2) *converges absolutely.*

Proof. We first prove that the series in (4.2) converges absolutely on ∂B_l. Observe that

$$|\nabla(R_\Theta H)(x)| \le |R_\Theta \nabla H(x)| \prod_{i=1}^{n} |DR_{k_i}(R_{k_{i-1}} \dots R_{k_1}(x))|. \tag{4.4}$$

Assuming (4.3) and using

$$|DR_j(x)| \leq \frac{r_j^2}{|x - z_j|^2}, \quad x \in \mathbb{R}^2 \setminus B_j, \quad j = 1, \dots, m, \tag{4.5}$$

it follows from (4.4) that for $x \in \partial B_l$ we have

$$|\nabla(R_\Theta H)(x)| \leq M \cdot \prod_{i=1}^{n} \frac{r_{k_i}^2}{(d + r_{k_i})^2} \leq M \cdot \left(\frac{r_{max}}{(d + r_{max})}\right)^{2n} < M \left(\frac{s}{m-1}\right)^n \tag{4.6}$$

for some $s < 1$, where

$$d = \min_{1 \leq i \neq j \leq m} \operatorname{dist}(B_i, B_j), \quad r_{max} = \max_{1 \leq i \leq m} r_i, \quad \text{and } M = \|\nabla H\|_{L^\infty(\bigcup_{k=1}^m \overline{B_k})}.$$

Note that the number of those Θ's which have n components is $(m-1)^n$. It can be deduced from (4.6) that for $x \in \partial B_l$,

$$\sum_{\Theta \in S_l} \left| \Lambda_\Theta \frac{\partial}{\partial \nu_l} (R_\Theta H)(x) \right| \leq M \sum_{n=1}^{\infty} \left(\frac{s}{m-1}\right)^n (m-1)^n < C$$

for some constant C independent of x.

We now prove that $\varphi^{(l)}$ satisfies (2.8). Let us first observe the following: for each $l = 1, \dots, m$,

$$\cup_{k \neq l} \{(k, \Theta),\ (k) \mid \Theta \in S_k\} = S_l. \tag{4.7}$$

Recalling that $\mathcal{K}_{B_l}^* \varphi^{(l)} = 0$, $l = 1, \dots, m$, and using (4.1), (4.2), and (4.7), we arrive at

$$\begin{aligned}
\sum_{k \neq l} \frac{\partial(\mathcal{S}_{B_k} \varphi^{(k)})}{\partial \nu^{(l)}}\Big|_{B_l} &= \sum_{k \neq l} \frac{\partial}{\partial \nu^{(l)}} \left(\frac{-1}{2\lambda_k} R_k \left[\sum_{\Theta \in S_k} \Lambda_\Theta (R_\Theta H) + H \right] \right) \\
&= \sum_{\Theta \in S_l} \Lambda_\Theta \frac{\partial}{\partial \nu^{(l)}} (R_\Theta H) \\
&= \lambda_l \varphi^{(l)} - \frac{\partial H}{\partial \nu^{(l)}}\Big|_{\partial B_l},
\end{aligned}$$

which is exactly the desired result. ∎

As an immediate application of the above theorem we obtain the following explicit form of the first-order polarization tensor.

THEOREM 4.3. *Suppose $d = 2$. The first-order polarization tensor m_{ij} is given by*

$$m_{ij} = \sum_{l=1}^{m} |B_l| \frac{1}{\lambda_l} \left[\sum_{\Theta \in S_l} \Lambda_\Theta \frac{\partial}{\partial x_j} (R_\Theta(x_i))(z_l) + \delta_{ij} \right], \quad i, j = 1, 2. \tag{4.8}$$

Proof. Let $H(x) = x_i$ and $\varphi_i^{(l)}$ be the corresponding solution of (2.8). Then by (4.2), we have

$$\varphi_i^{(l)} = \frac{1}{\lambda_l} \frac{\partial}{\partial \nu^{(l)}} \left[\sum_{\Theta \in S_l} \Lambda_\Theta (R_\Theta(x_i)) + H \right], \quad i = 1, 2,\ l = 1, \dots, m.$$

It then follows from the divergence theorem and the mean value property of harmonic functions that

$$\begin{aligned}\int_{\partial B_l} x_j \varphi_i^{(l)} d\sigma &= \frac{1}{\lambda_l}\left[\sum_{\Theta\in S_l}\Lambda_\Theta \int_{\partial B_l} x_j \frac{\partial}{\partial \nu^{(l)}}(R_\Theta(x_i))(x)d\sigma + \int_{\partial B_l} x_j \frac{\partial}{\partial \nu^{(l)}} x_i d\sigma\right]\\ &= \frac{1}{\lambda_l}\left[\sum_{\Theta\in S_l}\Lambda_\Theta \int_{B_l} \frac{\partial}{\partial x_j}(R_\Theta(x_i))(x)d\sigma + \delta_{ij}|B_l|\right]\\ &= |B_l|\frac{1}{\lambda_l}\left[\sum_{\Theta\in S_l}\Lambda_\Theta \frac{\partial}{\partial x_j}(R_\Theta(x_i))(z_l) + \delta_{ij}\right].\end{aligned}$$

Thus we get the explicit expression (4.8) as desired. ∎

Let us now write formulae (4.2) and (4.8) in a more explicit way assuming that there are only two inclusions. We note that in this case the assumption (4.3) is trivially fulfilled. If $m = 2$, then R_Θ for $\Theta \in S_1$ takes the form

$$R_\Theta = (R_2R_1)^k R_2 R_1^n \quad \text{for some } k = 0, 1, \dots, \text{ and } n = 0, 1, \tag{4.9}$$

and R_Θ for $\Theta \in S_2$ takes the form

$$R_\Theta = (R_1R_2)^k R_1 R_2^n \quad \text{for some } k = 0, 1, \dots, \text{ and } n = 0, 1. \tag{4.10}$$

Here $R_j^0 = I$, $j = 1, 2$.

COROLLARY 4.4. *If $m = 2$, then*

$$\begin{aligned}\varphi^{(1)} &= \frac{1}{\lambda_1}\sum_{k=0}^{\infty}\frac{1}{(4\lambda_1\lambda_2)^k}\frac{\partial}{\partial\nu^{(1)}}\left[(R_2R_1)^k\left(I - \frac{1}{2\lambda_2}R_2\right)H\right]\Big|_{\partial B_1},\\ \varphi^{(2)} &= \frac{1}{\lambda_2}\sum_{k=0}^{\infty}\frac{1}{(4\lambda_1\lambda_2)^k}\frac{\partial}{\partial\nu^{(2)}}\left[(R_1R_2)^k\left(I - \frac{1}{2\lambda_1}R_1\right)H\right]\Big|_{\partial B_2}.\end{aligned} \tag{4.11}$$

Proof. It follows from (4.2), (4.9), and (4.10) that

$$\begin{aligned}\varphi^{(1)} &= \frac{1}{\lambda_1}\frac{\partial}{\partial\nu^{(1)}}\left[\sum_{k=0}^{\infty}\frac{1}{(4\lambda_1\lambda_2)^k}(R_2R_1)^k\left(-\frac{1}{2\lambda_2}R_2 + \frac{1}{4\lambda_1\lambda_2}R_2R_1\right)H + H\right]\Big|_{\partial B_1},\\ \varphi^{(2)} &= \frac{1}{\lambda_2}\frac{\partial}{\partial\nu^{(2)}}\left[\sum_{k=0}^{\infty}\frac{1}{(4\lambda_1\lambda_2)^k}(R_1R_2)^k\left(-\frac{1}{2\lambda_1}R_1 + \frac{1}{4\lambda_1\lambda_2}R_1R_2\right)H + H\right]\Big|_{\partial B_2}.\end{aligned}$$

By rearranging the summations, we get (4.11). ∎

COROLLARY 4.5. *Let $m = 2$. Suppose that the centers of the disks B_1 and B_2 are on the x_1-axis. Then the polarization tensor m_{ij} is given by*

$$
\begin{aligned}
&m_{12} = m_{21} = 0,\\
&m_{11} = \frac{|B_1|}{\lambda_1} + \frac{|B_2|}{\lambda_2}\\
&+ \frac{|B_1|}{\lambda_1}\sum_{k=0}^{\infty}\frac{1}{(4\lambda_1\lambda_2)^k}\left[(R_2R_1)^k\left(\frac{1}{2\lambda_2}g_2 + \frac{1}{4\lambda_1\lambda_2}R_2(g_1)g_2\right)\prod_{i=0}^{k-1}(R_2R_1)^i(R_2(g_1)g_2)\right](z_1)\\
&+ \frac{|B_2|}{\lambda_2}\sum_{k=0}^{\infty}\frac{1}{(4\lambda_1\lambda_2)^k}\left[(R_1R_2)^k\left(\frac{1}{2\lambda_1}g_1 + \frac{1}{4\lambda_1\lambda_2}R_1(g_2)g_1\right)\prod_{i=0}^{k-1}(R_1R_2)^i(R_1(g_2)g_1)\right](z_2),\\
&m_{22} = \frac{|B_1|}{\lambda_1} + \frac{|B_2|}{\lambda_2}\\
&+ \frac{|B_1|}{\lambda_1}\sum_{k=0}^{\infty}\frac{1}{(4\lambda_1\lambda_2)^k}\left[(R_2R_1)^k\left(-\frac{1}{2\lambda_2}g_2 + \frac{1}{4\lambda_1\lambda_2}R_2(g_1)g_2\right)\prod_{i=0}^{k-1}(R_2R_1)^i(R_2(g_1)g_2)\right](z_1)\\
&+ \frac{|B_2|}{\lambda_2}\sum_{k=0}^{\infty}\frac{1}{(4\lambda_1\lambda_2)^k}\left[(R_1R_2)^k\left(-\frac{1}{2\lambda_1}g_1 + \frac{1}{4\lambda_1\lambda_2}R_1(g_2)g_1\right)\prod_{i=0}^{k-1}(R_1R_2)^i(R_1(g_2)g_1)\right](z_2),
\end{aligned}
$$

where the functions g_1 and g_2 are defined by

$$g_j(x) := \frac{r_j^2}{|x - z_j|^2}, \quad x \in \mathbb{R}^2 \setminus \overline{B_j},\ j = 1, 2.$$

Proof. By Theorem 4.3, (4.9), and (4.10), we have

$$
\begin{aligned}
m_{ij} =& \frac{|B_1|}{\lambda_1}\left[\sum_{k=0}^{\infty}(4\lambda_1\lambda_2)^{-k}\frac{\partial}{\partial x_j}\left((R_2R_1)^k\left(-\frac{1}{2\lambda_2}R_2 + \frac{1}{4\lambda_1\lambda_2}R_2R_1\right)(x_i)\right)(z_1) + \delta_{ij}\right]\\
&+ \frac{|B_2|}{\lambda_2}\left[\sum_{k=0}^{\infty}(4\lambda_1\lambda_2)^{-k}\frac{\partial}{\partial x_j}\left((R_1R_2)^k\left(-\frac{1}{2\lambda_1}R_1 + \frac{1}{4\lambda_1\lambda_2}R_1R_2\right)(x_i)\right)(z_2) + \delta_{ij}\right].
\end{aligned}
$$

Easy computations show that for x on the x_1-axis,

$$DR_{B_j}(x) = g_j(x)\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad j = 1, 2,$$

and

$$\nabla R_j f(x) = (R_j \nabla f)(x) g_j(x)\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad j = 1, 2.$$

Therefore, we get for $H = x_i$

$$
\begin{aligned}
&\nabla\Big((R_2R_1)^kR_2(H)\Big)(x) = \nabla H \cdot \left[(R_2R_1)^k g_2(x)\prod_{i=0}^{k-1}(R_2R_1)^i(R_2(g_1)g_2)(x)\right]\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},\\
&\nabla\Big((R_2R_1)^kR_2R_1(H)\Big)(x) = \nabla H \cdot \left[\prod_{i=0}^{k}(R_2R_1)^i(R_2(g_1)g_2)(x)\right]\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.
\end{aligned}
$$

One can get similar formulae for $\nabla((R_1R_2)^kR_1(H))$ and $\nabla((R_1R_2)^kR_1R_2(H))$. By substituting these formulae into the first equation of the proof, we obtain Corollary 4.5. ∎

5. Representation by equivalent ellipses. Let m_{ij} be the first-order polarization tensor of the inclusions $\cup_{l=1}^m B_l$. We define the *overall conductivity* $\bar{k}$ of $B = \cup_{l=1}^m B_l$ by

$$\frac{\bar{k}-1}{\bar{k}+1}\sum_{l=1}^{m}|B_l| := \sum_{l=1}^{m}\frac{k_l-1}{k_l+1}|B_l|, \tag{5.1}$$

and its *center* $\bar{z}$ by

$$\frac{\bar{k}-1}{\bar{k}+1}\bar{z}\sum_{l=1}^{m}|B_l| = \sum_{l=1}^{m}\frac{k_l-1}{k_l+1}\int_{B_l} x dx. \tag{5.2}$$

Note that if k_l is the same for all l, then $\bar{k} = k_l$ and $\bar{z}$ is the center of mass of B. Figure 7.1, at the end of the last section, shows that the center $\bar{z}$ is a very good match with the center reconstructed by the boundary measurements.

In this section we represent and visualize the multiple inclusions $\cup_{l=1}^m B_l$ by means of an ellipse, $\mathcal{E}$, of center $\bar{z}$ with the same polarization tensor. We call $\mathcal{E}$ the *equivalent ellipse* of $\cup_{l=1}^m B_l$. It will turn out that if we consider the problem of determining the collection D of the closely spaced small inhomogeneities D_l, the only information of real interest that could be reconstructed from boundary measurements is exactly $\mathcal{E}$. This will be shown in the last section.

At this point let us review a method to find an ellipse from a given first-order polarization tensor. This method is due to Brühl et al. [8]. Let $\mathcal{E}'$ be an ellipse whose focal line is on either the x_1- or the x_2-axis. We suppose that its semimajor axis is of length a and its semiminor axis is of length b. Let $\mathcal{E} = R\mathcal{E}'$, where $R = \left(\begin{smallmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{smallmatrix}\right)$. Let M be the polarization tensor of $\mathcal{E}$. We want to recover a, b, and θ from M knowing the conductivity $k = \bar{k}$.

The polarization tensor M' for $\mathcal{E}'$ takes the form

$$M' = (k-1)|\mathcal{E}'|\begin{pmatrix}\frac{a+b}{a+kb} & 0 \\ 0 & \frac{a+b}{b+ka}\end{pmatrix}, \tag{5.3}$$

and that of $\mathcal{E}$ is given by $M = RM'R^T$. Suppose that the eigenvalues of M are λ_1 and λ_2 and the corresponding eigenvectors of unit length are $(e_{11}, e_{12})^T$ and $(e_{21}, e_{22})^T$. Then it is shown in [8] that

$$a = \sqrt{\frac{p}{\pi q}}, \quad b = \sqrt{\frac{pq}{\pi}}, \quad \theta = \arctan\frac{e_{21}}{e_{11}}, \tag{5.4}$$

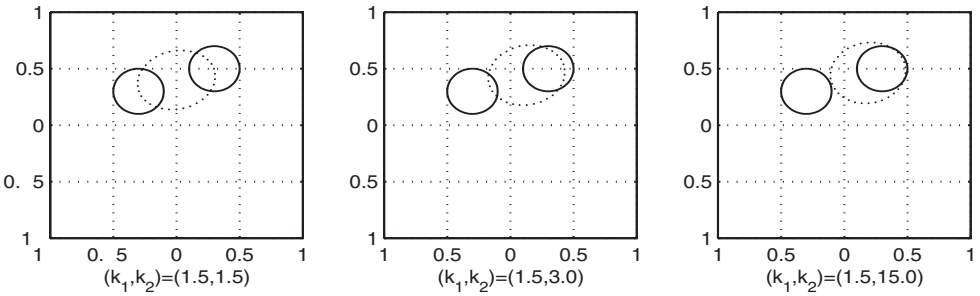
where

$$\frac{1}{p} = \frac{k-1}{k+1}\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right) \quad \text{and} \quad q = \frac{\lambda_2 - k\lambda_1}{\lambda_1 - k\lambda_2}. \tag{5.5}$$

We now show some numerical examples of equivalent ellipses. We represent the set of inclusions $B = \cup_{l=1}^m B_l$ by an equivalent ellipse of center $\bar{z}$ and conductivity $\bar{k}$. We assume that the inclusion B_l takes the following form:

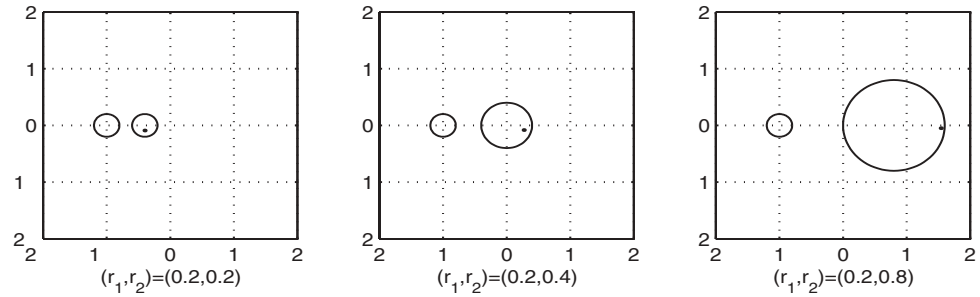
$$\partial B_l = \{\big(a_0^l + a_1^l\cos(t) + a_2^l\cos(2t), b_0^l + b_1^l\sin(t) + b_2^l\sin(2t)\big)\,|0 \le t < 2\pi\}.$$

In order to evaluate the first-order polarization tensor of multiple inclusions, we solve



$a_0^i, a_1^i, a_2^i, b_0^i, b_1^i, b_2^i$	k_i	$\bar{k}$	a	b	θ	$\bar{z}$
-0.3,0.2,0, 0.3,0.2,0 0.3,0.2,0, 0.5,0.2,0	1.5 1.5	1.5	0.313	0.256	0.322	(-0.000, 0.400)
	1.5 3	2.077	0.307	0.261	0.322	(0.129, 0.443)
	1.5 15	3.324	0.301	0.266	0.322	(0.188, 0.463)

FIG. 5.1. *When the two disks have the same radius and the conductivity of the one on the right-hand side is increasing, the equivalent ellipse is moving toward the right inclusion. In the table,* $\bar{k}$ *and* $\bar{z}$ *are the overall conductivity and center defined by* (5.1) *and* (5.2) *and* a, b, θ *are the semiaxes' lengths and angle of orientation measured in radian of the equivalent ellipse.*



k_i	$a_0^i, a_1^i, a_2^i, b_0^i, b_1^i, b_2^i$	$\bar{k}$	a	b	θ	$\bar{z}$
1.5 1.5	-1,0.2,0, 0,0.2,0 -0.4,0.2,0, 0,0.2,0	1.5	0.317	0.254	0	(-0.700, 0.000)
	-1,0.2,0, 0,0.2,0 0,0.2,0, 0,0.2,0	1.5	0.478	0.420	0	(-0.200, 0.000)
	-1,0.2,0, 0,0.2,0 0.8,0.2,0, 0,0.2,0	1.5	0.844	0.806	0	(0.694, 0.000)

FIG. 5.2. *When the conductivities of the two disks are the same and the radius of the disk on the right-hand side is increasing, the equivalent ellipse is moving toward the right inclusion.*

the integral equation (6.4) with $H(x) = x_i$ to find $\varphi_i^{(l)}$ for $i = 1, 2$ and $l = 1, \ldots, m$ and then calculate $m_{ij} = \sum_{l=1}^{m} \int_{\partial B_l} x_j \varphi_i^{(l)}(x) d\sigma$.

Figures 5.1 and 5.2 show how the equivalent ellipse changes as the conductivities and the sizes of the inhomogeneities B_l vary. The solid line represents the actual inhomogeneities and the dash lines are the effective ellipses. Figure 7.1 exhibits equivalent ellipses of other configurations.

6. Derivation of the asymptotic formula. In the remainder of this paper, we consider the problem of determining the location and the polarization tensor of a set of closely spaced small inclusions from boundary measurements. Our algorithm makes use of a new asymptotic expansion of the voltage potentials in the presence of a set of close-to-touching small conductivity inhomogeneities. Since this formula can be obtained following the arguments presented in [1], we only outline its derivation leaving the details to the reader.

In this section, we use the notation stated at the beginning of this paper and suppose that assumptions (H1), (H2), and (H3) hold.

Based on the arguments given in [19, 20], the following theorem was proved in [22].

THEOREM 6.1. *The solution u of the problem*

$$\begin{cases} \nabla \cdot \left(\chi\left(\Omega \setminus \bigcup_{j=1}^{m} \overline{D_j} \right) + \sum_{j=1}^{m} k_j \chi(D_j) \right) \nabla u = 0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu}\Big|_{\partial\Omega} = g \end{cases} \tag{6.1}$$

can be represented as

$$u(x) = H(x) + \sum_{j=1}^{m} \mathcal{S}_{D_j} \psi^{(j)}(x), \quad x \in \Omega, \tag{6.2}$$

where the harmonic function H is given by

$$H(x) = -\mathcal{S}_\Omega(g)(x) + \mathcal{D}_\Omega(f)(x), \quad x \in \Omega, \quad f := u|_{\partial\Omega}, \tag{6.3}$$

and $\psi^{(j)} \in L^2_0(\partial D_j)$, $j = 1, \dots, m$, satisfies the integral equation

$$(\lambda_j I - \mathcal{K}^*_{D_j})\psi^{(j)} - \sum_{k \neq j} \frac{\partial(\mathcal{S}_{D_k}\psi^{(k)})}{\partial \nu^{(j)}}\Big|_{\partial D_j} = \frac{\partial H}{\partial \nu^{(j)}}\Big|_{\partial D_j} \quad \text{on } \partial D_j, \quad j = 1, \dots, m. \tag{6.4}$$

Moreover, for all $n \in \mathbb{N}$, there exist a constant $C_n = C(n, \Omega, dist(D, \partial\Omega))$ independent of $|D|$ and the conductivities $k_j, j = 1, \dots, m$, such that

$$\|H\|_{\mathcal{C}^n(\overline{D})} \leq C_n \|g\|_{L^2(\partial\Omega)}. \tag{6.5}$$

Let $N(x, z)$ be the Neumann function for Δ in Ω corresponding to a Dirac mass at z; that is, N is the solution to

$$\begin{cases} \Delta_x N(x, z) = -\delta_z \quad \text{in } \Omega, \\ \dfrac{\partial N}{\partial \nu}\Big|_{\partial\Omega} = -\dfrac{1}{|\partial\Omega|}. \end{cases} \tag{6.6}$$

In addition, we assume that

$$\int_{\partial\Omega} N(x, y) d\sigma(x) = 0 \quad \text{for } y \in \Omega. \tag{6.7}$$

Let us fix one more notation. For D, a subset of Ω, let

$$N_D f(x) := \int_{\partial D} N(x, y) f(y) d\sigma(y).$$

Then following the same lines of [1], one can prove the following theorem.

THEOREM 6.2. *Let U be the background solution, i.e., the solution of*

$$\begin{cases} \Delta U = 0 \quad \text{in } \Omega, \\ \dfrac{\partial U}{\partial \nu}\big|_{\partial\Omega} = g \in L^2_0(\partial\Omega), \\ \displaystyle\int_{\partial\Omega} U(x)d\sigma(x) = 0. \end{cases} \tag{6.8}$$

Then the solution u of (6.1) *can be represented as*

$$u(x) = U(x) - \sum_{j=1}^{m} N_{D_j}\psi^{(j)}(x), \quad x \in \partial\Omega, \tag{6.9}$$

where $\psi^{(j)}$, $j = 1, \dots, m$, is defined by (6.4).

For $x \in \partial\Omega$, by using the change of variables $y = \frac{x-z}{\epsilon}$ we may write

$$\sum_{j=1}^{m} N_{D_j}\psi^{(j)}(x) = \epsilon^{d-1}\sum_{j=1}^{m}\int_{\partial B_j} N(x, \epsilon y + z)\psi^{(j)}(\epsilon y + z)d\sigma(y). \tag{6.10}$$

As in [1], we expand the Neumann function as

$$N(x, \epsilon y + z) = \sum_{|\beta|=0}^{\infty} \frac{1}{\beta!}\epsilon^{|\beta|}\partial_z^\beta N(x, z)y^\beta. \tag{6.11}$$

We then use the uniqueness of the solution to the integral equation (3.1) and the expansion of the harmonic function H,

$$H(x) := H(z) + \sum_{|\alpha|=1}^{\infty} \frac{1}{\alpha!}(\partial^\alpha H)(z)(x-z)^\alpha, \quad x \in \overline{D},$$

to show that

$$\psi^{(j)}(\epsilon y + z) = \sum_{|\alpha|=1}^{\infty} \frac{\epsilon^{|\alpha|-1}}{\alpha!}(\partial^\alpha H)(z)\varphi_\alpha^{(j)}(y), \quad y \in \partial B_j, \tag{6.12}$$

where $\varphi_\alpha^{(j)}$ is the solution of (3.1). Substituting (6.11) and (6.12) into (6.10), we obtain

$$\sum_{j=1}^{m} N_{D_j}\psi^{(j)}(x) = \sum_{|\alpha|=1}^{\infty}\sum_{|\beta|=0}^{\infty} \frac{\epsilon^{|\alpha|+|\beta|+d-2}}{\alpha!\beta!}(\partial^\alpha H)(z)\partial_z^\beta N(x, z)\sum_{j=1}^{m}\int_{\partial B_j} y^\beta \varphi_\alpha^{(j)}(y)d\sigma(y).$$

If $\beta = 0$, then $\int_{\partial B_j} \varphi_\alpha^{(j)}(y)d\sigma(y) = 0$ for $j = 1, \dots, m$, and hence we get

$$\sum_{j=1}^{m} N_{D_j}\psi^{(j)}(x) = \sum_{|\alpha|=1}^{\infty}\sum_{|\beta|=1}^{\infty} \frac{\epsilon^{|\alpha|+|\beta|+d-2}}{\alpha!\beta!}(\partial^\alpha H)(z)\partial_z^\beta N(x, z)m_{\alpha\beta}, \tag{6.13}$$

where $m_{\alpha\beta}$ is the polarization tensor.

We now convert the formula (6.13) to the one given solely by U and its derivatives, not H. Using formula (4.10) of [1], one can show that

$$|\partial^\alpha H(z) - \partial^\alpha U(z)| \leq C\epsilon^d \|g\|_{L^2(\partial\Omega)} \quad \text{for } \alpha \in \mathbb{N},$$

where C is independent of ϵ and g. We finally have the following theorem.

THEOREM 6.3. *The following pointwise asymptotic expansion holds uniformly in $x \in \partial\Omega$ for $d = 2$ or 3:*

$$u(x) = U(x) - \sum_{|\alpha|=1}^{d} \sum_{|\beta|=1}^{d} \frac{\epsilon^{|\alpha|+|\beta|+d-2}}{\alpha!\beta!} (\partial^\alpha U)(z) \partial_z^\beta N(x,z) m_{\alpha\beta} + O(\epsilon^{2d}), \tag{6.14}$$

where the remainders $O(\epsilon^{2d})$ are dominated by $C\epsilon^{2d}\|g\|_{L^2(\partial\Omega)}$ for some constant C independent of $x \in \partial\Omega$.

7. Detection of closely spaced inclusions. In this section, we present an algorithm to reconstruct the first-order polarization tensor and the center of closely spaced small inclusions from a finite number of boundary measurements and show some results of numerical experiments. The algorithm is based on the asymptotic expansion formula (6.14) and represents the generalization of the numerical methods derived in [6] and [17] for determining well-separated conductivity inhomogeneities. For $g \in L_0^2(\partial\Omega)$, define the harmonic function $H[g](x)$, $x \in \mathbb{R}^d \setminus \overline{\Omega}$, by

$$H[g](x) := -\mathcal{S}_\Omega(g)(x) + \mathcal{D}_\Omega(u|_{\partial\Omega})(x), \quad x \in \mathbb{R}^d \setminus \bar{\Omega}, \tag{7.1}$$

where u is the solution of (6.1). Then by substituting (6.14) into (7.1) and using a simple formula $\mathcal{D}_\Omega(N(\cdot - z))(x) = \Gamma(x - z)$ for $z \in \Omega$ and $x \in \mathbb{R}^d \setminus \overline{\Omega}$, we get

$$H[g](x) = -\sum_{|\alpha|=1}^{d} \sum_{|\beta|=1}^{d} \frac{\epsilon^{|\alpha|+|\beta|+d-2}}{\alpha!\beta!} (\partial^\alpha U)(z) \partial_z^\beta \Gamma(x - z) m_{\alpha\beta} + O(\epsilon^{2d}). \tag{7.2}$$

Assume for the sake of simplicity that $d = 2$. Our reconstruction procedure is the following.

DETECTION ALGORITHM.

(D1) For $g_j = \frac{\partial x_j}{\partial \nu}$, $j = 1, 2$, measure $u|_{\partial\Omega}$.

(D2) Compute the first-order polarization tensor $\epsilon^2 M = \epsilon^2 (m_{ij})$ for D by

$$\epsilon^2 m_{ij} = \lim_{t\to\infty} 2\pi t H[g_i](te_j). \tag{7.3}$$

(D3) Compute $h_j = \lim_{t\to\infty} 2\pi t H[g_3](te_j)$ for $g_3 = \frac{\partial(x_1 x_2)}{\partial \nu}$, $j = 1, 2$. Then the center is estimated by solving

$$z = (h_1, h_2)(\epsilon^2 M)^{-1}. \tag{7.4}$$

(D4) Let the overall conductivity $\bar{k} = \infty$ if the polarization tensor M is positive definite. Otherwise assume $\bar{k} = 0$. Using the similar method in [8] as explained in section 5, we obtain the shape of the equivalent ellipse.

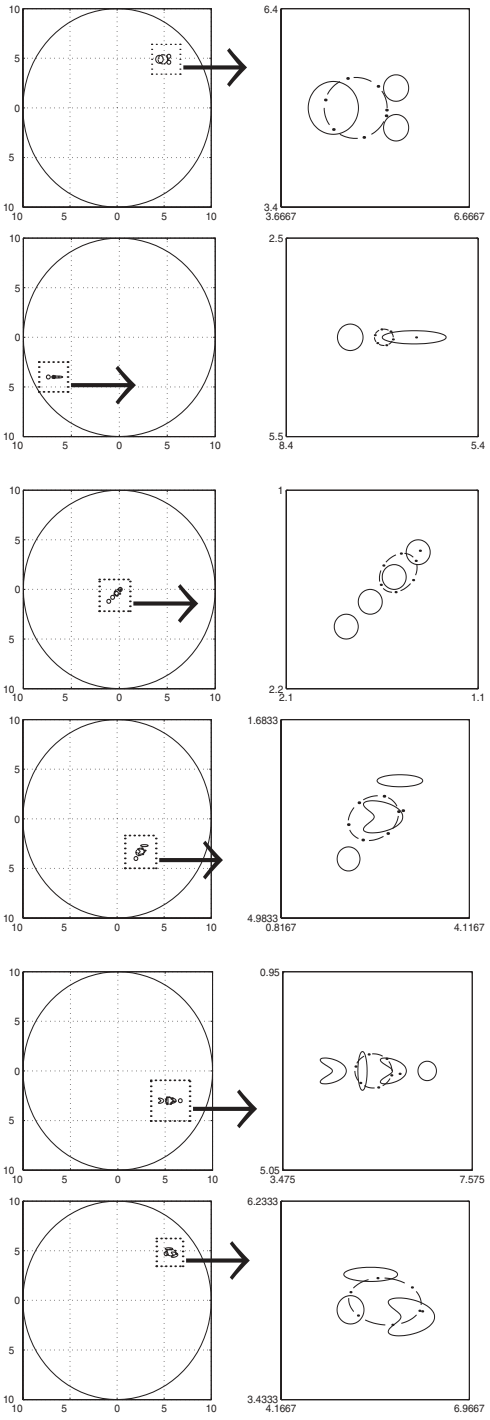


FIG. 7.1. *Reconstruction of closely spaced small inhomogeneities. The dash line is the equivalent ellipse and the dash-dot line is the detected ellipse. The table for numerical values is given in Table* 7.1.

TABLE 7.1

Table for Figure 7.1. Here $\bar{k}, \bar{z}$ are the overall conductivity and center defined by (5.1) and (5.2). $\bar{a}$, $\bar{b}$, and $\bar{\theta}$ are semiaxis lengths and the angle of orientation of the equivalent ellipse while a, b, and θ are those of the detected ellipse, assuming $k = \infty$. z is the detected center.

k_i	$a_0^i, a_1^i, a_2^i, b_0^i, b_1^i, b_2^i$	$\bar{k}$	$\bar{a}$	$\bar{b}$	$\bar{\theta}$	$\bar{z}$
		k	a	b	θ	z
100	5.5, 0.2, 0, 5.2, 0.2, 0	60.079	0.511	0.468	0.000	(4.838, 4.900)
100	5.5, 0.2, 0, 4.6, 0.2, 0	∞	0.502	0.461	0.000	(4.856, 4.899)
50	4.5, 0.4, 0, 4.9, 0.4, 0					
1.5	-7.4, 0.2, 0, -4, 0.2, 0	1.5	0.474	0.190	0.000	(-6.844, -4.000)
1.5	-6.4, 0.5, 0, -4, 0.1, 0	∞	0.146	0.123	0.000	(-6.875, -4.000)
100	0.1, 0.2, 0, 0, 0.2, 0					
100	-0.3, 0.2, 0, -0.4, 0.2, 0	3.88	0.511	0.315	0.785	(-0.236, -0.336)
1.5	-0.7, 0.2, 0, -0.8, 0.2, 0	∞	0.355	0.267	0.785	(-0.233, -0.333)
1.5	-1.1, 0.2, 0,-1.2, 0.2, 0					
5	2.9, 0.4, 0, -2.7, 0.1, 0	18.655	0.491	0.365	0.443	(2.494, -3.375)
100	2.5, 0.25, 0.2, -3.3, 0.25, 0.05	∞	0.458	0.351	0.443	(2.434, -3.321)
50	2.0, 0.2, 0, -4.0, 0.2, 0					
5	4.5, 0.15, 0.2, -3, 0.25, 0.05					
5	5.2, 0.1, 0, -3, 0.4, 0	5	0.507	0.419	-0.000	(5.502, -3.000)
5	5.8, 0.15, 0.2, -3, 0.25,0.05	∞	0.401	0.353	-0.000	(5.436, -3.000)
5	6.6, 0.2, 0, -3, 0.2, 0					
100	6.0, 0.25, 0.2, 4.6, 0.25, 0.05	100	0.549	0.331	-0.089	(5.728, 4.772)
100	5.5, 0.4, 0, 5.2, 0.1, 0	∞	0.540	0.329	-0.089	(5.712, 4.817)
100	5.2, 0.2, 0, 4.7, 0.2, 0					

In order to collect data $u|_{\partial\Omega}$ in step (D1), we solve the direct problem (6.1) as follows: Using the formula (6.2) and the jump relations (2.5) and (2.6), we have the following equation:

(7.5)
$$\begin{cases} u = \dfrac{u}{2} + \mathcal{K}_\Omega u - \mathcal{S}_\Omega g + \displaystyle\sum_{i=1}^{m} \mathcal{S}_{D_i}\psi^{(i)} \quad \text{on } \partial\Omega, \\ (\lambda_j I - \mathcal{K}_{D_j}^*)\psi^{(j)} - \displaystyle\sum_{k\neq j} \frac{\partial(\mathcal{S}_{D_k}\psi^{(k)})}{\partial\nu^{(j)}}\bigg|_{\partial D_j} = \frac{\partial H}{\partial\nu^{(j)}}\bigg|_{\partial D_j} \quad \text{on } \partial D_j, \quad j = 1, \dots, m. \end{cases}$$

We solve the integral equation using the collocation method and obtain $u|_{\partial\Omega}$ on $\partial\Omega$ for given data g.

A few words are required for the step (D4). In order to find the overall conductivity, it is necessary to know the individual conductivity k_l and the size of B_l, $l = 1, \dots, m$, which seems impossible. Thus we assume a priori that $\bar{k}$ is either ∞ or 0 depending upon the sign of the detected polarization tensor. Therefore, it is natural that our algorithm gives better information when the conductivity contrast between the background and inclusions is high.

We illustrate in Figure 7.1 the viability of this algorithm. For rigorous justification of the validity of this algorithm the reader is referred to [6, 17]. It should also be noted that although our algorithm is only described and tested in the two-dimensional case, we are confident that it works in the three-dimensional case as well.

Acknowledgment. We would like to thank June-Yub Lee for helpful conversation on numerical computations.

REFERENCES

[1] H. Ammari and H. Kang, *High-order terms in the asymptotic expansions of the steady-state voltage potentials in the presence of conductivity inhomogeneities of small diameter*, SIAM J. Math. Anal., 34 (2003), pp. 1152–1166.

[2] H. Ammari and H. Kang, *Properties of the generalized polarization tensors*, Multiscale Model. Simul., 1 (2003), pp. 335–348.

[3] H. Ammari, H. Kang, G. Nakamura, and K. Tanuma, *Complete asymptotic expansions of solutions of the system of elastostatics in the presence of an inclusion of small diameter and detection of an inclusion*, J. Elasticity, 67 (2002), pp. 97–129.

[4] H. Ammari and A. Khelifi, *Electromagnetic scattering by small dielectric inhomogeneities*, J. Math. Pures Appl., 82 (2003), pp. 749–842.

[5] H. Ammari, S. Moskow, and M. Vogelius, *Boundary integral formulas for the reconstruction of electric and electromagnetic inhomogeneities of small volume*, ESAIM Cont. Optim. Calc. Var., 9 (2003), pp. 49–66.

[6] H. Ammari and J.K. Seo, *An accurate formula for the reconstruction of conductivity inhomogeneities*, Adv. in Appl. Math., 30 (2003), pp. 679–705.

[7] H. Ammari, M. Vogelius, and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter. II. The full Maxwell equations*, J. Math. Pures Appl., 80 (2001), pp. 769–814.

[8] M. Brühl, M. Hanke, and M. Vogelius, *A direct impedance tomography algorithm for locating small inhomogeneities*, Numer. Math., 93 (2003), pp. 635–654.

[9] Y. Capdeboscq and M. Vogelius, *A general representation formula for the boundary voltage perturbations caused by internal conductivity inhomogeneities of low volume fraction*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 159–173.

[10] D.J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of conductivity imperpections of small diameter by boundary measurements. Continuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[11] H. Cheng and L. Greengard, *A method of images for the evaluation of electrostatic fields in systems of closely spaced conducting cylinders*, SIAM J. Appl. Math., 58 (1998), pp. 122–141.

[12] R.R. Coifman, A. McIntosh, and Y. Meyer, *L'integrale de Cauchy définit un opérateur borné sur L^2 pour les courbes Lipschitziennes*, Ann. of Math. (2), 116 (1982), pp. 361–387.

[13] L. Escauriaza, E.B. Fabes, and G. Verchota, *On a regularity theorem for weak solutions to transmission problems with internal Lipschitz boundaries*, Proc. Amer. Math. Soc., 115 (1992), pp. 1069–1076.

[14] E.B. Fabes, M. Jodeit, and N.M. Riviére, *Potential techniques for boundary value problems on $\mathcal{C}^1$ domains*, Acta Math., 141 (1978), pp. 165–186.

[15] G.B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, 1976.

[16] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[17] H. Kang, E. Kim, and K. Kim, *Anisotropic polarization tensors and detection of an anisotropic inclusion*, SIAM J. Appl. Math. 63 (2003), pp. 1276–1291.

[18] H. Kang, E. Kim, and J. Lee, *Identification of elastic inclusions and elastic moment tensors by boundary measurements*, Inverse Problems, 19 (2003), pp. 703–724.

[19] H. Kang and J.K. Seo, *The layer potential technique for the inverse conductivity problem*, Inverse Problems, 12 (1996), pp. 267–278.

[20] H. Kang and J.K. Seo, *Recent progress in the inverse conductivity problem with single measurement*, in Inverse Problems and Related Fields, CRC Press, Boca Raton, FL, 2000, pp. 69–80.

[21] R.E. Kleinman and T.B.A. Senior, *Rayleigh Scattering*. Low and High Frequency Asymptotics, V.K. Varadan and V.V. Varadan, eds., North–Holland, Amsterdam, 1986, pp. 1–70.

[22] O. Kwon and J.-K. Seo, *Total size estimation and identification of multiple anomalies in the inverse conductivity problem*, Inverse Problems, 17 (2001), pp. 59–75.

[23] O. Kwon, J.K. Seo, and J.R. Yoon, *A real-time algorithm for the location search of discontinuous conductivities with one measurement*, Comm. Pure Appl. Math., 55 (2002), pp. 1–29.

[24] A. Nachmann, *Reconstructions from boundary measurements*, Ann. of Math. (2), 128 (1988), pp. 531–576.

[25] G. Pólya and G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Annals of Mathematical Studies, 27, Princeton University Press, Princeton, 1951.
[26] M. Schiffer and G. Szegö, *Virtual mass and polarization*, Trans. Amer. Math. Soc., 67 (1949), pp. 130–205.
[27] G.C. Verchota, *Layer potentials and regularity for the Dirichlet problem for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.
[28] M. Vogelius and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 723–748.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2429–2451

QUADRILATERAL $H(\mathrm{div})$ FINITE ELEMENTS*

DOUGLAS N. ARNOLD†, DANIELE BOFFI‡, AND RICHARD S. FALK§

Abstract. We consider the approximation properties of quadrilateral finite element spaces of vector fields defined by the Piola transform, extending results previously obtained for scalar approximation. The finite element spaces are constructed starting with a given finite dimensional space of vector fields on a square reference element, which is then transformed to a space of vector fields on each convex quadrilateral element via the Piola transform associated to a bilinear isomorphism of the square onto the element. For affine isomorphisms, a necessary and sufficient condition for approximation of order $r+1$ in L^2 is that each component of the given space of functions on the reference element contain all polynomial functions of total degree at most r. In the case of bilinear isomorphisms, the situation is more complicated and we give a precise characterization of what is needed for optimal order L^2-approximation of the function and of its divergence. As applications, we demonstrate degradation of the convergence order on quadrilateral meshes as compared to rectangular meshes for some standard finite element approximations of $\boldsymbol{H}(\mathrm{div})$. We also derive new estimates for approximation by quadrilateral Raviart–Thomas elements (requiring less regularity) and propose a new quadrilateral finite element space which provides optimal order approximation in $\boldsymbol{H}(\mathrm{div})$. Finally, we demonstrate the theory with numerical computations of mixed and least squares finite element approximations of the solution of Poisson's equation.

Key words. quadrilateral, finite element, approximation, mixed finite element

AMS subject classifications. 65N30, 41A10, 41A25, 41A27, 41A63

DOI. 10.1137/S0036142903431924

1. Introduction.

Many mixed finite element methods are based on variational principles employing the space $\boldsymbol{H}(\mathrm{div},\Omega)$ consisting of $\boldsymbol{L}^2$ vector fields with divergence in L^2. For such methods, finite element subspaces of $\boldsymbol{H}(\mathrm{div},\Omega)$ are generally constructed starting from a space of reference shape functions on a reference element, typically the unit simplex or unit square in two dimensions. See, e.g., [4] for numerous examples. These shape functions are then transformed to general triangular, rectangular, or quadrilateral elements via polynomial diffeomorphisms and the Piola transform. For the case of triangular and rectangular (or more generally parallelogram) elements, i.e., the case of affine isomorphisms, the order of approximation so achieved can be easily determined from the highest degree of complete polynomial space contained in the space of reference shape functions. In the case of arbitrary convex quadrilaterals with bilinear diffeomorphisms, the situation is less well understood. In this paper, we determine precisely what reference shape functions are needed to obtain a given order of approximation in $\boldsymbol{L}^2$ and $\boldsymbol{H}(\mathrm{div},\Omega)$ by such elements. It turns out that the accuracy of some of the standard $\boldsymbol{H}(\mathrm{div},\Omega)$ finite elements is lower for general quadrilateral elements than for rectangular elements.

Let $\hat{K}$ be a reference element, the closure of an open set in $\mathbb{R}^2$, and let $\boldsymbol{F}:\hat{K}\to\mathbb{R}^2$ be a diffeomorphism of $\hat{K}$ onto an actual element $K=\boldsymbol{F}(\hat{K})$. For functions in

*Received by the editors July 21, 2003; accepted for publication (in revised form) April 2, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/sinum/42-6/43192.html
†Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN 55455 (arnold@ima.umn.edu). The research of this author was supported by NSF grant DMS-0107233.
‡Dipartimento di Matematica, Università di Pavia, 27100 Pavia, Italy (boffi@dimat.unipv.it). The research of this author was supported by IMATI-CNR, Italy and by MIUR/PRIN2001, Italy.
§Department of Mathematics, Rutgers University, Piscataway, NJ 08854 (falk@math.rutgers.edu). The research of this author was supported by NSF grant DMS-0072480.

$\boldsymbol{H}(\mathrm{div},\Omega)$ the natural way to transform functions from $\hat{K}$ to K is via the *Piola transform*. Namely, given a function $\hat{\boldsymbol{u}} : \hat{K} \to \mathbb{R}^2$, we define $\boldsymbol{u} = \boldsymbol{P}_F\hat{u} : K \to \mathbb{R}^2$ by

$$\boldsymbol{u}(\boldsymbol{x}) = J\boldsymbol{F}(\hat{\boldsymbol{x}})^{-1} D\boldsymbol{F}(\hat{\boldsymbol{x}})\hat{\boldsymbol{u}}(\hat{\boldsymbol{x}}), \tag{1.1}$$

where $\boldsymbol{x} = \boldsymbol{F}(\hat{\boldsymbol{x}})$, and $D\boldsymbol{F}(\hat{\boldsymbol{x}})$ is the Jacobian matrix of the mapping $\boldsymbol{F}$ and $J\boldsymbol{F}(\hat{\boldsymbol{x}})$ its determinant. The transform has the property that if $\boldsymbol{u} = \boldsymbol{P}_F\hat{\boldsymbol{u}}$, $p = \hat{p} \circ \boldsymbol{F}^{-1}$ for some $\hat{p} : \hat{K} \to \mathbb{R}$, and $\boldsymbol{n}$ and $\hat{\boldsymbol{n}}$ denote the unit outward normals on ∂K and $\partial\hat{K}$, respectively, then

$$\int_K \operatorname{div} \boldsymbol{u} p \, d\boldsymbol{x} = \int_{\hat{K}} \hat{\operatorname{div}}\, \hat{\boldsymbol{u}}\hat{p} \, d\hat{\boldsymbol{x}}, \qquad \int_{\partial K} \boldsymbol{u}\cdot\boldsymbol{n} p \, ds = \int_{\partial\hat{K}} \hat{\boldsymbol{u}}\cdot\hat{\boldsymbol{n}}\hat{p} \, d\hat{s}.$$

Since continuity of $\boldsymbol{u}\cdot\boldsymbol{n}$ is necessary for finite element subspaces of $\boldsymbol{H}(\mathrm{div},\Omega)$, use of the Piola transform facilitates the definition of finite element subspaces of $\boldsymbol{H}(\mathrm{div},\Omega)$ by mapping from a reference element. Another important property of the Piola transform, which follows directly from the chain rule and which we shall use frequently below, is that if $\boldsymbol{G}$ is a diffeomorphism whose domain is K, then

$$\boldsymbol{P}_{G\circ F} = \boldsymbol{P}_G \circ \boldsymbol{P}_F. \tag{1.2}$$

Using the Piola transform, a standard construction of a finite element subspace proceeds as follows. Let $\hat{K}$ be a fixed reference element, typically either the unit simplex or the unit square. Let $\hat{\boldsymbol{V}} \subset \boldsymbol{H}(\mathrm{div},\hat{K})$ be a finite-dimensional space of vector fields on $\hat{K}$, typically polynomial, the space of reference *shape functions*. Now suppose we are given a mesh $\mathcal{T}_h$ consisting of elements K, each of which is the image of $\hat{K}$ under some given diffeomorphism: $K = \boldsymbol{F}_K(\hat{K})$. Via the Piola transform we then obtain the space $\boldsymbol{P}_{F_K}\hat{\boldsymbol{V}}$ of shape functions on K. Finally we define the finite element space as

$$\boldsymbol{S}_h = \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div},\Omega) \mid \boldsymbol{v}|_K \in \boldsymbol{P}_{F_K}\hat{\boldsymbol{V}} \ \forall K \in \mathcal{T}_h\}.$$

Recall that $\boldsymbol{S}_h$ may be characterized as the subspace of

$$\boldsymbol{V}_h := \{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega) \mid \boldsymbol{v}|_K \in \boldsymbol{P}_{F_K}\hat{\boldsymbol{V}} \ \forall K \in \mathcal{T}_h\},$$

consisting of vector fields whose normal component is continuous across interelement edges.

We now recall a few examples of this construction in the case where $\hat{K}$ is the unit square. If we restrict to linear diffeomorphisms $\boldsymbol{F}$, the resulting finite elements $K = \boldsymbol{F}(\hat{K})$ will be parallelograms (or, with the further restriction to diagonal linear diffeomorphisms, rectangles). If we allow general bilinear diffeomorphisms, the resulting finite elements can be arbitrary convex quadrilaterals. The best known example of shape functions on the reference square for construction of $\boldsymbol{H}(\mathrm{div},\Omega)$ finite element spaces is the Raviart–Thomas space of index $r \geq 0$ for which $\hat{\boldsymbol{V}}$ is taken to be $\boldsymbol{\mathcal{RT}}_r := \mathcal{P}_{r+1,r}(\hat{K}) \times \mathcal{P}_{r,r+1}(\hat{K})$. Here and below $\mathcal{P}_{s,t}(\hat{K})$ denotes the space of polynomial functions on $\hat{K}$ of degree at most s in $\hat{x}_1$ and at most t in $\hat{x}_2$. Thus a basis for $\boldsymbol{\mathcal{RT}}_r$ is given by the $2(r+1)(r+2)$ vector fields

$$(\hat{x}_1^i\hat{x}_2^j, 0), \quad (0, \hat{x}_1^j\hat{x}_2^i), \quad 0 \leq i \leq r+1, \ 0 \leq j \leq r. \tag{1.3}$$

A second example is given by choosing $\hat{\boldsymbol{V}}$ to be the Brezzi–Douglas–Marini space of index $r \geq 1$, $\hat{\boldsymbol{V}} = \boldsymbol{\mathcal{BDM}}_r$, which is the span of $\boldsymbol{\mathcal{P}}_r(\hat{K})$ and the two additional vector fields $\mathrm{curl}(\hat{x}_1^{r+1}\hat{x}_2)$ and $\mathrm{curl}(\hat{x}_1\hat{x}_2^{r+1})$. Another possibility is the Brezzi–Douglas–Fortin–Marini space $\hat{\boldsymbol{V}} = \boldsymbol{\mathcal{BDFM}}_{r+1}$, $r \geq 0$, which is the subspace of codimension 2 of $\boldsymbol{\mathcal{P}}_{r+1}(\hat{K})$ spanned by $(\hat{x}_1^i\hat{x}_2^j, 0)$ and $(0, \hat{x}_1^j\hat{x}_2^i)$ for nonnegative i and j with $i+j \leq r+1$ and $j \leq r$. We note that for each of these choices $\hat{\boldsymbol{V}}$ strictly contains $\boldsymbol{\mathcal{P}}_r(\hat{K})$ but does not contain $\boldsymbol{\mathcal{P}}_{r+1}(\hat{K})$. Note that $\boldsymbol{\mathcal{BDM}}_0$ is not defined, $\boldsymbol{\mathcal{BDFM}}_1 = \boldsymbol{\mathcal{RT}}_0$, and $\boldsymbol{\mathcal{BDM}}_r \subsetneq \boldsymbol{\mathcal{BDFM}}_{r+1} \subsetneq \boldsymbol{\mathcal{RT}}_r$ for $r \geq 1$. More information about these spaces can be found in [4, section III.3.2].

One of the basic issues in finite element theory concerns the approximation properties of finite element spaces. Namely, under certain regularity assumptions on the mesh $\mathcal{T}_h$, for a given smooth vector field $\boldsymbol{u} : \Omega \to \mathbb{R}^2$ one usually estimates the error (in some norm to be made more precise) in the best approximation of $\boldsymbol{u}$ by vector fields in $\boldsymbol{S}_h$ as a quantity involving powers of h, the maximum element diameter. For instance, given a shape-regular sequence of triangular or parallelogram meshes $\mathcal{T}_h$ of Ω with $\boldsymbol{S}_h$ the corresponding Raviart–Thomas spaces of index $r \geq 0$, then for any vector field $\boldsymbol{u}$ smooth enough that the right-hand sides of the next expressions make sense, there exists $\boldsymbol{\pi}_h\boldsymbol{u} \in \boldsymbol{S}_h$ such that (cf. [4])

$$\|\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u}\|_{L^2(\Omega)} \leq Ch^{r+1}|\boldsymbol{u}|_{H^{r+1}(\Omega)},$$
$$\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u})\|_{L^2(\Omega)} \leq Ch^{r+1}|\operatorname{div}\boldsymbol{u}|_{H^{r+1}(\Omega)}.$$

In the case of more general shape-regular convex quadrilaterals, the best known estimate appears to be the one obtained by Thomas in [12]:

$$\|\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u}\|_{L^2(\Omega)} \leq Ch^{r+1}[|\boldsymbol{u}|_{H^{r+1}(\Omega)} + h|\operatorname{div}\boldsymbol{u}|_{H^{r+1}(\Omega)}],$$
$$\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u})\|_{L^2(\Omega)} \leq Ch^{r}|\operatorname{div}\boldsymbol{u}|_{H^{r+1}(\Omega)}.$$

Note that the order in h for the L^2 estimate on $\boldsymbol{u}$ is the same as for the parallelogram meshes, but additional regularity is required, while the estimate for $\operatorname{div}\boldsymbol{u}$ is one order lower in h. As we shall see below, the latter estimate cannot be improved. However, in section 4 of this paper, we use a modification of the usual scaling argument to obtain the improved L^2 estimate

$$\|\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u}\|_{L^2(\Omega)} \leq Ch^{r+1}|\boldsymbol{u}|_{H^{r+1}(\Omega)}.$$

We restrict our presentation to two-dimensional domains, the three-dimensional case being considerably more complicated. We hope to address this issue in future work. We observe that in [9] the construction of $\boldsymbol{H}(\mathrm{div}, \Omega)$ elements on hexahedrons has been considered. The point of view of [9] is somewhat different from ours in that the elements are not obtained by applying the Piola transform starting from a fixed set of basis functions on the unit cube. Other papers dealing with modifications of standard shape functions for the approximation of vector fields are [11, 8]; in the first paper a simple lowest-order two-dimensional element is proposed (which is not obtained via the Piola transform), while in the second paper a construction based on macroelements is presented.

In this paper, we adapt the theory presented in [1] to the case of vector elements defined by the Piola transform, seeking necessary conditions for L^2-approximation of order $r+1$ for $\boldsymbol{u}$ and $\operatorname{div}\boldsymbol{u}$. More specifically, we shall prove in section 3 that in order for the $\boldsymbol{L}^2$ error in the best approximation of $\boldsymbol{u}$ by functions in $\boldsymbol{V}_h$ to be of order

$r+1$, the space $\hat{\boldsymbol{V}}$ must contain $\boldsymbol{\mathcal{S}}_r$, where $\boldsymbol{\mathcal{S}}_r$ is the subspace of codimension one of $\boldsymbol{\mathcal{RT}}_r$ spanned by the vector fields in (1.3) except that the two fields $(\hat{x}_1^{r+1}\hat{x}_2^r, 0)$ and $(0, \hat{x}_1^r\hat{x}_2^{r+1})$ are replaced by the single vector field $(\hat{x}_1^{r+1}\hat{x}_2^r, -\hat{x}_1^r\hat{x}_2^{r+1})$. To establish this result, we shall exhibit a domain Ω and a sequence $\mathcal{T}_h$ of meshes of it, and prove that whenever $\boldsymbol{\mathcal{S}}_r$ is not contained in $\hat{\boldsymbol{V}}$, there exists a smooth vector field $\boldsymbol{u}$ on Ω such that

$$\inf_{\boldsymbol{v}\in\boldsymbol{V}_h} \|\boldsymbol{u}-\boldsymbol{v}\|_{L^2(\Omega)} \neq o(h^r).$$

The example is far from pathological. The domain is simply a square, the mesh sequence does not degenerate in any sense—in fact all the elements of all the meshes in the sequence are similar to a single right trapezoid—and the function $\boldsymbol{u}$ is a polynomial. We use the same mesh sequence to establish a necessary condition for order $r+1$ approximation to $\operatorname{div}\boldsymbol{u}$, namely that $\operatorname{div}\hat{\boldsymbol{V}} \supseteq \mathcal{R}_r$, where $\mathcal{R}_r$ is the subspace of codimension one of $\mathcal{Q}_{r+1}$, the space of polynomials of degree $\leq r+1$ in each variable separately, spanned by the monomials in $\mathcal{Q}_{r+1}$ except $\hat{x}_1^{r+1}\hat{x}_2^{r+1}$. A consequence of these results, also discussed in section 3, is that while the Raviart–Thomas space of index r achieves order $r+1$ approximation in $\boldsymbol{L}^2$ for quadrilateral meshes as for rectangular meshes, the order of approximation of the divergence is only of order r in the quadrilateral case (but of order $r+1$ for rectangular meshes). Thus, in the case $r=0$, there is no convergence in $\boldsymbol{H}(\operatorname{div},\Omega)$. For the Brezzi–Douglas–Marini and Brezzi–Douglas–Fortin–Marini spaces, the order of convergence is severely reduced on general quadrilateral meshes not only for $\operatorname{div}\boldsymbol{u}$ but also for $\boldsymbol{u}$.

In section 4, we show that the necessary conditions for order $r+1$ approximation of $\boldsymbol{u}$ and $\operatorname{div}\boldsymbol{u}$ established in section 3 are also sufficient. The argument used allows us to obtain the previously mentioned improved estimate for approximation by quadrilateral Raviart–Thomas elements. In section 5, we devise a new finite element subspace of $\boldsymbol{H}(\operatorname{div},\Omega)$ which gives optimal order approximation in both $\ll^2$ and $\boldsymbol{H}(\operatorname{div},\Omega)$ on general convex quadrilaterals. In sections 6 and 7, we present applications of these results to the approximation of second order elliptic partial differential equations by mixed and least squares finite element methods. In particular, we show that despite the lower order of approximation of the divergence by Raviart–Thomas quadrilateral elements, the mixed method approximation of the scalar and vector variable retain optimal order convergence orders in L^2. By contrast, error estimates for the least squares method indicate a possible loss of convergence for both the scalar and vector variable. In the final section, we illustrate the positive results with some numerical examples and confirm the degradation of accuracy on quadrilateral meshes in the cases predicted by our theory.

2. Approximation theory of vector fields on rectangular meshes. In this preliminary section of the paper we adapt to vector fields the results presented in the corresponding section of [1] for scalar functions. Although the Piola transform is used in the definition of the finite elements, its simple expression on rectangular meshes requires only minor changes in the proof given in [1] and so we give only a statement of the results.

Let K be any square with edges parallel to the axes, namely $K = \boldsymbol{F}_K(\hat{K})$ with

$$\boldsymbol{F}_K(\hat{\boldsymbol{x}}) = \boldsymbol{x}_K + h_K\hat{\boldsymbol{x}}, \tag{2.1}$$

where $\boldsymbol{x}_K \in \mathbb{R}^2$ is the lower left corner of K and $h_K > 0$ is its side length. The Piola transform of $\hat{\boldsymbol{u}} \in \boldsymbol{L}^2(\hat{K})$ is simply given by $(\boldsymbol{P}_{F_K}\hat{\boldsymbol{u}})(\boldsymbol{x}) = h_K^{-1}\hat{\boldsymbol{u}}(\hat{\boldsymbol{x}})$ where

$\boldsymbol{x} = \boldsymbol{F}_K\hat{\boldsymbol{x}}$. We also have the simple expressions $\operatorname{div}(\boldsymbol{P}_{F_K}\hat{\boldsymbol{u}})(\boldsymbol{x}) = h_K^{-2}\,\hat{\operatorname{div}}\,\hat{\boldsymbol{u}}(\hat{\boldsymbol{x}})$ and $\|\boldsymbol{P}_{F_K}\hat{\boldsymbol{u}}\|_{L^2(K)} = \|\hat{\boldsymbol{u}}\|_{L^2(\hat{K})}$.

Let Ω denote the unit square (Ω and $\hat{K}$ both denote the unit square, but we use the notation Ω when we think of it as a domain, while we use $\hat{K}$ when we think of it as a reference element), and for n a positive integer, let $\mathcal{U}_h$ be the uniform mesh of Ω into n^2 subsquares of side length $h = 1/n$. Given a subspace $\hat{\boldsymbol{V}}$ of $\boldsymbol{L}^2(\hat{K})$ we define

$$\boldsymbol{V}_h = \{\boldsymbol{u} : \Omega \to \mathbb{R}^2 \mid \boldsymbol{u}|_K \in \boldsymbol{P}_{F_K}\hat{\boldsymbol{V}}\ \forall K \in \mathcal{U}_h\}. \tag{2.2}$$

In this definition, when we write $\boldsymbol{u}|_K \in \boldsymbol{P}_{F_K}\hat{\boldsymbol{V}}$ we mean only that $\boldsymbol{u}|_K$ agrees with a function in $\boldsymbol{P}_{F_K}\hat{\boldsymbol{V}}$ almost everywhere, and so do not impose any interelement continuity. Then we have the following approximation results.

THEOREM 2.1. *Let $\hat{\boldsymbol{V}}$ be a finite-dimensional subspace of $\boldsymbol{L}^2(\hat{K})$ and r be a nonnegative integer. The following conditions are equivalent:*

(i) *There is a constant C such that $\inf_{\boldsymbol{v}\in\boldsymbol{V}_h}\|\boldsymbol{u}-\boldsymbol{v}\|_{L^2(\Omega)} \le Ch^{r+1}|\boldsymbol{u}|_{H^{r+1}(\Omega)}$ for all $\boldsymbol{u} \in \boldsymbol{H}^{r+1}(\Omega)$.*

(ii) $\inf_{\boldsymbol{v}\in\boldsymbol{V}_h}\|\boldsymbol{u}-\boldsymbol{v}\|_{L^2(\Omega)} = o(h^r)$ *for all $\boldsymbol{u} \in \boldsymbol{\mathcal{P}}_r(\Omega)$.*

(iii) $\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{P}}_r(\hat{K})$.

THEOREM 2.2. *Let $\hat{\boldsymbol{V}}$ be a finite-dimensional subspace of $\boldsymbol{L}^2(\hat{K})$ and r be a nonnegative integer. The following conditions are equivalent:*

(i) *There is a constant C such that*

$$\inf_{\boldsymbol{v}\in\boldsymbol{V}_h}\|\operatorname{div}\boldsymbol{u}-\operatorname{div}\boldsymbol{v}\|_{L^2(\Omega)} \le Ch^{r+1}|\operatorname{div}\boldsymbol{u}|_{H^{r+1}(\Omega)}$$

for all $\boldsymbol{u} \in \boldsymbol{H}^{r+1}(\Omega)$ with $\operatorname{div}\boldsymbol{u} \in H^{r+1}(\Omega)$.

(ii) $\inf_{\boldsymbol{v}\in\boldsymbol{V}_h}\|\operatorname{div}\boldsymbol{u}-\operatorname{div}\boldsymbol{v}\|_{L^2(\Omega)} = o(h^r)$ *for all $\boldsymbol{u}$ with $\operatorname{div}\boldsymbol{u} \in \mathcal{P}_r(\Omega)$.*

(iii) $\hat{\operatorname{div}}\,\hat{\boldsymbol{V}} \supseteq \mathcal{P}_r(\hat{K})$.

Remark. Since we do not impose interelement continuity in the definition of $\boldsymbol{V}_h$, in Theorem 2.2 $\operatorname{div}\boldsymbol{v}$ should be interpreted as the divergence applied elementwise to $\boldsymbol{v} \in \boldsymbol{V}_h$.

3. A necessary condition for optimal approximation of vector fields on general quadrilateral meshes. In this section, we determine the properties of the finite element approximating spaces that are necessary for order $r+1$ L^2-approximation of a vector field and its divergence on quadrilateral meshes. The construction of the finite element spaces proceeds as in the previous section. We start with the reference shape functions, a finite-dimensional space $\hat{\boldsymbol{V}}$ of vector fields on the unit square $\hat{K} = [0,1]\times[0,1]$ (typically $\hat{\boldsymbol{V}}$ consists of polynomials). Given an arbitrary convex quadrilateral K and a bilinear isomorphism $\boldsymbol{F}_K$ of the reference element $\hat{K}$ onto K, the shape functions on K are then taken to be $\boldsymbol{P}_{F_K}\hat{\boldsymbol{V}}$. (Note that there are eight possible choices for the bilinear isomorphism $\boldsymbol{F}_K$, but the space $\boldsymbol{P}_{F_K}\hat{\boldsymbol{V}}$ does not depend on the particular choice whenever $\hat{\boldsymbol{V}}$ is invariant under the symmetries of the square, which is usually the case in practice. When that is not the case, which we shall allow, it is necessary to specify not only the elements K but for each a choice of bilinear isomorphism from the reference element to K.) Finally, given a quadrilateral mesh $\mathcal{T}$ of a two-dimensional domain Ω, we can then construct the space of vector fields $\boldsymbol{V}(\mathcal{T})$ consisting of functions on Ω which belong to $\boldsymbol{P}_{F_K}\hat{\boldsymbol{V}}$ when restricted to a generic quadrilateral $K \in \mathcal{T}$.

It follows from the results of the previous section that if we consider the sequence $\mathcal{T}_h = \mathcal{U}_h$ of meshes of the unit square into congruent subsquares of side length $h = 1/n$,

then the approximation estimate

$$\inf_{\boldsymbol{v}\in\boldsymbol{V}(\mathcal{T}_h)} \|\boldsymbol{u}-\boldsymbol{v}\|_{L^2(\Omega)} = o(h^r) \quad \forall\ \boldsymbol{u}\in\boldsymbol{\mathcal{P}}_r(\Omega) \tag{3.1}$$

is valid only if $\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{P}}_r(\hat{K})$ and the estimate

$$\inf_{\boldsymbol{v}\in\boldsymbol{V}(\mathcal{T}_h)} \|\operatorname{div}\boldsymbol{u} - \operatorname{div}\boldsymbol{v}\|_{L^2(\Omega)} = o(h^r) \quad \forall\ \boldsymbol{u} \text{ with } \operatorname{div}\boldsymbol{u}\in\mathcal{P}_r(\Omega) \tag{3.2}$$

is valid only if $\hat{\operatorname{div}}(\hat{\boldsymbol{V}}) \supseteq \mathcal{P}_r(\hat{K})$. In this section we show that for these estimates to hold for more general quadrilateral mesh sequences $\mathcal{T}_h$, stronger conditions on $\hat{\boldsymbol{V}}$ are required.

Before stating the main results of this section, we briefly recall a measure for the shape regularity of a convex quadrilateral K, cf. [7, A.2, pp. 104–105] or [13]. From the quadrilateral K we obtain four triangles by the four possible choices of three vertices from the vertices of K, and we define ρ_K as the smallest diameter of the inscribed circles to these four triangles. The *shape constant* of K is then $\sigma_K := h_K/\rho_K$, where $h_K = \operatorname{diam}(K)$. A bound on σ_K implies a bound on the ratio of any two sides of K and also a bound away from 0 and π for its angles (and conversely such bounds imply an upper bound on σ_K). It also implies bounds on the Lipschitz constant of $h_K^{-1}\boldsymbol{F}_K$ and its inverse. The shape constant of a mesh $\mathcal{T}_h$ consisting of convex quadrilaterals is then defined to be the supremum of the shape constants σ_K for $K\in\mathcal{T}_h$, and a family $\mathcal{T}_h$ of such meshes is called shape-regular if the shape constants for the meshes can be uniformly bounded.

The following two theorems give necessary conditions on the shape functions in order to ensure estimates like (3.1) and (3.2) on arbitrary quadrilateral mesh sequences. The spaces $\boldsymbol{\mathcal{S}}_r$ and $\mathcal{R}_r$ were defined in section 1.

THEOREM 3.1. *Suppose that the estimate* (3.1) *holds whenever $\mathcal{T}_h$ is a shape-regular sequence of quadrilateral meshes of a two-dimensional domain Ω. Then $\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{S}}_r$.*

THEOREM 3.2. *Suppose that the estimate* (3.2) *holds whenever $\mathcal{T}_h$ is a shape-regular sequence of quadrilateral meshes of a two-dimensional domain Ω. Then $\hat{\operatorname{div}}\,\hat{\boldsymbol{V}} \supseteq \mathcal{R}_r$.*

In order to establish the theorems, we shall make use of two results analogous to Theorem 4 of [1]. To state these results, we introduce some specific bilinear mappings. For $\alpha > 0$, let $\boldsymbol{F}^\alpha$ and $\boldsymbol{G}^\alpha$ denote the mappings

$$\boldsymbol{F}^\alpha(\hat{\boldsymbol{x}}) = (\hat{x}_1, (\alpha+\hat{x}_1)\hat{x}_2), \qquad \boldsymbol{G}^\alpha(\hat{\boldsymbol{x}}) = \boldsymbol{F}^\alpha(\hat{x}_2,\hat{x}_1), \tag{3.3}$$

each of which maps the unit square $\hat{K}$ to the quadrilateral K^α with vertices $(0,0)$, $(1,0)$, $(1,\alpha+1)$, and $(0,\alpha)$.

LEMMA 3.3. *Let $\hat{\boldsymbol{V}}$ be a space of vector fields on $\hat{K}$ such that $\boldsymbol{P}_F\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{P}}_r(\boldsymbol{F}(\hat{K}))$ when $\boldsymbol{F}$ is any of the four bilinear isomorphisms $\boldsymbol{F}^1$, $\boldsymbol{F}^2$, $\boldsymbol{G}^1$, and $\boldsymbol{G}^2$. Then $\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{S}}_r$.*

LEMMA 3.4. *Let $\hat{\boldsymbol{V}}$ be a space of vector fields on $\hat{K}$ such that $\operatorname{div}\boldsymbol{P}_F\hat{\boldsymbol{V}} \supseteq \mathcal{P}_r(\boldsymbol{F}(\hat{K}))$ when $\boldsymbol{F}$ is any of the four bilinear isomorphisms $\boldsymbol{F}^1$, $\boldsymbol{F}^2$, $\boldsymbol{G}^1$, and $\boldsymbol{G}^2$. Then $\hat{\operatorname{div}}\,\hat{\boldsymbol{V}} \supseteq \mathcal{R}_r$.*

We postpone the proof of these lemmas to the end of the section. Now, based on Lemma 3.3 and Theorem 2.1, we establish Theorem 3.1.

Proof of Theorem 3.1. To establish the theorem, we assume that $\hat{\boldsymbol{V}} \not\supseteq \boldsymbol{\mathcal{S}}_r$ and exhibit a sequence $\mathcal{T}_h$ of shape regular meshes ($h = 1, 1/2, 1/3, \dots$) of the unit square

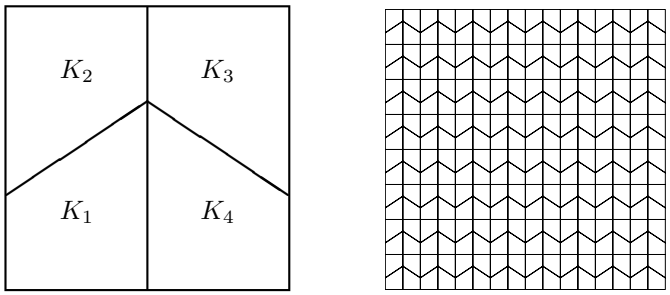


FIG. 1. (a) *The mesh* $\mathcal{T}_1$ *of the unit square into four trapezoids.* (b) *The mesh* $\mathcal{T}_h$ *(here* $h = 1/8$*) composed of translated dilates of* $\mathcal{T}_1$.

for which the estimate (3.1) does not hold. We know, by Lemma 3.3, that for either $\alpha = 1$ or $\alpha = 2$ either $\boldsymbol{P}_{F^\alpha}\hat{\boldsymbol{V}}$ or $\boldsymbol{P}_{G^\alpha}\hat{\boldsymbol{V}}$ does not contain $\boldsymbol{\mathcal{P}}_r(K^\alpha)$. We fix this value of α and, without loss of generality, suppose that

$$\boldsymbol{P}_{F^\alpha}\hat{\boldsymbol{V}} \not\supseteq \boldsymbol{\mathcal{P}}_r(K^\alpha). \tag{3.4}$$

Set $\beta = \alpha/(1+2\alpha)$. As shown in Figure 1(a), we define a mesh $\mathcal{T}_1$ consisting of four congruent elements $K_1, \ldots, K_4$, with the vertices of K_1 given by $(0,0)$, $(1/2,0)$, $(1/2, 1-\beta)$, and $(0,\beta)$. For $h = 1/n$, we construct the mesh $\mathcal{T}_h$ by partitioning the unit square into n^2 subsquares K and meshing each subsquare K with the mesh obtained by applying $\boldsymbol{F}_K$, given by (2.1), to $\mathcal{T}_1$ as shown in Figure 1(b). For each element T of the mesh $\mathcal{T}_h$ there is a natural way to construct a bilinear mapping $\boldsymbol{F}$ from the unit square onto T based on the mapping $\boldsymbol{F}^\alpha$. The first step is to compose $\boldsymbol{F}^\alpha$ with the linear isomorphism $\boldsymbol{E}(\boldsymbol{x}) = (x_1/2, x_2/(1+2\alpha))$ to obtain a bilinear map from the unit square onto the trapezoid K_1. Composing further with the natural isometries of K_1 onto K_2, K_3, and K_4, we obtain bilinear maps $\boldsymbol{F}_j$ from the unit square onto each of the trapezoids K_j, $j = 1, \ldots, 4$. Finally, further composition with the map $\boldsymbol{F}_K$ (consisting of dilation and translation) taking the unit square onto the subsquare K containing T, defines a bilinear diffeomorphism of the unit square onto T.

Having specified the mesh $\mathcal{T}_h$ and a bilinear map from the unit square onto each element of the mesh, we have determined the space $\boldsymbol{V}(\mathcal{T}_h)$ based on the shape functions in $\hat{\boldsymbol{V}}$. We need to show that the estimate (3.1) does not hold. To do so, we observe that $\boldsymbol{V}(\mathcal{T}_h)$ coincides precisely with the space $\boldsymbol{V}_h$ constructed at the start of section 2 (see (2.2)) if we use $\boldsymbol{V}(\mathcal{T}_1)$ as the space of shape functions on the unit square to begin the construction. This observation is easily verified in view of the composition property (1.2) of the Piola transform. Thus we may invoke Theorem 2.1 to conclude that (3.1) does not hold if we can show that $\boldsymbol{V}(\mathcal{T}_1) \not\supseteq \boldsymbol{\mathcal{P}}_r(\hat{K})$. Now, by construction, the functions in $\boldsymbol{V}(\mathcal{T}_1)$ restrict to functions in $\boldsymbol{P}_{F_1}\hat{\boldsymbol{V}}$ on $K_1 = \boldsymbol{F}_1\hat{K}$, so it is enough to show that $\boldsymbol{P}_{F_1}\hat{\boldsymbol{V}} \not\supseteq \boldsymbol{\mathcal{P}}_r(K_1)$. But $\boldsymbol{F}_1 = \boldsymbol{E} \circ \boldsymbol{F}^\alpha$ and, hence, $\boldsymbol{P}_{F_1}\hat{\boldsymbol{V}} = \boldsymbol{P}_E(\boldsymbol{P}_{F^\alpha}\hat{\boldsymbol{V}})$. Now $\boldsymbol{E}$ is a linear isomorphism of K^α onto K_1, and so $\boldsymbol{P}_E$ is a linear isomorphism of $\boldsymbol{\mathcal{P}}_r(K^\alpha)$ onto $\boldsymbol{\mathcal{P}}_r(K_1)$. Thus $\boldsymbol{P}_{F_1}\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{P}}_r(K_1)$ if and only if $\boldsymbol{P}_{F^\alpha}\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{P}}_r(K^\alpha)$ and so the theorem is complete in view of (3.4). ∎

Proof of Theorem 3.2. The proof is essentially identical to the preceding one, except that Lemma 3.4 and Theorem 2.2 are used in place of Lemma 3.3 and Theorem 2.1. ∎

Before turning to the proof of Lemmas 3.3 and 3.4, we draw some implications from Theorems 3.1 and 3.2 for the approximation properties of the extensions of stan-

dard finite element subspaces of $\boldsymbol{H}(\mathrm{div},\Omega)$ from rectangular meshes to quadrilateral meshes. By definition, $\boldsymbol{\mathcal{S}}_r \subseteq \boldsymbol{\mathcal{RT}}_r$, so Theorem 3.1 does not contradict the possibility that the Raviart–Thomas space of index r achieves order $r+1$ approximation in $\boldsymbol{L}^2$ on quadrilateral meshes, just as for rectangular meshes. This is indeed the case (see the discussion in section 1). But $\mathrm{div}\,\boldsymbol{\mathcal{RT}}_r = \mathcal{Q}_r$ which contains $\mathcal{R}_{r-1}$ but not $\mathcal{R}_r$. Thus we may conclude from Theorem 3.2 that the best possible order of approximation to the divergence in L^2 for the Raviart–Thomas space of index r is only r on quadrilateral meshes, one degree lower than for rectangular meshes, and, in particular, there is no convergence for $r=0$. (This lower order is achieved, as discussed in section 1.) In contrast to the Raviart–Thomas spaces, for the Brezzi–Douglas–Marini and Brezzi–Douglas–Fortin–Marini spaces there is a loss of L^2-approximation order on quadrilateral meshes. Both $\boldsymbol{\mathcal{BDM}}_r$ and $\boldsymbol{\mathcal{BDFM}}_{r+1}$ contain $\boldsymbol{\mathcal{P}}_r$, which is enough to ensure order $r+1$ approximation in $\boldsymbol{L}^2$ on rectangular meshes. However, it is easy to check that $\boldsymbol{\mathcal{BDM}}_r$ contains $\boldsymbol{\mathcal{S}}_{\lfloor (r-1)/2\rfloor}$ but not $\boldsymbol{\mathcal{S}}_{\lfloor (r+1)/2\rfloor}$ so that the best possible order of approximation for the Brezzi–Douglas–Marini space of index r on general quadrilateral meshes is $\lfloor (r+1)/2\rfloor$, a substantial loss of accuracy in comparison to the rectangular case. For the divergence, we have $\hat{\mathrm{div}}\,\boldsymbol{\mathcal{BDM}}_r = \mathcal{P}_{r-1}(\hat{K})$ which contains $\mathcal{R}_{\lfloor (r-2)/2\rfloor}$ but not $\mathcal{R}_{\lfloor r/2\rfloor}$. Therefore the best possible order of approximation for the divergence for the Brezzi–Douglas–Marini space of index r on general quadrilateral meshes is $\mathcal{R}_{\lfloor r/2\rfloor}$. Similarly, the best possible order of L^2-approximation for the Brezzi–Douglas–Fortin–Marini space of index $r+1$ on general quadrilateral meshes is $\lfloor (r+2)/2\rfloor$, while since $\hat{\mathrm{div}}\,\boldsymbol{\mathcal{BDFM}}_{r+1} = \mathcal{P}_r(\hat{K})$, the best possible rate for the divergence is $\lfloor (r+1)/2\rfloor$. We specifically note that in the lowest index cases, namely when $\hat{\boldsymbol{V}} = \boldsymbol{\mathcal{RT}}_0$, $\boldsymbol{\mathcal{BDM}}_1$, or $\boldsymbol{\mathcal{BDFM}}_1$ (which is identical to $\boldsymbol{\mathcal{RT}}_0$), *the best approximation in* $\boldsymbol{H}(\mathrm{div},\Omega)$ *does not converge in* $\boldsymbol{H}(\mathrm{div},\Omega)$ *for general quadrilateral mesh sequences.* Section 8 of this paper contains a numerical confirmation of this result.

We conclude this section with the proofs of Lemmas 3.3 and 3.4.

Proof of Lemma 3.3. By hypothesis $\boldsymbol{P}_F\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{P}}_r(\boldsymbol{F}(\hat{K}))$ or, equivalently, $\hat{\boldsymbol{V}} \supseteq \boldsymbol{P}_F^{-1}[\boldsymbol{\mathcal{P}}_r(\boldsymbol{F}(\hat{K}))]$, for $\boldsymbol{F} = \boldsymbol{F}^1$, $\boldsymbol{F}^2$, $\boldsymbol{G}^1$, and $\boldsymbol{G}^2$. Thus it is sufficient to prove that

$$\boldsymbol{\mathcal{S}}_r \subseteq \boldsymbol{\Sigma}_r := \boldsymbol{P}_{F^1}^{-1}[\boldsymbol{\mathcal{P}}_r(K^1)] + \boldsymbol{P}_{F^2}^{-1}[\boldsymbol{\mathcal{P}}_r(K^2)] + \boldsymbol{P}_{G^1}^{-1}[\boldsymbol{\mathcal{P}}_r(K^1)] + \boldsymbol{P}_{G^2}^{-1}[\boldsymbol{\mathcal{P}}_r(K^2)]. \tag{3.5}$$

We will prove this using induction on r.

Now for any diffeomorphism $\boldsymbol{F}:\hat{K}\to K$ and any $\boldsymbol{u}:K\to\mathbb{R}^2$, we have, directly from the definition of the Piola transform, that

$$(\boldsymbol{P}_F^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}}) = J\boldsymbol{F}(\hat{\boldsymbol{x}})DF(\hat{\boldsymbol{x}})^{-1}\boldsymbol{u}(\boldsymbol{x}) = \begin{pmatrix} \dfrac{\partial F_2}{\partial \hat{x}_2}(\hat{\boldsymbol{x}}) & -\dfrac{\partial F_1}{\partial \hat{x}_2}(\hat{\boldsymbol{x}}) \\ -\dfrac{\partial F_2}{\partial \hat{x}_1}(\hat{\boldsymbol{x}}) & \dfrac{\partial F_1}{\partial \hat{x}_1}(\hat{\boldsymbol{x}}) \end{pmatrix} \boldsymbol{u}(\boldsymbol{x}). \tag{3.6}$$

Specializing to the case where $\boldsymbol{F} = \boldsymbol{F}^\alpha$ or $\boldsymbol{G}^\alpha$ given by (3.3), we have

$$(\boldsymbol{P}_{F^\alpha}^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}}) = \begin{pmatrix} \alpha+\hat{x}_1 & 0 \\ -\hat{x}_2 & 1 \end{pmatrix}\boldsymbol{u}(\boldsymbol{x}), \qquad (\boldsymbol{P}_{G^\alpha}^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}}) = \begin{pmatrix} \hat{x}_1 & -1 \\ -\alpha-\hat{x}_2 & 0 \end{pmatrix}\boldsymbol{u}(\boldsymbol{x}).$$

Thus, when $u(\boldsymbol{x})$ is the constant vector field $(1,0)$, $(\boldsymbol{P}_{F^1}^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}}) = (1+\hat{x}_1, -\hat{x}_2)$, and when $u(\boldsymbol{x}) \equiv (0,1)$, $(\boldsymbol{P}_{F^1}^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}}) = (0,1)$ and $(\boldsymbol{P}_{G^1}^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}}) = (-1,0)$. These three vector fields span $\boldsymbol{\mathcal{S}}_0$, which establishes (3.5) in the case $r=0$.

Suppose now that $\boldsymbol{\mathcal{S}}_{r-1} \subseteq \boldsymbol{\Sigma}_{r-1}$ for some $r \geq 1$. To complete the induction we need to show that $\boldsymbol{\mathcal{S}}_r \subseteq \boldsymbol{\Sigma}_r$. Now $\boldsymbol{\mathcal{S}}_r$ is spanned by $\boldsymbol{\mathcal{S}}_{r-1}$ plus the $4r+4$ additional

vector fields

$$\begin{array}{llll}(\hat{x}_1^i\hat{x}_2^r,0) & \text{and} & (0,\hat{x}_1^r\hat{x}_2^i), & 0\le i\le r,\\ (\hat{x}_1^{r+1}\hat{x}_2^j,0) & \text{and} & (0,\hat{x}_1^j\hat{x}_2^{r+1}), & 0\le j\le r-1,\\ (\hat{x}_1^r\hat{x}_2^{r-1},0) & \text{and} & (\hat{x}_1^{r+1}\hat{x}_2^r,-\hat{x}_1^r\hat{x}_2^{r+1}).\end{array}$$

Pick $0\le i\le r$, and set $\boldsymbol{F}=\boldsymbol{G}^\alpha$ and $\boldsymbol{u}(\boldsymbol{x})=(0,-x_1^{r-i}x_2^i)\in\boldsymbol{\mathcal{P}}_r(K^\alpha)$. Note that $\boldsymbol{x}=\boldsymbol{G}^\alpha\hat{\boldsymbol{x}}=(\hat{x}_2,(\alpha+\hat{x}_2)\hat{x}_1)$. Then

$$(\boldsymbol{P}_{G^\alpha}^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}})=(x_1^{r-i}x_2^i,0)=(\hat{x}_2^{r-i}(\alpha+\hat{x}_2)^i\hat{x}_1^i,0)=(\hat{x}_1^i\hat{x}_2^r,0)+i\alpha(\hat{x}_1^i\hat{x}_2^{r-1},0)\pmod{\boldsymbol{\mathcal{S}}_{r-1}}.$$

Since $\boldsymbol{\mathcal{S}}_{r-1}\subseteq\boldsymbol{\Sigma}_r$ by the inductive hypothesis, and since we may take both $\alpha=1$ and $\alpha=2$, we conclude that $(\hat{x}_1^i\hat{x}_2^r,0)\in\boldsymbol{\Sigma}_r$ (for $0\le i\le r$) and also that $(\hat{x}_1^r\hat{x}_2^{r-1},0)\in\boldsymbol{\Sigma}_r$.

In a similar way, setting $\boldsymbol{F}=\boldsymbol{F}^\alpha$ and $\boldsymbol{u}(\boldsymbol{x})=(0,x_1^{r-i}x_2^i)$, we conclude that $(0,\hat{x}_1^r\hat{x}_2^i)\in\boldsymbol{\Sigma}_r$, $0\le i\le r$. The choice $\boldsymbol{F}=\boldsymbol{F}^\alpha$ and $\boldsymbol{u}(\boldsymbol{x})=(x_1^{r-j}x_2^j,0)$ together with the fact that $\boldsymbol{\Sigma}_r\supseteq\mathcal{Q}_r\times\mathcal{Q}_r$, which is a consequence of the proof thus far, implies that $(\hat{x}_1^{r+1}\hat{x}_2^j,0)\in\boldsymbol{\Sigma}_r$ for $0\le j\le r-1$. The choice $\boldsymbol{F}=\boldsymbol{G}^\alpha$ with the same choice of $\boldsymbol{u}$ similarly implies that $(0,\hat{x}_1^j\hat{x}_2^{r+1})\in\boldsymbol{\Sigma}_r$ for $0\le j\le r-1$.

Finally, with $\boldsymbol{u}(\boldsymbol{x})=(x_2^r,0)$, we find that $(\boldsymbol{P}_{F^1}^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}})=(\hat{x}_1^{r+1}\hat{x}_2^r,-\hat{x}_1^r\hat{x}_2^{r+1})$ (mod $\mathcal{Q}_r\times\mathcal{Q}_r$), which completes the proof of (3.5) and so the lemma. ∎

Proof of Lemma 3.4. The hypothesis is that $\operatorname{div}\boldsymbol{P}_F\hat{\boldsymbol{V}}\supseteq\mathcal{P}_r(\boldsymbol{F}(\hat{K}))$ for $\boldsymbol{F}=\boldsymbol{F}^1$, $\boldsymbol{F}^2$, $\boldsymbol{G}^1$, and $\boldsymbol{G}^2$. Now $\hat{\operatorname{div}}\,\hat{\boldsymbol{u}}(\hat{\boldsymbol{x}})=J\boldsymbol{F}(\hat{\boldsymbol{x}})\operatorname{div}(\boldsymbol{P}_F\hat{\boldsymbol{u}})(\boldsymbol{F}\hat{\boldsymbol{x}})$, so $\hat{\operatorname{div}}\,\hat{\boldsymbol{V}}$ contains all functions on $\hat{K}$ of the form $\hat{\boldsymbol{x}}\mapsto J\boldsymbol{F}(\hat{\boldsymbol{x}})p(\boldsymbol{F}\hat{\boldsymbol{x}})$ with $p\in\mathcal{P}_r(\boldsymbol{F}(\hat{K}))$ and $\boldsymbol{F}\in\{\boldsymbol{F}^1,\boldsymbol{F}^2,\boldsymbol{G}^1,\boldsymbol{G}^2\}$. To prove the lemma, it suffices to show that the span of such functions, call it Σ_r, contains $\mathcal{R}_r$. Note that $J\boldsymbol{F}^\alpha(\hat{\boldsymbol{x}})=\alpha+\hat{x}_1$ and $J\boldsymbol{G}^\alpha(\hat{\boldsymbol{x}})=-\alpha-\hat{x}_2$.

For $r=0$, we take $p\equiv1$ and $\boldsymbol{F}=\boldsymbol{F}^1$, $\boldsymbol{F}^2$, and $\boldsymbol{G}^1$, and find that Σ_r contains $1+\hat{x}_1$, $2+\hat{x}_1$, and $-1-\hat{x}_2$. These three functions span $\mathcal{R}_0$, so $\Sigma_0\supseteq\mathcal{R}_0$.

We continue the proof that $\Sigma_r\supseteq\mathcal{R}_r$ by induction on r. Now $\mathcal{R}_r$ is the span of $\mathcal{R}_{r-1}$ and the $2r+3$ additional functions $\hat{x}_1^{r+1}\hat{x}_2^i$ and $\hat{x}_1^i\hat{x}_2^{r+1}$, $0\le i\le r$, and $\hat{x}_1^r\hat{x}_2^r$. Taking $p(\boldsymbol{x})=x_1^{r-i}x_2^i$ and $\boldsymbol{F}=\boldsymbol{F}^\alpha$ we find that the function $\hat{\boldsymbol{x}}\mapsto\hat{x}_1^{r-i}(\alpha+\hat{x}_1)^{i+1}\hat{x}_2^i$ belongs to Σ_r. Modulo $\mathcal{R}_{r-1}$ (which is contained in Σ_r by the inductive hypothesis), this is equal to the function $\hat{\boldsymbol{x}}\mapsto\hat{x}_1^{r+1}\hat{x}_2^i+(i+1)\alpha\hat{x}_1^r\hat{x}_2^i$. Using both $\alpha=1$ and 2, we conclude that $\hat{x}_1^{r+1}\hat{x}_2^i$ belongs to Σ_r for $0\le i\le r$ and that $\hat{x}_1^r\hat{x}_2^r$ does as well. The same choice of p with $\boldsymbol{F}=\boldsymbol{G}^\alpha$ shows that Σ_r contains the functions $\hat{x}_1^i\hat{x}_2^{r+1}$, $0\le i\le r$, and completes the proof. ∎

4. Sufficient conditions for optimal order approximation. In this section we show that the necessary conditions we have obtained in the previous section are also sufficient for approximation of order $r+1$ in $\boldsymbol{L}^2$ and $\boldsymbol{H}(\mathrm{div},\Omega)$. To state this more precisely, we recall the construction of projection operators for $\boldsymbol{H}(\mathrm{div})$ finite elements. We suppose that we are given a bounded projection $\hat{\boldsymbol{\pi}}:\boldsymbol{H}^{r+1}(\hat{K})\to\hat{\boldsymbol{V}}$ (typically this operator is specified via a unisolvent set of degrees of freedom for $\hat{\boldsymbol{V}}$). We then define the corresponding projection $\boldsymbol{\pi}_K:\boldsymbol{H}^{r+1}(K)\to\boldsymbol{P}_F\hat{\boldsymbol{V}}$ for an arbitrary element $K=\boldsymbol{F}(\hat{K})$ via the Piola transform, as expressed in this commuting diagram:

$$\begin{array}{ccc}\boldsymbol{H}^{r+1}(\hat{K}) & \xrightarrow{\hat{\boldsymbol{\pi}}} & \hat{\boldsymbol{V}}\\ \Big\downarrow{\scriptstyle\boldsymbol{P}_F} & & \Big\downarrow{\scriptstyle\boldsymbol{P}_F}\\ \boldsymbol{H}^{r+1}(K) & \xrightarrow[\boldsymbol{\pi}_K]{} & \boldsymbol{P}_F\hat{\boldsymbol{V}}\end{array}.$$

That is, $\boldsymbol{\pi}_K = \boldsymbol{P}_F \circ \hat{\boldsymbol{\pi}} \circ \boldsymbol{P}_F^{-1}$. Finally a global projection operator $\boldsymbol{\pi}_h : \boldsymbol{H}^{r+1}(\Omega) \to \boldsymbol{V}(\mathcal{T}_h)$ is defined piecewise: $(\boldsymbol{\pi}_h \boldsymbol{u})|_K = \boldsymbol{\pi}_K(\boldsymbol{u}|_K)$. (The degrees of freedom used to define $\hat{\boldsymbol{\pi}}$ will determine the degree of interelement continuity enjoyed by $\boldsymbol{\pi}_h \boldsymbol{u}$. In particular, for the standard $\boldsymbol{H}(\mathrm{div})$ finite element spaces discussed previously, the degrees of freedom ensure that on any edge $\hat{e}$ of $\hat{K}$, $(\hat{\boldsymbol{\pi}}\hat{\boldsymbol{u}}) \cdot \hat{\boldsymbol{n}}$ on $\hat{e}$ depends only on $\boldsymbol{u} \cdot \boldsymbol{n}$ on $\hat{e}$. From this it results that $\boldsymbol{\pi}_h \boldsymbol{u} \in \boldsymbol{H}(\mathrm{div})$.)

The following two theorems contain the main results of this section.

THEOREM 4.1. *Let $\hat{\boldsymbol{\pi}} : \boldsymbol{H}^{r+1}(\hat{K}) \to \hat{\boldsymbol{V}}$ be a bounded projection operator. Given a quadrilateral mesh $\mathcal{T}_h$ of a domain Ω, let $\boldsymbol{\pi}_h : \boldsymbol{H}^{r+1}(\Omega) \to \boldsymbol{V}(\mathcal{T}_h)$ be defined as above. Suppose that $\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{S}}_r$. Then there exists a constant C depending only on the bound for $\hat{\boldsymbol{\pi}}$ and on the shape regularity of $\mathcal{T}_h$, such that*

$$\|\boldsymbol{u} - \boldsymbol{\pi}_h \boldsymbol{u}\|_{L^2(\Omega)} \le C h^{r+1} |\boldsymbol{u}|_{H^{r+1}(\Omega)} \tag{4.1}$$

for all $\boldsymbol{u} \in \boldsymbol{H}^{r+1}(\Omega)$.

THEOREM 4.2. *Let $\hat{\boldsymbol{\pi}} : \boldsymbol{H}^{r+1}(\hat{K}) \to \hat{\boldsymbol{V}}$ be a bounded projection operator. Given a quadrilateral mesh $\mathcal{T}_h$ of a domain Ω, let $\boldsymbol{\pi}_h : \boldsymbol{H}^{r+1}(\Omega) \to \boldsymbol{V}(\mathcal{T}_h)$ be defined as above. Suppose that $\hat{\mathrm{div}}\, \hat{\boldsymbol{V}} \supseteq \mathcal{R}_r$. Suppose also that there exists a bounded projection operator $\hat{\Pi} : H^{r+1}(\hat{K}) \to \hat{\mathrm{div}}\, \hat{\boldsymbol{V}}$ such that*

$$\hat{\mathrm{div}}\, \hat{\boldsymbol{\pi}} \hat{\boldsymbol{u}} = \hat{\Pi}\, \hat{\mathrm{div}}\, \hat{\boldsymbol{u}} \quad \forall \hat{\boldsymbol{u}} \in \boldsymbol{H}^{r+1}(\hat{K}) \text{ with } \hat{\mathrm{div}}\, \hat{\boldsymbol{u}} \in H^{r+1}(\hat{K}). \tag{4.2}$$

Then there exists a constant C depending only on the bounds for $\hat{\boldsymbol{\pi}}$ and $\hat{\Pi}$ and on the shape regularity of $\mathcal{T}_h$, such that

$$\|\mathrm{div}\, \boldsymbol{u} - \mathrm{div}\, \boldsymbol{\pi}_h \boldsymbol{u}\|_{L^2(\Omega)} \le C h^{r+1} |\mathrm{div}\, \boldsymbol{u}|_{H^{r+1}(\Omega)} \tag{4.3}$$

for all $\boldsymbol{u} \in \boldsymbol{H}^{r+1}(\Omega)$ with $\mathrm{div}\, \boldsymbol{u} \in H^{r+1}(\Omega)$.

Remarks. 1. It follows immediately that if the hypotheses of both theorems are met, then $\boldsymbol{\pi}_h$ furnishes order $r+1$ approximation in $\boldsymbol{H}(\mathrm{div}, \Omega)$:

$$\|\boldsymbol{u} - \boldsymbol{\pi}_h \boldsymbol{u}\|_{\boldsymbol{H}(\mathrm{div},\Omega)} \le C h^{r+1} (|\boldsymbol{u}|_{H^{r+1}(\Omega)} + |\mathrm{div}\, \boldsymbol{u}|_{H^{r+1}(\Omega)})$$

for all $\boldsymbol{u} \in \boldsymbol{H}^{r+1}(\Omega)$ with $\mathrm{div}\, \boldsymbol{u} \in H^{r+1}(\Omega)$.

2. The commutativity hypothesis involving the projection $\hat{\Pi}$ plays a major role in the theory of $\boldsymbol{H}(\mathrm{div}, \Omega)$ finite elements. It is satisfied in the case of the Raviart–Thomas, Brezzi–Douglas–Marini, and Brezzi–Douglas–Fortin–Marini elements, as well as for the new elements introduced in the next sections, with $\hat{\Pi}$ equal to the L^2 projection onto $\hat{\mathrm{div}}\, \hat{\boldsymbol{V}}$.

3. When applied to the Raviart–Thomas elements of index r, Theorem 4.1 gives

$$\|\boldsymbol{u} - \boldsymbol{\pi}_h \boldsymbol{u}\|_{L^2(\Omega)} \le C h^{r+1} |\boldsymbol{u}|_{H^{r+1}(\Omega)}$$

and Theorem 4.2 gives

$$\|\mathrm{div}\, \boldsymbol{u} - \mathrm{div}\, \boldsymbol{\pi}_h \boldsymbol{u}\|_{L^2(\Omega)} \le C h^{r} |\mathrm{div}\, \boldsymbol{u}|_{H^{r+1}(\Omega)}.$$

The latter estimate is proved in [12], but the former estimate appears to be new. It improves on the estimate given in [12]:

$$\|\boldsymbol{u} - \boldsymbol{\pi}_h \boldsymbol{u}\|_{L^2(\Omega)} \le C h^{r+1} [|\boldsymbol{u}|_{H^{r+1}(\Omega)} + h |\mathrm{div}\, \boldsymbol{u}|_{H^{r+1}(\Omega)}].$$

The proofs of the theorems depend on the following two lemmas which are strengthened converses of Lemmas 3.3 and 3.4.

LEMMA 4.3. *Let $\hat{\boldsymbol{V}}$ be a space of vector fields on $\hat{K}$ containing $\boldsymbol{\mathcal{S}}_r$. Then $\boldsymbol{P}_F\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{P}}_r(K)$ for all bilinear isomorphisms $\boldsymbol{F}$ of $\hat{K}$ onto convex quadrilaterals $K = \boldsymbol{F}(\hat{K})$.*

Proof. It is sufficient to show that $\boldsymbol{\mathcal{S}}_r \supseteq \boldsymbol{P}_F^{-1}[\boldsymbol{\mathcal{P}}_r(K)]$, since then the hypothesis $\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{S}}_r$ implies that

$$\boldsymbol{P}_F\hat{\boldsymbol{V}} \supseteq \boldsymbol{P}_F\boldsymbol{\mathcal{S}}_r \supseteq \boldsymbol{P}_F\boldsymbol{P}_F^{-1}[\boldsymbol{\mathcal{P}}_r(K)] = \boldsymbol{\mathcal{P}}_r(K).$$

Now (3.6) tells us that

$$\boldsymbol{P}_F^{-1}\boldsymbol{u} = \begin{pmatrix} \partial F_2/\partial\hat{x}_2 & -\partial F_1/\partial\hat{x}_2 \\ -\partial F_2/\partial\hat{x}_1 & \partial F_1/\partial\hat{x}_1 \end{pmatrix} (\boldsymbol{u}\circ\boldsymbol{F}).$$

Since $\boldsymbol{u} \in \boldsymbol{\mathcal{P}}_r(K)$ and $\boldsymbol{F}$ is bilinear, $\boldsymbol{u}\circ\boldsymbol{F} \in \boldsymbol{\mathcal{Q}}_r(\hat{K})$. Also, again in view of the bilinearity of $\boldsymbol{F}$, the matrix appearing in this equation is the sum of a constant matrix field and one of the form $(\hat{x}_1, -\hat{x}_2)^T(a_2, -a_1)$ (where $a_i \in \mathbb{R}$ is the coefficient of $\hat{x}_1\hat{x}_2$ in F_i). It follows immediately that $\boldsymbol{P}_F^{-1}\boldsymbol{u} \in \boldsymbol{\mathcal{S}}_r$. ∎

LEMMA 4.4. *Let $\hat{\boldsymbol{V}}$ be a space of vector fields on $\hat{K}$ such that $\hat{\operatorname{div}}\,\hat{\boldsymbol{V}} \supseteq \mathcal{R}_r$. Then $\operatorname{div}\boldsymbol{P}_F\hat{\boldsymbol{V}} \supseteq \mathcal{P}_r(K)$ for all bilinear isomorphisms $\boldsymbol{F}$ of $\hat{K}$ onto convex quadrilaterals $K = \boldsymbol{F}(\hat{K})$.*

Proof. Let $p \in \mathcal{P}_r(K)$ be arbitrary. Choose any $\boldsymbol{u} \in \boldsymbol{H}(\operatorname{div},\Omega)$ such that $\operatorname{div}\boldsymbol{u} = p$. From the identity

$$(\hat{\operatorname{div}}\,\boldsymbol{P}_F^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}}) = J\boldsymbol{F}(\hat{\boldsymbol{x}})(\operatorname{div}\boldsymbol{u})(\boldsymbol{x}),$$

we have $\hat{\operatorname{div}}\,\boldsymbol{P}_F^{-1}\boldsymbol{u} = J\boldsymbol{F}\cdot(p\circ\boldsymbol{F})$. Now $p \in \mathcal{P}_r(K)$ and $\boldsymbol{F}$ is bilinear, so $p\circ\boldsymbol{F}$ belongs to $\mathcal{Q}_r(\hat{K})$ and $J\boldsymbol{F}$ is linear. Thus $\hat{q} := \hat{\operatorname{div}}\,\boldsymbol{P}_F^{-1}\boldsymbol{u} \in \mathcal{R}_r$.

Invoking the hypothesis that $\mathcal{R}_r \subseteq \hat{\operatorname{div}}\,\hat{\boldsymbol{V}}$, we can find $\hat{\boldsymbol{v}} \in \hat{\boldsymbol{V}}$ such that $\hat{\operatorname{div}}\,\hat{\boldsymbol{v}} = \hat{q}$. Then

$$\begin{aligned} p(\boldsymbol{x}) &= \operatorname{div}\boldsymbol{u}(\boldsymbol{x}) = J\boldsymbol{F}(\hat{\boldsymbol{x}})^{-1}(\hat{\operatorname{div}}\,\boldsymbol{P}_F^{-1}\boldsymbol{u})(\hat{\boldsymbol{x}}) \\ &= J\boldsymbol{F}(\hat{\boldsymbol{x}})^{-1}\hat{q}(\hat{\boldsymbol{x}}) = J\boldsymbol{F}(\hat{\boldsymbol{x}})^{-1}\,\hat{\operatorname{div}}\,\hat{\boldsymbol{v}}(\hat{\boldsymbol{x}}) = \operatorname{div}\boldsymbol{P}_F\hat{\boldsymbol{v}}(\boldsymbol{x}). \end{aligned}$$

This shows that $p \in \operatorname{div}\boldsymbol{P}_F\hat{\boldsymbol{V}}$ as required. ∎

Proof of Theorem 4.1. We will show that if $\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{S}}_r$ and K is any convex quadrilateral, then

$$\|\boldsymbol{u} - \boldsymbol{\pi}_K\boldsymbol{u}\|_{L^2(K)} \le Ch_K^{r+1}|\boldsymbol{u}|_{H^{r+1}(K)} \quad \forall\boldsymbol{u} \in \boldsymbol{H}^{r+1}(K), \tag{4.4}$$

where $h_K = \operatorname{diam}(K)$ and the constant C depends only on $\hat{\boldsymbol{\pi}}$ and the shape constant for K. The theorem follows easily by squaring both sides and summing over the elements.

We establish (4.4) in two steps. First we prove it under the additional assumption that $h_K = 1$, and then we use a simple scaling argument to obtain it for arbitrary K.

For the first part we use the Bramble–Hilbert lemma. In view of Lemma 4.3 and the fact that $\hat{\boldsymbol{\pi}}$ is a projection onto $\hat{\boldsymbol{V}}$, it follows that $\boldsymbol{\pi}_K\boldsymbol{u} = \boldsymbol{u}$ for all $\boldsymbol{u} \in \boldsymbol{\mathcal{P}}_r(K)$. Now under the assumption that $h_K = 1$, the Piola transform $\boldsymbol{P}_{F_K}$ is bounded and invertible both from $\boldsymbol{L}^2(\hat{K})$ to $\boldsymbol{L}^2(K)$ and from $\boldsymbol{H}^{r+1}(\hat{K})$ to $\boldsymbol{H}^{r+1}(K)$ with bounds in both norms depending only on the shape constant. A similar statement holds for $\boldsymbol{P}_{F_K}^{-1}$.

Since $\hat{\boldsymbol{\pi}}$ is bounded from $\boldsymbol{H}^{r+1}(\hat{K})$ to $\boldsymbol{L}^2(\hat{K})$, it follows that $\boldsymbol{\pi}_K = \boldsymbol{P}_{F_K} \circ \hat{\boldsymbol{\pi}} \circ \boldsymbol{P}_{F_K}^{-1}$ is bounded from $\boldsymbol{H}^{r+1}(K)$ to $\boldsymbol{L}^2(K)$ with bound depending only on the bound for $\hat{\boldsymbol{\pi}}$ and the shape constant for K. The map $\boldsymbol{u} \mapsto \boldsymbol{u} - \boldsymbol{\pi}_K \boldsymbol{u}$ is then similarly bounded and moreover vanishes on $\boldsymbol{\mathcal{P}}_r(K)$. Therefore,

$$\|\boldsymbol{u} - \boldsymbol{\pi}_K \boldsymbol{u}\|_{L^2(K)} \le \|\boldsymbol{I} - \boldsymbol{\pi}_K\|_{\mathcal{L}(\boldsymbol{H}^{r+1}(K), \boldsymbol{L}^2(K))} \inf_{\boldsymbol{p} \in \boldsymbol{\mathcal{P}}_r(K)} \|\boldsymbol{u} - \boldsymbol{p}\|_{H^{r+1}(K)}.$$

Now the Bramble–Hilbert lemma states that the last infimum can be bounded by $c|\boldsymbol{u}|_{H^{r+1}(K)}$, where c depends only on r and the shape regularity of K (see, e.g., [2, Lemma 4.3.8]). The estimate (4.4) then follows for $h_K = 1$ with $C = c\|\boldsymbol{I} - \boldsymbol{\pi}_K\|_{\mathcal{L}(\boldsymbol{H}^{r+1}(K), \ll^2(K))}$.

To complete the proof, let K be an arbitrary convex quadrilateral, and denote by $\boldsymbol{M} : K \to \tilde{K} := h_K^{-1} K$ the dilation $\boldsymbol{M}(\boldsymbol{x}) = h_K^{-1} \boldsymbol{x}$. Then the bilinear maps $\boldsymbol{F}_K$ and $\boldsymbol{F}_{\tilde{K}}$ of the reference element $\hat{K}$ onto K and $\tilde{K}$, respectively, are related by the equation $\boldsymbol{F}_{\tilde{K}} = \boldsymbol{M} \circ \boldsymbol{F}_K$, from which it follows easily that $\boldsymbol{\pi}_{\tilde{K}} = \boldsymbol{P}_M \circ \boldsymbol{\pi}_K \circ \boldsymbol{P}_M^{-1}$. Of course, $\boldsymbol{P}_M$ has a very simple form:

$$\boldsymbol{P}_M \boldsymbol{u}(\tilde{\boldsymbol{x}}) = h_K \boldsymbol{u}(h_K \tilde{\boldsymbol{x}}).$$

Now for any $\boldsymbol{u} \in \boldsymbol{H}^{r+1}(K)$, let $\tilde{\boldsymbol{u}} = \boldsymbol{P}_M \boldsymbol{u} \in \boldsymbol{H}^{r+1}(\tilde{K})$. It is then easy to check that

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{\pi}_K \boldsymbol{u}\|_{L^2(K)} &= \|\boldsymbol{P}_M^{-1}(\tilde{\boldsymbol{u}} - \boldsymbol{\pi}_{\tilde{K}} \tilde{\boldsymbol{u}})\|_{L^2(K)} = \|\tilde{\boldsymbol{u}} - \boldsymbol{\pi}_{\tilde{K}} \tilde{\boldsymbol{u}}\|_{L^2(\tilde{K})} \\ &\le C|\tilde{\boldsymbol{u}}|_{H^{r+1}(\tilde{K})} = C h_K^{r+1} |\boldsymbol{u}|_{H^{r+1}(K)}, \end{aligned}$$

where we obtained the inequality from the already established result for elements of unit diameter. ∎

Proof of Theorem 4.2. As for the previous theorem, it suffices to prove a local result:

$$\begin{aligned} \|\operatorname{div} \boldsymbol{u} - \operatorname{div} \boldsymbol{\pi}_K \boldsymbol{u}\|_{L^2(K)} &\le C h_K^{r+1} |\operatorname{div} \boldsymbol{u}|_{H^{r+1}(K)} \\ &\qquad \forall \boldsymbol{u} \in \boldsymbol{H}^{r+1}(K) \text{ with } \operatorname{div} \boldsymbol{u} \in H^{r+1}(K), \end{aligned} \tag{4.5}$$

where C depends only on the bounds for $\hat{\boldsymbol{\pi}}$ and $\hat{\Pi}$ and the shape constant of K.

Define $\Lambda_K : L^2(K) \to L^2(K)$ by

$$\Lambda_K p(\boldsymbol{x}) = J\boldsymbol{F}(\hat{\boldsymbol{x}})^{-1} \hat{\Pi}[J\boldsymbol{F} \cdot (p \circ \boldsymbol{F})](\hat{\boldsymbol{x}}), \tag{4.6}$$

i.e., $\Lambda_K p = \{J\boldsymbol{F}^{-1} \cdot \hat{\Pi}[J\boldsymbol{F} \cdot (p \circ \boldsymbol{F})]\} \circ \boldsymbol{F}^{-1}$. Then

$$\begin{aligned} \operatorname{div} \boldsymbol{\pi}_K \boldsymbol{u}(\boldsymbol{x}) &= \operatorname{div}(\boldsymbol{P}_{F_K} \hat{\boldsymbol{\pi}} \boldsymbol{P}_{F_K}^{-1} \boldsymbol{u})(\boldsymbol{x}) = J\boldsymbol{F}(\hat{\boldsymbol{x}})^{-1} \hat{\operatorname{div}}(\hat{\boldsymbol{\pi}} \boldsymbol{P}_{F_K}^{-1} \boldsymbol{u})(\hat{\boldsymbol{x}}) \\ &= J\boldsymbol{F}(\hat{\boldsymbol{x}})^{-1} \hat{\Pi}(\hat{\operatorname{div}} \boldsymbol{P}_{F_K}^{-1} \boldsymbol{u})(\hat{\boldsymbol{x}}) = J\boldsymbol{F}(\hat{\boldsymbol{x}})^{-1} \hat{\Pi}[J\boldsymbol{F} \cdot (\operatorname{div} \boldsymbol{u}) \circ \boldsymbol{F}](\hat{\boldsymbol{x}}). \end{aligned}$$

That is, $\operatorname{div} \boldsymbol{\pi}_K \boldsymbol{u} = \Lambda_K(\operatorname{div} \boldsymbol{u})$. Thus,

$$\|\operatorname{div} \boldsymbol{u} - \operatorname{div} \boldsymbol{\pi}_K \boldsymbol{u}\|_{L^2(K)} = \|\operatorname{div} \boldsymbol{u} - \Lambda_K(\operatorname{div} \boldsymbol{u})\|_{L^2(K)}$$

and (4.5) will hold if we can prove that

$$\|p - \Lambda_K p\|_{L^2(K)} \le C h_K^{r+1} |p|_{H^{r+1}(K)} \quad \forall p \in H^{r+1}(K). \tag{4.7}$$

The proof of (4.7) is again given first in the case of elements of unit diameter. Then Λ_K is bounded uniformly from $H^{r+1}(K)$ to $L^2(K)$ for elements K with uniformly

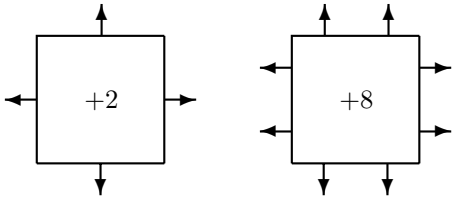


FIG. 2. *Element diagrams indicating the degrees of freedom for* $\boldsymbol{\mathcal{ABF}}_0$ *and* $\boldsymbol{\mathcal{ABF}}_1$.

bounded shape constant. Now, as noted in the proof of Lemma 4.4, if $p \in \mathcal{P}_r(K)$, then $J\boldsymbol{F} \cdot (p \circ \boldsymbol{F}) \in \mathcal{R}_r \subseteq \hat{\mathrm{div}}\, \hat{\boldsymbol{V}}$. Since $\hat{\Pi}$ is a projection onto $\hat{\mathrm{div}}\, \hat{\boldsymbol{V}}$, it follows that $\Lambda_K p = p$ for $p \in \mathcal{P}_r(K)$. Thus the Bramble–Hilbert lemma implies (4.7) under the restriction $h_K = 1$. To extend to elements of arbitrary diameter, we again use a dilation. ∎

5. Construction of spaces with optimal order $\boldsymbol{H}(\mathrm{div}, \Omega)$ approximation. We have previously shown that none of the standard finite element approximations of $\boldsymbol{H}(\mathrm{div}, \Omega)$ (i.e., the Raviart–Thomas, Brezzi–Douglas–Marini, or Brezzi–Douglas–Fortin–Marini spaces) maintain the same order of approximation on general convex quadrilaterals as they do on rectangles. In this section, we use the conditions determined in the previous sections to construct finite element subspaces of $\boldsymbol{H}(\mathrm{div}, \Omega)$ which do have this property. To obtain approximation of order $r+1$ in $\boldsymbol{H}(\mathrm{div}, \Omega)$ on general convex quadrilaterals, we require that the space of reference shape functions $\hat{\boldsymbol{V}} \supseteq \boldsymbol{\mathcal{S}}_r$ and $\hat{\mathrm{div}}\, \hat{\boldsymbol{V}} \supseteq \mathcal{R}_r$. A space with this property is $\boldsymbol{\mathcal{ABF}}_r := \mathcal{P}_{r+2,r}(\hat{K}) \times \mathcal{P}_{r,r+2}(\hat{K})$, for which $\hat{\mathrm{div}}\, \boldsymbol{\mathcal{ABF}}_r = \mathcal{R}_r$.

As degrees of freedom for $\boldsymbol{\mathcal{ABF}}_r$ on the reference element, we take

$$\int_{\hat{e}} \hat{\boldsymbol{u}} \cdot \hat{\boldsymbol{n}} \hat{q}\, d\hat{s}, \quad \hat{q} \in \mathcal{P}_r(\hat{e}) \text{ for each edge } \hat{e} \text{ of } \hat{K} \tag{5.1}$$

$$\int_{\hat{K}} \hat{\boldsymbol{u}} \cdot \hat{\boldsymbol{\phi}}\, d\hat{\boldsymbol{x}}, \quad \hat{\boldsymbol{\phi}} \in \mathcal{P}_{r-1,r}(\hat{K}) \times \mathcal{P}_{r,r-1}(\hat{K}), \tag{5.2}$$

$$\int_{\hat{K}} \hat{\mathrm{div}}\, \hat{\boldsymbol{u}} \hat{x}_1^{r+1} \hat{x}_2^i\, d\hat{\boldsymbol{x}}, \quad \int_{\hat{K}} \hat{\mathrm{div}}\, \hat{\boldsymbol{u}} \hat{x}_1^i \hat{x}_2^{r+1}\, d\hat{\boldsymbol{x}}, \quad i = 0, \ldots, r. \tag{5.3}$$

Note that (5.1) and (5.2) are the standard degrees of freedom for the Raviart–Thomas elements on the reference square. In all we have specified $4(r+1)+2r(r+1)+2(r+1) = 2(r+3)(r+1) = \dim \boldsymbol{\mathcal{ABF}}_r$ degrees of freedom. Since the new degrees of freedom, with respect to the standard Raviart–Thomas elements, are local, we remark that the implementation of the new space $\boldsymbol{\mathcal{ABF}}_r$ should not be more expensive than that of $\boldsymbol{\mathcal{RT}}_r$. Figure 2 indicates the degrees of freedom for the first two cases $r = 0$ and 1.

In order to see that these choices of $\hat{\boldsymbol{V}}$ and degrees of freedom determine a finite element subspace of $\boldsymbol{H}(\mathrm{div}, \Omega)$, we need to show that the degrees of freedom are unisolvent, and that if the degrees of freedom on an edge $\hat{e}$ vanish, then $\hat{\boldsymbol{u}} \cdot \hat{\boldsymbol{n}}$ vanishes on e (this will ensure that the assembled finite element space belongs to $\boldsymbol{H}(\mathrm{div}, \Omega)$). The second point is immediate. On any edge $\hat{e}$ of $\hat{K}$, $\boldsymbol{u} \cdot \boldsymbol{n} \in \mathcal{P}_r(\hat{e})$, so the vanishing of the degrees of freedom (5.1) associated to $\hat{e}$ does indeed ensure that $\hat{\boldsymbol{u}} \cdot \hat{\boldsymbol{n}} \equiv 0$.

We now verify unisolvence by showing that if $\hat{\boldsymbol{u}} \in \boldsymbol{\mathcal{ABF}}_r$ and all the quantities (5.1)–(5.3) vanish, then $\hat{\boldsymbol{u}} = 0$. If $\hat{q} \in \mathcal{Q}_r(\hat{K})$, then $\hat{q}|_{\hat{e}} \in \mathcal{P}_r(\hat{e})$ for any edge $\hat{e}$ of

$\hat{K}$, and $\hat{\nabla}\hat{q} \in \mathcal{P}_{r-1,r}(\hat{K}) \times \mathcal{P}_{r,r-1}(\hat{K})$. Therefore,

$$\int_{\hat{K}} \hat{\operatorname{div}}\, \hat{\boldsymbol{u}}\hat{q}\, d\hat{\boldsymbol{x}} = \int_{\partial\hat{K}} \hat{\boldsymbol{u}} \cdot \hat{\boldsymbol{n}}\hat{q}\, ds - \int_{\hat{K}} \hat{\boldsymbol{u}} \cdot \hat{\nabla}\hat{q}\, d\hat{\boldsymbol{x}} = 0, \quad \hat{q} \in \mathcal{Q}_r(\hat{K}).$$

In view of (5.3) we then have that

$$\int_{\hat{K}} \hat{\operatorname{div}}\, \hat{\boldsymbol{u}}\hat{q}\, d\hat{\boldsymbol{x}} = 0, \quad \hat{q} \in \mathcal{R}_r.$$

Since $\hat{\operatorname{div}}\, \hat{\boldsymbol{u}} \in \mathcal{R}_r$ we conclude that $\hat{\operatorname{div}}\, \hat{\boldsymbol{u}} = 0$. Now we may write

$$\hat{\boldsymbol{u}} = \sum_{i=0}^{r} [a_i(\hat{x}_1^{r+2}\hat{x}_2^i, 0) + b_i(0, \hat{x}_1^i\hat{x}_2^{r+2})] + \hat{\boldsymbol{v}}$$

with $\hat{\boldsymbol{v}} \in \boldsymbol{\mathcal{RT}}_r$. Since

$$0 = \hat{\operatorname{div}}\, \hat{\boldsymbol{u}} = \sum_{i=1}^{r} (r+2)(a_i\hat{x}_1^{r+1}\hat{x}_2^i + b_i\hat{x}_1^i\hat{x}_2^{r+1}) + \hat{\operatorname{div}}\, \hat{\boldsymbol{v}},$$

and $\hat{\operatorname{div}}\, \hat{\boldsymbol{v}} \in \mathcal{Q}_r$, it follows that $a_i = b_i = 0$ and so $\hat{\boldsymbol{u}} = \hat{\boldsymbol{v}} \in \boldsymbol{\mathcal{RT}}_r$. Since (5.1), (5.2) are unisolvent degrees of freedom for $\boldsymbol{\mathcal{RT}}_r$ [4, Proposition III.3.4], we conclude that $\hat{\boldsymbol{u}} = 0$.

We also note that a small variant of the first part of this argument establishes the commutativity property (4.2) with $\hat{\boldsymbol{\pi}} : \boldsymbol{H}^1(\hat{K}) \to \boldsymbol{\mathcal{ABF}}_r$ the projection determined by the degrees of freedom (5.1)–(5.3) and $\hat{\Pi}$ the L^2-projection onto $\mathcal{R}_r = \hat{\operatorname{div}}\, \boldsymbol{\mathcal{ABF}}_r$. Thus, all the hypotheses of Theorems 4.1 and 4.2 are satisfied and the estimates (4.1) and (4.3) hold on general quadrilateral meshes for finite element spaces based on $\boldsymbol{\mathcal{ABF}}_r$.

6. Application to mixed finite element methods. One of the main applications of finite element subspaces of $\boldsymbol{H}(\operatorname{div}, \Omega)$ is to the approximation of second order elliptic boundary value problems by mixed finite element methods. For the model problem $\Delta p = f$ in Ω, $p = 0$ on $\partial\Omega$, the mixed formulation is the following: Find $\boldsymbol{u} \in \boldsymbol{H}(\operatorname{div}, \Omega)$ and $p \in L^2(\Omega)$ such that

$$\begin{aligned} (\boldsymbol{u}, \boldsymbol{v}) + (p, \operatorname{div} \boldsymbol{v}) &= 0 \quad \forall \boldsymbol{v} \in \boldsymbol{H}(\operatorname{div}, \Omega), \\ (\operatorname{div} \boldsymbol{u}, q) &= (f, q) \quad \forall q \in L^2(\Omega), \end{aligned}$$

where $(\cdot, \cdot)$ denotes the $L^2(\Omega)$ inner product. For $\boldsymbol{S}_h \subseteq \boldsymbol{H}(\operatorname{div}, \Omega)$ and $W_h \subseteq L^2(\Omega)$, the mixed finite element approximation seeks $\boldsymbol{u}_h \in \boldsymbol{S}_h$ and $p_h \in W_h$ such that

$$\begin{aligned} (\boldsymbol{u}_h, \boldsymbol{v}) + (p_h, \operatorname{div} \boldsymbol{v}) &= 0 \quad \forall \boldsymbol{v} \in \boldsymbol{S}_h, \\ (\operatorname{div} \boldsymbol{u}_h, q) &= (f, q) \quad \forall q \in W_h. \end{aligned}$$

The pair $(\boldsymbol{S}_h, W_h)$ is said to be stable if the following conditions are satisfied:

$$(\boldsymbol{v}, \boldsymbol{v}) \geq c\|\boldsymbol{v}\|^2_{\boldsymbol{H}(\operatorname{div},\Omega)} \quad \forall \boldsymbol{v} \in \boldsymbol{Z}_h = \{\boldsymbol{v} \in \boldsymbol{S}_h : (\operatorname{div} \boldsymbol{v}, q) = 0\ \forall q \in W_h\}, \tag{6.1}$$

$$\sup_{\boldsymbol{v} \in \boldsymbol{S}_h} \frac{(\operatorname{div} \boldsymbol{v}, q)}{\|\boldsymbol{v}\|_{\boldsymbol{H}(\operatorname{div},\Omega)}} \geq c\|q\|_{L^2(\Omega)} \quad \forall q \in W_h. \tag{6.2}$$

By Brezzi's theorem [3], if $(\boldsymbol{S}_h, W_h)$ is a stable pair, then the quasioptimality estimate

(6.3)
$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{H}(\mathrm{div},\Omega)}+\|p-p_h\|_{L^2(\Omega)}\le C\Big(\inf_{\boldsymbol{v}\in\boldsymbol{S}_h}\|\boldsymbol{u}-\boldsymbol{v}\|_{\boldsymbol{H}(\mathrm{div},\Omega)}+\inf_{q\in W_h}\|p-q\|_{L^2(\Omega)}\Big)$$

holds with C depending only on Ω and the constant c entering into the stability conditions.

For the space $\boldsymbol{S}_h$ we will take $\boldsymbol{V}(\mathcal{T}_h)\cap\boldsymbol{H}(\mathrm{div},\Omega)$, where $\mathcal{T}_h$ is an arbitrary quadrilateral mesh and $\boldsymbol{V}(\mathcal{T}_h)$ is constructed as described at the start of section 3 starting from a space of reference shape functions $\hat{\boldsymbol{V}}$ on the unit square. To specify the corresponding space W_h, we first define a space of reference shape functions $\hat{W}=\hat{\mathrm{div}}\,\hat{\boldsymbol{V}}$, next define the space of shape functions on K by $W_K=\{\hat{w}\circ\boldsymbol{F}_K^{-1}\mid\hat{w}\in\hat{W}\}$, and then set

$$W_h=\{w\in L^2(\Omega)\mid w|_K\in W_K\}.$$

Now suppose that $\hat{\boldsymbol{V}}$ is any one of the previously considered spaces $\boldsymbol{\mathcal{RT}}_r$, $\boldsymbol{\mathcal{BDM}}_r$, $\boldsymbol{\mathcal{BDFM}}_{r+1}$, or $\boldsymbol{\mathcal{ABF}}_r$. Associated with each of these spaces is a unisolvent set of degrees of freedom. These are given in (5.1) and (5.2) for $\boldsymbol{\mathcal{RT}}_r$, by (5.1)–(5.3) for $\boldsymbol{\mathcal{ABF}}_r$, and, for $\boldsymbol{\mathcal{BDM}}_r$ and $\boldsymbol{\mathcal{BDFM}}_{r+1}$, by (5.1) and $\int_{\hat{K}}\hat{\boldsymbol{u}}\cdot\hat{\boldsymbol{\phi}}\,d\hat{\boldsymbol{x}}$ with $\hat{\boldsymbol{\phi}}$ in $\boldsymbol{\mathcal{P}}_{r-2}(\hat{K})$ or $\boldsymbol{\mathcal{P}}_{r-1}(\hat{K})$, respectively. These degrees of freedom determine the projection $\hat{\boldsymbol{\pi}}:\boldsymbol{H}^1(\hat{K})\to\hat{\boldsymbol{V}}$ and then, by the construction described at the start of section 4, the projection $\boldsymbol{\pi}_h:\boldsymbol{H}^1(\Omega)\to\boldsymbol{S}_h$. Moreover, the degrees of freedom ensure the commutativity property (4.2) where $\hat{\Pi}$ is the $L^2(\hat{K})$ projection onto $\hat{W}$. From these observations it is straightforward to derive the stability conditions (6.1) and (6.2), as we shall now do.

Given $\boldsymbol{v}\in\boldsymbol{Z}_h$ and $K\in\mathcal{T}_h$, let $\hat{\boldsymbol{v}}=\boldsymbol{P}_{F_K}^{-1}(\boldsymbol{v}|_K)\in\hat{\boldsymbol{V}}$, $\hat{q}=\hat{\mathrm{div}}\,\hat{\boldsymbol{v}}\in\hat{W}$, and $q=\hat{q}\circ\boldsymbol{F}_K^{-1}\in W_K$. Then $(\mathrm{div}\,\boldsymbol{v},q)_{L^2(K)}=0$ (because we can extend q to Ω by zero and obtain a function in W_h and $\mathrm{div}\,\boldsymbol{v}$ is orthogonal to W_h since $\boldsymbol{v}\in\boldsymbol{Z}_h$). But $(\mathrm{div}\,\boldsymbol{v},q)_{L^2(K)}=(\hat{\mathrm{div}}\,\hat{\boldsymbol{v}},\hat{q})_{L^2(\hat{K})}=\|\hat{\mathrm{div}}\,\hat{\boldsymbol{v}}\|^2_{L^2(K)}$, so $\hat{\mathrm{div}}\,\hat{\boldsymbol{v}}=0$ and therefore $\mathrm{div}\,\boldsymbol{v}=[(J\boldsymbol{F}_K)^{-1}\hat{\mathrm{div}}\,\hat{\boldsymbol{v}}]\circ\boldsymbol{F}_K^{-1}=0$. Thus, if $\boldsymbol{v}\in\boldsymbol{Z}_h$, then $\mathrm{div}\,\boldsymbol{v}=0$, and (6.1) follows immediately with $c=1$.

To prove (6.2), we shall show that for any given $q\in W_h$ there exists $\boldsymbol{v}\in\boldsymbol{S}_h$ with

$$(\mathrm{div}\,\boldsymbol{v},q)=\|q\|_{L^2(\Omega)}\tag{6.4}$$

and

$$\|\boldsymbol{v}\|_{\boldsymbol{H}(\mathrm{div},\Omega)}\le C\|q\|_{L^2(\Omega)}.\tag{6.5}$$

As usual, we start by noting that there exists $\boldsymbol{u}\in\boldsymbol{H}^1(\Omega)$ with $\mathrm{div}\,\boldsymbol{u}=q$ and $\|\boldsymbol{u}\|_{H^1(\Omega)}\le C\|q\|_{L^2(\Omega)}$ and letting $\boldsymbol{v}=\boldsymbol{\pi}_h\boldsymbol{u}$. Now $(\mathrm{div}\,\boldsymbol{\pi}_h\boldsymbol{u},q)=(\mathrm{div}\,\boldsymbol{u},q)$ whenever $q\in W_h$, as follows directly from the construction of $\boldsymbol{\pi}_h$, the commutativity property (4.2), and the properties of the Piola transform. Therefore (6.4) holds. To prove (6.5) we note that in each case $\hat{\boldsymbol{V}}\supseteq\boldsymbol{\mathcal{S}}_0$, so Theorem 4.1 gives the estimate $\|\boldsymbol{u}-\boldsymbol{\pi}_h\boldsymbol{u}\|_{L^2(\Omega)}\le Ch\|\boldsymbol{u}\|_{H^1(\Omega)}$, and so, by the triangle inequality, $\|\boldsymbol{v}\|_{L^2(\Omega)}\le C\|q\|_{L^2(\Omega)}$. Also, on any element K, $\mathrm{div}\,\boldsymbol{v}=\mathrm{div}\,\boldsymbol{\pi}_K\boldsymbol{u}=\Lambda_K(\mathrm{div}\,\boldsymbol{u})=\Lambda_K q$, where Λ_K is defined by (4.6), which implies that $\|\mathrm{div}\,\boldsymbol{v}\|_{L^2(\Omega)}\le C\|q\|_{L^2(\Omega)}$. This establishes (6.5) and completes the proof of stability.

Remark. Note that we do not have $W_h = \operatorname{div} \boldsymbol{S}_h$ on general quadrilateral meshes, although this is the case on rectangular meshes. With that choice of W_h it would be easy to prove (6.1) but the proof of (6.2) would not be clear.

We now turn our attention to error estimates for mixed methods. Having established stability, we can combine the quasioptimality estimate (6.3) with the bounds for the approximation error given by Theorems 4.1 and 4.2 (and Theorem 1 of [1] for the approximation error for p) to obtain error bounds. For the $\boldsymbol{\mathcal{ABF}}_r$ method this gives

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{H}(\operatorname{div},\Omega)} + \|p - p_h\|_{L^2(\Omega)} \le Ch^{r+1}(|u|_{H^{r+1}(\Omega)} + |\operatorname{div} \boldsymbol{u}|_{H^{r+1}(\Omega)} + |p|_{H^{r+1}(\Omega)}).$$

But for the $\boldsymbol{\mathcal{RT}}_r$ method it gives only an $O(h^r)$ bound, and no convergence at all for $r = 0$, because of the decreased approximation for the divergence (and the approximation orders are even lower for $\boldsymbol{\mathcal{BDM}}_r$ and $\boldsymbol{\mathcal{BDFM}}_{r+1}$).

It is possible to improve on this by following the approach of [6] and [5], as we now do. First we define $\Pi_K : L^2(K) \to W_K$ by $\Pi_K p = (\hat{\Pi}\hat{p}) \circ \boldsymbol{F}_K^{-1}$ with $\hat{p} = p \circ \boldsymbol{F}_K$, and then we define $\Pi_h : L^2(\Omega) \to W_h$ by $\Pi_h p|_K = \Pi_K(p|_K)$. It follows that $(p - \Pi_K p, \operatorname{div} \boldsymbol{v})_{L^2(K)} = (\hat{p} - \hat{\Pi}\hat{p}, \hat{\operatorname{div}}\, \boldsymbol{P}_{F_K}^{-1}\boldsymbol{v})_{L^2(K)}$, so

$$(p - \Pi_h p, \operatorname{div} \boldsymbol{v}) = 0 \quad \forall \boldsymbol{v} \in \boldsymbol{S}_h.$$

We then have the following error estimates.

THEOREM 6.1.

$$\begin{aligned}
\|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^2(\Omega)} &\le \|\boldsymbol{u} - \boldsymbol{\pi}_h \boldsymbol{u}\|_{L^2(\Omega)},\\
\|\operatorname{div} \boldsymbol{u}_h\|_{L^2(\Omega)} &\le C\|\operatorname{div} \boldsymbol{u}\|_{L^2(\Omega)},\\
\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_{L^2(\Omega)} &\le C\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{\pi}_h \boldsymbol{u})\|_{L^2(\Omega)},\\
\|\Pi_h p - p_h\|^2_{L^2(\Omega)} &= (\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{U} - \boldsymbol{\pi}_h \boldsymbol{U}) + (\operatorname{div}[\boldsymbol{u} - \boldsymbol{u}_h], P - \Pi_h P),
\end{aligned}$$

where P is the solution to the Dirichlet problem $-\Delta P = \Pi_h p - p_h$ in Ω, $P = 0$ on $\partial\Omega$ and $\boldsymbol{U} = \operatorname{grad} P$.

Proof. Using the error equations

$$(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{v}) + (p - p_h, \operatorname{div} \boldsymbol{v}) = 0 \quad \forall \boldsymbol{v} \in \boldsymbol{S}_h, \qquad (\operatorname{div}[\boldsymbol{u} - \boldsymbol{u}_h], q) = 0 \quad \forall q \in W_h,$$

we obtain

$$\begin{aligned}
(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{\pi}_h \boldsymbol{u} - \boldsymbol{u}_h) &= (p - p_h, \operatorname{div}[\boldsymbol{u}_h - \boldsymbol{\pi}_h \boldsymbol{u}]) = (\Pi_h p - p_h, \operatorname{div}[\boldsymbol{u}_h - \boldsymbol{\pi}_h \boldsymbol{u}])\\
&= (\Pi_h p - p_h, \operatorname{div}[\boldsymbol{u}_h - \boldsymbol{u}]) = 0.
\end{aligned}$$

Hence, $\|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{L^2(\Omega)} = (\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{\pi}_h \boldsymbol{u})$ and it easily follows that

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^2(\Omega)} \le \|\boldsymbol{u} - \boldsymbol{\pi}_h \boldsymbol{u}\|_{L^2(\Omega)}.$$

To estimate $\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_{L^2(\Omega)}$, we observe that if $v \in \boldsymbol{S}_h$ and we define

$$q(\boldsymbol{x}) = \begin{cases} |J\boldsymbol{F}_K(\hat{\boldsymbol{x}})| \operatorname{div} \boldsymbol{v}(\boldsymbol{x}), & \boldsymbol{x} \in K,\\ 0, & \boldsymbol{x} \in \Omega \setminus K, \end{cases}$$

then $q \in W_h$. Therefore, from the error equation we have

$$(\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h), |J\boldsymbol{F}_K| \operatorname{div} \boldsymbol{v})_K = 0.$$

Choosing $\boldsymbol{v} = \boldsymbol{u}_h$, it easily follows that

$$\||J\boldsymbol{F}_K|^{1/2}\operatorname{div}\boldsymbol{u}_h\|_{L^2(K)} \le \||J\boldsymbol{F}_K|^{1/2}\operatorname{div}\boldsymbol{u}\|_{L^2(K)},$$

and so $\|\operatorname{div}\boldsymbol{u}_h\|_{L^2(K)} \le C\|\operatorname{div}\boldsymbol{u}\|_{L^2(K)}$ with C depending on the shape constant for K. Choosing $\boldsymbol{v} = \boldsymbol{\pi}_h\boldsymbol{u} - \boldsymbol{u}_h$, it also follows that

$$\||J\boldsymbol{F}_K|^{1/2}\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_{L^2(K)} \le \||J\boldsymbol{F}_K|^{1/2}\operatorname{div}(\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u})\|_{L^2(K)},$$

so $\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_{L^2(K)} \le C\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u})\|_{L^2(K)}$. Summing over all quadrilaterals, we obtain

$$\|\operatorname{div}\boldsymbol{u}_h\|_{L^2(\Omega)} \le C\|\operatorname{div}\boldsymbol{u}\|_{L^2(\Omega)}, \qquad \|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_{L^2(\Omega)} \le C\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u})\|_{L^2(\Omega)}.$$

To estimate $\|p - p_h\|_{L^2(\Omega)}$, we define P as the solution to the Dirichlet problem $\Delta P = \Pi_h p - p_h$ in Ω, $P = 0$ on $\partial\Omega$, and set $\boldsymbol{U} = \operatorname{grad} P$. Then

$$\begin{aligned}\|\Pi_h p - p_h\|^2_{L^2(\Omega)} &= (\operatorname{div}\boldsymbol{U}, \Pi_h p - p_h) = (\operatorname{div}\boldsymbol{\pi}_h\boldsymbol{U}, \Pi_h p - p_h) = -(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{\pi}_h\boldsymbol{U})\\ &= (\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{U} - \boldsymbol{\pi}_h\boldsymbol{U}) - (\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{U})\\ &= (\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{U} - \boldsymbol{\pi}_h\boldsymbol{U}) + (\operatorname{div}[\boldsymbol{u} - \boldsymbol{u}_h], P)\\ &= (\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{U} - \boldsymbol{\pi}_h\boldsymbol{U}) + (\operatorname{div}[\boldsymbol{u} - \boldsymbol{u}_h], P - \Pi_h P). \qquad \square\end{aligned}$$

To obtain order of convergence estimates, one needs to apply the approximation properties of a particular space. For the Raviart–Thomas elements of index r we obtain the following estimates.

THEOREM 6.2. *Suppose $(\boldsymbol{u}_h, p_h)$ is the mixed method approximation to $(\boldsymbol{u}, p)$ obtained when $\hat{\boldsymbol{V}}$ is the Raviart–Thomas reference space of index r and suppose that the domain Ω is convex. Then for $p \in H^{r+2}(\Omega)$,*

$$\begin{aligned}\|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^2(\Omega)} &\le Ch^{r+1}\|\boldsymbol{u}\|_{H^{r+1}(\Omega)},\\ \|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_{L^2(\Omega)} &\le Ch^{r}\|\operatorname{div}\boldsymbol{u}\|_{H^{r}(\Omega)},\\ \|p - p_h\|_{L^2(\Omega)} &\le \begin{cases} Ch^{r+1}\|p\|_{H^{r+1}(\Omega)} & (r \ge 1),\\ Ch\|p\|_{H^2(\Omega)} & (r = 0).\end{cases}\end{aligned}$$

Proof. It follows from [7, section I.A.2] that $\|p - \Pi_h p\|_{L^2(\Omega)} \le Ch^{k+1}\|p\|_{k+1,\Omega}$, $0 \le k \le r$, and it follows from Theorems 4.1 and 4.2 that, for $0 \le k \le r$,

$$\|\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u}\|_{L^2(\Omega)} \le Ch^{k+1}\|\boldsymbol{u}\|_{H^{k+1}(\Omega)}, \qquad \|\operatorname{div}[\boldsymbol{u} - \boldsymbol{\pi}_h\boldsymbol{u}]\|_{L^2(\Omega)} \le Ch^{k}\|\operatorname{div}\boldsymbol{u}\|_{H^{k}(\Omega)}.$$

Inserting these results in Theorem 6.1, we immediately obtain the first two estimates of Theorem 6.2. From the last estimate of Theorem 6.1, we also obtain

$$\|\Pi_h p - p_h\|_{L^2(\Omega)} \le C(h\|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^2(\Omega)} + h^{\min(1+r,2)}\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_{L^2(\Omega)}).$$

Here we have used elliptic regularity, which holds under the assumption that Ω is convex, to bound $\|\boldsymbol{U}\|_{H^1(\Omega)} = \|P\|_{H^2(\Omega)}$ by $\|\Pi_h p - p_h\|_{L^2(\Omega)}$. Hence, for $r \ge 1$ and $0 \le k \le r$, we obtain

$$\|\Pi_h p - p_h\|_{L^2(\Omega)} \le Ch^{k+2}(\|\boldsymbol{u}\|_{H^{k+1}(\Omega)} + \|\operatorname{div}\boldsymbol{u}\|_{H^k(\Omega)}) \le Ch^{k+2}\|\boldsymbol{u}\|_{H^{k+1}(\Omega)}.$$

Choosing $k = r - 1$ and $k = r$, we obtain for $r \ge 1$

$$\|\Pi_h p - p_h\|_{L^2(\Omega)} \le Ch^{r+1}\|\boldsymbol{u}\|_{H^r(\Omega)}, \qquad \|\Pi_h p - p_h\|_{L^2(\Omega)} \le Ch^{r+2}\|\boldsymbol{u}\|_{H^{r+1}(\Omega)},$$

and for $r = 0$, $\|\Pi_h p - p_h\|_{L^2(\Omega)} \le Ch\|\boldsymbol{u}\|_{1,\Omega}$. The final estimates of the theorem now follow directly by the triangle inequality. $\square$

7. Application to least squares methods. A standard finite element least squares approximation of the Dirichlet problem $\Delta p = f$ in Ω, $p = 0$ on $\partial\Omega$ seeks $p_h \in W_h \subseteq H_0^1(\Omega)$ and $\boldsymbol{u}_h \in \boldsymbol{S}_h \subseteq \boldsymbol{H}(\operatorname{div}, \Omega)$ minimizing

$$J(q, \boldsymbol{v}) = \|\boldsymbol{v} - \operatorname{grad} q\|^2_{L^2(\Omega)} + \|\operatorname{div} \boldsymbol{v} + f\|^2_{L^2(\Omega)}$$

over $W_h \times \boldsymbol{S}_h$. For any choices of subspaces this satisfies the quasioptimality estimate (cf. [10])

$$\|p - p_h\|_{H^1(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{H}(\operatorname{div},\Omega)} \le C\Big(\inf_{p \in W_h} \|p - q\|_{H^1(\Omega)} + \inf_{\boldsymbol{v} \in \boldsymbol{S}_h} \|\boldsymbol{u} - \boldsymbol{v}\|_{\boldsymbol{H}(\operatorname{div},\Omega)} \Big).$$

If we take W_h to be the standard H^1 finite element space based on reference shape functions $\mathcal{Q}_{r+1}$ and use the $\boldsymbol{\mathcal{ABF}}_r$ space for $\boldsymbol{S}_h$, we immediately obtain

$$\|p - p_h\|_{H^1(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{H}(\operatorname{div},\Omega)} \le C h^{r+1} (\|p\|_{H^{r+1}(\Omega)} + \|\boldsymbol{u}\|_{H^{r+1}(\Omega)} + \|\operatorname{div} \boldsymbol{u}\|_{H^{r+1}(\Omega)}).$$

However, the quasioptimality estimate suggests that if we choose the same W_h but use the $\boldsymbol{\mathcal{RT}}_r$ elements for $\boldsymbol{S}_h$, the lower rate of approximation of $\operatorname{div} \boldsymbol{u}$ may negatively influence the approximation of both variables.

Next, we use a duality argument to obtain a second estimate, which provides improved convergence for p in L^2 when the $\boldsymbol{\mathcal{ABF}}$ spaces are used, but again suggests difficulties for the $\boldsymbol{\mathcal{RT}}$ spaces. We shall henceforth assume that the domain Ω is convex so that we have 2-regularity for the Dirichlet problem for the Laplacian. Define $\boldsymbol{w} \in \boldsymbol{H}(\operatorname{div}, \Omega)$ and $r \in H_0^1(\Omega)$ as solution of the dual problem

$$\int_\Omega (\boldsymbol{w} - \nabla r) \cdot \boldsymbol{v}\, d\boldsymbol{x} + \int_\Omega \operatorname{div} \boldsymbol{w} \operatorname{div} \boldsymbol{v}\, d\boldsymbol{x} = 0 \quad \forall \boldsymbol{v} \in \boldsymbol{H}(\operatorname{div}, \Omega), \tag{7.1}$$

$$\int_\Omega (\boldsymbol{w} - \nabla r) \cdot \nabla q\, d\boldsymbol{x} = -\int_\Omega (p - p_h) q\, d\boldsymbol{x} \quad \forall q \in H_0^1(\Omega). \tag{7.2}$$

This problem has a unique solution, since if $p - p_h$ were to vanish, then we could take $\boldsymbol{v} = \boldsymbol{w}$ and $q = r$, subtract the equations, and conclude that $\boldsymbol{w} = \nabla r$, $\operatorname{div} \boldsymbol{w} = 0$ with $r \in H_0^1(\Omega)$, which implies that $\boldsymbol{w}$ and r vanish. For general $p - p_h$, the solution of the dual problem may be written as $\boldsymbol{w} = \nabla(r + g)$ where $g \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfies $\Delta g = p - p_h$ and $r \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfies $\Delta r = g - p + p_h$ (so $\operatorname{div} \boldsymbol{w} = g$). Note that $\|r\|_{H^2(\Omega)} + \|\boldsymbol{w}\|_{H^1(\Omega)} + \|\operatorname{div} \boldsymbol{w}\|_{H^2(\Omega)} \le C \|p - p_h\|_{L^2(\Omega)}$. Choosing $q = p - p_h$, $\boldsymbol{v} = \boldsymbol{u} - \boldsymbol{u}_h$, subtracting (7.2) from (7.1), and using the error equations

$$\int_\Omega (\boldsymbol{u} - \boldsymbol{u}_h - \nabla[p - p_h]) \cdot \boldsymbol{v}\, d\boldsymbol{x} + \int_\Omega \operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h) \operatorname{div} \boldsymbol{v}\, d\boldsymbol{x} = 0 \quad \forall \boldsymbol{v} \in \boldsymbol{S}_h,$$

$$\int_\Omega (\boldsymbol{u} - \boldsymbol{u}_h - \nabla[p - p_h]) \cdot \nabla q\, d\boldsymbol{x} = 0 \quad \forall q \in W_h,$$

one obtains the estimate

$$\|p - p_h\|^2_{L^2(\Omega)} \le C(\|r - r_I\|_{H^1(\Omega)} + \|\boldsymbol{w} - \boldsymbol{w}_I\|_{\boldsymbol{H}(\operatorname{div},\Omega)})(\|p - p_h\|_{H^1(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{H}(\operatorname{div},\Omega)})$$

for all $\boldsymbol{w}_I \in \boldsymbol{S}_h$ and $r_I \in W_h$. This estimate will furnish an improved order of convergence for p in L^2 as compared to H^1 if $\boldsymbol{S}_h$ has good approximation properties in $\boldsymbol{H}(\operatorname{div}, \Omega)$. For the $\boldsymbol{\mathcal{ABF}}_r$ space (still with W_h based on $\mathcal{Q}_{r+1}$) we obtain

$$\|p - p_h\|_{L^2(\Omega)} \le C h^{r+2} (\|p\|_{H^{r+1}(\Omega)} + \|\boldsymbol{u}\|_{H^{r+r}(\Omega)} + \|\operatorname{div} \boldsymbol{u}\|_{H^{r+1}(\Omega)}).$$

But for the $\boldsymbol{\mathcal{RT}}_0$ space we obtain no convergence whatsoever. Numerical computations reported in the next section verify these findings for both the scalar and the vector variable: with W_h taken to be the usual four node H^1 elements based on $\mathcal{Q}_1$ and $\boldsymbol{S}_h$ based on $\boldsymbol{\mathcal{ABF}}_0$, we obtain convergence of order 1 for $\boldsymbol{u}$ in $\boldsymbol{H}(\mathrm{div},\Omega)$ and of order 2 for p in $L^2(\Omega)$, but if we use $\boldsymbol{\mathcal{RT}}_0$ elements instead there is no L^2 convergence for $\boldsymbol{u}$ or p.

The numerical computations of the next section also exhibit second order convergence for $\operatorname{div}\boldsymbol{u}$ in $\boldsymbol{L}^2(\Omega)$ when approximated by the $\boldsymbol{\mathcal{ABF}}_0$ method on square meshes. We close this section by showing that

$$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)} = O(h^{r+2})$$

when the $\boldsymbol{\mathcal{ABF}}_r$ elements are used on rectangular meshes. Now define $\boldsymbol{w}\in\boldsymbol{H}(\mathrm{div},\Omega)$ and $r\in H_0^1(\Omega)$ by

$$\int_\Omega (\boldsymbol{w}-\nabla r)\cdot\boldsymbol{v}\,d\boldsymbol{x} + \int_\Omega \operatorname{div}\boldsymbol{w}\operatorname{div}\boldsymbol{v}\,d\boldsymbol{x} = \int_\Omega \operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\operatorname{div}\boldsymbol{v}\,d\boldsymbol{x} \quad \forall \boldsymbol{v}\in\boldsymbol{H}(\mathrm{div},\Omega),$$
$$\int_\Omega (\boldsymbol{w}-\nabla r)\cdot\nabla q\,d\boldsymbol{x} = 0 \quad \forall q\in H_0^1(\Omega).$$

Then $\Delta r = \operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)$ and $\boldsymbol{w}=\nabla r$, and so $\|r\|_{H^2(\Omega)} + \|\boldsymbol{w}\|_{H^1(\Omega)} \le C\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$. Taking $\boldsymbol{v}=\boldsymbol{u}-\boldsymbol{u}_h$, $q=p-p_h$ and using the error equations, we obtain

$$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|^2_{L^2(\Omega)} = \int_\Omega (\boldsymbol{w}-\boldsymbol{w}_I-\nabla[r-r_I])\cdot(\boldsymbol{u}-\boldsymbol{u}_h-\nabla[p-p_h])\,d\boldsymbol{x} + \int_\Omega \operatorname{div}(\boldsymbol{w}-\boldsymbol{w}_I)\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\,d\boldsymbol{x}$$

for any $\boldsymbol{w}_I\in\boldsymbol{S}_h$ and $r_I\in W_h$. Taking $\boldsymbol{w}_I=\boldsymbol{\pi}_h\boldsymbol{w}$ and r_I a standard interpolant of r, the first integral on the right-hand side is bounded by

$$\begin{aligned} &Ch\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}(\|p-p_h\|_{H^1(\Omega)} + \|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{H}(\mathrm{div},\Omega)}) \\ &\quad\le Ch^{r+2}\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}(\|p\|_{H^{r+1}(\Omega)} + \|\boldsymbol{u}\|_{H^{r+1}(\Omega)} + \|\operatorname{div}\boldsymbol{u}\|_{H^{r+1}(\Omega)}). \end{aligned}$$

To bound the second integral, we note that, in the rectangular case, $\operatorname{div}\boldsymbol{\pi}_h\boldsymbol{w} = \Pi_h\operatorname{div}\boldsymbol{w}$ with Π_h the L^2-projection into $\operatorname{div}\boldsymbol{V}_h$, and also, in the rectangular case, $\operatorname{div}\boldsymbol{V}_h$ contains all piecewise polynomials of degree at most $r+1$, so $\|q-\Pi_h q\|_{L^2(\Omega)} \le Ch^{r+2}\|q\|_{H^{r+2}(\Omega)}$ for all q. Therefore,

$$\begin{aligned} \int_\Omega \operatorname{div}(\boldsymbol{w}-\boldsymbol{w}_I)\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\,d\boldsymbol{x} &= \int_\Omega \operatorname{div}\boldsymbol{w}[\operatorname{div}\boldsymbol{u}-\Pi_h(\operatorname{div}\boldsymbol{u})]\,d\boldsymbol{x} \\ &\le C\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}h^{r+2}\|\operatorname{div}\boldsymbol{u}\|_{H^{r+2}(\Omega)}. \end{aligned}$$

Combining these estimates, we conclude that

$$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)} \le Ch^{r+2}(\|p\|_{H^{r+1}(\Omega)} + \|\boldsymbol{u}\|_{H^{r+1}(\Omega)} + \|\operatorname{div}\boldsymbol{u}\|_{H^{r+2}(\Omega)}).$$

8. Numerical results. In this section, we illustrate our results with several numerical examples using two sequences of meshes. The first is a uniform mesh of the unit square into n^2 subsquares and the second is a mesh of trapezoids as shown in Figure 1(b) (with the notation of Theorem 3.1, here $\alpha=1$ and $\beta=1/3$). In the first of

TABLE 1

Errors and orders of convergence for the piecewise $\boldsymbol{H}(\mathrm{div},\Omega)$ projection into discontinuous $\boldsymbol{\mathcal{BDM}}_1$ and discontinuous $\boldsymbol{\mathcal{BDFM}}_2$.

Piecewise $\boldsymbol{H}(\mathrm{div},\Omega)$ projection into $\boldsymbol{\mathcal{BDM}}_1$ on square meshes

	$\|\boldsymbol{u}-\pi_h\boldsymbol{u}\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\pi_h\boldsymbol{u})\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order
2	1.94e−02	13.010		2.11e−01	30.151	
4	5.08e−03	3.405	1.9	1.15e−01	16.428	0.9
8	1.28e−03	0.861	2.0	5.86e−02	8.375	1.0
16	3.22e−04	0.216	2.0	2.94e−02	4.207	1.0
32	8.05e−05	0.054	2.0	1.47e−01	2.106	1.0
64	2.01e−05	0.013	2.0	7.36e−03	1.053	1.0

Piecewise $\boldsymbol{H}(\mathrm{div},\Omega)$ projection into $\boldsymbol{\mathcal{BDM}}_1$ on trapezoidal meshes

	$\|\boldsymbol{u}-\pi_h\boldsymbol{u}\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\pi_h\boldsymbol{u})\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order
2	2.57e−02	17.243		2.63e−01	37.646	
4	7.89e−03	5.291	1.7	1.83e−01	26.109	0.5
8	2.80e−03	1.879	1.5	1.50e−01	21.430	0.3
16	1.21e−03	0.811	1.2	1.40e−01	20.031	0.1
32	5.78e−04	0.387	1.1	1.37e−01	19.662	0.0
64	2.85e−04	0.191	1.0	1.37e−01	19.568	0.0

Piecewise $\boldsymbol{H}(\mathrm{div},\Omega)$ projection into $\boldsymbol{\mathcal{BDFM}}_2$ on square meshes

	$\|\boldsymbol{u}-\pi_h\boldsymbol{u}\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\pi_h\boldsymbol{u})\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order
2	1.52e−02	10.206		5.27e−02	7.538	
4	3.80e−03	2.552	2.0	1.32e−02	1.884	2.0
8	9.51e−04	0.638	2.0	3.29e−03	0.471	2.0
16	2.38e−04	0.159	2.0	8.24e−04	0.118	2.0
32	5.94e−05	0.040	2.0	2.06e−04	0.029	2.0
64	1.49e−05	0.010	2.0	5.15e−05	0.007	2.0

Piecewise $\boldsymbol{H}(\mathrm{div},\Omega)$ projection into $\boldsymbol{\mathcal{BDFM}}_2$ on trapezoidal meshes

	$\|\boldsymbol{u}-\pi_h\boldsymbol{u}\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\pi_h\boldsymbol{u})\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order
2	1.86e−02	12.502		6.85e−02	9.791	
4	5.07e−03	3.399	1.9	3.52e−02	5.040	1.0
8	1.38e−03	0.926	1.9	1.77e−02	2.538	1.0
16	4.29e−04	0.288	1.7	8.89e−03	1.271	1.0
32	1.66e−04	0.111	1.4	4.45e−03	0.636	1.0
64	7.56e−05	0.051	1.1	2.22e−03	0.318	1.0

these examples (see Table 1), we demonstrate the decreased orders of convergence of the $\boldsymbol{\mathcal{BDM}}_1$ and $\boldsymbol{\mathcal{BDFM}}_2$ spaces by computing the piecewise $\boldsymbol{H}(\mathrm{div},\Omega)$ projection of a simple smooth function, $\boldsymbol{u}=\operatorname{grad}[x_1(1-x_1)x_2(1-x_2)]$, into the discontinuous versions of these spaces. On a rectangular mesh, the space $\boldsymbol{\mathcal{BDFM}}_2$ gives second order approximation of both components of the vector and of its divergence. This is

TABLE 2
Errors and orders of convergence for the mixed approximation to Poisson's equation.

$\boldsymbol{\mathcal{RT}}_0$ *on square meshes*

	$\|p-p_h\|_{L^2(\Omega)}$			$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order	err.	%	order
2	1.84e−02	55.28		6.09e−02	40.83		2.11e−01	30.15	
4	1.04e−02	31.07	0.8	3.32e−02	22.24	0.9	1.15e−01	16.43	0.9
8	5.33e−03	15.99	1.0	1.69e−02	11.34	1.0	5.86e−02	8.38	1.0
16	2.68e−03	8.05	1.0	8.49e−03	5.70	1.0	2.94e−02	4.21	1.0
32	1.34e−03	4.03	1.0	4.25e−03	2.85	1.0	1.47e−02	2.11	1.0

$\boldsymbol{\mathcal{RT}}_0$ *on trapezoidal meshes*

	$\|p-p_h\|_{L^2(\Omega)}$			$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order	err.	%	order
2	1.84e−02	55.08		6.34e−02	42.55		2.67e−01	38.14	
4	1.08e−02	32.37	0.8	3.63e−02	24.38	0.8	1.85e−01	26.51	0.5
8	5.60e−03	16.80	0.9	1.91e−02	12.83	0.9	1.53e−01	21.82	0.3
16	2.83e−03	8.48	1.0	9.81e−03	6.58	1.0	1.43e−01	20.42	0.1
32	1.42e−03	4.25	1.0	4.97e−03	3.33	1.0	1.40e−01	20.05	0.0

$\boldsymbol{\mathcal{ABF}}_0$ *on square meshes*

	$\|p-p_h\|_{L^2(\Omega)}$			$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order	err.	%	order
2	2.49e−02	74.59		6.89e−02	64.21		5.27e−02	7.54	
4	1.36e−02	40.65	0.9	3.42e−02	22.97	1.0	1.32e−02	1.88	2.0
8	7.03e−03	21.08	1.0	1.70e−02	11.43	1.0	3.29e−03	0.47	2.0
16	3.70e−03	11.10	0.9	8.51e−03	5.71	1.0	8.24e−04	0.12	2.0
32	1.93e−03	5.78	0.9	4.25e−03	2.85	1.0	2.06e−04	0.03	2.0

$\boldsymbol{\mathcal{ABF}}_0$ *on trapezoidal meshes*

	$\|p-p_h\|_{L^2(\Omega)}$			$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order	err.	%	order
2	2.31e−02	69.38		6.59e−02	44.20		6.91e−02	9.89	
4	1.33e−02	39.98	0.8	3.58e−02	24.04	0.9	3.58e−02	5.12	0.9
8	7.22e−03	21.66	0.9	1.85e−02	12.41	1.0	1.81e−02	2.58	1.0
16	3.84e−03	11.51	0.9	9.43e−03	6.33	1.0	9.05e−03	1.30	1.0
32	2.00e−03	5.99	0.9	4.77e−03	3.20	1.0	4.53e−03	0.65	1.0

confirmed in the approximation of the piecewise $\boldsymbol{H}(\mathrm{div},\Omega)$ projection. On a trapezoidal mesh, $\boldsymbol{\mathcal{BDFM}}_2$ gives only first order approximation of both components of the vector and of its divergence, and this is also confirmed in the approximation of the piecewise $\boldsymbol{H}(\mathrm{div},\Omega)$ projection. On a rectangular mesh, the space $\boldsymbol{\mathcal{BDM}}_1$ gives second order approximation of both components of the vector, but only first order approximation of its divergence. On a trapezoidal mesh these orders of convergence are reduced to first order for the approximation of both components of the vector and the approximation of the divergence shows no convergence. These theoretical convergence orders are also confirmed in the computations. Although we do not in-

TABLE 3
Errors and orders of convergence for the least squares approximation to Poisson's equation.

$\boldsymbol{\mathcal{RT}}_0$ *on square meshes*

	$\|p-p_h\|_{L^2(\Omega)}$			$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order	err.	%	order
2	2.61e−01	52.28		1.07e+00	48.03		5.78e+00	58.58	
4	7.71e−02	15.42	1.8	5.15−01	23.19	1.1	3.09e+00	31.34	0.9
8	2.01e−02	4.01	1.9	2.53e−01	11.41	1.0	1.57e+00	15.94	1.0
16	5.07e−03	1.01	2.0	1.26e−01	5.68	1.0	7.90e−01	8.00	1.0
32	1.27e−03	0.25	2.0	6.30e−02	2.84	1.0	3.95e−01	4.01	1.0
64	3.18e−04	0.06	2.0	3.15e−02	1.42	1.0	1.98e−01	2.00	1.0

$\boldsymbol{\mathcal{RT}}_0$ *on trapezoidal meshes*

	$\|p-p_h\|_{L^2(\Omega)}$			$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order	err.	%	order
2	2.95e−01	58.96		1.24e+00	55.74		6.03e+00	61.07	
4	1.08e−01	21.67	1.4	6.05−01	27.26	1.0	3.68e+00	37.25	0.7
8	4.29e−02	8.58	1.3	3.10e−01	13.97	1.0	2.50e+00	25.37	0.6
16	2.51e−02	5.01	0.8	1.72e−01	7.74	0.9	2.09e+00	21.16	0.3
32	2.06e−02	4.12	0.3	1.13e−01	5.09	0.6	1.97e+00	19.96	0.1
64	1.95e−02	3.89	0.1	9.27e−02	4.17	0.3	1.94e+00	19.64	0.0

$\boldsymbol{\mathcal{ABF}}_0$ *on square meshes*

	$\|p-p_h\|_{L^2(\Omega)}$			$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order	err.	%	order
2	1.42e−01	28.46		1.04e+00	46.77		2.19e+00	22.18	
4	3.35e−02	6.70	2.1	5.10e−01	22.98	1.0	5.88e−01	9.96	1.9
8	8.22e−03	1.64	2.0	2.53e−01	11.38	1.0	1.50e−01	1.52	2.0
16	2.04e−03	0.41	2.0	1.26e−01	5.67	1.0	3.76e−02	0.38	2.0
32	5.10e−04	0.10	2.0	6.30e−02	2.84	1.0	8.41e−03	0.10	2.0

$\boldsymbol{\mathcal{ABF}}_0$ *on trapezoidal meshes*

	$\|p-p_h\|_{L^2(\Omega)}$			$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}$			$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_{L^2(\Omega)}$		
n	err.	%	order	err.	%	order	err.	%	order
2	1.89e−01	37.74		1.17e+00	52.86		3.0e+00	31.39	
4	5.49e−02	10.98	1.8	5.61e−01	25.24	1.1	1.12e+00	11.32	1.5
8	1.45e−02	2.89	1.9	2.80e−01	12.62	1.0	5.00e−01	5.07	1.2
16	3.67e−03	0.73	1.9	1.40e−01	6.32	1.0	2.42e−01	2.45	1.0
32	9.20e−04	0.18	2.0	7.02e−02	3.16	1.0	1.20e−01	1.21	1.0

clude the details of the computations, the same convergence orders are observed in computations of the $\boldsymbol{L}^2(\Omega)$, rather than the piecewise $\boldsymbol{H}(\operatorname{div},\Omega)$ projection.

The second computation, reported in Table 2, illustrates our results on the convergence orders of $\boldsymbol{\mathcal{RT}}_0$ and $\boldsymbol{\mathcal{ABF}}_0$ for the approximation of Poisson's equation by the standard mixed finite element method. The exact solution is $p = x_1(1-x_1)x_2(1-x_2)$. As expected, on a trapezoidal mesh, $\boldsymbol{\mathcal{RT}}_0$ gives a first order approximation to the scalar and vector variable (the same as on a rectangular mesh), but there is no con-

vergence of the approximation of the divergence of the vector variable in contrast to the standard first order approximation seen on rectangles. When $\boldsymbol{\mathcal{ABF}}_0$ is used instead, there is an improvement in the convergence order of the divergence of the vector variable.

Finally, Table 3 shows the difference in the convergence orders of $\boldsymbol{\mathcal{RT}}_0$ and $\boldsymbol{\mathcal{ABF}}_0$ coupled with $\mathcal{Q}_1$ for the scalar variable for the approximation of Poisson's equation by a standard least squares finite element method. Again the exact solution is $p = x_1(1-x_1)x_2(1-x_2)$. When $\boldsymbol{\mathcal{RT}}_0$ is used, the poor approximation of the divergence on trapezoidal meshes results in poor approximation of both the scalar and vector variable, while on a rectangle the scalar variable is approximated to second order and the vector variable and its divergence to first order. When $\boldsymbol{\mathcal{ABF}}_0$ is used instead, one achieves second order convergence for the scalar variable and first order convergence for the vector variable on both rectangular and quadrilateral meshes. The divergence of the vector variable is approximated to second order on rectangles and to first order on trapezoids, as predicted by the theory.

REFERENCES

[1] D. N. Arnold, D. Boffi, and R. S. Falk, *Approximation by quadrilateral finite elements*, Math. Comp., 71 (2002), pp. 909–922.

[2] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[3] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 129–151.

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

[5] J. Douglas Jr. and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.

[6] R. S. Falk and J. E. Osborn, *Error estimates for mixed methods*, RAIRO Anal. Numér., 14 (1980), pp. 249–277.

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

[8] Y. Kuznetsov and S. Repin, *New mixed finite element method on polygonal and polyhedral meshes*, Russian J. Numer. Anal. Math. Modelling, 18 (2003), pp. 261–278.

[9] R. L. Naff, T. F. Russel, and J. D. Wilson, *Shape functions for velocity interpolation in general hexahedral cells*, Comput. Geosci., 6 (2002), pp. 285–314.

[10] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[11] J. Shen, *Mixed Finite Element Methods on Distorted Rectangular Grids*, Technical report, Institute for Scientific Computation, Texas A&M University, College Station, TX, 1994.

[12] J. M. Thomas, *Sur l'analyse numerique des methodes d'elements finis hybrides et mixtes*, Ph.D. Thesis, Université Pierre et Marie Curie, Paris, France, 1977.

[13] Z. Zhang, *Analysis of some quadrilateral nonconforming elements for incompressible elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 640–663.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2452–2475

LEAST-SQUARES GALERKIN METHODS FOR PARABOLIC PROBLEMS III: SEMIDISCRETE CASE FOR SEMILINEAR PROBLEMS*

MOHAMMAD MAJIDI†

Abstract. This is the third part of a series of papers on least-squares Galerkin methods for parabolic initial-boundary value problems. These methods are based on the minimization of a least-squares functional for an equivalent first-order system over space and time with respect to suitable discrete spaces. This paper presents the derivation and analysis of one-step methods for semidiscretization in time from least-squares principles for semilinear parabolic problems. One of the most important features of the least-squares methodology is a built-in a posteriori estimate for the approximation error. For the presentation in this paper, we focus our attention on the specific combination of piecewise linear, not necessarily continuous, functions in time with continuous piecewise linear for the flux and scalar variables, respectively. For the resulting method, a convergence result is shown for the scalar variable.

Key words. least-squares Galerkin method, first-order system, a posteriori error estimate, adaptive time-step control, semilinear parabolic problems

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/S0036142902409185

1. Introduction.

Efficient computation of accurate approximations to initial-boundary value problems for parabolic partial differential equations requires adaptive techniques. Both the step-size for the discretization in time as well as the choice of mesh for the spatial discretization need to be adapted to local features of the solution using appropriate error estimators. One of the strengths of the least-squares Galerkin approach presented in this series of papers is that it provides such an a posteriori error estimator automatically by evaluating the associated least-squares functional. The purpose of this third paper is to prove convergence of the least-squares method for the semilinear case with respect to the time discretization. For this semilinear case we will observe an order reduction. Adaptive strategies for time-step control, based on a posteriori error estimators, were derived for the backward Euler discretization in [13] and in the framework of discontinuous Galerkin methods in a series of papers [6, 7, 8, 9, 10] for linear and nonlinear parabolic problems. See also [25, section 12] for the linear case. A different approach to adaptive time-stepping based on extrapolation principles was established for the linear case in [2, 3, 4] and for the nonlinear case in [15]. In [26], four least-squares Galerkin approaches are used to approximate a numerical solution of a first-order convection-diffusion system in $H^1(\Omega) \times H(\mathrm{div}, \Omega)$.

The least-squares Galerkin methodology presented here is conceptually different from all these previous approaches to adaptivity for parabolic problems. Our approach is based on the reformulation of the parabolic problem as a first-order system by introducing the flux u as an additional variable. A suitable least-squares functional is then minimized with respect to discrete spaces in order to construct approximations for the scalar variable p and for the flux u simultaneously. In [18] the equivalence

*Received by the editors June 11, 2002; accepted for publication (in revised form) April 7, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/sinum/42-6/40918.html
†Institute for Applied Mathematics, University of Hannover, D-30177 Hannover, Germany (majidi@ifam.uni-hannover.de).

of the least-squares functional to the consistency error for the linear case is proved, which implies that it provides an a posteriori error estimator to be used for the adaptive control of the time steps. In the semilinear case, the equivalence of the least-squares functional to the consistency error is an open problem. In this case we have only numerical evidence (cf. [17]), which implies that it provides an a posteriori error estimator to be used for the adaptive control of the time steps. In the fully discrete case, which will be considered in the fourth part of this series of papers, the least-squares functional provides an a posteriori error estimator for the approximation error in time and space. The use of least-squares functionals associated with first-order systems for a posteriori error estimation for elliptic boundary value problems is studied in [1] and, in particular for nonlinear problems, in [23]. At first sight, it may seem that the extension to more variables by changing to the first-order system formulation leads to an increase in computational work which would make the least-squares approach inefficient. However, a fair comparison of different methods is not so simple since the overall goal is to gain a certain required accuracy at minimal cost. For example, the discontinuous Galerkin method using piecewise linear functions has similar approximation properties in time for p (second order in the semilinear case) to the method discussed in this paper. The discontinuous Galerkin method requires two spatial approximations for p in this case, while the least-squares method described below needs one for p and two for u. A detailed comparison of the amount of work involved in the computation would also need to account for the availability of fast solvers for the elliptic subproblems. However, such a detailed study is far beyond the scope of this paper. In many practical applications, the flux u is actually of interest in itself and the approximation property of the least-squares Galerkin method for this variable then becomes important.

In the rest of this paper, we proceed as follows. First we introduce in section 2 the spaces that will be used in this work. To define the derivative for L^2-functions we need to introduce a tool called *distributions*. Section 3 presents the least-squares Galerkin framework and derives the associated variational formulations. In section 4, the representation of the discrete evolution in terms of solution of a nonlinear equation is derived which is the basis for the convergence analysis. The convergence analysis of the approximation error in time with respect to the H^1-norm is carried out in section 5. Finally, section 6 addresses the issues related to the solution of the elliptic problems at each time step.

2. Distributions and other spaces. Distributions give a useful tool to define and handle with Sobolev space and derivation in the weak form. Next, we will use the same notation as that of [22]. In this section, we give some definition of standard spaces using distributions. For $d \in \mathbb{N}$, let $\Omega \subset \mathbb{R}^d$ be a bounded domain with smooth boundary and let us define, for short,

$$\mathcal{D}(\Omega) = C_0^\infty(\Omega).$$

In $\mathcal{D}(\Omega)$ we observe the following concept of convergence: for a sequence $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{D}(\Omega)$ and $\varphi \in \mathcal{D}(\Omega)$ we say the sequence φ_n *converges* in $\mathcal{D}(\Omega)$ to φ or $\varphi_n \xrightarrow{\mathcal{D}} \varphi$ iff there is a compact $K \subset \Omega$, where $\mathrm{supp}\varphi_n \subset K$ for all $n \in \mathbb{N}$, and for all α,

$$\partial^\alpha \varphi_n \longrightarrow \partial^\alpha \varphi \text{ uniformly in } \Omega.$$

Now we are in the position to define $\mathcal{D}'(\Omega)$ as the space of linear and continuous functionals on $\mathcal{D}(\Omega)$. The functionals contained in $\mathcal{D}'(\Omega)$ are called *distributions*. We

call $F : \mathcal{D}(\Omega) \to \mathbb{R}$ *continuous* iff for all $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{D}(\Omega)$ and $\varphi \in \mathcal{D}(\Omega)$ with $\varphi_n \xrightarrow{\mathcal{D}} \varphi$ it follows that $F(\varphi_n) \xrightarrow{\mathbb{R}} F(\varphi)$. Now we observe the following concept of convergence on $\mathcal{D}'(\Omega)$: For $(F_n)_{n\in\mathbb{N}} \subset \mathcal{D}'(\Omega)$ and $F \in \mathcal{D}'(\Omega)$, we say $F_n \xrightarrow{\mathcal{D}'} F$ iff $F_n(\varphi) \xrightarrow{\mathbb{R}} F(\varphi)$ for all $\varphi \in \mathcal{D}(\Omega)$. Obviously each $f \in L^p(\Omega) \cup C^m(\Omega), 1 \le p \le \infty$, $m \in \mathbb{N}_0 \cup \{\infty\}$, could generate the following distribution:

$$F_f : \mathcal{D}(\Omega) \to \mathbb{R},$$

where

$$F_f(\varphi) := \langle F_f, \varphi\rangle := \int_\Omega f(x)\varphi(x)\,dx.$$

For such a function we identify f with its generated distribution F_f. In the above equation, F_f operates on the left-hand side like a functional and on the middle and right-hand sides like a function in the inner product $\langle\cdot,\cdot\rangle$. We will prefer the inner product notation $\langle\cdot,\cdot\rangle$, but we will use both notations whenever the situation could be better understood.

For $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$, we set $|\alpha| := \sum_{i=1}^d \alpha_i$ and for $F \in \mathcal{D}'(\Omega)$, we define the (*weak*) *derivation* as the following distribution:

$$\partial^\alpha F : \mathcal{D}(\Omega) \to \mathbb{R},$$
$$\langle \partial^\alpha F, \varphi\rangle = (-1)^{|\alpha|} \langle F, \partial^\alpha \varphi\rangle .$$

Notice that this "new definition" is an expansion of the definition of derivation in the classical sense. For $F \in \mathcal{D}'(\Omega)$ we say $\partial^\alpha F \in L^p(\Omega)$ iff there is $g_\alpha \in L^p(\Omega)$, where

$$\partial^\alpha F(\varphi) = \int_\Omega g_\alpha(x)\varphi(x)\,dx \quad \forall \varphi \in \mathcal{D}(\Omega).$$

2.1. Sobolev spaces. Now we are able to define Sobolev spaces: For $k \in \mathbb{N}_0$ let

$$H^k(\Omega) = \left\{ v \in L^2(\Omega) \;\middle|\; \partial^\alpha v \in L^2(\Omega) \quad \forall \alpha \in \mathbb{N}_0^d,\ |\alpha| \le k \right\}$$

be the Sobolev space of order k.

2.2. H(div)-space. Now we define for a distribution $F \in (\mathcal{D}'(\Omega))^d$ the *distributive divergence* $\mathrm{div}F \in \mathcal{D}'(\Omega)$,

$$\mathrm{div}F(\varphi) = -F(\nabla\varphi) \quad \forall \varphi \in \mathcal{D}(\Omega),$$

that will be used in the next sections. We can now introduce the $H(\mathrm{div})$-space as

$$H(\mathrm{div}, \Omega) = \left\{ v \in (L^2(\Omega))^d \;\middle|\; \mathrm{div} v \in L^2(\Omega) \right\}.$$

3. Least-squares Galerkin formulation in time. In this section we introduce the parabolic semilinear problem. Furthermore we formulate the Galerkin approximation in time. For detailed information, the reader is referred to [18].

Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain and assume the boundary of Ω is divided in Γ_D and Γ_N. Furthermore let n be the exterior unit normal on Γ_N. Assume $a,\ c \in L^\infty(\Omega)$ and that $a(x) \ge \underline{a},\ c(x) \ge \underline{c}$ for all $x \in \Omega$ with positive constants $\underline{a}, \underline{c}$.

For $\hat{f} : \Omega \times \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$ and $\Lambda : \Omega \to \mathbb{R}$, we consider the first-order system formulation of the semilinear parabolic equation,

$$\begin{aligned} c(x)\,\partial_t p(t,x) + \operatorname{div} u(t,x) + \hat{f}(x, p(t,x), \nabla p(t,x)) &= 0 \quad \forall (t,x) \in (0,T)\times\Omega, \\ u(t,x) + a(x)\,\nabla p(t,x) &= 0 \quad \forall (t,x) \in (0,T)\times\Omega, \\ u(t,x)\cdot n(x) &= 0 \quad \forall (t,x) \in (0,T)\times\Gamma_N, \\ p(t,x) &= 0 \quad \forall (t,x) \in (0,T)\times\Gamma_D, \\ p(0,x) &= \Lambda(x) \quad \forall x \in \Omega, \end{aligned} \tag{3.1}$$

with some $T > 0$. We are interested in a scalar function $p : [0,T]\times\Omega \to \mathbb{R}$ and a vector function $u : [0,T]\times\Omega \to \mathbb{R}^2$ that solve (3.1). More general boundary conditions can be handled in the standard way by a suitable modification of the right-hand side $\hat{f}$. Obviously, the system in (3.1) does not always have a solution. In the following observation we restrict $\hat{f}$ and Λ such that the existence of a unique solution can be guaranteed. Let us define the following spaces:

$$\begin{aligned} H_{\Gamma_N}(\operatorname{div},\Omega) &= \{ v \in H(\operatorname{div},\Omega) \mid n\cdot v = 0 \text{ on } \Gamma_N \}, \\ H^1_{\Gamma_D}(\Omega) &= \{ q \in H^1(\Omega) \mid q = 0 \text{ on } \Gamma_D \}, \end{aligned}$$

where $H^1(\Omega)$ and $H(\operatorname{div},\Omega)$ were defined in the last section. Furthermore we define for a Hilbert space H the following spaces:

$$\begin{aligned} L^2((0,T);H) &= \left\{ v : (0,T) \to H \;\middle|\; \int_0^T \|v\|_H^2 < \infty \right\}, \\ H^1((0,T);H) &= \left\{ v : (0,T) \to H \;\middle|\; \partial_t v \in L^2((0,T);H) \right\}. \end{aligned}$$

For each $q \in H^1_{\Gamma_D}(\Omega)$, we define

$$\begin{aligned} &\mathcal{I}_q : \Omega \to \mathbb{R}, \\ &\mathcal{I}_q(x) = \hat{f}(x, q(x), \nabla q(x)), \end{aligned}$$

for short, and set

$$f(q) := \mathcal{I}_q.$$

Assuming $f \in C^1(H^1_{\Gamma_D}(\Omega); L^2(\Omega))$ and $\Lambda \in H^1_{\Gamma_D}(\Omega)$, we now replace (3.1) by

$$\begin{aligned} c\partial_t p + \operatorname{div} u + f(p) &= 0 \quad \text{in } (0,T)\times\Omega, \\ u + a\,\nabla p &= 0 \quad \text{in } (0,T)\times\Omega, \\ u\cdot n &= 0 \quad \text{on } (0,T)\times\Gamma_N, \\ p &= 0 \quad \text{on } (0,T)\times\Gamma_D, \\ p(0) &= \Lambda \quad \text{in } \Omega, \end{aligned} \tag{3.2}$$

which has a unique solution $(u,p) \in L^2((0,T); H_{\Gamma_N}(\operatorname{div},\Omega)) \times H^1((0,T); H^1_{\Gamma_D}(\Omega))$ depending continuously on the given initial and boundary data (see, e.g., [24, section 15.1]). For symmetry, we replace the first and second equations of (3.2) with the equivalent first-order system

$$\begin{aligned} c^{1/2}\partial_t p + c^{-1/2}\operatorname{div} u + c^{-1/2} f(p) &= 0, \\ a^{-1/2} u + a^{1/2}\,\nabla p &= 0. \end{aligned} \tag{3.3}$$

Obviously, the solution $(u,p) \in L^2\left((0,T); H_{\Gamma_N}(\mathrm{div},\Omega)\right) \times H^1\left((0,T); H^1_{\Gamma_D}(\Omega)\right)$ of (3.3) minimizes the least-squares functional

$$\tilde{\mathcal{F}}(u,p) = \int_0^T \left(\alpha \left\| c^{1/2}\partial_t p + c^{-1/2}(\mathrm{div} u + f(p)) \right\|^2_{0,\Omega} + \left\| a^{-1/2}\, u + a^{1/2}\nabla p \right\|^2_{0,\Omega} \right) ds\,,$$

where $\alpha > 0$ has to be chosen later. The choice of the weight α was in the first part of this series of papers the main decision to get a functional that could be used as an a posteriori error estimator (see [18, 19]). On the other hand, if $\tilde{\mathcal{F}}(u,p) = 0$, then (u,p) is a solution of (3.3).

For the time discretization, we choose subspaces

$$V_\tau((0,T); H_{\Gamma_N}(\mathrm{div},\Omega)) \subset L^2((0,T); H_{\Gamma_N}(\mathrm{div},\Omega))\,, \tag{3.4}$$

$$Q_\tau((0,T); H^1_{\Gamma_D}(\Omega)) \subset H^1((0,T); H^1_{\Gamma_D}(\Omega)) \tag{3.5}$$

and perform the minimization of the least-squares functional in these spaces. The subindex τ emphasizes that these spaces are semidiscrete spaces. Appropriate choices include the combination of piecewise polynomial, not necessarily continuous, functions for V_τ with piecewise polynomial continuous functions for Q_τ on a subdivision $\triangle = \{0 = t_0 < t_1 < \cdots < t_M = T\}$ of $[0,T]$ with step-sizes $\tau_j = t_j - t_{j-1}, j = 1,\ldots,M$. However, due to the continuity requirement for Q_τ, minimizing the least-squares functional with respect to these spaces generally leads to a coupling of all $M+1$ time-levels. Now we set $\alpha = \tau$ and consider one-step methods that consist of stepping only from one time-level t to the next $t+\tau$ by minimizing the least-squares functional

$$\begin{aligned} \hat{\mathcal{F}}(u_\tau, p_\tau) = \int_t^{t+\tau} \Big(& \tau \Big\| c^{1/2}\partial_t(p_\tau(s)) + c^{-1/2}\Big[\mathrm{div}(u_\tau(s)) + f(p_\tau(s))\Big] \Big\|^2_{0,\Omega} \\ & + \Big\| a^{-1/2} u_\tau(s) + a^{1/2}\nabla(p_\tau(s)) \Big\|^2_{0,\Omega} \Big) ds. \end{aligned} \tag{3.6}$$

Throughout the rest of this paper, the symbols $\lesssim$ and $\gtrsim$ indicate that the inequality holds up to constants that are independent of the discretization parameter τ. Similarly, the symbol $\eqsim$ stands for equivalence independently of discretization parameters.

4. Variational formulation for the nonlinear least-squares functional. In this section, the representation of the discrete evolution in terms of solution of a nonlinear equation is derived which is the basis for the convergence analysis. Our aim is to minimize the least-squares functional (3.6) in $\tilde{V}_\tau \times \tilde{Q}_\tau$, where

$$\begin{aligned} \tilde{V}_\tau &= \left\{ \left(1 - \frac{\sigma}{\tau}\right) v_1 + \frac{\sigma}{\tau} v_2 \;\middle|\; v_1, v_2 \in H_{\Gamma_N}(\mathrm{div},\Omega) \right\}, \\ \tilde{Q}_\tau &= \left\{ \frac{\sigma}{\tau}\, q \;\middle|\; q \in H^1_{\Gamma_D}(\Omega) \right\}. \end{aligned}$$

For short, let us set

$$p_\tau(t+\sigma) = \left(1 - \frac{\sigma}{\tau}\right) p_\tau(t) + \hat{p}_\tau(\sigma)$$

and

$$q_\tau(\sigma) = \left(1 - \frac{\sigma}{\tau}\right) p_\tau(t) + \hat{q}_\tau(\sigma),$$

where $\hat{p}_\tau, \hat{q}_\tau \in \tilde{Q}_\tau$.

Our aim is to find $(u_\tau, \hat{p}_\tau) \in \tilde{V}_\tau \times \tilde{Q}_\tau$ with

$$\hat{\mathcal{F}}\,(u_\tau, p_\tau) = \min_{(v_\tau, \hat{q}_\tau) \in \tilde{V}_\tau \times \tilde{Q}_\tau} \hat{\mathcal{F}}\,(v_\tau, q_\tau)\,. \tag{4.1}$$

Roughly speaking, our one-step method assumes a given $p_\tau(t)$ (for $t = 0$, we start with $p_\tau(0) = p(0) = \Lambda$) and computes in each step a $p_\tau(t+\tau) = (1-\frac{\tau}{\tau})p_\tau(t)+\hat{p}_\tau(\tau) = \hat{p}_\tau(\tau)$, $\hat{p}_\tau \in \tilde{Q}_\tau$, for the next step. Actually, this assumption leads to a coupling of all $M+1$ time-levels with respect to the scalar variable p_τ. With the choice of the above spaces, we can simplify the minimization problem (4.1) to analyze the observed method. We are interested in the exact evolution operator. Let us define the operator A as

$$\begin{aligned} &A : D_A \subset H^1_{\Gamma_D}(\Omega) \to L^2(\Omega), \\ &\quad (c^{1/2}\, A\, c^{1/2} p, q)_{0,\Omega} = (a\, \nabla p, \nabla q)_{0,\Omega} \qquad \forall q \in H^1_{\Gamma_D}(\Omega)\,. \end{aligned} \tag{4.2}$$

Formally, we may write $A = -c^{-1/2}\mathrm{div}\,(a\,\nabla)\,c^{-1/2}$. Note that the operator A is bijective as a mapping from

$$D_A = \{ q \in H^1_{\Gamma_D}(\Omega)\,|\, a\,\nabla q \in H_{\Gamma_N}(\mathrm{div}, \Omega)\,\}$$

to $L^2(\Omega)$. Now we can define the evolution operator. We deduce from [24, Chapter 15.1] that

$$\begin{aligned} &E(t+\tau;t) : H^1_{\Gamma_D}(\Omega) \to H^1_{\Gamma_D}(\Omega), \\ &E(t+\tau;t)\, p(t) = p(t+\tau) = c^{-1/2} \exp(-\tau A)\, c^{1/2} p(t) \\ &\qquad\qquad - c^{-1/2} \int_0^\tau (\exp(-(\tau-\sigma)A)) c^{-1/2} f(p(t+\sigma))\, d\sigma\,. \end{aligned} \tag{4.3}$$

Next, we define the discrete evolution operator

$$\begin{aligned} E_d(t+\tau;t) :& H^1_{\Gamma_D}(\Omega) \to H^1_{\Gamma_D}(\Omega), \\ & E_d(t+\tau;t) p_\tau(t) = p_\tau(t+\tau), \end{aligned} \tag{4.4}$$

where $p_\tau(t)$ is the initial value of our one-step method and $p_\tau(t+\tau)$ its solution using the minimization method (4.1). We will see later that a unique solution exists if τ is small enough. The following theorem gives $p_\tau(t+\tau) = E_d(t+\tau;t)p_\tau(t)$ as a solution of a nonlinear equation depending on $p_\tau(t)$.

THEOREM 4.1. *The solution $E_d(t+\tau;t)p_\tau(t)$ of the minimization method* (4.1) *fulfills the following nonlinear equation (in the sense of distributions):*

$$\begin{aligned} \int_0^\tau c^{1/2} \Big[I + \sigma \Big(c^{-1/2} f'(p_\tau(t+\sigma)) c^{-1/2} + A \Big) \Big] \Big[& c^{1/2} \frac{p_\tau(t+\tau) - p_\tau(t)}{\tau} \\ & + A c^{1/2} p_\tau(t+\sigma) + c^{-1/2} f(p_\tau(t+\sigma)) \Big]\, d\sigma = 0, \end{aligned}$$

where f' is the derivation of f.

Proof. We will separate the minimization problem (4.1) in two different variational problems. From the first variational problem we gain a formula for $c^{-1/2}\mathrm{div}u$. Setting that formula in the second variational problem and simplifying leads finally to our result.

The minimization problem (4.1) is equivalent to the variational problem: Find $(u_\tau, p_\tau) \in \tilde{V}_\tau \times \tilde{Q}_\tau$, such that

$$\begin{aligned}\int_0^\tau \Big[\tau\Big(&c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}\mathrm{div}u_\tau(t+\sigma) + c^{-1/2}f(p_\tau(t+\sigma)),\\ &c^{1/2}\partial_t q(\sigma) + c^{-1/2}\,\mathrm{div}v(\sigma) + c^{-1/2}f'(p_\tau(t+\sigma))q(\sigma)\Big)_{0,\Omega}\\ &+\Big(a^{-1/2}u_\tau(t+\sigma) + a^{1/2}\nabla p_\tau(t+\sigma), a^{-1/2}v(\sigma) + a^{1/2}\nabla q(\sigma)\Big)_{0,\Omega}\Big]d\sigma = 0\end{aligned} \tag{4.5}$$

for all $(v, q) \in \tilde{V}_\tau \times \tilde{Q}_\tau$. This formula includes two different variational formulations in spaces $\tilde{V}_\tau$ and $\tilde{Q}_\tau$, which we will separate. We deduce the first equation from (4.5) and (3.2) as

$$\begin{aligned}\int_0^\tau \Big[\tau\Big(&c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}\mathrm{div}u_\tau(t+\sigma) + c^{-1/2}f(p_\tau(t+\sigma)), c^{-1/2}\mathrm{div}v(\sigma)\Big)_{0,\Omega}\\ &+\Big(a^{-1/2}u_\tau(t+\sigma) + a^{1/2}\nabla p_\tau(t+\sigma), a^{-1/2}v(\sigma)\Big)_{0,\Omega}\Big]d\sigma = 0\end{aligned}$$

for all $v \in \tilde{V}_\tau$. If we define

$$D_{A^2} = \{q \in D_A \mid Aq \in D_A\},$$

$$\hat{D}_\tau = \Big\{\frac{\sigma}{\tau}\, q_1 + \Big(1 - \frac{\sigma}{\tau}\Big) q_2 \;\mid q_1, q_2 \in D_{A^2}\Big\}$$

and set $v = -a\nabla q$, where $q \in \hat{D}_\tau$, then we get $c^{-1/2}\mathrm{div}v = Ac^{1/2}q$. Hence, we obtain

$$\begin{aligned}\int_0^\tau \Big[\tau\Big(&c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}[\mathrm{div}u_\tau(t+\sigma) + f(p_\tau(t+\sigma))], Ac^{1/2}q(\sigma)\Big)_{0,\Omega}\\ &-\Big(a^{-1/2}u_\tau(t+\sigma) + a^{1/2}\nabla p_\tau(t+\sigma), a^{1/2}\nabla q(\sigma)\Big)_{0,\Omega}\Big]d\sigma = 0\end{aligned} \tag{4.6}$$

for all $q \in \hat{D}_\tau$. Again using the definition of A, we write

$$\begin{aligned}\int_0^\tau \Big[\Big(&c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}f(p_\tau(t+\sigma)) + \frac{1}{\tau}\, c^{1/2}p_\tau(t+\sigma), \tau Ac^{1/2}q(\sigma)\Big)_{0,\Omega}\\ &+\Big(c^{-1/2}\mathrm{div}u_\tau(t+\sigma), (I+\tau A)c^{1/2}q(\sigma)\Big)_{0,\Omega}\Big]d\sigma = 0\end{aligned}$$

for all $q \in \hat{D}_\tau$. Furthermore, we use the fact that the operators $(I + \tau A)$ and τA are bijective and self-adjoint to make the equivalent reformulation

$$\begin{aligned}\int_0^\tau \Big[\Big(&(I+\tau A)^{-1}\Big[c^{1/2}\partial_t p_\tau(t+\sigma) + c^{-1/2}f(p_\tau(t+\sigma))\\ &+\frac{1}{\tau}\, c^{1/2}p_\tau(t+\sigma)\Big], (I+\tau A)(\tau A)c^{1/2}q(\sigma)\Big)_{0,\Omega}\\ &+\Big((\tau A)^{-1}c^{-1/2}\mathrm{div}u_\tau(t+\sigma), (\tau A)(I+\tau A)c^{1/2}q(\sigma)\Big)_{0,\Omega}\Big]d\sigma = 0\end{aligned}$$

to obtain

$$\int_0^\tau \Big[\Big((I+\tau A)^{-1}\Big[c^{1/2}\partial_t p_\tau(t+\sigma)+c^{-1/2}f(p_\tau(t+\sigma))+\frac{1}{\tau}\,c^{1/2}p(t+\sigma)\Big] + (\tau A)^{-1}c^{-1/2}\mathrm{div}u_\tau(t+\sigma),(I+\tau A)\tau A c^{1/2}q(\sigma)\Big)_{0,\Omega}\Big]d\sigma=0 \tag{4.7}$$

for all $q\in\hat{D}_\tau$. Defining

$$\hat{L}_\tau=\Big\{\frac{\sigma}{\tau}\,q_1+\Big(1-\frac{\sigma}{\tau}\Big)\,q_2 \;\Big|\; q_1,q_2\in L^2(\Omega)\Big\},$$

we observe that $(I+\tau A)\tau Ac^{1/2}:\hat{D}_\tau\to\hat{L}_\tau$ is one-to-one. Hence, (4.7) leads to

$$c^{-1/2}\mathrm{div}u_\tau=-(I+\tau A)^{-1}\Big(\tau A[c^{1/2}\partial_t p_\tau+c^{-1/2}f(p_\tau)]+Ac^{1/2}p_\tau\Big) \tag{4.8}$$

in $\hat{L}_\tau$.

Now, we obtain the second equation from (4.5):

$$\int_0^\tau\Big[\tau\Big(c^{1/2}\partial_t p_\tau(t+\sigma)+c^{-1/2}[\mathrm{div}u_\tau(t+\sigma)+f(p_\tau(t+\sigma))],c^{-1/2}\partial_t q(\sigma) + f'(p_\tau(t+\sigma))\,c^{1/2}q(\sigma)\Big)_{0,\Omega}\Big]d\sigma + \underbrace{\int_0^\tau\Big[\Big(a^{-1/2}u_\tau(t+\sigma)+a^{1/2}\nabla p_\tau(t+\sigma),a^{1/2}\nabla q(\sigma)\Big)_{0,\Omega}\Big]d\sigma}_{=:K}=0 \tag{4.9}$$

for all $q\in\tilde{Q}_\tau$. Restricting (4.9) to $q\in\tilde{Q}_\tau\cap\hat{D}_\tau$ and using (4.6) lead to

$$K=\int_0^\tau\Big[\tau\Big(c^{1/2}\partial_t p_\tau(t+\sigma)+c^{-1/2}[\mathrm{div}u_\tau(t+\sigma)+f(p_\tau(t+\sigma))],Ac^{1/2}q(\sigma)\Big)_{0,\Omega}\Big]d\sigma$$

for all $q\in\tilde{Q}_\tau\cap\hat{D}_\tau$. Hence, we obtain

$$\int_0^\tau\Big[\tau\Big(c^{1/2}\partial_t p_\tau(t+\sigma)+c^{-1/2}[\mathrm{div}u_\tau(t+\sigma)+f(p_\tau(t+\sigma))],c^{1/2}\partial_t q(\sigma) + c^{-1/2}f'(p_\tau(t+\sigma))q(\sigma)+Ac^{1/2}q(\sigma)\Big)_{0,\Omega}\Big]d\sigma=0$$

for all $q\in\tilde{Q}_\tau\cap\hat{D}_\tau$. If we replace $c^{-1/2}\mathrm{div}u$ by the right-hand side of (4.8) and note that $1-\frac{z}{1+z}=\frac{1}{1+z}$, we deduce

$$\int_0^\tau\Big[\tau\Big((I+\tau A)^{-1}\Big[c^{1/2}\partial_t p_\tau(t+\sigma)+Ac^{1/2}p_\tau(t+\sigma)+c^{-1/2}f(p_\tau(t+\sigma))\Big],c^{1/2}\partial_t q(\sigma) + c^{-1/2}f'(p_\tau(t+\sigma))q(\sigma)+A\,c^{1/2}q(\sigma)\Big)_{0,\Omega}\Big]d\sigma=0$$

for all $q\in\tilde{Q}_\tau\cap\hat{D}_\tau$. Since $q\in\tilde{Q}_\tau\cap\hat{D}_\tau$ (i.e., $q(\sigma)=\frac{\sigma}{\tau}\hat{q},\hat{q}\in D_{A^2}$), we have

$$\int_0^\tau\Big[\tau\Big((I+\tau A)^{-1}\Big[c^{1/2}\partial_t p_\tau(t+\sigma)+Ac^{1/2}p_\tau(t+\sigma)+c^{-1/2}f(p_\tau(t+\sigma))\Big],\frac{1}{\tau}\,c^{1/2}\hat{q} + c^{-1/2}f'(p_\tau(t+\sigma))c^{-1/2}\frac{\sigma}{\tau}\,c^{1/2}\hat{q}+A\,\frac{\sigma}{\tau}\,c^{1/2}\hat{q}\Big)_{0,\Omega}\Big]d\sigma=0$$

for all $\hat{q} \in D_{A^2}$. Equivalence reformulation leads to

$$\left\langle \int_0^\tau c^{1/2}\Big[I+\sigma\Big(c^{-1/2}f'(p_\tau(t+\sigma))c^{-1/2}+A\Big)\Big]\Big[c^{1/2}\partial_t p_\tau(t+\sigma) + Ac^{1/2}p_\tau(t+\sigma)+c^{-1/2}f(p_\tau(t+\sigma))\Big]d\sigma, (I+\tau A)^{-1}\hat{q}\Big)\right\rangle = 0$$

for all $\hat{q} \in D_{A^2}$. Since $C_0^\infty(\Omega) \subset (I+\tau A)^{-1}D_{A^2} = D_{A^3}$, we obtain

$$\left\langle \int_0^\tau c^{1/2}\Big[I+\sigma\Big(c^{-1/2}f'(p_\tau(t+\sigma))c^{-1/2}+A\Big)\Big]\Big[c^{1/2}\partial_t p_\tau(t+\sigma) + Ac^{1/2}p_\tau(t+\sigma)+c^{-1/2}f(p_\tau(t+\sigma))\Big]d\sigma, \varphi\Big\rangle\right\rangle = 0$$

for all $\varphi \in C_0^\infty(\Omega)$. □

To construct a numerical method, we should replace the integral

$$\int_0^\tau c^{1/2}\Big[I+\sigma\Big(c^{-1/2}f'(p_\tau(t+\sigma))c^{-1/2}+A\Big)\Big]\Big[c^{1/2}\partial_t p_\tau(t+\sigma) + Ac^{1/2}p_\tau(t+\sigma)+c^{-1/2}f(p_\tau(t+\sigma))\Big]d\sigma$$

by computable terms. For simplicity, we assume $a \equiv c \equiv 1$. We could get the general case by modifying this special case. We deduce from Theorem 4.1 that

$$\int_0^\tau \big[I+\sigma(A+f'(p_\tau(t+\sigma)))\big]\big[\partial_t p_\tau(t+\sigma)+Ap_\tau(t+\sigma)+f(p_\tau(t+\sigma))\big]\,d\sigma = 0.$$

Computing the integral, we obtain

$$\begin{aligned} p_\tau(t+\tau) = {} & r_{2,2}(\tau A)p_\tau(t) - r_{0,2}(\tau A)\Bigg[\int_0^\tau (I+\sigma A)f(p_\tau(t+\sigma))\,d\sigma \\ & - \int_0^\tau \sigma f'(p_\tau(t+\sigma))(\partial_t p_\tau(t+\sigma)+Ap_\tau(t+\sigma)+f(p_\tau(t+\sigma)))\,d\sigma\Bigg], \end{aligned} \tag{4.10}$$

where

$$r_{2,2}(z) = \frac{1-\frac{z^2}{6}}{1+z+\frac{z^2}{3}}$$

and

$$r_{0,2}(z) = \frac{1}{1+z+\frac{z^2}{3}}.$$

The above operators are defined in the sense of the standard definition by eigenvalues of A. For further information the reader is referred to [18, 19]. If we now set $f(p) \equiv$ const, we observe the same result as in [18, Lemma 3.1]. Hence, Theorem 4.1 is a generalization of [18, Lemma 3.1].

5. Convergence theory. In this section, the convergence analysis of the approximation error in time with respect to the H^1-norm is observed, which is the main result of this paper. Define for a mesh $\triangle = \{0 = t_0^\triangle < t_1^\triangle < \cdots < t_{M_\triangle}^\triangle = T\}$ with step-sizes $\tau_j^\triangle = t_j^\triangle - t_{j-1}^\triangle$, $j = 1, \ldots, M_\triangle$:

$$\tau_{\min,\triangle} := \min_{j=1}^{M_\triangle} \tau_j^\triangle$$

and

$$\tau_\triangle := \max_{j=1}^{M_\triangle} \tau_j^\triangle.$$

We use $\triangle$ as an index for the mesh points t_j, step-sizes τ_j, and the number of mesh points M of an observed mesh $\triangle$ whenever we would like to emphasize that the choice of these parameters depends on the mesh $\triangle$.

Furthermore, we assume that

$$\frac{\tau_\triangle}{\tau_{\min,\triangle}} \eqsim 1 \tag{5.1}$$

for each observed $\triangle \subset [0,T]$. The following theorem shows the second-order convergence of the minimizing method (4.1).

Let us define for $\tilde{\alpha} \geq 0$

$$H_{\tilde{\alpha}} := D_{A^{\tilde{\alpha}/2}} \subset L^2(\Omega).$$

$H_{\tilde{\alpha}}$ is a Hilbert space with respect to the inner product

$$\langle q_1, q_2 \rangle_{\tilde{\alpha}} := (A^{\frac{\tilde{\alpha}}{2}} q_1, A^{\frac{\tilde{\alpha}}{2}} q_2)_{0,\Omega}$$

(cf. [21, p. 195]). Let $\|\cdot\|_{\tilde{\alpha}} := (\langle \cdot, \cdot \rangle_{\tilde{\alpha}})^{\frac{1}{2}}$ be the associated norm. Furthermore, put

$$H_{-\tilde{\alpha}} := (H_{\tilde{\alpha}})' = \left\{ \tilde{q} \in \mathcal{D}' \;\middle|\; \sup_{q \in H_{\tilde{\alpha}}} \frac{\langle \tilde{q}, q \rangle}{\|q\|_{\tilde{\alpha}}} < \infty \right\}$$

as the dual space of $H_{\tilde{\alpha}}$. Our assumption $f \in C^1(H^1_{\Gamma_D}(\Omega); L^2(\Omega))$ leads to the fact that f is local Lipschitz. Next, consider the exact solution $p \in H^1((0,T); H^1_{\Gamma_D}(\Omega))$ of (3.1). Let L_t be the Lipschitz constant of f in a neighborhood $B(p(t), \delta_t) \subset H^1_{\Gamma_D}(\Omega)$ (i.e., the open ball with center $p(t)$ and radius δ_t in $H^1_{\Gamma_D}(\Omega)$) of the exact solution $p(t)$ of (3.1). Furthermore set $L = \sup_{t \in [0,T]} L_t < \infty$ and $\delta = \inf_{t \in [0,T]} \delta_t > 0$. To prepare a short notation for the next theorem, we define

$$g(\tau, \sigma, q) = r_{0,2}(\tau A)\, \sigma f'(q)\, [Aq \;\; + f(q)] \quad \forall q \in H^1_{\Gamma_D}(\Omega). \tag{5.2}$$

Generally, with our current assumptions, $g(\tau, \sigma, \cdot)$ is not well-posed. If we assume that $f : H_{-1} \to H_{-2}$ and $f(H_1) \subset H_0 = L^2(\Omega)$, we receive a well-posed $g(\tau, \sigma, \cdot)$: Starting with

$$q \in H_1 \underbrace{=}_{\text{cf. [14, Theorem 2.23]}} H^1_{\Gamma_D}(\Omega),$$

we get

$$\underbrace{Aq}_{\in H_{-1}} + \underbrace{f(q)}_{\in H_0} \in H_{-1}.$$

For each $\hat{q} \in H_{-1}$ the linear operator $f'(\hat{q})$ maps H_{-1} into H_{-2}, since we assume $f : H_{-1} \to H_{-2}$, and hence $\sigma f'(q)\,[Aq \ + f(q)] \in H_{-2}$. Finally we have $r_{0,2}(\tau A) : H_{-2} \to H_2 \subset H_1 = H^1_{\Gamma_D}(\Omega)$, which leads to $g(\tau,\sigma,q) = r_{0,2}(\tau A)\,\sigma f'(q)\,[Aq \ + f(q)] \in H^1_{\Gamma_D}(\Omega)$. Furthermore, we assume that the function $g(\tau,\sigma,\cdot)$ is local Lipschitz for all $\sigma,\tau \in [0,T],\ \sigma \le \tau$.

Let $\tilde{L}_{t,\tau,\sigma}$ be the Lipschitz constant of $g(\tau,\sigma,\cdot)$ in a neighborhood $B(p(t),\tilde{\delta}_{t,\tau,\sigma}) \subset H^1_{\Gamma_D}(\Omega)$ of the solution $p(t)$ of (3.2). Defining

$$\tilde{L} = \max_{t,\tau,\sigma\in[0,T]}\{\tilde{L}_{t,\tau,\sigma}\} < \infty,$$

$$\tilde{\delta} = \min_{t,\tau,\sigma\in[0,T]}\{\tilde{\delta}_{t,\tau,\sigma}\} > 0,$$

$$\hat{\delta} = \min\{\tilde{\delta}, \delta\},$$

and

$$\hat{L} = \max\{\tilde{L}, L\},$$

we deduce

$$\|f(q) - f(p(t))\|_{0,\Omega} \le \hat{L}\|q - p(t)\|_{1,\Omega}$$

and

$$\|g(\tau,\sigma,q) - g(\tau,\sigma,p(t))\|_{1,\Omega} \le \hat{L}\|q - p(t)\|_{1,\Omega}$$

for all $q \in H^1_{\Gamma_D}(\Omega)$ with $\|q - p(t)\|_{1,\Omega} < \hat{\delta}$.

Before we formulate the theorem, we recall the well-known Gronwall lemma.

LEMMA 5.1 (Gronwall's lemma (for weakly singular kernels)). *Assume $\hat{a}, \hat{b}, \hat{\alpha}, \hat{\beta}$ are nonnegative constants, with $\hat{\alpha}, \hat{\beta} < 1$, and $0 < T < \infty$. There is a constant $C = C(\hat{a}, \hat{b}, \hat{\alpha}, \hat{\beta}) < \infty$ such that for any integrable function $v : [0,T] \to \mathbb{R}$ satisfying*

$$0 \le v(t) \le \hat{a}t^{-\hat{\alpha}} + \hat{b}\int_0^t (t-s)^{-\hat{\beta}} v(s)ds$$

for t a.e. in $[0,T]$, we have

$$v(t) \le \hat{a}Ct^{-\hat{\alpha}}$$

a.e. on $0 < t \le T$.

Proof. This is a special case of [12, Lemma 7.1.1]. The proof is an elementary iteration argument, followed by Lebesgue's dominated convergence theorem. ∎

Finally, we present the main convergence theorem of this paper.

THEOREM 5.2. *Assume that $f : H_{-1} \to H_{-2}$, where its restriction on $H^1_{\Gamma_D}(\Omega)$ is in $C^1(H^1_{\Gamma_D}(\Omega); L^2(\Omega))$, i.e.,*

$$f_{\big|_{H^1_{\Gamma_D}(\Omega)}} \in C^1(H^1_{\Gamma_D}(\Omega); L^2(\Omega)).$$

Furthermore, let $p \in C^2([0,T]; H^1_{\Gamma_D}(\Omega))$ be the exact solution of (3.2) and assume that the function $g(\tau,\sigma,\cdot)$ defined in (5.2) is local Lipschitz for all $\sigma,\tau \in [0,T]$, $\sigma \le \tau$. There are a τ^* and a constant C, such that for all observed meshes $\triangle$ which satisfy (5.1) $\big($remember that $\triangle = \{0 < t_1^\triangle < \cdots < t^\triangle_{M_\triangle - 1} < T\} \subset [0,T]$ with step-sizes $\tau_j^\triangle = t_j^\triangle - t_{j-1}^\triangle$, $j = 1,\ldots,M_\triangle$, and $\tau_\triangle \le \tau^*\big)$, the solution of our minimizing problem (4.1) $p_\tau(t_n^\triangle)$ satisfies

$$\|p_\tau(t_n^\triangle) - p(t_n^\triangle)\|_{1,\Omega} \le C\,\tau_\triangle^2 \tag{5.3}$$

for all $0 \le n \le M_\triangle$.

Proof. We will separate the proof in three parts: In parts (a) and (c) we show

$$\max_{t\in[0,T]} \|q_\tau(t) - p(t)\|_{1,\Omega} < \hat\delta$$

for sufficiently small $\tau_\triangle$. We need this result to use the Lipschitz property of f and $g(\tau,\sigma,\cdot)$. Due to this Lipschitz property we show in part (b) an upper estimation

$$\|p_\tau(t_n^\triangle) - p(t_n^\triangle)\|_{1,\Omega} \le \text{const.}\ \tau_\triangle^2 + \text{const.}\ \tau_\triangle \sum_{j=1}^{n-1} ((n-j+1)\tau_\triangle)^{-\frac12} \|p_\tau(t_j^\triangle) - p(t_j^\triangle)\|_{1,\Omega}$$

and use Gronwall's lemma (Lemma 5.1) to get the main result.

(a) In this part of the proof we will show that there is a $q_\tau \in Q_\tau$ such that

$$\max_{t\in[0,T]} \|q_\tau(t) - p(t)\|_{1,\Omega} < \hat\delta. \tag{5.4}$$

Remember that Q_τ defined in (3.5) is a subset of $H^1((0,T); H^1_{\Gamma_D}(\Omega))$ with piecewise linear continuous functions on the observed subdivision $\triangle \subset [0,T]$. For a given mesh $\{0 < t_1^\triangle < \cdots < t^\triangle_{M_\triangle - 1} < T\} = \triangle \subset [0,T]$, we define $I_\triangle p$ to be the linear spline of p. It is known from the interpolation theory (cf. [5, Theorem 7.3]) that $\max_{t\in[0,T]} \|I_\triangle p(t) - p(t)\|_{1,\Omega} \le C_1 \tau_\triangle^2$, where C_1 is a constant which only depends on $\sup_{t\in[0,T]} \|\partial_t^2 p(t)\|_{1,\Omega} < \infty$. Setting $\tau^* \le \min\{1, \frac{\hat\delta}{2C_1}\} =: \varepsilon_0$, we have

$$\max_{t\in[0,T]} \|I_\triangle p(t) - p(t)\|_{1,\Omega} \le C_1\varepsilon_0^2 \le C_1\varepsilon_0 \le \frac{\hat\delta C_1}{2C_1} \le \frac{\hat\delta}{2}.$$

Let $q_j \in H^1_{\Gamma_D}(\Omega)$ such that $\|q_j - p(t_j^\triangle)\|_{1,\Omega} < \frac{\hat\delta}{2}$ for $j = 1,\ldots,M_\triangle$ and $q_0 = p_0$. If we define

$$q_\tau(t) = \frac{t - t_{j-1}^\triangle}{\tau_j^\triangle}\, q_j + \left(1 - \frac{t - t_{j-1}^\triangle}{\tau_j^\triangle}\right) q_{j-1}$$

for given $t \in [t_{j-1}^\triangle, t_j^\triangle]$, $1 \le j \le M_\triangle$, then we get

$$\max_{t\in[0,T]} \|q_\tau(t) - p(t)\|_{1,\Omega} \le \max_{t\in[0,T]} \|I_\triangle p(t) - p(t)\|_{1,\Omega} + \max_{t\in[0,T]} \|q_\tau(t) - I_\triangle p(t)\|_{1,\Omega} < \delta.$$

If we assume that τ^* is chosen, such that

$$\|p_\tau(t_j^\triangle) - p(t_j^\triangle)\|_{1,\Omega} \le j\frac{\hat{\delta}}{3M_\triangle} \tag{5.5}$$

for $j = 0, \dots, M_\triangle$, then we get, together with the above considerations,

$$\max_{t\in[0,T]} \|p_\tau(t) - p(t)\|_{1,\Omega} < \hat{\delta} \tag{5.6}$$

for all observed $\triangle$. We will prove (5.5) later in part (c). We should note that the result in (5.6) allows us to use the local Lipschitz property of f and g.

(b) In this part, our goal is to prove the main result using the so-called defect equation.

(i) First we reformulate the defect equation in a form where estimation with powers of $\tau_\triangle$ becomes obvious. Set $\hat{p}_n = p(t_n^\triangle)$, $n = 0, \dots, M_\triangle$, for short. For $0 < n \le M_\triangle$, we get

$$\hat{p}_n = E_d(t_n^\triangle; t_{n-1}^\triangle)\hat{p}_{n-1} \underbrace{-E_d(t_n^\triangle; t_{n-1}^\triangle)\hat{p}_{n-1} + \hat{p}_n}_{=:d_n}, \tag{5.7}$$

where d_n is the so-called *defect* and the discrete evolution operator $E_d(t_n^\triangle; t_{n-1}^\triangle)$ was defined in (4.4). Hence, using Theorem 4.1 and (4.10) to reformulate the defect, we obtain

$$\begin{aligned} d_n = &- r_{2,2}(\tau_n^\triangle A)\hat{p}_{n-1} + \hat{p}_n \\ &- r_{0,2}(\tau_n^\triangle A)\bigg[\int_0^{\tau_n^\triangle} (I + \sigma A)(\partial_t p(t_{n-1}^\triangle + \sigma) + Ap(t_{n-1}^\triangle + \sigma))\, d\sigma \\ &- \int_0^{\tau_n^\triangle} \sigma f'(p(t_n^\triangle + \sigma)) \underbrace{(\partial_t p(t_n^\triangle + \sigma) + Ap(t_n^\triangle + \sigma) + f(p(t_n^\triangle + \sigma)))}_{=0}\, d\sigma\bigg]. \end{aligned} \tag{5.8}$$

Integration by parts leads to

$$\begin{aligned} & r_{0,2}(\tau_n^\triangle A)\int_0^{\tau_n^\triangle} \underbrace{(I + \sigma A)}_{=:\mathcal{U}(\sigma)} \underbrace{\partial_t p(t_n^\triangle + \sigma)}_{=:\mathcal{V}'(\sigma)}\, d\sigma \\ &\quad = r_{0,2}(\tau_n^\triangle A)\left[\mathcal{U}(\sigma)\mathcal{V}(\sigma)\Big|_0^{\tau_n^\triangle} - \int_0^{\tau_n^\triangle} \mathcal{U}'(\sigma)\mathcal{V}(\sigma)d\sigma\right] \\ &\quad = r_{0,2}(\tau_n^\triangle A)\left[(I + \tau_n^\triangle A)\hat{p}_n - \hat{p}_{n-1} - \int_0^{\tau_n^\triangle} Ap(t_{n-1}^\triangle + \sigma)\, d\sigma\right]. \end{aligned} \tag{5.9}$$

Using (5.9) in (5.8) and simplifying, we get

$$\begin{aligned} d_n = &(r_{0,2}(\tau_n^\triangle A) - r_{2,2}(\tau_n^\triangle A))\hat{p}_{n-1} + (I - (I + \tau_n^\triangle A)\, r_{0,2}(\tau_n^\triangle A))\hat{p}_n \\ &- r_{0,2}(\tau_n^\triangle A)\int_0^{\tau_n^\triangle} \sigma A^2 p(t_{n-1}^\triangle + \sigma)\, d\sigma. \end{aligned} \tag{5.10}$$

If we notice that

(5.11)
$$r_{0,2}(\tau_n^{\triangle}A)\int_0^{\tau_n^{\triangle}}\sigma A^2 I_{\triangle}p(t_{n-1}^{\triangle}+\sigma)\,d\sigma = r_{0,2}(\tau_n^{\triangle}A)\left(\frac{{\tau_n^{\triangle}}^2}{3}A^2\hat{p}_n + \frac{{\tau_n^{\triangle}}^2}{6}A^2\hat{p}_{n-1}\right),$$

we can expand (5.10) with $I_{\triangle}p$ as follows:

$$d_n = -\int_0^{\tau_n^{\triangle}}\sigma A^2 r_{0,2}(\tau_n^{\triangle}A)(p(t_{n-1}^{\triangle}+\sigma) - I_{\triangle}p(t_{n-1}^{\triangle}+\sigma))\,d\sigma. \tag{5.12}$$

Again, using [5, Theorem 7.3], for $\sigma\in[0,\tau_n^{\triangle}]$, we have

(5.13)
$$p(t_{n-1}^{\triangle}+\sigma) - I_{\triangle}p(t_{n-1}^{\triangle}+\sigma) = \frac{1}{2}\,\partial_t^2 p(t_{n-1}^{\triangle}+\theta_\sigma)\sigma(\sigma-\tau_n^{\triangle}), \quad \text{where } \theta_\sigma\in(0,\tau_n^{\triangle})\ .$$

(ii) Our next goal is to get an estimation like

$$\|e_n\|_{1,\Omega} \le \text{const.}\ \tau_{\triangle}^2 + \text{const.}\ \tau_{\triangle}\sum_{j=1}^{n-1}((n-j+1)\tau_{\triangle})^{-\frac{1}{2}}\|e_j\|_{1,\Omega},$$

and use Gronwall's lemma (Lemma 5.1) to get the main result.

First we use the chain rule to write $\partial_t f(q(t)) = f'(q(t))\ \partial_t q(t)$ for all $q\in H^1((0,T)\,;H^1_{\Gamma_D}(\Omega))$. Hence, integration by parts leads to

$$\begin{aligned}
&\int_0^{\tau_n^{\triangle}} \underbrace{\sigma}_{=:\mathcal{U}(\sigma)} \underbrace{[f'(p_\tau(t_{n-1}^{\triangle}+\sigma))\ \partial_t p_\tau(t_{n-1}^{\triangle}+\sigma) - f'(p(t_{n-1}^{\triangle}+\sigma))\ \partial_t p(t_{n-1}^{\triangle}+\sigma)]}_{=:\mathcal{V}'(\sigma)}\,d\sigma \\
&\quad = \mathcal{U}(\sigma)\mathcal{V}(\sigma)\Big|_0^{\tau_n^{\triangle}} - \int_0^{\tau_n^{\triangle}}\mathcal{U}'(\sigma)\mathcal{V}(\sigma)d\sigma \\
&\quad = \tau_n^{\triangle}\,(f(p_n)-f(\hat{p}_n)) - \int_0^{\tau_n^{\triangle}} f(p_\tau(t_{n-1}^{\triangle}+\sigma)) - f(p(t_{n-1}^{\triangle}+\sigma))\,d\sigma.
\end{aligned} \tag{5.14}$$

Now define

$$R(t_j^{\triangle},t_n^{\triangle}) = \prod_{i=j+1}^{n} r_{2,2}(\tau_i^{\triangle}A),$$

$$\Sigma_{j+1}(\sigma) = f(p_\tau(t_j^{\triangle}+\sigma)) - f(I_{\triangle}p(t_j^{\triangle}+\sigma)),$$

and

$$\hat{\Sigma}_{j+1}(\sigma) = f(I_{\triangle}p(t_j^{\triangle}+\sigma)) - f(p(t_j^{\triangle}+\sigma)).$$

Let $e_j = p_\tau(t_j^{\triangle}) - p(t_j^{\triangle})$ be the error for $j=0,\dots,M_{\triangle}$. Hence, we deduce the

following from (5.14), (4.10), the defect equation (5.7), and the definition of g in (5.2):

$$\begin{aligned} e_n = {} & r_{2,2}(\tau_n^{\triangle} A)e_{n-1} - \tau_n^{\triangle} r_{0,2}(\tau_n^{\triangle} A)(f(p_n) - f(\hat{p}_n)) \\ & - \int_0^{\tau_n^{\triangle}} r_{0,2}(\tau_n^{\triangle} A)\sigma A \Sigma_n(\sigma)\, d\sigma - \int_0^{\tau_n^{\triangle}} r_{0,2}(\tau_n^{\triangle} A)\sigma A \hat{\Sigma}_n(\sigma)\, d\sigma \\ & - \int_0^{\tau_n^{\triangle}} \underbrace{g(\tau_n^{\triangle}, \sigma, p_\tau(t_{n-1}^{\triangle} + \sigma)) - g(\tau_n^{\triangle}, \sigma, p(t_{n-1}^{\triangle} + \sigma))}_{=:\mathcal{G}_n(\sigma)} d\sigma + d_n. \end{aligned}$$

This way we get recursively (noting that $e_0 = 0$)

$$\begin{aligned} e_n = {} & \sum_{j=0}^{n-1} R(t_{j+1}^{\triangle}, t_n^{\triangle}) \Big[d_{j+1} + \tau_n^{\triangle} r_{0,2}(\tau_n^{\triangle} A)(f(p_n) - f(\hat{p}_n)) \\ & - \int_0^{\tau_{j+1}^{\triangle}} r_{0,2}(\tau_{j+1}^{\triangle} A)(\sigma A)(\Sigma_{j+1}(\sigma) + \hat{\Sigma}_{j+1}(\sigma)) + \mathcal{G}_{j+1}(\sigma) d\sigma \Big]. \end{aligned} \tag{5.15}$$

For short, we set

$$S(t_{j+1}^{\triangle}, t_n^{\triangle}) = (I + \tau_{j+1}^{\triangle} A) r_{0,2}(\tau_{j+1}^{\triangle} A) R(t_{j+1}^{\triangle}, t_n^{\triangle})$$

and notice that

$$\|\sigma A\, r_{0,2}(\tau A) q\|_{1,\Omega} \le \|(I + \sigma A)\, r_{0,2}(\tau A) q\|_{1,\Omega} \le \|(I + \tau A)\, r_{0,2}(\tau A) q\|_{1,\Omega}$$

for all $j = 1, \dots, M_{\triangle}$, $q \in H^1_{\Gamma_D}(\Omega)$, and $\sigma \in [0, \tau]$. Hence, we can deduce from (5.15) and the local Lipschitz property of $g(\tau_j^{\triangle}, \sigma, \cdot)$ that

$$\begin{aligned} \|e_n\|_{1,\Omega} \lesssim {} & \left\| \sum_{j=0}^{n-1} S(t_{j+1}^{\triangle}, t_n^{\triangle}) \left(\int_0^{\tau_{j+1}^{\triangle}} \hat{\Sigma}_{j+1}(\sigma) d\sigma + \tau_{j+1}^{\triangle} \hat{\Sigma}_{j+1}(\tau_{j+1}^{\triangle}) \right) \right\|_{1,\Omega} \\ & + \left\| \sum_{j=0}^{n-1} R(t_{j+1}^{\triangle}, t_n^{\triangle}) d_{j+1} \right\|_{1,\Omega} \\ & + \left\| \sum_{j=0}^{n-1} S(t_{j+1}^{\triangle}, t_n^{\triangle}) \left(\int_0^{\tau_{j+1}^{\triangle}} \Sigma_{j+1}(\sigma)\, d\sigma + \tau_{j+1}^{\triangle} \Sigma_{j+1}(\tau_{j+1}^{\triangle}) \right) \right\|_{1,\Omega} \\ & + \left\| \sum_{j=0}^{n-1} \int_0^{\tau_{j+1}^{\triangle}} p_\tau(t_j^{\triangle} + \sigma) - p(t_j^{\triangle} + \sigma) d\sigma \right\|_{1,\Omega} \\ =: {} & \|I_1\|_{1,\Omega} + \|I_2\|_{1,\Omega} + \|I_3\|_{1,\Omega} + \|I_4\|_{1,\Omega}. \end{aligned}$$

Now, our aim is to estimate each of the terms $\|I_i\|_{1,\Omega}$, $i = 1, 2, 3, 4$. To estimate $\|I_1\|_{1,\Omega}$, let us define

$$\rho_{2,2}(z) = \frac{1 + \frac{5}{6} z + \frac{z^2}{6}}{1 + z + \frac{z^2}{3}},$$

$$\tilde{\rho}_{2,2}(z) = \frac{1 - \frac{z^2}{6}}{1 + \frac{5}{6}z + \frac{z^2}{6}},$$

and

$$\xi(t_j^{\triangle}, t_n^{\triangle}) = \prod_{i=j+1}^{n} \rho_{2,2}(\tau_i^{\triangle} A), \quad j = 0, \ldots, M_{\triangle} - 1.$$

The Poincaré–Friedrichs inequality leads to

$$\|I_1\|_{1,\Omega} \lesssim \|\nabla I_1\|_{0,\Omega} \lesssim \|AI_1\|_{0,\Omega}.$$

The local Lipschitz property of f, along with (5.4) and (5.13), leads to

$$\max_{\sigma \in [0, t_n^{\triangle}]} \|\hat{\Sigma}_{j+1}(\sigma)\|_{0,\Omega} \leq L \max_{\sigma \in [0, t_n^{\triangle}]} \|I_{\triangle} p(t_j^{\triangle} + \sigma) - p(t_j^{\triangle} + \sigma)\|_{1,\Omega} \lesssim \tau_{\triangle}^2. \tag{5.16}$$

For Hilbert spaces $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $(\hat{\mathcal{H}}, \|\cdot\|_{\hat{\mathcal{H}}})$, let $\mathcal{L}(\mathcal{H}, \hat{\mathcal{H}})$ be the space of linear continuous operators $\mathcal{T} : \mathcal{H} \to \hat{\mathcal{H}}$ and let

$$\|\mathcal{T}\|_{\mathcal{H} \to \hat{\mathcal{H}}} = \sup_{0 \neq q \in \mathcal{H}} \frac{\|\mathcal{T} q\|_{\hat{\mathcal{H}}}}{\|q\|_{\mathcal{H}}}$$

be the underlying operator norm on $\mathcal{L}(\mathcal{H}, \hat{\mathcal{H}})$. Observing that $\rho_{2,2}$ is positive and decreases on $[0, \infty)$ and that $\sup_{x \in (0,\infty)} |\tilde{\rho}_{2,2}(x)| = 1$, together with the fact that $r_{2,2} = \rho_{2,2}\tilde{\rho}_{2,2}$, we could easily deduce the following by the formula of the geometric series:

(5.17)

$$\begin{aligned}
\|I_1\|_{1,\Omega} \lesssim & \|AI_1\|_{0,\Omega} \lesssim \tau_{\triangle}^2 \left\| \sum_{j=0}^{n-1} S(t_{j+1}^{\triangle}, t_n^{\triangle}) \tau_{j+1}^{\triangle} A \right\|_{L^2 \to L^2} \\
& \lesssim \tau_{\triangle}^2 \left\| \sum_{j=0}^{n-1} \xi(t_{j+1}^{\triangle}, t_n^{\triangle}) r_{0,2}(\tau_{\min,\triangle}\, A) \tau_{\triangle}\, A \right\|_{L^2 \to L^2} \\
& \lesssim \tau_{\triangle}^2 \left\| \sum_{j=0}^{n-1} \rho_{2,2}(\tau_{\min,\triangle}\, A)^{n-j-1} r_{0,2}(\tau_{\min,\triangle}\, A) \tau_{\triangle}\, A \right\|_{L^2 \to L^2} \\
& \lesssim \tau_{\triangle}^2 \left\| (I - \rho_{2,2}(\tau_{\min,\triangle}\, A))^{-1} r_{0,2}(\tau_{\min,\triangle}\, A) \tau_{\triangle} A \right\|_{L^2 \to L^2} \\
& = \tau_{\triangle}^2 \left\| 6(\tau_{\min,\triangle}\, A + \tau_{\min,\triangle}^2\, A^2)^{-1} r_{0,2}(\tau_{\min,\triangle}\, A)^{-1} r_{0,2}(\tau_{\min,\triangle}\, A) \tau_{\triangle} A \right\|_{L^2 \to L^2} \\
& \lesssim \tau_{\triangle}^2 \left\| (I + \tau_{\min,\triangle}\, A)^{-1} \right\|_{L^2 \to L^2} \lesssim \tau_{\triangle}^2.
\end{aligned}$$

To estimate $\|I_2\|_{1,\Omega}$, we define

$$\varphi_{2,2}(z) = \frac{1 + z + \frac{z^2}{6}}{1 + z + \frac{z^2}{3}}$$

and

$$\tilde{\varphi}_{2,2}(z) = \frac{1 - \frac{z^2}{6}}{1 + z + \frac{z^2}{6}},$$

and we note that $\varphi_{2,2}$ is a positive decreasing function on $[0,\infty)$, that $\sup_{x\in(0,\infty)}|\tilde{\varphi}_{2,2}(x)| = 1$, and that $r_{2,2} = \varphi_{2,2}\tilde{\varphi}_{2,2}$. Again, using the geometric series, (5.12) and (5.13) lead to

$$\begin{aligned}
\|I_2\|_{1,\Omega} &\lesssim \left\| \sum_{j=0}^{n-1} R(t_{j+1}^{\triangle}, t_n^{\triangle})\, r_{0,2}(\tau_{j+1}^{\triangle} A) \int_0^{\tau_{j+1}^{\triangle}} A^2\sigma^2(\sigma - \tau_{j+1}^{\triangle}) d\sigma \right\|_{H^1_{\Gamma_D}\to H^1_{\Gamma_D}} \\
&= \left\| \sum_{j=0}^{n-1} R(t_{j+1}^{\triangle}, t_n^{\triangle})\, r_{0,2}(\tau_{j+1}^{\triangle} A)\, \frac{1}{12}\, {\tau_{j+1}^{\triangle}}^4 A^2 \right\|_{H^1_{\Gamma_D}\to H^1_{\Gamma_D}} \\
&\underbrace{\leq}_{\sup\limits_{x\in(0,\infty)} |\tilde{\varphi}_{2,2}(x)|=1} \left\| \sum_{j=0}^{n-1} \varphi_{2,2}(\tau_{\min,\triangle}\, A)^{n-j-1}\, r_{0,2}(\tau_{\min,\triangle}\, A)\tau_{\triangle}^4 A^2 \right\|_{H^1_{\Gamma_D}\to H^1_{\Gamma_D}} \\
&\lesssim \tau_{\triangle}^2 \left\| (I - r_{2,2}((\tau_{\min,\triangle}\, A))^{-1}\, r_{0,2}(\tau_{\min,\triangle}\, A)\tau_{\triangle}^2 A^2 \right\|_{H^1_{\Gamma_D}\to H^1_{\Gamma_D}} \\
&= \tau_{\triangle}^2 \left\| 6(\tau_{\min,\triangle}\, A)^{-2} r_{0,2}(\tau_{\min,\triangle}\, A)^{-1}\, r_{0,2}(\tau_{\min,\triangle}\, A)\tau_{\triangle}^2 A^2 \right\|_{H^1_{\Gamma_D}\to H^1_{\Gamma_D}} \\
&= \tau_{\triangle}^2\, 6 \left(\frac{\tau_{\triangle}}{\tau_{\min,\triangle}} \right)^2 \lesssim \tau_{\triangle}^2.
\end{aligned} \tag{5.18}$$

To estimate $\|I_3\|_{1,\Omega}$ we proceed as in (5.18) with

$$\|I_3\|_{1,\Omega} \lesssim \left\| \sum_{j=0}^{n-1} Op_j \left(\int_0^{\tau_{j+1}^{\triangle}} \Sigma_{j+1}(\sigma) d\sigma + \tau_{j+1}^{\triangle} \Sigma_{j+1}(\tau_{j+1}^{\triangle}) \right) \right\|_{1,\Omega}, \tag{5.19}$$

where we set

$$Op_j = \varphi_{2,2}(\tau_{\min,\triangle} A)^{n-j-1}(I + \tau_{\triangle} A) r_{0,2}(\tau_{\triangle} A), \quad j = 0, \ldots, n-1,$$

for short. Again, using the local Lipschitz property of f and the inequality (5.4), we deduce

$$\begin{aligned}
\int_0^{\tau_{j+1}^{\triangle}} \|\Sigma_{j+1}(\sigma)\|_{0,\Omega}\, d\sigma &\leq \hat{L} \int_0^{\tau_{j+1}^{\triangle}} \|p_\tau(t_j^{\triangle} + \sigma) - I_{\triangle} p(t_j^{\triangle} + \sigma)\|_{1,\Omega}\, d\sigma \\
&= \hat{L}\, \frac{\tau_{j+1}^{\triangle}}{2} (\|e_j\|_{1,\Omega} + \|e_{j+1}\|_{1,\Omega}).
\end{aligned} \tag{5.20}$$

Combining (5.19) and (5.20) leads to

$$\|I_3\|_{1,\Omega} \lesssim \tau_{\triangle} \sum_{j=0}^{n-1} \|Op_j\|_{L^2\to H^1_{\Gamma_D}} (\|e_{j+1}\|_{1,\Omega} + \|e_j\|_{1,\Omega}). \tag{5.21}$$

For $j = n-1$, we get

$$\|Op_{n-1}\|_{L^2\to H^1_{\Gamma_D}} = \|(I + \tau_{\triangle} A) r_{0,2}(\tau_{\triangle} A)\|_{L^2\to H^1_{\Gamma_D}} \lesssim \tau_{\triangle}^{-1/2}.$$

Assumption (5.1), the inequality $n\tau_{\min,\triangle} \le t_n^{\triangle} \le n\tau_{\triangle}$, and the fact that $\varphi_{2,2}(\infty) = \frac{1}{2}$ lead to

$$\begin{aligned}\|Op_j\|_{L^2\to H^1_{\Gamma_D}} &\lesssim \|\varphi_{2,2}(\tau_{\min,\triangle}A)^{n-j-1} - 2^{j+1-n}\|_{L^2\to H^1_{\Gamma_D}} \\ &\quad + \|2^{j+1-n}(I+\tau_{\triangle}A)r_{0,2}(\tau_{\triangle}A)\|_{L^2\to H^1_{\Gamma_D}} \\ &\lesssim {t_{n-j+1}^{\triangle}}^{-1/2} + \| \underbrace{2^{j+1-n}}_{\lesssim (n-j+1)^{-1/2}} (I+\tau_{\triangle}A)r_{0,2}(\tau_{\triangle}A)\|_{L^2\to H^1_{\Gamma_D}} \lesssim {t_{n-j+1}^{\triangle}}^{-1/2}\end{aligned}$$

for $j = 0, \dots, n-2$ (cf. [11, Theorem 1.1]). For the last term $\|I_4\|_{1,\Omega}$, we immediately see the following:

$$\|I_4\|_{1,\Omega} \lesssim \tau_{\triangle}^2 + \tau_{\triangle} \sum_{j=0}^{n-1} (\|e_j\|_{1,\Omega} + \|e_{j+1}\|_{1,\Omega}).$$

Putting all our estimations together, we deduce

$$\begin{aligned}\|e_n\|_{1,\Omega} &\lesssim \|I_1\|_{1,\Omega} + \|I_2\|_{1,\Omega} + \|I_3\|_{1,\Omega} + \|I_4\|_{1,\Omega} \\ &\lesssim \tau_{\triangle}^2 + \tau_{\triangle} \sum_{j=0}^{n-1} {t_{n-j+1}^{\triangle}}^{-1/2} (\|e_j\|_{1,\Omega} + \|e_{j+1}\|_{1,\Omega}) \\ &\lesssim \tau_{\triangle}^2 + \tau_{\triangle} \sum_{j=1}^{n} {t_{n-j+1}^{\triangle}}^{-1/2} \|e_j\|_{1,\Omega}.\end{aligned}$$

This means that the error satisfies

$$\|e_n\|_{1,\Omega} \le C_2\tau_{\triangle}^2 + C_2\tau_{\triangle} \sum_{j=1}^{n} \hat{t}_{n-j+1}^{-1/2} \|e_j\|_{1,\Omega},$$

with $\hat{t}_j = j\tau_{\triangle}$, $j = 0, \dots, n$, where C_2 is a constant. Now we get $(1 - C_2\tau_{\triangle}^{1/2}) \ge \frac{1}{2} > 0$ if we choose τ^* small enough, say $\tau^* \le \min\{\varepsilon_3, \frac{1}{4C_2^2}\}$, where we will give ε_3 in part (c) of the proof. Setting $C = 2C_2$ we easily get

$$\begin{aligned}\|e_n\|_{1,\Omega} &\le \underbrace{\frac{C_2}{1 - C_2\tau_{\triangle}^{1/2}}}_{\le \frac{C_2}{1-\frac{C_2}{2C_2}} = 2C_2} \tau_{\triangle}^2 + \frac{C_2\tau_{\triangle}}{1 - C_2\tau_{\triangle}^{1/2}} \sum_{j=1}^{n-1} \hat{t}_{n-j+1}^{-1/2} \|e_j\|_{1,\Omega} \\ &\le C\tau_{\triangle}^2 + C\tau_{\triangle} \sum_{j=1}^{n-1} \hat{t}_{n-j+1}^{-1/2} \|e_j\|_{1,\Omega}.\end{aligned} \tag{5.22}$$

Now, let us define $v(t) = \|e_j\|_{1,\Omega}$ for $t \in [\hat{t}_{j-1}, \hat{t}_j)$, $j = 1, \dots, n$. One may deduce

$$\begin{aligned}\int_0^t (t-s)^{-1/2} v(s)\, ds &\ge \sum_{j=1}^{m-1} \int_{\hat{t}_{j-1}}^{\hat{t}_j} (\underbrace{t-s}_{\le \hat{t}_m - \hat{t}_{j-1}})^{-1/2}\, ds\, \|e_j\|_{1,\Omega} \\ &\ge \tau_{\triangle} \sum_{j=1}^{m-1} \hat{t}_{m-j+1}^{-1/2} \|e_j\|_{1,\Omega}\end{aligned} \tag{5.23}$$

for $m = 1, \dots, n$ and $t \in [\hat{t}_{m-1}, \hat{t}_m)$. For each $t \in [0, \hat{t}_n)$, we choose $m \in \{1, \dots, n\}$, such that $t \in [\hat{t}_{m-1}, \hat{t}_m)$. From (5.22) and (5.23) we obtain

$$v(t) = \|e_m\|_{1,\Omega} \le C\,\tau_\triangle^2 + C\,\tau_\triangle \sum_{j=1}^{n-1} \hat{t}_{n-j+1}^{-1/2} \|e_j\|_{1,\Omega} \le C\,\tau_\triangle^2 + C \int_0^t (t-s)^{-1/2} v(s)\,ds.$$

Setting $\hat{a} = C\,\tau_\triangle^2$, $\hat{b} = C$, $\hat{\alpha} = 0$, and $\hat{\beta} = \frac{1}{2}$, Gronwall's lemma (Lemma 5.1) leads to

$$\|e_n\|_{1,\Omega} \lesssim \tau_\triangle^2.$$

(c) Our aim now is to choose τ^* to prove (5.5) for $j = 0, \dots, M_\triangle$. We show this by induction on $j \in \{0, 1, \dots, M_\triangle\}$.

Initial step. For $j = 0$, we have $p_\tau(0) = p(0) = p_0$, and hence (5.5) is true.

Inductive step. To show (5.5) for $j+1$, set $\gamma_1 := \frac{\hat{\delta}}{6M_\triangle} < \frac{\hat{\delta}}{2}$ and $\gamma_2 := \frac{(2j+1)\hat{\delta}}{6M_\triangle} < \frac{\hat{\delta}}{2}$. Consider the function

$$D : \overline{B(p(t_{j+1}^\triangle), \gamma_1)} \subset H^1_{\Gamma_D}(\Omega) \to \overline{B(p(t_{j+1}^\triangle), \gamma_1)},$$

$$D(q) := r_{2,2}(\tau_{j+1}^\triangle A) p(t_j^\triangle) - \int_0^{\tau_{j+1}^\triangle} r_{0,2}(\tau_{j+1}^\triangle A)\sigma A f(\tilde{q}(\sigma)) + g(\tau_{j+1}^\triangle, \sigma, \tilde{q}(\sigma))d\sigma,$$

where $\tilde{q}(\sigma) := (1 - (\sigma/\tau_{j+1}^\triangle))p(t_j^\triangle) + (\sigma/\tau_{j+1}^\triangle)\, q$ ($\overline{B(p(t_{j+1}^\triangle), \gamma_1)}$ is the closed ball with center $p(t_{j+1}^\triangle)$ and radius γ_1). From

$$\begin{aligned}\|p(t_j^\triangle + \sigma) - \tilde{q}(\sigma)\|_{1,\Omega} &\le \|p(t_j^\triangle + \sigma) - I_\triangle p(t_j^\triangle + \sigma)\|_{1,\Omega} + \|\tilde{q}(\sigma) - I_\triangle p(t_j^\triangle + \sigma)\|_{1,\Omega} \\ &\le \underbrace{C_1 \tau_\triangle^2}_{\le \frac{\hat{\delta}}{2}\text{ , because } \tau^* \le \varepsilon_0} + \underbrace{\|q - p(t_{j+1}^\triangle)\|_{1,\Omega}}_{\le \gamma_1 < \frac{\hat{\delta}}{2}} < \hat{\delta}\end{aligned}$$

we obtain by basic estimations

$$\begin{aligned}&\|p(t_{j+1}^\triangle) - D(q)\|_{1,\Omega} \\ &\quad\le \frac{c_1}{2}\,{\tau_{j+1}^\triangle}^{-1/2} \int_0^{\tau_{j+1}^\triangle} \|f(p(t_j^\triangle + \sigma)) - f(\tilde{q}(\sigma))\|_{0,\Omega} \\ &\qquad\qquad + \|g(\tau_{j+1}^\triangle, \sigma, p(t_j^\triangle + \sigma)) - g(\tau_{j+1}^\triangle, \sigma, \tilde{q}(\sigma))\|_{1,\Omega}\, d\sigma + \underbrace{\|d_{j+1}\|_{1,\Omega}}_{\le \hat{C}{\tau_{j+1}^\triangle}^2 \text{ cf. (5.18)}} \\ &\quad\le c_1 \hat{L} \tau_{j+1}^\triangle {}^{-1/2} \int_0^{\tau_{j+1}^\triangle} \underbrace{\|p(t_j^\triangle + \sigma) - \tilde{q}(\sigma)\|_{1,\Omega}}_{<\hat{\delta}} + \hat{C}{\tau_{j+1}^\triangle}^{1/2} \le {\tau_{j+1}^\triangle}^{1/2}(c_1 \hat{L}\hat{\delta} + \hat{C}),\end{aligned}$$

where c_1 and $\hat{C}$ are constants. On the one hand, for $\tau^* \le \min\{\varepsilon_0, (\gamma_1/(c_1\hat{L}\hat{\delta} + \hat{C}))^2\} =: \varepsilon_1$, $D(q) \in \overline{B(p(t_{j+1}^\triangle), \gamma_1)}$. So D is well-defined. One the other hand, we get

$$\|D(q_1) - D(q_2)\|_{1,\Omega} \le c_1 \hat{L} {\tau_{j+1}^\triangle}^{1/2} \|q_1 - q_2\|_{1,\Omega}.$$

For $\tau^* \le \min\{\varepsilon_1, (1/2c_1\hat{L})^2\} =: \varepsilon_2$ and all $q_1, q_2 \in \overline{B(p(t_{j+1}^{\triangle}), \gamma_1)}$ we obtain

$$\|D(q_1) - D(q_2)\|_{1,\Omega} \le \frac{1}{2}\,\|q_1 - q_2\|_{1,\Omega}.$$

Hence, Banach's fixed-point theorem leads to convergence of the sequence $q_{m+1} := D(q_m)$, where $q_0 \in \overline{B(p(t_{j+1}^{\triangle}), \gamma_1)}$.

Setting $p^{\bullet} := \lim_{m\to\infty} q_m$, we deduce

$$\|p^{\bullet} - p(t_{j+1}^{\triangle})\|_{1,\Omega} \le \gamma_1. \tag{5.24}$$

Furthermore, define the function

$$\tilde{D} : \overline{B(p^{\bullet}, \gamma_2)} \subset H^1_{\Gamma_D}(\Omega) \to \overline{B(p^{\bullet}, \gamma_2)},$$

where

$$\tilde{D}(q) = r_{2,2}(\tau_{j+1}^{\triangle}A)p_\tau(t_j^{\triangle}) - \int_0^{\tau_{j+1}^{\triangle}} r_{0,2}(\tau_{j+1}^{\triangle}A)\sigma A f(\hat{q}(\sigma)) + g(\tau_{j+1}^{\triangle}, \sigma, \hat{q}(\sigma))d\sigma,$$

and $\hat{q}(\sigma) := (1 - (\sigma/\tau_{j+1}^{\triangle}))p_\tau(t_j^{\triangle}) + (\sigma/\tau_{j+1}^{\triangle})\,q$. Using the assumption of the induction for j, we obtain

$$\|\tilde{p}^{\bullet}(\sigma) - \hat{q}(\sigma)\|_{1,\Omega} \le \max\{\underbrace{\|p_\tau(t_j^{\triangle}) - p(t_j^{\triangle})\|_{1,\Omega}}_{\le \frac{j\hat{\delta}}{3M_{\triangle}}}, \|p^{\bullet} - q\|_{1,\Omega}\} < \frac{\hat{\delta}}{2},$$

which leads to

$$\begin{aligned}
\|p^{\bullet} - \tilde{D}(q)\|_{1,\Omega} = \ & \|D(p^{\bullet}) - \tilde{D}(q)\|_{1,\Omega} \le \ \|p(t_j^{\triangle}) - p_\tau(t_j^{\triangle})\|_{1,\Omega} \\
& + \frac{c_1}{2}{\tau_{j+1}^{\triangle}}^{-1/2} \int_0^{\tau_{j+1}^{\triangle}} \|f(\tilde{p}^{\bullet}(\sigma)) - f(\hat{q}(\sigma))\|_{0,\Omega} \\
& + \|g(\tau_{j+1}^{\triangle}, \sigma, \tilde{p}^{\bullet}(\sigma)) - g(\tau_{j+1}^{\triangle}, \sigma, \hat{q}(\sigma))\|_{1,\Omega}\, d\sigma \\
\le \ & \frac{j\delta}{3M_{\triangle}} + c_1\hat{L}{\tau_{j+1}^{\triangle}}^{1/2}\gamma_2.
\end{aligned}$$

For $\tau^* \le \min\{\varepsilon_2, (\hat{\delta}/c_1\hat{L}\gamma_2 6M_{\triangle})^2\} =: \varepsilon_3$, we obtain

$$\frac{j\hat{\delta}}{3M_{\triangle}} + c_1\hat{L}{\tau_{j+1}^{\triangle}}^{1/2}\gamma_2 \le \frac{j\hat{\delta}}{3M_{\triangle}} + \frac{\hat{\delta}}{6M_{\triangle}} = \frac{(2j+1)\hat{\delta}}{6M_{\triangle}} = \gamma_2.$$

Hence, $\tilde{D}$ is well-defined. On the other hand, for $\tau^* \le \varepsilon_2$ and all $q_1, q_2 \in \overline{B(p^{\bullet}, \gamma_2)}$ we get

$$\|\tilde{D}(q_1) - \tilde{D}(q_2)\|_{1,\Omega} \le \frac{1}{2}\,\|q_1 - q_2\|_{1,\Omega}.$$

This leads to convergence of the sequence $q_{m+1} := \tilde{D}(q_m)$ using Banach's fixed-point theorem, where again $q_0 \in \overline{B(p^{\bullet}, \gamma_2)}$. Hence, for the limit of the fixed-point iteration

we know $p_\tau(t^\triangle_{j+1}) = \lim_{m\to\infty} q_{m+1}$, and so we have $\|p^\bullet - p_\tau(t^\triangle_{j+1})\|_{1,\Omega} \le \gamma_2$. Putting this result together with (5.24), we get

$$\|p(t^\triangle_{j+1}) - p_\tau(t^\triangle_{j+1})\|_{1,\Omega} \le \|p^\bullet - p(t^\triangle_{j+1})\|_{1,\Omega} + \|p^\bullet - p_\tau(t^\triangle_{j+1})\|_{1,\Omega} \le \gamma_1 + \gamma_2 = \frac{(j+1)\hat{\delta}}{3M_\triangle}. \quad \square$$

Now we are interested in a result for the whole interval $[0,T]$.

COROLLARY 5.3. *With the conditions of Theorem* 5.2*, we get*

$$\max_{t\in[0,T]} \|p(t) - p_\tau(t)\|_{1,\Omega} \le C\tau^2_\triangle.$$

Proof. Let $\hat{t} \in [0,T]$ such that

$$\max_{t\in[0,T]} \|p(t) - p_\tau(t)\|_{1,\Omega} = \|p(\hat{t}) - p_\tau(\hat{t})\|_{1,\Omega}.$$

From the triangle inequality we obtain

$$\|p(\hat{t}) - p_\tau(\hat{t})\|_{1,\Omega} \le \|p(\hat{t}) - I_\triangle p(\hat{t})\|_{1,\Omega} + \|I_\triangle p(\hat{t}) - p_\tau(\hat{t})\|_{1,\Omega}. \tag{5.25}$$

To estimate the first term on the right-hand side, we use the interpolation theory to get

$$\|p(\hat{t}) - I_\triangle p(\hat{t})\|_{1,\Omega} \le C_1\tau^2_\triangle.$$

Now choose $0 \le n \le M_\triangle$ such that $\hat{t} \in [t^\triangle_n, t^\triangle_{n+1}]$. The functions $I_\triangle p$ and p_τ are linear on $[t^\triangle_n, t^\triangle_{n+1}]$; hence we get

$$\|I_\triangle p(\hat{t}) - p_\tau(\hat{t})\|_{1,\Omega} \le \max\{\|e_n\|_{0,\Omega}, \|e_{n+1}\|_{1,\Omega}\} \underbrace{\le}_{\text{cf. Theorem 5.2}} C^*\tau^2_\triangle. \quad \square$$

6. The minimization problem. This section addresses the issues related to the solution of the elliptic problems at each time step. We approximate now the least-squares functional $\hat{\mathcal{F}}$ with Simpson's rule and get

$$\begin{aligned}
\hat{\mathcal{F}}(u_\tau, p_\tau) \approx{}& \frac{1}{6}\|p^+_\tau - p_\tau(t) + \tau \operatorname{div} u^-_\tau + \tau f(p_\tau(t))\|^2_{0,\Omega} \\
&+ \frac{2}{3}\left\|p^+_\tau - p_\tau(t) + \tau\frac{1}{2}(\operatorname{div} u^-_\tau + \operatorname{div} u^+_\tau) + \tau f\left(\frac{p^+_\tau + p_\tau(t)}{2}\right)\right\|^2_{0,\Omega} \\
&+ \frac{1}{6}\|p^+_\tau - p_\tau(t) + \tau \operatorname{div} u^+_\tau + \tau f(p^+_\tau)\|^2_{0,\Omega} \\
&+ \frac{\tau}{6}\|u^-_\tau + a^{1/2}\nabla p_\tau(t)\|^2_{0,\Omega} + \frac{\tau}{6}\|u^+_\tau + \nabla p^+_\tau\|^2_{0,\Omega} \\
&+ \frac{\tau}{6}\|(u^-_\tau + u^+_\tau) + (\nabla p_\tau(t) + \nabla p^+_\tau)\|^2_{0,\Omega} = \mathcal{F}(u^-_\tau, u^+_\tau, p^+_\tau; p_\tau(t)).
\end{aligned} \tag{6.1}$$

Now our goal is to minimize the least-squares functional $\mathcal{F}$ in $H_{\Gamma_N}(\operatorname{div},\Omega)^2 \times H^1_{\Gamma_D}(\Omega)$.

Obviously we see

$$\mathcal{F}(u^-_\tau, u^+_\tau, p^+_\tau; p_\tau(t)) = \|\mathcal{R}(u^-_\tau, u^+_\tau, p^+_\tau; p_\tau(t))\|^2_{0,\Omega}, \tag{6.2}$$

where

$$\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) = \begin{pmatrix} \frac{1}{\sqrt{6}}\,(p_\tau^+ - p_\tau(t) + \tau \mathrm{div} u_\tau^- + \tau\, f(p_\tau(t))) \\ \sqrt{\frac{2}{3}}\,(p_\tau^+ - p_\tau(t) + \frac{\tau}{2}\mathrm{div}(u_\tau^- + u_\tau^+) + \tau\, f(\frac{p_\tau^+ + p_\tau(t)}{2})) \\ \frac{1}{\sqrt{6}}\,(p_\tau^+ - p_\tau(t) + \tau \mathrm{div} u_\tau^+ + \tau f(p_\tau^+)) \\ \sqrt{\frac{\tau}{6}}\,(\,u_\tau^- + \nabla p_\tau(t)) \\ \sqrt{\frac{\tau}{6}}\,(\,u_\tau^- + u_\tau^+ + \nabla p_\tau(t) + \nabla p_\tau^+) \\ \sqrt{\frac{\tau}{6}}\,(\,u_\tau^+ + \nabla p_\tau^+) \end{pmatrix}.$$

Considering $\mathcal{F}$ on $H_{\Gamma_N}(\mathrm{div}, \Omega)^2 \times H^1_{\Gamma_D}(\Omega)$ and (4.1) lead to

$$\mathcal{F}'(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) \equiv 0 \quad \text{in } H_{\Gamma_N}(\mathrm{div}, \Omega)^2 \times H^1_{\Gamma_D}(\Omega), \tag{6.3}$$

where

$$\mathcal{F}'(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) \,:\, H_{\Gamma_N}(\mathrm{div}, \Omega)^2 \times H^1_{\Gamma_D}(\Omega) \longrightarrow \mathbb{R},$$

$$\begin{aligned} &\mathcal{F}'(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)) \begin{pmatrix} v_\tau^- \\ v_\tau^+ \\ q_\tau \end{pmatrix} \\ &\quad = \left(\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)), \mathcal{J}_\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+) \begin{pmatrix} v_\tau^- \\ v_\tau^+ \\ q_\tau \end{pmatrix} \right)_{0,\Omega}, \end{aligned} \tag{6.4}$$

and

$$\mathcal{J}_\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+) = (\partial_j \mathcal{R}_i)_{i=1,\ldots,6, j=1,\ldots,3}$$

is the *Jacobi matrix* of $\mathcal{R}$ at the point $(u_\tau^-, u_\tau^+, p_\tau^+)$. Elementary calculus leads to

$$\mathcal{J}_\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+) \,:\, H_{\Gamma_N}(\mathrm{div}, \Omega)^2 \times H^1_{\Gamma_D}(\Omega) \longrightarrow (L^2(\Omega))^9,$$

where

$$\mathcal{J}_\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+) = \frac{1}{\sqrt{6}} \begin{pmatrix} \tau\mathrm{div} & 0 & 1 \\ 2\,\tau\mathrm{div} & 2\,\tau\mathrm{div} & 4 + 2\tau f'(\frac{p_\tau^+ + p_\tau(t)}{2}) \\ 0 & \tau\mathrm{div} & 1 + \tau f'(p_\tau^+) \\ \sqrt{\tau} & 0 & 0 \\ \sqrt{\tau} & \sqrt{\tau} & \sqrt{\tau}\nabla \\ 0 & \sqrt{\tau} & \sqrt{\tau}\nabla \end{pmatrix}.$$

Now we obtain from (6.3) and (6.4) that (4.1) is equivalent to

$$\left(\mathcal{R}(u_\tau^-, u_\tau^+, p_\tau^+; p_\tau(t)), \mathcal{J}_{\mathcal{R}}(u_\tau^-, u_\tau^+, p_\tau^+) \begin{pmatrix} v_\tau^- \\ v_\tau^+ \\ q_\tau \end{pmatrix} \right)_{0,\Omega} = 0 \tag{6.5}$$

for all $(v_\tau^-, v_\tau^+, q_\tau) \in V_\tau^2 \times Q_\tau$.

There are several solution methods which directly use this nonlinear variational formulation. One of them is the nonlinear conjugate gradient method (cf. [20, section 5.2]), which in each step constructs a descent direction with the help of the gradient (6.4). The study of this solution is beyond the scope of this paper. Alternatively, one can first linearize this nonlinear minimization problem to use methods of Newton or Gauss–Newton type. We will choose this way to receive the fully discrete approximation in the next part of this paper.

The minimization problem of this paper could also be generalized for nonautonomous problems. The proof of the convergence result in this case is a little different from the technique observed here. Interested readers are referred to [16, Chapters 3, 4].

Acknowledgments. The author would like to thank Travis Austin, Michael Florig, and Gerhard Starke for reading this manuscript and for their suggestions. Furthermore the author is very thankful to the anonymous referees for helpful comments and suggestions.

REFERENCES

[1] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Local error estimates and adaptive refinement for first-order system least squares*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43.
[2] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations* I, Impact Comput. Sci. Engrg., 2 (1990), pp. 279–317.
[3] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations* II, Impact Comput. Sci. Engrg., 3 (1991), pp. 93–122.
[4] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations* III, Impact Comput. Sci. Engrg., 4 (1992), pp. 1–45.
[5] P. Deuflhard and A. Hohmann, *Numerische Mathematik* I, De Gruyter, Berlin, 1993.
[6] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* I: *A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.
[7] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* II: *Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.
[8] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* IV: *Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.
[9] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* V: *Long-time integration*, SIAM J. Numer. Anal., 32 (1995), pp. 1750–1763.
[10] K. Eriksson, C. Johnson, and S. Larsson, *Adaptive finite element methods for parabolic problems* VI: *Analytic semigroups*, SIAM J. Numer. Anal., 35 (1998), pp. 1315–1325.
[11] C. González and C. Palencia, *Stability of Runge-Kutta methods for abstract time-dependent parabolic problems: The Hölder case*, Math. Comp., 68 (1999), pp. 73–89.
[12] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer, New York, 1981.
[13] C. Johnson, Y.-Y. Nie, and V. Thomée, *An a posteriori error estimate and adaptive timestep control for a backward Euler discretization of a parabolic problem*, SIAM J. Numer. Anal., 27 (1990), pp. 277–291.
[14] T. Kato, *Perturbation Theory for Linear Operators*, Springer, New York, 1984.
[15] M. Majidi, *Adaptive Rothe-Verfahren für ein nichtlineares parabolisches Anfangs-Randwert-problem in der Hydrologie*, Diplomarbeit, 1999.

[16] M. Majidi, *Adaptive Finite-Element-Ausgleichsformulierungen für Parabolische Anfangs-Randwertaufgaben*, Ph.D. thesis, Universität Hannover, Hannover, Germany, 2002.
[17] M. Majidi, *Least-squares Galerkin methods for parabolic problems* IV: *The fully discrete case for semilinear problems*, 2005, in preparation.
[18] M. Majidi and G. Starke, *Least-squares Galerkin methods for parabolic problems* I: *Semi-discretization in time*, SIAM J. Numer. Anal., 39 (2001), pp. 1302–1323.
[19] M. Majidi and G. Starke, *Least-squares Galerkin methods for parabolic problems* II: *The fully discrete case and adaptive algorithms*, SIAM J. Numer. Anal., 39 (2002), pp. 1648–1666.
[20] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.
[21] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, Berlin, Heidelberg, New York, 1983.
[22] I. Richards and H. Youn, *Theory of Distributions: A Non-Technical Introduction*, Cambridge University Press, New York, Port Chester, Melbourne, Sydney, 1990.
[23] G. Starke, *Least-squares mixed finite element solution of variably saturated subsurface flow problems*, SIAM J. Sci. Comput., 21 (2000), pp. 1869–1885.
[24] M. E. Taylor, *Partial Differential Equations* III *(Nonlinear Equations)*, Appl. Math. Sci. 117, Springer, New York, 1997.
[25] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer, Berlin, 1997.
[26] D.-P. Yang, *Some least-squares Galerkin procedures for first-order time-dependent convection-diffusion system*, Comput. Methods Appl. Mech. Engrg., 180 (1999), pp. 81–95.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2476–2495

A $\mathcal{V}$-CYCLE MULTIGRID APPROACH FOR MORTAR FINITE ELEMENTS*

BARBARA I. WOHLMUTH†

Abstract. Mortar methods, based on dual Lagrange multipliers, provide a flexible tool for the numerical approximation of partial differential equations. The associated finite element spaces are, in general, nonconforming and nonnested. Optimal multigrid results have previously been established for $\mathcal{W}$-cycle and the variable $\mathcal{V}$-cycle multigrid methods. In this paper, we introduce a new multigrid method based on a nested sequence of modified mortar spaces for which we can establish that the $\mathcal{V}$-cycle with one smoothing step has contraction numbers uniformly bounded away from one. To obtain nested mortar spaces, we apply a product form of certain corrections at the interfaces. Numerical results demonstrate the efficiency of the resulting multigrid solver.

Key words. dual space, mortar finite elements, multigrid methods, nonmatching triangulations

AMS subject classifications. 65N30, 65N55

DOI. 10.1137/S003614290343092X

1. Introduction.

Nonconforming domain decomposition techniques, such as mortar methods, provide a more flexible approach than standard conforming approximations of elliptic problems. Different physical models, discretization schemes, and nonmatching triangulations of the domain can be coupled across interior interfaces by using mortar methods. There has also been substantial progress in establishing a theoretical foundation. The central issue is the choice of interface conditions that ensure stable and optimal discretization schemes for the problem as a whole. Within the framework of mortar methods, the information transfer across the interfaces is given in terms of a weak continuity condition. The jump of the solution at the interfaces is controlled by suitable Lagrange multiplier spaces [6, 7] satisfying an approximation property and a uniform inf-sup condition [5].

In this paper, we will use dual Lagrange multipliers which yield locally defined basis functions; see, e.g., [26]. Dual Lagrange multipliers provide the same accuracy as standard multipliers [6, 7] and give rise to more efficient iterative solvers and sparser local stiffness matrices. In particular, no mass matrices have to be inverted at the interfaces, and the mortar projection can be computed locally. The numerical solution of the resulting linear systems can be based on a symmetric positive definite system or on an equivalent saddle point formulation. Many different iterative solvers have been introduced and analyzed, among them iterative substructuring methods, which have been applied successfully to mortar finite elements; see [1, 2, 3, 17]. The Schur complement plays an important role in the construction of efficient iterative solvers based on the saddle point formulation. Using the techniques introduced in [12], multigrid methods for the indefinite system have been successfully analyzed in the mortar setting; see [8, 9]. The exact solution of an approximative Schur complement can be replaced by an inexact solution; see [25, 30]. A cascadic multigrid method is analyzed

*Received by the editors July 1, 2003; accepted for publication (in revised form) June 22, 2004; published electronically March 31, 2005. This work was supported in part by the Deutsche Forschungsgemeinschaft, SFB 404, C12.

http://www.siam.org/journals/sinum/42-6/43092.html

†Institut für Angewandte Analysis und Numerische Simulation, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany (wohlmuth@mathematik.uni-stuttgart.de, http://www.ians.uni-stuttgart.de/nmh).

in [10]. A first optimal multigrid convergence result for the positive definite formulation on the constrained space can be found in [18, 19]. Convergence rates which are uniformly bounded away from one are established for a variable $\mathcal{V}$-cycle result and numerical results illustrate the performance. In [28], a positive definite mesh dependent bilinear form is considered on the unconstrained product space. Based on a modified transfer operator and a modified smoother, level independent $\mathcal{W}$-cycle convergence results can be established if the number of smoothing steps is large enough. General results for multigrid algorithms for nonconforming finite elements are also given in [14, 15].

In this paper, we will work with the positive definite system and consider a modified multigrid approach. Our new iterative solver is based on a nested sequence of finite element spaces. To obtain a hierarchy of nested spaces, the nonconforming mortar spaces have to be modified at the interfaces. To preserve the complexity of a standard multigrid algorithm, such modifications must be kept local. The construction of the nested spaces is based on a multiplicative correction on each level. A suitable modification of the spaces at the crosspoints yields optimal best approximations results and a condition number which is bounded as in the standard conforming case. As a result we obtain contraction numbers uniformly bounded away from one with respect to the refinement level for the $\mathcal{V}$-cycle with one smoothing step. For simplicity of notation, we will talk about level independent convergence rates. The algebraic formulation of the prolongation can be easily obtained from the standard transfer operator and a local postprocessing of smaller complexity.

The rest of this paper is organized as follows: In section 2, we briefly review the mortar finite element method. We introduce a hierarchy of nested spaces in section 3. These spaces satisfy the mortar condition with respect to the finest triangulation. In section 4, we establish optimal multigrid convergence results in terms of a best approximation property and a bound on the condition number. In section 5, we consider the algebraic formulation of the problem in more detail and show that the modifications can easily be carried out in a local postprocessing step. Finally in section 6, we illustrate the numerical performance of our new multigrid approach.

2. The elliptic problem and the mortar method. We will consider the following elliptic second order boundary value problem:

$$\begin{aligned} -\mathrm{div}(a\nabla u) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

Here, $0 < a_0 \leq a \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, and $\Omega \subset \mathbb{R}^2$ is a bounded polygonal domain. Let Ω be decomposed into K nonoverlapping polygonal subdomains Ω_k such that $\overline{\Omega} = \bigcup_{k=1}^{K} \overline{\Omega}_k$. The intersection between the boundaries of any pair of subdomains $\partial\Omega_l \cap \partial\Omega_k$, $k \neq l$, is assumed to be either empty, a vertex, or a common edge. We will talk about an interface only in the latter case. A point is a crosspoint if it is in Ω and an endpoint of an interface. The set of crosspoints is denoted by $\mathcal{C}$. On each subdomain, we define a quasi-uniform simplicial triangulation $\mathcal{T}_k$ and denote the finite-dimensional space of conforming piecewise linear finite elements on $\mathcal{T}_k$, with homogeneous Dirichlet boundary conditions on $\partial\Omega \cap \partial\Omega_k$, by X_k. The mortar method can now be characterized by discrete Lagrange multiplier spaces M_m defined on the interfaces γ_m, $1 \leq m \leq M$, of the decomposition. For each interface, there exists a pair $1 \leq l(m) < k(m) \leq K$ such that $\overline{\gamma}_m = \partial\Omega_{l(m)} \cap \partial\Omega_{k(m)}$. Moreover, each interface γ_m is associated with a one-dimensional mesh $\mathcal{S}_m$, inherited from either $\mathcal{T}_{k(m)}$ or $\mathcal{T}_{l(m)}$. Thus, the elements of $\mathcal{S}_m$ are boundary edges of either $\mathcal{T}_{l(m)}$ or

$\mathcal{T}_{k(m)}$. The subdomain from which the interface γ_m inherits its mesh is called the slave subdomain $\Omega_{s(m)}$ and the one opposite is the master subdomain $\Omega_{m(m)}$. We denote the trace space, $X_{s(m)}$, restricted to γ_m, and with zero boundary conditions by W_m; i.e., $W_m := \{w \in H_0^1(\gamma_m),\ w = v_{|\gamma_m} \text{ with } v \in X_{s(m)}\}$.

Following the approach in [6, 7], the mortar approximation u_h can be obtained as the solution of a positive definite nonconforming variational problem. It is defined on a subspace V_h of the unconstrained product space $X_h := \{v \in L^2(\Omega) \mid \ v_{|\Omega_k} \in X_k,\ 1 \le k \le K\}$. The elements of V_h satisfy a weak matching condition across the interfaces, and V_h is given by

$$V_h := \big\{v \in X_h \mid b(v,\mu) = 0,\ \mu \in M_h\big\},$$

where $M_h := \prod_{m=1}^{M} M_m$ and

$$b(v,\mu) := \sum_{m=1}^{M} \int_{\gamma_m} [v]\ \mu\ d\sigma, \quad \mu \in M_h, v \in X_h.$$

Here, the jump is defined as $[v]_{|\gamma_m} := v_{|\Omega_{m(m)}} - v_{|\Omega_{s(m)}}$; see, e.g., [6, 7]. The nonconforming formulation of the mortar method can then be written in terms of the constrained space V_h: find $u_h \in V_h$ such that

$$a(u_h, v_h) = (f, v_h)_0, \quad v_h \in V_h, \tag{2.2}$$

where the bilinear form $a(\cdot,\cdot)$ is defined as

$$a(v,w) := \sum_{k=1}^{K} \int_{\Omega_k} a \nabla v \cdot \nabla w\ dx, \quad v, w \in X := \prod_{k=1}^{K} H^1(\Omega_k).$$

It is obvious that the quality of the nonconforming approach (2.2) and the properties of V_h depend on the discrete Lagrange multiplier space. We will work with the dual Lagrange multiplier space introduced in [26] which satisfy $n_m := \dim M_m = \dim W_m$. The basis functions $\{\mu_i\}_{1 \le i \le n_m}$ of M_m and the nodal basis functions $\{\varphi_i\}_{1 \le i \le n_m}$ of the trace space W_m satisfy the following biorthogonality relation:

$$\int_{\gamma_m} \mu_i\ \varphi_j\, d\sigma = \delta_{ij} \int_{\gamma_m} \varphi_j\, d\sigma, \quad 1 \le i, j \le n_m.$$

Such locally defined dual Lagrange multiplier spaces can also be defined for higher order finite elements; see [24]. In this paper, we use piecewise linear but discontinuous Lagrange multiplier basis functions, but we note that there also exist continuous Lagrange multiplier basis functions; see [27].

3. New nested spaces. In the following, we denote the unconstrained product spaces associated with a nested sequence of quasi-uniform and shape regular triangulations $\mathcal{T}^l$, $l = 0, \dots, L$, by $X^l := \prod_{k=1}^{K} X_k^l$. We assume that $\mathcal{T}^{l+1}$ is obtained from $\mathcal{T}^l$ by uniform refinement and that the mesh sizes satisfy $2\, h_{l+1} = h_l$. The constrained mortar spaces are denoted by V^l, the dual Lagrange multiplier spaces by $M^l := \prod_{m=1}^{M} M_m^l$, and the trace spaces having zero boundary conditions by $W^l := \prod_{m=1}^{M} W_m^l$. We note that, in general, the global triangulations $\mathcal{T}^l$, $l = 0, \dots, L$, do not match across the interfaces. The corresponding nodal basis functions of X^l,

V^l, and M^l on level l are denoted by $\{\theta_p^l\}_{p\in\mathcal{N}_u^l}$, $\{\phi_p^l\}_{p\in\mathcal{N}_c^l}$, and $\{\lambda_p^l\}_{p\in\mathcal{N}_s^l}$, respectively, where $\mathcal{N}_u^l$, $\mathcal{N}_c^l$, and $\mathcal{N}_s^l$ denote the sets of nodes. We note that $\mathcal{N}_u^l = \mathcal{N}_c^l \cup \mathcal{N}_s^l$ and that $\{\theta_p^l\}_{p\in\mathcal{N}_s^l}$ restricted to the interfaces defines a basis of W_h which is biorthogonal to $\{\lambda_p^l\}_{p\in\mathcal{N}_s^l}$. In contrast to the unconstrained product spaces X^l, the nonconforming mortar spaces V^l are nonnested; i.e., $V^l \not\subset V^{l+1}$. In our multigrid analysis, we will assume that the problem is H^2-regular.

The central idea is to replace V^l in a first step by $\widetilde{V}^l$ yielding a nested sequence of finite element spaces, i.e., with $\widetilde{V}^0 \subset \widetilde{V}^1 \subset \cdots \subset \widetilde{V}^L := V^L$, while maintaining the same approximation properties. Then, in a second step, we replace $\widetilde{V}^l$, $0 \le l < L$, by a smaller space $\widehat{V}^l$ that is continuous at the crosspoints. We note that each element v in V^l with $[v] = 0$ for all interfaces is an element in V^{l+1}. But, generally, $v \in V^l$ will not guarantee that $b(v,\mu) = 0$ for all $\mu \in M^{l+1}$ and thus $v \notin V^{l+1}$.

On each level, we therefore introduce a locally defined linear operator $Q^l : X^l \longrightarrow X^l$ by

$$Q^l v := v - \sum_{p\in\mathcal{N}_s^l} \frac{b(v,\lambda_p^l)}{b(\theta_p^l,\lambda_p^l)}\theta_p^l. \tag{3.1}$$

Due to the duality of the Lagrange multiplier space, we find for each λ_q^l, $q \in \mathcal{N}_s^l$,

$$b(Q^l v,\lambda_q^l) = b(v,\lambda_q^l) - \sum_{p\in\mathcal{N}_s^l} \frac{b(v,\lambda_p^l)}{b(\theta_p^l,\lambda_p^l)}\, b(\theta_p^l,\lambda_q^l) = b(v,\lambda_q^l) - b(v,\lambda_q^l) = 0,$$

and thus $Q^l v \in V^l$. Moreover, Q^l restricted to V^l is the identity, and the kernel of Q^l is span $\{\theta_p^l, p \in \mathcal{N}_s^l\}$. We now define our new spaces by $\widetilde{V}^l := \text{span}\,\{\varphi_p^l, p \in \mathcal{N}_c^l\}$, where the basis functions φ_p^l are given by

$$\varphi_p^l := Q^L\ Q^{L-1}\ \ldots\ Q^{l+1}\ \phi_p^l, \quad p \in \mathcal{N}_c^l. \tag{3.2}$$

By construction, we have $\widetilde{V}^L = V^L$ and $\dim \widetilde{V}^l \le \dim V^l$. Moreover, we find that $\varphi_p^l(q) = \delta_{pq}$, $p,q \in \mathcal{N}_c^l$, and thus the φ_p^l are linearly independent yielding $\dim \widetilde{V}^l = \dim V^l$. We note that the elements of $\widetilde{V}^l$ are, in general, not level l functions. Observing that $\phi_p^l = Q^l\theta_p^l$, $p \in \mathcal{N}_c^l$, the new basis functions φ_p^l can also be written in terms of a subset of the nodal basis functions θ_p^l

$$\varphi_p^l = P^l\ \theta_p^l := Q^L\ Q^{l-1}\ \ldots\ Q^{l+1}\ Q^l\ \theta_p^l, \quad p \in \mathcal{N}_c^l.$$

Using the definition of P^l, it is easy to see that $\widetilde{V}^l = P^l V^l$. By construction, the elements of $\widetilde{V}^l$ satisfy the mortar condition with respect to the finest level L and thus $\widetilde{V}^l \subset V^L$. The application of P^l is illustrated in Figure 3.1. For each level $k = l, \ldots, L$, we add a local level k correction to $v \in X^l$ on the slave sides. These level k corrections are nonzero only in a strip of width h_k in the slave subdomains; see Figure 3.1

To see that the spaces $\widetilde{V}^l$ are nested, we consider the basis functions in more detail. Using the definition (3.1), we obtain $Q^l\theta_q^l = \sum_{p\in\mathcal{N}_u^{l+1}} \alpha_p^q \theta_p^{l+1}$. Observing that $Q^{l+1}\theta_p^{l+1} = 0$ for $p \in \mathcal{N}_s^{l+1}$, we find

$$\varphi_q^l = P^l\theta_q^l = P^{l+1} \sum_{p\in\mathcal{N}_u^{l+1}} \alpha_p^q \theta_p^{l+1} = \sum_{p\in\mathcal{N}_c^{l+1}} \alpha_p^q P^{l+1}\theta_p^{l+1} = \sum_{p\in\mathcal{N}_c^{l+1}} \alpha_p^q \varphi_p^{l+1},$$

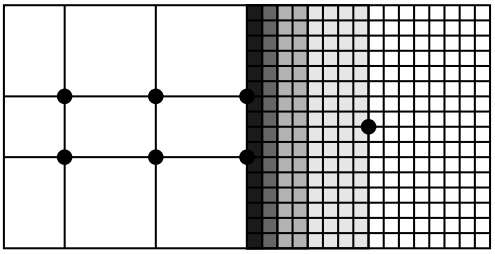


FIG. 3.1. *Action of P^l, where the master side is on the left and $l = L - 3$.*

and thus $\widetilde{V}^l \subset \widetilde{V}^{l+1}$.

The definition of P^l as a product provides the natural framework within a multigrid approach. We find that P^l is closely related to the mortar projection. Introducing the mortar projection $\pi_m^l : L^2(\gamma_m) \longrightarrow W_m^l$, by

$$\int_{\gamma_m} \pi_m^l v \mu \, d\sigma = \int_{\gamma_m} v \mu \, d\sigma, \quad \mu \in M_m^l, \tag{3.3}$$

and $\Pi^l : X^l \longrightarrow W^l$, by $(\Pi^l v)_{|\gamma_m} := \pi_m^l([v]_{|\gamma_m})$, we can rewrite the projection Q^l as

$$Q^l = \mathrm{Id} \; - E^l \Pi^l,$$

where $E^l : W^l \longrightarrow X^l$ is defined by $E^l := \sum_{m=1}^M E_m^l$, and E_m^l denotes the trivial extension by zero from W_m^l into $X_{s(m)}^l \subset X^l$; i.e., all coefficients in the basis representation being not associated with nodes in the interior of γ_m are set to be zero. We note that π_m^l restricted to W_m^l is the identity. The operator Q^l plays a crucial role in the multigrid analysis. In the next lemma, we briefly recall the essential properties of the mortar projection (3.3). For a proof in the dual Lagrange multiplier setting, we refer to [26].

LEMMA 3.1. *The mortar projection π_m^l satisfies the following properties with constants independent of the refinement level* l*:*

- *It is $H_{00}^{1/2}$-stable, i.e.,*

$$\|\pi_m^l v\|_{H_{00}^{1/2}(\gamma_m)} \leq C \|v\|_{H_{00}^{1/2}(\gamma_m)}, \quad v \in H_{00}^{1/2}(\gamma_m).$$

- *It satisfies an approximation property*

$$\|v - \pi_m^l v\|_{0;\gamma_m} \leq C h_l |v|_{1;\gamma_m}, \quad v \in H_0^1(\gamma_m).$$

To establish level independent multigrid convergence rates, we have to consider the spaces $\widetilde{V}^l$ in more detail. In particular, appropriate approximation properties are essential. As a preliminary step in this direction, we give an equivalent formulation of P^l.

LEMMA 3.2. *The operator P^l, defined in terms of a product, can also be expressed as a sum over different levels,*

$$P^l = \mathrm{Id} \; - \sum_{k=l+1}^{L} E^k (\Pi^k - \Pi^{k-1}) - E^l \Pi^l.$$

Proof. The proof is obtained by induction. By definition, we have $Q^l = \mathrm{Id} - E^l\Pi^l$. Let us assume that $Q^j \dots Q^l = \mathrm{Id} - \sum_{k=l+1}^{j} E^k(\Pi^k - \Pi^{k-1}) - E^l\Pi^l$ holds. We then get

$$Q^{j+1}\, Q^j \dots Q^l = (\mathrm{Id} - E^{j+1}\Pi^{j+1}) \left(\mathrm{Id} - \sum_{k=l+1}^{j} E^k(\Pi^k - \Pi^{k-1}) - E^l\Pi^l \right)$$

$$= \mathrm{Id} - \sum_{k=l+1}^{j} E^k(\Pi^k - \Pi^{k-1}) - E^l\Pi^l - E^{j+1}\Pi^{j+1} \tag{3.4}$$

$$+ \sum_{k=l+1}^{j} E^{j+1}\Pi^{j+1}E^k(\Pi^k - \Pi^{k-1}) + E^{j+1}\Pi^{j+1}E^l\Pi^l.$$

We recall that $E^k\Pi^k$ and $E^k\Pi^{k-1}$ when restricted to the interfaces are Π^k and Π^{k-1}, respectively. Moreover, the trace spaces are nested; i.e., $W^k \subset W^{k+1}$, and thus $\Pi^{k+i}E^k\Pi^k = \Pi^k$, $i \geq 0$. Using these observations in (3.4), we find

$$Q^{j+1}\, Q^j \dots Q^l = \mathrm{Id} - \sum_{k=l+1}^{j} E^k(\Pi^k - \Pi^{k-1}) - E^l\Pi^l - E^{j+1}\Pi^{j+1}$$

$$+ \sum_{k=l+1}^{j} E^{j+1}(\Pi^k - \Pi^{k-1}) + E^{j+1}\Pi^l$$

$$= \mathrm{Id} - \sum_{k=l+1}^{j} E^k(\Pi^k - \Pi^{k-1}) - E^{j+1}(\Pi^{j+1} - \Pi^j) - E^l\Pi^l. \qquad \square$$

4. $\boldsymbol{\mathcal{V}}$-cycle multigrid convergence analysis. The two basic tools to establish level independent multigrid convergence rates are an approximation and a smoothing property. The approximation property for the mortar space V^l is well known; see, e.g., [26]. Here, we have to consider the modified spaces $\widetilde{V}^l$. In the following, we work with the broken H^1-norm which is defined by $\|\cdot\|_1^2 := \sum_{k=1}^{K} \|\cdot\|_{1;\Omega_k}^2$.

LEMMA 4.1. *The modified spaces $\widetilde{V}^l$, $0 \leq l \leq L$, satisfy an approximation property, i.e.,*

$$\inf_{v \in \widetilde{V}^l} \|u - v\|_1 \leq C h_l |u|_2, \quad u \in H^2(\Omega) \cap H_0^1(\Omega),$$

where the constant $C < \infty$ does not depend on L and h_l.

Proof. The starting point for the proof is Lemma 3.2. We define $w := Q^l I^l u$, where I^l is the broken Lagrange interpolation operator onto X^l. Then, it is well known that

$$\|u - w\|_1 \leq C h_l |u|_2, \quad u \in H_0^1(\Omega) \cap H^2(\Omega);$$

see [26]. An analogous result for standard Lagrange multipliers can be found in [6, 7]. We recall that $w \in V^l$ and thus $\Pi^l w = 0$. Defining $v := P^l w \in \widetilde{V}^l$ and using the additive structure of P^l, we find

$$\|u - v\|_1 = \|u - w + \sum_{k=l+1}^{L} E^k(\Pi^k - \Pi^{k-1})w\|_1$$

$$\leq \|u - w\|_1 + \sum_{k=l+1}^{L} \|E^k(\Pi^k - \Pi^{k-1})w\|_1.$$

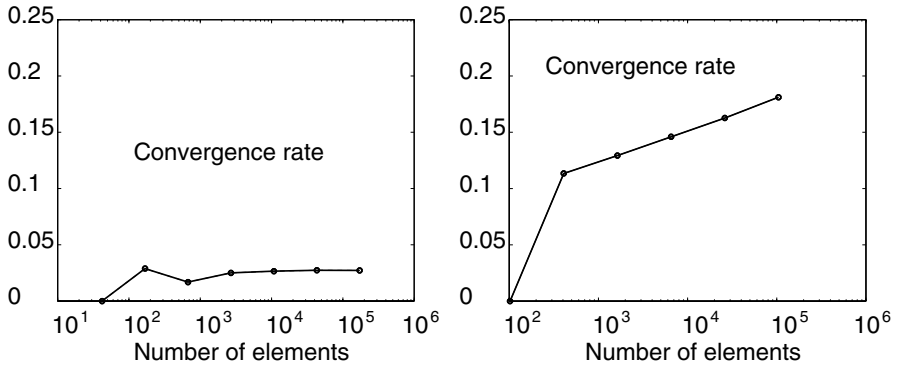


FIG. 4.1. *Convergence rates based on $\widetilde{V}^l$, no crosspoints (left) and crosspoints (right).*

A standard inverse estimate provides for the trivial extension on level k

$$\|E^k(\Pi^k - \Pi^{k-1})w\|_1 \leq \frac{C}{\sqrt{h_k}}\|(\Pi^k - \Pi^{k-1})w\|_{0;\mathcal{S}},$$

where $\mathcal{S} := \cup_{m=1}^M \gamma_m$. In what follows, we denote by $\|\cdot\|_{t;\mathcal{S}}$ and $|\cdot|_{t;\mathcal{S}}$ the broken H^t-norm and H^t-seminorm on $\mathcal{S}$; i.e.,

$$\|v\|_{t;\mathcal{S}}^2 := \sum_{m=1}^M \|v\|_{t;\gamma_m}^2, \text{ and } \quad |v|_{t;\mathcal{S}}^2 := \sum_{m=1}^M |v|_{t;\gamma_m}^2,$$

respectively. The definition of w guarantees that $[w]_{|\gamma_m} \in H_0^1(\gamma_m)$, and thus we are in the setting of Lemma 3.1. In terms of the approximation property, we find

$$\begin{aligned}\|(\Pi^k - \Pi^{k-1})w\|_{0;\mathcal{S}} &\leq \|\Pi^k w - [w]\|_{0;\mathcal{S}} + \|\Pi^{k-1}w - [w]\|_{0;\mathcal{S}} \\ &\leq Ch_k|[w]|_{1;\mathcal{S}} \;=\; Ch_k|[w-u]|_{1;\mathcal{S}}.\end{aligned}$$

Combining these results and using $2h_k = h_{k-1}$, we obtain

$$\begin{aligned}\|u-v\|_1 &\leq C\left(\|u-w\|_1 + \sum_{k=l+1}^L \sqrt{h_k}|[w-u]|_{1;\mathcal{S}}\right) \\ &\leq C\left(\|u-w\|_1 + \sqrt{h_l}|[w-u]|_{1;\mathcal{S}}\right) \\ &\leq C\left(\|u-w\|_1 + \sqrt{h_l}|[I^l u - u]|_{1;\mathcal{S}} + \sqrt{h_l}|[\Pi^l I^l u]|_{1;\mathcal{S}}\right) \\ &\leq C\left(\|u-w\|_1 + h_l|u|_{3/2;\mathcal{S}}\right) \;\leq\; Ch_l|u|_2.\end{aligned}$$

To bound $|[\Pi^l I^l u]|_{1;\mathcal{S}}$, we have used an inverse estimate and the stability of the mortar projection in the $H_{00}^{1/2}$-norm. ∎

Unfortunately, the nested spaces $\widetilde{V}^l$ do not, in general, provide optimal multigrid results. Figure 4.1 shows the asymptotic convergence rates of a $\mathcal{V}$-cycle with one smoothing step.

In the left picture, the domain is decomposed into two subdomains and there are no crosspoints. For this situation, we observe level independent convergence rates.

However, in the more general case of many subdomains, this does not hold. This is illustrated in the right picture. For a decomposition into 13 subdomains and with six crosspoints, we find that the convergence rates depend on the refinement level. To advance our understanding, we consider the stiffness matrix $\widetilde{A}^l$ associated with $\widetilde{V}^l$ in more detail. As in the standard case, its smallest eigenvalue is of order $1/(h_l)^2$ but its largest eigenvalue is not uniformly bounded independently of the refinement level. This is caused by the modified basis functions φ_p^l, associated with the crosspoints, for which the energy is not bounded by a constant. As a consequence, the condition number of $\widetilde{A}^l$ is not of order $1/(h_l)^2$, and we do not obtain level independent convergence rates.

To obtain better numerical results, we have to work with different spaces. We introduce a subspace of $\widetilde{V}^l$. Let $\widehat{V}^l$ be defined by

$$\begin{aligned}\widehat{V}^L &:= V^L,\\ \widehat{V}^l &:= \{v \in \widetilde{V}^l \mid [v]_{|\gamma_m} \in H_{00}^{1/2}(\gamma_m), 1 \le m \le M\}, \quad 0 \le l \le L-1.\end{aligned}$$

The condition $[v]_{|\gamma_m} \in H_{00}^{1/2}(\gamma_m)$ for all interfaces is equivalent to making v continuous at the crosspoints. We note that $\widehat{V}^l \not\subset H_0^1(\Omega)$ and as a consequence, the Galerkin orthogonality does not hold. The proof of Lemma 4.1 gives rise to the following corollary. By definition $w := Q^l I^l u$ is continuous at the crosspoints and thus $v := P^l w \in \widehat{V}^l$.

COROLLARY 4.2. *The nested spaces $\widehat{V}^l$, $0 \le l \le L$, satisfy an approximation property, i.e.,*

$$\inf_{v \in \widehat{V}^l} \|u - v\|_1 \le C h_l |u|_2, \quad u \in H^2(\Omega) \cap H_0^1(\Omega),$$

where the constant $C < \infty$ does not depend on L and h_l.

We now define $\widehat{u}^l \in \widehat{V}^l$ as the solution of the following positive definite variational problem: find $\widehat{u}^l \in \widehat{V}^l$ such that

$$a(\widehat{u}^l, v) = (f, v)_0, \quad v \in \widehat{V}^l.$$

To obtain optimal a priori estimates, the consistency error has to be of order h_l.

LEMMA 4.3. *The discrete finite element solution $\widehat{u}^l$, $0 \le l \le L$, satisfies the following a priori estimate for $u \in H_0^1(\Omega) \cap H^2(\Omega)$:*

$$\|u - \widehat{u}^l\|_1 + \frac{1}{h_l}\|u - \widehat{u}^l\|_0 \le C h_l |u|_2,$$

where the constant does not depend on L and h_l.

Proof. We start with the jump of an element $w \in \widehat{V}^l$. Using the orthogonality of $[w]$ to M^L, we find

$$\|[w]\|_{0;\mathcal{S}}^2 = ([w],[w])_{0;\mathcal{S}} \le \|[w]\|_{0;\mathcal{S}} \inf_{\mu \in M^L} \|[w] - \mu\|_{0;\mathcal{S}} \le C\sqrt{h_L}\, |[w]|_{1/2;\mathcal{S}}\, \|[w]\|_{0;\mathcal{S}}.$$

In terms of the approximation property of M^L and the orthogonality of $[w]$ to M^L, the upper bound for the jump results in a bound on the consistency error given by

$$\sup_{\substack{w \in \widehat{V}^l \\ w \neq 0}} \frac{\int_{\mathcal{S}} \frac{\partial u}{\partial n}[w]\, d\sigma}{\|w\|_1} \le C\sqrt{h_L} \inf_{\mu \in M^L} \|\frac{\partial u}{\partial n} - \mu\|_{0;\mathcal{S}} \sup_{\substack{w \in \widehat{V}^l \\ w \neq 0}} \frac{|[w]|_{1/2;\mathcal{S}}}{\|w\|_1} \le C h_L |u|_2.$$

We note that the consistency error of an element in $\widehat{V}^l$ is, in general, smaller than for an element in V^l. It is of order h_L for all $v \in \widehat{V}^l$, $0 \le l \le L$, whereas it is of order h_l for $v \in V^l$. This is due to the fact that the jump is orthogonal to M^L and not only to M^l. The space M^l is associated with the mesh on level l whereas M^L is associated with that on level L and $\dim M^L > \dim M^l$. Then, the a priori estimate in the energy norm follows from Corollary 4.2. The a priori estimate in the L^2-norm can easily be obtained by using the H^2-regularity and the Aubin–Nitsche trick for nonconforming elements. In that case, the consistency error of the dual problem enters additionally in the upper bound. ∎

Although the spaces $\widehat{V}^l$ are nested, we are in a nonconforming setting. In contrast to standard conforming approaches, no Galerkin orthogonality holds, and we cannot bound $\|u - \widehat{u}^l\|_0$ by $Ch_l\|u - \widehat{u}^l\|_1$. However, a weaker result can be established.

COROLLARY 4.4. *There exists a constant independent of the level L such that*

$$\|\widehat{u}^l - \widehat{u}^{l-1}\|_0 \le Ch_l\|\widehat{u}^l - \widehat{u}^{l-1}\|_1, \quad 0 \le l \le L.$$

Proof. We start with a discrete Galerkin orthogonality

$$a(\widehat{u}^l - \widehat{u}^{l-1}, v) = 0, \quad v \in \widehat{V}^{l-1}.$$

Introducing $w \in H_0^1(\Omega)$ by $a(w, v) = (\widehat{u}^l - \widehat{u}^{l-1}, v)_0$, $v \in H_0^1(\Omega)$ and $\widehat{w}^k \in \widehat{V}^k$ by $a(\widehat{w}^k, v) = (\widehat{u}^l - \widehat{u}^{l-1}, v)_0$, $v \in \widehat{V}^k$, we find

$$\begin{aligned}\|\widehat{u}^l - \widehat{u}^{l-1}\|_0^2 &= a(\widehat{w}^l, \widehat{u}^l - \widehat{u}^{l-1}) = a(\widehat{w}^l - \widehat{w}^{l-1}, \widehat{u}^l - \widehat{u}^{l-1}) \\ &\le C\|\widehat{u}^l - \widehat{u}^{l-1}\|_1\|\widehat{w}^l - \widehat{w}^{l-1}\|_1 \le C\|\widehat{u}^l - \widehat{u}^{l-1}\|_1(\|\widehat{w}^l - w\|_1 + \|\widehat{w}^{l-1} - w\|_1).\end{aligned}$$

Then, the a priori estimate in the energy norm and the H^2-regularity yield the upper bound for $\|\widehat{u}^l - \widehat{u}^{l-1}\|_0$. ∎

Following the approach in [16], we have to consider the condition number of the modified stiffness matrix $\widehat{A}^l$ associated with the nested spaces $\widehat{V}^l$ to obtain level independent convergence rates. The ellipticity of the bilinear form $a(\cdot, \cdot)$ on $\widehat{V}^l \times \widehat{V}^l$ guarantees

$$c\|v\|_0^2 \le a(v, v) \le C\|v\|_1^2, \quad v \in \widehat{V}^l, \quad 0 \le l \le L.$$

We define $\widehat{\mathcal{N}}_c^l$ as a subset of $\mathcal{N}_c^l$ such that $p \in \widehat{\mathcal{N}}_c^l$ if p does not coincide geometrically with a crosspoint. We now use the following basis of $\widehat{V}^l \subset \widetilde{V}^l$:

$$\{\varphi_p^l, \ p \in \widehat{\mathcal{N}}_c^l\} \cup \left\{ \sum_{p \in \mathcal{C}_q} \varphi_p^l, \ q \in \mathcal{C} \right\},$$

where $\mathcal{C}$ is the set of crosspoints of the domain decomposition, and $p \in \mathcal{C}_q$ if and only if $p \in \mathcal{N}_c^l$ coincides geometrically with the crosspoint $q \in \mathcal{C}$. In what follows, we use the same symbol for the function v as an element in the finite element space S, $S \in \{X^l, \widetilde{V}^l, \widehat{V}^l\}$, and its vector representation $v \in \mathbb{R}^{\dim S}$ with respect to the specified basis functions of S.

We denote by $\|\cdot\|_S$ the Euclidean vector norm. Then for the L^2-norm of $v \in \widehat{V}^l$, we find the standard estimate

$$\|v\|_0 \ge \frac{c}{h_l}\|v\|_{\widehat{V}^l}, \quad 0 \le l \le L,$$

where the constant $c > 0$ does not depend on the level L. We recall that the operator P^l does not modify the values at the interior and master nodes on level l. The following lemma provides an upper bound for the condition number of the stiffness matrix $\widehat{A}^l$.

LEMMA 4.5. *There exists a constant $C < \infty$ independent of the level L such that*

$$\kappa(\widehat{A}^l) \leq \frac{C}{h_l^2}, \quad 0 \leq l \leq L.$$

Proof. It is sufficient to establish a suitable upper bound for the energy norm of $v \in \widehat{V}^l$. Each $v \in \widehat{V}^l$ can be written as

$$v = Q^L \; Q^{L-1} \; \ldots \; Q^{l+1} \; Q^l w,$$

where $w \in X^l$ is continuous at the crosspoints and vanishes at all nodes $p \in \mathcal{N}_s^l$. We note that w is uniquely defined. Lemma 3.2 and the triangle inequality yield

$$\|v\|_1 \leq \|w - E^l \Pi^l w\|_1 + \sum_{j=l+1}^{L} \|E^j(\Pi^j - \Pi^{j-1})w\|_1. \tag{4.1}$$

In a first step, we establish an upper bound for $\|w - E^l\Pi^l w\|_1$. The definition of E^l and Π^l gives $z := w - E^l\Pi^l w \in V^l$. Applying an inverse estimate and using the continuity of z, we obtain

$$\begin{aligned} \|z\|_1^2 \leq \frac{C}{h_l^2}\|z\|_0^2 &\leq C \sum_{p \in \mathcal{N}_u^l} z(p)^2 \leq C\left(\sum_{p \in \mathcal{N}_c^l} z(p)^2 + \sum_{p \in \mathcal{N}_s^l} z(p)^2 \right) \\ &\leq C\left(\|z\|_{V^l}^2 + \sum_{p \in \mathcal{N}_s^l} z(p)^2 \right). \end{aligned}$$

The second term on the right-hand side can be bounded as follows: For each node $p \in \mathcal{N}_s^l$, we find due to the mortar condition on level l that

$$z(p) = \sum_{q \in \mathcal{N}_c^l} \frac{\int_{\mathcal{S}} z(q)[\phi_q^l]\mu_p^l \, d\sigma}{\int_{\mathcal{S}} \phi_p^l \, d\sigma}.$$

We denote the set of nodes $q \in \mathcal{N}_c^l$ such that supp $\phi_q^l \cap$ supp μ_p^l has a nonzero measure by $\mathcal{I}_p$. Moreover, the number of elements in $\mathcal{I}_p$ is bounded independently of p and L. Now, $z(p)$ can be bounded by

$$z(p)^2 \leq C \sum_{q \in \mathcal{I}_p} z(q)^2,$$

and as a result, we find $\sum_{p \in \mathcal{N}_s^l} z(p)^2 \leq C\|z\|_{V^l}^2$. Observing that Q^j, $j > l$, does not modify the values at the mortar nodes on level l, we get $z(p) = v(p)$ for all $p \in \mathcal{N}_c^l$. We note that dim $V^l \geq$ dim $\widehat{V}^l$ and thus $\|z\|_{V^l} \geq \|v\|_{\widehat{V}^l}$. However, by construction, z is continuous at the crosspoints and therefore $\|z\|_{V^l} \leq C\|v\|_{\widehat{V}^l}$. In a second step, we have to consider the second term on the right-hand side in (4.1). A standard inverse inequality shows

$$\begin{aligned} \|E^j(\Pi^j - \Pi^{j-1})w\|_1 &\leq \frac{C}{h_j}\|E^j(\Pi^j - \Pi^{j-1})w\|_0 \leq \frac{C}{\sqrt{h_j}}\|(\Pi^j - \Pi^{j-1})w\|_{0;\mathcal{S}} \\ &\leq \frac{C}{\sqrt{h_j}}\left(\|\Pi^j w - [w]\|_{0;\mathcal{S}} + \|\Pi^{j-1}w - [w]\|_{0;\mathcal{S}} \right). \end{aligned}$$

The continuity of w at the crosspoints guarantees that $[w]_{|\gamma_m} \in H_{00}^{1/2}(\gamma_m)$ for all interfaces γ_m, $1 \le m \le M$. Thus, we can apply Lemma 3.1 and get

$$\|E^j(\Pi^j - \Pi^{j-1})w\|_1 \le C\sqrt{h_j}|[w]|_{1;\mathcal{S}} \le C\frac{\sqrt{h_j}}{h_l}|[w]|_{0;\mathcal{S}} \le C\frac{\sqrt{h_j}}{\sqrt{h_l}}\|v\|_{\widehat{V}^l}.$$

Here, we have used $w \in X^l$ with $w(p) = v(p)$ for all interior and master nodes on level l and $w(p) = 0$ for all slave nodes $p \in \mathcal{N}_s^l$. In terms of (4.1), we obtain

$$\|v\|_1 \le C\sum_{j=l}^{L}\frac{\sqrt{h_j}}{\sqrt{h_l}}\|v\|_{\widehat{V}^l} \le C\sum_{j=0}^{\infty}\sqrt{2}^{-j}\|v\|_{\widehat{V}^l} \le C\|v\|_{\widehat{V}^l}. \qquad \square$$

We can now formulate our main result. Here, u_n^L stands for the nth multigrid iterate, and u^L is the mortar finite element solution on level L. Our multigrid method is based on the nested sequence $\widehat{V}^0 \subset \widehat{V}^1 \subset \cdots \subset \widehat{V}^{L-1} \subset V^L$. Moreover, we use the natural embedding to define the prolongation operator, and the restriction matrix is its transpose. We apply a damped Richardson iteration as the smoothing operator, and the damping factor ω on level l is bounded by $c/\widehat{\lambda}_{\max}^l \le \omega \le 1/\widehat{\lambda}_{\max}^l$, where $\widehat{\lambda}_{\max}^l$ is the maximal eigenvalue of the stiffness matrix associated with the nodal basis of $\widehat{V}^l$.

THEOREM 4.6. *There exists a constant independent of the refinement level L such that the $\mathcal{V}$-cycle multigrid convergence rate is given by*

$$a(u^L - u_n^L, u^L - u_n^L) \le \left(\frac{C}{C+m}\right)^2 a(u^L - u_{n-1}^L, u^L - \widehat{u}_{n-1}^L),$$

where m denotes the number of smoothing steps.

The condition number bound in combination with the approximation property and the construction of the nested sequence of spaces yields level independent convergence rates for the $\mathcal{V}$-cycle multigrid method with one smoothing step; see [11, 13, 20]. Using the approach given in [16], the proof follows from Corollary 4.4 and Lemma 4.5. The results can also be extended to other multigrid variants such as, e.g., the $\mathcal{W}$-cycle.

Remark 4.7. We note that on the finest level no continuity at the crosspoints is required.

5. Algebraic formulation. The multigrid method proposed in the previous section is based on the natural embedding of the spaces. In this section, we consider the algebraic representation of the prolongation and restriction and show how the stiffness matrix $\widehat{A}^l$ from Lemma 4.5 can be recursively assembled. We start by describing how to obtain the stiffness matrix associated with the nodal basis of V^L. Let us denote by A the stiffness matrix associated with the unconstrained product space X^L. For simplicity, we suppress the level index L. Then A has a block diagonal structure. We use a decomposition of the matrix into block matrices and three different sets of nodes: $\mathcal{N}_s^l$, $\mathcal{N}_m^l$, and $\mathcal{N}_i^l$. The index set $\mathcal{N}_i^l$ contains all nodes not on the interfaces $\overline{\gamma}_m$, $1 \le m \le M$. All nodes being on the master sides and all nodes coinciding geometrically with a crosspoint are in $\mathcal{N}_m^l$; i.e., $\mathcal{N}_m^l = \mathcal{N}_c^l \setminus \mathcal{N}_i^l$. We recall that the set $\mathcal{N}_s^l$ stands for the nodes on level l associated with the vertices in the interior of the slave sides. Now, each vector $x \in \mathbb{R}^{n_L}$, $n_L := \dim X^L$, can be decomposed according to the three different blocks in $x = (x_i, x_m, x_s)^T$. Then, each

element in $v \in V^L$ can be written in terms of the nodal basis of X^L, and we find for the algebraic representation

$$x = \begin{pmatrix} x_i \\ x_m \\ x_s \end{pmatrix} = \begin{pmatrix} \mathrm{Id} & 0 \\ 0 & \mathrm{Id} \\ 0 & M \end{pmatrix} \begin{pmatrix} v_i \\ v_m \end{pmatrix} =: Bv,$$

where M is a sparse mass matrix associated with the mortar projection. We note that x and v as functions are the same, but their algebraic representations differ. We observe that each $x \in X^L$ is in V^L if and only if it has the algebraic form $x = (x_i, x_m, Mx_m)^T$ with respect to the nodal basis of X^L. We now define the stiffness matrix A^{c} by

$$A^{\mathrm{c}} := \begin{pmatrix} A_{ii} & A_{im} + A_{is}M \\ A_{mi} + M^T A_{si} & A_{\mathrm{mod}} \end{pmatrix},$$

where A_{kl}, $k, l \in \{s, m, i\}$, are the block matrices of the stiffness matrix A associated with the nodal basis functions of X^L, and $A_{\mathrm{mod}} := A_{mm} + M^T A_{ss} M + A_{ms} M + M^T A_{sm}$.

LEMMA 5.1. *The variational problem* (2.1) *on level* L *is equivalent to the algebraic system*

$$A^{\mathrm{c}} u_L = f^{\mathrm{c}} := \begin{pmatrix} f_i \\ f_m + M^T f_s \end{pmatrix}.$$

Proof. Observing that each test function $v \in V^L$ has the algebraic form $v = (v_i, v_m, Mv_m)^T$ with respect to the nodal basis of X^L, we obtain $v^T f = (v_i, v_m)(f_i, f_m + M^T f_s)^T$, where f is the algebraic representation of the right-hand side with respect to the nodal basis of X^L. Thus, the variational problem (2.1) can be written as $(Bv)^T A(Bu_L) = v^T (B^T AB) u_L = (Bv)^T f = v^T (B^T f)$. It is now easy to verify that $A^{\mathrm{c}} = (B^T AB)$ and $f^{\mathrm{c}} = (B^T f)$. ∎

Remark 5.2. In contrast to standard mortar approaches the matrix A^{c} can easily be assembled from A by local post processing. A crucial role is played by the structure of M. Only in the case of a dual Lagrange multiplier space is M a sparse scaled mass matrix. In the case of standard Lagrange multipliers, M has the form $M = M_s^{-1} M_m$, where M_m is a rectangular mass matrix and M_s is tridiagonal.

In the rest of this section, we focus on the prolongation operator. Let I_l^{l+1} be the algebraic representation of the natural embedding of X^l in X^{l+1}, and let W_l be the matrix representation of the projection operator Q^l. Using the definition of $Q^l : X^l \longrightarrow X^l$, we find

$$W_l := \begin{pmatrix} \mathrm{Id} & 0 & 0 \\ 0 & \mathrm{Id} & 0 \\ 0 & M & 0 \end{pmatrix}. \tag{5.1}$$

The following lemma shows that the natural embedding of $\widetilde{V}^l$ in $\widetilde{V}^{l+1}$ can be expressed in terms of I_l^{l+1} and W_l.

LEMMA 5.3. *Let the interpolation matrix* Z_l^{l+1} *be defined by*

$$Z_l^{l+1} := \begin{pmatrix} (I_l^{l+1})_{ii} & (I_l^{l+1})_{im} + (I_l^{l+1})_{is} M_l \\ 0 & (I_l^{l+1})_{mm} \end{pmatrix},$$

where $(I_l^{l+1})_{jk}$, $j,k \in \{i,m,s\}$, *denote the block components of* I_l^{l+1}. *Then,* Z_l^{l+1} *is the matrix representation of the natural embedding of* $\widetilde{V}^l$ *in* $\widetilde{V}^{l+1}$.

Proof. We start by considering the nodal basis functions of $\widetilde{V}^l$ and $\widetilde{V}^{l+1}$ in more detail. The algebraic representation of φ_p^l in the nodal basis of X^L is given by

$$W_L I_{L-1}^L W_{L-1} I_{L-2}^{L-1} \dots I_l^{l+1} W_l e_p^l,$$

where e_p^l denotes the unit vector of $\mathbb{R}^{n_l}$, $n_l := \dim X^l$, associated with the node p on level l. Analogously, the algebraic presentation of φ_p^{l+1} in the nodal basis of X^L is given by

$$W_L I_{L-1}^L W_{L-1} I_{L-2}^{L-1} \dots I_{l+1}^{l+2} W_{l+1} e_p^{l+1},$$

where e_p^{l+1} denotes the unit vector of $\mathbb{R}^{n_{l+1}}$, $n_{l+1} := \dim X^{l+1}$, associated with the node p on level $l+1$. We now consider the natural embedding $\widetilde{V}^l \subset \widetilde{V}^{l+1}$. Each element $v \in \widetilde{V}^l \subset \widetilde{V}^{l+1} \subset X^L$ can be written uniquely as a linear combination of the nodal basis functions φ_p^l, φ_p^{l+1}, and θ_p^L

$$v = \sum_{p \in \mathcal{N}_c^l} v_p^l \varphi_p^l = \sum_{p \in \mathcal{N}_c^{l+1}} v_p^{l+1} \varphi_p^{l+1} = \sum_{p \in \mathcal{N}_u^L} x_p^L \theta_p^L.$$

We define $w^l \in \mathbb{R}^{n_l}$ by $w_p^l := v_p^l$, $p \in \mathcal{N}_c^l$ and $w_p^l := 0$, $p \in \mathcal{N}_s^l$, and $w^{l+1} \in \mathbb{R}^{n_{l+1}}$ by $w_p^{l+1} = v_p^{l+1}$, $p \in \mathcal{N}_c^{l+1}$ and $w_p^{l+1} = 0$, $p \in \mathcal{N}_s^{l+1}$. We then find

$$W_L I_{L-1}^L W_{L-1} I_{L-2}^{L-1} \dots I_l^{l+1} W_l w^l = W_L I_{L-1}^L W_{L-1} I_{L-2}^{L-1} \dots I_{l+1}^{l+2} W_{l+1} w^{l+1}.$$

In the next step, we decompose $I_l^{l+1} W_l w^l = c^{l+1} + y^{l+1}$, where $c_p^{l+1} = 0$, $p \in \mathcal{N}_s^{l+1}$ and $y_p^{l+1} = 0$, $p \in \mathcal{N}_c^{l+1}$. We note that this decomposition is unique. Moreover, the definition of W_{l+1} yields $W_{l+1} y^{l+1} = 0$ and thus $W_L I_{L-1}^L W_{L-1} I_{L-2}^{L-1} \dots I_l^{l+1} W_l w^l = W_L I_{L-1}^L W_{L-1} I_{L-2}^{L-1} \dots I_{l+1}^{l+2} W_{l+1} c^{l+1}$. Observing that the kernel of W_j, $1 \le j \le L$, is associated with the nodes $p \in \mathcal{N}_s^j$, we find $c^{l+1} = w^{l+1}$. A straightforward computation shows that

$$\begin{aligned} c^{l+1} &= \begin{pmatrix} \mathrm{Id} & 0 & 0 \\ 0 & \mathrm{Id} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} (I_l^{l+1})_{ii} & (I_l^{l+1})_{im} & (I_l^{l+1})_{is} \\ 0 & (I_l^{l+1})_{mm} & 0 \\ 0 & (I_l^{l+1})_{sm} & (I_l^{l+1})_{ss} \end{pmatrix} \begin{pmatrix} \mathrm{Id} & 0 & 0 \\ 0 & \mathrm{Id} & 0 \\ 0 & M_l & 0 \end{pmatrix} w^l \\ &= \begin{pmatrix} (I_l^{l+1})_{ii} & (I_l^{l+1})_{im} + (I_l^{l+1})_{is} M_l & 0 \\ 0 & (I_l^{l+1})_{mm} & 0 \\ 0 & 0 & 0 \end{pmatrix} w^l. \end{aligned}$$

Using the structure of w^{l+1} and w^l, we find

$$\begin{pmatrix} v^{l+1} \\ 0 \end{pmatrix} = \begin{pmatrix} Z_l^{l+1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v^l \\ 0 \end{pmatrix},$$

and thus $v^{l+1} = Z_l^{l+1} v^l$. ∎

Figure 5.1 illustrates the prolongation operator. For simplicity, we restrict ourselves to a function in $\widetilde{V}^l$ which vanishes at all interior vertices. The support of such a function is marked by the shadowed region. Then, the function, as an element in $\widetilde{V}^l$, is uniquely defined by its values at the vertices on the master side, which are

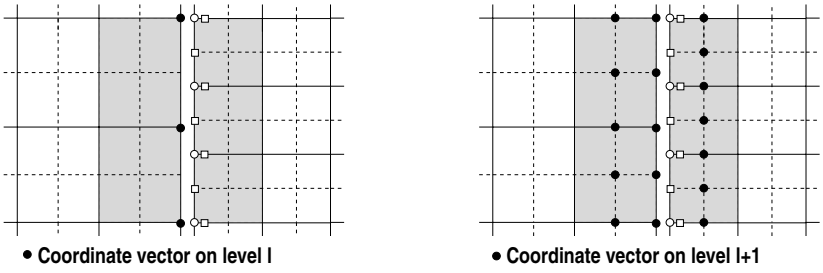


FIG. 5.1. *Prolongation operator from* $\widetilde{V}^l$ *onto* $\widetilde{V}^{l+1}$, $L := l+1$.

marked by filled circles in the left picture of Figure 5.1. The values on the master side are extended on the slave side in the prescribed multiplicative way such that the constraints at the interface on level L are satisfied. The vertices on the slave side on level l are marked by empty circles and on level $L := l+1$ by empty squares. We now interpret the function as an element in $\widetilde{V}^{l+1}$. In the right picture, the relevant vertices to specify the function are shown by filled circles. The values at the filled circles in the interior of the slave subdomain are obtained from the values at the empty circles and the standard prolongation. We note that the values at the empty squares do not contribute.

Remark 5.4. Although the spaces M^l are nonnested, we can assemble the scaled mass matrices M_l, $0 \le l \le L-1$, recursively. We note that M^l can be embedded in a higher dimensional but locally defined space. These macro spaces are nested and can be used for the assembly process. Thus, we have to compute the intersection of the edges on the slave and master sides only on the highest level.

Additionally, we have to consider the spaces $\widehat{V}^l$ which are continuous at the crosspoints. By definition, we have $\widehat{V}^l \subset \widetilde{V}^l$. We start by introducing new sets of nodes. Let $\mathcal{C}^l$ be the set of nodes which coincide geometrically with one crosspoint. For each crosspoint $c \in \mathcal{C}$, we define one master node p_c such that $p_c \in \mathcal{C}^l$ coincides geometrically with the crosspoint c. The choice is arbitrary but should be fixed. Moreover, we set $\mathcal{N}_g^l := \mathcal{N}_m^l \setminus \mathcal{C}^l$ and $\mathcal{C}_m^l := \cup_{c\in\mathcal{C}}\ p_c$. Now, each element $v \in \widetilde{V}^l$ and $v \in \widehat{V}^l$ can be written as (v_i, v_g, v_d) and (v_i, v_g, v_c), respectively. Here, the block vectors are connected to the basis functions associated with the nodes in $\mathcal{N}_i^l$, $\mathcal{N}_g^l$, $\mathcal{C}^l$ and $\mathcal{N}_i^l$, $\mathcal{N}_g^l$, $\mathcal{C}_m^l$, respectively. We note that $\mathcal{N}_m^l = \mathcal{N}_g^l \cup \mathcal{C}^l$. Then, the natural embedding $\widehat{V}^l \subset \widetilde{V}^l$ has the algebraic form

$$\begin{pmatrix} v_i \\ v_g \\ v_d \end{pmatrix} = \begin{pmatrix} \mathrm{Id} & 0 & 0 \\ 0 & \mathrm{Id} & 0 \\ 0 & 0 & C \end{pmatrix} \begin{pmatrix} v_i \\ v_g \\ v_c \end{pmatrix} =: R_l \begin{pmatrix} v_i \\ v_g \\ v_c \end{pmatrix},$$

where C is a block diagonal matrix. Each block diagonal entry C_{ii} corresponds to one crosspoint $c_i \in \mathcal{C}$ and has the form $C_{ii} := (1, 1, \ldots, 1)^T$, where the number of ones is equal to the number of nodes coinciding geometrically with the crosspoint c_i. We note that the number of columns of C is equal to the number of crosspoints and the number of rows is equal to the number of elements in $\mathcal{C}^l$. Thus the size of C does not depend on the level. On the other hand, each element $v \in \widehat{V}^l$, $l < L$, can be written

as

$$\begin{pmatrix} v_i \\ v_g \\ v_c \end{pmatrix} = \begin{pmatrix} \mathrm{Id} & 0 & 0 \\ 0 & \mathrm{Id} & 0 \\ 0 & 0 & DC^T \end{pmatrix} \begin{pmatrix} v_i \\ v_g \\ v_d \end{pmatrix} =: \widehat{R}_l^T \begin{pmatrix} v_i \\ v_g \\ v_d \end{pmatrix},$$

where $D \in \mathbb{R}^{n_c \times n_c}$ is a diagonal matrix, and n_c is the number of crosspoints. The entries are defined by $d_{ii} := 1/k_i$, where k_i is the number of nodes coinciding geometrically with the crosspoint c_i. We point out that DC^T does not depend on the level, and we set formally $\widehat{R}_L^T := \mathrm{Id}$. Then, the algebraic representation of the embedding of $\widehat{V}^l \subset \widehat{V}^{l+1}$, $0 \le l \le L-1$, is given by

$$\widehat{Z}_l^{l+1} = \widehat{R}_{l+1}^T Z_l^{l+1} R_l.$$

The coarse stiffness matrices $\widehat{A}_l$ are obtained recursively by a Galerkin assembly

$$\widehat{A}_l := (\widehat{Z}_l^{l+1})^T \widehat{A}_{l+1} \widehat{Z}_l^{l+1}, \quad 0 \le l \le L-1,$$

where $\widehat{A}_L := A^c$. We recall that A^c is the stiffness matrix associated with the constrained space V_L. The restriction of the defect d_{l+1} is given by $d_l := \widehat{Z}_l^T d_{l+1}$. We note that all modifications can be carried out locally. Additionally, they are of smaller complexity than one matrix vector multiplication, and only the nodes on the interfaces are involved.

Remark 5.5. In contrast to the approach given in [19], we work with a nested sequence of nonconforming spaces. The coarse spaces satisfy the constraints with respect to the highest level. As a result, we can use the natural embedding operator as prolongation. Using a Gauß–Seidel smoother, this yields a monotone variant. A combination of monotone multigrid techniques for Signorini problems [21, 22, 23] with our approach gives a globally convergent monotone multigrid strategy for variational inequalities based on interface constraints; see [29].

Remark 5.6. The construction of the spaces $\widehat{V}^l$ is restricted to the two-dimensional situation and has to be generalized in three dimensions (3D). In 3D, we cannot, in general, require continuity on the wirebasket. However, as in two dimensions, we can impose continuity at the crosspoints. Additionally, we have to impose a weak matching condition on the edges of the wire basket. This can be realized in terms of one master edge.

6. Numerical results. In this section, we present some numerical results for our multigrid method. They confirm the theoretical results and illustrate the performance of the method. In particular, we consider three different examples with several crosspoints. The numerical realization is based on the finite element toolbox ug [4],and uniform refinement is applied in each refinement step. In all our examples, we compare two different smoothers and show the influence of the number of smoothing steps on the convergence rates. On each level, the initial iterate is set to be zero. As stopping criteria, we use a relative tolerance of 5e−15.

In our first example, we consider a decomposition of the unit square $\Omega = (0,1) \times (0,1)$ into nine squares $\Omega_{ij} := ((i-1)/3, i/3) \times ((j-1)/3, j/3)$, $1 \le i,j \le 3$. The right-hand side f and the boundary conditions of $-\Delta u = f$ are chosen such that the exact solution is given by $u(x,y) = \sin(2\pi y)\exp(-x^{-2})\exp(-0.1(1-x)^{-2}) + \sin(\pi x)\exp(-1.25y^{-2})\exp(-0.1(1-y)^{-2})$. Figure 6.1 shows the decomposition into subdomains, the nonmatching triangulations and the isolines of the solution. The initial triangulation has 72 elements and is nonmatching at the interface.

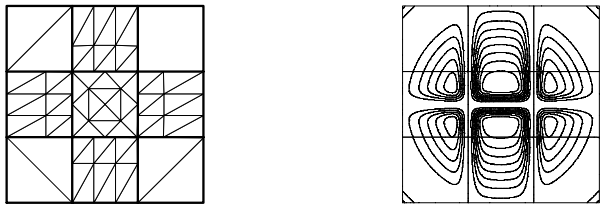


FIG. 6.1. *Decomposition into nine subdomains and initial triangulation (left), isolines of the solution (right) (Example* 1*).*

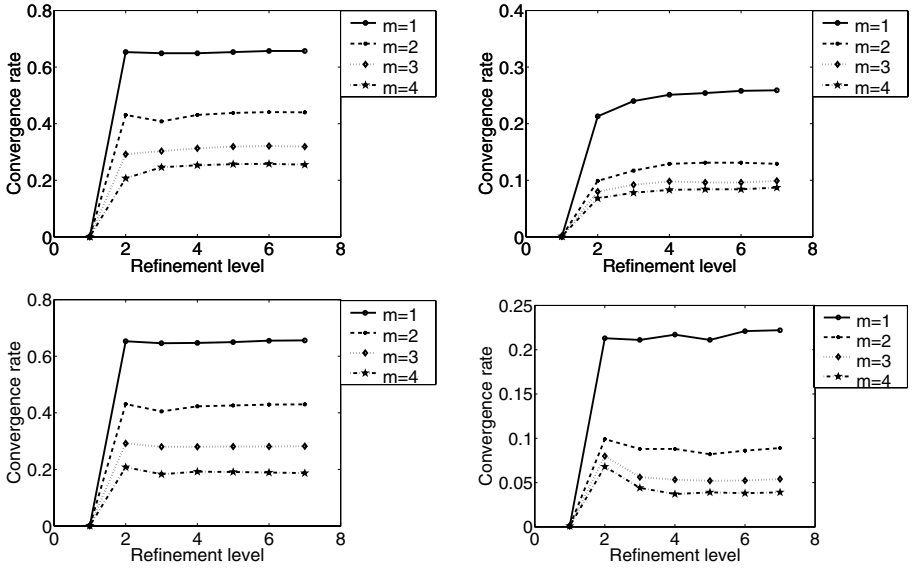


FIG. 6.2. *Asymptotic convergence rates for a damped Jacobi smoother (left) and a symmetric Gauß–Seidel one (right) (Example* 1*). In the first row the $\mathcal{V}(m,m)$-cycle results are shown, and in the second row those of the $\mathcal{W}(m,m)$-cycle results.*

We observe level independent convergence rates for all our tests. In particular, we compare two different types of smoothers. As expected the symmetric Gauß–Seidel smoother yields considerably better results than the damped Jacobi smoother, for $\omega = 0.8$. In each case, we apply m pre- and m postsmoothing steps, $1 \leq m \leq 4$. Increasing the number of smoothing steps gives better results. The first row in Figure 6.2 shows our numerical results for the $\mathcal{V}$-cycle and in the second row those of the $\mathcal{W}$-cycle. There is no big difference between the $\mathcal{V}$-cycle and $\mathcal{W}$-cycle results.

In our second example, we consider a nonconvex domain which is decomposed into 13 subdomains; see Figure 6.3 for the decomposition and the isolines of the solution. We find six crosspoints having four adjacent subdomains.

As in our first example, we observe asymptotically constant convergence rates. In the case of the Jacobi smoother the asymptotic starts later than in the case of the symmetric Gauß–Seidel smoother. We set the numerical convergence rate to be the reduction factor of the residual in the last iteration step. At the beginning of the iteration, $\|r^{l+1}\|/\|r^l\|$ is increasing. However, asymptotically this ratio tends to be a constant value on each refinement level. This constant value is not yet reached on

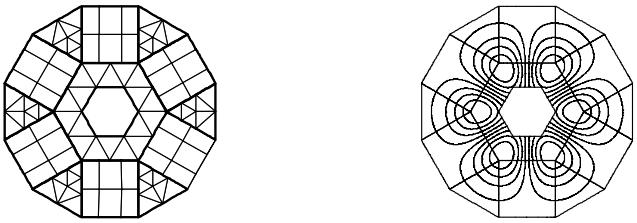


FIG. 6.3. *Decomposition into* 13 *subdomains and initial triangulation (left) and isolines of the solution (right) (Example* 2*).*

the higher levels for the $\mathcal{V}(1,1)$-cycle with the Gauß–Seidel smoother. As a result, we observe, in the right upper picture of Figure 6.4, a decreasing convergence rate. Comparing the multigrid convergence rates of $m = 1$ and $m = 2$, we find a big difference in the case of the Gauß–Seidel smoother. This is also true if we compare $m = 2$ and $m = 4$ in the case of the Jacobi smoother. Taking into account that the symmetric Gauß–Seidel smoother is more expensive than the Jacobi smoother, we cannot compare the results for the same m directly. Comparing the symmetric Gauß–Seidel smoother with $m = 2$ and the Jacobi smoother with $m = 4$, we find that the Gauß–Seidel smoother yields better convergence rates.

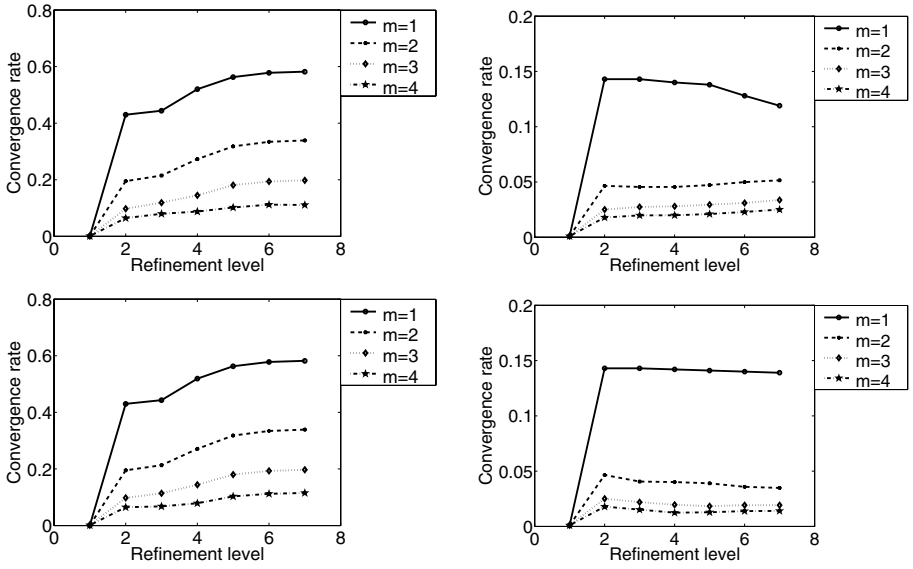


FIG. 6.4. *Asymptotic convergence rates for a damped Jacobi smoother (left) and a symmetric Gauß–Seidel one (right) (Example* 2*). In the first row the $\mathcal{V}(m,m)$-cycle results are shown and in the second row those of the $\mathcal{W}(m,m)$-cycle results.*

In our third example, we consider a decomposition of the unit square $(-0.5, 0.5)^2$ into three subdomains. Two of the subdomains are nonconvex. Figure 6.5 shows the decomposition into the subdomains, the nonmatching triangulations, and the isolines of the solution. The slave sides are chosen to be on the middle subdomain. The right-hand side f and the Dirichlet boundary conditions of $-\Delta u = f$ are chosen such that the exact solution is given by $x(x-y)\exp(-10(x^2+0.6y^2))$.

In this example, the differences between the $\mathcal{V}(m,m)$-cycle and $\mathcal{W}(m,m)$-cycle results are extremely small for the Jacobi smoother; see Figure 6.6. Moreover, the number of iteration steps is the same. The Gauß–Seidel smoother results in better convergence rates for the $\mathcal{W}(m,m)$-cycle. The $\mathcal{V}(1,1)$-cycle and the $\mathcal{W}(2,2)$-cycle show the same effect as the $\mathcal{V}(1,1)$-cycle in Example 2. We observe for all our examples considerably better results if the number of smoothing steps is greater than one.

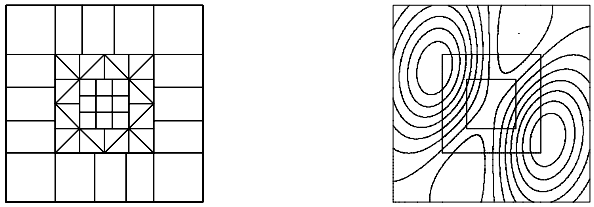


FIG. 6.5. *Decomposition into three subdomains and initial triangulation (left) and isolines of the solution (right) (Example* 3*).*

In contrast to the right picture in Figure 4.1, we observe level independent convergence rates for the $\mathcal{V}$-cycle. A comparison of Figures 4.1 and 6.4 shows the influence of the modification of the nonconforming spaces. The numerical results of Figure 4.1 are based on the nested sequence $\widetilde{V}^0 \subset \widetilde{V}^1 \subset \cdots \subset \widetilde{V}^L$. The condition number of the stiffness matrix associated with these spaces is not bounded by Ch_l^{-2}. As a result, we do not observe level independent convergence rates for the $\mathcal{V}(1,1)$-cycle.

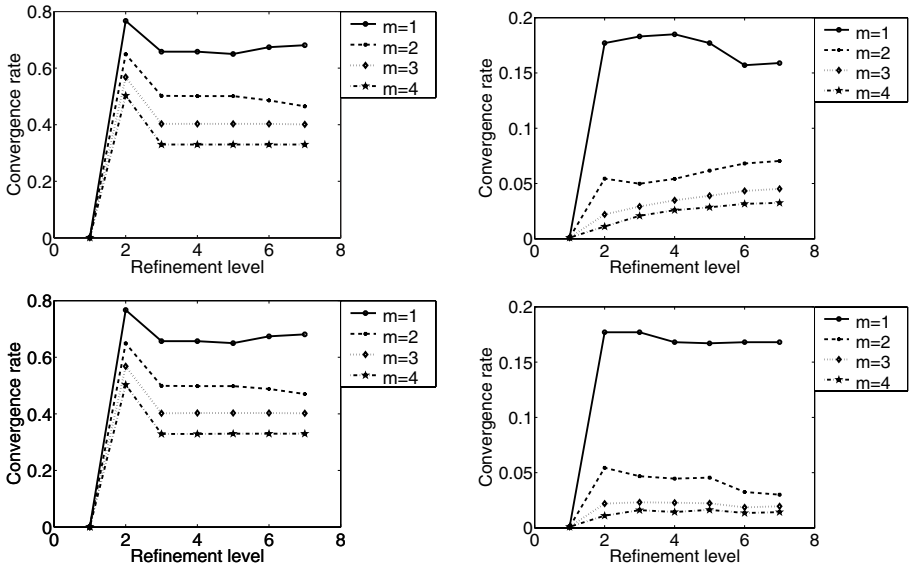


FIG. 6.6. *Asymptotic convergence rates for a damped Jacobi smoother (left) and a symmetric Gauß–Seidel one (right) (Example* 3*). In the first row the $\mathcal{V}(m,m)$-cycle results are shown and in the second row those of the $\mathcal{W}(m,m)$-cycle results.*

Compared to the more general saddlepoint multigrid method analyzed in [9, 25], the performance of this multigrid variant applied to a mortar discretization with dual Lagrange multipliers is considerably better. This is due to the fact that we can exploit the biorthogonality of the Lagrange multiplier and the trace space for the construction

of the solver. In case of the saddlepoint variant this cannot be done, and we have to include an inner iteration loop in the multigrid scheme. Applying this new variant to standard Lagrange multipliers increases the computational cost. The modified stiffness matrices can be easily obtained by local static condensation in the case of dual Lagrange multipliers. This does not hold for standard Lagrange multipliers. In this situation the multiplication with a diagonal matrix has to be replaced by the multiplication with the inverse of a mass matrix, and the resulting stiffness matrix is not as sparse. In both cases, we obtain a condition number which is bounded by Ch_l^{-2}. Compared to [28], the $\mathcal{V}(1,1)$-cycle of this multigrid variant is more robust. In contrast to the two variants discussed in [25, 28], this multigrid variant can also be used in combination with monotone techniques to solve nonlinear multibody contact problems [29]. The monotonicity and thus the convergence of the nonlinear algorithm cannot be guaranteed if we do not work with nested spaces and a positive definite formulation.

REFERENCES

[1] Y. Achdou and Y. Kuznetsov, *Substructuring preconditioners for finite element methods on nonmatching grids*, East-West J. Numer. Math., 3 (1995), pp. 1–28.

[2] Y. Achdou, Y. Kuznetsov, and O. Pironneau, *Substructuring preconditioners for the Q_1 mortar element method*, Numer. Math., 71 (1995), pp. 419–449.

[3] Y. Achdou, Y. Maday, and O. Widlund, *Iterative substructuring preconditioners for mortar element methods in two dimensions*, SIAM J. Numer. Anal., 36 (1999), pp. 551–580.

[4] P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuss, H. Rentz-Reichert, and C. Wieners, *UG—a flexible software toolbox for solving partial differential equations*, Comput. Vis. Sci., 1 (1997), pp. 27–40.

[5] F. Ben Belgacem, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[6] C. Bernardi, Y. Maday, and A. Patera, *Domain decomposition by the mortar element method*, in Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters, H. Kaper et al., eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 269–286.

[7] C. Bernardi, Y. Maday, and A. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, H. Brezis et al., eds., Paris, 1994, pp. 13–51.

[8] D. Braess and W. Dahmen, *Stability estimates of the mortar finite element method for 3-dimensional problems*, East-West J. Numer. Math., 6 (1998), pp. 249–263.

[9] D. Braess, W. Dahmen, and C. Wieners, *A multigrid algorithm for the mortar finite element method*, SIAM J. Numer. Anal., 37 (1999), pp. 48–69.

[10] D. Braess, P. Deuflhard, and K. Lipnikov, *A cascadic conjugate gradient method for domain decomposition with non-matching grids*, Computing, 69 (2002), pp. 205–225.

[11] D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the $\mathcal{V}$-cycle*, SIAM J. Numer. Anal., 20 (1983), pp. 967–975.

[12] D. Braess and R. Sarazin, *An efficient smoother for the Stokes problem*, Appl. Numer. Math., 23 (1997), pp. 3–19.

[13] J. Bramble, *Multigrid Methods*, Pitman Research Notes in Mathematics Series 294, Longman Scientific & Technical, Harlow, UK, 1993.

[14] S. Brenner, *Convergence of nonconforming $\mathcal{V}$-cycle and $\mathcal{F}$-cycle multigrid algorithms for second order elliptic boundary value problems*, Math. Comput. 73 (2004), pp. 1041–1066.

[15] S. Brenner, *Convergence of nonconforming multigrid methods without full elliptic regularity*, Math. Comput., 68 (1999), pp. 25–53.

[16] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts in Applied Mathematics, Springer-Verlag, New York, 1994.

[17] M. Casarin and O. Widlund, *A hierarchical preconditioner for the mortar finite element method*, Electron Trans. Numer. Anal. 4 (1996), pp. 75–88.

[18] J. Gopalakrishnan, *On the Mortar Finite Element Method*, Ph.D. thesis, Texas A&M University, College Station, TX, 1999.

[19] J. Gopalakrishnan and J. Pasciak, *Multigrid for the mortar finite element method*, SIAM

J. Numer. Anal., 37 (2000), pp. 1029–1052.
[20] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[21] R. Kornhuber, *Monotone multigrid methods for elliptic variational inequalities* I, Numer. Math., 69 (1994), pp. 167–184.
[22] R. Kornhuber and R. Krause, *Adaptive multigrid methods for Signorini's problem in linear elasticity*, Comput. Vis. Sci., 4 (2001), pp. 9–20.
[23] R. Krause, *Monotone Multigrid Methods for Signorini's Problem with Friction*, Ph.D. thesis, FU Berlin, Berlin, 2001.
[24] P. Oswald and B. Wohlmuth, *On polynomial reproduction of dual FE bases*, in Proceedings of the 13th International Conference on Domain Decomposition Methods, N. Debit, M. Garbey, R. Hoppe, J. Pèriaux, D. Keyes, and Y. Kuznetsov, eds., International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 2001, pp. 85–96.
[25] C. Wieners and B. Wohlmuth, *Duality estimates and multigrid analysis for saddle point problems arising from mortar discretizations*, SIAM J. Sci. Comput., 24 (2003), pp. 2163–2184.
[26] B. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Lecture Notes Comput. Sci. Eng. 17, Springer-Verlag, Berlin, 2001.
[27] B. Wohlmuth, *Comparison of dual lagrange multiplier spaces for mortar finite element discretizations*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 995–1012.
[28] B. Wohlmuth and R. Krause, *Multigrid methods based on the unconstrained product space arising from mortar finite element discretizations*, SIAM J. Numer. Anal., 39 (2001), pp. 192–213.
[29] B. Wohlmuth and R. Krause, *Monotone multigrid methods on nonmatching grids for nonlinear multibody contact problems*, SIAM J. Sci. Comput., 25 (2004), pp. 324–347.
[30] W. Zulehner, *A class of smoothers for saddle point problems*, Computing, 65 (2000), pp. 227–246.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2496–2521

EXPLICIT AND AVERAGING A POSTERIORI ERROR ESTIMATES FOR ADAPTIVE FINITE VOLUME METHODS*

C. CARSTENSEN†, R. LAZAROV‡, AND S. TOMOV§

Abstract. Local mesh-refining algorithms known from adaptive finite element methods are adopted for locally conservative and monotone finite volume discretizations of boundary value problems for steady-state convection-diffusion-reaction equations. The paper establishes residual-type explicit error estimators and averaging techniques for a posteriori finite volume error control with and without upwind in global H^1- and L^2-norms. Reliability and efficiency are verified theoretically and confirmed empirically with experimental support for the superiority of the suggested adaptive mesh-refining algorithms over uniform mesh refining. A discussion of adaptive computations in the simulation of contaminant concentration in a nonhomogeneous water reservoir concludes the paper.

Key words. convection-diffusion-reaction equations, 3-D problems, finite volume approximation, a posteriori error estimators, residual estimators, averaging estimators, ZZ refinement indicator

AMS subject classifications. 65N30, 65M35

DOI. 10.1137/S0036142903425422

1. Introduction.

We consider the following convection-diffusion-reaction problem: Find $u = u(x)$ such that

$$\begin{cases} Lu \equiv \nabla \cdot (-A\nabla u + \underline{b}u) + \gamma u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ (-A\nabla u + \underline{b}u) \cdot \underline{n} = g & \text{on } \Gamma_N^{\text{in}}, \\ -(A\nabla u) \cdot \underline{n} = 0 & \text{on } \Gamma_N^{\text{out}}. \end{cases} \tag{1.1}$$

Here Ω is a bounded polygonal domain in R^d, $d = 2, 3$; $A = A(x)$ is $d \times d$ symmetric, bounded, and uniformly positive definite matrix in Ω; $\underline{b}$ is a given vector function; $\underline{n}$ is the unit outer vector normal to $\partial\Omega$; and f is a given source function. We have also used the notation ∇u for the gradient of a scalar function u and $\nabla \cdot \underline{b}$ for the divergence of a vector function $\underline{b}$ in R^d. The boundary of Ω, $\partial\Omega$ is split into Dirichlet, Γ_D and Neumann, Γ_N parts. Further, the Neumann boundary is divided into two parts: $\Gamma_N = \Gamma_N^{\text{in}} \cup \Gamma_N^{\text{out}}$, where $\Gamma_N^{\text{in}} = \{x \in \Gamma_N : \underline{n}(x) \cdot \underline{b}(x) < 0\}$ and $\Gamma_N^{\text{out}} = \{x \in \Gamma_N : \underline{n}(x) \cdot \underline{b}(x) \geq 0\}$. We assume that Γ_D has positive surface measure.

This problem is a prototype for flow and transport in porous media. For example, $u(x)$ can represent the pressure head in an aquifer or the concentration of a chemical that is dissolved and distributed in groundwater due to the processes of diffusion, dispersion, and absorption. In many cases $A = \epsilon I$, where I is the identity matrix in

*Received by the editors April 1, 2003; accepted for publication (in revised form) May 27, 2004; published electronically March 31, 2005. This work was partially supported by NSF grant DMS–9973328. It was finalized while the first author was a guest at the Isaac Newton Institute for Mathematical Science, Cambridge, UK.

http://www.siam.org/journals/sinum/42-6/42542.html

†Institute of Mathematics, Humboldt University of Berlin, S172: Rudower Chausee 25, Unter den Linden 6, 1099 Berlin, Germany (cc@math.hu-berlin.de).

‡Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (lazarov@math.tamu.edu).

§Information Technology Division, Brookhaven National Laboratory, Bldg. 515, Upton, NY 11973 (tomov@bnl.gov).

R^d and $\epsilon > 0$ is a small parameter. This corresponds to the important and difficult class of singularly perturbed convection-diffusion problems (see, e.g., the monograph of Ross, Stynes, and Tobiska [31]). In our computations we have used our approach for grid adaptation for this type of problem as well. However, we do not claim that the developed theory in this paper covers this important practical case. Further, $u(x)$ can be viewed as a limit for $t = \infty$ of the solution $u = u(x,t)$ of the corresponding time-dependent problem

$$u_t + Lu = f,\ t > 0,\ x \in \Omega \tag{1.2}$$

with boundary conditions as above and an initial condition $u(x,0) = u_0(x)$, where u_0 is a given function in Ω. Various generalizations, mostly considering nonlinear terms, are possible and widely used in the applications. For example, γu is replaced by a nonlinear reaction term $\gamma(u)$, or the linear convective term $\underline{b}u$ is replaced by a nonlinear flux $\underline{b}(u)$. In this work we follow the framework of the model problem (1.1) and focus on its 3-D setting.

The development of efficient solution methods featuring error control is important for various applications. Our study has been motivated by the research in groundwater modeling and petroleum reservoir simulations (see, e.g., [19]). The solutions of problems in that area exhibit steep gradients and rapid changes due to localized boundary data, discontinuities in the coefficients of the differential equation, and/or other local phenomena (for example, extraction/injection wells, faults, etc.). In order to accurately resolve such local behavior, the numerical method should be able to detect the regions in which the solution changes significantly and to refine the grid locally in a balanced manner so that the overall accuracy is uniform in the whole domain.

Equation (1.1) expresses conservation of the properly scaled quantity u over any subdomain contained in Ω. In the context of groundwater, fluid flow $u(x)$ is in general either the water mass or the mass of the chemical dissolved in the water. Numerical methods that have this property over a number of nonoverlapping subdomains that cover the whole domain are called locally conservative. Finite volumes (control volumes, box schemes), mixed finite elements, and discontinuous Galerkin methods have this highly desirable property. The simplicity of the finite volume approximations combined with their local conservation property and flexibility motivated our study.

There are few works related to a posteriori error estimates for finite volume methods. In [2] Angermann studied a balanced a posteriori error estimate for finite volume discretizations for convection-diffusion equations in two dimensions on Voronoi meshes. The derivation of the error estimator is based on the idea of his previous work [3] on the finite element method. The estimator for the finite volume method contains two new terms which have been studied previously. Some extensions to Angermann's work related to more general situations in respect to space dimension and type of control volumes can be found in Thiele's dissertation [35]. Again, the ideas from the finite element method were exploited in deriving an upper error estimate for the space discretization of parabolic problems. In our paper we use a similar approach; namely, the error estimates for the finite volume method are derived by using the relation between the finite volume and finite element methods (see, e.g., [8]). We note that, despite recent progress (see, e.g., the monographs [23, 26]), the theory of finite volume methods is still under development. This in turn raises certain difficulties in establishing an independent and sharp a posteriori error analysis for the finite volume approximations.

A posteriori error indicators and estimators for the finite element method have been used and studied in the past two decades. Since the pioneering paper of Babuška and Rheinboldt [6], the research in this field has expanded in various directions that include the *residual-based* method (see the survey paper of Verfürth [36]), *hierarchical-based* error estimators [9], estimators based on postprocessing of the approximate solution gradient [37, 38], error estimators that control the error or its gradient in the maximum norm, etc. One popular approach is to evaluate certain local residuals and obtain the a posteriori error indicator by solving local Dirichlet or Neumann problems by taking the local residuals as data [6, 9]. Another variation of the method that controls the global L^2- and H^1-norms of the error uses the Galerkin orthogonality, a priori interpolation estimates, and global stability (see, for example, [21]). Furthermore, solving appropriate dual problems, instead of using the a priori interpolation estimates, leads to error estimators controlling various kinds of error functionals [11]. Solving finite element problems in a space enriched by hierarchical bases functions gives rise to *hierarchical-based* error estimators [9]. There are error estimators based on optimal a priori estimates in a maximum norm [22]. Another type of error estimator/indicator, widely (and in most cases heuristically) used in many adaptive finite element codes, is based on postprocessing (averaging) of the approximate solution gradient (see [37, 38]). In the context of the finite element method for elliptic partial differential equations, averaging or recovery techniques are justified in [10, 14, 30]. Finally, for an extensive study of the efficiency and the reliability of the local estimators and indicators for finite element approximations, we refer to the recent monograph of Babuska and Strouboulis [7].

In this paper we adapt the finite element local error estimation techniques to the case of finite volume approximations. We consider mainly the *residual-based* a posteriori error estimators and analyze the one that uses Galerkin orthogonality, a priori interpolation estimates, and global stability in L^2- and H^1-norms. Our theoretical and experimental findings are similar to those in [2] and could be summarized as follows. The a posteriori error estimates in the finite volume element method are quite close to those in the finite element method, and the mathematical tools from finite element theory can be successfully applied for their analysis. Our computational experiments with various model problems confirm this conclusion. For more computational examples we refer to [25].

The paper is organized as follows. We start with the finite volume element formulation in section 2. The section defines the used notation and approximations and gives some general results from the finite volume approximations. Section 3 studies the *residual-based* error estimator, followed by a short description of the used adaptive refinement strategy (in section 4). Finally, in section 5, we present numerous computational results for 2-D and 3-D test problems which illustrate the adaptive strategy and support our theoretical findings.

2. Finite volume element approximation. Subsection 2.1 introduces the notation used in the paper. In subsection 2.2 we define the finite volume element approximations and give an a priori estimate for the error.

2.1. Notation. We denote by $L^2(K)$ the square-integrable real-valued functions over $K \subset \Omega$, by $(\cdot,\cdot)_{L^2(K)}$ the inner product in $L^2(K)$, and by $|\cdot|_{H^1(K)}$ and $||\cdot||_{H^1(K)}$, respectively, the seminorm and norm of the Sobolev space $H^1(K)$, namely,

$$||u||_{L^2(K)} := (u,u)^{1/2}_{L^2(K)}, \qquad |u|_{H^1(K)} := (\nabla u, \nabla u)^{1/2}_{L^2(K)},$$
$$||u||^2_{H^1(K)} := ||u||^2_{L^2(K)} + |u|^2_{H^1(K)}.$$

In addition, if $K = \Omega$, we suppress the index K and also write $(\cdot,\cdot)_{L^2(\Omega)} := (\cdot,\cdot)$ and $||\cdot||_{L^2} := ||\cdot||$. Further, we use the Hilbert space $H^1_D(\Omega) = \{v \in H^1(\Omega) :\ v|_{\Gamma_D} = 0\}$. Finally, we denote by $H^{1/2}(\partial K)$ the space of the traces of functions in $H^1(K)$ on the boundary ∂K.

To avoid writing unknown constants we use the notation $a \lesssim b$ instead of the inequality $a \le Cb$, where the constant C is independent of the mesh size h.

In our analysis we shall use the following simple inequality valid for $\Omega \subset R^d$, $d > 1$, with Lipschitz continuous boundary $\partial\Omega$ (called Ilin's inequality; cf., e.g., [28]): Let Ω_δ be a strip along $\partial\Omega$ of width δ. Then

$$
\begin{aligned}
&||u||_{L^2(\Omega_\delta)} \lesssim \delta^{1/2}||u||_{H^1(\Omega)} \quad \text{for all } u \in H^1(\Omega);\\
&||u||_{L^2(\Omega_\delta)} \lesssim \delta^{s}||u||_{H^s(\Omega)} \quad \text{for all } u \in H^s(\Omega),\ 0 < s < 1/2.
\end{aligned}
\tag{2.1}
$$

The first inequality is trivial in the case where Ω is a half-space and u has a compact support. The proof in the general case will follow easily by using partition of unity and transforming each subdomain into half-space. The second inequality is obtained using the fact that $||u||_{L^2(\Omega_\delta)} \lesssim \delta||u||_{H^1(\Omega)}$ for all $u \in H^1_0(\Omega)$ and interpolation of Banach spaces (cf., e.g., [1]).

Next, we introduce the bilinear form $a(\cdot,\cdot)$ defined on $H^1_D(\Omega) \times H^1_D(\Omega)$ as

$$
a(u,v) := (A\nabla u - \underline{b}u, \nabla v) + (\gamma u, v) + \int_{\Gamma_N^{\text{out}}} \underline{b}\cdot\underline{n}\ u\ v\, ds. \tag{2.2}
$$

We assume that the coefficients of problem (1.1) are such that

(a) the form is $H^1_D(\Omega)$-elliptic (coercive); i.e., there is a constant $c_0 > 0$ such that

$$
c_0||u||_{H^1} \le a(u,u) \quad \text{for all } u \in H^1_D(\Omega); \tag{2.3}
$$

(b) the form is bounded (continuous) on $H^1_D(\Omega)$; i.e., there is a constant $c_1 > 0$ such that

$$
a(u,v) \le c_1||u||_{H^1}||v||_{H^1} \quad \text{for all } u,\ v \in H^1_D(\Omega). \tag{2.4}
$$

The above two conditions guarantee that the expression $a(u,u)$ is equivalent to the norm in $H^1_D(\Omega)$. Further, we shall use the notation $||u||_a^2 = a(u,u)$ and call this expression the "energy" norm.

A sufficient condition for the coercivity of the bilinear form is $\gamma(x)+0.5\nabla\cdot\underline{b}(x) \ge 0$ for all $x \in \Omega$, while a sufficient condition for the continuity is boundedness of the coefficients $A(x)$, $\underline{b}(x)$, and $\gamma(x)$ in Ω. Further in the paper we assume that these conditions are satisfied. Then (1.1) has the following weak form: Find $u \in H^1_D(\Omega)$ such that

$$
a(u,v) = F(v) := (f,v) - \int_{\Gamma_N^{\text{in}}} gv\, ds \quad \text{for all } v \in H^1_D(\Omega). \tag{2.5}
$$

2.2. Approximation method. The domain Ω is partitioned into triangular (for the 2-D case) or tetrahedral (for the 3-D case) finite elements denoted by K. The

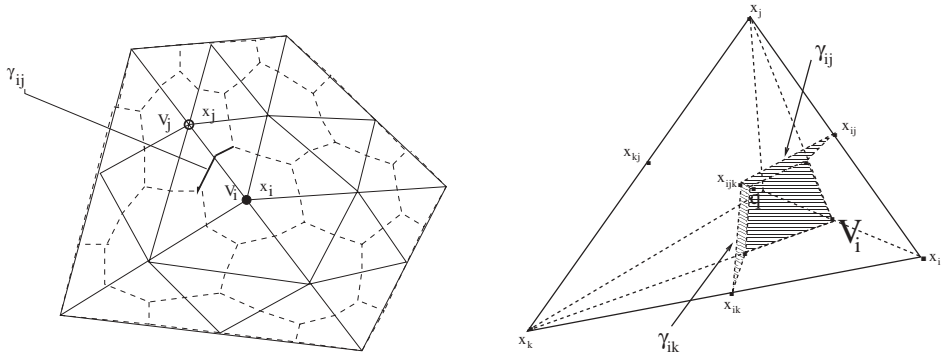


FIG. 1. *Left: Finite element and finite volume partitions in two dimensions. Right: Contribution from one element to control volume V_i, γ_{ij}, and γ_{ik} in three dimensions; point q is the element's medicenter. Internal points for the faces are the medicenters of the faces.*

elements are considered to be closed sets and the splitting, often called triangulation of Ω, is denoted by $\mathcal{T}$. We assume that the mesh is aligned with the discontinuities of the coefficients of the differential equation (if any), with the data f and g, and with the interfaces between Γ_D, Γ_N^{out}, and Γ_N^{in}.

We note that our analysis will be valid also for domains with smooth boundaries. In this case we have to modify the triangulation so that the methods do not lose accuracy due to approximation of the domain. Such schemes have been discussed in [18].

We introduce the set $N_h = \{x_i : x_i \text{ is a vertex of element } K \in \mathcal{T}\}$ and denote by N_h^0 the set of all vertices in N_h except those on Γ_D. For a given vertex x_i we denote by $\Pi(i)$ the index set of all neighbors of x_i in N_h, i.e., all vertices that are connected to x_i by an edge.

For a given finite element triangulation $\mathcal{T}$, we construct a dual mesh $\mathcal{T}^*$ (based upon $\mathcal{T}$), whose elements are called control volumes (boxes, finite volumes, etc.). There are various ways to introduce the control volumes. Almost all approaches can be described in the following general scheme. In each element $K \in \mathcal{T}$ a point q is selected. For the 3-D case, on each of the four faces $\overline{x_i x_j x_k}$ of K a point x_{ijk} is selected and on each of the six edges $\overline{x_i x_j}$ a point x_{ij} is selected. Then q is connected to the points x_{ijk}, and in the corresponding faces, the points x_{ijk}, are connected to the points x_{ij} by straight lines (see Figure 1). Control volume associated with a vertex x_i is denoted by V_i and defined as the union of the "quarter" elements $K \in \mathcal{T}$, which have x_i as a vertex (see Figure 1). The interface between two control volumes, V_i and V_j, is denoted by γ_{ij}, i.e., $\overline{V}_i \cap \overline{V}_j = \gamma_{ij}$.

We assume that $\mathcal{T}$ is locally quasi uniform, that is, for $K \in \mathcal{T}$, $|K| \lesssim \rho(K)^d$, where $\rho(K)$ is the radius of the largest ball contained in K and $|K|$ denotes the area or volume of K. In the context of locally refined grids, this means that the smallest interior angle is bounded away from zero and any two neighboring finite elements are of approximately the same size, whereas elements that are far away may have quite different sizes.

In our 3-D computations q is the center of gravity of the element K, x_{ijk} are the centers of gravity of the corresponding faces, and x_{ij} are the mid-points (centers of gravity) of the corresponding edges (as shown on Figure 1).

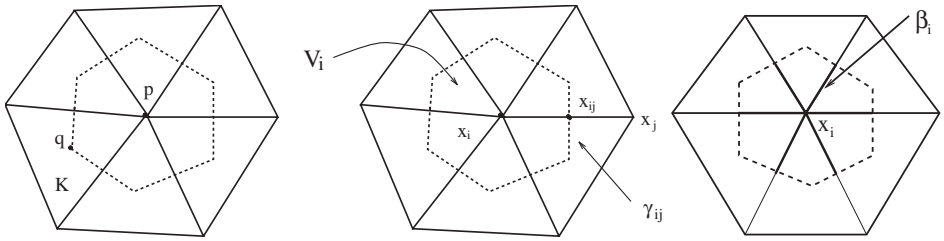


FIG. 2. *Control volumes with circumcenters as internal points (Voronoi meshes) and interface γ_{ij} of V_i and V_j. The rightmost picture shows the segments β_i in bold.*

In two dimensions, another possibility is to choose q to be the center of the circumscribed circle of K. These types of control volumes form Voronoi or perpendicular bisector (PEBI) meshes (see, e.g. [23, pp. 764, 825]). Then obviously, γ_{ij} are the PEBIs of the three edges of K (see Figure 2). This construction requires that all finite elements are triangles of acute type, which we shall assume whenever such triangulation is used.

We define the linear finite element space S_h as

$$S_h = \{v \in C(\Omega) : v|_K \text{ is affine for all } K \in \mathcal{T} \text{ and } v|_{\Gamma_D} = 0\}$$

and its dual volume element space S_h^* by

$$S_h^* = \{v \in L^2(\Omega) : v|_V \text{ is constant for all } V \in \mathcal{T}^* \text{ and } v|_{\Gamma_D} = 0\}.$$

Obviously, $S_h = \text{span}\{\phi_i : x_i \in N_h^0\}$ and $S_h^* = \text{span}\{\chi_i : x_i \in N_h^0\}$, where ϕ_i denotes the standard nodal linear basis function associated with the node x_i, and χ_i denotes the characteristic function of the volume V_i. Let $I_h : C(\Omega) \cap H_D^1(\Omega) \to S_h$ be the interpolation operator and $I_h^* : C(\Omega) \cap H_D^1(\Omega) \to S_h^*$ and $P_h^* : C(\Omega) \cap H_D^1(\Omega) \to S_h^*$ be the piecewise constant interpolation and projection operators:

$$I_h u = \sum_{x_i \in N_h} u(x_i)\phi_i(x),\ I_h^* u = \sum_{x_i \in N_h} u(x_i)\chi_i(x),\ \text{and } P_h^* u = \sum_{x_i \in N_h} \bar{u}_i \chi_i(x).$$

Here $\bar{u}_i$ is the averaged value of u over the volume V_i for $x_i \in N_h^0$, i.e., $\bar{u}_i = \int_{V_i} u\,dx/|V_i|$, and $\bar{u}_i = 0$ for $x_i \in \Gamma_D$. In fact, I_h also makes sense as an interpolation operator from S_h^* to S_h. Namely, if $v^* \in S_h^*$, then $I_h v^* \in S_h$ and $I_h v^*(x_i) = v^*(x_i)$.

Further, for $v^* \in S_h^*$, we use the notation $v_i^* = v^*(x_i)$. We also define the "total flux" and its approximation by

$$\underline{\sigma} := -A\nabla u + \underline{b}u, \qquad \underline{\sigma}_h := -A\nabla_h u_h + \underline{b}u_h$$

and assume that the coefficients $A(x)$ and $\underline{b}(x)$ are elementwise smooth. Also, we denote by $\nabla_h \cdot$ the $\mathcal{T}$-piecewise divergence and by ∇_h the $\mathcal{T}$-piecewise gradient. Integrals involving piecewise quantities are considered as sums over the pieces where the quantities are defined.

The finite volume element approximation u_h of (1.1) is the solution to the following problem: Find $u_h \in S_h$ such that

$$a_h(u_h, v_h^*) := A(u_h, v_h^*) + C(u_h, v_h^*) = F(v_h^*) \quad \text{for all } v_h^* \in S_h^*. \tag{2.6}$$

Here the bilinear forms $A(u_h, v^*)$ and $C(u_h, v^*)$ are defined on $S_h \times S_h^*$ and the linear form $F(v^*)$ is defined on S_h^*. They are given by

$$A(u_h, v^*) = \sum_{x_i \in N_h^0} v_i^* \left(- \int_{\partial V_i \setminus \Gamma_N} (A \nabla_h u_h) \cdot \underline{n} ds + \int_{V_i} \gamma u_h dx \right), \tag{2.7}$$

$$C(u_h, v^*) = \sum_{x_i \in N_h^0} v_i^* \int_{\partial V_i \setminus \Gamma_N^{\text{in}}} (\underline{b} \cdot \underline{n})\ u_h \, ds, \tag{2.8}$$

$$F(v^*) = \sum_{x_i \in N_h^0} v_i^* \left\{ \int_{V_i} f dx - \int_{\partial V_i \cap \Gamma_N^{\text{in}}} g \, ds \right\}. \tag{2.9}$$

Obviously, $\nabla \cdot \underline{\sigma}_h$ is well defined over $V_i \cap K$ for all $V_i \in \mathcal{T}^*$ and $K \in \mathcal{T}$. This ensures, in particular, that the surface integrals in (2.7) and (2.8) exist.

In addition to $C(u_h, v^*)$ we introduce the form $C^{\text{up}}(u_h, v^*)$ that uses upwind approximation. Approximation (2.7)–(2.9) can be used for moderate convection fields and dominating diffusion. For small diffusion, for example, when $A = \epsilon I$ with ϵ small, approximation (2.7)–(2.9) gives oscillating numerical results, which we would like to avoid. We are interested in approximation methods that produce solutions satisfying the maximum principle and are locally conservative. Such schemes are also known as monotone schemes (see, e.g., [24, 31]). A well-known sufficient condition for a scheme to be monotone is that the corresponding stiffness matrix be an M-matrix (see [33, pp. 182, 260] and [31, p. 202]).

The upwind approximation that we use for problems with large convection (or small diffusion) is locally mass conservative and gives the desired stabilization. We split the integral over ∂V_i on integrals over $\gamma_{ij} = \partial V_i \cap \partial V_j$ (see Figure 1) and introduce out-flow and in-flow parts of the boundary of the volume V_i. This splitting can be characterized by the quantities $(\underline{b} \cdot \underline{n})_+ = \max(0,\ \underline{b} \cdot \underline{n})$ and $(\underline{b} \cdot \underline{n})_- = \min(0,\ \underline{b} \cdot \underline{n})$, where $\underline{n}$ is the outer unit vector normal to ∂V_i. Then we introduce

$$C^{\text{up}}(u_h, v^*) = \sum_{x_i \in N_h^0} v_i^* \Bigg\{ \sum_{j \in \Pi(i)} \int_{\gamma_{ij}} \left((\underline{b} \cdot \underline{n})_+ u_h(x_i) + (\underline{b} \cdot \underline{n})_- u_h(x_j) \right) ds + \int_{\Gamma_N^{\text{out}} \cap \partial V_i} (\underline{b} \cdot \underline{n})\, u_h(x_i)\ ds \Bigg\}. \tag{2.10}$$

This approximation is well defined for any $\underline{b}$. In order to avoid technicalities in our analysis we assume that the vector field $\underline{b}$ is piecewise smooth and has small variation over each finite element. Thus, the quantity $\underline{b} \cdot \underline{n}$ does not change sign over γ_{ij}.

The upwind finite volume element approximation u_h of (1.1) becomes the following: Find $u_h \in S_h$ such that

$$a_h^{\text{up}}(u_h, v^*) := A(u_h, v^*) + C^{\text{up}}(u_h, v^*) = F(v^*) \text{ for all } v^* \in S_h^*. \tag{2.11}$$

This is an extension of the classical upwind approximation of the convection term and is closely related to the discontinuous Galerkin approximation (see, e.g., [22]) or to the Tabata scheme for the Galerkin finite element method [34]. It is also related to the scheme on Voronoi meshes derived by Mishev [27]. A different type of weighted upwind approximation on Voronoi meshes in two dimensions has been studied by Angermann [2].

3. A posteriori error analysis. This section is devoted to the mathematical derivation of computable error bounds in the energy norm. Throughout this section, $u \in H^1_D$ denotes the exact solution of (2.5) and $u_h \in S_h$ denotes the discrete solution of either (2.6) or (2.11). Then, $e := u - u_h \in H^1_D(\Omega)$ is the (unknown) error and $\overline{e} := P^*_h e \in S^*_h$ is its $\mathcal{T}^*$-piecewise integral mean. We denote by $\mathcal{E}$ the set of all interior edges/faces in $\mathcal{T}$, respectively, in two/three dimensions. Also, for a vertex $x_i \in N^0_h$ let $\beta_i := V_i \cap \mathcal{E}$ (see Figure 2). For any $E \in \mathcal{E}$ let $[\underline{\sigma}_h] \cdot \underline{n}$ denote the jump of $\underline{\sigma}_h$ across E in normal to E direction $\underline{n}$. The orientation of $\underline{n}$ is not important as long as the jump is in the same direction. In general, if $\underline{n}$ is present in a boundary integral, it will denote the outward unit vector normal to the boundary. With every element $K \in \mathcal{T}$, edge/face $E \in \mathcal{E}$, and volume $V_i \in \mathcal{T}^*$ we associate local mesh size denoted correspondingly by h_K, h_E, and h_i. Since the mesh is locally quasi uniform the introduced mesh sizes are locally equivalent, i.e., bound each other from above and below with constants independent of the mesh size. Then, we introduce a global discontinuous mesh size function $h(x)$, $x \in \Omega$, that assumes value h_K, h_E, and h_i depending on $x \in K \setminus \partial K$, $x \in E$, or $x = x_i$, respectively. Finally, we use the following shorthand notation for integration over all faces E in $\mathcal{E}$:

$$\int_{\mathcal{E}} v ds := \sum_{E \in \mathcal{E}} \int_E v ds, \qquad \|v\|_{L^2(\mathcal{E})} := \sum_{E \in \mathcal{E}} \int_E v^2 ds.$$

3.1. Energy-norm a posteriori error estimate of the scheme without upwind. We consider problem (2.6) and begin our analysis with the case when the form $C(\cdot,\cdot)$ is evaluated by (2.8). We first give a representation of the error and introduce some locally computable quantities. In Theorem 3.1 we show that these quantities give a reliable estimate for the error. Further, we introduce the error estimator, based on local "averaging" of the "total flux" $\underline{\sigma}$ over the control volumes, and show that this estimator is reliable up to higher order terms.

The following lemma gives a representation of the error.

LEMMA 3.1. *Assume that the bilinear form $a(\cdot,\cdot)$ satisfies* (2.3) *and* (2.4). *Then for the error $e = u - u_h$, where u is the solution of* (2.5) *and u_h is the solution of* (2.6), *we have*

$$\begin{aligned} \|e\|_a^2 &= (f - \nabla_h \cdot \underline{\sigma}_h - \gamma u_h,\ e - \overline{e}) - \int_{\mathcal{E}} [\underline{\sigma}_h] \cdot \underline{n}\, (e - \overline{e})\, ds \\ &\quad - \int_{\Gamma_N^{\mathrm{in}}} (g - \underline{\sigma}_h \cdot \underline{n})\, (e - \overline{e})\, ds - \int_{\Gamma_N^{\mathrm{out}}} (A \nabla_h u_h) \cdot \underline{n}\, (e - \overline{e})\, ds. \end{aligned} \tag{3.1}$$

Proof. We take $v = e \in H^1_D(\Omega)$ in (2.5) and use the definition of $a(\cdot,\cdot)$ by (2.2) to get

$$\begin{aligned} a(e,e) &= a(u,e) - a(u_h,e) \\ &= (f - \gamma u_h, e) + (\underline{\sigma}_h, \nabla e) - \int_{\Gamma_N^{\mathrm{in}}} g e\, ds - \int_{\Gamma_N^{\mathrm{out}}} (\underline{b} \cdot \underline{n})\, u_h\, e\, ds. \end{aligned}$$

We integrate the second term on the right-hand side by parts on each element $K \in \mathcal{T}$:

$$\int_K \underline{\sigma}_h \cdot \nabla e\, ds = \int_{\partial K} (\underline{\sigma}_h \cdot \underline{n}) e\, ds - \int_K e \nabla \cdot \underline{\sigma}_h\, dx.$$

The sum over all elements yields the jump contributions $[\underline{\sigma}_h]\cdot\underline{n}$ along $\mathcal{E}$ and eventually proves

$$
\begin{aligned}
a(e,e) &= (f-\nabla_h\cdot\underline{\sigma}_h-\gamma u_h,e)-\int_{\mathcal{E}}[\underline{\sigma}_h]\cdot\underline{n}\,e\,ds\\
&\quad-\int_{\Gamma_N^{\mathrm{in}}}(g-\underline{\sigma}_h\cdot\underline{n})e\,ds-\int_{\Gamma_N^{\mathrm{out}}}(A\nabla_h u_h)\cdot\underline{n}\,e\,ds.
\end{aligned}
\tag{3.2}
$$

It remains to be shown that the preceding right-hand side vanishes if e is replaced by $\overline{e}$. For each control volume V_i we have from (2.6)–(2.8) that

$$
\int_{\partial V_i\setminus\Gamma_N}\underline{\sigma}_h\cdot\underline{n}\,ds=\int_{V_i}(f-\gamma u_h)\,dx-\int_{\partial V_i\cap\Gamma_N^{\mathrm{out}}}(\underline{b}\cdot\underline{n})\,u_h\,ds-\int_{\partial V_i\cap\Gamma_N^{\mathrm{in}}}g\,ds.
$$

The Gauss divergence theorem is applied to each nonvoid $K\cap V_i$, $K\in\mathcal{T}$, so that the left-hand side of the above inequality becomes

$$
\int_{\partial V_i\setminus\Gamma_N}\underline{\sigma}_h\cdot\underline{n}\,ds=\int_{V_i}\nabla_h\cdot\underline{\sigma}_h\,dx+\int_{\beta_i}[\underline{\sigma}_h]\cdot\underline{n}\,ds-\int_{\partial V_i\cap\Gamma_N}\underline{\sigma}_h\cdot\underline{n}\,ds.
$$

The difference of the preceding two identities is multiplied by $\overline{e}(x_i)$ and summed over all control volumes. This results in

$$
0=(f-\nabla_h\cdot\underline{\sigma}_h-\gamma u_h,\overline{e})-\int_{\mathcal{E}}[\underline{\sigma}_h]\cdot\underline{n}\,\overline{e}\,ds-\int_{\Gamma_N^{\mathrm{in}}}(g-\underline{\sigma}_h\cdot\underline{n})\,\overline{e}\,ds-\int_{\Gamma_N^{\mathrm{out}}}A\nabla_h u_h\cdot\underline{n}\,\overline{e}\,ds.
$$

Subtracting this identity from (3.2) concludes the proof of (3.1). ∎

Motivated by the above considerations we introduce the following locally computable quantities that play a major role in the design of adaptive algorithms and their a posteriori error analysis.

DEFINITION 3.1. *Set*

$$
\begin{aligned}
R_K(x)&:=(f-\nabla\cdot\underline{\sigma}_h-\gamma u_h)(x),\ x\in K,\\
R_E(x)&:=([\underline{\sigma}_h]\cdot\underline{n})(x),\ x\in E,\ \textit{for } E\cap\Gamma_N=\emptyset,\\
R_E^{\mathrm{in}}(x)&:=(g-\underline{\sigma}_h\cdot\underline{n})(x),\ x\in E,\ \textit{for } E\subset\Gamma_N^{\mathrm{in}},\\
R_E^{\mathrm{out}}(x)&:=(A\nabla u_h\cdot\underline{n})(x),\ x\in E,\ \textit{for } E\subset\Gamma_N^{\mathrm{out}}
\end{aligned}
$$

and define

$$
\begin{aligned}
\eta_R&:=\|h\,R_K\|_{L^2(\Omega)},\ \eta_E:=\|h^{1/2}\,R_E\|_{L^2(\mathcal{E})},\\
\eta_N&:=\|h^{1/2}\,R_E^{\mathrm{in}}\|_{L^2(\Gamma_N^{in})}+\|h^{1/2}\,R_E^{\mathrm{out}}\|_{L^2(\Gamma_N^{\mathrm{out}})}.
\end{aligned}
$$

LEMMA 3.2. *Suppose that $R_E\in L^2(\mathcal{E})$ and that the partitioning $\mathcal{T}$ of Ω is locally quasi uniform. Then*

$$
\int_{\mathcal{E}}[\underline{\sigma}_h]\cdot\underline{n}(e-\overline{e})\,ds\lesssim\eta_E\,\|\nabla e\|\ \textit{for any } e\in H^1_D(\Omega),
$$

where the constant in the notation $\lesssim$ depends only on the shape of the elements in $\mathcal{T}$ and the volumes in $\mathcal{T}^*$.

Proof. A well-established trace inequality (cf., e.g., [12, Theorem 1.6.6] or [15, Theorem 1.4]) and scaling argument lead to

$$h_E^{1/2}\|v\|_{L^2(E)} \lesssim \|v\|_{L^2(K)} + h_E\|\nabla v\|_{L^2(K)} \tag{3.3}$$

for all $v \in H^1(K)$ and edges E of an element $K \in \mathcal{T}$. An application to $v := e - \overline{e}$ on each $K \cap V_i$, where $K \in \mathcal{T}$ and $x_i \in N_h$, leads to

$$\begin{aligned}\int_{\beta_i} [\underline{\sigma}_h]\cdot\underline{n}(e-\overline{e})\,ds &\le \|[\underline{\sigma}_h]\cdot\underline{n}\|_{L^2(\beta_i)}\|e-\overline{e}\|_{L^2(\beta_i)}\\ &\lesssim h_i^{1/2}\|[\underline{\sigma}_h]\cdot\underline{n}\|_{L^2(\beta_i)}(h_i^{-1}\|e-\overline{e}\|_{L^2(V_i)} + \|\nabla e\|_{L^2(V_i)}).\end{aligned}$$

Further, Poincaré's inequality for $x_i \in N_h^0$ (in which case $\int_{V_i}(e-\overline{e})\,dx = 0$) or Friedrichs's inequality for $x_i \in N_h\backslash N_h^0$ (in which case $\overline{e} = 0$ on V_i and $e = 0$ on $\partial V_i \cap \Gamma_D$) shows that

$$h_i^{-1}\|e-\overline{e}\|_{L^2(V_i)} \lesssim \|\nabla e\|_{L^2(V_i)}. \tag{3.4}$$

Poincaré's, respectively, Friedrichs's, inequality is valid in this case because the volumes V_i are star shaped w.r.t. a ball of radius $\sim h_i$, which follows from the quasi uniformity of $\mathcal{T}$ and our choice of $\mathcal{T}^*$. Substituting the last result into the preceding inequality yields

$$\int_{\beta_i} [\underline{\sigma}_h]\cdot\underline{n}(e-\overline{e})\,ds \lesssim \|h^{1/2}\ [\underline{\sigma}_h]\cdot\underline{n}\|_{L^2(\beta_i)}\|\nabla e\|_{L^2(V_i)}$$

for all $x_i \in N_h$. A summation over all vertices yields the assertion. ∎

Below we establish that the sum of the quantities η_R, η_E, and η_N gives a reliable estimate for the error in the global energy norm.

THEOREM 3.1. *Assume that the coefficients of the bilinear form $a(\cdot,\cdot)$ are such that* (2.3) *and* (2.4) *are satisfied, and that the partitioning $\mathcal{T}$ of Ω is locally quasi uniform. Then*

$$\|e\|_a \lesssim \eta_R + \eta_E + \eta_N.$$

The constant in this inequality depends on the constants c_0 in (2.3) *and c_1 in* (2.4)*, and on the shape of the elements in $\mathcal{T}$ and $\mathcal{T}^*$, but is independent of h.*

Proof. The identity (3.1) of Lemma 3.1 represents $\|e\|_a^2$ as a sum of four terms. We bound the first term using Cauchy's inequality, the second one using Lemma 3.2, and the remaining two terms using again Cauchy's inequality:

$$\|e\|_a^2 \lesssim \eta_R\|h^{-1}(e-\overline{e})\| + \eta_E\|\nabla e\| + \eta_N\|h^{-1/2}(e-\overline{e})\|_{L^2(\Gamma_N)}.$$

Inequality (3.4) is combined with the trace inequality (3.3) to obtain

$$\|h^{-1/2}(e-\overline{e})\|_{L^2(\Gamma_N)}^2 + \|h^{-1}(e-\overline{e})\|^2 \lesssim \sum_{x_i\in N_h}(h_i^{-2}\|e-\overline{e}\|_{L^2(V_i)}^2 + \|\nabla e\|_{L^2(V_i)}^2) \lesssim \|\nabla e\|^2.$$

Condition (2.3) yields $\|\nabla e\| \lesssim \|e\|_a$ and this concludes the proof of the theorem. ∎

Now we introduce an error estimator that is based on local averaging (post-processing) of the "total flux" $\underline{\sigma}_h$. For finite element approximations this estimator,

often called the ZZ-estimator, has been justified by Carstensen and Bartels [10, 14] and Rodriguez [30].

DEFINITION 3.2. *Let P_i be the L^2-projection onto the affine functions on V_i. We define the error indicator η_Z for $A(x)$ and $\underline{b}(x)$ smooth over the volumes $V_i \in \mathcal{T}^*$ as*

$$\eta_Z := \left(\sum_{x_i \in N_h} \|\underline{\sigma}_h - P_i \underline{\sigma}_h\|^2_{L^2(V_i)} \right)^{1/2}.$$

Remark 3.1. In our numerical experiments we have allowed $A(x)$ to have jumps that are aligned with the partition $\mathcal{T}$. In such cases we have changed the projection P_i. For example, if $V_i = V_i^1 \cup V_i^2$ and $A(x)$ is smooth on V_i^1 and V_i^2 but has jumps across their interface, then P_i is defined in a piecewise way as

$$\|\underline{\sigma}_h - P_i \underline{\sigma}_h\|^2_{L^2(V_i)} = \|\underline{\sigma}_h - P_i^1 \underline{\sigma}_h\|^2_{L^2(V_i^1)} + \|\underline{\sigma}_h - P_i^2 \underline{\sigma}_h\|^2_{L^2(V_i^2)},$$

where P_i^1 and P_i^2 are the L^2-projections on the affine functions on V_i^1 and V_i^2, respectively.

To simplify our notation we shall use the concept of "higher order terms" (h.o.t.). Since the finite volume scheme at hand is of first order for $u \in H^2(\Omega)$, i.e., $\|e\|_a \lesssim h$, then it is reasonable to denote all terms that tend to zero faster than $O(h)$ by h.o.t. Below, we shall refer to the following quantities as h.o.t.:

(a) $\|h^2\ \nabla(\gamma u_h)\|_{L^2(\Omega)}$ for $\gamma \in H^1(\Omega)$;
(b) $\|h^2\ \nabla f\|_{L^2(\Omega)}$ if $f \in H^1(\Omega)$;
(c) $\|hf\|_{L^2(\Omega_D)}$ if $f \in H^s(\Omega)$, $s > 0$, and $\Omega_D := \cup\{V_i : x_i \in N_h \cap \Gamma_D\}$ is a strip of width h around Γ_D (to show that this quantity is h.o.t. we apply Ilin's inequality (2.1) and get $\|hf\|_{L^2(\Omega_D)} \lesssim h^{1+s}\|f\|_{H^s(\Omega)}$, $s < 1/2$);
(d) $h_E^{1/2}\|g - \bar{g}\|_{L^2(E)}$ for $\bar{g} = \int_E g ds/|E|$ and $g \in H^1(E)$ for $E \subset \Gamma_N^{\text{out}}$;
(e) denote by $\tilde{r}(x)$ a linear approximation of $r(x)$ on K. Thus, $\widetilde{\nabla \cdot A}$ and $\widetilde{\nabla \cdot \underline{b}}$ are linear approximations on K of $\nabla \cdot A$ and $\nabla \cdot \underline{b}$, respectively. Here $\nabla \cdot A$ is understood as a vector with components divergence of the rows of $A(x)$. If A and $\underline{b}$ are sufficiently smooth on K, then $\nabla \cdot A - \widetilde{\nabla \cdot A}$ and $\nabla \cdot \underline{b} - \widetilde{\nabla \cdot \underline{b}}$ are h.o.t.

More generally, if functions $\alpha(h)$, $\beta(h)$, and $\gamma(h)$ satisfy $\alpha(h) \le \beta(h) + \gamma(h)$ and $\gamma(h)/\beta(h) \to 0$ as $h \to 0$, we will denote $\gamma(h)$ as h.o.t. compared to $\beta(h)$. In the case above we have $\beta(h) = h$.

In the analysis that follows we derive a posteriori error estimates based on averaging techniques. In the estimates derived the constants in $\lesssim$ depend only on c_0 from (2.3), c_1 from (2.4), and the shape of the elements in $\mathcal{T}$ and $\mathcal{T}^*$. The h.o.t. will account for the smoothness of the coefficients of the differential equation. The smoothness requirements, as stated in the theorems below, yield h.o.t. of order $O(h^2)$, i.e., one order higher than needed. Using standard results from interpolation of Banach spaces (cf., e.g., [1]) we can weaken the assumptions, requiring smoothness of order $\epsilon > 0$ less than that stated.

LEMMA 3.3. *Let the coefficients A and $\underline{b}$ be $C^1(\Omega)$-functions and let P_i be the L^2-projection onto the affine functions on $V_i \in \mathcal{T}^*$. Then*

$$h_i^{1/2}\|[\underline{\sigma}_h] \cdot \underline{n}\|_{L^2(\beta_i)} \lesssim \|\underline{\sigma}_h - P_i \underline{\sigma}_h\|_{L^2(V_i)} + \text{h.o.t.} \quad \text{for all } V_i \in \mathcal{T}^*. \tag{3.5}$$

The multiplicative constants in the notation $\lesssim$ depend on the shape of the elements in $\mathcal{T}$ and the shape of the control volumes in $\mathcal{T}^*$, while the h.o.t. depend on the smoothness of the coefficients A and $\underline{b}$.

Proof. If A and $\underline{b}$ are polynomials, then $\underline{\sigma}_h|_K$ is in a finite dimensional space for any $K \in \mathcal{T}$. In this case we easily prove (3.5) without h.o.t. by an equivalence-of-norm argument on finite dimensional spaces. Namely, both sides of (3.5) define seminorms for finite dimensional $\underline{\sigma}_h$. If $\|\underline{\sigma}_h - P_i\underline{\sigma}_h\|_{L^2(V_i)} = 0$ for some $\underline{\sigma}_h$, then $\underline{\sigma}_h = P_i\underline{\sigma}_h$ on V_i. Since $P_i\underline{\sigma}_h$ is linear on V_i, this shows that $\underline{\sigma}_h$ is also linear. Therefore, the jump $[\underline{\sigma}_h]$ is zero on β_i, i.e., the left-hand side of (3.5) vanishes as well. This proves that the seminorm on the right-hand side is stronger than the seminorm on the left-hand side and so proves (3.5). A scaling argument shows that the multiplicative constant behind $\lesssim$ is independent of h_i.

The case when A and $\underline{b}$ are smooth functions but $\underline{\sigma}_h|_K$ is not finite dimensional over $K \in \mathcal{T}$ is treated using approximation. Namely, we introduce polynomial approximations $\overline{\underline{\sigma}}_h$ of $\underline{\sigma}_h$ for any $K \in \mathcal{T}$ based on approximations of A and $\underline{b}$, taking into account that

$$\|\underline{\sigma}_h - \overline{\underline{\sigma}}_h\|_{L^2(V_i)} = \text{h.o.t.} \quad \text{and} \quad \|[\underline{\sigma}_h - \overline{\underline{\sigma}}_h] \cdot \underline{n}\|_{L^2(V_i)} = \text{h.o.t.},$$

and use the result for the finite dimensional case to get (3.5). ∎

As a corollary we get the following inequality.

COROLLARY 3.1. *Let the assumptions of Lemma* 3.3 *be satisfied. Then*

$$\eta_E \lesssim \eta_Z + \textit{h.o.t.} \tag{3.6}$$

The above inequality follows directly by squaring (3.5) and summing over all $x_i \in N_h$.

Recall that η_Z is defined for internal vertex nodes. Below we show that η_Z together with η_N can be used as an estimator for the H^1-norm of the error modulus of h.o.t.

THEOREM 3.2. *Let the assumptions of Lemma* 3.3 *be satisfied and let* $f \in H^1(\Omega)$. *Then*

$$\|e\|_a \lesssim \eta_Z + \eta_N + \textit{h.o.t.} \tag{3.7}$$

Proof. We use again the error representation from Lemma 3.1. In Theorem 3.1 we have bounded the third and fourth sums from the error representation by $\eta_N\|\nabla e\|$ and the second sum by $\eta_E\|\nabla e\|$. Further, η_E was bounded in Lemma 3.3 by η_Z + h.o.t., so it remains to establish the bound

$$(f - \nabla \cdot \underline{\sigma}_h - \gamma u_h,\ e - \overline{e}) \lesssim (\eta_Z + \text{h.o.t.})\|\nabla e\|.$$

For $x_i \in N_h^0$ denote by $\overline{f}$ and $\overline{\gamma u_h}$ the integral means over V_i of f and γu_h, respectively. Then we have

$$\begin{aligned}\int_{V_i} (f - \nabla_h \cdot \underline{\sigma}_h - \gamma u_h)(e - \overline{e})\,dx &= \int_{V_i} (f - \overline{f})(e - \overline{e})\,dx \\ &\quad - \int_{V_i} \nabla_h \cdot (\underline{\sigma}_h - P_i\underline{\sigma}_h)(e - \overline{e})\,dx - \int_{V_i} (\gamma u_h - \overline{\gamma u_h})(e - \overline{e})\,dx \\ &\le \|e - \overline{e}\|_{L^2(V_i)}\big(\|f - \overline{f}\|_{L^2(V_i)} + \|\nabla_h \cdot (\underline{\sigma}_h - P_i\underline{\sigma}_h)\|_{L^2(V_i)} \\ &\qquad + \|\gamma u_h - \overline{\gamma u_h}\|_{L^2(V_i)}\big).\end{aligned} \tag{3.8}$$

Poincaré's inequality gives

$$\begin{aligned}\|e - \overline{e}\|_{L^2(V_i)} &\lesssim h_i\|\nabla e\|_{L^2(V_i)}, \\ \|f - \overline{f}\|_{L^2(V_i)} &\lesssim h_i\|\nabla f\|_{L^2(V_i)}, \\ \|\gamma u_h - \overline{\gamma u_h}\|_{L^2(V_i)} &\lesssim h_i\|\nabla(\gamma u_h)\|_{L^2(V_i)}.\end{aligned} \tag{3.9}$$

The term $\|\nabla_h \cdot (\underline{\sigma}_h - P_i\underline{\sigma}_h)\|_{L^2(V_i)}$ is treated by the inverse estimate

$$\|\nabla_h \cdot (\underline{\sigma}_h - P_i\underline{\sigma}_h)\|_{L^2(V_i)} \lesssim h_i^{-1}\|\underline{\sigma}_h - P_i\underline{\sigma}_h\|_{L^2(V_i)} + \text{h.o.t.} \tag{3.10}$$

As in the proof of Lemma 3.3, we first prove (3.10) when $\underline{\sigma}_h$ is finite dimensional by equivalence of norms followed by a scaling argument and then, for the general case, by a perturbation analysis. The combination of (3.8)–(3.10) shows

$$\int_{V_i} (f - \nabla_h \cdot \underline{\sigma}_h - \gamma u_h)(e - \overline{e})\, dx \lesssim \|\nabla e\|_{L^2(V_i)} \left(\|\underline{\sigma}_h - P_i\underline{\sigma}_h\|_{L^2(V_i)} + \text{h.o.t.}\right). \tag{3.11}$$

So far (3.11) holds for $x_i \in N_h^0$. For $x_i \in N_h \cap \Gamma_D$ we replace $\overline{e}$, $\overline{f}$, and $\overline{\gamma u_h}$ by zero and deduce the first and third inequalities of (3.9) from Friedrichs's inequality (notice that e and γu_h vanish on $\Gamma_D \cap V_i$). The inverse estimate (3.10) holds for $x_i \in N_h \cap \Gamma_D$ as well. The aforementioned arguments prove (3.11) with $\|h^2\ \nabla f\|_{L^2(V_i)}$ replaced by $\|h\, f\|_{L^2(V_i)}$. This shows

$$(f - \nabla_h \cdot \underline{\sigma}_h - \gamma u_h, e - \overline{e}) \lesssim (\eta_Z + \|hf\|_{L^2(\Omega_D)} + \text{h.o.t.})\|\nabla e\|.$$

The last result, the discussion at the beginning of the theorem, Ilin's inequality (2.1), and the ellipticity assumption conclude the proof of the theorem. ∎

THEOREM 3.3. *Suppose that the coefficients A and $\underline{b}$ are $C^1(\Omega)$-functions, $f \in H^1(\Omega)$, $\gamma \in H^1(\Omega)$, $g \in H^{1/2}(\mathcal{E})$, and that the partitioning $\mathcal{T}$ of Ω is locally quasi uniform. Then*

$$\eta_Z + \eta_R + \eta_E + \eta_N \lesssim \|e\|_a + h.o.t.$$

Proof. We will prove that the quantities η_R, η_E, η_N, and η_Z are bounded by $C\ \|e\|_a + \text{h.o.t.}$ The h.o.t. appear by applying averaging techniques as in the proof of Lemma 3.3 and therefore we will consider only the case when $\underline{\sigma}_h$ is finite dimensional. First, we will bound the contributions to η_N due to Γ_N^{in}, namely, we will prove

$$\|h^{1/2}\,(g - \underline{\sigma}_h \cdot \underline{n})\|_{L^2(\Gamma_N^{\text{in}})} \lesssim \|e\|_a + \text{h.o.t.} \tag{3.12}$$

We consider an element $K \in \mathcal{T}$ that has an edge/face $E \subset \Gamma_N^{\text{in}}$. We will use the pair (K, E) in the rest of the proof (see Figure 3).

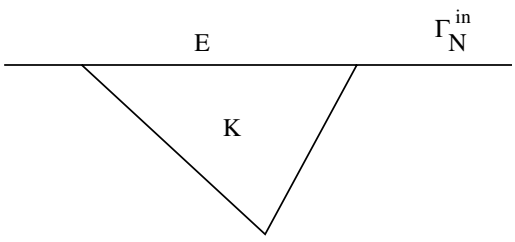


FIG. 3. *The pair (K, E) of edge $E \subset \Gamma_N^{\text{in}}$ and element K used in the proof of inequality* (3.12).

First, we note that

$$h_E^{1/2}\|g-\overline{g}\|_{L^2(E)} = \text{h.o.t.} \quad \text{for} \quad \overline{g} := \int_E g\,ds/|E|.$$

Then

$$\|g-\underline{\sigma}_h\cdot\underline{n}\|_{L^2(E)} \le \|g-\overline{g}\|_{L^2(E)} + \|\overline{g}-\underline{\sigma}_h\cdot\underline{n}\|_{L^2(E)} \lesssim \|\overline{g}-\underline{\sigma}_h\cdot\underline{n}\|_{L^2(E)} + \text{h.o.t.}$$

We prove below that

$$\|\overline{g}-\underline{\sigma}_h\cdot\underline{n}\|_{L^2(E)} \lesssim h_E^{-1/2}\|\underline{\sigma}-\underline{\sigma}_h\|_{L^2(K)} + \text{h.o.t.}$$

so that summation over all $E\subset\Gamma_N^{\text{in}}$ yields (3.12).

Consider an edge-bubble function $b_E\in H^1(\Omega)$, $b_E\ge 0$, $b_E(x)=0$ on $\Omega\setminus K$ and $\partial K\setminus E$, with properties

$$\int_E b_E\,ds = \int_E ds,\ \|b_E\|_{L^\infty(K)}\lesssim 1,\ \|\nabla b_E\|_{L^\infty(K)}\lesssim 1/h_E. \tag{3.13}$$

A 2-D example of such a bubble is $b_E = 6\phi_1\phi_2$, where ϕ_1 and ϕ_2 are the standard linear nodal basis functions associated with the end points of the edge E. Let $z\in H^1(K)$ be the harmonic extension of $(\overline{g}-\underline{\sigma}_h\cdot\underline{n})b_E$ from ∂K to K. The extension is bounded in H^1 [29, Theorem 4.1.1] on a reference element $\hat{K}$ by the $H^{1/2}(\hat{E})$-norm of the extended quantity and, since all norms are equivalent on a finite dimensional space, by its $L^2(\hat{E})$-norm. Therefore, a scaling argument gives

$$h_E^{1/2}\|\nabla z\|_{L^2(K)} + h_E^{-1/2}\|z\|_{L^2(K)} \lesssim \|b_E(\overline{g}-\underline{\sigma}_h\cdot\underline{n})\|_{L^2(E)}. \tag{3.14}$$

We define the linear operator P_K into the space of polynomials of degree 2 on an element $K\in\mathcal{T}$ as

$$(b_K\ P_K z, p_h)_{L^2(K)} = (z,p_h)_{L^2(K)}$$

for all polynomials p_h of degree 2. Here $b_K\in H^1(\Omega)$, $b_K\ge 0$, is an element-bubble function with properties

$$\text{supp } b_K\subset K,\ \int_K b_K\,ds = \int_K ds,\ \|b_K\|_{L^\infty(K)}\lesssim 1,\ \|\nabla b_K\|_{L^\infty(K)}\lesssim 1/h_K.$$

A 2-D example of such a bubble is $b_K = 60\phi_1\phi_2\phi_3$, where ϕ_1, ϕ_2, and ϕ_3 are the standard linear nodal basis functions associated with the vertices of the element K. Then $\tilde{z} := z - b_K\,P_K z$ by construction has the properties

$$\tilde{z} = (\overline{g}-\underline{\sigma}_h\cdot\underline{n})b_E \quad \text{on } E, \qquad \tilde{z}=0 \quad \text{on } \partial K\backslash E,$$
$$(\tilde{z},\,p_h)_{L^2(K)} = 0 \quad \text{for all polynomials } p_h \text{ of degree 2.}$$

Inequality (3.14) remains valid for z replaced by $\tilde{z}$ because of the following. Choosing $p_h = P_K z$ in the definition of P_K yields

$$\|b_K^{1/2}P_K z\|_{L^2(K)}^2 = (z,P_K z)_{L^2(K)} \lesssim \|z\|_{L^2(K)}\|P_K z\|_{L^2(K)}.$$

We use norm equivalence on finite dimensional spaces on a reference element and scaling to K to get that the quantities $||b_K P_K z||_{L^2(K)}$, $\|b_K^{1/2}P_K z\|_{L^2(K)}$, and $\|P_K z\|_{L^2(K)}$

are equivalent up to constants independent of h, and therefore $\|b_K P_K z\|_{L^2(K)} \lesssim \|z\|_{L^2(K)}$. We use again the equivalence-of-norms argument, inverse inequality, and the properties of z to get that

$$\begin{aligned}\|\nabla(b_K P_K z)\|_{L^2(K)} &\lesssim \|\nabla b_K\|_{L^2(K)}\|P_K z\|_{L^2(K)} + \|b_K \nabla(P_K z)\|_{L^2(K)}\\ &\lesssim h_E^{-1}\|P_K z\|_{L^2(K)} + h_E^{-1}\|z\|_{L^2(K)}\\ &\lesssim h_E^{-1/2}\|b_E(\overline{g} - \underline{\sigma}_h\cdot\underline{n})\|_{L^2(E)}.\end{aligned}$$

Combined with the bound for $\|b_K P_K z\|_{L^2(K)}$, this completes the proof of (3.12) for $z = \tilde{z}$.

Given a polynomial p_h of degree 2, using the Gauss divergence theorem and the properties of $\tilde{z}$, we deduce

$$\begin{aligned}&\int_E b_E(\overline{g} - \underline{\sigma}_h\cdot\underline{n})(\underline{\sigma} - \underline{\sigma}_h)\cdot\underline{n}\,ds = \int_{\partial K}\tilde{z}(\underline{\sigma} - \underline{\sigma}_h)\cdot\underline{n}\,ds\\ &\quad= \int_K(\underline{\sigma} - \underline{\sigma}_h)\cdot\nabla\tilde{z}\,dx + \int_K \tilde{z}(\nabla\cdot(\underline{\sigma} - \underline{\sigma}_h) - p_h)\,dx\\ &\quad\lesssim \big(\|\underline{\sigma} - \underline{\sigma}_h\|_{L^2(K)} + h_E\|\nabla\cdot(\underline{\sigma} - \underline{\sigma}_h) - p_h\|_{L^2(K)}\big)\, h_E^{-1/2}\|b_E(\overline{g} - \underline{\sigma}_h\cdot\underline{n})\|_{L^2(E)}.\end{aligned}$$

Choosing proper p_h in the second term of the last inequality makes that term h.o.t. Indeed, write down first the equality (see the basic problem (1.1))

$$\nabla\cdot(\underline{\sigma} - \underline{\sigma}_h) - p_h = \gamma u - f - (\nabla\cdot A)\cdot\nabla u_h + u_h\,\nabla\cdot\underline{b} + \underline{b}\cdot\nabla u_h - p_h. \tag{3.15}$$

Here $\nabla\cdot A$ is understood as a vector with component divergence on the rows of $A(x)$. Let $\tilde{f}$, $\widetilde{\gamma u}$, $\widetilde{\nabla\cdot\underline{b}}$, $\tilde{\underline{b}}$, and $\widetilde{\nabla\cdot A}$ be the linear approximations on K of f, γu, $\nabla\cdot\underline{b}$, $\underline{b}$, and $\nabla\cdot A$, respectively.

Now, we choose p_h to be the following polynomial of degree 2 on K,

$$p_h = \widetilde{\gamma u} - \tilde{f} - (\widetilde{\nabla\cdot A})\cdot\nabla u_h + u_h\widetilde{\nabla\cdot\underline{b}} + \tilde{\underline{b}}\cdot\nabla u_h,$$

take the $L^2(K)$-norm of (3.15), and use the triangle's inequality to get

$$\begin{aligned}\|\nabla\cdot(\sigma - \underline{\sigma}_h) - p_h\|_{L^2(K)} &\le \|f - \tilde{f}\|_{L^2(K)} + \|\gamma u - \widetilde{\gamma u}\|_{L^2(K)} + \|u_h(\nabla\cdot\underline{b} - \widetilde{\nabla\cdot\underline{b}})\|_{L^2(K)}\\ &\quad + \|(\underline{b} - \tilde{\underline{b}})\cdot\nabla u_h\|_{L^2(K)} + \|\nabla u_h\cdot(\nabla\cdot A - \widetilde{\nabla\cdot A})\|_{L^2(K)}\\ &\lesssim \big(\|u\|_{H^2(K)} + \|u_h\|_{H^1(K)}\big)\ \text{h.o.t.} + \|h_K\nabla f\|_{L^2(K)}.\end{aligned}$$

Therefore (note that $g = \underline{\sigma}\cdot\underline{n}$ on Γ_N^{in}),

$$\begin{aligned}\|b_E^{1/2}(\overline{g} - \underline{\sigma}_h\cdot\underline{n})\|_{L^2(E)}^2 &= \int_E \tilde{z}(\overline{g} - g)\,ds + \int_E\tilde{z}(\underline{\sigma} - \underline{\sigma}_h)\cdot\underline{n}\,ds\\ &\lesssim h_E^{-1/2}\big(\|\underline{\sigma} - \underline{\sigma}_h\|_{L^2(K)} + \text{h.o.t.}\big)\ \|b_E^{1/2}(\overline{g} - \underline{\sigma}_h\cdot\underline{n})\|_{L^2(E)}\end{aligned}$$

and so

$$\|b_E^{1/2}(\overline{g} - \underline{\sigma}_h\cdot\underline{n})\|_{L^2(E)} \lesssim h_E^{-1/2}\|\underline{\sigma} - \underline{\sigma}_h\|_{L^2(K)} + \text{h.o.t.}$$

Using again the equivalence-of-norms estimate (equivalence of norms on finite dimensional spaces on reference element and scaling)

$$\|\overline{g} - \underline{\sigma}_h\cdot\underline{n}\|_{L^2(E)} \lesssim \|b_E^{1/2}(\overline{g} - \underline{\sigma}_h\cdot\underline{n})\|_{L^2(E)}$$

we finally prove that

$$\begin{aligned}\|g-\underline{\sigma}_h\cdot\underline{n}\|_{L^2(E)} &\le \|g-\overline{g}\|_{L^2(E)}+\|\overline{g}-\underline{\sigma}_h\cdot\underline{n}\|_{L^2(E)}\\ &\lesssim \|b_E^{1/2}(\overline{g}-\underline{\sigma}_h\cdot\underline{n})\|_{L^2(E)}+\text{h.o.t.}\\ &\lesssim h_E^{-1/2}\|\underline{\sigma}-\underline{\sigma}_h\|_{L^2(K)}+\text{h.o.t.}\end{aligned}$$

Similarly, $\|A\nabla u_h\cdot\underline{n}\|_{L^2(E)}\lesssim h_E^{-1/2}\|\underline{\sigma}-\underline{\sigma}_h\|_{L^2(K)}+\text{h.o.t.}$ for $E\subset\Gamma_N^{\text{out}}$, which, combined with the result for $E\subset\Gamma_N^{\text{in}}$, proves that $\eta_N\lesssim ||e||_a+\text{h.o.t.}$

A similar technique shows that $\eta_E\lesssim\|e\|_a+\text{h.o.t.}$

The inequality $\eta_R\lesssim\|e\|_a+\text{h.o.t.}$ can be proved in the following way. Take the average $\bar{R}_K$ of the residual $R_K:=f-\nabla\cdot\underline{\sigma}_h-\gamma u_h$ over an element K to derive

$$\|\bar{R}_K\|_{L^2(K)}\le\|R_K-\bar{R}_K\|_{L^2(K)}+\|R_K\|_{L^2(K)}=\ \text{h.o.t.}+\|R_K\|_{L^2(K)}.$$

Further, apply the technique from Lemma 3.1 to deduce the equality $(R_K,b_K\bar{R}_K)_{L^2(K)}=a(e,b_K\bar{R}_K)$ and therefore

$$\begin{aligned}(R_K,b_K\bar{R}_K)_{L^2(K)}&=\|b_K^{1/2}R_K\|^2_{L^2(K)}-(R_K,b_K(R_K-\bar{R}_K))_{L^2(K)}=a(e,b_K\bar{R}_K)\\ &\lesssim\|e\|_{H^1(K)}\|b_K\bar{R}_K\|_{H^1(K)}\lesssim\|e\|_{H^1(K)}h_K^{-1}\|\bar{R}_K\|_{L^2(K)}\\ &\lesssim h_K^{-1}\|e\|_{H^1(K)}\|R_K\|_{L^2(K)}+\text{h.o.t.}\end{aligned}$$

Here we used the inverse inequality and the boundedness of the coefficients of the differential equation (1.1). Then we take the term $(R_K,b_K(R_K-\bar{R}_K))_{L^2(K)}$ to the right-hand side and consider it as h.o.t. Finally, use that $\|b_K^{1/2}R_K\|_{L^2(K)}\approx\|R_K\|_{L^2(K)}$ to obtain

$$\|R_K\|_{L^2(K)}\lesssim h_K^{-1}\|e\|_{H^1(K)}+\text{h.o.t.}$$

A summation over all $K\in\mathcal{T}$ yields the inequality $\eta_R\lesssim\|e\|_a+\text{h.o.t.}$

Now we prove the remaining inequality, $\eta_Z\lesssim\|e\|_a+\text{h.o.t.}$ Since P_i is a linear $L^2(V_i)$ projector, we have that

$$\|\underline{\sigma}_h-P_i\underline{\sigma}_h\|_{L^2(V_i)}\le\|\underline{\sigma}_h-P_i\underline{\sigma}\|_{L^2(V_i)}.$$

Adding and subtracting $\underline{\sigma}$ in the right-hand side and applying the triangle's inequality we get

$$\|\underline{\sigma}_h-P_i\underline{\sigma}\|_{L^2(V_i)}\le\|\underline{\sigma}_h-\underline{\sigma}\|_{L^2(V_i)}+\|\underline{\sigma}-P_i\underline{\sigma}\|_{L^2(V_i)}=\|\underline{\sigma}_h-\underline{\sigma}\|_{L^2(V_i)}+\text{h.o.t.}$$

since $\|\underline{\sigma}-P_i\underline{\sigma}\|_{L^2(V_i)}=\text{h.o.t.}$ for $\underline{\sigma}$ smooth. The summation over all x_i concludes the proof of the theorem. ∎

3.2. Analysis of the upwind scheme in the H^1-norm. This section is devoted to the case when an upwind approximation is applied to the convection term, namely, we consider problem (2.11).

DEFINITION 3.3. *For an element $K\in\mathcal{T}$ we denote by $\gamma_K:=\cup_{\gamma_{ij}}(K\cap\gamma_{ij})$ and set*

$$\eta_E^{\text{up}}:=\left(\sum_{K\in\mathcal{T}}\sum_{\gamma_{ij}\subset\gamma_K}\|h^{1/2}\underline{b}\cdot\underline{n}\,(u_h(x_i)-u_h)\|^2_{L^2(\gamma_{ij})}\right)^{1/2},$$

$$\eta_N^{\text{up}}:=\|h^{1/2}\underline{b}\cdot\underline{n}\,\nabla u_h\|_{L^2(\Gamma_N^{\text{out}})}.$$

THEOREM 3.4. *Let the assumptions of Theorems* 3.1 *and* 3.2 *be satisfied, and let the upwind approximation be applied to the convection term. Then*

$$\|e\|_a \lesssim \eta_Z + \eta_N + \eta_E^{\mathrm{up}} + \eta_N^{\mathrm{up}} + h.o.t. \tag{3.16}$$

Proof. Since $a_h^{\mathrm{up}}(u_h, v^*) = F(v^*)$ and $a_h(u, v^*) = F(v^*)$ for $v^* \in S_h^*$ we have the orthogonality condition $a_h(u, v^*) - a_h^{\mathrm{up}}(u_h, v^*) = 0$. Choosing $v^* = \bar{e}$ we get the following representation for the energy norm of the error:

$$\begin{aligned}\|e\|_a^2 &= a(e,e) - a_h(u,\bar{e}) + a_h^{\mathrm{up}}(u_h,\bar{e}) \\ &= \{a(e,e) - a_h(e,\bar{e})\} + \{a_h^{\mathrm{up}}(u_h,\bar{e}) - a_h(u_h,\bar{e})\} \\ &= \{a(e,e) - a_h(e,\bar{e})\} + \{C_h^{\mathrm{up}}(u_h,\bar{e}) - C_h(u_h,\bar{e})\}.\end{aligned}$$

For the first term, $a(e,e) - a_h(e,\bar{e})$, we use the same approach as in the analysis of the scheme without upwind (see Lemma 3.1) and show that

$$\begin{aligned}a(e,e) - a_h(u,\bar{e}) &= (f - \nabla_h \cdot \underline{\sigma}_h - \gamma u_h,\ e - \bar{e}) - \int_{\mathcal{E}} [\underline{\sigma}_h] \cdot \underline{n}\,(e - \bar{e})\,ds \\ &\quad - \int_{\Gamma_N^{\mathrm{in}}} (g - \underline{\sigma}_h \cdot \underline{n})\,(e - \bar{e})\,ds - \int_{\Gamma_N^{\mathrm{out}}} (A\nabla_h u_h) \cdot \underline{n}\,(e - \bar{e})\,ds.\end{aligned} \tag{3.17}$$

This presentation allows us to use estimate (3.7) of Theorem 3.2.

For the second term, $C_h^{\mathrm{up}}(u_h,\bar{e}) - C_h(u_h,\bar{e})$, we get

$$\begin{aligned}C_h^{\mathrm{up}}(u_h,\bar{e}) - C_h(u_h,\bar{e}) = \sum_{x_i \in N_h^0} \bar{e}_i \Bigg\{ &\sum_{j \in \Pi(i)} \int_{\gamma_{ij}} ((\underline{b}\cdot\underline{n})_+ u_h(x_i) + (\underline{b}\cdot\underline{n})_- u_h(x_j) - \underline{b}\cdot\underline{n} u_h)\,ds \\ &+ \int_{\partial V_i \cap \Gamma_N^{\mathrm{out}}} (\underline{b}\cdot\underline{n}\ u_h(x_i) - \underline{b}\cdot\underline{n}\ u_h)\ ds \Bigg\}.\end{aligned}$$

Here the unit normal vector $\underline{n}$ on γ_{ij} is oriented in such a way that $\underline{b}\cdot\underline{n} \geq 0$. We want to express the above sum as a sum over the elements. To do so we specify that the indexes (ij) are oriented so that $(x_i - x_j)\cdot\underline{n} \leq 0$. We get that

$$\begin{aligned}C_h^{\mathrm{up}}(u_h,\bar{e}) - C_h(u_h,\bar{e}) = \sum_{K \in \mathcal{T}_h} \Bigg\{ &\sum_{\gamma_{ij} \subset K} (\bar{e}_i - \bar{e}_j) \int_{\gamma_{ij}} \underline{b}\cdot\underline{n}\,(u_h(x_i) - u_h)\,ds \\ &+ \sum_{V_i \cap K} \bar{e}_i \int_{\partial V_i \cap \Gamma_N^{\mathrm{out}}} \underline{b}\cdot\underline{n}\ (u_h(x_i) - u_h)\ ds \Bigg\}.\end{aligned}$$

We denote by $[\bar{e}] := e_i - \bar{e}_j$ the jump of $\bar{e}$ across γ_{ij} and take into account that $[\bar{e} - e] = [\bar{e}]$. Then, by the Schwarz inequality, the term involving the integral over γ_{ij} is bounded by $C\|[e - \bar{e}]\|_{L^2(\gamma_{ij})}\ \|\underline{b}\cdot\underline{n}\,(u_h(x_i) - u_h)\|_{L^2(\gamma_{ij})}$. As before, using trace, Poincaré's, and/or Friedrichs's inequalities we get

$$\|[e - \bar{e}]\|_{L^2(\gamma_{ij})} \lesssim h_i^{1/2} \|\nabla e\|_{L^2(V_i)},$$

which bounds the integrals over γ_{ij} in the error representation with η_E^{up}.

For the terms involving integration over Γ_N^{out} we have

$$|u_h(x_i) - u_h(x)| \leq |\nabla u_h \cdot \underline{t}(x)|.|x_i - x|.$$

Here $\underline{t}(x)$ is a unit vector along $\partial V_i \cap K$, an edge in two dimensions, or a face in three dimensions. Then in two dimensions $\underline{t}$ is simply a unit vector perpendicular to $\underline{n}$, while in three dimensions $\underline{t}(x)$ depends on the position of x on the face and is again perpendicular to $\underline{n}$. In both cases $|u_h(x_i) - u_h(x)| \leq |h_K \nabla u_h|$. Using the Schwarz inequality we bound the term involving integration over Γ_N^{out} in the following way:

$$\sum_{V_i \cap K} \bar{e}_i \int_{\partial V_i \cap \Gamma_N^{\text{out}}} \underline{b} \cdot \underline{n}\, (u_h(x_i) - u_h)\, ds \leq C \|\nabla e\|.\|h^{1/2} \underline{b} \cdot \underline{n}\, \nabla u_h\|_{L^2(\Gamma_N^{\text{out}})},$$

which eventually gives the term η_N^{up} in (3.16) and completes the proof. ∎

3.3. Error estimates in L^2. We use duality techniques to get error estimators for different quantities of the error. In this subsection we will show how to use the duality technique in order to derive an error estimator in the global $L^2(\Omega)$-norm for the scheme without upwinding. The main assumption in this section is that the solution of problem (1.1) is H^2 regular.

DEFINITION 3.4. *We define the residual L^2 a posteriori error estimator $\tilde{\rho}$ as*

$$\tilde{\rho} := (\tilde{\eta}_R^2 + \tilde{\eta}_E^2 + \tilde{\eta}_N^2)^{1/2}, \tag{3.18}$$

where

$$\begin{aligned}
\tilde{\eta}_R^2 &:= \|h\,(R_K - \bar{R}_K)\|^2 + \|h^2\, R_K\|^2, \\
\tilde{\eta}_E^2 &:= \|h^{1/2}\,(R_E - \bar{R}_E)\|_{L^2(\mathcal{E})}^2 + \|h^{3/2}\, R_E\|_{L^2(\mathcal{E})}^2, \\
\tilde{\eta}_N^2 &:= \|h^{1/2}\,(R_E^{\text{in}} - \bar{R}_E^{\text{in}})\|_{L^2(\Gamma_N^{\text{in}})}^2 + \|h^{3/2}\, R_E^{\text{in}}\|_{L^2(\Gamma_N^{\text{in}})}^2 \\
&\quad + \|h^{1/2}\,(R_E^{\text{out}} - \bar{R}_E^{\text{out}})\|_{L^2(\Gamma_N^{\text{out}})}^2 + \|h^{3/2}\, R_E^{\text{out}}\|_{L^2(\Gamma_N^{\text{out}})}^2,
\end{aligned}$$

and $\bar{R}_K$, $\bar{R}_E$, $\bar{R}_E^{\text{in}}$, and $\bar{R}_E^{\text{out}}$ are the $K \in \mathcal{T}$, $E \in \mathcal{E}$, $E \in \Gamma_N^{\text{in}}$, and $E \in \Gamma_N^{\text{out}}$ piecewise mean values of, correspondingly, R_K, R_E, R_E^{in}, and R_E^{out} introduced in Definition 3.1.

Our aim is to show that the estimator $\tilde{\rho}$ is reliable in the $L^2(\Omega)$-norm. The a posteriori $L^2(\Omega)$ error analysis involves the following continuous dual problem: Find $\tilde{e} \in H_D^1(\Omega)$ such that

$$a(v, \tilde{e}) = (e, v) \text{ for any } v \in H_D^1(\Omega), \tag{3.19}$$

where e is the exact error, defined as before.

THEOREM 3.5. *Let the solution $\tilde{e}$ of the dual problem* (3.19) *be $H^2(\Omega)$ regular. If the coefficients of our basic problem* (1.1) *are sufficiently regular, namely R_K, R_E, R_E^{in}, and R_E^{out} are correspondingly in $H^1(K)$, $H^{1/2}(E)$, $H^{1/2}(\Gamma_N^{\text{in}})$, and $H^{1/2}(\Gamma_N^{\text{out}})$, then the residual L^2 a posteriori error estimator* (3.18) *from Definition* 3.4 *is reliable, i.e., $\|e\| \lesssim \tilde{\rho}$.*

Proof. Let $v = e$ in (3.19) and argue as in the proof of Lemma 3.1 to show

$$\begin{aligned}
\|e\|^2 = a(e, \tilde{e}) &= (R_K, \tilde{e} - e^*) - (R_E, \tilde{e} - e^*)_{L^2(\mathcal{E})} \\
&\quad - (R_E^{\text{in}}, \tilde{e} - e^*)_{L^2(\Gamma_N^{\text{in}})} - (R_E^{\text{out}}, \tilde{e} - e^*)_{L^2(\Gamma_N^{\text{out}})}
\end{aligned} \tag{3.20}$$

for an arbitrary $e^* \in S_h^*$. To evaluate the right-hand side of this identity we use the nodal interpolation operator I_h and its properties. If $\tilde{e} \in H^2(\Omega)$, the Sobolev inequalities [12, Theorem 4.3.4] guarantee that $I_h\tilde{e}$ is well defined. The properties of the interpolant are well established in the finite element literature (see, for example, [12]), namely,

(3.21)
$$h_K^{-2}\|\tilde{e} - I_h\tilde{e}\|_{L^2(K)} + h_K^{-1}|\tilde{e} - I_h\tilde{e}|_{H^1(K)} + h_K^{-3/2}\|\tilde{e} - I_h\tilde{e}\|_{L^2(\partial K)} \leq C_{I,K}|\tilde{e}|_{H^2(K)}.$$

Now, in (3.20) we choose $e^* = I_h^* I_h\tilde{e}$ so that $\tilde{e} - e^* = (\tilde{e} - I_h\tilde{e}) + (I_h\tilde{e} - I_h^* I_h\tilde{e})$. Further, we apply the Schwarz inequality on the integrals involving $\tilde{e} - I_h\tilde{e}$ and use (3.21) to get the bound

$$\begin{aligned}(R_K, \tilde{e} - I_h\tilde{e}) - (R_E, \tilde{e} - I_h\tilde{e})_{L^2(\mathcal{E})} - (R_E^{\text{in}}, \tilde{e} - I_h\tilde{e})_{L^2(\Gamma_N^{\text{in}})} - (R_E^{\text{out}}, \tilde{e} - I_h\tilde{e})_{L^2(\Gamma_N^{\text{out}})}\\ \lesssim \big(\|h^2\ R_K\| + \|h^{3/2}\ R_E\|_{L^2(\mathcal{E})} + \|h^{3/2}\ R_E^{\text{in}}\|_{L^2(\Gamma_N^{\text{in}})} + \|h^{3/2}\ R_E^{\text{out}}\|_{L^2(\Gamma_N^{\text{out}})}\big)|\tilde{e}|_{H^2(\Omega)}.\end{aligned}$$

For the integrals involving $I_h\tilde{e} - I_h^* I_h\tilde{e}$ we first note that if K is a fixed element in $\mathcal{T}$, then for every vertex x_i of K, the quantities $|K \cap V_i|$ (volume in three dimensions and area in two dimensions) are equal. Also, for vertices x_i on the face/edge E we have that the boundary quantities $|E \cap V_i|$ (area in three dimensions and length in two dimensions) are also equal. Therefore,

$$\int_K (I_h\tilde{e} - I_h^* I_h\tilde{e})dx = 0, \qquad \int_E (I_h\tilde{e} - I_h^* I_h\tilde{e})ds = 0.$$

We apply the last fact to the integrals involving $I_h\tilde{e} - I_h^* I_h\tilde{e}$ in order to subtract from R_K, R_E, R_E^{in}, and R_E^{out} their mean values $\bar{R}_K$, $\bar{R}_E$, $\bar{R}_E^{\text{in}}$, and $\bar{R}_E^{\text{out}}$. Then, using Schwarz and Poincaré inequalities we bound the term involving $I_h\tilde{e} - I_h^* I_h\tilde{e}$, namely,

$$\begin{aligned}&|(R_K, I_h^* I_h\tilde{e} - I_h\tilde{e}) - (R_E, I_h^* I_h\tilde{e} - I_h\tilde{e})_{L^2(\mathcal{E})} - (R_E^{\text{in}}, I_h^* I_h\tilde{e} - I_h\tilde{e})_{L^2(\Gamma_N^{\text{in}})}\\ &\qquad - (R_E^{\text{out}}, I_h^* I_h\tilde{e} - I_h\tilde{e})_{L^2(\Gamma_N^{\text{out}})}|\\ &\lesssim \big(\|h\ (R_K - \bar{R}_K)\| + \|h^{1/2}(R_E - \bar{R}_E)\|_{L^2(\mathcal{E})}\\ &\qquad + \|h^{1/2}\ (R_E^{\text{in}} - \bar{R}_E^{\text{in}})\|_{L^2(\Gamma_N^{\text{in}})} + \|h^{1/2}\ (R_E^{\text{out}} - \bar{R}_E^{\text{out}})\|_{L^2(\Gamma_N^{\text{out}})}\big)\|\tilde{e}\|_{H^2(K)},\end{aligned}$$

where we have used the inequality

$$\begin{aligned}\|I_h\tilde{e} - I_h^* I_h\tilde{e}\|_{L^2(K)} &\lesssim h_K|I_h\tilde{e}|_{H^1(K)} \lesssim h_K|\tilde{e} - I_h\tilde{e}|_{H^1(K)} + h_K|\tilde{e}|_{H^1(K)}\\ &\lesssim h_K^2|\tilde{e}|_{H^2(K)} + h_K|\tilde{e}|_{H^1(K)} \lesssim h_K|\tilde{e}|_{H^2(K)}.\end{aligned}$$

Applying the above estimates, the stability of the dual problem with respect to the right-hand side, $\|\tilde{e}\|_{H^2(\Omega)} \leq C\|e\|$, and obvious manipulations, we get that the L^2 a posteriori error estimator $\tilde{\rho}$ is reliable. Moreover, since the coefficients of (1.1) are sufficiently regular we can apply Poincaré's inequality to the terms $\|R_K - \bar{R}_K\|_{L^2(K)}$, $\|R_E - \bar{R}_E\|_{L^2(E)}$, $\|R_E^{\text{in}} - \bar{R}_E^{\text{in}}\|_{L^2(E)}$, and $\|R_E^{\text{out}} - \bar{R}_E^{\text{out}}\|_{L^2(E)}$ to get one additional power of h that will make the error estimator of second order.

Note that we did not explicitly apply Poincaré's inequality in the definition of the error estimator in order to make it well defined for problems with less than that stated in the theorem regularity. ∎

4. Adaptive grid refinement and solution strategy. In this section we present the adaptive mesh refinement strategy that we use. It is based on the grid refinement approach in the finite element methods (see, e.g., [11, 36]). A different grid adaptation strategy, again in the finite element method, has been proposed, justified, and used in [20].

For a given finite element partitioning $\mathcal{T}$, desired error tolerance ρ, and a norm in which the tolerance to be achieved is, say $|||\cdot|||$, do the following:

- compute the finite volume approximation $u_h \in S_h$, as given in subsection 2.2;
- using the a posteriori error analysis, compute the errors ρ_K for all $K \in \mathcal{T}$;
- mark those finite elements K for which $\rho_K \geq \rho/\sqrt{N}$; here N is the number of elements in $\mathcal{T}$;
- if $\sum_{K\in\mathcal{T}} \rho_K^2 > \rho^2$, then refine the marked elements;
- additionally refine until a conforming mesh is reached;
- repeat the above process until no elements have been refined.

For the 2-D case we refine marked elements by uniformly splitting the marked triangles into four. The refinement to conformity is done by bisection through the longest edge. For the 3-D version of the code the elements (tetrahedrons) are refined using the algorithm described by Arnold, Mukherjee, and Pouly in [5].

The described procedure yields error control and optimal mesh (heuristics), which are the goals in the adaptive algorithm. The nested meshes obtained in the process are used to define multilevel preconditioners. The initial guess for every new level is taken to be the interpolation of u_h from the previous level.

5. Numerical examples. Here we present two sets of numerical examples to test the our theoretical results. The first two examples are simple 2-D elliptic problems while the remaining tests illustrate our approach on 3-D problems of flow and transport in porous media.

5.1. 2-D test problems. In Example 1 we consider problems with known solutions and compare the behavior of the error estimators with the exact errors. Example 2 is for discontinuous matrix $A(x)$ with an unknown solution.

Example 1. We consider three Dirichlet problems for the Poisson equation on an L-shaped domain with known exact solutions $u = r^{4/3}\sin\frac{4\theta}{3}$ (Problem 1), $u = r^{2/3}\sin\frac{2\theta}{3}$ (Problem 2), and $u = r^{1/2}\sin\frac{\theta}{2}$ (Problem 3). These functions belong to $H^{1+s}(\Omega)$ with s almost $4/3$, $2/3$, and $1/2$, respectively. In Figure 4 we show the mesh and the error for Problem 2 after four levels of local refinement.

The theory shows that the a posteriori error estimators η_E and η_Z are equivalent to the H^1-norm of the error. This theoretical result is confirmed by our computations, which are summarized in Figure 5. The left picture gives the exact error (solid line) and the a posteriori error estimators η_Z (dashed line) and η_E (dash-dotted line) for the three problems over the different levels of the mesh. The levels are obtained by uniform refinement (splitting every triangle into 4) and have 65, 255, 833, 3,201, 12,545, 49,665, and 197,633 nodes correspondingly for levels $1, \ldots, 7$. The errors are printed in logarithmic scale in order to demonstrate the linear behavior of the error as a function of the level. For exact solutions in $H^{1+1/2-\epsilon}$, $H^{1+2/3-\epsilon}$, and $H^{1+4/3-\epsilon}$ ($\epsilon > 0$) one can see the theoretically expected rate of error reduction over the levels of $1/2$, $2/3$, and 1 correspondingly. One can observe that both η_Z and η_E are equivalent to the exact error, as proved in the theoretical section. The same is true when the local refinement method from section 4 is applied. The numerical results are given in Figure 5, right. The y scale is again the error, and the x scale is the refinement level.

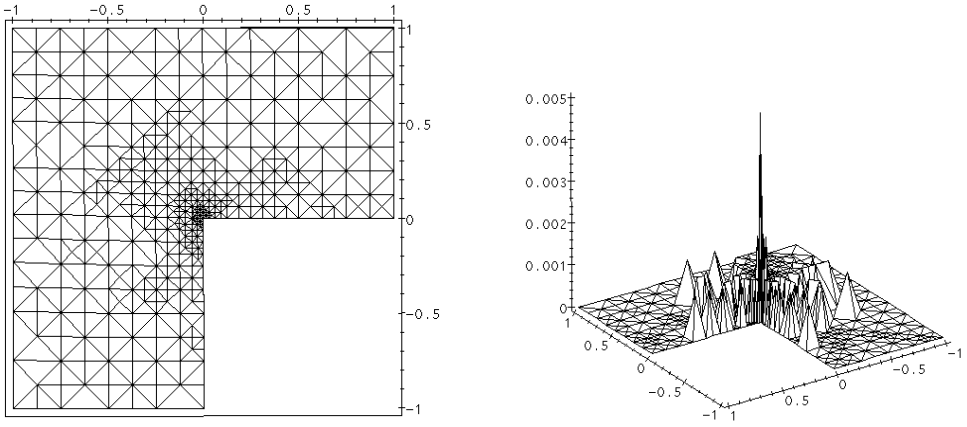


FIG. 4. *Locally refined mesh and the corresponding error after four levels of refinement.*

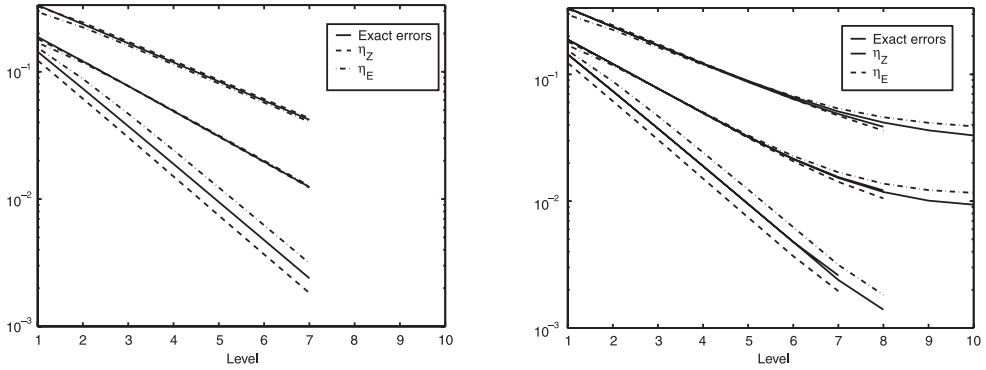


FIG. 5. *Comparison of the H^1-norm of the error for solutions $H^{1+4/3-\epsilon}$ (Problem 1), $H^{1+2/3-\epsilon}$ (Problem 2), and $H^{1+1/2-\epsilon}$ (Problem 3) on a sequence of uniformly refined grids and for grids refined locally by using the a posteriori error estimates. Left: Exact error, η_Z, and η_E for uniformly refined grids. Right: Exact error, η_Z, and η_E for locally refined grids.*

The error tolerances supplied to the refinement procedures are 0.0026 for Problem 1, 0.0122 for Problem 2, and 0.0385 for Problem 3. These are the exact errors for the problems considered on level 7 of the uniformly refined mesh. The result shows that, on the locally refined meshes, as in the uniform refinement case, both η_Z and η_E are equivalent to the exact error. Another observation is that, although the meshes are refined, only locally is the rate of error reduction over the refinement levels the same as on the uniformly refined meshes (compare the error reduction slopes with the ones in Figure 5, left).

Finally, we demonstrate the efficiency of the adaptive error control by giving the number of the degrees of freedom (DOF) on the locally refined mesh levels from Figure 5, right, and comparing them with the number of DOF on the uniformly refined mesh levels (see Table 5.1). Note the difference in the order of the mesh sizes for uniform refinement and local refinement for Problems 2 and 3. For Problem 1 we have full elliptic regularity, and η_Z/η_E are supposed to lead to uniform refinement, which is confirmed by the numerical experiment. The results demonstrate the efficiency of

TABLE 5.1

Number of DOF for the levels resulting from local refinement based on the η_Z and η_E error estimators. The error tolerances supplied to the refinement procedures are 0.0026 *for Problem* 1, 0.0122 *for Problem* 2, *and* 0.0385 *for Problem* 3 (*see Example* 1).

Level	Uniform mesh	Problem 1		Problem 2		Problem 3	
		η_Z	η_E	η_Z	η_E	η_Z	η_E
1	65	65	65	65	65	65	65
2	255	225	225	225	225	213	175
3	833	833	833	815	805	467	375
4	3201	3201	3201	2025	2080	940	695
5	12545	12545	12545	3990	4219	1461	1033
6	49665	49665	49665	5879	6249	1889	1357
7	197633	169618	197626	7322	7815	2183	1634
8			581852	8365	9034	2508	1776
9					9793		1892
10					10097		1986

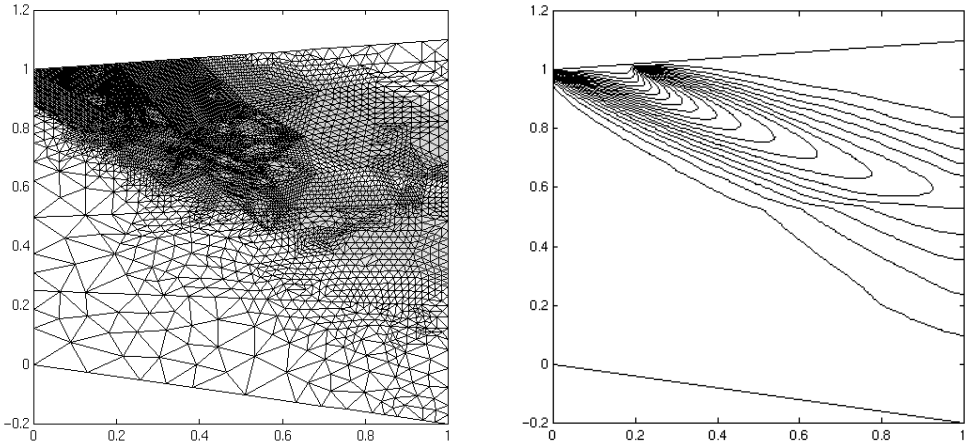


FIG. 6. *Convection-diffusion problem; the inhomogeneities are represented by three layers. Left: The locally refined mesh after four levels of adaptive refinement* (3,032 *nodes and* 5,910 *triangles*). *Right: The level curves of the solution.*

applying local refinement based on η_Z and η_E for problems with singular solutions.

Example 2. We consider problem (1.1) with Ω shown in Figure 6. In this problem Γ_D is the upper boundary, $\underline{b} = (1, -0.5)$, and $f = 0$. The domain is taken to have three layers (see Figure 6) with $A(x) = 0.01\ I$ in the top layer, $0.05\ I$ in the internal layer, and $0.001\ I$ at the bottom. The Dirichlet boundary value is 1 for $x < 0.2$ and 0 otherwise. On the Neumann boundary we take $g = 0$. In this problem we have used the upwind approximation (2.11) and the local refinement procedures based on η_Z and η_E.

Since the exact solution is not known we judge the quality of the error estimators η_Z and η_E by comparing the results with the ones on uniformly refined meshes. Also, when choosing problems with known local behavior we expect the estimators to lead to refinement that closely follows the local behavior of the solution profile. This is a standard testing approach (see, for example, [4]).

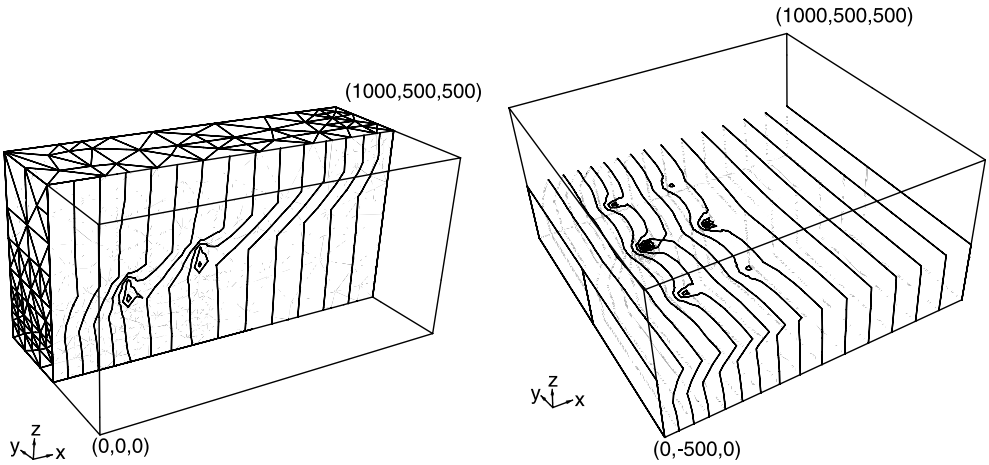


FIG. 7. *Pressure computations for a nonhomogeneous reservoir. Left: Contour curves of the pressure for the cross section* $x_2 = 250$. *Right: Contour curves of the pressure for the cross section* $x_3 = 200$.

Figure 6 shows the mesh on level 4 (left) with 3,032 nodes and 5,910 triangles. On the right are the solution level curves. This particular mesh was obtained by refinement based on η_Z with $\rho = 4\%$ of $|u_h|_1$ (≈ 0.1616, i.e., $\rho = 0.006464$). The mesh obtained by four levels of uniform refinement has 38,257 DOF. The discrete solutions have the same qualitative behavior in both cases. As expected, the mesh refinement follows the discrete solution profile. Refinement based on η_E, compared to η_Z, leads to slightly different, but qualitatively and quantitatively similar, meshes.

5.2. 3-D problems of flow and transport in porous media. This test is very similar to the 2-D Example 2. Here we test the error estimators η_Z and η_E on a real 3-D application in fluid flow and transport in porous media. Again, the exact solution is unknown but we know its local behavior, which is due to boundary layers, discontinuities of coefficients, and localized sources. The problem is described as follows.

A steady-state flow, with Darcy velocity $\underline{v}$ measured in ft/yr, has been established in a parallelepiped-shaped reservoir $\Omega = [0, 1000] \times [-500, 500] \times [0, 500]$ (see Figure 7, right). First, we determine the pressure $p(x)$ in Ω as the solution $u(x)$ of problem (1.1) with $\underline{b} = 0$, $\gamma = 0$, and $A(x) = D(x)$, where $D(x)$ is the permeability tensor. The pressure at faces $x_1 = 0$ and $x_1 = 1000$ is constant (correspondingly, 2,000 and 0). The rest of the boundary is subject to a no-flow condition. We take the permeability $D(x)$ to be 32 I everywhere in Ω except in the layer (see Figure 7, middle) where $D(x)$ is taken to be 10 times smaller than in the rest of the domain, i.e., in the layer $D(x) = 3.2\ I$.

Also, we have six production wells. For all of them x_3 is in the range $0, \ldots, 400$. Their (x_1, x_2) coordinates are correspondingly (200, -250), (400, -250), (200, 0), (400, 0), (200, 250), and (400, 250). We treat a well simply as a line-delta function (sink) along the well axis. Production rates $Q = 16{,}000$ l/yr for wells in plane $x_2 = 0$, and $Q = 8{,}000$ l/yr for the rest, are the intensities of the sink. Figure 7 shows half of the mesh and the contour curves of the pressure for the cross section $x_2 = 250$ (left) after five levels of local refinement. It has 19,850 tetrahedrons and 3,905 nodes. The right picture shows the contour curves for the cross section $x_3 = 200$.

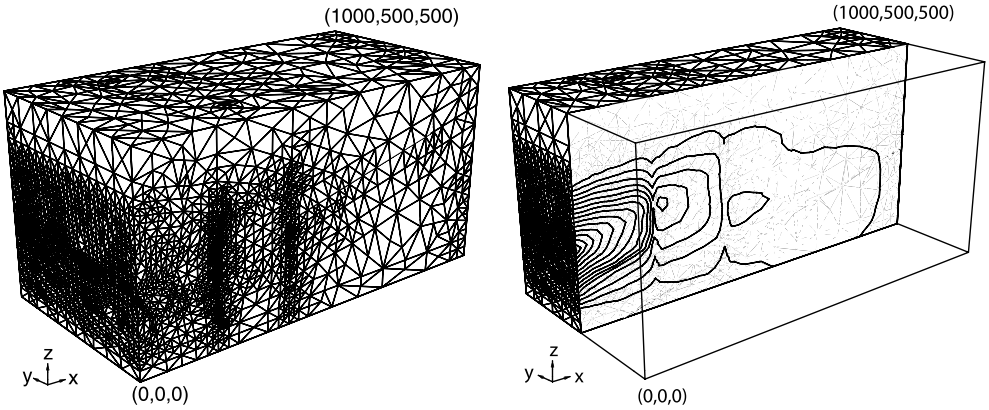


FIG. 8. *Concentration computations for a nonhomogeneous reservoir. Left: The* 3*-D mesh on refinement level* 11 *with* 219,789 *tetrahedrons and* 39,752 *nodes. Right: Concentration contour curves for cross section* $x_2 = 250$.

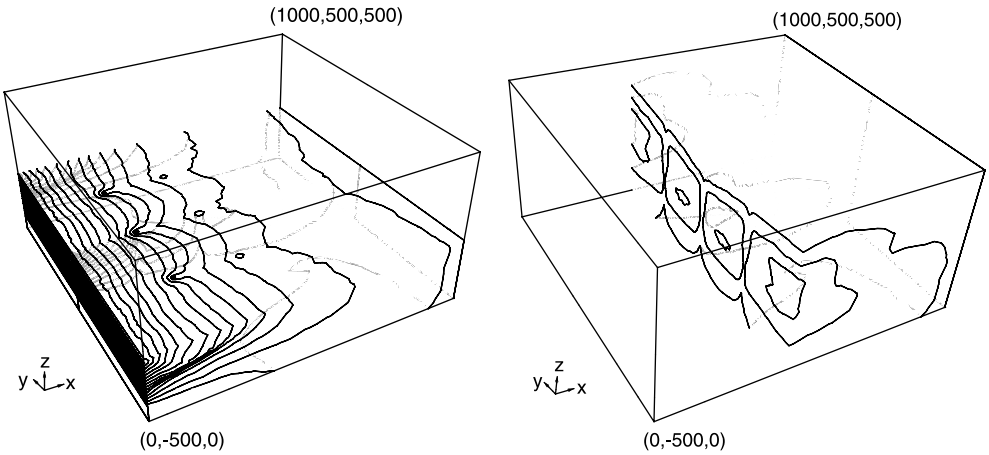


FIG. 9. *Concentration level curves at cross sections* $x_3 = 200$ (*left*) *and* $x_1 = 400$ (*right*).

The weighted pressure gradient $-D\nabla p$ forces the groundwater to flow. The transport of a contaminant dissolved in the water (in our case, benzene) is described by the convection-diffusion-reaction equation (1.1), where $u(x)$ represents the benzene concentration, $\underline{b}$ is the Darcy velocity $\underline{v} = -D\nabla p$, γ is the biodegradation rate, and $A(x)$ is the diffusion-dispersion tensor:

$$A(x) = k_{\text{diff}} I + k_t \underline{v}^T \underline{v}/|\underline{v}| + k_l(|\underline{v}|^2 I - \underline{v}^T \underline{v})/|\underline{v}|.$$

Here $k_{\text{diff}} = 0.0001$, $k_t = 21$, and $k_l = 2.1$ are the coefficients of diffusion, transverse, and longitudinal dispersions, respectively. A steady piecewise linear in x_3 and constant in x_2 leakage of benzene of maximum 30 mg/l is applied on the boundary strip $x_1 = 0$ and $50 \le x_3 \le 350$. The leakage is 30 mg/l at $x_3 = 200$ and drops linearly to 0 at $x_3 =$ 50 and 350. The rest of the boundary is subject to a homogeneous Neumann boundary condition. The dispersion/convection process causes the dissolved benzene to disperse in the reservoir. The biodegradation transforms it into a solid substance which is absorbed by the soil. This leads to a decrease in the benzene. The computations are

for the case of low absorption rate $\gamma = 0.05$. We approximate the convection term using the upwind approximation (2.10).

Figure 8 shows the obtained mesh in half of the domain (left) on refinement level 11. The mesh has 219,789 tetrahedrons and 39,752 nodes. The first five level of refinement are for the pressure equation, the rest for the concentration. Figure 8 (right) shows the level curves for the concentration in the reservoir cross section $x_2 = 250$ on the same refinement level. Figure 9 gives the level curves at two more cross sections, $x_3 = 200$ (left) and $x_1 = 400$ (right).

REFERENCES

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

[2] L. Angermann, *Balanced a posteriori error estimates for finite-volume type discretizations of convection-dominated elliptic problems*, Computing, 55 (1995), pp. 305–324.

[3] L. Angermann, *An a-posteriori estimation for the solution of an elliptic singularly perturbed problem*, IMA J. Numer. Anal., 12 (1992), pp. 201–215.

[4] L. Angermann, P. Knabner, and K. Thiele, *An error estimator for a finite volume discretization of density driven flow in porous media*, Appl. Numer. Math., 26 (1998), pp. 179–191.

[5] D. N. Arnold, A. Mukherjee, and L. Pouly, *Locally adapted tetrahedral meshes using bisection*, SIAM J. Sci. Comput., 22 (2000), pp. 431–448.

[6] I. Babuška and W. C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[7] I. Babuska and T. Strouboulis, *The Finite Element Method and Its Reliability*, Oxford University Press, London, 2001.

[8] R. E. Bank and D. J. Rose, *Some error estimates for the box method*, SIAM J. Numer. Anal., 24 (1987), pp. 777–787.

[9] R. E. Bank and R. K. Smith, *A posteriori error estimates based on hierarchical bases*, SIAM J. Numer. Anal., 30 (1993), pp. 921–935.

[10] S. Bartels and C. Carstensen, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. Part* II: *Higher order FEM*, Math. Comp., 71 (2002), pp. 971–994.

[11] R. Becker and R. Rannacher, *A feed-back approach to error control in finite element methods: Basic analysis and examples*, East-West J. Numer. Math., 4 (1996), pp. 237–264.

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

[13] Z. Cai, *On the finite volume element method*, Numer. Math., 58 (1991), pp. 713–735.

[14] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. Part* I: *Low order conforming, nonconforming, and mixed FEM*, Math. Comp., 71 (2002), pp. 945–969.

[15] C. Carstensen and S. A. Funken, *Constants in Clément-interpolation error and residual-based a posteriori estimates in finite element methods*, East-West J. Numer. Anal., 8 (2000), pp. 153–175.

[16] C. Carstensen and S. A. Funken, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput., 21 (2000), pp. 1465–1484.

[17] C. Carstensen and S. A. Funken, *A posteriori error control in low-order finite element discretizations of incompressible stationary flow problems*, Math. Comp., 70 (2001), pp. 1353–1381.

[18] S. H. Chou and Q. Li, *Error estimates in L^2, H^1, and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach*, Math. Comp., 69 (2000), pp. 103–120.

[19] G. Dagan, *Flow and Transport in Porous Formations*, Springer-Verlag, Berlin, Heidelberg, 1989.

[20] W. Dörfler and O. Wilderotter, *An adaptive finite element method for a linear elliptic equation with variable coefficients*, ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 481–491.

[21] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[22] K. Eriksson and C. Johnson, *An adaptive finite element method for linear elliptic problems*, Math. Comp., 50 (1988), pp. 361–382.

[23] R. Eymard, T. Gallouët, and R. Herbin, *Finite Volume Methods*, in Handbook of Numerical Analysis VII, North-Holland, Amsterdam, 2000, pp. 713–1020.
[24] T. Ikeda, *Maximum Principle in Finite Element Models for Convection-Diffusion Phenomena*, Lecture Notes in Numer. Appl. Anal. 4, North-Holland Math. Stud. 6, Kinokuniya Book Store Co., Ltd., Tokyo, 1983.
[25] R. D. Lazarov and S. Z. Tomov, *A posteriori error estimates for finite volume element approximations of convection-diffusion-reaction equations*, Comput. Geosci., 6 (2002), pp. 483–503. Available online at http://www.isc.tamu.edu/iscpubs/0107.ps.
[26] R. H. Li, Z. Y. Chen, and W. Wu, *Generalized Difference Method for Differential Equations. Numerical Analysis of Finite Volume Methods*, Marcel Dekker, Inc., New York, Basel, 2000.
[27] I. D. Mishev, *Finite volume methods on Voronoi meshes*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 193–212.
[28] L. Oganesian and V. L. Ruhovetz, *Variational-Difference Methods for Solving Elliptic Equations*, Publishing House of Armenian Academy of Sciences, Erevan, Armenia, 1979.
[29] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Clarendon Press, Oxford, UK, 1999.
[30] R. Rodriguez, *Some remarks on Zienkiewicz-Zhu estimator*, Numer. Methods Partial Differential Equations, 10 (1994), pp. 625–635.
[31] H.-O. Ross, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.
[32] R. Rannacher and R. Scott, *Some optimal error estimates for piecewise linear element approximations*, Math. Comp., 38 (1982), pp. 437–445.
[33] A. A. Samarskii, *The Theory of Difference Schemes*, Marcel Dekker, Inc., New York, 2001.
[34] M. Tabata, *A finite element approximation corresponding to the upwind finite differencing*, Mem. Numer. Math., 4 (1977), pp. 47–63.
[35] K. Thiele, *Adaptive Finite Volume Discretization of Density Driven Flows in Porous Media*, Ph.D. Dissertation, Naturwissenschaftliche Facultät I, Universität Erlangen-Nürnberg, Nuremberg, Germany, 1999.
[36] R. Verfürth, *A posteriori error estimation and adaptive mesh-refinement techniques*, J. Comput. Appl. Math., 50 (1994), pp. 67–83.
[37] O. C. Zienkiewicz and J. Z. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, Int. J. Numer. Methods Engrg., 24 (1987), pp. 337–357.
[38] O. C. Zienkiewicz and J. Z. Zhu, *Adaptivity and mesh generation*, Int. J. Numer. Methods Engrg., 32 (1991), pp. 783–810.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2522–2568

FLOQUET THEORY AS A COMPUTATIONAL TOOL*

GERALD MOORE†

Abstract. We describe how classical Floquet theory may be utilized, in a continuation framework, to construct an efficient Fourier spectral algorithm for approximating periodic orbits. At each continuation step, only a single square matrix, whose size equals the dimension of the phase-space, needs to be factorized; the rest of the required numerical linear algebra just consists of back-substitutions with this matrix. The eigenvalues of this key matrix are the Floquet exponents, whose crossing of the imaginary axis indicates bifurcation and change-in-stability. Hence we also describe how the new periodic orbits created at a period-doubling bifurcation point may be efficiently computed using our approach.

Key words. Floquet, Fourier, periodic orbit, continuation, period doubling

AMS subject classifications. 37C27, 37G15, 65L10, 37M20, 65P30, 65T40, 65T50

DOI. 10.1137/S0036142903434175

1. Introduction.

Floquet theory is the mathematical theory of linear, periodic systems of ordinary differential equations (ODEs) and as such appears in every standard book on ODEs, e.g., [1, 20]. (We especially recommend, however, the extensive elementary discussion in [19].) In this paper we wish to utilize Floquet theory in order to efficiently compute approximations to periodic orbits of nonlinear autonomous systems. The basic equations defining a periodic orbit are nonlinear, but applying a Newton-like method for their solution will lead to linear, periodic systems. It is for this reason that Floquet theory is so important for us.

The nonlinear, autonomous system we shall consider is

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t), \lambda), \qquad \boldsymbol{F} : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}^n; \tag{1.1}$$

i.e., $\boldsymbol{F}$ is a smooth function on $\mathbb{R}^n$ and depends on a parameter λ. For the rest of this section (and sections 2 and 4), however, we shall temporarily consider

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t)), \qquad \boldsymbol{F} : \mathbb{R}^n \mapsto \mathbb{R}^n. \tag{1.2}$$

The simplest solutions of (1.2) are the stationary points $\boldsymbol{x}^\star \in \mathbb{R}^n$ defined by

$$\boldsymbol{F}(\boldsymbol{x}^\star) = \boldsymbol{0}.$$

We will only be interested in stationary points at which periodic orbits are created, i.e., the famous Hopf bifurcation points considered in section 3. It is the next simplest solution of (1.2) that this paper is concerned with.

DEFINITION. *$\boldsymbol{u}^\star : \mathbb{R} \mapsto \mathbb{R}^n$ is a* periodic orbit *for* (1.2)*, with (minimal) period $2\pi T^\star > 0$, if*

$$\begin{aligned} &\dot{\boldsymbol{u}}^\star(t) = \boldsymbol{F}(\boldsymbol{u}^\star(t)) \quad \forall t \in \mathbb{R}, \\ &\boldsymbol{u}^\star(0) = \boldsymbol{u}^\star(2\pi T^\star), \\ &\boldsymbol{u}^\star(t) \neq \boldsymbol{u}^\star(0) \quad \forall t \in (0, 2\pi T^\star). \end{aligned}$$

*Received by the editors August 28, 2003; accepted for publication (in revised form) June 19, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/sinum/42-6/43417.html
†Department of Mathematics, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London SW7 2AZ, UK (g.moore@imperial.ac.uk).

In order to set up an appropriate set of equations for computing a periodic orbit, two key facts should be kept in mind:

- $T^\star$ that defines the period is also unknown;
- for any $c \in \mathbb{R}$, $\boldsymbol{u}^\star(t+c)$ describes the "same" periodic orbit.

(More precisely, $\boldsymbol{u}^\star(t)$ and $\boldsymbol{u}^\star(t+c)$ differ only by *phase*.)

The problem of working with an unknown period is dealt with by defining

$$\boldsymbol{v}^\star(\theta) \equiv \boldsymbol{u}^\star(T^\star\theta).$$

Hence $\boldsymbol{v}^\star$ has period 2π and satisfies

$$\dot{\boldsymbol{v}}^\star(\theta) = T^\star \boldsymbol{F}(\boldsymbol{v}^\star(\theta)).$$

Thus we should solve

$$\dot{\boldsymbol{v}}(\theta) = T\boldsymbol{F}(\boldsymbol{v}(\theta)), \qquad \boldsymbol{v}(0) = \boldsymbol{v}(2\pi) \tag{1.3}$$

for both the function $\boldsymbol{v}$ and the scalar T. Since there is an extra unknown, i.e., T, we must have an extra scalar equation. This fits in with the fact that $\boldsymbol{v}^\star(\theta+c)$ is a solution for any $c \in \mathbb{R}$. The modern way of "fixing the phase," i.e., constructing an extra scalar equation which determines a unique c, is as follows.

- We assume that we know a nearby periodic orbit $\boldsymbol{v}^0(\theta)$ of period 2π. This is a natural assumption to make in a continuation framework.
- We *fix* the phase by seeking the value of c which makes $\boldsymbol{v}^\star(\theta+c)$ as close as possible to $\boldsymbol{v}^0(\theta)$, e.g.,

$$\min_{c\in\mathbb{R}} \int_0^{2\pi} \|\boldsymbol{v}^\star(\theta+c) - \boldsymbol{v}^0(\theta)\|_2^2 \,\mathrm{d}\theta.$$

- Setting the derivative with respect to c equal to zero gives

$$\int_0^{2\pi} \boldsymbol{v}^0(\theta) \cdot \dot{\boldsymbol{v}}^\star(\theta+c) \,\mathrm{d}\theta = 0,$$

but it is convenient to integrate-by-parts and write

$$\int_0^{2\pi} \dot{\boldsymbol{v}}^0(\theta) \cdot \boldsymbol{v}^\star(\theta+c) \,\mathrm{d}\theta = 0.$$

Hence our final set of equations for a periodic orbit is

$$\begin{aligned} &\dot{\boldsymbol{v}}(\theta) = T\boldsymbol{F}(\boldsymbol{v}(\theta)), \qquad \boldsymbol{v}(0) = \boldsymbol{v}(2\pi), \\ &\int_0^{2\pi} \dot{\boldsymbol{v}}^0(\theta) \cdot \boldsymbol{v}(\theta) \,\mathrm{d}\theta = 0. \end{aligned} \tag{1.4}$$

We shall follow this idea later in section 4, but using a more appropriate inner product.

As we shall see, Floquet theory is best combined with a Fourier method to approximate periodic orbits. This leads to the question, why aren't Fourier spectral methods more popular, compared to the completely dominant collocation with piecewise-polynomials [11, 14, 23, 24]? (Of course, Fourier approximation of periodic orbits has been considered in a few papers, e.g., [9, 10], but *not* using the present approach. For example, in [30] it is suggested that an approximation to the monodromy matrix be

computed using an initial-value algorithm, which we regard as generally unacceptable.) So long as the solution one is trying to approximate is smooth, the standard advantage of spectral methods has always been their approximation power, and the standard disadvantage has been that they require the solution of linear equations with nonsparse coefficient matrices. The key task of the present paper is to emphasize that this disadvantage is *not* present when using Fourier spectral methods to approximate periodic orbits *in a continuation framework*. The underlying reason can be split into three parts.

- Fourier mode decoupling occurs for linear, constant-coefficient differential equations, as described in section 2.
- Floquet theory transforms linear, periodic differential equations into constant-coefficient form, as described in section 4.
- The Floquet variables required for this transformation can be updated efficiently as part of the continuation process, as described in section 5.

The practical difference between piecewise-polynomial and Fourier approximation becomes clear if one considers the linear, periodic system

$$-\dot{\boldsymbol{y}}(\theta) + \mathsf{A}(\theta)\boldsymbol{y}(\theta) = \boldsymbol{f}(\theta),$$

where $\boldsymbol{f} : [0, 2\pi] \mapsto \mathbb{R}^n$ and $\mathsf{A} : [0, 2\pi] \mapsto \mathbb{R}^{n\times n}$ is given periodic data, and we wish to determine a periodic solution $\boldsymbol{y} : [0, 2\pi] \mapsto \mathbb{R}^n$. Applying a piecewise-polynomial collocation method at the N mesh-points

$$0 = \theta_0 < \theta_1 < \cdots < \theta_{N-1} < \theta_N = 2\pi$$

leads to an $Nn \times Nn$ coefficient matrix with the almost block-bidiagonal structure

$$\begin{pmatrix} \square & \square & & & & \\ & \square & \square & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \square & \square \\ \square & & & & & \square \end{pmatrix}, \tag{1.5}$$

where each $\square$ is of size $n \times n$. (We assume that *condensation of local parameters* has been applied, as in [3].) With a spectral method based on Fourier modes, however, just a *single* $n \times n$ matrix needs to be factorized; all other matrix operations are just back-substitutions with quasi-upper triangular matrices. We must admit, however, that it *is* necessary to use a small multiple of Nn^2 locations for storing Floquet information, i.e., roughly the same number of nonzero elements as in (1.5).

To conclude this introduction, we describe the contents of this paper. In section 2, we show how mode-decoupling occurs for Fourier approximation of constant-coefficient problems, not only for 2π-periodic equations but also for another form of periodicity that is (in general) required by the Floquet theory. Some of these results are then applied in section 3 in order to efficiently calculate periodic orbits created at a Hopf bifurcation point. In section 4, we describe Floquet theory in the practical form that we shall utilize it. It is then immediately employed, in section 5, to advance a periodic orbit over a single continuation step. To illustrate the performance of the previous algorithms, we present some numerical results in section 6. Finally, section 7 shows how the Floquet theory makes it very easy to move onto the new periodic orbits created at a period-doubling point. For a background to many of the topics discussed in this paper, we recommend [12, 13].

2. Constant-coefficient equations. For periodic, constant-coefficient equations, Fourier analysis is especially simple because the Fourier modes decouple. We shall require two spaces of periodic functions, which we denote by $\mathcal{Y}_+$ and $\mathcal{Y}_-$, respectively. The first of these is just the usual space of 2π-periodic functions, spanned by

$$1,\ \cos\theta,\ \sin\theta,\ \cos 2\theta,\ \sin 2\theta,\ \ldots.$$

The second of these is the subspace of 4π-periodic functions which satisfy

$$y(\theta) = -y(\theta + 2\pi)$$

and is therefore spanned by

$$\cos\tfrac{\theta}{2},\ \sin\tfrac{\theta}{2},\ \cos\tfrac{3\theta}{2},\ \sin\tfrac{3\theta}{2},\ \ldots.$$

Of course, the direct sum

$$\mathcal{Y}_+ \oplus \mathcal{Y}_-$$

gives *all* 4π-periodic functions. The key reason why the space $\mathcal{Y}_-$ is important to us is that the product of two of its elements lies in $\mathcal{Y}_+$, and such products will arise naturally in section 4. Similarly, the product of an element of $\mathcal{Y}_+$ with an element of $\mathcal{Y}_-$ lies in $\mathcal{Y}_-$.

We mention here that, throughout this paper, Fourier series will be described using real trigonometric functions rather than the mathematically more elegant complex exponentials. This is solely because we wish to remain close to the practical implementation of our algorithms.

2.1. Equations in $\mathcal{Y}_+^n$. For a constant $n \times n$ matrix A, consider the linear, periodic, homogeneous differential equation

$$-\dot{\boldsymbol{z}}(\theta) + \mathsf{A}\boldsymbol{z}(\theta) = \boldsymbol{0}, \qquad \boldsymbol{z} \in \mathcal{Y}_+^n,$$

which means that each component of $\boldsymbol{z}$ is in $\mathcal{Y}_+$. If the set $\{m\mathrm{i} : m \in \mathbb{Z}\}$ does *not* contain any eigenvalue of A, then this equation only has the trivial solution $\boldsymbol{z}(\theta) \equiv \boldsymbol{0}$. Under this assumption, consider the linear, periodic, inhomogeneous differential equation

$$-\dot{\boldsymbol{z}}(\theta) + \mathsf{A}\boldsymbol{z}(\theta) = \boldsymbol{f}(\theta), \qquad \boldsymbol{z} \in \mathcal{Y}_+^n, \tag{2.1}$$

for a given $\boldsymbol{f} \in \mathcal{Y}_+^n$. If

$$\boldsymbol{f}(\theta) \equiv \boldsymbol{f}_0^c + \sum_{m=1}^{\infty} \bigl\{\boldsymbol{f}_m^c \cos m\theta + \boldsymbol{f}_m^s \sin m\theta\bigr\},$$

then the Fourier coefficients of

$$\boldsymbol{z}(\theta) \equiv \boldsymbol{z}_0^c + \sum_{m=1}^{\infty} \bigl\{\boldsymbol{z}_m^c \cos m\theta + \boldsymbol{z}_m^s \sin m\theta\bigr\}$$

are given by

$$\mathsf{A}\boldsymbol{z}_0^c = \boldsymbol{f}_0^c \tag{2.2a}$$

and

$$\begin{pmatrix} \mathsf{A} & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{A} \end{pmatrix} \begin{pmatrix} \boldsymbol{z}_m^c \\ \boldsymbol{z}_m^s \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_m^c \\ \boldsymbol{f}_m^s \end{pmatrix}, \quad m = 1, 2, \ldots. \tag{2.2b}$$

Since the eigenvalues of $\left(\begin{smallmatrix} \mathsf{A} & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{A} \end{smallmatrix}\right)$ are related to those of A through the mapping

$$\mu \to \mu \pm m\mathrm{i},$$

the above eigenvalue assumption means that $\left(\begin{smallmatrix} \mathsf{A} & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{A} \end{smallmatrix}\right)$ is nonsingular for all $m \in \mathbb{N}$.

2.2. Equations in $\mathcal{Y}_-^n$. Similarly, if we ask the question whether the linear, periodic, homogeneous differential equation

$$-\dot{\boldsymbol{z}}(\theta) + \mathsf{A}\boldsymbol{z}(\theta) = \boldsymbol{0}, \qquad \boldsymbol{z} \in \mathcal{Y}_-^n,$$

has any nontrivial solution, then the answer is that it does *not*, provided that the set $\left\{[m - \frac{1}{2}]\mathrm{i} : m \in \mathbb{Z}\right\}$ contains no eigenvalue of A. Furthermore, if $\boldsymbol{f} \in \mathcal{Y}_-^n$ is given, the linear, periodic, inhomogeneous differential equation

$$-\dot{\boldsymbol{z}}(\theta) + \mathsf{A}\boldsymbol{z}(\theta) = \boldsymbol{f}(\theta), \qquad \boldsymbol{z} \in \mathcal{Y}_-^n, \tag{2.3}$$

has a unique solution under this assumption. If

$$\boldsymbol{f}(\theta) \equiv \sum_{m=1}^{\infty} \left\{ \boldsymbol{f}_m^c \cos[m - \tfrac{1}{2}]\theta + \boldsymbol{f}_m^s \sin[m - \tfrac{1}{2}]\theta \right\},$$

then the Fourier coefficients of

$$\boldsymbol{z}(\theta) \equiv \sum_{m=1}^{\infty} \left\{ \boldsymbol{z}_m^c \cos[m - \tfrac{1}{2}]\theta + \boldsymbol{z}_m^s \sin[m - \tfrac{1}{2}]\theta \right\}$$

are given by

$$\begin{pmatrix} \mathsf{A} & -[m - \frac{1}{2}]\mathsf{I} \\ [m - \frac{1}{2}]\mathsf{I} & \mathsf{A} \end{pmatrix} \begin{pmatrix} \boldsymbol{z}_m^c \\ \boldsymbol{z}_m^s \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_m^c \\ \boldsymbol{f}_m^s \end{pmatrix}, \quad m = 1, 2, \ldots. \tag{2.4}$$

2.3. Computational algorithm in $\mathcal{Y}_+^n$. Now we show how to efficiently compute spectral approximations to the solution of (2.1). This is achieved by restricting to a finite number of Fourier modes ($m \le M$) and using accurate numerical quadrature to approximate the Fourier coefficients of $\boldsymbol{f}$.

Thus $\boldsymbol{f}$ in (2.1) is approximated by

$$\boldsymbol{f}(\theta) \approx \tilde{\boldsymbol{f}}_0^c + \sum_{m=1}^{M} \left\{ \tilde{\boldsymbol{f}}_m^c \cos m\theta + \tilde{\boldsymbol{f}}_m^s \sin m\theta \right\}, \tag{2.5}$$

where

- $\tilde{\boldsymbol{f}}_0^c \equiv \dfrac{1}{N} \displaystyle\sum_{j=1}^{N} \boldsymbol{f}(\theta_j),$
- $\tilde{\boldsymbol{f}}_m^c \equiv \dfrac{2}{N} \displaystyle\sum_{j=1}^{N} \boldsymbol{f}(\theta_j) \cos m\theta_j, \qquad m = 1, 2, \ldots, M,$
- $\tilde{\boldsymbol{f}}_m^s \equiv \dfrac{2}{N} \displaystyle\sum_{j=1}^{N} \boldsymbol{f}(\theta_j) \sin m\theta_j, \qquad m = 1, 2, \ldots, M,$

and $N \equiv 2M+1$ with $\theta_j = \frac{2\pi j}{N}, \quad j = 1, \ldots, N$. In other words, we are using the fact that

$$\frac{1}{2\pi}\int_0^{2\pi} f(\theta)g(\theta)\,\mathrm{d}\theta = \frac{1}{N}\sum_{j=1}^{N} f(\theta_j)g(\theta_j) \tag{2.6}$$

when f and g are both in the subspace of $\mathcal{Y}_+$ spanned by

$$1,\ \cos\theta,\ \sin\theta,\ \cos 2\theta,\ \sin 2\theta,\ \ldots\ ,\ \cos M\theta,\ \sin M\theta.$$

Also the above transformation from function values to approximate Fourier coefficients can be written in matrix form

$$\mathsf{Q}_+ \boldsymbol{g}^p = \sqrt{N/2}\,\tilde{\boldsymbol{g}}^m,$$

where

$$\begin{aligned}\boldsymbol{g}^p &\equiv (g(\theta_1), g(\theta_2), \ldots g(\theta_N))^T,\\ \tilde{\boldsymbol{g}}^m &\equiv (\sqrt{2}\tilde{g}_0^c, \tilde{g}_1^c, \tilde{g}_1^s, \ldots \tilde{g}_M^c, \tilde{g}_M^s)^T\end{aligned}$$

with

$$g(\theta) \approx \tilde{g}_0^c + \sum_{m=1}^{M}\left\{\tilde{g}_m^c \cos m\theta + \tilde{g}_m^s \sin m\theta\right\},$$

and Q_+ is the $N \times N$ orthogonal matrix with (i,j)th component

$i = 1$	odd $i \geq 3$	even i
$1/\sqrt{N}$	$\sqrt{2/N}\sin\frac{i-1}{2}\theta_j$	$\sqrt{2/N}\cos\frac{i}{2}\theta_j$

.

Thus it can be applied to the point values of each component of $\boldsymbol{f}$ to obtain (2.5). (For large M, it is more efficient to map between point values and approximate modal values of $\boldsymbol{f}$ using the fast Fourier transform [17, 29]. In this case it may be preferable to choose N to be slightly greater than $2M+1$, i.e., so as to be a highly composite integer. The standard techniques for dealing with this situation are described in [6, 29].) Now the approximate Fourier coefficients for $\boldsymbol{z}$ in (2.1),

$$\boldsymbol{z}(\theta) \approx \tilde{\boldsymbol{z}}_0^c + \sum_{m=1}^{M}\left\{\tilde{\boldsymbol{z}}_m^c \cos m\theta + \tilde{\boldsymbol{z}}_m^s \sin m\theta\right\},$$

may be computed as in (2.2). Note that this is especially efficient when A has already been reduced to Schur form [17]; i.e.,

$$\mathsf{A} = \mathsf{Q}\hat{\mathsf{U}}\mathsf{Q}^T, \tag{2.7}$$

where $\hat{\mathsf{U}} \in \mathbb{R}^{n\times n}$ is a quasi-upper triangular matrix and $\mathsf{Q} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix. Then, if $\tilde{\boldsymbol{z}}_0^c = \mathsf{Q}\hat{\boldsymbol{z}}_0^c$, $\tilde{\boldsymbol{f}}_0^c = \mathsf{Q}\hat{\boldsymbol{f}}_0^c$, and

$$\left.\begin{aligned}\tilde{\boldsymbol{z}}_m^{c/s} &= \mathsf{Q}\hat{\boldsymbol{z}}_m^{c/s}\\ \tilde{\boldsymbol{f}}_m^{c/s} &= \mathsf{Q}\hat{\boldsymbol{f}}_m^{c/s}\end{aligned}\right\}, \quad m = 1, \ldots, M,$$

we have only to solve the systems

$$\hat{\mathsf{U}}\hat{\boldsymbol{z}}_0^c = \hat{\boldsymbol{f}}_0^c \quad \text{and} \quad \begin{pmatrix} \hat{\mathsf{U}} & -m\mathsf{I} \\ m\mathsf{I} & \hat{\mathsf{U}} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{z}}_m^c \\ \hat{\boldsymbol{z}}_m^s \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{f}}_m^c \\ \hat{\boldsymbol{f}}_m^s \end{pmatrix}, \quad m = 1, \ldots, M.$$

The first is just back-substitution, starting with $\hat{u}_{nn}$ or the 2×2 block

$$\begin{pmatrix} \hat{u}_{n-1,n-1} & \hat{u}_{n-1,n} \\ \hat{u}_{n,n-1} & \hat{u}_{nn} \end{pmatrix},$$

while the second only requires solving 2×2 or 4×4 systems like

$$\begin{pmatrix} \hat{u}_{nn} & -m \\ m & \hat{u}_{nn} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \hat{u}_{n-1,n-1} & \hat{u}_{n-1,n} & -m & 0 \\ \hat{u}_{n,n-1} & \hat{u}_{nn} & 0 & -m \\ m & 0 & \hat{u}_{n-1,n-1} & \hat{u}_{n-1,n} \\ 0 & m & \hat{u}_{n,n-1} & \hat{u}_{nn} \end{pmatrix}.$$

2.4. Computational algorithm in $\mathcal{Y}_-^n$. Now we show how to efficiently compute spectral approximations to the solution of (2.3). Again this is achieved by restricting to a finite number of Fourier modes ($m \le M$) and using accurate numerical quadrature to approximate the Fourier coefficients of $\boldsymbol{f}$. Thus $\boldsymbol{f}$ in (2.3) is approximated by

$$\boldsymbol{f}(\theta) \approx \sum_{m=1}^{M} \left\{ \tilde{\boldsymbol{f}}_m^c \cos\left[m - \tfrac{1}{2}\right]\theta + \tilde{\boldsymbol{f}}_m^s \sin\left[m - \tfrac{1}{2}\right]\theta \right\}, \tag{2.8}$$

where

- $\tilde{\boldsymbol{f}}_m^c \equiv \dfrac{2}{N} \displaystyle\sum_{j=1}^{N} \boldsymbol{f}(\theta_j) \cos\left[m - \tfrac{1}{2}\right]\theta_j, \qquad m = 1, 2, \ldots, M,$
- $\tilde{\boldsymbol{f}}_m^s \equiv \dfrac{2}{N} \displaystyle\sum_{j=1}^{N} \boldsymbol{f}(\theta_j) \sin\left[m - \tfrac{1}{2}\right]\theta_j, \qquad m = 1, 2, \ldots, M,$

and $\theta_j = \frac{2\pi j}{N}, \quad j = 1, \ldots, N$. In other words, we are now using the fact that (2.6) also holds when f and g are both in the subspace of $\mathcal{Y}_-$ spanned by

$$\cos\tfrac{\theta}{2},\ \sin\tfrac{\theta}{2},\ \cos\tfrac{3\theta}{2},\ \sin\tfrac{3\theta}{2},\ \ldots\ ,\ \cos\left[M - \tfrac{1}{2}\right]\theta,\ \sin\left[M - \tfrac{1}{2}\right]\theta.$$

Again the above transformation from function values to approximate Fourier coefficients can be written in matrix form

$$\mathsf{Q}_- \boldsymbol{g}^p = \sqrt{N/2}\, \tilde{\boldsymbol{g}}^m,$$

where

$$\boldsymbol{g}^p \equiv (g(\theta_1), g(\theta_2), \ldots g(\theta_N))^T,$$
$$\tilde{\boldsymbol{g}}^m \equiv (\tilde{g}_1^c, \tilde{g}_1^s, \ldots \tilde{g}_M^c, \tilde{g}_M^s)^T$$

with

$$g(\theta) \approx \sum_{m=1}^{M} \left\{ \tilde{g}_m^c \cos\left[m - \tfrac{1}{2}\right]\theta + \tilde{g}_m^s \sin\left[m - \tfrac{1}{2}\right]\theta \right\},$$

and Q_- is the $2M \times N$ matrix with orthonormal rows and (i,j)th component

i even	i odd
$\sqrt{2/N}\sin\frac{i-1}{2}\theta_j$	$\sqrt{2/N}\cos\frac{i}{2}\theta_j$

;

the inverse transformation similarly being

$$\boldsymbol{g}^p = \sqrt{N/2}\,\mathsf{Q}_-^T\,\tilde{\boldsymbol{g}}^m.$$

Thus it can be applied to the point values of each component of $\boldsymbol{f}$ to obtain (2.8). (Again, for large M, the mapping between point values and approximate modal values of $\boldsymbol{f}$ is more efficiently carried out by a variant of the fast Fourier transform.) Now the approximate Fourier coefficients for $\boldsymbol{z}$ in (2.3),

$$\boldsymbol{z}(\theta) \approx \sum_{m=1}^{M} \left\{\tilde{\boldsymbol{z}}_m^c \cos[m-\tfrac{1}{2}]\theta + \tilde{\boldsymbol{z}}_m^s \sin[m-\tfrac{1}{2}]\theta\right\},$$

may be computed as in (2.4) above. Note that this is again especially efficient when A has already been reduced to Schur form as in (2.7). Then, if

$$\left.\begin{aligned}\tilde{\boldsymbol{z}}_m^{c/s} &= \mathsf{Q}\hat{\boldsymbol{z}}_m^{c/s}\\ \tilde{\boldsymbol{f}}_m^{c/s} &= \mathsf{Q}\hat{\boldsymbol{f}}_m^{c/s}\end{aligned}\right\}, \quad m = 1,\ldots,M,$$

we have only to solve the systems

$$\begin{pmatrix}\hat{\mathsf{U}} & -[m-\frac{1}{2}]\mathsf{I}\\ [m-\frac{1}{2}]\mathsf{I} & \hat{\mathsf{U}}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{z}}_m^c\\ \hat{\boldsymbol{z}}_m^s\end{pmatrix} = \begin{pmatrix}\hat{\boldsymbol{f}}_m^c\\ \hat{\boldsymbol{f}}_m^s\end{pmatrix}, \quad m = 1,\ldots,M.$$

This just involves back-substitution, e.g., starting with the 2×2 system

$$\begin{pmatrix}\hat{u}_{nn} & -[m-\frac{1}{2}]\\ [m-\frac{1}{2}] & \hat{u}_{nn}\end{pmatrix}$$

or the 4×4 system

$$\begin{pmatrix}\hat{u}_{n-1,n-1} & \hat{u}_{n-1,n} & -[m-\frac{1}{2}] & 0\\ \hat{u}_{n,n-1} & \hat{u}_{nn} & 0 & -[m-\frac{1}{2}]\\ [m-\frac{1}{2}] & 0 & \hat{u}_{n-1,n-1} & \hat{u}_{n-1,n}\\ 0 & [m-\frac{1}{2}] & \hat{u}_{n,n-1} & \hat{u}_{nn}\end{pmatrix}.$$

3. Hopf bifurcation. In this section we look at a straightforward application of the ideas in sections 2.1 and 2.3. We were surprised that this does not seem to have appeared in the literature before, the closest example we have found being [16], which does not, however, take full advantage of the mode-decoupling.

Hopf bifurcation refers to the creation of small amplitude periodic orbits at a particular point on a curve of stationary points. For the parameter-dependent equation (1.1), we shall consider a stationary point $(\boldsymbol{x}^\star, \lambda^\star)$ satisfying the following properties.[1]

[1]For simplicity, we assume that the curve of stationary points is parametrizable by λ; cf. [18].

(a) $\boldsymbol{F}(\boldsymbol{x}^\star, \lambda^\star) = \boldsymbol{0}$ and the Jacobian matrix $\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)$ is nonsingular. Thus the implicit function theorem tells us that there is a locally unique curve of stationary points passing through $(\boldsymbol{x}^\star, \lambda^\star)$. This may be parametrized by λ, and so we denote it by $(\boldsymbol{x}^\star(\lambda), \lambda)$.

(b) $\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)$ has a pair of simple purely imaginary eigenvalues

$$\pm \mathrm{i}\omega^\star, \quad \omega^\star > 0,$$

and no other eigenvalues of the form $\{m\mathrm{i}\omega^\star : m \in \mathbb{Z}\}$. Hence the right eigenvector pair $\boldsymbol{\varphi}^\star_{\Re} \pm \mathrm{i}\boldsymbol{\varphi}^\star_{\Im}$ satisfies

$$\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)[\boldsymbol{\varphi}^\star_{\Re} \pm \mathrm{i}\boldsymbol{\varphi}^\star_{\Im}] = \pm \mathrm{i}\omega^\star[\boldsymbol{\varphi}^\star_{\Re} \pm \mathrm{i}\boldsymbol{\varphi}^\star_{\Im}],$$

i.e.,

$$\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)\boldsymbol{\varphi}^\star_{\Re} = -\omega^\star \boldsymbol{\varphi}^\star_{\Im}, \qquad \mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)\boldsymbol{\varphi}^\star_{\Im} = \omega^\star \boldsymbol{\varphi}^\star_{\Re},$$

and the left eigenvector pair $\boldsymbol{\psi}^\star_{\Re} \pm \mathrm{i}\boldsymbol{\psi}^\star_{\Im}$ satisfies

$$\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)^T[\boldsymbol{\psi}^\star_{\Re} \pm \mathrm{i}\boldsymbol{\psi}^\star_{\Im}] = \mp \mathrm{i}\omega^\star[\boldsymbol{\psi}^\star_{\Re} \pm \mathrm{i}\boldsymbol{\psi}^\star_{\Im}],$$

i.e.,

$$\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)^T\boldsymbol{\psi}^\star_{\Re} = \omega^\star \boldsymbol{\psi}^\star_{\Im}, \qquad \mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)^T\boldsymbol{\psi}^\star_{\Im} = -\omega^\star \boldsymbol{\psi}^\star_{\Re}.$$

A suitable choice of normalization is that

$$\begin{aligned} \boldsymbol{\varphi}^\star_{\Re} \cdot \boldsymbol{\varphi}^\star_{\Im} &= 0, & \|\boldsymbol{\varphi}^\star_{\Re}\|_2^2 + \|\boldsymbol{\varphi}^\star_{\Im}\|_2^2 &= 1, \\ \boldsymbol{\psi}^\star_{\Re} \cdot \boldsymbol{\varphi}^\star_{\Re} &= 1, & \boldsymbol{\psi}^\star_{\Re} \cdot \boldsymbol{\varphi}^\star_{\Im} &= 0, \\ \boldsymbol{\psi}^\star_{\Im} \cdot \boldsymbol{\varphi}^\star_{\Re} &= 0, & \boldsymbol{\psi}^\star_{\Im} \cdot \boldsymbol{\varphi}^\star_{\Im} &= 1. \end{aligned}$$

(c) If the $n \times n$ matrix $\mathsf{K}^\star$ is defined by

$$\mathsf{K}^\star \equiv \frac{\mathrm{d}}{\mathrm{d}\lambda}\Big\{\mathsf{J}(\boldsymbol{x}^\star(\lambda), \lambda)\Big\}\Big|_{\lambda=\lambda^\star},$$

then

$$\boldsymbol{\psi}^\star_{\Re} \cdot \mathsf{K}^\star \boldsymbol{\varphi}^\star_{\Re} + \boldsymbol{\psi}^\star_{\Im} \cdot \mathsf{K}^\star \boldsymbol{\varphi}^\star_{\Im} \neq 0. \tag{3.1}$$

This means that, as λ moves away from $\lambda^\star$, the eigenvalues of $\mathsf{J}(\boldsymbol{x}^\star(\lambda), \lambda)$ corresponding to $\pm\mathrm{i}\omega^\star$ are no longer purely imaginary; i.e., if these eigenvalues are denoted

$$\mu^\star(\lambda) \pm \mathrm{i}\omega^\star(\lambda),$$

then (3.1) is equivalent to $\frac{\mathrm{d}\mu^\star}{\mathrm{d}\lambda}(\lambda^\star) \neq 0$, since (from [22])

$$\begin{aligned} \gamma^\star_{\Re} &\equiv \frac{\mathrm{d}\mu^\star}{\mathrm{d}\lambda}(\lambda^\star) = \frac{\boldsymbol{\psi}^\star_{\Re} \cdot \mathsf{K}^\star \boldsymbol{\varphi}^\star_{\Re} + \boldsymbol{\psi}^\star_{\Im} \cdot \mathsf{K}^\star \boldsymbol{\varphi}^\star_{\Im}}{2}, \\ \gamma^\star_{\Im} &\equiv \frac{\mathrm{d}\omega^\star}{\mathrm{d}\lambda}(\lambda^\star) = \frac{\boldsymbol{\psi}^\star_{\Re} \cdot \mathsf{K}^\star \boldsymbol{\varphi}^\star_{\Im} - \boldsymbol{\psi}^\star_{\Im} \cdot \mathsf{K}^\star \boldsymbol{\varphi}^\star_{\Re}}{2}. \end{aligned} \tag{3.2}$$

If, for λ near $\lambda^\star$, we look for a periodic orbit of (1.1) near $\boldsymbol{x}^\star$, then first we make our usual change-of-variable

$$\boldsymbol{v}(\theta) \equiv \boldsymbol{u}(T\theta),$$

which switches from $\boldsymbol{u}(t)$ with unknown period $2\pi T$ to $\boldsymbol{v}(\theta)$ with period 2π, and (1.1) transforms to

$$\dot{\boldsymbol{v}}(\theta) = T\boldsymbol{F}(\boldsymbol{v}(\theta), \lambda), \qquad \boldsymbol{v} \in \mathcal{Y}^n_+. \tag{3.3}$$

Since $\boldsymbol{F}(\boldsymbol{x}^\star, \lambda^\star) = \boldsymbol{0}$,

$$\boldsymbol{v}(\theta) \equiv \boldsymbol{x}^\star + \boldsymbol{z}(\theta), \qquad \boldsymbol{z} \in \mathcal{Y}^n_+,$$

will be an *approximate* periodic orbit (of period $2\pi T$) if $\boldsymbol{z}(\theta)$ is "small" and satisfies

$$-\dot{\boldsymbol{z}}(\theta) + T\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)\boldsymbol{z}(\theta) = \boldsymbol{0}. \tag{3.4}$$

But for $T = \frac{1}{\omega^\star}$, we know that (3.4) has solutions

$$\boldsymbol{z}(\theta) = C\boldsymbol{a}^\star(\theta + c)$$

for arbitrary constants C & c, where

$$\boldsymbol{a}^\star(\theta) \equiv \boldsymbol{\varphi}^\star_{\Re} \sin\theta + \boldsymbol{\varphi}^\star_{\Im} \cos\theta.$$

Thus we seek solutions of (3.3) in the form

$$\boldsymbol{v}(\theta) \equiv \boldsymbol{x}^\star(\lambda) + \varepsilon[\boldsymbol{a}^\star(\theta) + \boldsymbol{z}(\theta)] \tag{3.5}$$

for small nonzero ε, with

$$\frac{1}{\pi}\int_0^{2\pi} \boldsymbol{a}^\star(\theta) \cdot \boldsymbol{z}(\theta)\,\mathrm{d}\theta = 0 \tag{3.6a}$$

and

$$\frac{1}{\pi}\int_0^{2\pi} \boldsymbol{p}^\star(\theta) \cdot \boldsymbol{z}(\theta)\,\mathrm{d}\theta = 0, \tag{3.6b}$$

where

$$\boldsymbol{p}^\star(\theta) \equiv \frac{\mathrm{d}\boldsymbol{a}^\star}{\mathrm{d}\theta}(\theta) = \boldsymbol{\varphi}^\star_{\Re} \cos\theta - \boldsymbol{\varphi}^\star_{\Im} \sin\theta.$$

Equation (3.6a) fixes the amplitude of the periodic orbit, so we are using ε as a parametrization, and (3.6b) fixes the phase.

We shall soon require properties of the constant coefficient differential operator

$$-\frac{\mathrm{d}}{\mathrm{d}\theta} + \mathsf{A}^\star \tag{3.7}$$

on $\mathcal{Y}^n_+$, where

$$\mathsf{A}^\star \equiv T^\star \mathsf{J}(\boldsymbol{x}^\star, \lambda^\star) \quad \text{and} \quad T^\star \equiv \frac{1}{\omega^\star}.$$

The right null-space of (3.7) is spanned by $\boldsymbol{p}^\star(\theta)$ and $\boldsymbol{a}^\star(\theta)$, while its left null-space is spanned by

$$\tilde{\boldsymbol{a}}^\star(\theta) \equiv \boldsymbol{\psi}^\star_{\Re} \sin\theta + \boldsymbol{\psi}^\star_{\Im} \cos\theta \quad \text{and} \quad \tilde{\boldsymbol{p}}^\star(\theta) \equiv \boldsymbol{\psi}^\star_{\Re} \cos\theta - \boldsymbol{\psi}^\star_{\Im} \sin\theta.$$

If we consider the augmented linear equation

$$\begin{aligned} -\dot{\boldsymbol{z}}(\theta) + \mathsf{A}^\star \boldsymbol{z}(\theta) + \tfrac{T}{T^\star} \boldsymbol{p}^\star(\theta) + \lambda T^\star \mathsf{K}^\star \boldsymbol{a}^\star(\theta) &= \boldsymbol{0}, \\ \frac{1}{\pi} \int_0^{2\pi} \boldsymbol{a}^\star(\theta) \cdot \boldsymbol{z}(\theta)\, \mathrm{d}\theta &= 0, \\ \frac{1}{\pi} \int_0^{2\pi} \boldsymbol{p}^\star(\theta) \cdot \boldsymbol{z}(\theta)\, \mathrm{d}\theta &= 0 \end{aligned} \tag{3.8}$$

for unknowns $(\boldsymbol{z}(\theta), T, \lambda)$, as a mapping from $\mathcal{Y}^n_+ \times \mathbb{R}^2$ to itself; then (3.8) will only have the trivial solution $(\boldsymbol{0}, 0, 0)$ if the determinant of

$$\begin{bmatrix} \frac{1}{T^\star} \int_0^{2\pi} \tilde{\boldsymbol{a}}^\star(\theta) \cdot \boldsymbol{p}^\star(\theta)\, \mathrm{d}\theta & T^\star \int_0^{2\pi} \tilde{\boldsymbol{a}}^\star(\theta) \cdot \mathsf{K}^\star \boldsymbol{a}^\star(\theta)\, \mathrm{d}\theta \\ \frac{1}{T^\star} \int_0^{2\pi} \tilde{\boldsymbol{p}}^\star(\theta) \cdot \boldsymbol{p}^\star(\theta)\, \mathrm{d}\theta & T^\star \int_0^{2\pi} \tilde{\boldsymbol{p}}^\star(\theta) \cdot \mathsf{K}^\star \boldsymbol{a}^\star(\theta)\, \mathrm{d}\theta \end{bmatrix}$$

is nonzero. Since

$$\int_0^{2\pi} \tilde{\boldsymbol{a}}^\star(\theta) \cdot \boldsymbol{p}^\star(\theta)\, \mathrm{d}\theta = 0 \quad \text{and} \quad \int_0^{2\pi} \tilde{\boldsymbol{p}}^\star(\theta) \cdot \boldsymbol{p}^\star(\theta)\, \mathrm{d}\theta \neq 0,$$

this depends only on

$$\int_0^{2\pi} \tilde{\boldsymbol{a}}^\star(\theta) \cdot \mathsf{K}^\star \boldsymbol{a}^\star(\theta)\, \mathrm{d}\theta \neq 0,$$

which is equivalent to (3.1).

If the smooth mapping

$$\boldsymbol{G} : (\mathcal{Y}^n_+ \times \mathbb{R}^2) \times \mathbb{R} \mapsto \mathcal{Y}^n_+ \times \mathbb{R}^2$$

is constructed by the following:

- for $\varepsilon \neq 0$, $\boldsymbol{G}(\boldsymbol{z}, T, \lambda; \varepsilon)$ is defined by

$$\begin{aligned} -[\dot{\boldsymbol{a}}^\star(\theta) + \dot{\boldsymbol{z}}(\theta)] + \frac{1}{\varepsilon} T \boldsymbol{F}(\boldsymbol{x}^\star(\lambda) + \varepsilon[\boldsymbol{a}^\star(\theta) + \boldsymbol{z}(\theta)], \lambda) \\ \frac{1}{\pi} \int_0^{2\pi} \boldsymbol{a}^\star(\theta) \cdot \boldsymbol{z}(\theta)\, \mathrm{d}\theta \\ \frac{1}{\pi} \int_0^{2\pi} \boldsymbol{p}^\star(\theta) \cdot \boldsymbol{z}(\theta)\, \mathrm{d}\theta, \end{aligned}$$

- $\boldsymbol{G}(\boldsymbol{z}, T, \lambda; 0)$ is defined by

$$\begin{aligned} -[\dot{\boldsymbol{a}}^\star(\theta) + \dot{\boldsymbol{z}}(\theta)] + T \mathsf{J}(\boldsymbol{x}^\star(\lambda), \lambda)[\boldsymbol{a}^\star(\theta) + \boldsymbol{z}(\theta)] \\ \frac{1}{\pi} \int_0^{2\pi} \boldsymbol{a}^\star(\theta) \cdot \boldsymbol{z}(\theta)\, \mathrm{d}\theta \\ \frac{1}{\pi} \int_0^{2\pi} \boldsymbol{p}^\star(\theta) \cdot \boldsymbol{z}(\theta)\, \mathrm{d}\theta; \end{aligned}$$

then the zeroes $(\boldsymbol{z}(\theta), T, \lambda)$ of $\boldsymbol{G}$ for nonzero ε correspond to periodic orbits of (3.3) through (3.5). It is immediate, however, that $\boldsymbol{G}(\boldsymbol{0}, T^\star, \lambda^\star; 0) = \boldsymbol{0}$, and (3.8) also tells us that, at $\varepsilon = 0$, the linearization of $\boldsymbol{G}$ with respect to $(\boldsymbol{z}, T, \lambda)$ at $(\boldsymbol{0}, T^\star, \lambda^\star)$ has no nontrivial solution. Hence the implicit function theorem applies to $\boldsymbol{G}$ at $(\boldsymbol{0}, T^\star, \lambda^\star; 0)$ and tells us there is a locally unique solution curve of periodic orbits for (3.3), parametrized by ε. (We note that, from a practical point of view, it is more efficient to replace $\boldsymbol{x}^\star(\lambda)$ above with the first-order approximation

$$\boldsymbol{x}^\star + [\lambda - \lambda^\star]\boldsymbol{\ell}^\star,$$

where $\boldsymbol{\ell}^\star$ is defined by

$$\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)\boldsymbol{\ell}^\star = -\boldsymbol{F}_\lambda(\boldsymbol{x}^\star, \lambda^\star).$$

For simplicity, however, we do not include this extra trick.)

Rearranging the equation for zeroes of $\boldsymbol{G}$ enables us to define the following Newton-chord iteration for obtaining these periodic orbits.

- Choose small $\varepsilon \neq 0$ and set

$$\boldsymbol{y}^{(0)}(\theta) = \boldsymbol{a}^\star(\theta), \quad T^{(0)} = T^\star, \quad \lambda^{(0)} = \lambda^\star.$$

- Solve

$$\begin{aligned} \left[-\frac{\mathrm{d}}{\mathrm{d}\theta} + \mathsf{A}^\star\right]\boldsymbol{z}(\theta) + \tfrac{\delta T}{T^\star}\boldsymbol{p}^\star(\theta) + \delta\lambda\, T^\star \mathsf{K}^\star \boldsymbol{a}^\star(\theta) &= \tfrac{1}{\varepsilon}\boldsymbol{r}^{(k)}(\theta), \\ \frac{1}{\pi}\int_0^{2\pi} \boldsymbol{a}^\star(\theta)\cdot\boldsymbol{z}(\theta)\,\mathrm{d}\theta &= 0, \\ \frac{1}{\pi}\int_0^{2\pi} \boldsymbol{p}^\star(\theta)\cdot\boldsymbol{z}(\theta)\,\mathrm{d}\theta &= 0 \end{aligned} \tag{3.9}$$

  for $\boldsymbol{z} \in \mathcal{Y}_+^n$, δT, and $\delta\lambda$, where

$$\boldsymbol{r}^k(\theta) \equiv \dot{\boldsymbol{y}}^{(k)}(\theta) - T^{(k)}\boldsymbol{F}(\boldsymbol{x}^\star(\lambda^{(k)}) + \varepsilon\boldsymbol{y}^{(k)}(\theta), \lambda^{(k)}).$$

- Set

$$\begin{aligned} \boldsymbol{y}^{(k+1)}(\theta) &= \boldsymbol{y}^{(k)} + \boldsymbol{z}(\theta), \\ T^{(k+1)} &= T^{(k)} + \delta T, \\ \lambda^{(k+1)} &= \lambda^{(k)} + \delta\lambda. \end{aligned}$$

Note that only the same augmented constant-coefficient differential equation, with varying right-hand sides, needs to be solved at each iteration.

3.1. Fourier approximation. Finally, we show how to efficiently compute accurate approximations to the periodic orbits of (3.3), using the above Newton-chord iteration and the results of section 2.3. The key step is how to calculate the approximate Fourier coefficients

$$\boldsymbol{z}(\theta) \approx \tilde{\boldsymbol{z}}_0^c + \sum_{m=1}^{M}\left\{\tilde{\boldsymbol{z}}_m^c \cos m\theta + \tilde{\boldsymbol{z}}_m^s \sin m\theta\right\}$$

from the right-hand side

$$\begin{aligned}\boldsymbol{f}(\theta) &\equiv \tfrac{1}{\varepsilon}\boldsymbol{r}^{(k)}(\theta)\\ &\approx \tilde{\boldsymbol{f}}_0^c + \sum_{m=1}^{M}\left\{\tilde{\boldsymbol{f}}_m^c \cos m\theta + \tilde{\boldsymbol{f}}_m^s \sin m\theta\right\}\end{aligned}$$

in (3.9), and we see that matching Fourier coefficients gives the modal equations

- for $m = 0$

$$\mathsf{A}^\star \tilde{\boldsymbol{z}}_0^c = \tilde{\boldsymbol{f}}_0^c,$$

- for $m = 1$

$$\begin{bmatrix} \mathsf{A}^\star & -\mathsf{I} & \frac{1}{T^\star}\boldsymbol{\varphi}_{\Re}^\star & T^\star \mathsf{K}^\star \boldsymbol{\varphi}_{\Im}^\star \\ \mathsf{I} & \mathsf{A}^\star & -\frac{1}{T^\star}\boldsymbol{\varphi}_{\Im}^\star & T^\star \mathsf{K}^\star \boldsymbol{\varphi}_{\Re}^\star \\ \boldsymbol{\varphi}_{\Im}^{\star\,T} & \boldsymbol{\varphi}_{\Re}^{\star\,T} & 0 & 0 \\ \boldsymbol{\varphi}_{\Re}^{\star\,T} & -\boldsymbol{\varphi}_{\Im}^{\star\,T} & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{z}}_1^c \\ \tilde{\boldsymbol{z}}_1^s \\ \delta T \\ \delta\lambda \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{f}}_1^c \\ \tilde{\boldsymbol{f}}_1^s \\ 0 \\ 0 \end{bmatrix},$$

- for $2 \le m \le M$

$$\begin{bmatrix} \mathsf{A}^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{A}^\star \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{z}}_m^c \\ \tilde{\boldsymbol{z}}_m^s \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{f}}_m^c \\ \tilde{\boldsymbol{f}}_m^s \end{bmatrix}.$$

Hence the extra scalar unknowns δT and $\delta\lambda$ are solved for as part of the $m = 1$ system, while the other modal equations remain the same as in section 2.3. Thus, by applying

$$\begin{bmatrix} \boldsymbol{\psi}_{\Im}^\star \\ \boldsymbol{\psi}_{\Re}^\star \\ 0 \\ 0 \end{bmatrix}^T \qquad \text{and} \qquad \begin{bmatrix} \boldsymbol{\psi}_{\Re}^\star \\ -\boldsymbol{\psi}_{\Im}^\star \\ 0 \\ 0 \end{bmatrix}^T$$

to the $m = 1$ equation, we immediately determine δT and $\delta\lambda$ from

$$\begin{bmatrix} 0 & T^\star\gamma_{\Re}^\star \\ \frac{1}{T^\star} & T^\star\gamma_{\Im}^\star \end{bmatrix} \begin{bmatrix} \delta T \\ \delta\lambda \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \boldsymbol{\psi}_{\Im}^\star \cdot \tilde{\boldsymbol{f}}_1^c + \boldsymbol{\psi}_{\Re}^\star \cdot \tilde{\boldsymbol{f}}_1^s \\ \boldsymbol{\psi}_{\Re}^\star \cdot \tilde{\boldsymbol{f}}_1^c - \boldsymbol{\varphi}_{\Im}^\star \cdot \tilde{\boldsymbol{f}}_1^s \end{bmatrix};$$

i.e.,

$$\delta\lambda = \frac{1}{T^\star} d_\lambda \quad \text{and} \quad \delta T = T^\star d_T,$$

where

$$\begin{aligned} d_\lambda &\equiv \frac{\boldsymbol{\psi}_{\Im}^\star \cdot \tilde{\boldsymbol{f}}_1^c + \boldsymbol{\psi}_{\Re}^\star \cdot \tilde{\boldsymbol{f}}_1^s}{2\gamma_{\Re}^\star},\\ d_T &\equiv \frac{\gamma_{\Re}^\star\left(\boldsymbol{\psi}_{\Re}^\star \cdot \tilde{\boldsymbol{f}}_1^c - \boldsymbol{\varphi}_{\Im}^\star \cdot \tilde{\boldsymbol{f}}_1^s\right) - \gamma_{\Im}^\star\left(\boldsymbol{\psi}_{\Im}^\star \cdot \tilde{\boldsymbol{f}}_1^c + \boldsymbol{\psi}_{\Re}^\star \cdot \tilde{\boldsymbol{f}}_1^s\right)}{2\gamma_{\Re}^\star} \end{aligned}$$

and $\gamma_{\Re}^\star, \gamma_{\Im}^\star$ are defined in (3.2).

If $\mathsf{A}^\star$ has already been reduced to Schur form by

$$\mathsf{A}^\star \mathsf{Q}^\star = \mathsf{Q}^\star \mathsf{U}^\star,$$

where $\mathsf{U}^\star \in \mathbb{R}^{n\times n}$ is a quasi-upper triangular matrix and $\mathsf{Q}^\star \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, then it can be arranged that $\mathsf{U}^\star$ has its eigenvalues $\pm\mathrm{i}$ in the top left corner, i.e.,

$$\mathsf{U}^\star = \begin{bmatrix} 0 & \beta & \cdots & & & \\ -\beta^{-1} & 0 & \cdots & & & \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots \end{bmatrix}$$

for some nonzero β, as in the LAPACK standard form [2]. This means that

$$\boldsymbol{\varphi}^\star_{\Re} \equiv \frac{\beta}{\sqrt{1+\beta^2}}\boldsymbol{q}^\star_1 \qquad \text{and} \qquad \boldsymbol{\varphi}^\star_{\Im} \equiv \frac{1}{\sqrt{1+\beta^2}}\boldsymbol{q}^\star_2,$$

where $\boldsymbol{q}^\star_j$ denotes the jth column of $\mathsf{Q}^\star$. Hence, under the transformations $\tilde{\boldsymbol{z}}^c_0 = \mathsf{Q}^\star\hat{\boldsymbol{z}}^c_0$, $\tilde{\boldsymbol{f}}^c_0 = \mathsf{Q}^\star\hat{\boldsymbol{f}}^c_0$, and

$$\left.\begin{aligned} \tilde{\boldsymbol{z}}^{c/s}_m &= \mathsf{Q}^\star\hat{\boldsymbol{z}}^{c/s}_m \\ \tilde{\boldsymbol{f}}^{c/s}_m &= \mathsf{Q}^\star\hat{\boldsymbol{f}}^{c/s}_m \end{aligned}\right\}, \quad m = 1,\ldots,M,$$

we have only to solve the systems

- for $m = 0$

$$\mathsf{U}^\star\hat{\boldsymbol{z}}^c_0 = \hat{\boldsymbol{f}}^c_0;$$

- for $m = 1$

$$\begin{bmatrix} \mathsf{U}^\star & -\mathsf{I} \\ \mathsf{I} & \mathsf{U}^\star \\ \boldsymbol{e}_2^T & \beta\boldsymbol{e}_1^T \\ \beta\boldsymbol{e}_1^T & -\boldsymbol{e}_2^T \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{z}}^c_1 \\ \hat{\boldsymbol{z}}^s_1 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{f}}^c_1 - d_T\frac{\beta}{\sqrt{1+\beta^2}}\boldsymbol{e}_1 - d_\lambda\frac{1}{\sqrt{1+\beta^2}}\boldsymbol{c}^\star_2 \\ \hat{\boldsymbol{f}}^s_1 + d_T\frac{1}{\sqrt{1+\beta^2}}\boldsymbol{e}_2 - d_\lambda\frac{\beta}{\sqrt{1+\beta^2}}\boldsymbol{c}^\star_1 \\ 0 \\ 0 \end{bmatrix},$$

  where the components of $\boldsymbol{c}^\star_1$ and $\boldsymbol{c}^\star_2$ are the coefficients of $\mathsf{K}^\star\boldsymbol{q}^\star_1$ and $\mathsf{K}^\star\boldsymbol{q}^\star_2$, respectively, with respect to the orthonormal basis of $\mathbb{R}^n$ formed by the columns of $\mathsf{Q}^\star$, i.e.,

$$\mathsf{Q}^\star\boldsymbol{c}^\star_1 \equiv \mathsf{K}^\star\boldsymbol{q}^\star_1 \quad \text{and} \quad \mathsf{Q}^\star\boldsymbol{c}^\star_2 \equiv \mathsf{K}^\star\boldsymbol{q}^\star_2;$$

- for $2 \le m \le M$

$$\begin{bmatrix} \mathsf{U}^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{U}^\star \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{z}}^c_m \\ \hat{\boldsymbol{z}}^s_m \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{f}}^c_m \\ \hat{\boldsymbol{f}}^s_m \end{bmatrix}$$

by back-substitution. The systems for $m = 0$ and $m \ge 2$ are the same as in section 2.3 and nonsingular because of the eigenvalue conditions satisfied by $\mathsf{J}(\boldsymbol{x}^\star, \lambda^\star)$ in (b) on page 2530. The system for $m = 1$ is overdetermined, but consistent by construction,

and the last $n-2$ components for $\hat{\boldsymbol{z}}_1^{c/s}$ can again be solved by back-substitution. We are then left with the simple system

$$\begin{bmatrix} 0 & \beta & -1 & 0 \\ 1 & 0 & 0 & \beta \\ 0 & 1 & \beta & 0 \\ \beta & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \hat{z}_{11}^c \\ \hat{z}_{12}^c \\ \hat{z}_{11}^s \\ \hat{z}_{12}^s \end{bmatrix} = \begin{bmatrix} \star \\ \star \\ 0 \\ 0 \end{bmatrix},$$

where $\hat{z}_{11}^{c/s}$ and $\hat{z}_{12}^{c/s}$ refer to the first two components of $\hat{\boldsymbol{z}}_1^{c/s}$.

4. Practical Floquet theory. Now let $\mathsf{A}(\theta)$ be a 2π-periodic $n \times n$ matrix, i.e.,

$$\mathsf{A}(\theta + 2\pi) = \mathsf{A}(\theta) \qquad \forall \theta \in \mathbb{R}.$$

At first glance the periodic differential equation

$$-\dot{\boldsymbol{v}}(\theta) + \mathsf{A}(\theta)\boldsymbol{v}(\theta) = \boldsymbol{f}(\theta), \qquad \boldsymbol{v} \in \mathcal{Y}_+^n, \tag{4.1}$$

for a given $\boldsymbol{f} \in \mathcal{Y}_+^n$, seems much more difficult to analyze and solve than (2.1). It is the fundamental result of *Floquet theory*, however, that there is a change-of-variable

$$\boldsymbol{v}(\theta) = \mathsf{P}(\theta)\boldsymbol{w}(\theta)$$

which transforms (4.1) to constant-coefficient form. The price one has to pay to remain within real arithmetic, however, is that some of the components of the solution to the constant-coefficient problem may lie in $\mathcal{Y}_-$ rather than $\mathcal{Y}_+$: i.e., some of the components of $\boldsymbol{w}(\theta)$ may be in $\mathcal{Y}_-$, with the corresponding columns of the $n \times n$ matrix $\mathsf{P}(\theta)$ in $\mathcal{Y}_-^n$, but this still means that the product $\mathsf{P}(\theta)\boldsymbol{w}(\theta)$ is in $\mathcal{Y}_+^n$. In conclusion then, our constant-coefficient equations may be a combination of (2.1) and (2.3).

To see how this occurs, let $\mathsf{X}(\theta)$ be the principal fundamental solution matrix for the differential operator

$$-\frac{\mathrm{d}}{\mathrm{d}\theta} + \mathsf{A}(\theta); \tag{4.2}$$

i.e., $\mathsf{X}(0) = \mathsf{I}$ and

$$\dot{\mathsf{X}}(\theta) = \mathsf{A}(\theta)\mathsf{X}(\theta) \quad \forall \theta \in \mathbb{R},$$

and thus the jth column of $\mathsf{X}(\theta)$ solves the homogeneous initial value problem formed from (4.2), with $\boldsymbol{e}_j$ as the initial value. The columns of $\mathsf{X}(\theta)$ remain linearly independent, and so $\mathsf{X}(\theta)$ is always nonsingular. (There is, of course, no necessity for X to have any periodicity property!) $\mathsf{X}(2\pi)$ is called the *monodromy* matrix, and solutions of the fundamental algebraic eigenproblem

$$\mathsf{X}(2\pi)\boldsymbol{y} = \lambda \boldsymbol{y}$$

lead to the following three possibilities for solutions of the differential eigenproblem

$$-\dot{\boldsymbol{p}}(\theta) + \mathsf{A}(\theta)\boldsymbol{p}(\theta) = \mu \boldsymbol{p}(\theta),$$

with either $\boldsymbol{p} \in \mathcal{Y}_+^n$ or $\boldsymbol{p} \in \mathcal{Y}_-^n$.

1. For real $\lambda > 0$, we may define $\mu \in \mathbb{R}$ by $\lambda = \mathrm{e}^{2\pi\mu}$ and set

$$\boldsymbol{p}(\theta) \equiv \mathrm{e}^{-\mu\theta}\mathsf{X}(\theta)\boldsymbol{y},$$

so $\boldsymbol{p} \in \mathcal{Y}^n_+$ and

$$-\dot{\boldsymbol{p}}(\theta) + \mathsf{A}(\theta)\boldsymbol{p}(\theta) = \mu\boldsymbol{p}(\theta).$$

2. For real $\lambda < 0$, we may define $\mu \in \mathbb{R}$ by $-\lambda = \mathrm{e}^{2\pi\mu}$ and set

$$\boldsymbol{p}(\theta) \equiv \mathrm{e}^{-\mu\theta}\mathsf{X}(\theta)\boldsymbol{y},$$

so $\boldsymbol{p} \in \mathcal{Y}^n_-$ and

$$-\dot{\boldsymbol{p}}(\theta) + \mathsf{A}(\theta)\boldsymbol{p}(\theta) = \mu\boldsymbol{p}(\theta).$$

3. For a complex conjugate pair $\lambda_\Re \pm \mathrm{i}\lambda_\Im$ and $\boldsymbol{y}_\Re \pm \mathrm{i}\boldsymbol{y}_\Im$, so that

$$\mathsf{X}(2\pi)[\boldsymbol{y}_\Re, \boldsymbol{y}_\Im] = [\boldsymbol{y}_\Re, \boldsymbol{y}_\Im] \begin{bmatrix} \lambda_\Re & \lambda_\Im \\ -\lambda_\Im & \lambda_\Re \end{bmatrix},$$

we can do either of the following. (We describe below how to make a sensible choice!)

- Define $\mu \equiv \mu_\Re + \mathrm{i}\mu_\Im \in \mathbb{C}$ by $\lambda = \mathrm{e}^{2\pi\mu}$ and set

$$[\boldsymbol{p}_\Re(\theta), \boldsymbol{p}_\Im(\theta)] \equiv \mathrm{e}^{-\mu_\Re\theta}\mathsf{X}(\theta)[\boldsymbol{y}_\Re, \boldsymbol{y}_\Im] \begin{bmatrix} \cos\mu_\Im\theta & -\sin\mu_\Im\theta \\ \sin\mu_\Im\theta & \cos\mu_\Im\theta \end{bmatrix}.$$

Then $\boldsymbol{p}_\Re, \boldsymbol{p}_\Im \in \mathcal{Y}^n_+$, with $\boldsymbol{p}_\Re(0) = \boldsymbol{y}_\Re$ and $\boldsymbol{p}_\Im(0) = \boldsymbol{y}_\Im$, and

$$-\frac{\mathrm{d}}{\mathrm{d}\theta}[\boldsymbol{p}_\Re(\theta), \boldsymbol{p}_\Im(\theta)] + \mathsf{A}(\theta)[\boldsymbol{p}_\Re(\theta), \boldsymbol{p}_\Im(\theta)] = [\boldsymbol{p}_\Re(\theta), \boldsymbol{p}_\Im(\theta)] \begin{bmatrix} \mu_\Re & \mu_\Im \\ -\mu_\Im & \mu_\Re \end{bmatrix}.$$

- Define $\mu \equiv \mu_\Re + \mathrm{i}\mu_\Im \in \mathbb{C}$ by $-\lambda = \mathrm{e}^{2\pi\mu}$ and again set

$$[\boldsymbol{p}_\Re(\theta), \boldsymbol{p}_\Im(\theta)] \equiv \mathrm{e}^{-\mu_\Re\theta}\mathsf{X}(\theta)[\boldsymbol{y}_\Re, \boldsymbol{y}_\Im] \begin{bmatrix} \cos\mu_\Im\theta & -\sin\mu_\Im\theta \\ \sin\mu_\Im\theta & \cos\mu_\Im\theta \end{bmatrix}.$$

Now $\boldsymbol{p}_\Re, \boldsymbol{p}_\Im \in \mathcal{Y}^n_-$, but we still have $\boldsymbol{p}_\Re(0) = \boldsymbol{y}_\Re$ and $\boldsymbol{p}_\Im(0) = \boldsymbol{y}_\Im$, and

$$-\frac{\mathrm{d}}{\mathrm{d}\theta}[\boldsymbol{p}_\Re(\theta), \boldsymbol{p}_\Im(\theta)] + \mathsf{A}(\theta)[\boldsymbol{p}_\Re(\theta), \boldsymbol{p}_\Im(\theta)] = [\boldsymbol{p}_\Re(\theta), \boldsymbol{p}_\Im(\theta)] \begin{bmatrix} \mu_\Re & \mu_\Im \\ -\mu_\Im & \mu_\Re \end{bmatrix}.$$

The eigenvalues λ of $\mathsf{X}(2\pi)$ are called the *Floquet multipliers* for (4.2), while the corresponding μ are called the *Floquet exponents*. (This is *not quite* the standard terminology but is certainly what we require for a practical algorithm!) Note that each μ, and hence also the corresponding $\boldsymbol{p}(\theta)$, is not uniquely defined by the above construction; i.e., for any $\ell \in \mathbb{Z}$ we may set $\mu \to \mu + \mathrm{i}\ell$ and $\boldsymbol{p}(\theta) \to \boldsymbol{p}(\theta)\mathrm{e}^{\mathrm{i}\ell\theta}$. We shall always choose the size of the imaginary parts of the Floquet exponents to be as small

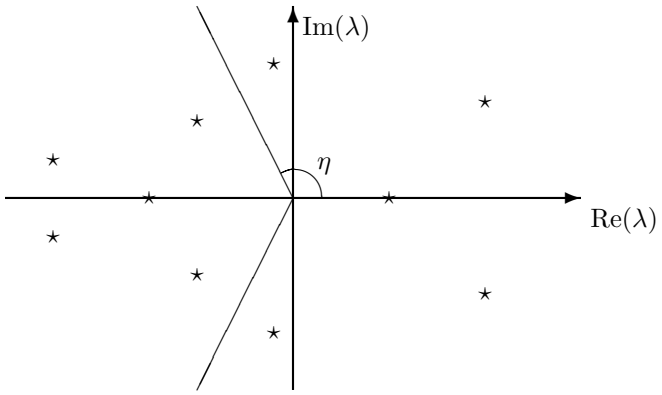


FIG. 1. *Splitting of the complex plane for Floquet multipliers.*

as possible in modulus. In conclusion, if the Floquet multipliers and exponents are denoted by

$$\lambda \equiv |\lambda| \mathrm{e}^{\mathrm{i}\theta} \quad \text{and} \quad \mu \equiv \mu_{\Re} + \mathrm{i}\mu_{\Im},$$

respectively, with $-\pi < \theta \leq \pi$, then the following table gives the mappings between them for $\mathcal{Y}_+$ and $\mathcal{Y}_-$.

$\mathcal{Y}_+$	$\lambda = \mathrm{e}^{2\pi\mu}$	$\mu_{\Re} = \frac{1}{2\pi}\ln\lvert\lambda\rvert, \quad \mu_{\Im} = \frac{\theta}{2\pi}$
$\mathcal{Y}_-$	$-\lambda = \mathrm{e}^{2\pi\mu}$	$\mu_{\Re} = \frac{1}{2\pi}\ln\lvert\lambda\rvert, \quad \mu_{\Im} = \begin{cases} \frac{\theta-\pi}{2\pi}, & \theta > 0, \\ \frac{\theta+\pi}{2\pi}, & \theta < 0 \end{cases}$

Instead of considering individual eigenvalues and eigenvectors for the monodromy matrix $\mathsf{X}(2\pi)$, the above argument can be better applied to well-conditioned invariant subspaces. Thus we choose some $0 < \eta < \pi$, such that no eigenvalue λ of $\mathsf{X}(2\pi)$ has $\arg\lambda = \eta$, and split the spectrum of $\mathsf{X}(2\pi)$ into two parts (see Figure 1), i.e.,

$$\mathsf{X}(2\pi)\mathsf{Y}_+ = \mathsf{Y}_+\Lambda_+ \quad \text{and} \quad \mathsf{X}(2\pi)\mathsf{Y}_- = \mathsf{Y}_-\Lambda_-,$$

where, for some $n_\pm \geq 0$ with $n_+ + n_- = n$, $\mathsf{Y}_+ \in \mathbb{R}^{n\times n_+}$, $\Lambda_+ \in \mathbb{R}^{n_+\times n_+}$, $\mathsf{Y}_- \in \mathbb{R}^{n\times n_-}$, $\Lambda_- \in \mathbb{R}^{n_-\times n_-}$. Here the columns of Y_+ span the invariant subspace of $\mathsf{X}(2\pi)$ corresponding to all the eigenvalues λ of $\mathsf{X}(2\pi)$ satisfying $-\eta < \arg\lambda < \eta$, while the columns of Y_- span the invariant subspace of $\mathsf{X}(2\pi)$ corresponding to all the eigenvalues λ of $\mathsf{X}(2\pi)$ satisfying $\eta - \pi < \arg[-\lambda] < \pi - \eta$. Therefore, defining $\mathsf{E}_+ \in \mathbb{R}^{n_+\times n_+}$ and $\mathsf{P}_+(\theta) \in \mathbb{R}^{n\times n_+}$ by

$$\Lambda_+ = \mathrm{e}^{2\pi\mathsf{E}_+}, \qquad \mathsf{P}_+(\theta) = \mathsf{X}(\theta)\mathsf{Y}_+\mathrm{e}^{-\mathsf{E}_+\theta}$$

and $\mathsf{E}_- \in \mathbb{R}^{n_-\times n_-}$ and $\mathsf{P}_-(\theta) \in \mathbb{R}^{n\times n_-}$ by

$$-\Lambda_- = \mathrm{e}^{2\pi\mathsf{E}_-}, \qquad \mathsf{P}_-(\theta) = \mathsf{X}(\theta)\mathsf{Y}_-\mathrm{e}^{-\mathsf{E}_-\theta},$$

we finally have

$$-\dot{\mathsf{P}}(\theta) + \mathsf{A}(\theta)\mathsf{P}(\theta) = \mathsf{P}(\theta)\mathsf{E}, \tag{4.3}$$

where

$$\mathsf{P}(\theta) = [\mathsf{P}_+(\theta), \mathsf{P}_-(\theta)]$$

and

$$\mathsf{E} = \begin{pmatrix} \mathsf{E}_+ & \mathsf{O} \\ \mathsf{O} & \mathsf{E}_- \end{pmatrix}.$$

Each column of $\mathsf{P}_+(\theta)$ is in $\mathcal{Y}_+^n$ and each column of $\mathsf{P}_-(\theta)$ is in $\mathcal{Y}_-^n$, but the $n \times n$ matrix $\mathsf{P}(\theta)$ must be nonsingular for all θ. The choice of η is not critical but, in order for our linear algebra problems to be uniformly well-posed, we shall

- keep η away from 0 and π, so that the moduli of the imaginary parts of the eigenvalues of $\mathsf{E}_\pm$ are all less than $\frac{1}{2}$,
- let η correspond to a "gap" in the arguments of the Floquet multipliers, so that the sum of the largest imaginary part of an eigenvalue from E_+ with the largest imaginary part of an eigenvalue from E_- stays below $\frac{1}{2}$.

We shall see later that it is easy to adapt η within a continuation framework for periodic orbits.

The above differential equation for P, i.e., (4.3), should be regarded as an eigenproblem. This makes it clear that there is an indeterminacy in the choice of $\mathsf{P}(\theta)$ and E, although the eigenvalues of E are invariants. Thus, for any orthogonal matrix $\mathsf{Q} \in \mathbb{R}^{n\times n}$, we have

$$-\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\mathsf{P}(\theta)\mathsf{Q}\right] + \mathsf{A}(\theta)\left[\mathsf{P}(\theta)\mathsf{Q}\right] = \left[\mathsf{P}(\theta)\mathsf{Q}\right]\left[\mathsf{Q}^T\mathsf{E}\mathsf{Q}\right],$$

and therefore we can choose

$$\mathsf{Q} \equiv \begin{pmatrix} \mathsf{Q}_+ & \mathsf{O} \\ \mathsf{O} & \mathsf{Q}_- \end{pmatrix}, \qquad \mathsf{Q}_+ \in \mathbb{R}^{n_+\times n_+}, \qquad \mathsf{Q}_- \in \mathbb{R}^{n_-\times n_-},$$

so that the transformations

$$\begin{aligned} \mathsf{P}(\theta) &\mapsto \mathsf{P}(\theta)\mathsf{Q}, \\ \mathsf{E} &\mapsto \mathsf{Q}^T\mathsf{E}\mathsf{Q} \end{aligned}$$

mean that we can assume E_+ and E_- are in real Schur (i.e., quasi-upper triangular) form. It would also be possible to ensure that

$$\frac{1}{4\pi}\int_0^{4\pi} \mathsf{P}(\theta)^T\mathsf{P}(\theta)\,\mathrm{d}\theta = \mathsf{I},$$

and so (4.3) could be regarded as a *dynamic* Schur factorization; but this is of more theoretical than practical interest.

If we now return to the problem of solving (4.1), then we can use the change-of-variable

$$\boldsymbol{v}(\theta) = \mathsf{P}(\theta)\boldsymbol{w}(\theta) \equiv \mathsf{P}(\theta)\begin{pmatrix} \boldsymbol{w}_+(\theta) \\ \boldsymbol{w}_-(\theta) \end{pmatrix},$$

where $\boldsymbol{w}_+ \in \mathcal{Y}_+^{n_+}$ and $\boldsymbol{w}_- \in \mathcal{Y}_-^{n_-}$, and obtain

$$-\mathsf{P}(\theta)\dot{\boldsymbol{w}}(\theta) + \left\{-\dot{\mathsf{P}}(\theta) + \mathsf{A}(\theta)\mathsf{P}(\theta)\right\}\boldsymbol{w}(\theta) = \boldsymbol{f}(\theta).$$

This simplifies to

$$-\dot{\boldsymbol{w}}(\theta) + \mathsf{E}\boldsymbol{w}(\theta) = \mathsf{P}(\theta)^{-1}\boldsymbol{f}(\theta),$$

and so we merely have to solve

$$\begin{aligned} -\dot{\boldsymbol{w}}_+(\theta) + \mathsf{E}_+\boldsymbol{w}_+(\theta) &= \boldsymbol{g}_+(\theta), \\ -\dot{\boldsymbol{w}}_-(\theta) + \mathsf{E}_-\boldsymbol{w}_-(\theta) &= \boldsymbol{g}_-(\theta), \end{aligned} \tag{4.4}$$

where

$$\mathsf{P}(\theta)\boldsymbol{g}(\theta) \equiv \mathsf{P}(\theta)\begin{pmatrix}\boldsymbol{g}_+(\theta)\\ \boldsymbol{g}_-(\theta)\end{pmatrix} = \boldsymbol{f}(\theta)$$

and thus $\boldsymbol{g}_+ \in \mathcal{Y}_+^{n_+}$ and $\boldsymbol{g}_- \in \mathcal{Y}_-^{n_-}$; i.e., we have reduced the problem of solving (4.1) to the simpler problem of solving (2.1) and (2.3). Equation (4.4) is nonsingular so long as E_+ has no eigenvalues of the form $m\mathrm{i}$ for $m \in \mathbb{Z}$ and E_- has no eigenvalues of the form $(m - \frac{1}{2})\mathrm{i}$ for $m \in \mathbb{Z}$. Hence, by our construction of E_+ and E_-, singularity can only occur when 0 is an eigenvalue of E_+, i.e., there is a Floquet multiplier equal to 1.

4.1. Application to periodic orbits. Just as for stationary points, it is important to distinguish between regular periodic orbits and singular ones. For stationary points $\boldsymbol{x}^\star$, we have only to look at the Jacobian matrix $\mathsf{J}(\boldsymbol{x}^\star)$: however, for a periodic orbit $\boldsymbol{u}^\star(t)$, with period $2\pi T^\star$, we must consider whether a linear differential equation has any nontrivial solutions. Looking at (1.4), we see that $(\boldsymbol{v}^\star(\theta), T^\star)$ satisfies

$$\begin{aligned} \dot{\boldsymbol{v}}(\theta) &= T\boldsymbol{F}(\boldsymbol{v}(\theta)), \qquad \boldsymbol{v} \in \mathcal{Y}_+^n, \\ \int_0^{2\pi} \dot{\boldsymbol{v}}^\star(\theta)\cdot\boldsymbol{v}(\theta)\,\mathrm{d}\theta &= 0. \end{aligned} \tag{4.5}$$

Hence we consider whether the linearization of this equation at the solution $(\boldsymbol{v}^\star, T^\star)$, i.e.,

$$\begin{aligned} -\dot{\boldsymbol{v}}(\theta) + T^\star\mathsf{J}(\boldsymbol{v}^\star(\theta))\boldsymbol{v}(\theta) + T\boldsymbol{F}(\boldsymbol{v}^\star(\theta)) &= \boldsymbol{0}, \qquad \boldsymbol{v} \in \mathcal{Y}_+^n, \\ \int_0^{2\pi} \dot{\boldsymbol{v}}^\star(\theta)\cdot\boldsymbol{v}(\theta)\,\mathrm{d}\theta &= 0, \end{aligned} \tag{4.6}$$

has any nonzero solutions $(\boldsymbol{v}(\theta), T)$.

We investigate (4.6) by applying the above Floquet theory to the linear periodic differential operator

$$-\dot{\boldsymbol{v}}(\theta) + \mathsf{A}^\star(\theta)\boldsymbol{v}(\theta), \qquad \boldsymbol{v} \in \mathcal{Y}_+^n, \tag{4.7}$$

where $\mathsf{A}^\star(\theta) \equiv T^\star\mathsf{J}(\boldsymbol{v}^\star(\theta))$; i.e., there exist $n \times n$ matrices

$$\mathsf{P}^\star(\theta) = \left[\mathsf{P}_+^\star(\theta), \mathsf{P}_-^\star(\theta)\right] \quad \text{and} \quad \mathsf{E}^\star = \begin{pmatrix}\mathsf{E}_+^\star & \mathsf{O}\\ \mathsf{O} & \mathsf{E}_-^\star\end{pmatrix}$$

so that the change-of-variable

$$\boldsymbol{v}(\theta) = \mathsf{P}^\star(\theta)\boldsymbol{w}(\theta) \equiv \mathsf{P}^\star(\theta)\begin{pmatrix}\boldsymbol{w}_+(\theta)\\ \boldsymbol{w}_-(\theta)\end{pmatrix}$$

transforms (4.6) to

$$-\dot{\boldsymbol{w}}(\theta) + \mathsf{E}^\star \boldsymbol{w}(\theta) + T\mathsf{P}^\star(\theta)^{-1}\boldsymbol{F}(\boldsymbol{v}^\star(\theta)) = \mathbf{0}, \qquad \boldsymbol{w}_\pm \in \mathcal{Y}_\pm^{n_\pm},$$
$$\int_0^{2\pi} \dot{\boldsymbol{v}}^\star(\theta) \cdot [\mathsf{P}^\star(\theta)\boldsymbol{w}(\theta)]\, \mathrm{d}\theta = 0.$$

Since we know that

$$-\dot{\boldsymbol{v}}(\theta) + T^\star \mathsf{J}(\boldsymbol{v}^\star(\theta))\boldsymbol{v}(\theta) = \mathbf{0}, \qquad \boldsymbol{v} \in \mathcal{Y}_+^n,$$

has the nontrivial solution $\boldsymbol{v}(\theta) = \dot{\boldsymbol{v}}^\star(\theta)$, we can choose the first column of $\mathsf{P}^\star(\theta)$ to be a normalization of $\dot{\boldsymbol{v}}^\star(\theta)$, i.e.,

$$\alpha^\star \mathsf{P}^\star(\theta)\boldsymbol{e}_1 \equiv \tfrac{1}{T^\star}\dot{\boldsymbol{v}}^\star(\theta)$$

for some nonzero $\alpha^\star \in \mathbb{R}$; hence, $\mathsf{E}_+^\star$ is in quasi-upper triangular form with zero first column and

$$\mathsf{P}^\star(\theta)^{-1}\boldsymbol{F}(\boldsymbol{v}^\star(\theta)) = \tfrac{1}{T^\star}\mathsf{P}^\star(\theta)^{-1}\dot{\boldsymbol{v}}^\star(\theta)$$
$$= \alpha^\star \boldsymbol{e}_1.$$

Thus to answer our question about (4.6), we only have to determine whether the much simpler problem

$$\begin{aligned} -\dot{\boldsymbol{w}}_+(\theta) + \mathsf{E}_+^\star \boldsymbol{w}_+(\theta) + T\alpha^\star \boldsymbol{e}_1 = \mathbf{0}, &\qquad \boldsymbol{w}_+ \in \mathcal{Y}_+^{n_+}, \\ -\dot{\boldsymbol{w}}_-(\theta) + \mathsf{E}_-^\star \boldsymbol{w}_-(\theta) = \mathbf{0}, &\qquad \boldsymbol{w}_- \in \mathcal{Y}_-^{n_-}, \\ \int_0^{2\pi} [\mathsf{P}^\star(\theta)\boldsymbol{e}_1] \cdot [\mathsf{P}^\star(\theta)\boldsymbol{w}(\theta)]\, \mathrm{d}\theta = 0 & \end{aligned} \tag{4.8}$$

has any nontrivial solutions $(\boldsymbol{w}(\theta), T)$. Since, by our construction of $\mathsf{E}_\pm^\star$, it is only possible for a nontrivial solution to appear in the constant Fourier mode for $\boldsymbol{w}_+$, this leads us to the following key definition.

DEFINITION. $\boldsymbol{u}^\star : \mathbb{R} \mapsto \mathbb{R}^n$ *is called a* nonsingular *or* regular *periodic orbit of* (1.2) *if zero is a* simple *eigenvalue of* $\mathsf{E}_+^\star$.

The justification for this definition is that if zero *is* a simple eigenvalue of $\mathsf{E}_+^\star$, with corresponding right and left eigenvectors $\boldsymbol{\varphi}^\star$ and $\boldsymbol{\psi}^\star$ say, then we know that the conditions [23]

$$\boldsymbol{\psi}^\star \cdot \boldsymbol{e}_1 \neq 0,$$
$$\int_0^{2\pi} [\mathsf{P}^\star(\theta)\boldsymbol{e}_1] \cdot [\mathsf{P}^\star(\theta)\boldsymbol{\varphi}^\star]\, \mathrm{d}\theta \neq 0$$

are both necessary and sufficient for (4.8) to have only the zero solution. However, since $\boldsymbol{\varphi}^\star \equiv \boldsymbol{e}_1$, both these conditions are immediately satisfied. Similarly, if zero is *not* a simple eigenvalue of $\mathsf{E}_+^\star$, then it is simple to check that (4.8) *does* have nontrivial solutions. Hence the basis of the above definition is the following conclusion.

CONCLUSION. *A necessary and sufficient condition for* (4.6) *to have only the trivial solution* $(\boldsymbol{v}, T) = (\mathbf{0}, 0)$ *is that zero is a simple eigenvalue of the matrix* $\mathsf{E}_+^\star$.

Of course, this is in complete agreement with the standard definition of a nonsingular periodic orbit, i.e., that 1 is a simple Floquet multiplier for (4.7) [1, 20].

Since, however, our algorithm in the next section works explicitly with $\mathsf{E}_+^\star$, the above definition is more appropriate for us.

Finally, it is also clear that, when using Floquet theory and working with $\boldsymbol{w}(\theta)$ instead of $\boldsymbol{v}(\theta)$, we can simplify our phase condition in (4.5); i.e., the second equation there can be replaced by

$$\frac{1}{2\pi}\int_0^{2\pi} \dot{\boldsymbol{v}}^\star(\theta)\cdot\left[\mathsf{P}^\star(\theta)\mathsf{P}^\star(\theta)^T\right]^{-1}\boldsymbol{v}(\theta)\,\mathrm{d}\theta = 0,$$

which is equivalent to the final equation of (4.8) being replaced by

$$\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{e}_1\cdot\boldsymbol{w}_+(\theta)\,\mathrm{d}\theta = 0. \tag{4.9}$$

(As we shall see in the next section, the 2π is a convenient normalization.) It is this phase condition that we shall use as part of our continuation algorithm in the next section, which connects with the fact that the bordered matrix

$$\begin{bmatrix}\mathsf{E}_+^\star & \alpha^\star\boldsymbol{e}_1\\ \boldsymbol{e}_1^T & 0\end{bmatrix}$$

is invertible for a nonsingular periodic orbit.

5. Continuation of periodic orbits. From our Floquet point of view, we describe a standard strategy for following a curve of periodic orbits for (1.1), commonly called pseudo-arclength [23]. Thus we assume that the equation

$$\dot{\boldsymbol{v}}(\theta) = T\boldsymbol{F}(\boldsymbol{v}(\theta),\lambda), \qquad \boldsymbol{v}\in\mathcal{Y}_+^n, \tag{5.1}$$

has a solution $\boldsymbol{v}^\star(\theta)$ with period $2\pi T^\star$ at $\lambda=\lambda^\star$. We also assume that the linearization

$$-\frac{\mathrm{d}}{\mathrm{d}\theta}+\mathsf{A}^\star(\theta),$$

where

$$\mathsf{A}^\star(\theta)\equiv T^\star\mathsf{J}(\boldsymbol{v}^\star(\theta),\lambda^\star),$$

has invariant subspaces defined by

$$-\dot{\mathsf{P}}^\star(\theta)+\mathsf{A}^\star(\theta)\mathsf{P}^\star(\theta)=\mathsf{P}^\star(\theta)\mathsf{E}^\star,$$

where

- $\mathsf{P}^\star(\theta)\equiv\left[\mathsf{P}_+^\star(\theta),\mathsf{P}_-^\star(\theta)\right]$ with each column of the $n\times n_+$ matrix $\mathsf{P}_+^\star$ in $\mathcal{Y}_+^n$ and each column of the $n\times n_-$ matrix $\mathsf{P}_-^\star$ in $\mathcal{Y}_-^n$;
- $\mathsf{E}^\star$ is quasi-upper triangular, with

$$\mathsf{E}^\star\equiv\begin{bmatrix}\mathsf{E}_+^\star & \mathsf{O}\\ \mathsf{O} & \mathsf{E}_-^\star\end{bmatrix},$$

  where $\mathsf{E}_+^\star\in\mathbb{R}^{n_+\times n_+}$ and $\mathsf{E}_-^\star\in\mathbb{R}^{n_-\times n_-}$;
- the sum of the largest imaginary part of an eigenvalue of $\mathsf{E}_+^\star$ with the largest imaginary part of an eigenvalue of $\mathsf{E}_-^\star$ is less than $\frac{1}{2}$;

- the first column of $\mathsf{P}^\star(\theta)$ is a normalization of $\dot{\boldsymbol{v}}^\star(\theta)$ and the first column of $\mathsf{E}^\star$ is zero.

If our periodic orbit $\boldsymbol{v}^\star(\theta)$ of period $2\pi T^\star$ is nonsingular at $\lambda = \lambda^\star$, according to the definition in section 4.1, then the implicit function theorem applies. This follows from considering the linearization of (5.1) with respect to $\boldsymbol{v}(\theta), T$ at $(\boldsymbol{v}^\star(\theta), T^\star, \lambda^\star)$ and adding the phase condition (4.9): thus the equation

$$\begin{aligned} -\dot{\boldsymbol{v}}(\theta) + \mathsf{A}^\star(\theta)\boldsymbol{v}(\theta) + T\boldsymbol{p}^\star(\theta) &= \boldsymbol{0}, \\ \frac{1}{2\pi}\int_0^{2\pi} \mathsf{P}^\star(\theta)^{-T}\boldsymbol{e}_1 \cdot \boldsymbol{v}(\theta)\,\mathrm{d}\theta &= 0 \end{aligned}$$

for $(\boldsymbol{v}(\theta), T) \in \mathcal{Y}^n_+ \times \mathbb{R}$, where

$$\boldsymbol{p}^\star(\theta) \equiv \tfrac{1}{T^\star}\dot{\boldsymbol{v}}^\star(\theta) = \boldsymbol{F}(\boldsymbol{v}^\star(\theta), \lambda^\star),$$

has only the zero solution; or equivalently, the equation

$$\begin{aligned} -\dot{\boldsymbol{w}}_+(\theta) + \mathsf{E}^\star_+\boldsymbol{w}_+(\theta) + \alpha^\star T\boldsymbol{e}_1 &= \boldsymbol{0}, \\ \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{e}_1 \cdot \boldsymbol{w}_+(\theta)\,\mathrm{d}\theta &= 0, \\ -\dot{\boldsymbol{w}}_-(\theta) + \mathsf{E}^\star_-\boldsymbol{w}_-(\theta) &= \boldsymbol{0} \end{aligned}$$

for $(\boldsymbol{w}_+(\theta), \boldsymbol{w}_-(\theta), T) \in \mathcal{Y}^{n_+}_+ \times \mathcal{Y}^{n_-}_- \times \mathbb{R}$, where

$$\alpha^\star \mathsf{P}^\star(\theta)\boldsymbol{e}_1 \equiv \boldsymbol{p}^\star(\theta),$$

has only the zero solution. Hence there is a unique curve of periodic orbits through $(\boldsymbol{v}^\star(\theta), T^\star)$, and this curve is parametrizable by λ.

We do not, however, wish to restrict ourselves to curves of periodic orbits which are parametrizable by λ. Thus we consider the full linearization of (5.1) with respect to $\boldsymbol{v}(\theta), T, \lambda$ at $(\boldsymbol{v}^\star(\theta), T^\star, \lambda^\star)$, and our basic assumption is that the equation

$$\begin{aligned} -\dot{\boldsymbol{v}}(\theta) + \mathsf{A}^\star(\theta)\boldsymbol{v}(\theta) + T\boldsymbol{p}^\star(\theta) + \lambda\boldsymbol{k}^\star(\theta) &= \boldsymbol{0}, \\ \frac{1}{2\pi}\int_0^{2\pi} \mathsf{P}^\star(\theta)^{-T}\boldsymbol{e}_1 \cdot \boldsymbol{v}(\theta)\,\mathrm{d}\theta &= 0 \end{aligned} \tag{5.2}$$

for $(\boldsymbol{v}(\theta), T, \lambda) \in \mathcal{Y}^n_+ \times \mathbb{R}^2$, where

$$\boldsymbol{k}^\star(\theta) \equiv T^\star \boldsymbol{F}_\lambda(\boldsymbol{v}^\star(\theta), \lambda^\star),$$

has a one-dimensional solution set spanned by $(\boldsymbol{v}^\star_t(\theta), T^\star_t, \lambda^\star_t)$. (This will be normalized in (5.7) below.) So the augmented equation

$$\begin{aligned} -\dot{\boldsymbol{v}}(\theta) + T\boldsymbol{F}(\boldsymbol{v}(\theta), \lambda) &= \boldsymbol{0}, \\ \frac{1}{2\pi}\int_0^{2\pi} \mathsf{P}^\star(\theta)^{-T}\boldsymbol{e}_1 \cdot \boldsymbol{v}(\theta)\,\mathrm{d}\theta &= 0, \\ \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{v}^\star_t(\theta)\cdot[\boldsymbol{v}(\theta) - \boldsymbol{v}^\star(\theta)]\,\mathrm{d}\theta & \\ + T^\star_t[T - T^\star] + \lambda^\star_t[\lambda - \lambda^\star] - \varepsilon &= 0 \end{aligned} \tag{5.3}$$

has the solution $[\boldsymbol{v}^\star(\theta), T^\star, \lambda^\star]$ for $\varepsilon = 0$; and the full linearization of (5.3) at this point gives the equation

$$\begin{aligned} -\dot{\boldsymbol{v}}(\theta) + \mathsf{A}^\star(\theta)\boldsymbol{v}(\theta) + T\boldsymbol{p}^\star(\theta) + \lambda\boldsymbol{k}^\star(\theta) &= \mathbf{0}, \\ \frac{1}{2\pi}\int_0^{2\pi} \mathsf{P}^\star(\theta)^{-T}\boldsymbol{e}_1 \cdot \boldsymbol{v}(\theta)\,\mathrm{d}\theta &= 0, \\ \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{v}_t^\star(\theta)\cdot\boldsymbol{v}(\theta)\,\mathrm{d}\theta + T_t^\star T + \lambda_t^\star\lambda &= 0, \end{aligned} \tag{5.4}$$

which has only the zero solution. Hence, the implicit function theorem tells us that (5.3) has a locally unique solution for all $|\varepsilon|$ sufficiently small, and this gives us a curve of periodic orbits $[\boldsymbol{v}(\theta), T, \lambda]$ parametrized by ε.

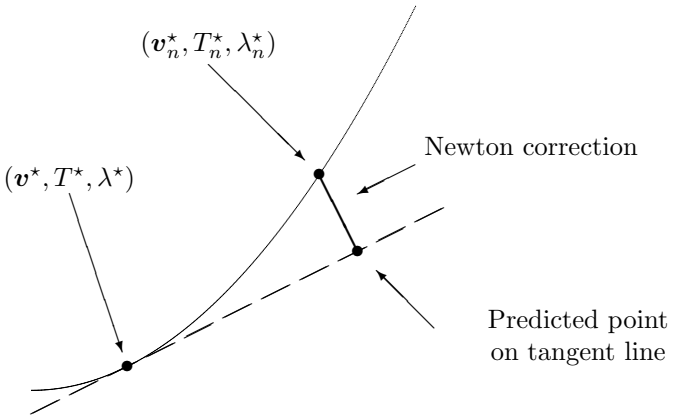


FIG. 2. *Continuing a curve of periodic orbits.*

Geometrically, our basic assumption on (5.2) is saying that the curve of periodic orbits has a unique tangent line. Thus, in section 5.1, we use our Floquet transformation to compute a Fourier approximation to this tangent line. Then, in section 5.2, we solve (5.3) using a Newton-chord iteration whose starting value is a point on this tangent line; see Figure 2. By using a simplified Newton's method, which keeps the linearization fixed, we sacrifice the quadratic convergence of the full Newton's method; this is more than compensated, however, by only having to apply our Floquet theory at $(\boldsymbol{v}^\star(\theta), T^\star, \lambda^\star)$. Thus we have an efficient algorithm for computing the Fourier approximation of a new point, $(\boldsymbol{v}_n^\star(\theta), T_n^\star, \lambda_n^\star)$ say, on the curve of periodic orbits. Having obtained this new point, we require the new Floquet variables $\mathsf{P}_n^\star(\theta), \mathsf{E}_n^\star$ there; i.e., we must efficiently update from $\mathsf{P}^\star(\theta), \mathsf{E}^\star$ to $\mathsf{P}_n^\star(\theta), \mathsf{E}_n^\star$. The algorithm for obtaining a Fourier approximation of $\mathsf{P}_n^\star(\theta), \mathsf{E}_n^\star$ is described in section 5.3. Finally, in section 5.4 we explain how the crucial bound on the size of the imaginary parts of the Floquet exponents is maintained, while in section 5.5 we show how to start the continuation method at a Hopf point.

5.1. Tangent predictor. First, we consider what the assumption (5.2) means in terms of our Floquet variables; i.e., if we set

$$\boldsymbol{v}(\theta) \equiv \mathsf{P}_+^\star(\theta)\boldsymbol{w}_+(\theta) + \mathsf{P}_-^\star(\theta)\boldsymbol{w}_-(\theta),$$

then we are assuming that

$$\begin{aligned} -\dot{\boldsymbol{w}}_+(\theta) + \mathsf{E}_+^\star \boldsymbol{w}_+(\theta) + \alpha^\star T \boldsymbol{e}_1 + \lambda \boldsymbol{k}_+^\star(\theta) &= \boldsymbol{0}, \\ \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{e}_1 \cdot \boldsymbol{w}_+(\theta)\,\mathrm{d}\theta &= 0, \\ -\dot{\boldsymbol{w}}_-(\theta) + \mathsf{E}_-^\star \boldsymbol{w}_-(\theta) + \lambda \boldsymbol{k}_-^\star(\theta) &= \boldsymbol{0}, \end{aligned} \tag{5.5}$$

where

$$\mathsf{P}^\star(\theta)\begin{bmatrix}\boldsymbol{k}_+^\star(\theta)\\ \boldsymbol{k}_-^\star(\theta)\end{bmatrix} \equiv \boldsymbol{k}^\star(\theta),$$

has a one-dimensional solution space. It is clear that this assumption rests on the equation for the constant mode of $\boldsymbol{w}_+(\theta)$; i.e., that the $(n_+ + 1)\times(n_+ + 2)$ coefficient matrix

$$\begin{bmatrix}\mathsf{E}_+^\star & \alpha^\star \boldsymbol{e}_1 & \boldsymbol{k}_0^c \\ \boldsymbol{e}_1^T & 0 & 0\end{bmatrix},$$

where

$$\boldsymbol{k}_0^c \equiv \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{k}_+^\star(\theta)\,\mathrm{d}\theta,$$

has full rank. Using the bordered matrix

$$\begin{bmatrix}\mathsf{E}_+^\star & \alpha^\star \boldsymbol{e}_1 \\ \boldsymbol{e}_1^T & 0\end{bmatrix}, \tag{5.6}$$

this can arise in three different ways.

- If (5.6) is nonsingular, then $\boldsymbol{v}^\star(\theta)$ is a nonsingular periodic orbit as described in section 4.1.
- If (5.6) has rank n_+, with $\mathsf{E}_+^\star$ having rank $n_+ - 1$, then this corresponds to zero being an eigenvalue of $\mathsf{E}_+^\star$ of geometric multiplicity one and algebraic multiplicity greater than one; in addition $\boldsymbol{k}_0^c$ is not in the range of $\mathsf{E}_+^\star$. (Generically, we would expect this zero eigenvalue of $\mathsf{E}_+^\star$ to have algebraic multiplicity two [15, 21].)
- If (5.6) has rank n_+, with $\mathsf{E}_+^\star$ having rank $n_+ - 2$, then this corresponds to zero being an eigenvalue of $\mathsf{E}_+^\star$ of geometric multiplicity two, but $\boldsymbol{e}_1$ is not in the range of $\mathsf{E}_+^\star$; in addition $\boldsymbol{k}_0^c$ is not in the range of $\mathsf{E}_+^\star$ and not parallel to $\boldsymbol{e}_1$. (Generically, we would not expect these conditions to appear [21].)

This analysis also shows that a suitable normalization for the solution of (5.2) or (5.5) is

$$\|\hat{\boldsymbol{t}}^\star\|^2 + [T_t^\star]^2 + [\lambda_t^\star]^2 = 1, \tag{5.7}$$

where $\hat{\boldsymbol{t}}^\star \in \mathbb{R}^{n_+}$ is the constant mode of $\boldsymbol{w}_+(\theta)$; i.e.,

$$\hat{\boldsymbol{t}}^\star \equiv \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{w}_+(\theta)\,\mathrm{d}\theta.$$

We shall also require $\boldsymbol{t}^\star \in \mathbb{R}^n$ later, which is just $\hat{\boldsymbol{t}}^\star$ padded out with zeroes.

Discretizing (5.5), our Fourier approximations

$$\boldsymbol{w}_+(\theta) \approx \boldsymbol{w}_0^c + \sum_{m=1}^{M} \left\{ \boldsymbol{w}_m^c \cos m\theta + \boldsymbol{w}_m^s \sin m\theta \right\},$$
$$\boldsymbol{k}_+^\star(\theta) \approx \boldsymbol{k}_0^c + \sum_{m=1}^{M} \left\{ \boldsymbol{k}_m^c \cos m\theta + \boldsymbol{k}_m^s \sin m\theta \right\}$$

give us

$$\begin{pmatrix} \mathsf{E}_+^\star & \alpha^\star \boldsymbol{e}_1 & \boldsymbol{k}_0^c \\ \boldsymbol{e}_1^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_0^c \\ T_t^\star \\ \lambda_t^\star \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ 0 \end{pmatrix} \tag{5.8}$$

for $m = 0$ and

$$\begin{pmatrix} \mathsf{E}_+^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{E}_+^\star \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_m^c \\ \boldsymbol{w}_m^s \end{pmatrix} = -\lambda_t^\star \begin{pmatrix} \boldsymbol{k}_m^c \\ \boldsymbol{k}_m^s \end{pmatrix} \tag{5.9}$$

for $m = 1, \ldots, M$; similarly, our Fourier approximations

$$\boldsymbol{w}_-(\theta) \approx \sum_{m=1}^{M} \left\{ \boldsymbol{w}_m^c \cos [m - \tfrac{1}{2}]\theta + \boldsymbol{w}_m^s \sin [m - \tfrac{1}{2}]\theta \right\},$$
$$\boldsymbol{k}_-^\star(\theta) \approx \sum_{m=1}^{M} \left\{ \boldsymbol{k}_m^c \cos [m - \tfrac{1}{2}]\theta + \boldsymbol{k}_m^s \sin [m - \tfrac{1}{2}]\theta \right\}$$

give us

$$\begin{pmatrix} \mathsf{E}_-^\star & -[m - \frac{1}{2}]\mathsf{I} \\ [m - \frac{1}{2}]\mathsf{I} & \mathsf{E}_-^\star \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_m^c \\ \boldsymbol{w}_m^s \end{pmatrix} = -\lambda_t^\star \begin{pmatrix} \boldsymbol{k}_m^c \\ \boldsymbol{k}_m^s \end{pmatrix} \tag{5.10}$$

for $m = 1, \ldots, M$. The one-dimensional null-space for the full-rank structured matrix in (5.8) may be efficiently obtained by employing algorithms from [17, Chap. 5].

- First, we construct the $n \times (n+1)$ quasi-upper triangular matrix

$$\begin{pmatrix} \alpha^\star \boldsymbol{e}_1 & \hat{\mathsf{E}}_+^\star & \boldsymbol{k}_0^c \end{pmatrix}, \tag{5.11}$$

  where $\hat{\mathsf{E}}_+^\star$ is obtained from $\mathsf{E}_+^\star$ by removing the zero first column.
- Second, Givens rotations are used to annihilate the nonzero elements below the diagonal and thus change (5.11) into an $n \times (n+1)$ upper triangular matrix.
- Third, post-multiplication with Householder matrices is used to annihilate the final column of (5.11).

We can then solve (5.9) and (5.10), since the restriction on the size of the imaginary parts of the eigenvalues of $\mathsf{E}_\pm^\star$ makes these systems nonsingular.

5.2. Newton correction. Our Newton correction will solve the system

$$\begin{aligned} -\dot{\boldsymbol{v}}(\theta) + T\boldsymbol{F}(\boldsymbol{v}(\theta), \lambda) &= \boldsymbol{0}, \\ \frac{1}{2\pi} \int_0^{2\pi} \mathsf{P}^\star(\theta)^{-T} \boldsymbol{e}_1 \cdot \boldsymbol{v}(\theta)\, \mathrm{d}\theta &= 0, \\ \frac{1}{2\pi} \int_0^{2\pi} \mathsf{P}^\star(\theta)^{-T} \boldsymbol{t}^\star \cdot [\boldsymbol{v}(\theta) - \boldsymbol{v}^\star(\theta)]\, \mathrm{d}\theta & \\ + T_t^\star [T - T^\star] + \lambda_t^\star [\lambda - \lambda^\star] - \varepsilon &= 0. \end{aligned} \tag{5.12}$$

Thus, for the same reason as in (4.9), we have replaced the final equation for the tangent line step in (5.3) by the more convenient equation above. Our Newton-chord iteration for solving (5.12) is defined from the starting values

$$\begin{aligned} \boldsymbol{v}^{(0)}(\theta) &= \boldsymbol{v}^\star(\theta) + \varepsilon \boldsymbol{v}_t^\star(\theta), \\ T^{(0)} &= T^\star + \varepsilon T_t^\star, \\ \lambda^{(0)} &= \lambda^\star + \varepsilon \lambda_t^\star \end{aligned}$$

for chosen small $|\varepsilon|$, and consists of

$$\begin{aligned} \boldsymbol{v}^{(k+1)}(\theta) &= \boldsymbol{v}^{(k)}(\theta) + \boldsymbol{z}(\theta), \\ T^{(k+1)} &= T^{(k)} + \delta T, \\ \lambda^{(k+1)} &= \lambda^{(k)} + \delta\lambda, \end{aligned}$$

where $(\boldsymbol{z}(\theta), \delta T, \delta\lambda) \in \mathcal{Y}_+^n \times \mathbb{R}^2$ satisfy

$$\begin{aligned} -\dot{\boldsymbol{z}}(\theta) + \mathsf{A}^\star(\theta)\boldsymbol{z}(\theta) + \delta T \boldsymbol{p}^\star(\theta) + \delta\lambda \boldsymbol{k}^\star(\theta) &= \boldsymbol{r}^{(k)}(\theta), \\ \frac{1}{2\pi}\int_0^{2\pi} \mathsf{P}^\star(\theta)^{-T}\boldsymbol{e}_1 \cdot \boldsymbol{z}(\theta)\,\mathrm{d}\theta &= 0, \\ \frac{1}{2\pi}\int_0^{2\pi} \mathsf{P}^\star(\theta)^{-T}\boldsymbol{t}^\star \cdot \boldsymbol{z}(\theta)\,\mathrm{d}\theta + T_t^\star \delta T + \lambda_t^\star \delta\lambda &= 0, \end{aligned} \tag{5.13}$$

with

$$\boldsymbol{r}^{(k)}(\theta) \equiv \dot{\boldsymbol{v}}^{(k)}(\theta) - T^{(k)}\boldsymbol{F}(\boldsymbol{v}^{(k)}(\theta), \lambda^{(k)}).$$

Hence, under the Floquet transformations

$$\begin{aligned} \boldsymbol{z}(\theta) &\equiv \mathsf{P}_+^\star(\theta)\boldsymbol{w}_+(\theta) + \mathsf{P}_-^\star(\theta)\boldsymbol{w}_-(\theta), \\ \boldsymbol{r}^{(k)}(\theta) &\equiv \mathsf{P}_+^\star(\theta)\boldsymbol{f}_+(\theta) + \mathsf{P}_-^\star(\theta)\boldsymbol{f}_-(\theta), \end{aligned}$$

our equation in $\mathcal{Y}_+^{n_+}$ becomes

$$\begin{aligned} -\dot{\boldsymbol{w}}_+(\theta) + \mathsf{E}_+^\star \boldsymbol{w}_+(\theta) + \delta T \alpha^\star \boldsymbol{e}_1 + \delta\lambda \boldsymbol{k}_+^\star(\theta) &= \boldsymbol{f}_+(\theta), \\ \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{e}_1 \cdot \boldsymbol{w}_+(\theta)\,\mathrm{d}\theta &= 0, \\ \frac{1}{2\pi}\int_0^{2\pi} \hat{\boldsymbol{t}}^\star \cdot \boldsymbol{w}_+(\theta)\,\mathrm{d}\theta + T_t^\star \delta T + \lambda_t^\star \delta\lambda &= 0, \end{aligned}$$

and our equation in $\mathcal{Y}_-^{n_-}$ is

$$-\dot{\boldsymbol{w}}_-(\theta) + \mathsf{E}_-^\star \boldsymbol{w}_-(\theta) + \delta\lambda \boldsymbol{k}_-^\star(\theta) = \boldsymbol{f}_-(\theta).$$

Hence our Fourier approximation in $\mathcal{Y}_+^{n_+}$ is

$$\begin{aligned} \boldsymbol{w}_+(\theta) &\approx \boldsymbol{w}_0^c + \sum_{m=1}^M \{\boldsymbol{w}_m^c \cos m\theta + \boldsymbol{w}_m^s \sin m\theta\}, \\ \boldsymbol{f}_+(\theta) &\approx \boldsymbol{f}_0^c + \sum_{m=1}^M \{\boldsymbol{f}_m^c \cos m\theta + \boldsymbol{f}_m^s \sin m\theta\}, \\ \boldsymbol{k}_+^\star(\theta) &\approx \boldsymbol{k}_0^c + \sum_{m=1}^M \{\boldsymbol{k}_m^c \cos m\theta + \boldsymbol{k}_m^s \sin m\theta\}, \end{aligned}$$

leading to

$$\begin{pmatrix} \mathsf{E}_+^\star & \alpha^\star \boldsymbol{e}_1 & \boldsymbol{k}_0^c \\ \boldsymbol{e}_1^T & 0 & 0 \\ (\hat{\boldsymbol{t}}^\star)^T & T_t^\star & \lambda_t^\star \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_0^c \\ \delta T \\ \delta\lambda \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_0^c \\ 0 \\ 0 \end{pmatrix} \tag{5.14}$$

for $m = 0$ and

$$\begin{pmatrix} \mathsf{E}_+^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{E}_+^\star \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_m^c \\ \boldsymbol{w}_m^s \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_m^c - \delta\lambda \boldsymbol{k}_m^c \\ \boldsymbol{f}_m^s - \delta\lambda \boldsymbol{k}_m^s \end{pmatrix} \tag{5.15}$$

for $m = 1, \ldots, M$; while our Fourier approximation in $\mathcal{Y}_-^{n_-}$ is

$$\begin{aligned} \boldsymbol{w}_-(\theta) &\approx \sum_{m=1}^{M} \left\{ \boldsymbol{w}_m^c \cos\left[m - \tfrac{1}{2}\right]\theta + \boldsymbol{w}_m^s \sin\left[m - \tfrac{1}{2}\right]\theta \right\}, \\ \boldsymbol{f}_-(\theta) &\approx \sum_{m=1}^{M} \left\{ \boldsymbol{f}_m^c \cos\left[m - \tfrac{1}{2}\right]\theta + \boldsymbol{f}_m^s \sin\left[m - \tfrac{1}{2}\right]\theta \right\}, \\ \boldsymbol{k}_-^\star(\theta) &\approx \sum_{m=1}^{M} \left\{ \boldsymbol{k}_m^c \cos\left[m - \tfrac{1}{2}\right]\theta + \boldsymbol{k}_m^s \sin\left[m - \tfrac{1}{2}\right]\theta \right\}, \end{aligned}$$

giving us

$$\begin{pmatrix} \mathsf{E}_-^\star & -[m - \frac{1}{2}]\mathsf{I} \\ [m - \frac{1}{2}]\mathsf{I} & \mathsf{E}_-^\star \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_m^c \\ \boldsymbol{w}_m^s \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_m^c - \delta\lambda \boldsymbol{k}_m^c \\ \boldsymbol{f}_m^s - \delta\lambda \boldsymbol{k}_m^s \end{pmatrix} \tag{5.16}$$

for $m = 1, \ldots, M$. By our construction of $(\hat{\boldsymbol{t}}^\star, T_t^\star, \lambda_t^\star)$, the coefficient matrix in (5.14) is nonsingular. Our linear algebra at the end of section 5.1 also means that we can solve (5.14) efficiently. We can then solve (5.15) and (5.16), since the restriction on the size of the imaginary parts of the eigenvalues of $\mathsf{E}_\pm^\star$ makes these systems nonsingular.

5.3. Floquet continuation. After having calculated $\boldsymbol{v}_n^\star(\theta)$, $T_n^\star$, and $\lambda_n^\star$, we need to update from

$$-\dot{\mathsf{P}}^\star(\theta) + \mathsf{A}^\star(\theta)\mathsf{P}^\star(\theta) = \mathsf{P}^\star(\theta)\mathsf{E}^\star \tag{5.17}$$

to

$$-\dot{\mathsf{P}}_n^\star(\theta) + \mathsf{A}_n^\star(\theta)\mathsf{P}_n^\star(\theta) = \mathsf{P}_n^\star(\theta)\mathsf{E}_n^\star, \tag{5.18}$$

where

$$\mathsf{A}_n^\star(\theta) \equiv T_n^\star \mathsf{J}(\boldsymbol{v}_n^\star(\theta), \lambda_n^\star).$$

We seek $\mathsf{P}_n^\star(\theta)$ in the Floquet form

$$\mathsf{P}_n^\star(\theta) \equiv \mathsf{P}^\star(\theta)\mathsf{P}(\theta),$$

with the normalization

$$\frac{1}{2\pi} \int_0^{2\pi} \mathsf{P}(\theta)\, \mathrm{d}\theta = \mathsf{I}. \tag{5.19}$$

Hence we can apply a Newton-chord method to solve (5.18) for $\mathsf{P}(\theta)$ and $\mathsf{E}_n^\star$, analogous to that described theoretically in [28] and practically in [7], and make use of our known Floquet variables for (5.17). Thus our Newton iteration starts from $\mathsf{P}^{(0)}(\theta) \equiv \mathsf{I}$ and $\mathsf{E}^{(0)} \equiv \mathsf{E}^\star$ and computes

$$\begin{aligned} \mathsf{P}^{(k+1)}(\theta) &= \mathsf{P}^{(k)}(\theta) + \mathsf{Z}(\theta), \\ \mathsf{E}^{(k+1)} &= \mathsf{E}^{(k)} + \delta\mathsf{E} \end{aligned}$$

from

$$-\dot{\mathsf{Z}}(\theta) + \mathsf{E}^\star\mathsf{Z}(\theta) - \mathsf{Z}(\theta)\mathsf{E}^\star - \delta\mathsf{E} = \mathsf{R}^{(k)}(\theta), \tag{5.20}$$

where

$$\mathsf{R}^{(k)}(\theta) \equiv \mathsf{P}^\star(\theta)^{-1}\left\{\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\mathsf{P}^\star(\theta)\mathsf{P}^{(k)}(\theta)\right] - \mathsf{A}_n^\star(\theta)\mathsf{P}^\star(\theta)\mathsf{P}^{(k)}(\theta) + \mathsf{P}^\star(\theta)\mathsf{P}^{(k)}(\theta)\mathsf{E}^{(k)}\right\}.$$

We now make use of the decompositions

$$\mathsf{E}^\star \equiv \begin{bmatrix} \mathsf{E}_+^\star & \mathsf{O} \\ \mathsf{O} & \mathsf{E}_-^\star \end{bmatrix}, \qquad \delta\mathsf{E} \equiv \begin{bmatrix} \delta\mathsf{E}_+ & \mathsf{O} \\ \mathsf{O} & \delta\mathsf{E}_- \end{bmatrix}$$

and the fact that $\mathsf{P}^\star(\theta) \equiv \left[\mathsf{P}_+^\star(\theta)|\mathsf{P}_-^\star(\theta)\right]$ leads to

$$\mathsf{Z}(\theta) \equiv \begin{bmatrix} \mathsf{Z}_{++}(\theta) & \mathsf{Z}_{+-}(\theta) \\ \mathsf{Z}_{-+}(\theta) & \mathsf{Z}_{--}(\theta) \end{bmatrix} \quad \text{and} \quad \mathsf{R}^{(k)}(\theta) \equiv \begin{bmatrix} \mathsf{R}_{++}^{(k)}(\theta) & \mathsf{R}_{+-}^{(k)}(\theta) \\ \mathsf{R}_{-+}^{(k)}(\theta) & \mathsf{R}_{--}^{(k)}(\theta) \end{bmatrix}.$$

Here $\mathsf{Z}_{++}(\theta) \in \mathbb{R}^{n_+ \times n_+}$ and $\mathsf{Z}_{--}(\theta) \in \mathbb{R}^{n_- \times n_-}$ have components in $\mathcal{Y}_+$, while $\mathsf{Z}_{+-}(\theta) \in \mathbb{R}^{n_+ \times n_-}$ and $\mathsf{Z}_{-+}(\theta) \in \mathbb{R}^{n_- \times n_+}$ have components in $\mathcal{Y}_-$, and similarly for the decomposition of $\mathsf{R}^{(k)}(\theta)$. Hence we can decompose (5.20) into two equations with components in $\mathcal{Y}_+$,

$$-\dot{\mathsf{Z}}_{++}(\theta) + \mathsf{E}_+^\star\mathsf{Z}_{++}(\theta) - \mathsf{Z}_{++}(\theta)\mathsf{E}_+^\star - \delta\mathsf{E}_+ = \mathsf{R}_{++}^{(k)}(\theta) \tag{5.21}$$

and

$$-\dot{\mathsf{Z}}_{--}(\theta) + \mathsf{E}_-^\star\mathsf{Z}_{--}(\theta) - \mathsf{Z}_{--}(\theta)\mathsf{E}_-^\star - \delta\mathsf{E}_- = \mathsf{R}_{--}^{(k)}(\theta), \tag{5.22}$$

and two equations with components in $\mathcal{Y}_-$,

$$-\dot{\mathsf{Z}}_{+-}(\theta) + \mathsf{E}_+^\star\mathsf{Z}_{+-}(\theta) - \mathsf{Z}_{+-}(\theta)\mathsf{E}_-^\star = \mathsf{R}_{+-}^{(k)}(\theta) \tag{5.23}$$

and

$$-\dot{\mathsf{Z}}_{-+}(\theta) + \mathsf{E}_-^\star\mathsf{Z}_{-+}(\theta) - \mathsf{Z}_{-+}(\theta)\mathsf{E}_+^\star = \mathsf{R}_{-+}^{(k)}(\theta). \tag{5.24}$$

Each of (5.21), (5.22), (5.23), and (5.24) can be replaced by the analogous Fourier approximation, which leads to mode-decoupling as we see below. We then obtain Sylvester equations [17], which can be solved by the Bartels–Stewart algorithm [4, 17]. Most of the work in this algorithm is devoted to reducing the appropriate matrices to Schur form, but here $\mathsf{E}_\pm^\star$ already have this form! Otherwise, only back-substitutions are required. Of course the product to form $\mathsf{R}^{(k)}(\theta)$ in (5.20) is carried out in point-space.

- To obtain $\mathsf{Z}_{++}(\theta)$, we set

$$\mathsf{Z}_{++}(\theta) \approx \sum_{m=1}^{M} \{\mathsf{Z}_m^c \cos m\theta + \mathsf{Z}_m^s \sin m\theta\},$$
$$\mathsf{R}_{++}^{(k)}(\theta) \approx \mathsf{R}_0^c + \sum_{m=1}^{M} \{\mathsf{R}_m^c \cos m\theta + \mathsf{R}_m^s \sin m\theta\},$$

and so (5.21) gives us

$$\delta \mathsf{E}_+ = -\mathsf{R}_0^c \tag{5.25}$$

for $m = 0$ and

$$\begin{bmatrix} \mathsf{E}_+^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{E}_+^\star \end{bmatrix} \begin{bmatrix} \mathsf{Z}_m^c \\ \mathsf{Z}_m^s \end{bmatrix} - \begin{bmatrix} \mathsf{Z}_m^c \\ \mathsf{Z}_m^s \end{bmatrix} \mathsf{E}_+^\star = \begin{bmatrix} \mathsf{R}_m^c \\ \mathsf{R}_m^s \end{bmatrix} \tag{5.26}$$

for $m = 1, \ldots, M$. The Sylvester equations in (5.26) are nonsingular, because the restriction on the size of the imaginary parts of the eigenvalues of $\mathsf{E}_+^\star$ means that

$$\begin{bmatrix} \mathsf{E}_+^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{E}_+^\star \end{bmatrix} \quad \text{and} \quad \mathsf{E}_+^\star$$

have no common eigenvalue.

- To obtain $\mathsf{Z}_{--}(\theta)$, we set

$$\mathsf{Z}_{--}(\theta) \approx \sum_{m=1}^{M} \{\mathsf{Z}_m^c \cos m\theta + \mathsf{Z}_m^s \sin m\theta\},$$
$$\mathsf{R}_{--}^{(k)}(\theta) \approx \mathsf{R}_0^c + \sum_{m=1}^{M} \{\mathsf{R}_m^c \cos m\theta + \mathsf{R}_m^s \sin m\theta\},$$

and so (5.22) gives us

$$\delta \mathsf{E}_- = -\mathsf{R}_0^c \tag{5.27}$$

for $m = 0$ and

$$\begin{bmatrix} \mathsf{E}_-^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{E}_-^\star \end{bmatrix} \begin{bmatrix} \mathsf{Z}_m^c \\ \mathsf{Z}_m^s \end{bmatrix} - \begin{bmatrix} \mathsf{Z}_m^c \\ \mathsf{Z}_m^s \end{bmatrix} \mathsf{E}_-^\star = \begin{bmatrix} \mathsf{R}_m^c \\ \mathsf{R}_m^s \end{bmatrix} \tag{5.28}$$

for $m = 1, \ldots, M$. The Sylvester equations in (5.28) are nonsingular, because the restriction on the size of the imaginary parts of the eigenvalues of $\mathsf{E}_-^\star$ means that

$$\begin{bmatrix} \mathsf{E}_-^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{E}_-^\star \end{bmatrix} \quad \text{and} \quad \mathsf{E}_-^\star$$

have no common eigenvalue.

- To obtain $\mathsf{Z}_{+-}(\theta)$, we set

$$\mathsf{Z}_{+-}(\theta) \approx \sum_{m=1}^{M} \left\{ \mathsf{Z}_m^c \cos\left[m-\tfrac{1}{2}\right]\theta + \mathsf{Z}_m^s \sin\left[m-\tfrac{1}{2}\right]\theta \right\},$$
$$\mathsf{R}_{+-}^{(k)}(\theta) \approx \sum_{m=1}^{M} \left\{ \mathsf{R}_m^c \cos\left[m-\tfrac{1}{2}\right]\theta + \mathsf{R}_m^s \sin\left[m-\tfrac{1}{2}\right]\theta \right\},$$

and so (5.23) gives us

$$\begin{bmatrix} \mathsf{E}_+^\star & -[m-\frac{1}{2}]\mathsf{I} \\ [m-\frac{1}{2}]\mathsf{I} & \mathsf{E}_+^\star \end{bmatrix} \begin{bmatrix} \mathsf{Z}_m^c \\ \mathsf{Z}_m^s \end{bmatrix} - \begin{bmatrix} \mathsf{Z}_m^c \\ \mathsf{Z}_m^s \end{bmatrix} \mathsf{E}_-^\star = \begin{bmatrix} \mathsf{R}_m^c \\ \mathsf{R}_m^s \end{bmatrix} \tag{5.29}$$

for $m = 1, \ldots, M$. The Sylvester equations in (5.29) are nonsingular, because of the restriction on the size of the imaginary parts of the eigenvalues of $\mathsf{E}_+^\star$ and $\mathsf{E}_-^\star$ means that

$$\begin{bmatrix} \mathsf{E}_+^\star & -[m-\frac{1}{2}]\mathsf{I} \\ [m-\frac{1}{2}]\mathsf{I} & \mathsf{E}_+^\star \end{bmatrix} \quad \text{and} \quad \mathsf{E}_-^\star$$

have no common eigenvalue.

- To obtain $\mathsf{Z}_{-+}(\theta)$, we set

$$\mathsf{Z}_{-+}(\theta) \approx \sum_{m=1}^{M} \left\{ \mathsf{Z}_m^c \cos\left[m-\tfrac{1}{2}\right]\theta + \mathsf{Z}_m^s \sin\left[m-\tfrac{1}{2}\right]\theta \right\},$$
$$\mathsf{R}_{-+}^{(k)}(\theta) \approx \sum_{m=1}^{M} \left\{ \mathsf{R}_m^c \cos\left[m-\tfrac{1}{2}\right]\theta + \mathsf{R}_m^s \sin\left[m-\tfrac{1}{2}\right]\theta \right\},$$

and so (5.24) gives us

$$\begin{bmatrix} \mathsf{E}_-^\star & -[m-\frac{1}{2}]\mathsf{I} \\ [m-\frac{1}{2}]\mathsf{I} & \mathsf{E}_-^\star \end{bmatrix} \begin{bmatrix} \mathsf{Z}_m^c \\ \mathsf{Z}_m^s \end{bmatrix} - \begin{bmatrix} \mathsf{Z}_m^c \\ \mathsf{Z}_m^s \end{bmatrix} \mathsf{E}_+^\star = \begin{bmatrix} \mathsf{R}_m^c \\ \mathsf{R}_m^s \end{bmatrix} \tag{5.30}$$

for $m = 1, \ldots, M$. The Sylvester equations in (5.30) are nonsingular, because the restriction on the size of the imaginary parts of the eigenvalues of $\mathsf{E}_-^\star$ and $\mathsf{E}_+^\star$ means that

$$\begin{bmatrix} \mathsf{E}_-^\star & -[m-\frac{1}{2}]\mathsf{I} \\ [m-\frac{1}{2}]\mathsf{I} & \mathsf{E}_-^\star \end{bmatrix} \quad \text{and} \quad \mathsf{E}_+^\star$$

have no common eigenvalue.

Finally, we note that our limit

$$\mathsf{E}_n^\star \equiv \lim_{k\to\infty} \begin{bmatrix} \mathsf{E}_+^{(k)} & \mathsf{O} \\ \mathsf{O} & \mathsf{E}_-^{(k)} \end{bmatrix}$$

will not generally be quasi-upper triangular, and so we will have to perform a last Schur factorization with

$$\mathsf{Q} \equiv \begin{pmatrix} \mathsf{Q}_+ & \mathsf{O} \\ \mathsf{O} & \mathsf{Q}_- \end{pmatrix}, \qquad \mathsf{Q}_+ \in \mathbb{R}^{n_+ \times n_+}, \qquad \mathsf{Q}_- \in \mathbb{R}^{n_- \times n_-},$$

so that the transformations

$$\begin{aligned}\mathsf{P}_n^\star(\theta) &\mapsto \mathsf{P}_n^\star(\theta)\mathsf{Q},\\ \mathsf{E}_n^\star &\mapsto \mathsf{Q}^T\mathsf{E}_n^\star\mathsf{Q}\end{aligned}$$

mean that the diagonal blocks of $\mathsf{E}_n^\star$ are now in real Schur form.

5.4. Controlling the Floquet exponents. On page 2539 we stated the conditions on our Floquet exponents, the eigenvalues of

$$\mathsf{E} \equiv \begin{bmatrix}\mathsf{E}_+ & \mathsf{O}\\ \mathsf{O} & \mathsf{E}_-\end{bmatrix},$$

that must be maintained during the continuation process. Basically, this means that the sum of any imaginary part of an eigenvalue of E_+ with any imaginary part of an eigenvalue of E_- must be less than $\frac{1}{2}$. We now show what to do if this condition is found either to fail or to be dangerously close to failing at the end of a continuation step.

Suppose we have

$$-\dot{\mathsf{P}}(\theta) + \mathsf{A}(\theta)\mathsf{P}(\theta) = \mathsf{P}(\theta)\mathsf{E},$$

with

$$\mathsf{P}(\theta) \equiv [\mathsf{P}_+(\theta), \mathsf{P}_-(\theta)] \quad \text{and} \quad \mathsf{E} \equiv \begin{bmatrix}\mathsf{E}_+ & \mathsf{O}\\ \mathsf{O} & \mathsf{E}_-\end{bmatrix}.$$

If the imaginary parts of a pair of complex conjugate eigenvalues of E_- are too large, we could [17]

- move them to the top left of E_-,
- block-diagonalize E_-

(altering $\mathsf{P}_-(\theta)$ in consequence), so that now E_- has the form

$$\mathsf{E}_- = \begin{bmatrix}\alpha & \beta_1 & 0 & \cdots & \cdots & 0\\ -\beta_2 & \alpha & 0 & \cdots & \cdots & 0\\ 0 & 0 & \ddots & \ddots & \ddots & \ddots\\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots\\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots\\ 0 & 0 & \ddots & \ddots & \ddots & \ddots\end{bmatrix},$$

where β_1, β_2 are positive. Now, if we denote the first two columns of $\mathsf{P}_-(\theta)$ by $\boldsymbol{p}_1(\theta)$ and $\boldsymbol{p}_2(\theta)$, then transforming them by

$$\begin{bmatrix}\boldsymbol{p}_1(\theta) & \boldsymbol{p}_2(\theta)\end{bmatrix} \to \begin{bmatrix}\boldsymbol{p}_1(\theta) & \boldsymbol{p}_2(\theta)\end{bmatrix} \begin{bmatrix}\sqrt{\beta_1}\cos\frac{1}{2}\theta & \sqrt{\beta_1}\sin\frac{1}{2}\theta\\ -\sqrt{\beta_2}\sin\frac{1}{2}\theta & \sqrt{\beta_2}\cos\frac{1}{2}\theta\end{bmatrix}$$

will transform the leading 2×2 block of E_- into

$$\begin{bmatrix}\alpha & \sqrt{\beta_1\beta_2} - \frac{1}{2}\\ \frac{1}{2} - \sqrt{\beta_1\beta_2} & \alpha\end{bmatrix},$$

thus decreasing the size of the imaginary parts of the eigenvalues by $\frac{1}{2}$. Now the new $\boldsymbol{p}_1(\theta), \boldsymbol{p}_2(\theta)$ are in $\mathcal{Y}_+^n$, and so we increase n_+ by 2 and decrease n_- by 2.

Similarly, if the imaginary parts of a pair of complex conjugate eigenvalues of E_+ are too large, we carry out the analogous procedure to transfer them to E_-; e.g.,

- moving them to the bottom right of E_+,
- block-diagonalizing E_+,
- transforming the final two columns of $\mathsf{P}_+(\theta)$, so that the size of the imaginary parts of the dangerous eigenvalues is decreased by $\frac{1}{2}$.

We have described above the simplest situations, where there is only a single pair of dangerous eigenvalues. We omit the obvious extensions, where a larger block of eigenvalues has to be controlled.

5.5. Starting at a Hopf point. Finally, we show how $\mathsf{E}^\star$ and $\mathsf{P}^\star(\theta)$ may be constructed at a Hopf bifurcation point in order to start the continuation process; so $(\boldsymbol{x}^\star, \lambda^\star)$ is a Hopf bifurcation point for (1.1), satisfying the conditions at the beginning of section 3. Hence we have a Schur factorization

$$\mathsf{A}^\star \mathsf{Q}^\star = \mathsf{Q}^\star \mathsf{U}^\star$$

for

$$\mathsf{A}^\star \equiv T^\star \mathsf{J}(\boldsymbol{x}^\star, \lambda^\star).$$

1. Block-diagonalize $\mathsf{A}^\star$ to obtain

$$\mathsf{A}^\star \mathsf{S}^\star = \mathsf{S}^\star \mathsf{D}^\star,$$

where $\mathsf{D}^\star$ is the $n \times n$ block-diagonal matrix

$$\mathsf{D}^\star \equiv \begin{bmatrix} \mathsf{U}_+^\star & \mathsf{O} \\ \mathsf{O} & \mathsf{U}_-^\star \end{bmatrix}$$

with quasi-upper triangular $\mathsf{U}_+^\star \in \mathbb{R}^{n_+ \times n_+}$ and $\mathsf{U}_-^\star \in \mathbb{R}^{n_- \times n_-}$. In fact each of $\mathsf{U}_+^\star$ and $\mathsf{U}_-^\star$ is itself block-diagonal with

$$\mathsf{U}_+^\star \equiv \begin{bmatrix} \mathsf{U}_0 & & & & \\ & \mathsf{U}_1 & & & \\ & & \mathsf{U}_2 & & \\ & & & \ddots & \\ & & & & \ddots \end{bmatrix}$$

and

$$\mathsf{U}_-^\star \equiv \begin{bmatrix} \mathsf{U}_{\frac{1}{2}} & & & & \\ & \mathsf{U}_{\frac{3}{2}} & & & \\ & & \mathsf{U}_{\frac{5}{2}} & & \\ & & & \ddots & \\ & & & & \ddots \end{bmatrix},$$

so that the eigenvalues of $\mathsf{U}_j, j \geq 0$, have an imaginary part "close to" $\pm j$ and the eigenvalues of $\mathsf{U}_{\frac{j}{2}}, j \geq 1$, have an imaginary part "close to" $\pm \frac{j}{2}$.

2. If $\left[\begin{smallmatrix} \alpha & \beta_1 \\ -\beta_2 & \alpha \end{smallmatrix}\right]$ is a 2×2 diagonal block of U_j or $\mathsf{U}_{\frac{j}{2}}$ for $j \geq 1$, with β_1, β_2 positive and $\boldsymbol{s}_\ell^\star, \boldsymbol{s}_{\ell+1}^\star$ denoting the corresponding columns of $\mathsf{S}^\star$, then we can transform these columns by

$$\widetilde{\boldsymbol{s}}_\ell^\star \to \sqrt{\beta_1}\boldsymbol{s}_\ell^\star - \sqrt{\beta_2}\boldsymbol{s}_{\ell+1}^\star,$$
$$\widetilde{\boldsymbol{s}}_{\ell+1}^\star \to \sqrt{\beta_1}\boldsymbol{s}_\ell^\star + \sqrt{\beta_2}\boldsymbol{s}_{\ell+1}^\star.$$

Thus we obtain

$$\mathsf{A}^\star\widetilde{\mathsf{S}}^\star = \widetilde{\mathsf{S}}^\star\widetilde{\mathsf{D}}^\star,$$

where

$$\widetilde{\mathsf{D}}^\star \equiv \begin{bmatrix} \widetilde{\mathsf{U}}_+^\star & \mathsf{O} \\ \mathsf{O} & \widetilde{\mathsf{U}}_-^\star \end{bmatrix}$$

with

$$\widetilde{\mathsf{U}}_+^\star \equiv \begin{bmatrix} \mathsf{U}_0 & & & & \\ & \widetilde{\mathsf{U}}_1 & & & \\ & & \widetilde{\mathsf{U}}_2 & & \\ & & & \ddots & \\ & & & & \ddots \end{bmatrix}$$

and

$$\widetilde{\mathsf{U}}_-^\star \equiv \begin{bmatrix} \widetilde{\mathsf{U}}_{\frac{1}{2}} & & & & \\ & \widetilde{\mathsf{U}}_{\frac{3}{2}} & & & \\ & & \widetilde{\mathsf{U}}_{\frac{5}{2}} & & \\ & & & \ddots & \\ & & & & \ddots \end{bmatrix},$$

so that the 2×2 diagonal blocks of $\widetilde{\mathsf{U}}_j, \widetilde{\mathsf{U}}_{\frac{j}{2}},\ j \geq 1$, have the form $\left[\begin{smallmatrix} \alpha & \sqrt{\beta_1\beta_2} \\ -\sqrt{\beta_1\beta_2} & \alpha \end{smallmatrix}\right]$.

3. Finally, if $\left[\begin{smallmatrix} \alpha & \sqrt{\beta_1\beta_2} \\ -\sqrt{\beta_1\beta_2} & \alpha \end{smallmatrix}\right]$ is now a 2×2 diagonal block of $\widetilde{\mathsf{U}}_j, j \geq 1$, with corresponding columns $\widetilde{\boldsymbol{s}}_\ell^\star, \widetilde{\boldsymbol{s}}_{\ell+1}^\star$ of $\widetilde{\mathsf{S}}^\star$, then $\mathsf{P}_+^\star(\theta) \in \mathbb{R}^{n\times n_+}$ is obtained by replacing these columns with

$$\widetilde{\boldsymbol{s}}_\ell^\star \cos j\theta - \widetilde{\boldsymbol{s}}_{\ell+1}^\star \sin j\theta \quad \text{and} \quad \widetilde{\boldsymbol{s}}_\ell^\star \sin j\theta + \widetilde{\boldsymbol{s}}_{\ell+1}^\star \cos j\theta.$$

Similarly, if $\left[\begin{smallmatrix} \alpha & \sqrt{\beta_1\beta_2} \\ -\sqrt{\beta_1\beta_2} & \alpha \end{smallmatrix}\right]$ is now a 2×2 diagonal block of $\widetilde{\mathsf{U}}_{\frac{j}{2}}, j \geq 1$, with corresponding columns $\widetilde{\boldsymbol{s}}_\ell^\star, \widetilde{\boldsymbol{s}}_{\ell+1}^\star$ of $\widetilde{\mathsf{S}}^\star$, then $\mathsf{P}_-^\star(\theta) \in \mathbb{R}^{n\times n_-}$ is obtained by replacing these columns with

$$\widetilde{\boldsymbol{s}}_\ell^\star \cos j\theta/2 - \widetilde{\boldsymbol{s}}_{\ell+1}^\star \sin j\theta/2 \quad \text{and} \quad \widetilde{\boldsymbol{s}}_\ell^\star \sin j\theta/2 + \widetilde{\boldsymbol{s}}_{\ell+1}^\star \cos j\theta/2.$$

Hence the columns of $\mathsf{P}_+^\star$ are in $\mathcal{Y}_+^n$ and the columns of $\mathsf{P}_-^\star$ are in $\mathcal{Y}_-^n$, with $\mathsf{P}^\star(\theta) \equiv \left[\mathsf{P}_+^\star(\theta), \mathsf{P}_-^\star(\theta)\right]$ satisfying

$$-\dot{\mathsf{P}}^\star(\theta) + \mathsf{A}^\star(\theta)\mathsf{P}^\star(\theta) = \mathsf{P}^\star(\theta)\mathsf{E}^\star.$$

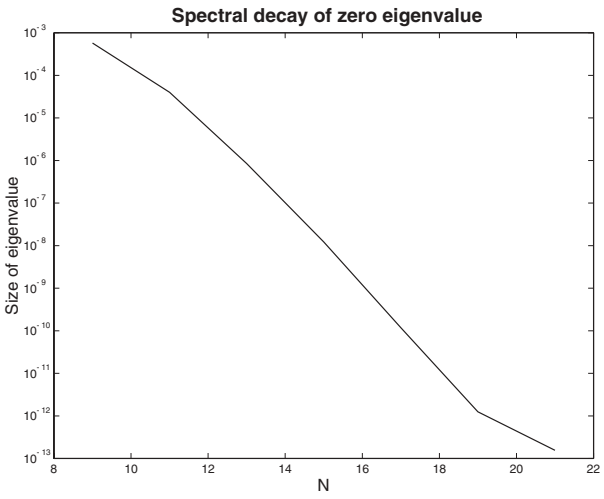


FIG. 3. *Approximating the zero Floquet exponent.*

Here

$$\mathsf{E}^\star \equiv \begin{bmatrix} \mathsf{E}^\star_+ & \mathsf{O} \\ \mathsf{O} & \mathsf{E}^\star_- \end{bmatrix}$$

is constructed from $\widetilde{\mathsf{U}}^\star_+$ and $\widetilde{\mathsf{U}}^\star_-$. The imaginary parts of the eigenvalues of $\mathsf{E}^\star$ are close to zero.

6. Numerical results. Now we illustrate the above algorithms with some well-known examples.

6.1. The Lorenz equations.

$$\begin{aligned} \dot{x} &= \sigma(y - x), \\ \dot{y} &= \lambda x - y - xz, \\ \dot{z} &= xy - bz. \end{aligned}$$

For $\sigma > b + 1$ there is a subcritical Hopf bifurcation from the stationary solution curves

$$\left(\pm\sqrt{b(\lambda-1)}, \pm\sqrt{b(\lambda-1)}, \lambda - 1\right), \quad \lambda > 1,$$

at

$$\lambda_H \equiv \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1}.$$

We use the parameter values $(\sigma, b) = (10, \frac{8}{3})$, which gives $\lambda_H \approx 24.74$, and follow the branch of periodic orbits in the range $\lambda_H \geq \lambda \geq 24$. For $N \equiv 2 * M + 1$, where M is the number of Fourier modes, we plot in Figure 3 the maximum size of the smallest Floquet exponent (which should be zero) over this range of λ. The exponential decay is clear. For this example, it is necessary only to work in $\mathcal{Y}_+$.

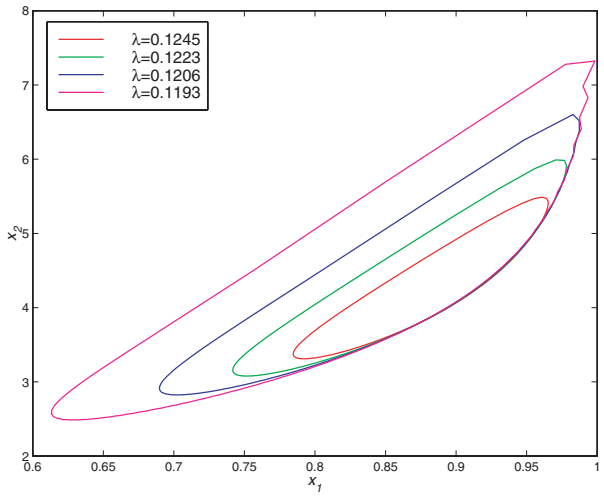


FIG. 4. *Approximating four periodic orbits.*

6.2. The A → B reaction equations.

$$\dot{x}_1 = -x_1 + \lambda(1 - x_1)\exp x_2,$$
$$\dot{x}_2 = -x_2 + \lambda a(1 - x_1)\exp x_2 - bx_2.$$

This system is used as an example in [14], in particular for the parameter values $a = 14$ and $b = 2$. In this case, the stationary solution curve through $(\boldsymbol{x}, \lambda) = (\mathbf{0}, 0)$ has two turning points as λ increases, and then there is a Hopf bifurcation point at

$$\lambda \approx 0.1309, \quad x_1 \approx 0.8951, \quad x_2 \approx 4.177.$$

The curve of periodic orbits created here exists, with λ decreasing, until $\lambda \approx 0.1055$, where it ends in a homoclinic orbit connected to the stationary solution curve. Again it is only necessary to work in $\mathcal{Y}_+$.

We use this example to illustrate how our algorithm may perform badly when applied to periodic orbits which lack smoothness or are poorly conditioned. In Figure 4, we show approximations to periodic orbits at four different values of λ, using the algorithm in section 5 with $M = 50$. We see oscillations appearing in the approximations as they try to cope with the lack of smoothness developing as $\boldsymbol{x} = (1, 10)$ is approached. As is shown in [14], this point is reached at $\lambda \approx 0.1055$, when the homoclinic orbit appears. (If results for smaller λ were shown, the oscillations would be more violent.) We exhibit this lack of smoothness differently in Figure 5, where the size of the Fourier modes is shown for $m = 1, \ldots, M$. It is clear that, as λ decreases, the exponential decay of these modes is gradually being lost. Similar conclusions can be drawn when we plot the "zero" Floquet exponent against λ in Figure 6. By the time $\lambda \approx 0.1193$ is reached, the two exponents are almost equal! Finally, and most tellingly, we examine $\mathrm{cond}\,(\mathsf{P}^\star(\theta))$ in Figure 7. Theoretically $\mathsf{P}^\star(\theta)$ can never be singular, and its condition number is a measure of the conditioning of the boundary value problem defining the periodic orbit. Here, however, we see $\mathrm{cond}\,(\mathsf{P}^\star(\theta))$ reaching 10^8! It is therefore no surprise that our algorithm, which *explicitly* works with $\mathsf{P}^\star(\theta)$, has difficulties in this situation. Collocation with piecewise-polynomials and an adaptive mesh, as used in [14], will obviously perform better in such cases.

6.3. The fourth order Lorenz equations.

$$\begin{aligned}
\dot{y}_1 &= -y_1 + 2\lambda - y_2^2 + (y_3^2 + y_4^2)/2,\\
\dot{y}_2 &= -y_2 + (y_1 y_2 - y_3 y_4) + (y_4^2 - y_3^2)/2,\\
\dot{y}_3 &= -y_3 + (y_2 - y_1)(y_3 + y_4)/2,\\
\dot{y}_4 &= -y_4 + (y_2 + y_1)(y_3 - y_4)/2.
\end{aligned}$$

This system is described in [27]. There is a stationary solution curve

$$\boldsymbol{y}(\lambda) \equiv (2\lambda, 0, 0, 0),$$

which has a supercritical bifurcation at $\lambda = \frac{1}{2}$ into the two stationary solution curves

$$\boldsymbol{y}_{\pm}(\lambda) \equiv (1, \pm\sqrt{2\lambda - 1}, 0, 0).$$

At $\lambda = 3$, the new solution curves bifurcate again, and there is then a supercritical Hopf bifurcation at

$$\lambda_H \approx 3.8531.$$

We follow the branch of periodic orbits for $\lambda_H \le \lambda \le 17$ and plot the behavior of the critical Floquet exponent in the three graphs of Figure 8. Of course one Floquet exponent is always zero, but another uninteresting one remains real and strictly negative. Thus it is the other two that we are concerned with. They are plotted in the complex plane for $\lambda_H \le \lambda \le 10$ in the first graph of Figure 8. (Since they become a complex conjugate pair for $\lambda \approx 4$, we only plot the one with positive imaginary part.) In this range of λ we can take $n_+ = n = 4$, but when λ reaches 10, this imaginary part has become greater than 0.3. Thus we decide to switch this pair of Floquet exponents from E_+ to E_-, as described in section 5.4, and so $n_+ = 2$ and $n_- = 2$. In the second graph of Figure 8, we continue to plot the critical Floquet exponent in the complex plane, but now for $10 \le \lambda \le 14.3$. Note that the imaginary part of the exponent at $\lambda = 10$ differs by $\frac{1}{2}$ in these two graphs ($\pm.33$ and $\pm.17$) as

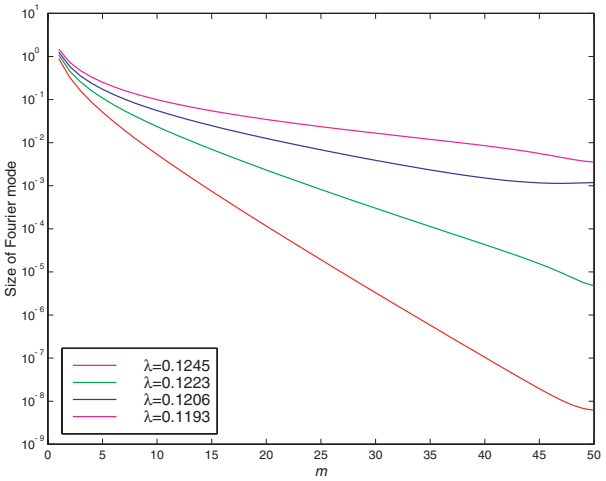


FIG. 5. *Decay of Fourier modes for periodic orbits.*

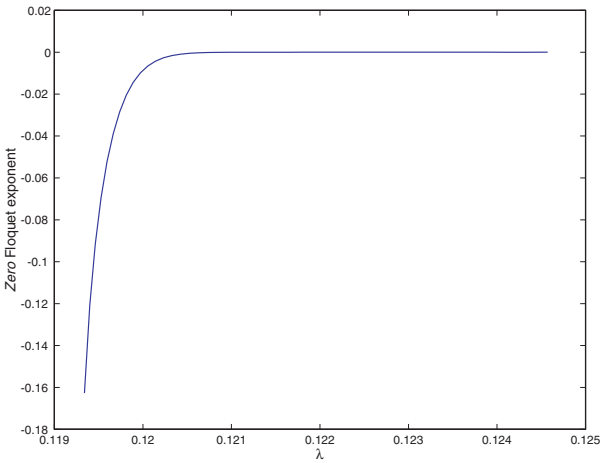


FIG. 6. *Degeneration of "zero" Floquet exponent.*

described in section 5.4. At $\lambda \approx 14.3$, this complex pair of critical Floquet exponents become real and negative, thus corresponding to negative real Floquet multipliers as described in section 4. We follow the interesting real Floquet exponent in the third graph of Figure 8, this time plotting it against λ. It quickly passes through zero, thus corresponding to a period-doubling bifurcation as described in section 7, and then continues to increase.

7. Period-doubling bifurcation. A further advantage of using Floquet theory to continue periodic orbits is that the occurrence of nonhyperbolic behavior is always obvious; i.e., because we always have $\mathsf{E}^\star$ available in quasi-upper triangular form, we can immediately see when a Floquet exponent crosses the imaginary axis. The three simplest types of bifurcation that can occur are as follows.

1. Saddle-node bifurcation (turning points): when $\mathsf{E}^\star_+$ has a geometrically single,

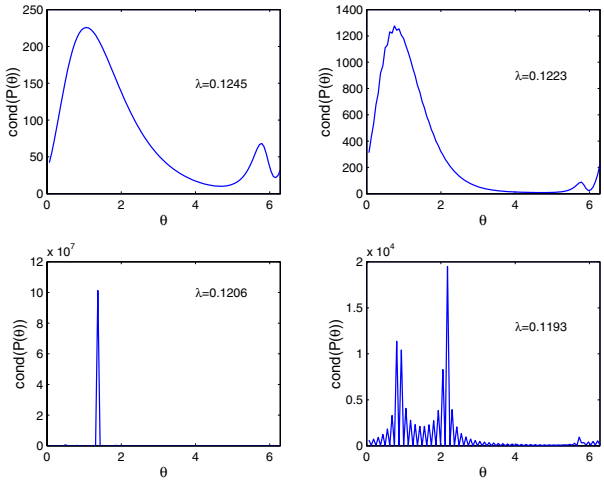


FIG. 7. *Ill-conditioning of periodic orbits.*

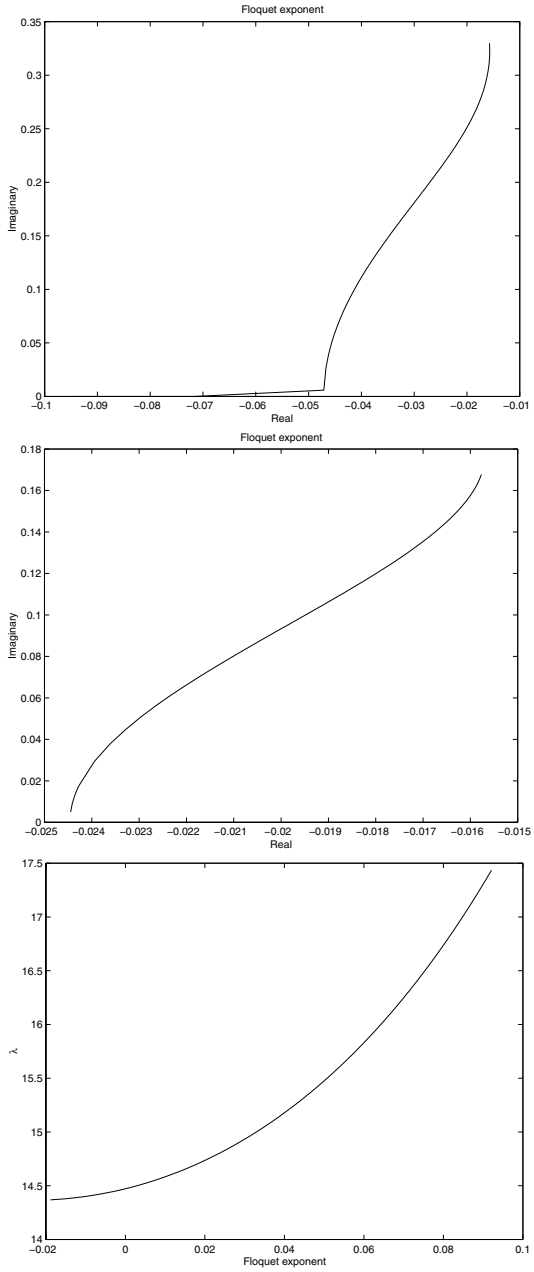


FIG. 8. *Movement of critical Floquet exponent.*

but algebraically double, zero eigenvalue, as mentioned in section 5.

2. Period-doubling bifurcation: when both $\mathsf{E}^{\star}_{+}$ and $\mathsf{E}^{\star}_{-}$ have a simple zero eigenvalue.
3. Neimark–Sacker (torus) bifurcation: when either $\mathsf{E}^{\star}_{+}$ or $\mathsf{E}^{\star}_{-}$ has a conjugate pair of purely imaginary simple eigenvalues.

If any of these occur during the continuation process, we may wish to determine the bifurcation point more precisely, which can be achieved by two different means.

- After having determined two values of the continuation parameter that bracket the bifurcation point, we may use a secant-like method to determine the parameter value at which the critical Floquet exponent is zero or has zero real part.
- We may set up a special augmented system of equations, whose solution will give the bifurcation point immediately! Modern algorithms for the above three types of bifurcation have recently been described in [15], and Floquet–Fourier versions of these algorithms could be constructed.

In addition, after having located a period-doubling or torus bifurcation point accurately, we may wish to follow these newly created objects. We do not consider Neimark–Sacker bifurcation here, since this is rather complicated [8, 21] and a treatment combining Floquet theory with the ideas in [25] would be quite lengthy. In this section, however, we do wish to present an algorithm for moving onto the new periodic orbits created at a period-doubling bifurcation point. In particular, we will see how this fits in very neatly with the spaces $\mathcal{Y}_\pm$ used in section 4 to describe our general Floquet theory and with the symmetry-breaking framework within which period-doubling is usually described [10].

The conditions that $(\boldsymbol{v}^\star(\theta), T^\star, \lambda^\star)$ must satisfy in order to be a period-doubling bifurcation point are as follows.[2]

(a) At $\lambda = \lambda^\star$ we have a nonsingular periodic orbit $\boldsymbol{v}^\star(\theta)$ of period $2\pi T^\star$, so that

$$\dot{\boldsymbol{v}}^\star(\theta) = T^\star \boldsymbol{F}(\boldsymbol{v}^\star(\theta), \lambda^\star)$$

and the solution space in $\mathcal{Y}^n_+$ for

$$-\dot{\boldsymbol{v}}(\theta) + T^\star \mathsf{J}(\boldsymbol{v}^\star(\theta), \lambda^\star)\boldsymbol{v}(\theta) = \boldsymbol{0} \tag{7.1}$$

is one-dimensional and spanned by $\dot{\boldsymbol{v}}^\star(\theta)$. Hence the implicit function theorem applies and there is a locally unique curve of periodic orbits through $\lambda^\star$, parametrized by λ, which we denote by $(\boldsymbol{v}^\star(\theta;\lambda), T^\star(\lambda))$ and which satisfies

$$\frac{\partial \boldsymbol{v}^\star}{\partial \theta}(\theta;\lambda) = T^\star(\lambda)\boldsymbol{F}(\boldsymbol{v}^\star(\theta;\lambda),\lambda). \tag{7.2}$$

(b) If $(\boldsymbol{v}^\star(\theta), T^\star, \lambda^\star)$ is a period-doubling bifurcation point; however, the solution space of (7.1) must also be one-dimensional in $\mathcal{Y}^n_-$, spanned by $\boldsymbol{\varphi}^\star(\theta)$ say.

(c) The final condition for period-doubling to occur is that

$$\int_0^{2\pi} \boldsymbol{\psi}^\star(\theta) \cdot \mathsf{K}^\star(\theta)\boldsymbol{\varphi}^\star(\theta)\,\mathrm{d}\theta \neq 0, \tag{7.3}$$

where

$$\mathsf{K}^\star(\theta) \equiv \frac{\mathrm{d}}{\mathrm{d}\lambda}\Big\{T^\star(\lambda)\mathsf{J}(\boldsymbol{v}^\star(\theta;\lambda),\lambda)\Big\}\,\Big|_{\lambda=\lambda^\star}$$

and $\boldsymbol{\psi}^\star \in \mathcal{Y}^n_-$ is the left null-vector satisfying

$$\dot{\boldsymbol{\psi}}^\star(\theta) + T^\star \mathsf{J}(\boldsymbol{v}^\star(\theta),\lambda^\star)^T \boldsymbol{\psi}^\star(\theta) = \boldsymbol{0}.$$

[2] As with Hopf bifurcation, we assume that the curve of periodic orbits is parametrizable by λ; cf. [15].

If $\mu^\star(\lambda)$ were a simple real eigenvalue equaling zero at $\lambda = \lambda^\star$ and satisfying

$$-\frac{\partial \boldsymbol{v}}{\partial \theta}(\theta;\lambda) + T^\star(\lambda)\mathsf{J}(\boldsymbol{v}^\star(\theta;\lambda),\lambda)\boldsymbol{v}(\theta;\lambda) = \mu^\star(\lambda)\boldsymbol{v}(\theta;\lambda)$$

for some nonzero $\boldsymbol{v}(\theta;\lambda) \in \mathcal{Y}^n_-$ such that $\boldsymbol{v}(\theta;\lambda^\star) \equiv \boldsymbol{\varphi}^\star(\theta)$, then (7.3) is equivalent to the transversal crossing condition

$$\frac{\mathrm{d}\mu^\star}{\mathrm{d}\lambda}(\lambda^\star) \neq 0,$$

since

$$\frac{\mathrm{d}\mu^\star}{\mathrm{d}\lambda}(\lambda^\star) \equiv \frac{\int_0^{2\pi} \boldsymbol{\psi}^\star(\theta)\cdot \mathsf{K}^\star(\theta)\boldsymbol{\varphi}^\star(\theta)\,\mathrm{d}\theta}{\int_0^{2\pi} \boldsymbol{\psi}^\star(\theta)\cdot \boldsymbol{\varphi}^\star(\theta)\,\mathrm{d}\theta}.$$

It is not necessary, however, for this eigenvalue to be simple.

In terms of our Floquet variables we have

$$-\dot{\mathsf{P}}^\star(\theta) + T^\star\mathsf{J}(\boldsymbol{v}^\star(\theta),\lambda^\star)\mathsf{P}^\star(\theta) = \mathsf{P}^\star(\theta)\mathsf{E}^\star;$$

where

$$\mathsf{P}^\star(\theta) \equiv [\mathsf{P}^\star_+(\theta), \mathsf{P}^\star_-(\theta)],$$

with $\mathsf{P}^\star_+(\theta) \in \mathbb{R}^{n\times n_+}$ and $\mathsf{P}^\star_-(\theta) \in \mathbb{R}^{n\times n_-}$, and

$$\mathsf{E}^\star \equiv \begin{bmatrix} \mathsf{E}^\star_+ & \mathsf{O} \\ \mathsf{O} & \mathsf{E}^\star_- \end{bmatrix},$$

with $\mathsf{E}^\star_+ \in \mathbb{R}^{n_+\times n_+}$ and $\mathsf{E}^\star_- \in \mathbb{R}^{n_-\times n_-}$. As usual, $n_+ + n_- = n$, with the columns of $\mathsf{P}^\star_+(\theta)$ in $\mathcal{Y}^n_+$ and the columns of $\mathsf{P}^\star_-(\theta)$ in $\mathcal{Y}^n_-$. Since $\mathsf{E}^\star_+$ has a simple zero eigenvalue, this matrix can be chosen to be in quasi-upper triangular form with first column zero. Hence the first column of $\mathsf{P}^\star_+(\theta)$ is a multiple of $\dot{\boldsymbol{v}}^\star(\theta)$, i.e., $\alpha^\star \boldsymbol{p}^\star(\theta) = \dot{\boldsymbol{v}}^\star(\theta)$. Similarly, since $\mathsf{E}^\star_-$ has a one-dimensional null-space, it too can be chosen to be in quasi-upper triangular form with first column zero, so that the first column of $\mathsf{P}^\star_-(\theta)$ is $\boldsymbol{\varphi}^\star(\theta)$. We thus have

$$\boldsymbol{p}^\star(\theta) = \mathsf{P}^\star_+(\theta)\boldsymbol{e}_1 \quad \text{and} \quad \boldsymbol{\varphi}^\star(\theta) = \mathsf{P}^\star_-(\theta)\boldsymbol{e}_1.$$

Hence the key period-doubling condition (7.3) is equivalent to the last n_- components of

$$\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{k}^\star(\theta)\,\mathrm{d}\theta$$

not lying in the range of $\mathsf{E}^\star_-$, where $\boldsymbol{k}^\star(\theta) \in \mathbb{R}^n$ is defined by

$$\mathsf{P}^\star(\theta)\boldsymbol{k}^\star(\theta) \equiv \mathsf{K}^\star(\theta)\boldsymbol{\varphi}^\star(\theta).$$

Note that $\mathsf{K}^\star(\theta)\boldsymbol{\varphi}^\star(\theta) \in \mathcal{Y}^n_-$ and so the first n_+ components of $\boldsymbol{k}^\star(\theta)$ lie in $\mathcal{Y}^{n_+}_-$ while the last n_- components lie in $\mathcal{Y}^{n_-}_+$.

From now on, we will be interested in a doubling of $\boldsymbol{v}^\star(\theta)$; i.e., we define

$$\boldsymbol{v}^\star_d(\theta) \equiv \boldsymbol{v}^\star(2\theta) \quad \text{and} \quad T^\star_d \equiv 2T^\star,$$

and so $(\boldsymbol{v}_d^\star(\theta), T_d^\star)$ satisfies

$$\dot{\boldsymbol{v}}(\theta) = T\boldsymbol{F}(\boldsymbol{v}(\theta), \lambda^\star).$$

(Thus we are offending against the minimality condition usually considered as part of the definition of a periodic orbit; cf. page 2522.) The Floquet variables for this period-doubled orbit are

$$\mathsf{P}_d^\star(\theta) \equiv \mathsf{P}^\star(2\theta) \equiv [\mathsf{P}_+^\star(2\theta), \mathsf{P}_-^\star(2\theta)] \quad \text{and} \quad \mathsf{E}_d^\star \equiv 2\,\mathsf{E}^\star \equiv 2\begin{bmatrix} \mathsf{E}_+^\star & \mathsf{O} \\ \mathsf{O} & \mathsf{E}_-^\star \end{bmatrix},$$

and thus they satisfy the equation

$$-\dot{\mathsf{P}}_d^\star(\theta) + \mathsf{A}_d^\star(\theta)\mathsf{P}_d^\star(\theta) = \mathsf{P}_d^\star(\theta)\mathsf{E}_d^\star,$$

where

$$\mathsf{A}_d^\star(\theta) \equiv T_d^\star \mathsf{J}(\boldsymbol{v}_d^\star(\theta), \lambda^\star). \tag{7.4}$$

Note that *all* the columns of $\mathsf{P}_d^\star(\theta)$ are now in $\mathcal{Y}_+^n$, but if $\boldsymbol{q}^\star(\theta)$ is one of the first n_+ columns, then it is *symmetric*, i.e.,

$$\boldsymbol{q}^\star(\theta) = \boldsymbol{q}^\star(\theta + \pi) \quad \forall \theta \in \mathbb{R},$$

while if $\boldsymbol{q}^\star(\theta)$ is one of the last n_- columns, then it is *antisymmetric*, i.e.,

$$\boldsymbol{q}^\star(\theta) = -\boldsymbol{q}^\star(\theta + \pi) \quad \forall \theta \in \mathbb{R}.$$

If we use (7.2) to define

$$\boldsymbol{v}_d^\star(\theta; \lambda) \equiv \boldsymbol{v}^\star(2\theta; \lambda) \quad \text{and} \quad T_d^\star(\lambda) \equiv 2T^\star(\lambda),$$

then $(\boldsymbol{v}_d^\star(\theta; \lambda), T_d^\star(\lambda), \lambda)$ is a curve of periodic orbits passing through $(\boldsymbol{v}_d^\star(\theta), T_d^\star, \lambda^\star)$. It is symmetric and parametrizable by λ and satisfies

$$\frac{\partial \boldsymbol{v}_d^\star}{\partial \theta}(\theta; \lambda) = T_d^\star(\lambda)\boldsymbol{F}(\boldsymbol{v}_d^\star(\theta; \lambda), \lambda).$$

It is important to realize that, since the imaginary parts of the eigenvalues of $\mathsf{E}^\star$ are bounded below $\frac{1}{2}$ in size, the imaginary parts of the eigenvalues of $\mathsf{E}_d^\star$ are correspondingly bounded below 1 in size. As we will only be using $\mathsf{E}_d^\star$ in connection with equations in $\mathcal{Y}_+^n$, as in section 2.1, this will be sufficient.

Now we look for a *new* curve of periodic orbits passing through $(\boldsymbol{v}_d^\star(\theta), T_d^\star, \lambda^\star)$ of the form

$$\boldsymbol{v}(\theta) \equiv \boldsymbol{v}_d^\star(\theta; \lambda) + \varepsilon\,[\boldsymbol{a}_d^\star(\theta) + \boldsymbol{z}(\theta)]$$

with period $T_d^\star(\lambda) + \varepsilon T$. Here ε is a small scalar which parametrizes the new curve, and $(\boldsymbol{z}(\theta), T, \lambda) \in (\mathcal{Y}_+^n, \mathbb{R}^2)$ are to be determined. $\boldsymbol{a}_d^\star(\theta)$ is the $(n_+ + 1)$th column of $\mathsf{P}_d^\star(\theta)$, and the amplitude of $\boldsymbol{v}(\theta)$ is fixed by insisting that

$$\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{a}_d^\star(\theta) \cdot \left[\mathsf{P}_d^\star(\theta)\mathsf{P}_d^\star(\theta)^T\right]^{-1} \boldsymbol{z}(\theta)\,\mathrm{d}\theta = 0. \tag{7.5a}$$

Note that $\boldsymbol{a}_d^\star(\theta) \equiv \boldsymbol{\varphi}^\star(2\theta)$ is antisymmetric according to the above definition, and so we are *breaking the symmetry* of $\boldsymbol{v}_d^\star(\theta;\lambda)$. Similarly, the phase of $\boldsymbol{v}(\theta)$ is fixed by insisting that

$$\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{p}_d^\star(\theta)\cdot\left[\mathsf{P}_d^\star(\theta)\mathsf{P}_d^\star(\theta)^T\right]^{-1}\boldsymbol{z}(\theta)\,\mathrm{d}\theta = 0, \tag{7.5b}$$

where $\boldsymbol{p}_d^\star(\theta) \equiv \boldsymbol{p}^\star(2\theta)$ is the first column of $\mathsf{P}_d^\star(\theta)$, i.e., $2\alpha^\star \boldsymbol{p}_d^\star(\theta) = \dot{\boldsymbol{v}}_d^\star(\theta)$. The system of equations $(\boldsymbol{z}(\theta), T, \lambda)$ must satisfy is therefore

$$-\dot{\boldsymbol{v}}(\theta) + [T_d^\star(\lambda) + \varepsilon T]\,\boldsymbol{F}(\boldsymbol{v}(\theta),\lambda) = \boldsymbol{0}, \tag{7.6}$$

together with (7.5). In order to apply the implicit function theorem, we may adopt the strategy used for Hopf bifurcation in section 3 and construct the smooth mapping

$$\boldsymbol{G} : (\mathcal{Y}_+^n \times \mathbb{R}^2) \times \mathbb{R} \mapsto \mathcal{Y}_+^n \times \mathbb{R}^2$$

by the following:

- for $\varepsilon \neq 0$, $\boldsymbol{G}(\boldsymbol{z}, T, \lambda; \varepsilon)$ is defined by

$$\begin{aligned} &-[\dot{\boldsymbol{a}}_d^\star(\theta) + \dot{\boldsymbol{z}}(\theta)] + \tfrac{1}{\varepsilon}\Big\{[T_d^\star(\lambda) + \varepsilon T] \\ &\qquad \boldsymbol{F}(\boldsymbol{v}_d^\star(\theta;\lambda) + \varepsilon[\boldsymbol{a}_d^\star(\theta) + \boldsymbol{z}(\theta)],\lambda) - T_d^\star(\lambda)\boldsymbol{F}(\boldsymbol{v}_d^\star(\theta;\lambda),\lambda)\Big\} \end{aligned} \tag{7.7}$$

  plus the two scalar conditions (7.5),

- $\boldsymbol{G}(\boldsymbol{z}, T, \lambda; 0)$ is defined by

$$-[\dot{\boldsymbol{a}}_d^\star(\theta) + \dot{\boldsymbol{z}}(\theta)] + T_d^\star(\lambda)\mathsf{J}(\boldsymbol{v}_d^\star(\theta;\lambda),\lambda)[\boldsymbol{a}_d^\star(\theta) + \boldsymbol{z}(\theta)] + T\boldsymbol{F}(\boldsymbol{v}_d^\star(\theta;\lambda),\lambda) \tag{7.8}$$

  plus the two scalar conditions (7.5).

Thus the zeroes $(\boldsymbol{z}(\theta), T, \lambda)$ of $\boldsymbol{G}$ for nonzero ε define nonsymmetric periodic orbits near $(\boldsymbol{v}_d^\star(\theta), T_d^\star, \lambda^\star)$. It is immediate, however, that $\boldsymbol{G}(\boldsymbol{0}, 0, \lambda^\star; 0) = \boldsymbol{0}$, and we can determine whether the linearization of $\boldsymbol{G}$ at this point has any nontrivial solutions by considering the system

$$\begin{aligned} &-\dot{\boldsymbol{z}}(\theta) + \mathsf{A}_d^\star(\theta)\boldsymbol{z}(\theta) + T\tfrac{\alpha^\star}{T^\star}\boldsymbol{p}_d^\star(\theta) + \lambda\mathsf{K}_d^\star(\theta)\boldsymbol{a}_d^\star(\theta) = \boldsymbol{0}, \\ &\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{a}_d^\star(\theta)\cdot\left[\mathsf{P}_d^\star(\theta)\mathsf{P}_d^\star(\theta)^T\right]^{-1}\boldsymbol{z}(\theta)\,\mathrm{d}\theta = 0, \\ &\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{p}_d^\star(\theta)\cdot\left[\mathsf{P}_d^\star(\theta)\mathsf{P}_d^\star(\theta)^T\right]^{-1}\boldsymbol{z}(\theta)\,\mathrm{d}\theta = 0 \end{aligned} \tag{7.9}$$

for $(\boldsymbol{z}(\theta), T, \lambda)$, where $\mathsf{A}_d^\star(\theta)$ is defined in (7.4) and

$$\mathsf{K}_d^\star(\theta) \equiv \frac{\mathrm{d}}{\mathrm{d}\lambda}\Big\{T_d^\star(\lambda)\mathsf{J}(\boldsymbol{v}_d^\star(\theta;\lambda),\lambda)\Big\}\,\Big|_{\lambda=\lambda^\star} \equiv \mathsf{K}^\star(2\theta).$$

Using the Floquet transformation

$$\boldsymbol{z}(\theta) \equiv \mathsf{P}_d^\star(\theta)\boldsymbol{w}(\theta) \quad \text{with } \boldsymbol{w} \in \mathcal{Y}_+^n,$$

(7.9) becomes

$$\begin{aligned} &-\dot{\boldsymbol{w}}(\theta) + \mathsf{E}_d^\star\boldsymbol{w}(\theta) + T\tfrac{\alpha^\star}{T^\star}\boldsymbol{e}_1 + \lambda\boldsymbol{k}_d^\star(\theta) = \boldsymbol{0}, \\ &\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{e}_1\cdot\boldsymbol{w}(\theta)\,\mathrm{d}\theta = 0, \\ &\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{e}_{n_+ +1}\cdot\boldsymbol{w}(\theta)\,\mathrm{d}\theta = 0, \end{aligned}$$

where

$$\mathsf{P}_d^\star(\theta)\boldsymbol{k}_d^\star(\theta) \equiv \mathsf{K}_d^\star(\theta)\boldsymbol{a}_d^\star(\theta) \equiv 2\mathsf{K}^\star(2\theta)\boldsymbol{\varphi}^\star(2\theta).$$

From (7.3) and the fact that zero is a simple eigenvalue of $\mathsf{E}_+^\star$, we know that neither of the two linearly independent vectors

$$\boldsymbol{e}_1 \quad \text{and} \quad \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{k}_d^\star(\theta)\,\mathrm{d}\theta$$

are in the range of $\mathsf{E}_d^\star$. Hence (7.9) has no nontrivial solution, and the implicit function theorem applies to $\boldsymbol{G}$ at $(\boldsymbol{0}, 0, \lambda^\star; 0)$ and tells us that $\boldsymbol{G}(\boldsymbol{z}, T, \lambda; \varepsilon) = \boldsymbol{0}$ has a locally unique solution curve parametrized by ε.

Thus we define the following Newton-chord iteration for obtaining these nonsymmetric period-doubled orbits.

- Set

$$\boldsymbol{y}^{(0)}(\theta) = \boldsymbol{a}_d^\star(\theta), \quad T^{(0)} = 0, \quad \lambda^{(0)} = \lambda^\star.$$

- Solve

$$\begin{aligned}
&\left[-\frac{\mathrm{d}}{\mathrm{d}\theta} + \mathsf{A}_d^\star(\theta)\right]\boldsymbol{z}(\theta) + \delta T\frac{\alpha^\star}{T^\star}\boldsymbol{p}_d^\star(\theta) + \delta\lambda\mathsf{K}_d^\star(\theta)\boldsymbol{a}_d^\star(\theta) = \frac{1}{\varepsilon}\boldsymbol{r}^{(k)}(\theta),\\
&\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{p}_d^\star(\theta)\cdot\left[\mathsf{P}_d^\star(\theta)\mathsf{P}_d^\star(\theta)^T\right]^{-1}\boldsymbol{z}(\theta)\,\mathrm{d}\theta = 0,\\
&\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{a}_d^\star(\theta)\cdot\left[\mathsf{P}_d^\star(\theta)\mathsf{P}_d^\star(\theta)^T\right]^{-1}\boldsymbol{z}(\theta)\,\mathrm{d}\theta = 0
\end{aligned} \tag{7.10}$$

  for $\boldsymbol{z} \in \mathcal{Y}_+^n$, δT, and $\delta\lambda$, where

$$\begin{aligned}
\boldsymbol{r}^k(\theta) \equiv\ &\dot{\boldsymbol{v}}_d^\star(\theta;\lambda^{(k)}) + \varepsilon\dot{\boldsymbol{y}}^{(k)}(\theta)\\
&- \left[T_d^\star(\lambda^{(k)}) + \varepsilon T^{(k)}\right]\boldsymbol{F}(\boldsymbol{v}_d^\star(\theta;\lambda^{(k)}) + \varepsilon\boldsymbol{y}^{(k)}(\theta), \lambda^{(k)}).
\end{aligned}$$

- Set

$$\begin{aligned}
\boldsymbol{y}^{(k+1)}(\theta) &= \boldsymbol{y}^{(k)} + \boldsymbol{z}(\theta),\\
T^{(k+1)} &= T^{(k)} + \delta T,\\
\lambda^{(k+1)} &= \lambda^{(k)} + \delta\lambda.
\end{aligned}$$

Note that only the same augmented linear periodic differential equation, with varying right-hand sides, needs to be solved, at each iteration.

7.1. Fourier approximation. Finally, we show how to efficiently compute accurate approximations to the periodic orbits of (7.6), using the above Newton-chord iteration and the results of section 2.3. Floquet transforming the basic linear iteration (7.10), we obtain

$$\begin{aligned}
&-\dot{\boldsymbol{w}}(\theta) + \mathsf{E}_d^\star\boldsymbol{w}(\theta) + \delta T\tfrac{\alpha^\star}{T^\star}\boldsymbol{e}_1 + \delta\lambda\boldsymbol{k}_d^\star(\theta) = \boldsymbol{f}(\theta),\\
&\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{e}_1\cdot\boldsymbol{w}(\theta)\,\mathrm{d}\theta = 0,\\
&\frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{e}_{n_++1}\cdot\boldsymbol{w}(\theta)\,\mathrm{d}\theta = 0,
\end{aligned} \tag{7.11}$$

where

$$\begin{aligned} \boldsymbol{z}(\theta) &\equiv \mathsf{P}_d^\star(\theta)\boldsymbol{w}(\theta), \\ \tfrac{1}{\varepsilon}\boldsymbol{r}^{(k)}(\theta) &\equiv \mathsf{P}_d^\star(\theta)\boldsymbol{f}(\theta) \end{aligned}$$

with $\boldsymbol{w}(\theta), \boldsymbol{f}(\theta) \in \mathcal{Y}_+^n$.

Using the approximate Fourier coefficients

$$\begin{aligned} \boldsymbol{w}(\theta) &\approx \tilde{\boldsymbol{w}}_0^c + \sum_{m=1}^{M} \left\{ \tilde{\boldsymbol{w}}_m^c \cos m\theta + \tilde{\boldsymbol{w}}_m^s \sin m\theta \right\}, \\ \boldsymbol{f}(\theta) &\approx \tilde{\boldsymbol{f}}_0^c + \sum_{m=1}^{M} \left\{ \tilde{\boldsymbol{f}}_m^c \cos m\theta + \tilde{\boldsymbol{f}}_m^s \sin m\theta \right\}, \end{aligned}$$

and

$$\boldsymbol{k}_d^\star(\theta) \approx \tilde{\boldsymbol{k}}_0^c + \sum_{m=1}^{M} \left\{ \tilde{\boldsymbol{k}}_m^c \cos m\theta + \tilde{\boldsymbol{k}}_m^s \sin m\theta \right\}$$

and then matching coefficients, gives us the modal equations

- for $m = 0$

$$\begin{bmatrix} \mathsf{E}_d^\star & \frac{\alpha^\star}{T^\star}\boldsymbol{e}_1 & \tilde{\boldsymbol{k}}_0^c \\ \boldsymbol{e}_1^T & 0 & 0 \\ \boldsymbol{e}_{n_+ + 1}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{w}}_0^c \\ \delta T \\ \delta\lambda \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{f}}_0^c \\ 0 \\ 0 \end{bmatrix},$$

- for $1 \le m \le M$

$$\begin{bmatrix} \mathsf{E}_d^\star & -m\mathsf{I} \\ m\mathsf{I} & \mathsf{E}_d^\star \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{w}}_m^c \\ \tilde{\boldsymbol{w}}_m^s \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{f}}_m^c - \delta\lambda\tilde{\boldsymbol{k}}_m^c \\ \tilde{\boldsymbol{f}}_m^s - \delta\lambda\tilde{\boldsymbol{k}}_m^s \end{bmatrix}.$$

Hence the extra scalar unknowns δT and $\delta\lambda$ are solved for as part of the nonsingular $m = 0$ system, while the other modal equations remain the same as in section 2.3, and nonsingular because of the eigenvalue conditions imposed on $\mathsf{E}^\star$.

The above iteration only moves us onto the curve of period-doubled orbits near $(\boldsymbol{v}^\star(\theta), T^\star, \lambda^\star)$. If we want to follow this new curve, using the techniques in section 5, then we must update the Floquet information as in section 5.3. Since the imaginary parts of the eigenvalues of $\mathsf{E}_d^\star$ may exceed $\frac{1}{2}$ in size, it may first be necessary to apply section 5.4 to $\mathsf{P}_d^\star(\theta)$, thus introducing some components in $\mathcal{Y}_-$.

7.2. Numerical example. As an illustration of period-doubling bifurcation, we continue using the fourth order Lorenz equations as an example. The third graph in Figure 8 shows the critical Floquet exponent passing through zero for λ slightly greater than 14.3, and a simple secant iteration quickly locates the bifurcation point at

$$\lambda^\star \approx 14.4722024 \quad \text{and} \quad T^\star \approx 0.1530674\,.$$

Now we apply the algorithm developed in this section to move onto the period-doubled orbit with $\varepsilon = \frac{1}{2}$. Since the phase-space is four-dimensional, Figure 9 shows two different three-dimensional projections. The number of Fourier modes used in the approximation is $M = 50$.

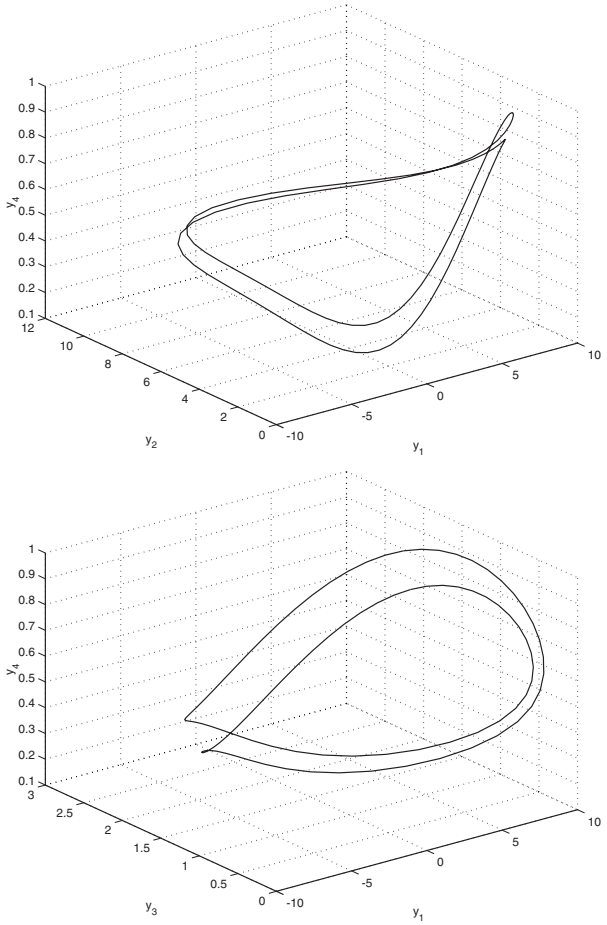


FIG. 9. *Period-doubled orbit with* $\varepsilon = \frac{1}{2}$.

8. Conclusion. We have shown how Floquet theory may be utilized in order to compute Fourier approximations of periodic orbits. The key result is that the size of the linear systems which must be solved is independent of the number M of Fourier modes used in the approximation. The overhead is that the Floquet variables $\mathsf{P}(\theta)$ must also be carried along in the continuation process, and this extra work is proportional to n, the dimension of the phase-space. Consequently, when $n \ll M$, we have a highly efficient and accurate algorithm for smooth and well-conditioned periodic orbits.

Of course, it would be silly to assert that the present algorithm can replace the AUTO package in [14], and we make no such claim! AUTO has been gradually refined over 20 years, and was developed from collocation and adaptive mesh ideas of an even earlier vintage, e.g., COLSYS [3]. What we do claim is that (just as in other areas of differential equations) spectral approximation of periodic orbits has its place alongside finite difference and piecewise polynomial approximation. Spectral methods are not trivial to implement efficiently and accurately on nonlinear problems, and further work is necessary before practical conclusions can be drawn. For example, if the basic linear algebra philosophy of [29] were used, Floquet theory could be thought of as an

efficient way of (approximately) factorizing the structured $Nn \times Nn$ matrix obtained from collocation [6].

Perhaps the most pleasing feature of the present algorithm is that it exploits the full mathematical structure of the periodic orbit problem. The fact that Floquet theory provides a *constant* matrix means that many algorithms for questions about stationary solutions (where the Jacobian matrix is constant of course) can easily be adapted for similar questions about periodic orbits.

- The algorithm for Hopf bifurcation in section 3 can be adapted to apply when invariant tori are created at a Neimark–Sacker bifurcation, as mentioned near the beginning of section 7.
- The algorithms for stable manifolds of stationary solutions and connecting orbits in [26] can be adapted to apply to stable manifolds of periodic orbits and periodic connections [5].
- Fourier approximation was recommended in [9, 10] to exploit spatial-temporal symmetries of periodic orbits. The analysis of such symmetries becomes even clearer when Floquet theory is utilized as well.

Papers devoted to these three applications of Floquet theory are currently being prepared.

REFERENCES

[1] H. Amann, *Ordinary Differential Equations*, Walter de Gruyter, Berlin, 1990.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, 2nd ed., SIAM, Philadelphia, 1995.

[3] U. M. Ascher, R. M. Mattheij, and R. D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, SIAM, Philadelphia, 1995.

[4] R. H. Bartels and G. W. Stewart, *Solution of the equation* $\mathsf{AX} + \mathsf{XB} = \mathsf{C}$, Comm. ACM, 15 (1972), pp. 820–826.

[5] W. J. Beyn, *On well-posed problems for connecting orbits in dynamical systems*, in Chaotic Numerics, Contemp. Math. 172, P. E. Kloeden and K. J. Palmer, eds., AMS, Providence, RI, 1994, pp. 131–168.

[6] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.

[7] F. Chatelin, *Eigenvalues of Matrices*, John Wiley, Chichester, UK, 1993.

[8] S-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[9] M. Dellnitz, *A computational method and path following for periodic solutions with symmetry*, in Continuation and Bifurcations: Numerical Techniques and Applications, D. Roose, B. De Dier, and A. Spence, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990, pp. 153–167.

[10] M. Dellnitz, *Computational bifurcation of periodic solutions in systems with symmetry*, IMA J. Numer. Anal., 12 (1992), pp. 429–455.

[11] E. J. Doedel, A. D. Jepson, and H. B. Keller, *Numerical methods for Hopf bifurcation and continuation of periodic solution paths*, in Computing Methods in Applied Sciences and Engineering, Vol. VI, R. Glowinski and J. L. Lions, eds., North-Holland, Amsterdam, 1984, pp. 127–138.

[12] E. J. Doedel, H. B. Keller, and J. P. Kernévez, *Numerical analysis and control of bifurcation problems. Part* 1: *Bifurcation in finite dimensions*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 493–520.

[13] E. J. Doedel, H. B. Keller, and J. P. Kernévez, *Numerical analysis and control of bifurcation problems. Part* 2: *Bifurcation in infinite dimensions*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 745–772.

[14] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Yu. A. Kuznetsov, B. Sandstede, and X. J. Wang, *Continuation and Bifurcation Software for Ordinary Differential Equations (with HomCont)*, Technical report, California Institute of Technology, Pasadena, CA, 1998; also available online from http://indy.cs.concordia.ca/auto/main.html.

[15] E. J. Doedel, W. Govaerts, and Yu. A. Kuznetsov, *Computation of periodic solution bifurcations in ODEs using bordered systems*, SIAM J. Numer. Anal., 41 (2003), pp. 401–

435.
[16] E. W. Gekeler, *On trigonometric collocation in Hopf bifurcation*, in Bifurcation and Symmetry, E. Allgower, K. Böhmer, and M. Golubitsky, eds., Internat. Ser. Numer. Math. 104, Birkhäuser, Basel, Switzerland, 1992, pp. 147–156.
[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[18] W. J. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equilibria*, SIAM, Philadelphia, 2000.
[19] R. Grimshaw, *Nonlinear Ordinary Differential Equations*, Blackwell, Oxford, UK, 1990.
[20] J. Hale, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.
[21] G. Iooss and D. D. Joseph, *Elementary Stability and Bifurcation Theory*, 2nd ed., Springer-Verlag, New York, 1990.
[22] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[23] H. B. Keller and A. D. Jepson, *Steady state and periodic solution paths: Their bifurcations and computations*, in Numerical Methods for Bifurcation Problems, T. Küpper, H. D. Mittelmann, and H. Weber, eds., Internat. Schriftenreihe Numer. Math. 70, Birkhäuser, Basel, Switzerland, 1984, pp. 219–246.
[24] G. Moore, *Computation and parametrization of periodic and connecting orbits*, IMA J. Numer. Anal., 15 (1995), pp. 245–263.
[25] G. Moore, *Computation and parametrisation of invariant curves and tori*, SIAM J. Numer. Anal., 33 (1996), pp. 2333–2358.
[26] G. Moore, *Laguerre approximation of stable manifolds with application to connecting orbits*, Math. Comp., 73 (2004), pp. 211–242.
[27] R. Seydel, *From Equilibrium to Chaos: Practical Bifurcation and Stability Analysis*, Elsevier, New York, 1988.
[28] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[29] C. F. Van Loan, *Computational Frameworks for the Fast Fourier Transform*, SIAM, Philadelphia, 1992.
[30] D. Viswanath, *The Lindstedt–Poincaré technique as an algorithm for computing periodic orbits*, SIAM Rev., 43 (2001), pp. 478–495.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2569–2589

NONLINEAR STABILITY ANALYSIS FOR THE METHOD OF TRANSPORT FOR THE ELASTIC-PLASTIC WAVE EQUATION*

GUIDO GIESE†

Abstract. In this paper we present an analytic assessment of the stability and convergence of the so-called method of transport for solving the elastic-plastic wave equation, which is a nonlinear hyperbolic partial differential equation. For the purely elastic wave equation, which is a linear hyperbolic conservation law, one can use von Neumann analysis for proving that stability and convergence immediately follow from the Lax equivalence theorem. For plastic deformation, however, nonlinear stability proofs require greater effort.

Key words. nonlinear stability, convergence, hyperbolic conservation laws, elasticity, plasticity

AMS subject classifications. 65M, 65N

DOI. 10.1137/S0036142902417881

1. Introduction.

There exist a large number of numerical schemes for hyperbolic conservation laws (cf. [2], [7], [9], [10], [22], [23]), some of which have been used for the simulation of waves in solids (e.g., [11], [12], [13], [14], [15], [16]). In [5] we followed the ansatz of Fey, who developed in [3], [4], and [8] a high order scheme called method of transport for solving the multidimensional Euler equations, to develop a numerical scheme for solving the elastic-plastic wave equation, which is not a pure conservation law anymore, since only the equations describing the conservation of momentum are in conservation form.

For linear conservation laws—as for linear PDEs in general—convergence proofs of numerical schemes are quite common, e.g., the Lax equivalence theorem (cf. [19]) for linear (schemes for) hyperbolic conservation laws, which also applies to the elastic wave equation due to its linearity. The Lax equivalence theorem basically states the equivalence of convergence of a linear numerical scheme on the one hand and consistency and stability on the other. For nonlinear PDEs, however, such as the elastic-plastic wave equation, convergence results are rare, mostly because a proof of stability is not possible.

In this paper, we will present analytic stability results for the method of transport for solving the elastic-plastic wave equation in one and two dimensions. We will start with an introduction to the method of transport for the elastic-plastic wave equation in the case of so-called antiplane shear in two space dimensions. Then we present the concepts of defining linear and nonlinear stability and their implications on the convergence of a numerical scheme. Afterwards, we analyze the stability of the linear scheme for elastic waves and the nonlinear stability of the scheme for plastic waves, including convergence considerations.

We will abstain from presenting computational examples, since they can be found in [3] or [5], reconfirming the convergence property for the scheme in one and two dimensions by various numerical computations.

*Received by the editors November 13, 2002; accepted for publication (in revised form) July 8, 2004; published electronically April 19, 2005.
http://www.siam.org/journals/sinum/42-6/41788.html
†Seminar for Applied Mathematics, Swiss Federal Institute of Technology, Zurich, Switzerland (guido_giese@yahoo.com).

2. Antiplane shear waves. We consider the propagation of shear waves in a plane. Let $\boldsymbol{u} \in \mathbb{R}^3$ denote the vector of displacements. Then

$$\epsilon = \frac{\partial u_3}{\partial x},$$
$$\gamma = \frac{\partial u_3}{\partial y},$$
$$w = \frac{\partial u_3}{\partial t}$$

are the strain components in x- and y-directions and the velocity component in the z-direction. Of the symmetric stress tensor $\underline{\boldsymbol{\sigma}}$ we need only the stress components $\sigma = (\underline{\boldsymbol{\sigma}})_{13}$ and $\tau = (\underline{\boldsymbol{\sigma}})_{23}$ in the x- and y-directions. The physical equations for the description of shear waves consist of three equations, the first one containing the conservation of momentum and two compatibility relations between velocity and strain,

$$\begin{pmatrix} w \\ \epsilon \\ \gamma \end{pmatrix}_t = \begin{pmatrix} \sigma/\rho \\ w \\ 0 \end{pmatrix}_x + \begin{pmatrix} \tau/\rho \\ 0 \\ w \end{pmatrix}_y, \tag{2.1}$$

where ρ denotes the density of the material. Our description of waves in solids is based on the model of small-strains, which is a linearization of the general flow-equations (cf. [18]). Equation (2.1) models an elastic medium that corresponds to a membrane with no coupling between longitudinal and transversal deformation, as would be the case in a solid with nonzero Poisson ratio.

Analogous to gas dynamics, we need an equation connecting stress and strain components, where we will distinguish between elastic and plastic deformation in the following.

2.1. Elastic shear waves. In the elastic case, Hook's law relates the stresses to the strains by

$$\sigma = \mu\epsilon,$$
$$\tau = \mu\gamma$$

with the elastic shear modulus μ. The wave-speed c is given by

$$c = \sqrt{\frac{\mu}{\rho}}. \tag{2.2}$$

Thus, we can write (2.1) in the well-known form of the wave equation as

$$\begin{pmatrix} w \\ \sigma \\ \tau \end{pmatrix}_t = \nabla \cdot c \begin{pmatrix} \sigma/(\rho c) & \tau/(\rho c) \\ \rho c w & 0 \\ 0 & \rho c w \end{pmatrix}. \tag{2.3}$$

Defining the vector of conserved quantities as $\boldsymbol{U}$ and the "flux" as $c\underline{\boldsymbol{L}}$, we get the simple form for (2.3):

$$\boldsymbol{U}_t + \nabla \cdot c\underline{\boldsymbol{L}} = 0. \tag{2.4}$$

2.2. Plastic shear waves. In regions of plastic deformation, however, the stress-strain relation contains only a relationship between infinitesimal changes of stress and strain components in time:

$$\begin{pmatrix} \dot{\epsilon} \\ \dot{\gamma} \end{pmatrix} = \underline{\mathcal{C}}(\sigma, \tau) \begin{pmatrix} \dot{\sigma} \\ \dot{\tau} \end{pmatrix}. \tag{2.5}$$

There exist quite a lot of different models for describing plastic deformation (cf. [18], [20]). For our further investigations we make use of the well-known Prandtl–Reuss equations, which we will present briefly in the following (for details, see [21] or [5], [11]).

The model uses the most commonly used yield function, i.e., the von Mises yield function, which reads in the special case of antiplane shear:

$$f(\sigma, \tau) = \sigma^2 + \tau^2 =: \kappa^2. \tag{2.6}$$

This yield function enables us to distinguish between plastic and elastic deformation, i.e., with

$$\kappa_0(t) = \max_{t_0 \le t' \le t} \kappa(t')$$

three different cases may occur:

- $\kappa(t) < \kappa_0$: elastic deformation.
- $\kappa(t) = \kappa_0$ and $\dot{\kappa} \le 0$: elastic unloading.
- $\kappa(t) = \kappa_0$ and $\dot{\kappa} > 0$: plastic loading.

In the elastic case, the stress-strain relationship is described by Hooke's law (cf. section 2.1). In the plastic region, the model of Prandtl–Reuss (for details, see [21]) yields the following matrix $\underline{\mathcal{C}}(\sigma, \tau)$ used in the stress-strain relationship (2.5):

$$\underline{\mathcal{C}}(\sigma, \tau) = \begin{pmatrix} \frac{1}{\mu} + \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma^2 & \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma\tau \\ \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma\tau & \frac{1}{\mu} + \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\tau^2 \end{pmatrix}, \tag{2.7}$$

where the function $\mu_p(\sigma)$ is the plastic shear modulus which can be measured in experiments. In the following we always assume that the occurrence of plasticity decreases the wave-speed, which is the case for almost all materials (cf. [18]), and hence

$$\mu_p(\sigma) \le \mu \quad \forall \sigma \in \mathbb{R}^+.$$

Using this stress-strain relationship in order to replace the time derivatives of strain variables in (2.1) by derivatives of stress variables, the wave equation describing plastic deformation has the form

$$\begin{pmatrix} w \\ \sigma \\ \tau \end{pmatrix}_t - \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\mu} + \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma^2 & \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma\tau \\ 0 & \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma\tau & \frac{1}{\mu} + \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\tau^2 \end{pmatrix}^{-1} \nabla \cdot c \begin{pmatrix} \sigma/(\rho c) & \tau/(\rho c) \\ w/c & 0 \\ 0 & w/c \end{pmatrix} = 0, \tag{2.8}$$

or simply

$$\begin{pmatrix} w \\ \sigma \\ \tau \end{pmatrix}_t - \underline{\boldsymbol{\mathcal{A}}}(\sigma,\tau)\nabla\cdot c \begin{pmatrix} \sigma/(\rho c) & \tau/(\rho c) \\ w/c & 0 \\ 0 & w/c \end{pmatrix} = 0. \tag{2.9}$$

Since $\underline{\boldsymbol{\mathcal{A}}}(\sigma,\tau)$ is not a Jacobian matrix, we no longer have a conservation law. Equation (2.8) can be written in divergence form with "source term":

$$\boldsymbol{U}_t + \nabla\cdot c\underline{\boldsymbol{L}} = \boldsymbol{P}(\sigma,\tau,\sigma_x,\sigma_y,\tau_x,\tau_y) \tag{2.10}$$

with

$$\begin{aligned} \underline{\boldsymbol{L}} &= -\underline{\boldsymbol{\mathcal{A}}}(\sigma,\tau)\begin{pmatrix} \sigma/(\rho c) & \tau/(\rho c) \\ w/c & 0 \\ 0 & w/c \end{pmatrix}, \\ \boldsymbol{P}(\sigma,\tau,\sigma_x,\sigma_y,\tau_x,\tau_y) &= \underline{\boldsymbol{\mathcal{A}}}(\sigma,\tau)\nabla\cdot c\begin{pmatrix} \sigma/(\rho c) & \tau/(\rho c) \\ w/c & 0 \\ 0 & w/c \end{pmatrix} \\ &\quad - \nabla\cdot c\underline{\boldsymbol{\mathcal{A}}}(\sigma,\tau)\begin{pmatrix} \sigma/(\rho c) & \tau/(\rho c) \\ w/c & 0 \\ 0 & w/c \end{pmatrix}. \end{aligned}$$

Another interpretation of the equation is to rewrite it as a system for five state variables:

$$\begin{pmatrix} w \\ \epsilon \\ \gamma \\ \sigma \\ \tau \end{pmatrix}_t - \nabla\cdot c\begin{pmatrix} \sigma/(\rho c) & \tau/(\rho c) \\ w/c & 0 \\ 0 & w/c \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \underline{\boldsymbol{\mathcal{C}}}^{-1}(\sigma,\tau)\begin{pmatrix} \epsilon \\ \gamma \end{pmatrix}_t \end{pmatrix}. \tag{2.11}$$

This formula will be the basis for the numerical scheme presented below. It shows that the velocity and strain variables have to be updated by computing fluxes, e.g., by using the method of transport and the stress variables σ and τ can be computed by integrating the stress-strain relationship, which will be done by applying a high order ODE solver.

The idea of the method of transport for solving a hyperbolic conservation law consists of decomposing the conservation law into advection equations which can be solved numerically. We will develop such decompositions for the wave equation under antiplane in the following sections, first for elastic waves where the system under consideration is a linear conservation law, and then for plastic waves, where not all variables are conserved.

2.3. Decomposition of the elastic wave equation. For elasticity the wave equation (2.3) (or the abbreviated version (2.4)) is a linear hyperbolic conservation law. We define new vectors $\boldsymbol{R}_i$,

$$\boldsymbol{R}_i = \boldsymbol{U} + \underline{\boldsymbol{L}}\vec{\boldsymbol{n}}_i, \quad i = 1,\dots,k, \tag{2.12}$$

where $\vec{\boldsymbol{n}}_i \in \mathbb{R}^2$ are k constant vectors, not necessarily of unit length. Then (2.4) can be rewritten in the strictly equivalent form

$$\boldsymbol{U}_t + \nabla\cdot c\underline{\boldsymbol{L}} = \frac{1}{k}\sum_{i=1}^{k}\left[(\boldsymbol{R}_i)_t + \nabla\cdot\left(\boldsymbol{R}_i\vec{\boldsymbol{n}}_i^T c\right)\right] = 0, \tag{2.13}$$

provided the vectors $\vec{\boldsymbol{n}}_i$ satisfy the following two conditions:

$$\sum_{i=1}^{k} \vec{\boldsymbol{n}}_i = 0, \tag{2.14}$$

$$\frac{1}{k} \sum_{i=1}^{k} \vec{\boldsymbol{n}}_i \vec{\boldsymbol{n}}_i^T = \underline{\boldsymbol{I}}.$$

The right-hand side of (2.13) can be reinterpreted as a coupled system of advection equations, transporting the quantity $\boldsymbol{R}_i$ at speed $c\vec{\boldsymbol{n}}_i$. Relation (2.14) is necessary and sufficient for this reinterpretation and the following numerical scheme to be consistent with the original equation.

Our numerical approximation for (2.3) consists of decoupling the system, i.e., solving each advection equation in (2.13) independently on the time interval $[t^n, t^{n+1}]$ of time-step n, which leads to

$$(\boldsymbol{R}_i)_t + \nabla \cdot \left(\boldsymbol{R}_i \vec{\boldsymbol{n}}_i^T c \right) = 0, \quad i = 1, \ldots, k, \tag{2.15}$$

and consequently our approximate update for the state vector $\boldsymbol{U}$ reads

$$\boldsymbol{U}(t^{n+1}) = \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{R}_i(t^{n+1}). \tag{2.16}$$

This decomposition into decoupled advection equations is a first order approximation in time, which is easy to prove:

$$\boldsymbol{U}_t = \frac{1}{k} \sum_{i=1}^{k} (\boldsymbol{R}_i)_t = -\frac{1}{k} \sum_{i=1}^{k} \nabla \cdot \left(c \boldsymbol{R}_i \vec{\boldsymbol{n}}_i^T \right) = -\nabla \cdot c\underline{\boldsymbol{L}}.$$

To obtain approximations of order two or higher in the time integration with the same one-step method, additional correction terms in the numerical fluxes are needed, which can be found by comparing the Taylor of the exact solution and the numerical solution (cf. [3], [4], [5]). We omit these correction terms in the following, since we focus our stability analysis on the first order scheme. The application of our analysis to higher order schemes including correction terms in the numerical fluxes is in principle analogous to the one-dimensional (1-D) case, but more complicated due to the more complex form of the numerical fluxes and hence beyond the scope of this paper.

It is noteworthy that the decomposition into advection equations according to (2.13) is a generalization of the decomposition of a linear conservation law into right eigenvectors in one dimension, which we will further analyze in our stability analysis for the scheme in one dimension.

2.4. Decomposition of the elastic-plastic wave equation. For plastic waves, the scheme described above with a decomposition of the form (2.13) is no longer possible since the plastic wave equation in the form (2.10) or (2.11) has no conservation form.

One approach is to use the same transported quantities $\boldsymbol{R}_i$ as defined in (2.12) and decompose the term on the right-hand side of (2.10) as well. Unfortunately, this formulation leads to severe numerical problems since the "source term" on the right-hand side of (2.10) depends on the spatial derivatives of $\boldsymbol{U}$. Moving derivatives

change the flux and thus the jump conditions, which means that the wave equation in divergence form with "source term" does not have the correct physical discontinuities.

Another idea is to start with (2.11). Our numerical approach consists of "extracting" the stress-strain relationship out of the flux computation, i.e., we write (2.1) as

$$\boldsymbol{V}_t + \nabla \cdot c\underline{\boldsymbol{L}}(\boldsymbol{U}) = 0 \tag{2.17}$$

with

$$\boldsymbol{V} = \begin{pmatrix} w \\ \epsilon \\ \gamma \end{pmatrix}, \qquad \boldsymbol{U} = \begin{pmatrix} w \\ \sigma \\ \tau \end{pmatrix}, \qquad \underline{\boldsymbol{L}} = -\frac{1}{c}\begin{pmatrix} \sigma/\rho & \tau/\rho \\ w & 0 \\ 0 & w \end{pmatrix}.$$

According to this ansatz, our numerical scheme will possess two steps:

- First, we decompose (2.17) into advection equations (analogously to the elastic wave equation) by defining

$$\boldsymbol{R}_i := \boldsymbol{V} + \underline{\boldsymbol{L}}(\boldsymbol{U})\vec{\boldsymbol{n}}_i. \tag{2.18}$$

The corresponding advection equations

$$(\boldsymbol{R}_i)_t + \nabla \cdot \left(\boldsymbol{R}_i \vec{\boldsymbol{n}}_i^T c\right) = 0 \tag{2.19}$$

are solved independently for the time-step $[t^n, t^{n+1}]$ to compute the update

$$\boldsymbol{V}^{n+1} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{R}_i(t^{n+1}) \tag{2.20}$$

of the velocity and strain variables.

- Afterwards, the stress variables contained in $\boldsymbol{U}$ have to be updated by integrating (2.5) in the stress space, which means integrating the ODE

$$\begin{pmatrix} \dot\sigma \\ \dot\tau \end{pmatrix} = \begin{pmatrix} \frac{1}{\mu} + \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma^2 & \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma\tau \\ \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\sigma\tau & \frac{1}{\mu} + \frac{1}{\kappa^2}\left(\frac{1}{\mu_p(\kappa)} - \frac{1}{\mu}\right)\tau^2 \end{pmatrix}^{-1} \begin{pmatrix} \dot\epsilon \\ \dot\tau \end{pmatrix}, \tag{2.21}$$

or in simple form

$$\begin{pmatrix} \dot\sigma \\ \dot\tau \end{pmatrix} = \underline{\boldsymbol{C}}^{-1}(\sigma,\tau)\begin{pmatrix} \dot\epsilon \\ \dot\tau \end{pmatrix}, \tag{2.22}$$

which is an equivalent formulation of (2.5) for $\dot\sigma$ and $\dot\tau$ on the given time interval $[t^n, t^{n+1}]$ of time-step n. The problem arises that $\dot\epsilon(t)$ and $\dot\gamma(t)$ are not known $\forall t \in [t^n, t^{n+1}]$. We only know $\Delta\epsilon := \epsilon^{n+1} - \epsilon^n$ and $\Delta\gamma := \gamma^{n+1} - \gamma^n$ from the flux updates.

In order to solve the ODE (2.22) we have to reconstruct the strain path, i.e., the evolution of the strain variables on the time interval $t \in [0, t^{n+1} - t^n]$:

$$\begin{pmatrix} \epsilon^{n+1} \\ \gamma^{n+1} \end{pmatrix} - \begin{pmatrix} \epsilon^{n} \\ \gamma^{n} \end{pmatrix} = \boldsymbol{a}t + \boldsymbol{b}t^2 + \boldsymbol{c}t^3 + \cdots. \tag{2.23}$$

Since we can compute the time derivatives of the strain variables at times t^n and t^{n+1} by using

$$\dot{\epsilon}(t^*) = \partial_x w(t^*),$$
$$\ddot{\epsilon}(t^n) = \partial_x \dot{w}(t^n) = \frac{1}{\rho}\left(\frac{\partial^2}{\partial x^2}\sigma + \frac{\partial^2}{\partial x \partial y}\tau\right)(t^n), \tag{2.24}$$
$$\dot{\gamma}(t^*) = \partial_y w(t^*),$$
$$\ddot{\gamma}(t^n) = \partial_y \dot{w}(t^n) = \frac{1}{\rho}\left(\frac{\partial^2}{\partial x \partial y}\sigma + \frac{\partial^2}{\partial y^2}\tau\right)(t^n) \tag{2.25}$$
$$\text{at} \quad t^* = t^n \quad \text{or} \quad t^{n+1}$$

(w is known at times t^n and t^{n+1} and σ and τ are known at time t^n), we can reconstruct the strain variables up to order four and the time derivatives of the strain variables (and thus the right-hand side of the ODE (2.22)) up to order three—higher order can be achieved by using higher spatial derivatives.

If the matrix function $\underline{\boldsymbol{\mathcal{C}}}(\sigma, \tau)$ is not smooth at the transition from elasticity to plasticity, one has to restart the integration on the yield surface. The exact point on the yield surface where the transition from elasticity to plasticity takes place can be found by iteration, e.g., bisection (cf. [5]).

2.5. Boundary conditions. Throughout this paper we will consider the Cauchy-type problem for the elastic-plastic wave equation (2.11), i.e., the computational domain is $\mathbb{R}$ for the 1-D equation and $\mathbb{R}^2$ for the two-dimensional (2-D) equation, and we prescribe an initial solution for the velocity, strain, and stress variables at a certain time t_0 and then advance the solution forward in time. Consequently, the physical domain under consideration has no spatial boundaries and does not require the prescription of spatial boundary conditions.

3. Definition of stability. The classical proof of convergence for numerical schemes consists of proving consistency and stability. For explicit numerical schemes for solving a PDE of the form

$$\boldsymbol{W}^{n+1} = \boldsymbol{\Psi}[\boldsymbol{W}^n] \tag{3.1}$$

with the discrete solution $\boldsymbol{W}^n$ calculated using a grid $\Delta x \times \Delta t$ at time t^n, we define linear stability as follows.

DEFINITION 3.1 (linear stability). *A linear scheme of the form* (3.1) *is called stable if there exist constants* $C(T)$, Δx_0, *and* Δt_0 *independent of* Δt, Δx *with*

$$\|\boldsymbol{W}^N\| \le C\|\boldsymbol{W}^0\| \quad \forall \Delta t \le \Delta t_0,\ \forall \Delta x \le \Delta x_0 \tag{3.2}$$

with $N = T/\Delta t$.

This definition of stability (together with the requirement that the physical problem is well-posed) is sufficient for the Lax equivalence (cf. [19]) theorem to hold and to ensure convergence of a consistent linear scheme. Of course, a sufficient condition for linear stability is that the update operator $\boldsymbol{\Psi}$ is bounded by one, i.e.,

$$\left\|\boldsymbol{W}^{n+1}\right\| \le \|\boldsymbol{W}^n\|, \tag{3.3}$$

which one can prove for linear schemes with von Neumann analysis.

However, for nonlinear schemes a reasonable definition of nonlinear stability is more difficult. Our goal is to apply the convergence results for nonlinear schemes

obtained in [1], [17]. The following definition of nonlinear stability is sufficient for the convergence proof obtained by [1], [17] to be applicable to a consistent numerical scheme of the form (3.1).

DEFINITION 3.2 (nonlinear stability). *A consistent numerical scheme of the form* (3.1) *is called stable if for two arbitrary state vectors* W_1 *and* W_2 *there exists a constant* C *independent of* Δx *and* Δt *such that*

$$\|\boldsymbol{\Psi}(\boldsymbol{W}_1) - \boldsymbol{\Psi}(\boldsymbol{W}_2)\| \leq (1 + C\Delta t)\|\boldsymbol{W}_1 - \boldsymbol{W}_2\| \quad \forall \boldsymbol{W}_1, \boldsymbol{W}_2. \tag{3.4}$$

The definition of nonlinear stability is constructed in such a way that consistency and stability will be sufficient for convergence to the exact solution. However, consistency and stability in this sense will not be necessary for convergence, as in the linear case (cf. [24]).

Simply put, we define a scheme to be nonlinearly stable if the evolution of the difference in time t of two states is bounded by an expression proportional to e^{Ct}, which is sufficient (together with consistency) to show convergence. The convergence follows from [24] using the L_2-norm of the approximate solution $\boldsymbol{W}^n$ at a grid level, which is defined as the sum over the L_2-norm of all grid cells:

$$\|\boldsymbol{W}^n\| := \Delta x \sum_i \|\boldsymbol{W}_i^n\|_2 \quad \text{in one dimension,}$$

$$\|\boldsymbol{W}^n\| := \Delta x \Delta y \sum_{i,j} \|\boldsymbol{W}_{i,j}^n\|_2 \quad \text{in two dimensions.}$$

The system under consideration in (2.11) contains physical constants such as ρ and μ, which make the application of the L_2-norm problematic, as we show in the following.

Remark 3.3 (variable transform). When using the L_2-norm for the state-vector containing velocity, stress, and strain variables for the physical system as presented in (2.11), the problem occurs that the different components of the state-vector are measured in different physical units (since we left physical constants such as ρ and μ in the system) and hence the L_2-norm is meaningless, since it contains the sum of (squared) physical quantities measured in different physical units. Hence, we either have to use a weighed L_2-norm (i.e., an energy norm) or apply a method often used in physics (cf. [11]): transforming physical quantities and time so that all components of the state-vector are denoted in the same unit. To be more precise, we apply the following transformation:

$$\begin{aligned}
\text{Transformed variables} \qquad & \text{Physical variables} \\
t' &= ct \Longrightarrow \frac{\partial}{\partial t} = c\frac{\partial}{\partial t'}, \\
w' &= cw, \\
(\sigma', \tau') &= \left(\frac{\sigma}{\rho}, \frac{\tau}{\rho}\right), \\
(\epsilon', \gamma') &= (c^2\epsilon, c^2\gamma), \\
\kappa' &= \frac{\kappa}{\rho}, \\
\mu_p'(\kappa') &:= \frac{\mu_p(\kappa)}{\rho c^2}.
\end{aligned}$$

Rewriting our physical system (2.11) in these "transformed variables" yields a system of the same form, but $\rho' = c' = \mu' = 1$, measuring all quantities in the same unit,

which will be used throughout the following stability analysis. For simplicity we will drop the $'$ used for the transformed variables.

4. Stability for elastic-plastic waves in 1-D. The physical behavior of a solid undergoing stress can be classified according to three categories—elastic loading, elastic unloading, and plastic loading. Analogous to the physical behavior, our stability analysis in one dimension consists of three steps. First, we assume that we have purely elastic deformation only, which means that the stress-strain relationship is described by the line $0 \to A$ according to Figure 4.1, and hence our system, as well as our numerical scheme, is linear, and stability in the sense of Definition 3.1 can be proved with von Neumann analysis (cf. [10]).

In the second step, we still consider elastic deformation, but this time the stress-strain values may lie on a line which does not necessarily pass through the origin, e.g., line $B \to C$ according to Figure 4.1. Physically speaking, the material is assumed to be in an elastic state everywhere, but having a different plasticity history in the past, explaining why different cells have a different linear stress-strain relationship (e.g., in Figure 4.1 one cell can be in an elastic unloading process, described by a straight line not passing through the origin).

In the last step, we allow plastic deformation. First of all, we will prove boundedness (i.e., (3.3)) of the solution for plastic deformation, provided $0 < \mu_p(\kappa) \leq \mu$. However, a severe problem occurs for plastic deformation, explained in the following remark, leading to restriction of our further analysis.

Remark 4.1 (restriction of analysis). In the general case of plasticity described by a hysteresis model as outlined in section 2.2, it is no longer possible to write the update operator in the form

$$\boldsymbol{W}^{n+1} = \boldsymbol{\Psi}[\boldsymbol{W}^n] \tag{4.1}$$

since the yield condition depends on $\kappa_0(t) = \max_{t_0 \leq t' \leq t} \kappa(t')$; i.e., it depends on all past state vectors $\boldsymbol{W}^0, \boldsymbol{W}^1, \ldots, \boldsymbol{W}^n$. In order to analyze nonlinear stability for plastic waves in the sense of (3.4), we will restrict our analysis to the following special cases:

- Nonlinear elastic waves (cf. Figure 4.2): The stress-strain relationship is described by a nonlinear function $\epsilon = h(\sigma)$.
- Ideal plastic deformation (cf. Figure 4.1): $\kappa_0 \equiv \text{const}$ and $\mu_p(\kappa) \equiv 0$.

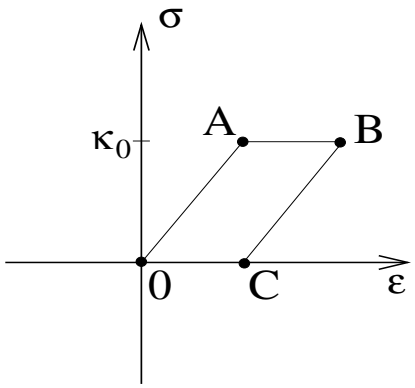


FIG. 4.1. *Hysteresis curve of an ideal plastic material. We consider three steps in our stability analysis: Elastic loading* $0 \to A$*, elastic unloading* $B \to C$*, and plastic loading along* $A \to B$*.*

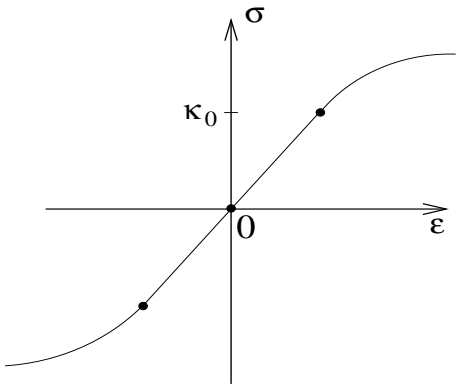


FIG. 4.2. *Stress-strain relationship of a nonlinear elastic material: The wave-speed decreases if* $|\sigma| > \kappa_0$.

In both cases the operator $\boldsymbol{\Psi}$ is of the form (4.1), as used in our definition of nonlinear stability (Definition 3.2).

4.1. Purely elastic waves. The first step of our analysis starts with the 1-D elastic wave equation (i.e., (2.3) restricted to one space dimension)

$$\boldsymbol{U}_t = \begin{pmatrix} w \\ \sigma \end{pmatrix}_t = \begin{pmatrix} \sigma \\ w \end{pmatrix}_x \quad \text{with } \mu = 1 \text{ (cf. Remark 3.3)}, \tag{4.2}$$

where the method of transport for elasticity is equivalent to the decomposition of the linear conservation law into right eigenvectors (cf. [10]). Hence, the 1-D right eigenvectors

$$\boldsymbol{R}^{\pm} = \frac{1}{2}\begin{pmatrix} w \pm \sigma \\ \sigma \pm w \end{pmatrix} \tag{4.3}$$

allow us to decompose the state- and flux-vectors as follows:

$$\begin{aligned} \boldsymbol{U} &= \boldsymbol{R}^{+} + \boldsymbol{R}^{-}, \\ \begin{pmatrix} \sigma \\ w \end{pmatrix}_x &= (\boldsymbol{R}^{+})_x - (\boldsymbol{R}^{-})_x. \end{aligned}$$

The cell update for a first order scheme with constant values in each cell reads

$$\boldsymbol{U}_i^{n+1} = (1-\lambda)\boldsymbol{U}_i^n + \lambda(\boldsymbol{R}_{i-1}^{+,n} + \boldsymbol{R}_{i+1}^{-,n}) \tag{4.4}$$

(λ = CFL-number), which can be written in the form

$$\boldsymbol{U}_i^{n+1} = (1-\lambda)\boldsymbol{U}_i^n + \lambda(\underline{\boldsymbol{A}}^{+}\boldsymbol{U}_{i-1}^n + \underline{\boldsymbol{A}}^{-}\boldsymbol{U}_{i+1}^n) \tag{4.5}$$

with the matrices

$$\underline{\boldsymbol{A}}^{\pm} := \frac{1}{2}\begin{pmatrix} 1 & \pm 1 \\ \pm 1 & 1 \end{pmatrix}.$$

Due to the linearity of the scheme, stability can be proven with von Neumann analysis, which consists of analyzing the norm of the solution $\boldsymbol{U}(x,t)$ in the Fourier space. The Fourier transform of the approximate solution reads

$$\mathcal{F}[\boldsymbol{U}](k,t^n) = \hat{\boldsymbol{U}}(k,t^n) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \boldsymbol{U}(x,t^n)e^{-ikx}\,dx. \tag{4.6}$$

Furthermore we need the identity

$$\mathcal{F}[\boldsymbol{U}(x + l\Delta x, t^n)](k, t^n) = \hat{\boldsymbol{U}}(k, t^n)e^{-il\xi} \tag{4.7}$$

with $\xi = k\Delta x$. Applying the Fourier transform to (4.4) leads to the amplification matrix

$$\underline{\boldsymbol{G}}(\xi) = (1 - \lambda)\underline{\boldsymbol{I}} + \lambda(e^{i\xi}\underline{\boldsymbol{A}}^+ + e^{-i\xi}\underline{\boldsymbol{A}}^-), \tag{4.8}$$

which has the eigenvalues $r_{1,2}$,

$$r_{1,2} = 1 - \lambda(1 - e^{\pm i\xi}). \tag{4.9}$$

Obviously, we have $|r_{1,2}| \leq 1$ for $\lambda \in [0, 1]$, which is necessary for stability.

Moreover, a sufficient condition for stability is that the L_2-norm of the matrix $\underline{\boldsymbol{G}}(\xi)$ is bounded by one; i.e., all eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ are bounded by one. We have the following.

LEMMA 4.2. *For $\lambda = 1$, the scheme* (4.4) *satisfies*

$$\|\boldsymbol{U}^{n+1}\|_2 = \|\boldsymbol{U}^n\|_2. \tag{4.10}$$

Proof. For $\lambda = 1$, the two eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ are

$$r_{1,2} = 1. \qquad \square$$

PROPOSITION 4.3. *Scheme* (4.4) *is stable according to Definition* 3.1 $\forall\lambda \in [0, 1]$.

Proof. We consider the L_2-norm of $\boldsymbol{U}^{n+1}$ computed according to (4.4). Obviously, the squared norm $\|\boldsymbol{U}^{n+1}\|_2^2$ is a polynomial of degree 2 in λ. Since

$$\begin{aligned} \|\boldsymbol{U}^{n+1}\|_2^2(\lambda) &\geq 0 \quad \forall\lambda \in \mathbb{R}, \\ \|\boldsymbol{U}^{n+1}\|_2^2(\lambda = 0) &= \|\boldsymbol{U}^n\|_2^2, \\ \|\boldsymbol{U}^{n+1}\|_2^2(\lambda = 1) &= \|\boldsymbol{U}^n\|_2^2, \end{aligned}$$

and because of $\lim_{\lambda\to\pm\infty} \|\boldsymbol{U}^{n+1}\|_2^2(\lambda) = \infty$, this polynomial has to be convex. Hence,

$$\|\boldsymbol{U}^{n+1}\|_2(\lambda) \leq \|\boldsymbol{U}^n\|_2 \quad \forall\lambda \in [0, 1]. \qquad \square \tag{4.11}$$

4.2. Elastic loading and unloading waves. The second step of the stability analysis, as sketched in Figure 4.1, focuses on the case when the stress-strain relationship is still linear, but not necessarily on a straight line through the origin (e.g., line $B \to C$); for example, a cell could be in an elastic unloading process.

Our numerical scheme consists of two steps now—updating velocity and strain by the method of transport and then updating the stress. To apply the method of transport (cf. (2.19) and (2.20) in two dimensions), we decompose the PDE in the 1-D case,

$$\boldsymbol{V}_t = \begin{pmatrix} w \\ \epsilon \end{pmatrix}_t = \begin{pmatrix} \sigma \\ w \end{pmatrix}_x, \tag{4.12}$$

into transported quantities $\boldsymbol{R}^\pm$ as follows:

$$\boldsymbol{R}^\pm := \frac{1}{2}\begin{pmatrix} w \pm \sigma \\ \epsilon \pm w \end{pmatrix}. \tag{4.13}$$

Thus, we update the velocity and strain variables in each cell by solving the advection equations,

$$\boldsymbol{R}_t^+ + (\boldsymbol{R}^+)_x = 0,$$
$$\boldsymbol{R}_t^- - (\boldsymbol{R}^-)_x = 0,$$

which yields an update formula for $\boldsymbol{V} = (w, \epsilon)$ with the CFL-number λ:

$$\boldsymbol{V}_i^{n+1} = (1-\lambda)\boldsymbol{V}_i^n + \lambda(\boldsymbol{R}_{i-1}^{+,n} + \boldsymbol{R}_{i+1}^{-,n}). \tag{4.14}$$

Afterwards, the stress is updated according to Hooke's law (with $\mu = 1$ due to Remark 3.3),

$$\sigma_i^{n+1} = \sigma_i^n + (\epsilon_i^{n+1} - \epsilon_i^n). \tag{4.15}$$

The stress-strain relationship is obviously linear but can be described by a different linear equation in each cell, which generalizes the analysis of the previous section, where each cell was described by the same linear stress-strain relationship. Physically speaking, each cell can be in a different elastic loading or unloading process. We will write these two steps (4.14), (4.15) as a one-step update for the vector $\boldsymbol{W}_i^n = (w_i^n, \epsilon_i^n, \sigma_i^n)$.

With the matrices $\underline{\boldsymbol{A}}^+$, $\underline{\boldsymbol{A}}^-$, $\underline{\boldsymbol{A}}^0$, and $\underline{\boldsymbol{E}}$ defined as follows:

$$\underline{\boldsymbol{A}}^\pm := \frac{1}{2}\begin{pmatrix} 1 & 0 & \pm 1 \\ \pm 1 & 1 & 0 \\ \pm 1 & 1 & 0 \end{pmatrix},$$
$$\underline{\boldsymbol{A}}^0 := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$
$$\boldsymbol{E} := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

We can write the scheme (4.14), (4.15) as one step:

$$\boldsymbol{W}_i^{n+1} = \lambda[\underline{\boldsymbol{A}}^+\boldsymbol{W}_{i-1}^n + \underline{\boldsymbol{A}}^-\boldsymbol{W}_{i+1}^n]\boldsymbol{W}_i^n + \underline{\boldsymbol{A}}^0\boldsymbol{W}_i^n + (1-\lambda)\underline{\boldsymbol{E}}\boldsymbol{W}_i^n. \tag{4.16}$$

Since the scheme (4.16) is still linear, we apply von Neumann analysis, which means we analyze the amplification matrix

$$\underline{\boldsymbol{G}}(\xi) = \lambda[\underline{\boldsymbol{A}}^+ e^{i\xi} + \underline{\boldsymbol{A}}^- e^{-i\xi}] + \underline{\boldsymbol{A}}^0 + (1-\lambda)\underline{\boldsymbol{E}}. \tag{4.17}$$

The eigenvalues r_i of the complex non-Hermitian matrix $\underline{\boldsymbol{G}}(\xi)$ turn out to be

$$r_1 = 1,$$
$$r_{2,3} = 1 - \lambda + \lambda e^{\pm i\xi},$$

which are all bounded by one $\forall \lambda \in [0,1]$, which is necessary for stability.

A sufficient condition for stability is that the eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ are bounded by one. The calculation of the eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ yields lengthy expressions for the eigenvalues, which are too complicated to be discussed analytically. Thus, we perform a computational analysis to verify the following.

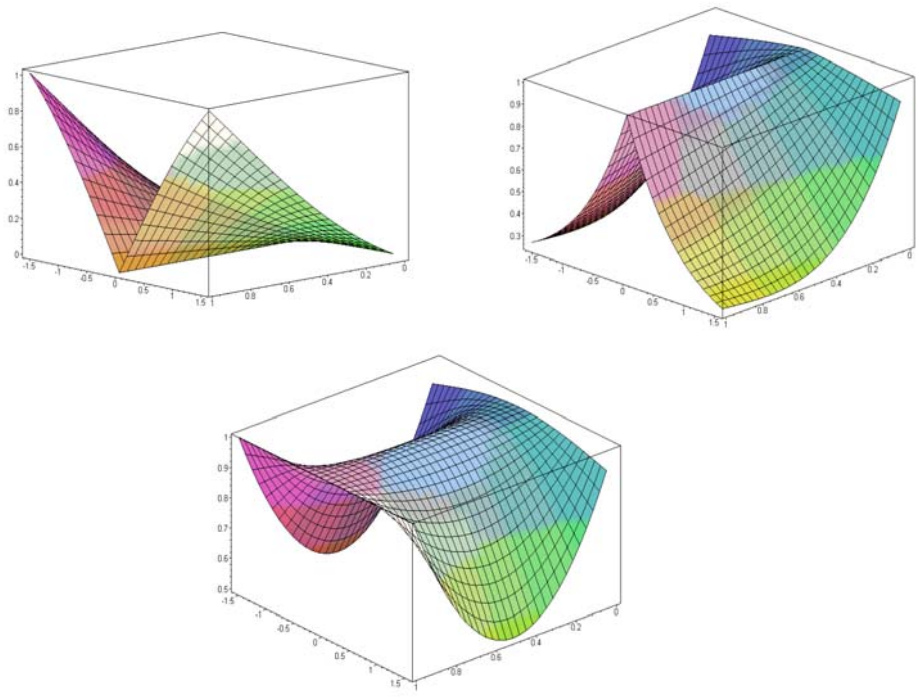


FIG. 4.3. *Three eigenvalues of the matrix* $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$*, plotted over the region* $(\lambda,\xi) \in [0,1] \times [-\pi/2,\pi/2]$*, showing their boundedness between zero and one.*

Conjecture 4.4 (boundedness of eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$). The eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ with the amplification matrix defined in (4.17) are bounded by one $\forall\lambda \in [0,1]$.

Proof. The three eigenvalues of the matrix $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ plotted in Figure 4.3 for $(\lambda,\xi) \in [0,1] \times [-\pi/2,\pi/2]$ show that these three eigenvalues are bounded by one. ∎

Consequently, we can conclude that

$$\|\boldsymbol{W}^{n+1}\|_2 \leq \|\boldsymbol{W}^n\|_2. \tag{4.18}$$

It is important to mention that our further investigations for plastic waves are independent of this result for the method of transport; i.e., the following results can be applied to any numerical scheme fulfilling the stability condition (4.18) for elastic waves.

4.3. Boundedness for elastic-plastic waves. The third step of our stability analysis generalizes to the case of plastic deformation, i.e., the stress update in (4.15) has to be replaced by the integral equation

$$\sigma_i^{n+1} = \sigma_i^n + \int_{\epsilon_i^n}^{\epsilon_i^{n+1}} \mu_p(\sigma(\epsilon))d\epsilon, \tag{4.19}$$

which yields an update for σ. For most materials (cf. [18]) $\mu_p(\sigma)$ is smaller for plasticity than for elasticity, i.e.,

$$\mu_p(\sigma) = \begin{cases} \mu = 1 & \text{if } |\sigma| \leq \kappa_0, \\ \leq \mu = 1 & \text{otherwise,} \end{cases} \tag{4.20}$$

with

$$\kappa_0(t) = \max_{0 \leq t' \leq t} |\sigma(t)|,$$

which will be used as an assumption in the following proposition. Equation (4.20) shows that the following analysis covers not only purely plastic deformation but also the case where deformation is first elastic and then becomes plastic.

PROPOSITION 4.5. *If* (4.18) *holds for elastic waves* $\forall \lambda \in [0,1]$ *and* $0 < \mu_p(\kappa) \leq \mu = 1$ $\forall \kappa \in \mathbb{R}^+$*, then the update operator* $\boldsymbol{\Psi}$ *of the scheme is also bounded by one for plasticity* $\forall \lambda \in [0,1]$*, i.e.,*

$$\|\boldsymbol{W}^{n+1}\|_2 \leq \|\boldsymbol{W}^n\|_2.$$

Proof. The basic idea of the proof is simple: With a given solution $\boldsymbol{W}^n$ at time level t^n we compute two different updates: one purely elastic update $\boldsymbol{W}_{el}^{n+1}$ computed according to (4.15) neglecting the existence of plasticity, and the plastic update $\boldsymbol{W}_{pl}^{n+1}$ computed according to (4.19). Since the elastic update of the stress σ_{el}^{n+1} cannot be smaller than the plastic update if $\mu_p(\kappa) \leq \mu = 1$, i.e.,

$$|\sigma_{pl}^{n+1}| \leq |\sigma_{el}^{n+1}|, \tag{4.21}$$

the norm of the plastic update of the vector $\boldsymbol{W}_{pl}^{n+1}$ cannot be larger than the elastic update $\boldsymbol{W}_{el}^{n+1}$:

$$\|\boldsymbol{W}_{pl}^{n+1}\|_2 \leq \|\boldsymbol{W}_{el}^{n+1}\|_2 \leq \|\boldsymbol{W}^n\|_2. \qquad \square$$

Hence, boundedness by one of the update operator for elasticity implies boundedness for plastic waves as well. It is noteworthy that this result is independent of the numerical scheme used for solving the fluxes, e.g., the method of transport.

4.4. Nonlinear stability for elastic-plastic waves. However, since we have a nonlinear scheme for plastic waves, boundedness of the operator is no longer sufficient for stability in the sense of Definition 3.2.

Therefore, we have to use (3.4) from Definition 3.2, which means that we have to investigate the difference between two solutions $\Delta \boldsymbol{W}_i^n = \bar{\boldsymbol{W}}_i^n - \boldsymbol{W}_i^n$ with

$$\begin{aligned} \boldsymbol{W}_i^n &:= (w_i^n, \epsilon_i^n, \sigma_i^n), \\ \bar{\boldsymbol{W}}_i^n &:= (\bar{w}_i^n, \bar{\epsilon}_i^n, \bar{\sigma}_i^n) \end{aligned}$$

in order to prove nonlinear stability.

As for the proof of boundedness, the basic idea is to prove that plasticity can only decrease the norm of the difference between two states relative to elasticity. Therefore, one computes an elastic update for both $\boldsymbol{W}_i$ and $\bar{\boldsymbol{W}}_i$ ignoring the existence of plasticity and then compares it to the plastic update. Provided the stability condition (4.18) holds for elastic waves, then we can already conclude a relation of the form

$$\|\Delta \boldsymbol{W}^{n+1}\|_2 \leq \|\Delta \boldsymbol{W}^n\|_2 \tag{4.22}$$

for elastic waves, since in that case $\boldsymbol{\Psi}[\boldsymbol{W}^n - \bar{\boldsymbol{W}}^n] = \boldsymbol{\Psi}[\boldsymbol{W}^n] - \boldsymbol{\Psi}[\bar{\boldsymbol{W}}^n]$ due to the linearity of the scheme for elastic waves. Thus, it is sufficient to show that plasticity can only reduce the norm of $\|\Delta \boldsymbol{W}^{n+1}\|_2$; i.e., we have to show

$$(\Delta\sigma_{pl}^{n+1})^2 \le (\Delta\sigma_{el}^{n+1})^2 \tag{4.23}$$

in each cell, where the indices *el* and *pl* distinguish between an elastic and a plastic stress update.

For reasons explained above in Remark 4.1, we will analyze only the special cases of an ideal plastic material and nonlinear elastic material.

4.4.1. Ideal plastic material. For such material, the stress-strain relationship can be described as follows:

$$\begin{aligned} \kappa(t) &= \sqrt{\sigma^2(t)}, \\ d\sigma &= \begin{cases} 0 & \text{if } \kappa = \kappa_0 \text{ and } \dot{\kappa} > 0, \\ d\epsilon & \text{otherwise,} \end{cases} \end{aligned} \tag{4.24}$$

which implies

$$\kappa_0 \equiv \text{const} \quad \text{and} \quad |\sigma(t)| \le \kappa_0. \tag{4.25}$$

We have the following proposition.

PROPOSITION 4.6. *The update operator* $\boldsymbol{\Psi}$ *for plastic waves in a material with the plasticity model described in* (4.24) *is nonlinearly stable in the sense of Definition* 3.2*; i.e., for the difference between two solutions* $\Delta \boldsymbol{W}^n$ *we have*

$$\|\Delta \boldsymbol{W}^{n+1}\|_2 \le \|\Delta \boldsymbol{W}^n\|_2.$$

Proof. If σ_{el}^{n+1} and $\bar{\sigma}_{el}^{n+1}$ denote the elastic stress update in a cell i neglecting the occurrence of plasticity, then the plastic update in this cell reads

$$\begin{aligned} \sigma_{pl}^{n+1} &= \operatorname{sgn}(\sigma_{el}^{n+1}) \min\{|\sigma_{el}^{n+1}|, \kappa_0\}, \\ \bar{\sigma}_{pl}^{n+1} &= \operatorname{sgn}(\bar{\sigma}_{el}^{n+1}) \min\{|\bar{\sigma}_{el}^{n+1}|, \kappa_0\}. \end{aligned}$$

Thus, one easily verifies

$$|\sigma_{pl}^{n+1} - \bar{\sigma}_{pl}^{n+1}| \le |\sigma_{el}^{n+1} - \bar{\sigma}_{el}^{n+1}|, \tag{4.26}$$

which implies

$$\|\Delta \boldsymbol{W}_{pl}^{n+1}\|_2 \le \|\Delta \boldsymbol{W}_{el}^{n+1}\|_2 \le \|\Delta \boldsymbol{W}^n\|_2. \qquad \square \tag{4.27}$$

4.4.2. Nonlinear elastic material. The second case we want to study is a nonlinear elastic material, where the stress-strain relationship is described by a nonlinear function $\sigma = h(\epsilon)$ (cf. Figure 4.2). If the elastic shear modulus $\mu = 1$ is a Lipschitz constant for the function f (which implies that the nonlinear wave-speed is smaller than the elastic speed), then we have the following.

PROPOSITION 4.7. *For stress-strain relationship of the form* $\sigma = h(\epsilon)$ *with* $\mu = 1$ *being a Lipschitz constant of the function* $h(\cdot)$*, the operator* $\boldsymbol{\Psi}$ *is stable in the sense of Definition* 3.2.

Proof. Let the index pl and el distinguish between the nonlinear and the linear elastic updates, whence in a cell i we have

$$|\sigma_{pl}^{n+1} - \bar{\sigma}_{pl}^{n+1}| = |h(\epsilon^{n+1}) - h(\bar{\epsilon}^{n+1})| \leq |\epsilon^{n+1} - \bar{\epsilon}^{n+1}| = |\sigma_{el}^{n+1} - \bar{\sigma}_{el}^{n+1}| \tag{4.28}$$

and thus

$$\|\Delta \boldsymbol{W}_{pl}^{n+1}\|_2 \leq \|\Delta \boldsymbol{W}_{el}^{n+1}\|_2 \leq \|\Delta \boldsymbol{W}^{n}\|_2. \tag{4.29}$$ ∎

It is noteworthy that these two results for nonlinear stability are independent of the method used for solving the PDE; i.e., they can be used for schemes other than the method of transport as well.

5. Stability for elastic-plastic waves in two dimensions. In this section, we will analyze the stability of our scheme in two dimensions. In principle, one can follow the same three steps and the same ideas as for one dimension. In two dimensions, the method of transport based on the decomposition of the PDE into advection equations is a generalization of the decomposition of a linear conservation law into right eigenvectors in one dimension.

5.1. Stability for elastic waves. As in one dimension, the first step of our stability analysis focuses on an purely elastic material. As mentioned above (cf. (2.13)–(2.16)), the idea of the method of transport is to decompose a PDE into advection equations describing the transport of some quantities $\boldsymbol{R}_i$. For elastic waves, these quantities can be written as

$$\boldsymbol{R}_i := \boldsymbol{U} + \underline{\boldsymbol{L}}\vec{\boldsymbol{n}}_i = \underline{\boldsymbol{A}}_i \boldsymbol{U} \tag{5.1}$$

with

$$\underline{\boldsymbol{A}}_i = \begin{pmatrix} 1 & n_i^1 & n_i^2 \\ n_i^1 & 1 & 0 \\ n_i^2 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \vec{\boldsymbol{n}}_i = \begin{pmatrix} n_i^1 \\ n_i^2 \end{pmatrix}. \tag{5.2}$$

We call λ the CFL-number of the scheme. In the following we use four diagonal waves, i.e.,

$$\vec{\boldsymbol{n}}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \vec{\boldsymbol{n}}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad \vec{\boldsymbol{n}}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \qquad \vec{\boldsymbol{n}}_4 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \tag{5.3}$$

which fulfill the consistency relation (2.14).

We assume a Cartesian grid, and consequently λ^2 is the part of the quantity $\boldsymbol{R}_i$ in a certain cell which is transported into the corner neighbors and $\lambda(1-\lambda)$ is the part which is transported into the direct neighbor cells (cf. Figure 5.1).

With the above assumptions, the update formula of the scheme for the state vector $\boldsymbol{U}_{ij}$ in the cell (i,j) can be formulated:

$$\begin{aligned} \boldsymbol{U}_{ij}^{n+1} = {} & [1 - \lambda^2 - 2\lambda(1-\lambda)]\boldsymbol{U}_{i,j}^n \\ & + \{\lambda^2[(\boldsymbol{R}_1^n)_{i-1,j-1} + (\boldsymbol{R}_3^n)_{i+1,j-1} + (\boldsymbol{R}_2^n)_{i-1,j+1} + (\boldsymbol{R}_4^n)_{i+1,j+1}] \\ & + \lambda(1-\lambda)[(\boldsymbol{R}_1^n)_{i-1,j} + (\boldsymbol{R}_4^n)_{i-1,j} + (\boldsymbol{R}_2^n)_{i-1,j} + (\boldsymbol{R}_3^n)_{i+1,j}] \\ & + \lambda(1-\lambda)[(\boldsymbol{R}_3^n)_{i,j-1} + (\boldsymbol{R}_4^n)_{i,j+1} + (\boldsymbol{R}_1^n)_{i,j-1} + (\boldsymbol{R}_2^n)_{i,j+1}]\}/4. \end{aligned} \tag{5.4}$$

Taking into account (5.1) and (5.2), the quantities $\boldsymbol{R}_k$ used in (5.4) can be expressed by the $\boldsymbol{U}_{ij}^n$.

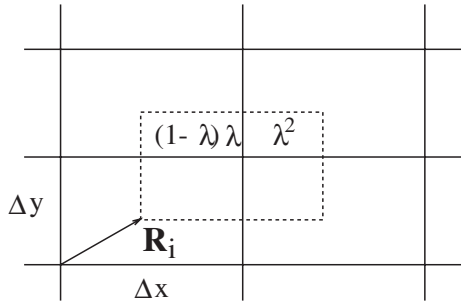


FIG. 5.1. *Advection of the quantity $\boldsymbol{R}_i$ in a cell into the direction $\vec{\boldsymbol{n}}_1 = (1,1)^T$. The part transported into the corner cell is λ^2, and the part transported into each of the two direct neighboring cells is $\lambda(1-\lambda)$, where λ denotes the CFL-number.*

Since the scheme is linear, we can apply the well-known von Neumann analysis to show stability for the scheme (5.4). Therefore, we analyze the Fourier transform of $\boldsymbol{U}(x,y,t)$,

$$\begin{aligned}\mathcal{F}[\boldsymbol{U}](k_1,k_2,t^n) &= \hat{\boldsymbol{U}}(k_1,k_2,t^n)\\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\boldsymbol{U}(x,y,t^n)e^{-i(k_1x+k_2y)}\,dx\,dy.\end{aligned} \tag{5.5}$$

Obviously we have

$$\mathcal{F}[\boldsymbol{U}(x+l\Delta x, y+m\Delta y, t^n)](k_1,k_2,t^n) = \hat{\boldsymbol{U}}(k_1,k_2,t^n)e^{-i(l\xi+m\eta)} \tag{5.6}$$

with $\xi = k_1\Delta x$ and $\eta = k_2\Delta y$. Hence, in the Fourier space the update scheme can be written as

$$\hat{\boldsymbol{U}}(k_1,k_2,t^{n+1}) = \underline{\boldsymbol{G}}(\xi,\eta)\hat{\boldsymbol{U}}(k_1,k_2,t^n). \tag{5.7}$$

In order to prove stability for the scheme (5.4) it is sufficient to show that the amplification matrix $\underline{\boldsymbol{G}}(\xi,\eta)$ is bounded in the L_2-norm; i.e., the eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ are less than or equal to one. We have the following.

PROPOSITION 5.1 (boundedness for $CFL = 1$). *For $CFL = 1$, the eigenvalues of $\underline{\boldsymbol{G}}^H(\xi,\eta)\underline{\boldsymbol{G}}(\xi,\eta)$ on a Cartesian grid are bounded by one, and hence the scheme is linearly stable according to Definition* 3.1.

Proof. Using (5.5) in (5.4), the eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ can be found:

$$\begin{aligned}r_1 &= -\frac{1}{8}\cos(2\xi+2\eta) + \frac{1}{4}\cos(2\eta) - \frac{1}{8}\cos(2\eta-2\xi) + \frac{1}{4}\cos(2\xi) + \frac{3}{4},\\ r_{2,3} &= \frac{1}{8}\cos(2\xi+2\eta) + \frac{1}{4}\cos(2\eta) + \frac{1}{8}\cos(2\eta-2\xi) + \frac{1}{4}\cos(2\xi) + \frac{1}{4},\end{aligned}$$

which are bounded by one. Hence, for $CFL = 1$ the scheme is stable.

Further, numerical plots of the eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ confirm that stability holds $\forall\lambda \in [0;1]$. We omit the plots of eigenvalues in this step, since we show them below in the more general context of the second step. ∎

5.2. Stability for elastic loading and unloading waves. As in one dimension, the second step of our analysis still focuses on elastic waves, but now allowing

different cells to be in a different hysteresis state (i.e., different cells have different linear stress-strain relationships, as indicated in Figure 4.1). Again, we can rewrite our two-step scheme:

- Updating w, ϵ, and γ using the method of transport, i.e., (2.18)–(2.20).
- Then updating σ and τ using the stress-strain relationship

$$\sigma_{ij}^{n+1} = \sigma_{ij}^{n} + (\epsilon_{ij}^{n+1} - \epsilon_{ij}^{n}),$$
$$\tau_{ij}^{n+1} = \tau_{ij}^{n} + (\gamma_{ij}^{n+1} - \gamma_{ij}^{n})$$

as a one-step update scheme for the vector $\boldsymbol{W} = (w, \epsilon, \gamma, \sigma, \tau)$. With the matrix

$$\underline{\boldsymbol{A}}_i := \frac{1}{4}\begin{pmatrix} 1 & 0 & 0 & n_i^1 & n_i^2 \\ n_i^1 & 1 & 0 & 0 & 0 \\ n_i^2 & 0 & 1 & 0 & 0 \\ n_i^1 & 1 & 0 & 0 & 0 \\ n_i^2 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{5.8}$$

we can define the transported quantities

$$\boldsymbol{R}_i := \underline{\boldsymbol{A}}_i \boldsymbol{W}, \tag{5.9}$$

where i is the index of one of the four diagonal waves according to (5.3). Furthermore, we define

$$\underline{\boldsymbol{A}}_0 := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

$$\underline{\boldsymbol{B}} := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix}.$$

Hence, our scheme for the update of $\boldsymbol{W}_{ij}^n$ in cell (i, j) reads

$$\begin{aligned} \boldsymbol{W}_{ij}^{n+1} = {} & [1 - \lambda^2 - 2\lambda(1-\lambda)]\underline{\boldsymbol{A}}_0 \boldsymbol{W}_{i,j}^n + \underline{\boldsymbol{B}} \boldsymbol{W}_{i,j}^n \\ & + \{\lambda^2[(\boldsymbol{R}_1^n)_{i-1,j-1} + (\boldsymbol{R}_3^n)_{i+1,j-1} + (\boldsymbol{R}_2^n)_{i-1,j+1} + (\boldsymbol{R}_4^n)_{i+1,j+1}] \\ & + \lambda(1-\lambda)[(\boldsymbol{R}_1^n)_{i-1,j} + (\boldsymbol{R}_4^n)_{i+1,j} + (\boldsymbol{R}_2^n)_{i-1,j} + (\boldsymbol{R}_3^n)_{i+1,j}] \\ & + \lambda(1-\lambda)[(\boldsymbol{R}_3^n)_{i,j-1} + (\boldsymbol{R}_4^n)_{i,j+1} + (\boldsymbol{R}_1^n)_{i,j-1} + (\boldsymbol{R}_2^n)_{i,j+1}]\}/4. \end{aligned} \tag{5.10}$$

Analogous to the 1-D case, we compute the amplification matrix as

$$\begin{aligned} \underline{\boldsymbol{G}}(\xi, \eta) = {} & [1 - \lambda^2 - 2\lambda(1-\lambda)]\underline{\boldsymbol{A}}_0 + \underline{\boldsymbol{B}} \\ & + \{\lambda^2[e^{i\xi+i\eta}\underline{\boldsymbol{A}}_1 + e^{-i\xi+i\eta}\underline{\boldsymbol{A}}_3 + e^{i\xi-i\eta}\underline{\boldsymbol{A}}_2 + e^{-i\xi-i\eta}\underline{\boldsymbol{A}}_4] \\ & + \lambda(1-\lambda)[e^{i\xi}\underline{\boldsymbol{A}}_1 + e^{-i\xi}\underline{\boldsymbol{A}}_4 + e^{i\xi}\underline{\boldsymbol{A}}_2 + e^{-i\xi}\underline{\boldsymbol{A}}_3] \\ & + \lambda(1-\lambda)[e^{i\eta}\underline{\boldsymbol{A}}_3 + e^{-i\eta}\underline{\boldsymbol{A}}_4 + e^{i\eta}\underline{\boldsymbol{A}}_1 + e^{-i\eta}\underline{\boldsymbol{A}}_2]\}/4, \end{aligned} \tag{5.11}$$

which is a complex non-Hermitian matrix. As in one dimension, we come to the following conjecture.

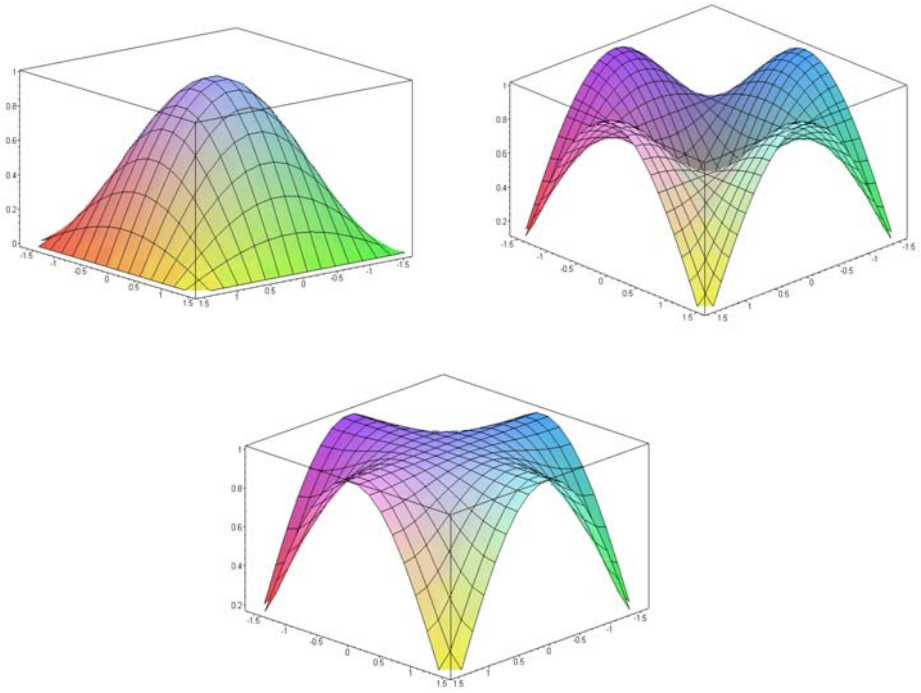


FIG. 5.2. *Three eigenvalues of the matrix* $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ *for* $CFL = \lambda = 1$*, plotted over the region* $(\xi, \eta) \in [-\pi/2, \pi/2]^2$*, showing their boundedness between zero and one. The remaining two eigenvalues are omitted, since they are constantly one.*

Conjecture 5.2 (boundedness of eigenvalues of amplification matrix). *The eigenvalues of the update matrix* $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ *with the amplification matrix defined in* (5.11) *on a Cartesian grid are bounded by one for* $CFL = \lambda = 1$.

Proof. The eigenvalues of the amplification matrix $\underline{\boldsymbol{G}}^H(\xi, \eta)\underline{\boldsymbol{G}}(\xi, \eta)$ plotted in Figure 5.2 show the eigenvalues for $\lambda = 1$, which are bounded by one.

It is important to mention that further computations of the eigenvalues of $\underline{\boldsymbol{G}}^H\underline{\boldsymbol{G}}$ for CFL-numbers between zero and one illustrates that stability holds $\forall\lambda \in [0, 1]$. □

5.3. Stability for elastic-plastic waves. Analogous to the 1-D case, one can argue that the norm of the difference of two solutions $\Delta\boldsymbol{W}$ can only decrease when plasticity occurs if we make similar assumptions for the stress-strain relationship as in 1-D, i.e., ideal plasticity. The argument used is the same as in one dimension: Plasticity yields smaller stress updates than elasticity as long as the plastic wavespeed is not greater than the elastic one.

6. Conclusion. The "classical" convergence proof for ODE and linear PDE solvers is based on two properties of a numerical scheme—consistency and stability. For our numerical scheme for solving the (nonlinear) elastic-plastic wave equation using the method of transport as underlying numerical scheme, we used a definition of nonlinear stability, which is based on a limit for the evolution of the difference between two numerical solutions in time. On the basis of this stability

definition it is straightforward to obtain convergence proofs for a consistent numerical scheme.

For purely elastic waves where our scheme is linear, and hence boundedness of the update operator is sufficient for stability, we were able to show stability with von Neumann analysis in one and two dimensions. For plastic waves, where the system to be simulated nonlinear, as, consequently, is our scheme, we were able to show boundedness of the update operator, provided that the plastic wave-speed is not higher than the elastic wave-speed in one dimension. Furthermore, we showed nonlinear stability for two special cases in one dimension—nonlinear elastic waves and ideal plastic waves, where the shear modulus reduces to zero in the plastic zone. The basic idea for the proof of boundedness and nonlinear stability is the observation that plasticity only decreases the value of the stress variables, provided the material "weakens" in plastic zones. Furthermore, it is straightforward to generalize this argument to two dimensions.

It is noteworthy that our stability results in the nonlinear (i.e., plastic) case are very general in the sense that they are independent of the numerical scheme used for solving the PDE, and they are based on the observation that the occurrence of plasticity only "improves" stability, provided the material becomes weaker in the plastic zone, which is the case for almost all materials. Moreover, the analysis performed is not restricted to one or two space dimensions—a generalization to the wave equation in three dimensions is straightforward.

REFERENCES

[1] R. Bodenmann and H. J. Schroll, *Compact difference methods applied to initial-boundary value problems for mixed systems*, Numer. Math., 73 (1996), pp. 291–309.

[2] P. Colella, *Multidimensional upwind methods for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 171–200.

[3] M. Fey, *Multidimensional upwinding. Part* I. *The method of transport for solving the Euler equations*, J. Comput. Phys., 143 (1998), pp. 159–180.

[4] M. Fey, *Multidimensional upwinding. Part* II. *Decomposition of the Euler equations into advection equations*, J. Comput. Phys., 143 (1998), pp. 181–199.

[5] G. Giese and M. Fey, *A genuinely multi-dimensional high-resolution scheme for the elastic-plastic wave equation*, J. Comput. Phys., 181 (2002), pp. 1–16.

[6] P. L. Gould, *Introduction to Linear Elasticity*, Springer-Verlag, New York, 1983.

[7] M. Fey, R. Jeltsch, and A.-T. Morel, *Multidimensional schemes for nonlinear systems of hyperbolic conservation laws*, in Numerical Analysis 1995, D. Griffiths and G. A. Watson, eds., Longman, Harlow, UK, 1996.

[8] M. Fey, R. Jeltsch, J. Maurer, and A.-T. Morel, *The method of transport for nonlinear systems of hyperbolic conservation laws in several space dimensions*, in Proceedings of the Conference on Numerical Analysis, VSP International Science, Zeist, The Netherlands, 1996.

[9] D. Kröner, *Numerical Schemes for Conservation Laws*, Wiley Teubner, Stuttgart, 1997.

[10] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.

[11] X. Lin, *Numerical Computation of Stress Waves in Solids*, Akademie Verlag, Berlin, 1996.

[12] X. Lin and J. Ballmann, *Improved bicharacteristic schemes for two-dimensional elastodynamic equations*, Quart. Appl. Math., 53 (1995), pp. 383–398.

[13] X. Lin and J. Ballmann, *A numerical scheme for axisymmetric elastic waves in solids*, Wave Motion, 21 (1995), pp. 115–126.

[14] X. Lin and J. Ballmann, *Numerical modelling of elastic-plastic waves in transversely isotropic composite materials*, ZAMM Z. Angew. Math. Mech., 75 (1995), pp. 267–268.

[15] X. Lin and J. Ballmann, *Elastic-plastic waves in cracked solids under plane stress*, in Nonlinear Waves in Solids, J. L. Wegner and F. R. Norwood, eds., Appl. Mech. Rev. 137, American Society of Mechanical Engineers, New York, 1995, pp. 155–160.

[16] X. Lin and J. Ballmann, *Numerical modelling of elastic-plastic deformation at crack tips in*

composite material under stress wave loading, J. Phys. III, 4 (1994), pp. 53–58.
[17] J. Lorenz and H. J. Schroll, *Stiff well-posedness for hyperbolic systems with large relaxation terms*, Adv. Differential Equations, 2 (1997), pp. 643–666.
[18] J. Lubliner, *Plasticity Theory*, Macmillan, New York, 1990.
[19] K. W. Morton and R. D. Richtmyer, *Difference Methods for Initial-Value Problems*, Wiley-Interscience, New York, 1967.
[20] W. K. Nowacki, *Stress Waves in Non-Elastic Solids*, Pergamon Press, Oxford, UK, 1978.
[21] L. Prandtl, *Spannungverteilung in plastischen Körpern*, in Proceedings of the First International Congress for Applied Mechanics, C. B. Biezeno and J. M. Burgers, eds., Delft, The Netherlands, 1924, p. 43.
[22] P. Roe, *Approximate Riemann solvers, parameter vectors and difference schemes*, J. Comput. Phys., 43 (1981), pp. 357–372.
[23] D. Serre, *Systems of Conservation Laws*, Cambridge University Press, Cambridge, UK, 1999.
[24] H. J. Schroll, *Convergence of implicit finite difference methods applied to nonlinear mixed systems*, SIAM J. Numer. Anal., 33 (1996), pp. 997–1013.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2590–2611

DUAL-PRIMAL FETI ALGORITHMS FOR EDGE ELEMENT APPROXIMATIONS: TWO-DIMENSIONAL H AND P FINITE ELEMENTS ON SHAPE-REGULAR MESHES*

ANDREA TOSELLI† AND XAVIER VASSEUR†

Abstract. A family of dual-primal finite element tearing and interconnecting (FETI) methods for edge element approximations in two dimensions is proposed and analyzed. The primal constraints here are averages over subdomain edges. It is shown that the condition number of the corresponding method is independent of the number of substructures and grows only polylogarithmically with the number of unknowns associated with individual substructures. The estimate is also independent of the jumps of both of the coefficients of the original problem. Numerical results validating our theoretical bounds are given.

Key words. edge elements, Maxwell's equations, finite elements, spectral elements, domain decomposition, FETI, preconditioners, heterogeneous coefficients

AMS subject classifications. 65F10, 65N22, 65N30, 65N55

DOI. 10.1137/S0036142903436915

1. Introduction.

In this paper, we consider the boundary value problem

$$\begin{aligned} L\mathbf{u} := \mathbf{curl}\,(a\ \mathrm{curl}\,\mathbf{u}) + A\ \mathbf{u} = \mathbf{f} \quad &\text{in } \Omega, \\ \mathbf{u}\cdot\mathbf{t} = 0 \quad &\text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

with Ω a bounded polygonal domain in $\mathbb{R}^2$. The domain Ω has unit diameter and $\mathbf{t}$ is its unit tangent. We have

$$\mathbf{curl}\, v := \left[\frac{\partial v}{\partial x_2}, -\frac{\partial v}{\partial x_1}\right]^T, \qquad \mathrm{curl}\,\mathbf{u} := \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2};$$

see, e.g., [19]. The coefficient matrix A is a symmetric uniformly positive definite matrix-valued function with entries $A_{ij} \in L^\infty(\Omega)$, $1 \le i,\, j \le 2$, and $a \in L^\infty(\Omega)$ is a positive function bounded away from zero.

The weak formulation of problem (1.1) requires the introduction of the Hilbert space $H(\mathrm{curl}\,;\Omega)$, defined by

$$H(\mathrm{curl}\,;\Omega) := \{\mathbf{v} \in (L^2(\Omega))^2 \mid \mathrm{curl}\,\mathbf{v} \in L^2(\Omega)\}.$$

The space $H(\mathrm{curl}\,;\Omega)$ is equipped with the inner product and graph norm

$$(\mathbf{u},\mathbf{v})_{\mathrm{curl}} := \int_\Omega \mathbf{u}\cdot\mathbf{v}\ dx + \int_\Omega \mathrm{curl}\,\mathbf{u}\,\mathrm{curl}\,\mathbf{v}\ dx, \qquad \|\mathbf{u}\|^2_{\mathrm{curl}} := (\mathbf{u},\mathbf{u})_{\mathrm{curl}},$$

and the tangential component $\mathbf{u}\cdot\mathbf{t}$, of a vector $\mathbf{u} \in H(\mathrm{curl}\,;\Omega)$ on the boundary $\partial\Omega$, belongs to the space $H^{-\frac{1}{2}}(\partial\Omega)$; see [6, 19]. The subspace of vectors in $H(\mathrm{curl}\,;\Omega)$ with vanishing tangential component on $\partial\Omega$ is denoted by $H_0(\mathrm{curl}\,;\Omega)$.

*Received by the editors October 30, 2003; accepted for publication (in revised form) August 19, 2004; published electronically April 19, 2005.
http://www.siam.org/journals/sinum/42-6/43691.html
†Seminar for Applied Mathematics (SAM), ETH Zürich, CH-8092 Zürich, Switzerland (toselli@sam.math.ethz.ch, vasseur@sam.math.ethz.ch). This work was partially supported by the Swiss National Science Foundation under project 20-63397.00.

For any $\mathcal{D} \subset \Omega$, we define the bilinear form

$$a_{\mathcal{D}}(\mathbf{u},\mathbf{v}) := \int_{\mathcal{D}} (a \operatorname{curl}\mathbf{u} \operatorname{curl}\mathbf{v} + A\, \mathbf{u}\cdot\mathbf{v})\, dx, \quad \mathbf{u},\, \mathbf{v} \in H(\operatorname{curl};\Omega). \tag{1.2}$$

The variational formulation of (1.1) is as follows.

Find $\mathbf{u} \in H_0(\operatorname{curl};\Omega)$ such that

$$a_{\Omega}(\mathbf{u},\mathbf{v}) = \int_{\Omega} \mathbf{f}\cdot\mathbf{v}\, dx, \quad \mathbf{v} \in H_0(\operatorname{curl};\Omega). \tag{1.3}$$

The purpose of this work is to construct and analyze a dual-primal finite element tearing and interconnecting (FETI-DP) preconditioner for h and p finite element approximations of problem (1.3). Neumann–Neumann (NN) and finite element tearing and interconnecting (FETI) algorithms are particular domain decomposition (DD) methods of iterative substructuring type; they rely on a nonoverlapping partition into subdomains. They are among the most popular and heavily tested DD methods and are now employed for the solution of huge problems on parallel architectures; see, e.g., [16, 9, 8, 31, 4]. The rate of convergence is often independent of possibly large jumps of the coefficients.

FETI methods rely on the reformulation of the original algebraic problem into an equivalent saddle point problem, involving discontinuous functions across the subdomain boundaries and a continuity constraint for the solution; see (3.3). In the original one-level FETI methods, completely discontinuous vectors are employed; the elimination of the primal variable thus requires the solution of local (generally singular) Neumann problems, and an equation for the Lagrange multipliers is then obtained. A first step consists in the elimination of the components belonging to a suitable coarse space, constructed from local subdomain kernels or suitable functions (constants for the Laplace equation and rigid body modes for linear elasticity, for example). This elimination employs a projection, constructed with a suitable scaling matrix. A preconditioner for the resulting equation is then constructed by solving local Dirichlet problems on the subdomains and by employing a set of scaling matrices which have the purpose of making convergence independent of possibly large jumps of the coefficients. We recall that one-level FETI methods actually consist of both coarse and local components; in the context of FETI methods the term "one-level" is in contrast to two-level methods that were developed primarily for biharmonic and shell element problems and that involve satisfying some of the continuity constraints at each step of the iterations.

The more recently developed FETI-DP methods employ a smaller space for the solution, where a certain number of degrees of freedom or linear functionals is continuous across the subdomains. These so-called *primal constraints* ensure that a nonsingular global problem needs to be solved in order to obtain an equation for the Lagrange multipliers; this step requires the solution of modified nonsingular Neumann problems and the solution of a coarse problem the size of which equals the number of primal constraints. As before, a preconditioner is constructed by solving local Dirichlet problems. FETI-DP algorithms present considerable advantages: the same code can now be employed for a wider class of problems, much less dense coarse matrices need to be inverted, they do not require the characterization of the kernels of local Neumann operators or the introduction of an additional scaling matrix for the construction of the coarse component of the preconditioner, and they may start the conjugate gradient (CG) iteration from an arbitrary initial guess. For these reasons,

they have now almost completely replaced one-level FETI methods for large scale computations. Connections between NN and FETI methods are being investigated; see [13].

The motivation of this work lies in the fact that no iterative substructuring methods (and, in particular, no NN or FETI preconditioners) that are robust with respect to the number of unknowns, the number of subdomains, and large jumps of the coefficients are presently available for edge element approximations of three-dimensional problems.

Some methods are available for two-dimensional approximations.

In [30], a DD preconditioner was proposed, which is based on a standard coarse space and local spaces associated to the subdomain edges. NN preconditioners with standard coarse spaces were studied in [24]. One-level FETI methods were developed in [26, 21], thanks to the introduction of suitable local functions which are the analogue of constants and rigid body modes for the Laplace equation and linear elasticity, respectively. These functions were then employed to construct a Balancing NN method in [25]. Standard coarse spaces, however, are not in general suitable for quasi-optimal preconditioners in three dimensions, and the search for suitable local functions in three dimensions for balancing NN and one-level FETI methods has produced no results so far. For these reasons we believe that FETI-DP algorithms will turn out to be easier to devise for three-dimensional problems. The scope of this work is then to begin to understand a good set of primal constraints in *two dimensions*, which have not been available so far. It turns out that the natural choice of edge averages on the subdomain edges leads to a robust preconditioner.

For the analysis, we employ the tools developed in [30]. We note that more general tools were later devised in [32], which consisted in decomposition results for trace functions in $H(\mathrm{curl}\,;\Omega)$ in two dimensions or $H(\mathrm{div}\,;\Omega)$ and stable curl/divergence-free extensions from the subdomain boundaries. Here, we have chosen to employ the results in [30], since we have in mind extensions to anisotropic meshes which are often needed for problems in conductor materials with high jumps in the conductivity. The work in [27, 28, 29] for scalar problems showed that when dealing with highly anisotropic meshes the analysis cannot employ trace norms or stable extensions, since the latter are not available in this case. The approach in [30], which does not rely on trace norms or stable extensions, but only on results for scalar problems (see Lemma 4.1 and its proof in section 5), appears more promising for edge element approximations on anisotropic meshes. This generalization is left to a future work.

We recall that effective multilevel strategies have also been developed for problems involving the curl-curl operator in three dimensions. We refer to, e.g., [1, 12, 22, 5].

This paper is organized as follows. In section 2, we introduce our discrete problems, the subdomain partition, and local and global finite element spaces. In section 3, we introduce our FETI-DP algorithms. Condition number bounds are given in section 4. First we give the technical tools necessary to prove them in subsection 4.1. These are the decomposition lemma, Lemma 4.1, and the abstract framework for the analysis of FETI-DP methods originally proposed in [15]. Our main result is the stability property in Lemma 4.6 in subsection 4.2. Lemma 4.1 is proven in [30] for h approximations, and we provide a proof for the case of p finite elements in section 5. A practical implementation of our algorithm is given in section 6 and some numerical results in section 7.

2. Discrete spaces. In this paper, we consider both h and p version finite elements. We discretize this problem using edge elements, which are also known as

Nédélec elements; see [20]. These are vector finite elements that ensure only the continuity of the tangential component across the elements, as is physically required for the electric and magnetic fields, solutions of Maxwell's equations. We refer to [19] for a general introduction of approximations of electromagnetic problems, the Sobolev space $H(\mathrm{curl}\,;\Omega)$, and edge elements.

2.1. Triangulations and subdomain partitions. We introduce a shape-regular triangulation $\mathcal{T} = \mathcal{T}_h$ of the domain Ω, made of affinely mapped quadrilaterals. In particular, if $\widehat{Q} = (-1,1)^2$ is a reference square, for each element $K \in \mathcal{T}$, there exists an affine mapping $F_K : \widehat{Q} \to K$, such that K is the image of $\widehat{Q}$. Here we consider only quadrilateral meshes for simplicity but note that our results are equally valid for h approximations on triangular meshes.

Let $\mathcal{E} = \mathcal{E}_h$ be the set of edges of $\mathcal{T}$. For every edge $e \in \mathcal{E}$, we fix a direction, given by a unit vector $\mathbf{t}_e$, tangent to e. The length of the edge e is denoted by $|e|$.

We next consider a nonoverlapping partition of the domain Ω,

$$\mathcal{F}_H = \left\{ \Omega_i \mid 1 \le i \le N,\ \bigcup_{i=1}^{N} \overline{\Omega_i} = \overline{\Omega} \right\},$$

such that each Ω_i is connected. The elements of $\mathcal{F}_H$ are called *subdomains* or *substructures*. For the h version, we take the substructures Ω_i as unions of fine elements. We denote the diameter of Ω_i by H_i and define H as the maximum of the diameters of the subdomains:

$$H := \max_{1 \le i \le N} \{H_i\}.$$

In this case $h < H$. For the p version we take $\mathcal{F}_H = \mathcal{T}_h$ and thus $H = h$.

We always assume that the substructures are images of a reference square under sufficiently regular maps, which effectively means that their aspect ratios remain uniformly bounded. In addition, we assume that the ratio of the diameters of two adjacent subregions is bounded away from zero and infinity. Further assumptions, necessary for the analysis but not for the definition of the algorithms, are made at the beginning of section 4.1.

We define the *edges* of the partition as the interior E_{ij} of the intersections

$$\overline{E}_{ij} := \partial\Omega_i \cap \partial\Omega_j, \quad i \ne j,\ |E_{ij}| > 0,$$

where $|E_{ij}|$ denotes the measure of E_{ij} and $\overline{E}_{ij}$ its closure. We note that $E_{ji} = E_{ij}$. We introduce a unit vector $\mathbf{t}_{E_{ij}}$ that is tangent to E_{ij}. Let $\mathcal{E}_H$ be the set of edges of $\mathcal{F}_H$, and let the interface Γ be the union of the edges of $\mathcal{F}_H$ or, equivalently, the parts of the subdomain boundaries that do not belong to $\partial\Omega$:

$$\Gamma := \bigcup_{i=1}^{N} \partial\Omega_i \setminus \partial\Omega.$$

For every subdomain Ω_i, let $\mathcal{I}_i$ be the set of indices j, such that E_{ij} is an edge of Ω_i:

$$\mathcal{I}_i := \{j \mid E_{ij} \subset \partial\Omega_i,\ E_{ij} \in \mathcal{E}_H\}.$$

Our assumptions on the partition $\mathcal{F}_H$ ensure that the the number of edges $|\mathcal{I}_i|$ is uniformly bounded.

We assume that the coefficients a and A are constant in each substructure Ω_i and denote them by a_i and A_i, respectively. We also assume that

$$0 < \beta_i |\mathbf{x}|^2 \leq \mathbf{x}^t A_i \mathbf{x} \leq \gamma_i |\mathbf{x}|^2, \quad \mathbf{x} \in \mathbb{R}^2, \tag{2.1}$$

for $i = 1, \ldots, N$, where $|\cdot|$ denotes the standard Euclidean norm.

2.2. Edge element functions. We next define the local spaces

$$\mathrm{H}_\star(\mathrm{curl}\,;\Omega_i) := \{\mathbf{u}_i \in \mathrm{H}(\mathrm{curl}\,;\Omega_i) \mid \mathbf{u}_i \cdot \mathbf{t} = 0 \text{ on } \partial\Omega \cap \partial\Omega_i\}$$

and the following polynomial spaces on the reference square for $k \geq 1$:

$$\mathcal{R}_k(\widehat{Q}) = \mathbb{Q}_{k-1,k}(\widehat{Q}) \otimes \mathbb{Q}_{k,k-1}(\widehat{Q}),$$

with $\mathbb{Q}_{k_1,k_2}(\widehat{Q})$ the space of polynomials of degree k_i in the ith variable. On an affinely mapped element $K \in \mathcal{T}$, we take

$$\mathcal{R}_k(K) = \{\mathbf{u} = J_{F_K}^{-T} \widehat{\mathbf{u}} \mid \widehat{\mathbf{u}} \in \mathcal{R}_k(\widehat{Q})\}, \tag{2.2}$$

with J_{F_K} the Jacobian of the transformation F_K. We note that the tangential component of a vector in $\mathcal{R}_k(K)$ is a function of $\mathbb{Q}_{k-1}$ over each edge of K.

For the h version, we employ the lowest-order Nédélec finite element spaces, originally introduced in [20], defined on each subdomain Ω_i:

$$X_i = X^h(\Omega_i) := \{\mathbf{u} \in \mathrm{H}_\star(\mathrm{curl}\,;\Omega_i) \mid \mathbf{u}_{|K} \in \mathcal{R}_1(K),\ K \in \mathcal{T}_h,\ K \subset \Omega_i\}.$$

Higher polynomial degrees can also be considered and our results and bounds will remain valid with constants that depend on the polynomial degree. See, e.g., [19] for more details. Functions in X_i have a constant tangential component over the fine edges in $\mathcal{E}$. The degrees of freedom for X_i are the constant values of the tangential component on the fine edges in $\mathcal{E}$ contained in $\overline{\Omega}_i$.

For the p version, we choose

$$X_i = X^k(\Omega_i) := \mathcal{R}_k(\Omega_i) \cap \mathrm{H}_\star(\mathrm{curl}\,;\Omega_i).$$

The basis functions can be associated to the (mapped) Gauss–Lobatto nodes on Ω_i, and the corresponding degrees of freedom are in this case the values at these nodes. Other basis functions are also possible, however. We refer to, e.g., [3, 11, 18, 7, 19] for more details on spectral and p finite element approximations of electromagnetic problems. The results in this paper for the p version are independent of the particular basis chosen.

We next consider the product space

$$X = X(\Omega) := \prod_{i=1}^{N} X_i \subset \prod_{i=1}^{N} \mathrm{H}_\star(\mathrm{curl}\,;\Omega_i),$$

which consists of vectors that have in general a discontinuous tangential component along the subdomain edges. The discrete solution is sought in the conforming space

$$\widehat{X} := X \cap \mathrm{H}_0(\mathrm{curl}\,;\Omega)$$

of vectors with a continuous tangential component along the edges in $\mathcal{E}_H$.

We now introduce some trace spaces consisting of tangential components on the boundaries of the substructures. A scalar function u, defined on $\partial\Omega_i \setminus \partial\Omega$, belongs to W_i if and only if there exists $\mathbf{u} \in X_i$ such that, for each edge,

$$u_{|_{E_{ij}}} = \mathbf{u} \cdot \mathbf{t}_{E_{ij}}, \quad E_{ij} \in \mathcal{E}_H,\ j \in \mathcal{I}(i).$$

For h approximations these are piecewise constant (or piecewise polynomial of degree $k-1$ if higher-order Nédélec elements are considered) along the edges E_{ij}. For p approximations they are polynomials of degree $k-1$ on each edge E_{ij}. We will employ the product space of functions defined on Γ, $W := \prod_i W_i$, and its continuous subspace $\widehat{W}$ consisting of tangential traces of vectors in $\widehat{X}$.

The scalar functions in the spaces W_i and W are uniquely defined by the degrees of freedom of the spaces X_i and X involving the tangential components along edges in $\mathcal{E}_H$. Throughout this paper, we will use the following notation: we denote a generic vector function in X_i using a bold letter with the superscript (i), e.g., $\mathbf{u}^{(i)}$, and employ the same notation for the corresponding column vector of degrees of freedom. Its tangential component $u^{(i)}$ is an element of W_i and is defined by

$$u^{(i)}{}_{|_{E_{ij}}} := \mathbf{u}^{(i)} \cdot \mathbf{t}_{E_{ij}}, \quad E_{ij} \in \mathcal{E}_H,\ j \in \mathcal{I}(i).$$

It is uniquely determined by the degrees of freedom $\mathbf{u}^{(i)}$ involving the tangential component along $\partial\Omega_i \setminus \partial\Omega$. We use the same notation $u^{(i)}$ for the column vector of these tangential degrees of freedom and the same notation for the spaces of functions X_i and W_i and for the corresponding spaces of degrees of freedom. We use similar notation for global functions in X and W.

We remark that a vector $\mathbf{u}$ belongs to the continuous space $\widehat{X}$ (and consequently its tangential component to $\widehat{W}$) if

$$u^{(i)}{}_{|_{E_{ij}}} = u^{(j)}{}_{|_{E_{ij}}}, \quad E_{ij} \in \mathcal{E}_H. \tag{2.3}$$

Finally, for $i = 1, \ldots, N$, we define the extensions into the interior of the Ω_i,

$$\mathcal{H}_i : W_i \longrightarrow X_i,$$

that are discrete harmonic with respect to the bilinear forms $a_{\Omega_i}(\cdot,\cdot)$. We recall that $\mathbf{u}^{(i)} = \mathcal{H}_i u^{(i)}$ minimizes the energy $a_{\Omega_i}(\mathbf{u}^{(i)}, \mathbf{u}^{(i)})$ among all the vectors of X_i with tangential component equal to $u^{(i)}$ on $\partial\Omega_i \setminus \partial\Omega$. We will refer to $\mathcal{H}_i$ as the *Maxwell* discrete harmonic extension.

2.3. Continuous finite element spaces. In the following we will also need the standard finite element spaces of scalar, continuous, piecewise polynomial functions. With

$$\mathrm{H}^1_\star(\Omega_i) := \{\phi \in \mathrm{H}^1(\Omega_i) \mid \phi = 0 \text{ on } \partial\Omega \cap \partial\Omega_i\},$$

we define, for the h version, the space of continuous piecewise bilinear functions

$$Q_i = Q^h(\Omega_i) := \{\phi \in \mathrm{H}^1_\star(\Omega_i)\ \phi \in \mathbb{Q}_{1,1}(K),\ K \in \mathcal{T}_h,\ K \subset \Omega_i\}.$$

For the p version, we employ

$$Q_i = Q^k(\Omega_i) := \mathbb{Q}_{k,k}(\Omega_i) \cap \mathrm{H}^1_\star(\Omega_i).$$

We note that in both cases $\operatorname{grad} Q_i \subset X_i$.

Discrete harmonic functions in Q_i will be referred to as *Laplace* discrete harmonic in the following.

3. FETI-DP methods. In this section, we introduce a FETI-DP method for the solution of the linear system arising from the edge element discretization of problem (1.3). In section 6, we give a practical implementation of the algorithm. Throughout the paper, we denote the Euclidean scalar product in l^2 by $\langle \cdot, \cdot \rangle$. We recall that FETI-DP methods were originally introduced in [8]. The first theoretical result was given in [17] for two-dimensional problems and then later in [15] for three dimensions. Some theoretical results for linear elasticity can be found in [14].

We first assemble the local stiffness matrices, relative to the bilinear forms $a_{\Omega_i}(\cdot, \cdot)$, and the local load vectors. The degrees of freedom that belong only to one substructure can be eliminated in parallel by block Gaussian elimination. We note that these are degrees of freedom associated to edges or nodes in the interior of the substructures, on $\partial\Omega$, and, in case polynomial spaces with $k > 0$ are employed, they also consist of values of the normal component on the subdomain boundaries. We are then left with the degrees of freedom involving the *tangential* component along the substructure boundaries. Let $f^{(i)}$ be the resulting right-hand sides and $S^{(i)}$ the Schur complement matrices

$$S^{(i)} : W_i \longrightarrow W_i,$$

relative to the tangential degrees of freedom on $\partial\Omega_i \setminus \partial\Omega$.

We recall that the local Schur complements satisfy the property

$$|u^{(i)}|^2_{S^{(i)}} := \langle u^{(i)}, S^{(i)} u^{(i)} \rangle = a_{\Omega_i}(\mathcal{H}_i u^{(i)}, \mathcal{H}_i u^{(i)}); \tag{3.1}$$

see, e.g., [23, 24]. Since the local bilinear forms are positive definite, so are the local Schur complements $S^{(i)}$.

We write

$$u := \begin{bmatrix} u^{(1)} \\ \vdots \\ u^{(N)} \end{bmatrix} \in W, \qquad S := \operatorname{diag}\{S^{(1)}, \ldots, S^{(N)}\}, \qquad f := \begin{bmatrix} f^{(1)} \\ \vdots \\ f^{(N)} \end{bmatrix}.$$

The solution $u \in W$ to the discrete problem can then be found by minimizing the energy

$$\frac{1}{2}\langle u, Su \rangle - \langle f, u \rangle$$

subject to the constraint that u is continuous, i.e., it belongs to $\widehat{W}$.

For FETI-DP methods we work in a subspace $\widetilde{W} \subset W$ of functions satisfying a certain number of continuity constraints. We have

$$\widetilde{W} = \widehat{W}_\Pi \oplus \widetilde{W}_\Delta.$$

Here the primal space $\widehat{W}_\Pi \subset \widehat{W}$ consists of continuous functions determined by degrees of freedom associated to the substructures. We choose a space of constant functions on the subdomain edges.

$$\widehat{W}_\Pi = \widehat{W}^H := \{u \in \widehat{W} \mid u_{|_{E_{ij}}} \in \mathbb{Q}_0,\ E_{ij} \in \mathcal{E}_H\}. \tag{3.2}$$

The degrees of freedom (*primal variables*) associated to this space are the averages of tangential components over the subdomain edges:

$$\bar{u}_{E_{ij}} = \frac{\int_{E_{ij}} u\, ds}{|E_{ij}|} = \frac{\int_{E_{ij}} \mathbf{u}\cdot\mathbf{t}_{E_{ij}}\, ds}{|E_{ij}|}.$$

These are the same degrees of freedom associated to a standard coarse space in case the substructures are elements of a coarse mesh; see [30, 24] and section 4.1.

The dual space $\widetilde{W}_\Delta$ is the product space of spaces associated to the substructures

$$\widetilde{W}_\Delta := \prod_{i=1}^{N} \widetilde{W}_{\Delta,i}$$

of functions for which the functional given by the primal variables vanish:

$$\widetilde{W}_{\Delta,i} := \{u \in W_i \mid \bar{u}_{E_{ij}} = 0,\ j \in \mathcal{I}(i)\}.$$

Therefore, $\widetilde{W}$ consists of functions that have a continuous average along the substructure edges; i.e., the averages are the same regardless of which substructure is considered for the calculation.

The primal degrees of freedom can then be eliminated together with the internal ones, at the expense of solving one coarse problem. We are then left with a problem involving interface functions with vanishing mean value along the substructure edges and, consequently, in the dual space, $\widetilde{W}_\Delta$. Let $\widetilde{S} : \widetilde{W}_\Delta \to \widetilde{W}_\Delta$ be the corresponding Schur complement and $\widetilde{f}_\Delta$ the corresponding load vector. We then look for $u_\Delta \in \widetilde{W}_\Delta$, such that

$$\frac{1}{2}\langle u_\Delta, \widetilde{S}u_\Delta\rangle - \langle \widetilde{f}_\Delta, u_\Delta\rangle \longrightarrow \min$$

subject to the constraint that u_Δ is continuous. The continuity constraint is expressed by the equation

$$B_\Delta u_\Delta = 0,$$

where B_Δ is constructed from $\{0, 1, -1\}$ and evaluates the difference between all the corresponding tangential degrees of freedom on Γ; cf. (2.3). We employ the same matrix as in our previous paper [26] and then enforce redundant conditions. The matrix B_Δ has the following block structure:

$$B_\Delta = [B_\Delta^{(1)} \quad B_\Delta^{(2)} \quad \cdots \quad B_\Delta^{(N)}],$$

where each block corresponds to a substructure.

We obtain the saddle point problem

$$\begin{aligned} \widetilde{S}u_\Delta + B_\Delta^T\lambda &= \widetilde{f}_\Delta, \\ B_\Delta u_\Delta &= 0, \end{aligned} \tag{3.3}$$

with $u_\Delta \in \widetilde{W}_\Delta$ and $\lambda \in V := \mathrm{Range}(B_\Delta)$.

We note that $\widetilde{S}$ can be obtained from the restriction of S to the space $\widetilde{W}$, by eliminating the primal degrees of freedom. We have therefore the minimization property

$$\langle u_\Delta, \widetilde{S}u_\Delta\rangle = \min\langle u, Su\rangle, \tag{3.4}$$

where the minimum is taken over all the functions $u = u_\Delta + w_\Pi$, $w_\Pi \in \widehat{W}_\Pi$. This property ensures that $\widetilde{S}$ is also positive definite.

Since the Schur complement $\widetilde{S}$ is invertible, an equation for λ can easily be found:

$$F\lambda = d, \tag{3.5}$$

with

$$F := B_\Delta \widetilde{S}^{-1} B_\Delta^T, \qquad d := B_\Delta \widetilde{S}^{-1} \widetilde{f}_\Delta. \tag{3.6}$$

In section 6, we provide explicit formulas for F and d. Once λ is found, the primal variables are given by

$$u_\Delta = \widetilde{S}^{-1}(\widetilde{f}_\Delta - B_\Delta^T \lambda) \in \widetilde{W}_\Delta.$$

In order to define a preconditioner for (3.5), we need to define scaling matrices and functions defined on the subdomain boundaries. As in our previous work they are constructed with the coefficient A only. For each substructure, we define $\delta_i^\dagger \in W_i$, such that on the edge E_{ij}, $j \in \mathcal{I}(i)$,

$$\delta_i^\dagger = \frac{\gamma_i^\chi}{\gamma_i^\chi + \gamma_j^\chi} \tag{3.7}$$

for an arbitrary but fixed $\chi \in [1/2, +\infty)$; see (2.1). By direct calculation, we find

$$\gamma_i {\delta_j^\dagger}^2 \leq \min(\gamma_i, \gamma_j). \tag{3.8}$$

For each substructure Ω_i, we next introduce a diagonal matrix $D_\Delta^{(i)} : V \to V$. The diagonal entry corresponding to the Lagrange multipliers that enforce the continuity along an edge E_{ij} is set equal to the (constant) value of $\delta_j^\dagger$ along E_{ji}

$$\delta_{ji}^\dagger := \delta_{j|_{E_{ji}}}^\dagger = \frac{\gamma_j^\chi}{\gamma_i^\chi + \gamma_j^\chi}.$$

We next define the scaled matrix

$$B_{D,\Delta} = [D_\Delta^{(1)} B_\Delta^{(1)} \quad D_\Delta^{(2)} B_\Delta^{(2)} \ \cdots \ D_\Delta^{(N)} B_\Delta^{(N)}] : \widetilde{W}_\Delta \to V.$$

We solve the dual system (3.5) using the preconditioned CG algorithm with the preconditioner

$$M^{-1} := B_{D,\Delta} S B_{D,\Delta}^T = \sum_{i=1}^N D_\Delta^{(i)} B_\Delta^{(i)} S^{(i)} B_\Delta^{(i)T} D_\Delta^{(i)}; \tag{3.9}$$

see [8, 17, 15].

4. Condition number bounds.

4.1. Technical tools. The analysis of the FETI-DP methods presented here relies on a decomposition result. We first need to introduce coarse spaces on the

subdomains. As is often customary in the analysis of iterative substructuring methods, we require that the substructures are elements of a shape-regular coarse mesh $\mathcal{T}_H$. This is always the case for p finite elements. We next define

$$X^H(\Omega_i) := \mathcal{R}_1(\Omega_i), \tag{4.1}$$

the lowest-order edge element space on the coarse element Ω_i; see (2.2). We note that the tangential traces of vectors in $X^H(\Omega_i)$ are restrictions of functions in the space $\widehat{W}^H$, defined in (3.2), to the boundary of Ω_i.

The following result can be found in [30, Lem. 4.2] for h approximations. The proof for the p finite element case is given in section 5. We need the scaled norm

$$\|\mathbf{u}\|^2_{\mathrm{curl},\Omega_i} := \|\mathbf{u}\|^2_{L^2(\Omega_i)} + H_i^2\|\mathrm{curl}\,\mathbf{u}\|^2_{L^2(\Omega_i)}, \quad \mathbf{u} \in X_i.$$

LEMMA 4.1. *Let Ω_i be a substructure. Then, for every $\mathbf{u} \in X_i$ there exists a unique decomposition*

$$\mathbf{u} = \mathbf{u}_H + \sum_{j\in\mathcal{I}(i)} \mathbf{u}_{ij} + \mathbf{u}^{int} \tag{4.2}$$

such that the following hold:

1. *$\mathbf{u}_H$ is a coarse function in $X^H(\Omega_i)$;*
2. *$\mathbf{u}_{ij} = \nabla\phi_{ij}$, with $\phi_{ij} \in Q_i$, is a Laplace discrete harmonic function that vanishes on $\partial\Omega_i \setminus E_{ij}$;*
3. *$\mathbf{u}^{int}$ has a vanishing tangential component on $\partial\Omega_i$.*

In addition, for $j \in \mathcal{I}(i)$,

$$\int_{E_{ij}} (\mathbf{u} - \mathbf{u}_H)\cdot\mathbf{t}_{E_{ij}}\,ds = \int_{E_{ij}} \nabla\phi_{ij}\cdot\mathbf{t}_{E_{ij}}\,ds = 0, \tag{4.3}$$

and

$$\|\nabla\phi_{ij}\|^2_{L^2(\Omega_i)} \le C\omega^2\|\mathbf{u}\|^2_{\mathrm{curl},\Omega_i}, \tag{4.4}$$

with $\omega = (1+\log(H/h))$ for h approximations and $\omega = (1+\log k)$ for p approximations.

We note that bounds for the components $\mathbf{u}_H$ and $\mathbf{u}^{int}$ can also be found, but they will not be necessary for the analysis in this paper.

The following result is a straightforward application of the existence of a stable finite element extension and of a trace theorem; see, in particular, [2] for the p version.

LEMMA 4.2. *Let Ω_i and Ω_j be two substructures that share an edge E_{ij}. Let $\phi^{(i)} \in Q_i$ and $\phi^{(j)} \in Q_j$ be two Laplace discrete harmonic functions that have a common trace on E_{ij} and vanish on $\partial\Omega_i \setminus E_{ij}$ and $\partial\Omega_j \setminus E_{ij}$, respectively. Then there exists a constant C, independent of h, k, H_i, and H_j, such that*

$$\|\nabla\phi^{(j)}\|^2_{L^2(\Omega_j)} \le C\|\nabla\phi^{(i)}\|^2_{L^2(\Omega_i)}.$$

We now recall an abstract framework for the analysis of FETI-DP algorithms, which was originally given in [15]. It turns out that condition number bounds rely on one stability estimate for the following jump operator:

$$P_\Delta := B^T_{D,\Delta}B_\Delta : \widetilde{W} \longrightarrow \widetilde{W}.$$

We summarize the properties of P_Δ proven in [15, sect. 6] in the following lemma.

LEMMA 4.3. *The operator P_Δ is a projection and preserves the jump of any function $w \in \widetilde{W}$, i.e.,*

$$B_\Delta P_\Delta w = B_\Delta w.$$

If $v := P_\Delta w$ for $w \in \widetilde{W}$, then on every edge E_{ij} of a substructure Ω_i, we have

$$v^{(i)} = \delta_j^\dagger (w^{(i)} - w^{(j)}). \tag{4.5}$$

Finally, $P_\Delta w = 0$ if $w \in \widehat{W}$.

The following fundamental result can be found in [15, Th. 1]. It employs the norms

$$|v|_S^2 := \langle v, Sv\rangle = \sum_{i=1}^N \langle v^{(i)}, S^{(i)} v^{(i)}\rangle, \quad |v|_{\widetilde{S}}^2 := \langle v, \widetilde{S} v\rangle. \tag{4.6}$$

THEOREM 4.4. *Let C_{P_Δ} be such that*

$$|P_\Delta w_\Delta|_S^2 \le C_{P_\Delta} |w_\Delta|_{\widetilde{S}}^2, \quad w_\Delta \in \widetilde{W}_\Delta. \tag{4.7}$$

Then, if $\widetilde{S}$ and M^{-1} are invertible,

$$\langle M\lambda, \lambda\rangle \le \langle F\lambda, \lambda\rangle \le C_{P_\Delta} \langle M\lambda, \lambda\rangle, \quad \lambda \in V. \tag{4.8}$$

4.2. Main results. We now present two lemmas. The first one is trivial for our approximations and ensures that the Schur complement $\widetilde{S}$ and the preconditioner M^{-1} are invertible. The second provides a key stability estimate in order to bound the largest eigenvalue of the preconditioned operator $M^{-1}F$. Our main result is given in Theorem 4.7.

LEMMA 4.5. *The Schur complement $\widetilde{S}$ and the preconditioner M^{-1} are invertible.*

Proof. The result for $\widetilde{S}$ is an immediate consequence of the fact that the local bilinear forms $a_{\Omega_i}(\cdot,\cdot)$ are positive definite. Indeed the Schur complement S is invertible and so is $\widetilde{S}$ thanks to (3.4).

In order to prove the invertibility of M^{-1}, we assume that there is a $\lambda = B_\Delta w_\Delta$, $w_\Delta \in \widetilde{W}$, such that

$$0 = M^{-1}\lambda = B_{D,\Delta} S B_{D,\Delta}^T B_\Delta w_\Delta.$$

This implies

$$0 = \langle \lambda, M^{-1}\lambda\rangle = |P_\Delta w_\Delta|_S^2.$$

Since the local Schur complements $S^{(i)}$ are invertible, this implies $P_\Delta w_\Delta = 0$. Lemma 4.3 then implies

$$\lambda = B_\Delta w_\Delta = B_\Delta P_\Delta w_\Delta = 0. \qquad \square$$

LEMMA 4.6. *There is a constant C, such that, for $w_\Delta \in \widetilde{W}_\Delta$,*

$$|P_\Delta w_\Delta|_S^2 \le C\, \eta\, \omega^2\, |w_\Delta|_{\widetilde{S}}^2,$$

where ω is the same as in Lemma 4.1 *and*

$$\eta := \max_{1\le i\le N} \frac{\gamma_i}{\beta_i}\left(1+\frac{H_i^2\beta_i}{a_i}\right).$$

Proof. Using the minimization property in (3.4), we consider the element $w = w_\Delta + w_\Pi$, $w_\Pi \in \widehat{W}_\Pi$, such that

$$|w_\Delta|_{\tilde{S}}^2 = |w|_S^2. \tag{4.9}$$

We note that, since w_Π is continuous,

$$v := P_\Delta w_\Delta = P_\Delta w.$$

We then need to calculate

$$|P_\Delta w|_S^2 = \sum_{i=1}^N |v^{(i)}|_{S^{(i)}}^2 = \sum_{i=1}^N a_{\Omega_i}(\mathcal{H}_i v^{(i)}, \mathcal{H}_i v^{(i)}).$$

On an edge E_{ij} of a substructure Ω_i, we employ the representation in (4.5). We recall that the function $\delta_j^\dagger$ is constant along an edge E_{ij} and $\delta_{ji}^\dagger$ is this value. We then decompose $v^{(i)}$ into contributions supported on single edges:

$$v^{(i)} = \sum_{j\in\mathcal{I}(i)} \theta_{E_{ij}}\delta_{ji}^\dagger(w^{(i)} - w^{(j)}), \tag{4.10}$$

where $\theta_{E_{ij}} \in W_i$ is identically one on E_{ij} and vanishes on $\partial\Omega_i \setminus E_{ij}$. We consider each contribution in this sum separately. Since, in addition, w is an element of $\widetilde{W}$, its average $\bar{w}_{E_{ij}}$ is the same whether it is calculated using $w^{(i)}$ or $w^{(j)}$. We can therefore write

$$\theta_{E_{ij}}\delta_{ji}^\dagger(w^{(i)} - w^{(j)}) = \theta_{E_{ij}}\delta_{ji}^\dagger(w^{(i)} - \bar{w}_{E_{ij}}) - \theta_{E_{ij}}\delta_{ji}^\dagger(w^{(j)} - \bar{w}_{E_{ij}}). \tag{4.11}$$

We consider the two terms in (4.11) separately.

In order to bound the first, we employ the decomposition in Lemma 4.1 for the vector $\mathbf{u} := \mathcal{H}_i w^{(i)}$. We recall that the tangential component of $\mathbf{u}_{ij} = \nabla\phi_{ij}$ vanishes on $\partial\Omega_i \setminus E_{ij}$ and, thanks to (4.3), it is equal to $\theta_{E_{ij}}(w^{(i)} - \bar{w}_{E_{ij}})$. Using (3.8), the minimizing property of the Maxwell discrete harmonic extension in (3.1), (2.1), and (4.4), we find

$$\begin{aligned} |\theta_{E_{ij}}\delta_{ji}^\dagger(w^{(i)} - \bar{w}_{E_{ij}})|_{S^{(i)}}^2 &\le \gamma_i \|\nabla\phi_{ij}\|_{L^2(\Omega_i)}^2 \\ &\le C\gamma_i\omega^2(\|\mathbf{u}\|_{L^2(\Omega_i)}^2 + H_i^2\|\operatorname{curl}\mathbf{u}\|_{L^2(\Omega_i)}^2) \\ &\le C\,\eta\,\omega^2\,a_{\Omega_i}(\mathcal{H}_i w^{(i)}, \mathcal{H}_i w^{(i)}) = C\,\eta\,\omega^2\,|w^{(i)}|_{S^{(i)}}^2. \end{aligned} \tag{4.12}$$

We then consider the second term in (4.11). The vector

$$\mathbf{u}^{(i)} := \mathcal{H}_i(\theta_{E_{ij}}(w^{(j)} - \bar{w}_{E_{ij}}))$$

can be decomposed according to Lemma 4.1, into the sum of two contributions $\mathbf{u}_{ij} = \nabla\phi_{ij}$ and $\mathbf{u}^{int}$. We next apply Lemma 4.1 to the function $\mathcal{H}_j w^{(j)}$ and obtain

$$\mathbf{u}^{(j)} := \mathcal{H}_j w^{(j)} = \mathbf{u}_H + \sum_{k\in\mathcal{I}(j)} \mathbf{u}_{jk} + \widetilde{\mathbf{u}}^{int}.$$

We note that the functions $\mathbf{u}_{ji} = \nabla\phi_{ji}$ and $\mathbf{u}_{ij} = \nabla\phi_{ij}$ have the same tangential component along the common edge E_{ij}, which is equal to $\theta_{E_{ij}}(w^{(j)} - \bar{w}_{E_{ij}})$. Using (3.8), the minimizing property of the Maxwell discrete harmonic extension, Lemma 4.2, (2.1), and (4.4), we find

$$
\begin{aligned}
|\theta_{E_{ij}}\delta_{ji}^{\dagger}(w^{(j)} - \bar{w}_{E_{ij}})|_{S^{(i)}}^2 &\leq \gamma_j \|\nabla\phi_{ij}\|_{L^2(\Omega_i)}^2 \\
&\leq C\gamma_j \|\nabla\phi_{ji}\|_{L^2(\Omega_j)}^2 \\
&\leq C\gamma_j\omega^2(\|\mathbf{u}^{(j)}\|_{L^2(\Omega_j)}^2 + H_j^2\|\operatorname{curl}\mathbf{u}^{(j)}\|_{L^2(\Omega_j)}^2) \\
&\leq C\,\eta\,\omega^2\, a_{\Omega_j}(\mathcal{H}_j w^{(j)}, \mathcal{H}_j w^{(j)}) = C\,\eta\,\omega^2\,|w^{(j)}|_{S^{(j)}}^2.
\end{aligned}
\tag{4.13}
$$

Combining (4.10), (4.12), and (4.13) and summing over the edges E_{ij}, we finally find

$$
|v^{(i)}|_{S^{(i)}}^2 \leq C\,\eta\,\omega^2\,|w^{(i)}|_{S^{(i)}}^2 + C\,\eta\,\omega^2 \sum_{j\in\mathcal{I}(i)} |w^{(j)}|_{S^{(j)}}^2.
$$

The proof is then concluded by summing over the substructures Ω_i and using (4.9). ∎

By combining Lemmas 4.6 and 4.5 and Theorem 4.4, we obtain our final result.

THEOREM 4.7. *The condition number of the preconditioned system $M^{-1}F$ satisfies*

$$
\kappa(M^{-1}F) \leq C\,\eta\,(1 + \log(H/h))^2
$$

for h finite element approximations and

$$
\kappa(M^{-1}F) \leq C\,\eta\,(1 + \log k)^2
$$

for p finite elements. Here, η is defined in Lemma 4.6.

5. Proof of Lemma 4.1. As already mentioned, the proof of Lemma 4.1 for the case of finite elements is given in [30, Lemma 4.2]. In this section we provide a proof for the case of p finite elements. The proof follows that of [30, Lemma 4.2] and is given here for completeness. It employs suitable orthogonal decomposition of edge element functions into gradients of scalar functions and *discrete curl free* functions.

Let $X_i^0 \subset X^k(\Omega_i)$ be the subspace of vectors with vanishing tangential component of $\partial\Omega_i$. If $Q_i^0 \subset Q_i$ is the subspace of functions that vanish on $\partial\Omega_i$, then $\operatorname{grad} Q_i^0 \subset X_i^0$ and the following orthogonal decomposition is well defined:

$$
X_i^0 = \operatorname{grad} Q_i^0 \oplus X_i^{0,\perp}. \tag{5.1}
$$

Proofs of the following fundamental result for a substructure of unit diameter can be found in [10, Th. 7.18] and in [18, sect. 4]. The case of a subdomain of diameter H_i can be treated by a scaling argument. We recall that its proof employs an interpolation operator on the edge element space.

LEMMA 5.1. *Let $\mathbf{u} \in X_i^{0,\perp}$. Then there is a constant, independent of H_i and k, such that*

$$
\|\mathbf{u}\|_{L^2(\Omega_i)} \leq C H_i \|\operatorname{curl}\mathbf{u}\|_{L^2(\Omega_i)}.
$$

We also need a decomposition result for polynomial functions. It is a classical result that is available in the literature in various forms. Since we did not find it in

exactly the form that we need, we have included a proof which employs the tools in [2] in order to help the reader.

LEMMA 5.2. *Let $\psi_H \in \mathbb{Q}_{1,1}(\Omega_i)$ and, for $j \in \mathcal{I}(i)$, let $\psi_{ij} \in Q^k(\Omega_i)$ be a Laplace discrete harmonic function that vanishes on $\partial\Omega_i \setminus E_{ij}$. If*

$$\psi := \psi_H + \sum_{j\in\mathcal{I}(i)} \psi_{ij},$$

then

$$|\psi_{ij}|^2_{H^1(\Omega_i)} \leq C(1+\log k)^2 |\psi|^2_{H^1(\Omega_i)},$$

with a constant that is independent of k and H_i.

Proof. We consider the case of a substructure of unit diameter. The more general case $H_i < 1$ can be treated by a scaling argument. The function ψ_{ij} belongs to $H^{1/2}_{00}(E_{ij})$, the subspace of $H^{1/2}(\partial\Omega)$ of functions that vanish on $\partial\Omega_i \setminus E_{ij}$; see, e.g., [2, sect. 2] for the definition of these spaces and the corresponding norms. Using the stable extension in [2, Th. 7.5], we find

$$|\psi_{ij}|^2_{H^1(\Omega_i)} \leq C\|\psi_{ij}\|^2_{H^{1/2}(\partial\Omega_i)} \leq C\|\psi_{ij}\|^2_{H^{1/2}_{00}(E_{ij})},$$

and, using [2, Th. 6.6],

$$\|\psi_{ij}\|^2_{H^{1/2}_{00}(E_{ij})} \leq \|\psi_{ij}\|^2_{H^{1/2}(E_{ij})} + C(1+\log k)\|\psi_{ij}\|^2_{L^\infty(E_{ij})}.$$

Combining these two inequalities yields

$$|\psi_{ij}|^2_{H^1(\Omega_i)} \leq C(1+\log k)\|\psi-\psi_H\|^2_{L^\infty(E_{ij})} + \|\psi-\psi_H\|^2_{H^{1/2}(E_{ij})}. \tag{5.2}$$

We note that ψ_H is the nodal interpolant of ψ on the linear space $\mathbb{Q}_{1,1}$, and therefore the inverse inequality in [2, Th. 6.2] can be employed. We obtain

$$(1+\log k)\|\psi-\psi_H\|^2_{L^\infty(E_{ij})} \leq C(1+\log k)^2\|\psi\|^2_{H^{1/2}(E_{ij})} \tag{5.3}$$

and

$$\|\psi-\psi_H\|^2_{H^{1/2}(E_{ij})} \leq C(1+\log k)\|\psi\|^2_{H^{1/2}(E_{ij})}. \tag{5.4}$$

Combining (5.2), (5.3), (5.4), and a trace estimate, we find

$$|\psi_{ij}|^2_{H^1(\Omega_i)} \leq C(1+\log k)^2\|\psi\|^2_{H^1(\Omega_i)}.$$

We note that if we add a constant to ψ, the left-hand side does not change. A quotient space argument then allows us to replace the full norm with the seminorm on the left-hand side. ∎

We recall that the coarse space $X^H(\Omega_i)$ was defined in (4.1). We now introduce the coarse interpolant

$$\rho_H : X^k(\Omega_i) \longrightarrow X^H(\Omega_i).$$

Here, $\rho_H\mathbf{u}$ is the unique vector that satisfies

$$\int_{E_{ij}} (\rho_H\mathbf{u}-\mathbf{u})\cdot\mathbf{t}_{E_{ij}}\,ds = 0, \quad j \in \mathcal{I}(i). \tag{5.5}$$

We also define $X_{ij} \subset X^k(\Omega_i)$ as the space of functions $\nabla\phi_{ij}$, where $\phi_{ij} \in Q^k(\Omega_i)$ is Laplace discrete harmonic and vanishes on $\partial\Omega_i \setminus E_{ij}$.

We are now ready to give a proof of Lemma 4.1. It is immediate to see that, for the substructure Ω_i and for $j \in \mathcal{I}(i)$, $l \in \mathcal{I}(i)$, $j \neq l$,

$$X^H(\Omega_i) \cap X_i^0 = X^H(\Omega_i) \cap X_{ij} = X_{ij} \cap X_i^0 = X_{ij} \cap X_{il} = \{0\}.$$

Counting the degrees of freedom, we see that

$$X^k(\Omega_i) = X^H(\Omega_i) \oplus \sum_{j \in \mathcal{I}(i)} X_{ij} \oplus X_i^0 \tag{5.6}$$

is a direct sum. We have therefore proved the existence and the uniqueness of the decomposition (4.2).

The first equality in (4.3) is a consequence of the fact that the tangential component of $\mathbf{u}_i := \mathbf{u}^{int}$ and of $\mathbf{u}_{il}$ for $l \neq j$ vanishes on the edge E_{ij}. The second one comes from the fact that ϕ_{ij} vanishes at the endpoints of E_{ij}.

We are then left with the proof of the stability property (4.4). Since the decomposition is unique, thanks to (5.5), we find $\mathbf{u}_H = \rho_H \mathbf{u}$. We now decompose each term into a gradient of a scalar function and a remainder. Since the coarse space $X^H(\Omega_i)$ is $\mathcal{R}_1(\Omega_i)$, we can write

$$\mathbf{u}_H = \nabla\phi_H + \alpha \begin{bmatrix} y - y_i \\ x_i - x \end{bmatrix} =: \nabla\phi_H + \mathbf{u}_H^\perp,$$

with $\phi_H \in \mathbb{Q}_{1,1}$ bilinear and (x_i, y_i) the center of gravity of Ω_i. By direct calculation, we find that this is an L^2 orthogonal decomposition and that

$$\|\mathbf{u}_H^\perp\|_{L^2(\Omega_i)} \leq C H_i \|\mathrm{curl}\, \mathbf{u}_H^\perp\|_{L^2(\Omega_i)}. \tag{5.7}$$

For the term $\mathbf{u}_i := \mathbf{u}^{int} \in X_i^0$, we employ the orthogonal decomposition in (5.1) and find

$$\mathbf{u}_i = \nabla\phi_i + \mathbf{u}_i^\perp.$$

Finally, by definition, $\mathbf{u}_{ij} = \nabla\phi_{ij}$ for each edge E_{ij}. We then group the gradient terms and the remainders and set

$$\phi := \phi_H + \sum_{j \in \mathcal{I}(i)} \phi_{ij} + \phi_i, \qquad \mathbf{u}^\perp := \mathbf{u}_H^\perp + \mathbf{u}_i^\perp.$$

We have therefore the decomposition

$$\mathbf{u} = \nabla\phi + \mathbf{u}^\perp. \tag{5.8}$$

We need to bound the $\nabla\phi_{ij}$ in terms of $\mathbf{u}$. Since ϕ_H and the $\{\phi_{il}\}$ are Laplace discrete harmonic, we can apply Lemma 5.2 and find

$$|\phi_{ij}|^2_{H^1(\Omega_i)} \leq C(1 + \log k)^2 \left| \phi_H + \sum_{l \in \mathcal{I}(i)} \phi_{il} \right|^2_{H^1(\Omega_i)} \leq C(1 + \log k)^2 |\phi|^2_{H^1(\Omega_i)}. \tag{5.9}$$

For the last step we also have used that fact that ϕ_i vanishes on $\partial\Omega_i$ and is thus orthogonal to Laplace discrete harmonic functions.

The last step is to bound $\nabla\phi$ in terms of $\mathbf{u}$. We first note that, using (5.7) and Lemma 5.1, we obtain

$$\|\mathbf{u}^{\perp}\|^2_{L^2(\Omega_i)} \leq CH_i^2(\|\operatorname{curl}\mathbf{u}_H^{\perp}\|^2_{L^2(\Omega_i)} + \|\operatorname{curl}\mathbf{u}_i^{\perp}\|^2_{L^2(\Omega_i)}).$$

Since $\operatorname{curl}\mathbf{u}_H^{\perp}$ is constant and $\operatorname{curl}\mathbf{u}_i^{\perp}$ has a vanishing mean value on Ω_i, these two functions are L^2 orthogonal and thus

$$\|\mathbf{u}^{\perp}\|^2_{L^2(\Omega_i)} \leq C_{\perp}H_i^2\|\operatorname{curl}\mathbf{u}^{\perp}\|^2_{L^2(\Omega_i)}. \tag{5.10}$$

Using (5.8), (5.10), and Young's inequality, we find

$$\begin{aligned}\|\mathbf{u}\|^2_{\operatorname{curl},\Omega_i} &= |\phi|^2_{H^1(\Omega_i)} + \|\mathbf{u}^{\perp}\|^2_{\operatorname{curl},\Omega_i} + 2\int_{\Omega_i}\nabla\phi\cdot\mathbf{u}^{\perp}\,dx \\ &\geq (1-\epsilon)|\phi|^2_{H^1(\Omega_i)} + (1+(1-\epsilon^{-1})C_{\perp})H_i^2\|\operatorname{curl}\mathbf{u}^{\perp}\|^2_{L^2(\Omega_i)}\end{aligned} \tag{5.11}$$

for $\epsilon\in(0,1)$. The choice $\epsilon = C_{\perp}/(C_{\perp}+1)$ ensures

$$|\phi|^2_{H^1(\Omega_i)} \leq (C_{\perp}+1)\|\mathbf{u}\|^2_{\operatorname{curl},\Omega_i},$$

which, combined with (5.9), concludes the proof.

6. Implementation aspects. In this section, we describe how we can efficiently implement the preconditioned algorithm described in this paper. Indeed, we need to construct the matrix F, the vector d (see (3.5) and (3.6)), and the preconditioner M^{-1} in (3.9).

In principle, a change of basis should be performed and the degrees of freedom partitioned into I (interior to the substructures), Π (common averages along the subdomain edges), and Δ. However, such change of basis is not trivial or advisable. Since the basis functions associated to the Δ block are not local in general, this would spoil the sparsity of certain matrices. In practice we will work with full vectors in the original product space consisting of all the degrees of freedom on Γ, satisfying no continuity constraint. We will then make sure that these degrees of freedom belong to the dual space $\widetilde{W}_{\Delta}$; i.e., the averages along all the subdomain edges vanish. For this reason, as already pointed out in section 3, the matrix $B_{\Delta} = B$ is the same as that of the one-level FETI method in [26]; it is constructed from $\{0,1,-1\}$ and evaluates the difference between all the corresponding tangential degrees of freedom on Γ.

We then consider an initial vector of Lagrange multipliers λ_0. We note that since we work with the matrix B which acts on the whole space W, in order to ensure that $\lambda_0 \in V = \operatorname{Range}(B_{\Delta})$, we need to choose $\lambda_0 = Bu_0$, with $u_0 \in \widetilde{W}$.

We work with the matrix $K : X \to X$, which acts on the product space and is block diagonal; each block $K^{(i)}$ corresponds to a substructure Ω_i and is the representation of the local bilinear form $a_{\Omega_i}(\cdot,\cdot)$. We also work with global vectors $\mathbf{u}\in X$ and the load vector, still denoted by $\mathbf{f}$, which represents the linear functional

$$\int_{\Omega}\mathbf{f}\cdot\mathbf{w}\,dx, \quad \mathbf{w}\in X. \tag{6.1}$$

In addition, if $w\in W$ is a vector of degrees of freedom on Γ, let

$$\widetilde{R}^T : W \to X$$

be the extension by zero from Γ into the whole of Ω.

We can then write the system for the solution $\mathbf{u} \in X$ as

$$\begin{aligned} K\mathbf{u} + C^T\mu + (B\widetilde{R})^T\lambda &= \mathbf{f}, \\ C\mathbf{u} &= 0, \\ (B\widetilde{R})\mathbf{u} &= 0. \end{aligned} \tag{6.2}$$

Here $C\mathbf{u} = 0$ imposes the constraint to a vector $\mathbf{u} \in X$ that it have vanishing averages along the subdomain edges and μ is a vector of Lagrange multipliers associated to these constraints. We note that the last condition imposes then redundant constraints. An equation for λ is obtained by eliminating $\mathbf{u}$ and μ. We obtain

$$\mathbf{u} = K^{-1}(I - C^T(CK^{-1}C^T)^{-1}CK^{-1})(\mathbf{f} - (B\widetilde{R})^T\lambda) =: H(\mathbf{f} - (B\widetilde{R})^T\lambda)$$

and thus

$$B\widetilde{R}\,H\,\widetilde{R}^T B^T\lambda = B\widetilde{R}\,H\,\mathbf{f}.$$

We finally find

$$\begin{aligned} F &= B\widetilde{R}\,H\,\widetilde{R}^T B^T, \\ d &= B\widetilde{R}\,H\,\mathbf{f}. \end{aligned}$$

We note that $\widetilde{R}\,H\,\widetilde{R}^T$ gives an expression for $\widetilde{S}^{-1}$. In addition, the application of H to a vector requires *two* applications of K^{-1} (and then the solution of two Neumann problems on each substructure) and *one* application of $(CK^{-1}C^T)^{-1}$. If we partition

$$C = [C^{(1)}\ C^{(2)}\ \cdots\ C^{(N)}],$$

with each block corresponding to a substructure, we can write

$$F_0 := CK^{-1}C^T = \sum_{i=1}^{N} C^{(i)}{K^{(i)}}^{-1}{C^{(i)}}^T.$$

Since the number of constraints (and thus of nonzero columns in ${C^{(i)}}^T$) is equal to the number of edges of Ω_i, we need to apply ${K^{(i)}}^{-1}$ only to these nonzero columns in order to calculate F_0. The matrix F_0 is then factored once and for all and its inversion provides a coarse problem, the size of which is equal to the number of edges of the subdomain partition. Finally, by construction, the operator H always returns a vector in the kernel of C, which therefore has vanishing mean value on the subdomain edges. We note that, in case the subdomain partition coincides with a coarse mesh the coarse matrix F_0 has the same size and stencil as the coarse one for the balancing NN method in [24], for which coarse degrees of freedom are also associated to the edges of a coarse mesh.

Concerning the preconditioner M^{-1} in (3.9), the local Schur complements $S^{(i)}$ are the same as those employed for the one-level FETI method in [26] and are obtained from the local stiffness matrices in the standard way; see, e.g., [24, Eq. 3.3].

With the exception of the coarse problem, which is here larger, the cost of the setup of the algorithm is the same as that in [26] since the matrices for the same local Dirichlet and Neumann problems need to be inverted. A similar consideration holds for the cost of each iteration. Each step of our algorithm requires the solution of one coarse and two local Neumann problems per subdomain for the application of F and one Dirichlet problem per subdomain for the application of the preconditioner. The same is true for the algorithm in [26].

TABLE 7.1

FETI-DP method. Estimated condition number and number of CG iterations necessary to obtain a relative preconditioned residual less than 10^{-12} (in parentheses), versus H/h and n. Case of $a = 1$, $b = 1$. The asterisks denote the cases for which we had not enough memory to run the corresponding algorithm.

H/h	32	16	8	4	2
$n = 32$	–	1.529 (5)	2.212 (11)	1.777 (11)	1.309 (8)
$n = 64$	1.801 (6)	2.950 (12)	2.446 (13)	1.806 (10)	1.312 (7)
$n = 128$	3.827 (13)	3.278 (15)	2.484 (12)	1.819 (10)	1.314 (7)
$n = 192$	4.154 (17)	3.329 (15)	2.496 (12)	1.816 (9)	*
$n = 256$	4.265 (17)	3.337 (14)	2.500 (12)	*	*

7. Numerical results. We consider the same mesh, partitions, and coefficient distribution as in [26, sect. 6] in order to allow a comparison with the one-level FETI. The domain $\Omega := (0,1)^2$ is partitioned into two uniform meshes $\mathcal{T}_h$ and $\mathcal{T}_H$. The fine triangulation is made of triangles and the coarse one of squares that are unions of fine triangles. The substructures Ω_i are the elements of the coarse triangulation $\mathcal{T}_H$. The fine triangulation $\mathcal{T}_h$ consists of $2 * n^2$ triangles, with $h = 1/n$. We choose

$$A = \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix}, \qquad \mathbf{f} = [\exp(-x/3 + y^2), -3\cos(2x - 5y - 10)]^T,$$

and we use the value $\chi = 1/2$ for the definition of the scaling matrices $D_\Delta^{(i)}$; see (3.7). We consider a CG algorithm and estimate the condition number of the preconditioned operator using the quantities provided by the CG. Since, however, convergence is much faster here, we employ a more restrictive stopping criterion than in [26] in order to obtain good condition number estimates: we stop the iteration when $\|z_k\|/\|f\|$ is less than 10^{-12} instead of 10^{-6}. Here, z_k is the kth preconditioned residual $M^{-1}(d - F\lambda_k)$. The estimated condition numbers here can then be compared with those in [26, sect. 6], while in order to compare the iteration counts we need to consider the double of those in [26].

In Table 7.1, we show the estimated condition number and the number of iterations as functions of the dimensions of the fine and coarse meshes for $a = b = 1$. For a fixed ratio H/h, the condition number and the number of iterations are quite insensitive to the dimension of the fine mesh and are consistent with a quadratic logarithmic growth; see Theorem 4.7. The condition numbers here can be compared with those in [26, Table 1]. Those for the FETI-DP method are generally slightly smaller than those for the one-level method. This is related to the fact that a coarse problem of larger size is solved here. For the uniform partition into square substructures, we have one coarse function for each substructure for one-level FETI and two for FETI-DP (four degrees of freedom for the four edges of each subdomain, shared by two substructures).

On the other hand, comparison of the iteration counts shows a faster convergence for the FETI-DP algorithm. This is related to the smaller condition number and also to the fact that there is basically no freedom for the initial guess of one-level FETI methods (cf., e.g., the algorithm in [21, p. 100]); this may often give a quite high initial residual. An arbitrary initial guess can be employed for FETI-DP and the null vector employed here provides a relatively small initial residual for our tests.

In Table 7.2, we show some results when the coefficient b has jumps across the interface. We consider a 4×4 checkerboard distribution, where b assumes two values,

TABLE 7.2
4×4 checkerboard distribution for $b : (b_1, b_2)$. Estimated condition number and number of CG iterations to obtain a relative preconditioned residual less than 10^{-12} (in parentheses), versus H/h and b_2. Case of $n = 128$, $a = 1$, and $b_1 = 100$.

H/h	4	8	16
$b2 = 1e-4$	3.777 (21)	5.395 (28)	7.633 (32)
$b2 = 1e-3$	3.760 (20)	5.382 (27)	7.606 (30)
$b2 = 1e-2$	3.713 (20)	5.308 (25)	7.504 (29)
$b2 = 1e-1$	3.561 (18)	5.089 (23)	7.196 (27)
$b2 = 1$	3.155 (16)	4.502 (20)	6.364 (25)
$b2 = 1e+1$	2.355 (13)	3.338 (17)	4.692 (20)
$b2 = 1e+2$	1.800 (10)	2.436 (13)	3.068 (15)
$b2 = 1e+3$	2.298 (13)	3.059 (15)	3.798 (17)
$b2 = 1e+4$	2.612 (14)	3.036 (16)	3.435 (17)
$b2 = 1e+5$	2.203 (12)	2.630 (14)	2.918 (15)
$b2 = 1e+6$	2.085 (12)	2.593 (13)	2.820 (14)

b_1 and b_2. For a fixed value of $n = 128$, $b_1 = 100$, and $a = 1$, the estimated condition number and the number of iterations are shown as a function of H/h and b_2. Similar behavior as in [26, Table 2] is observed here. For $b_2 = 100$, the coefficient b has a uniform distribution, and this corresponds to a local minimum for the condition number and the number of iterations. When b_2 decreases or increases, the condition number and the number of iterations normally increase, but they can still be bounded independently of b_2. We note, however, that for some very large values of b_2, convergence may be faster than in the uniform case. We also remark that, when b_2 is large, the local ratio b_2/a is also large; see η in Theorem 4.7. In this case, however, our results remain good and the condition number even appears to be less sensitive to H/h. We remark that condition numbers and iteration counts are smaller than the corresponding ones in [26, Table 2] for one-level FETI.

In Table 7.3, we show some results when the coefficient a has jumps. We consider the same 4×4 checkerboard distribution shown as for the previous tests. For a fixed value of $n = 128$, $a_1 = 0.01$, and $b = 1$, the estimated condition number and the number of iterations are shown as a function of H/h and a_2. For $a_2 = 0.01$, the

TABLE 7.3
4×4 checkerboard distribution for $a : (a_1, a_2)$. Estimated condition number and number of CG iterations to obtain a relative preconditioned residual less than 10^{-12} (in parentheses), versus H/h and a_2. Case of $n = 128$, $b = 1$, and $a_1 = 0.01$.

H/h	4	8	16
$a2 = 1.e-7$	2.668 (15)	4.342 (20)	7.097 (26)
$a2 = 1.e-6$	2.285 (14)	3.665 (19)	6.024 (25)
$a2 = 1.e-5$	1.769 (12)	2.418 (16)	3.869 (21)
$a2 = 1.e-4$	1.764 (12)	2.294 (15)	2.814 (17)
$a2 = 1.e-3$	1.791 (12)	2.353 (15)	2.814 (17)
$a2 = 1.e-2$	1.813 (13)	2.447 (16)	3.071 (18)
$a2 = 1.e-1$	1.816 (12)	2.467 (15)	3.173 (18)
$a2 = 1$	1.808 (10)	2.466 (14)	3.182 (16)
$a2 = 1.e+1$	1.801 (9)	2.454 (12)	3.172 (14)
$a2 = 1.e+2$	1.791 (8)	2.438 (10)	3.164 (12)
$a2 = 1.e+3$	1.771 (7)	2.427 (9)	3.159 (11)

TABLE 7.4
5×5 checkerboard distribution for $b : (b_1, b_2)$. Estimated condition number and number of CG iterations to obtain a relative preconditioned residual less than 10^{-12} (in parentheses), versus H/h and b_2. Case of $n = 128$, $a = 1$, and $b_1 = 100$.

H/h	4	8	16
$b2 = 1e-3$	$1.373e+04$ (>100)	42.26 (60)	9.056 (26)
$b2 = 1e-2$	1390 (>100)	37.22 (55)	8.824 (25)
$b2 = 1e-1$	145.1 (>100)	33.85 (53)	8.1 (24)
$b2 = 1$	87.32 (92)	23.6 (43)	5.99 (23)
$b2 = 1e+1$	20.91 (45)	6.342 (23)	3.588 (17)
$b2 = 1e+2$	1.800 (10)	2.436 (13)	3.068 (15)
$b2 = 1e+3$	3.115 (15)	2.664 (14)	2.639 (14)
$b2 = 1e+4$	10.53 (28)	2.847 (14)	1.855 (11)
$b2 = 1e+5$	41.41 (31)	2.857 (14)	1.761 (11)
$b2 = 1e+6$	107.7 (36)	2.87 (14)	1.82 (10)

coefficient a has a uniform distribution. A slight increase in the number of iterations and the condition number may be observed for some larger or smaller values of a_2 and when H/h is large. As for the previous table, when a_2 is small, the local ratio b/a_2 is large and our results remain good. The condition numbers and the iteration counts are smaller than the corresponding ones in [26, Table 3].

We finally present a test case where coefficient discontinuities do not coincide with the subdomain interfaces. We consider a 5×5 checkerboard distribution for the coefficient b. For a fixed value of $n = 128$, $b_1 = 100$, and $a = 1$, the estimated condition number and the number of iterations are shown in Table 7.4 as a function of $H/h = 4, 8, 16$ (corresponding to 32×32, 16×16, and 8×8 subdomains) and b_2. We see that there are cases corresponding to a large number of subdomains and small values of b for which convergence may deteriorate; the assumption that the coefficient b does not vary much in each subdomain may be therefore required in practice. The analogous case for jumps in the coefficient a, on the other hand, produces results that are similar to those in Table 7.3, and they are not presented here.

REFERENCES

[1] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Multigrid in H(div) and H(curl)*, Numer. Math., 85 (2000), pp. 197–217.

[2] I. BABUŠKA, A. CRAIG, J. MANDEL, AND J. PITKÄRANTA, *Efficient preconditioning for the p-version finite element method in two dimensions*, SIAM J. Numer. Anal., 28 (1991), pp. 624–661.

[3] F. B. BELGACEM AND C. BERNARDI, *Spectral element discretization of the Maxwell equations*, Math. Comp., 68 (1999), pp. 1497–1520.

[4] M. BHARDWAJ, K. PIERSON, G. REESE, T. WALSH, D. DAY, K. ALVIN, J. PEERY, C. FARHAT, AND M. LESOINNE, *Salinas: A scalable software for high-performance structural and solid mechanics simulations*, in Proceedings of the IEEE/ACM Supercomputing Conference, Baltimore, MD, 2002, p. 35.

[5] P. B. BOCHEV, C. J. GARASI, J. J. HU, A. C. ROBINSON, AND R. S. TUMINARO, *An improved algebraic multigrid method for solving Maxwell's equations*, SIAM J. Sci. Comput., 25 (2003), pp. 623–642.

[6] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[7] W. CECOT, W. RACHOWICZ, AND L. DEMKOWICZ, *An hp-adaptive finite element method for electromagnetics.* III. *A three-dimensional infinite element for Maxwell's equations*, Internat. J. Numer. Methods Engrg., 57 (2003), pp. 899–921.

[8] C. Farhat, M. Lesoinne, P. LeTallec, K. Pierson, and D. Rixen, *FETI-DP: A dual-primal unified FETI method.* I. *A faster alternative to the two-level FETI method*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.

[9] C. Farhat and F.-X. Roux, *Implicit parallel processing in structural mechanics*, Comput. Mech. Adv., 2 (1994), pp. 1–124.

[10] B. Hientzsch, *Fast Solvers and Domain Decomposition Preconditioners for Spectral Element Discretizations of Problems in H(curl)*, Ph.D. thesis, Tech. rep. 823, Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, 2001.

[11] B. Hientzsch, *Schwarz Preconditioners for Spectral Nédélec Elements for a Model Problem in H(curl)*, Tech. rep. 834, Courant Institute for Mathematical Sciences, New York University, New York, 2002.

[12] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.

[13] A. Klawonn and O. B. Widlund, *FETI and Neumann-Neumann iterative substructuring methods: Connections and new results*, Comm. Pure Appl. Math., 54 (2001), pp. 57–90.

[14] A. Klawonn and O. B. Widlund, *Selecting constraints in dual-primal FETI methods for elasticity in three dimensions*, in Proceedings of the 15th International Conference on Domain Decomposition Methods (Berlin, 2003), R. Kornhuber et al., eds., Lect. Notes Comput. Sci. Engrg., 40, Springer-Verlag; Berlin, 2004, pp. 67–81.

[15] A. Klawonn, O. B. Widlund, and M. Dryja, *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients*, SIAM J. Numer. Anal., 40 (2002), pp. 159–179.

[16] J. Mandel and M. Brezina, *Balancing domain decomposition for problems with large jumps in coefficients*, Math. Comp., 65 (1996), pp. 1387–1401.

[17] J. Mandel and R. Tezaur, *On the convergence of a dual-primal substructuring method*, Numer. Math., 88 (2001), pp. 543–558.

[18] P. Monk, *On the p- and hp-extension of Nédélec's curl-conforming elements*, J. Comput. Appl. Math., 53 (1994), pp. 117–137.

[19] P. Monk, *Finite Element Methods for Maxwell's Equations*, Numer. Math. Sci. Comput., Oxford University Press, New York, 2003.

[20] J.-C. Nédélec, *Mixed finite elements in R^3*, Numer. Math., 35 (1980), pp. 315–341.

[21] F. Rapetti and A. Toselli, *A FETI preconditioner for two dimensional edge element approximations of Maxwell's equations on nonmatching grids*, SIAM J. Sci. Comput., 23 (2001), pp. 92–108.

[22] S. Reitzinger and J. Schöberl, *An algebraic multigrid method for finite element discretizations with edge elements*, Numer. Linear Algebra Appl., 9 (2002), pp. 223–238.

[23] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[24] A. Toselli, *Neumann–Neumann methods for vector field problems*, Electron. Trans. Numer. Anal., 11 (2000), pp. 1–24.

[25] A. Toselli, *Two iterative substructuring methods for Maxwell's equations with discontinuous coefficients in two dimensions*, in Proceedings of the 12th International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. Chan, T. Kako, H. Kawarada, and O. Pironneau, eds., Chiba, Japan, 2002, pp. 215–222.

[26] A. Toselli and A. Klawonn, *A FETI domain decomposition method for edge element approximations in two dimensions with discontinuous coefficients*, SIAM J. Numer. Anal., 39 (2001), pp. 932–956.

[27] A. Toselli and X. Vasseur, *Neumann-Neumann and FETI Preconditioners for hp-Approximations on Geometrically Refined Boundary Layer Meshes in Two Dimensions*, Tech. rep. 02–15, Seminar für Angewandte Mathematik, ETH, Zürich, 2002; Numer. Math., submitted.

[28] A. Toselli and X. Vasseur, *A numerical study on Neumann-Neumann and FETI methods for hp-approximations on geometrically refined boundary layer meshes in two dimensions*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 4551–4579.

[29] A. Toselli and X. Vasseur, *Domain decomposition preconditioners of Neumann-Neumann type for hp-approximations on boundary layer meshes in three dimensions*, IMA J. Numer. Anal., 24 (2004), pp. 123–156.

[30] A. Toselli, O. B. Widlund, and B. I. Wohlmuth, *An iterative substructuring method for Maxwell's equations in two dimensions*, Math. Comp., 70 (2001), pp. 935–949.

[31] K. F. Traore, C. Farhat, M. Lesoinne, and D. Dureisseix, *A domain decomposition method with Lagrange multipliers for the massively parallel solution of large-scale contact problems*, in Proceedings of the 5th World Congress on Computational Mechanics, Vienna University of Technology, Austria, 2002.

[32] B. I. Wohlmuth, A. Toselli, and O. B. Widlund, *An iterative substructuring method for Raviart–Thomas vector fields in three dimensions*, SIAM J. Numer. Anal., 37 (2000), pp. 1657–1676.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2612–2632

FAST SWEEPING METHODS FOR STATIC HAMILTON–JACOBI EQUATIONS*

CHIU-YEN KAO†, STANLEY OSHER†, AND YEN-HSI TSAI‡

Abstract. We propose a new sweeping algorithm which discretizes the Legendre transform of the numerical Hamiltonian using an explicit formula. This formula yields the numerical solution at a grid point using only its immediate neighboring grid values and is easy to implement numerically. The minimization that is related to the Legendre transform in our sweeping scheme can either be solved analytically or numerically. We illustrate the efficiency and accuracy approach with several numerical examples in two and three dimensions.

Key words. fast sweeping method, Godunov Hamiltonian, static Hamilton–Jacobi equation

AMS subject classifications. 35F30, 65N06

DOI. 10.1137/S0036142902419600

1. Introduction.

The Hamilton–Jacobi equation

$$\psi_t(x,t) + H(x, \nabla\psi(x,t)) = 0 \tag{1.1}$$

arises in many applications ranging from classical mechanics to contemporary problems of optimal control. These include geometrical optics, crystal growth, etching, computer vision, obstacle navigation, path planning, photolithography, and seismology. In general, these nonlinear PDEs cannot be solved analytically. The solutions usually develop singularities in their derivatives even with smooth initial conditions. In these cases, the solutions do not satisfy the equation in the classical sense. The weak solution that is usually sought is called the viscosity solution [10]. Numerically, in general, one looks for a consistent and monotone scheme to construct approximate viscosity solutions [27].

In this paper, we focus on static Hamilton–Jacobi equations of the following form:

$$\begin{cases} H(x, \nabla\phi(x)) = R(x) & \text{for } x \in \Omega, \\ \phi(x) = q(x) & \text{for } x \in \Gamma \subset \partial\Omega, \end{cases} \tag{1.2}$$

where H, q, and $R > 0$ are Lipschitz continuous and H is also convex and homogeneous of degree one in $\nabla\phi(x)$. A special case of this type of equation is the eikonal equation,

$$|\nabla\phi| = r(x) \tag{1.3}$$

with the same type of Dirichlet boundary condition as in (1.2). Many numerical methods have been developed for this problem. Rouy and Tourin [24] used an iterative method to solve the discretized eikonal equation and proved that it converges to the viscosity solution. The key is to use an upwind, monotone, and consistent discretization for $|\nabla\phi|$. Instead of using iterative methods, Tsitsiklis [29], later Sethian [25],

*Received by the editors December 11, 2002; accepted for publication (in revised form) April 3, 2004; published electronically April 19, 2005. This research was supported by ONR grant N00014-03-1-0071 and ONR MURI grant N00014-02-1-0720.
http://www.siam.org/journals/sinum/42-6/41960.html
†Department of Mathematics, University of California Los Angeles, Los Angeles, CA 90095 (ckao@math.ucla.edu, sjo@math.ucla.edu).
‡Department of Mathematics and PACM, Princeton University, Princeton, NJ 08544 (ytsai@math.princeton.edu).

and Helmsen et. al. [14] proposed single-pass methods. Based on the monotonicity of the solution along the characteristics, they combined the heap-sort data structure with a variation of the classical Dijkstra algorithm to solve the steady state equation $|\nabla\phi| = r(\mathbf{x})$. This became known as the fast marching method whose complexity is $O(N \log N)$, where N is the total number of grid points in the domain. Later Sethian and Vladimirsky [26] generalized the method of [29] to solve (1.2).

Osher [18] provided a link between time-independent and time-dependent Hamilton–Jacobi equations. The zero level set of the viscosity solution ψ of (1.1) with suitable initial conditions at various time t is the solution $\phi(x, y) = t$ of (1.2). This gives an approach that one can try to solve the time-dependent equation by the level set formulation [19] with high order approximations on the partial derivatives [20], [15]. Falcone and Ferretti studied a class of semi-Lagrangian schemes which can be interpreted as a discrete version of the Hopf–Lax–Oleinik representation formula for first order time-dependent Hamilton–Jacobi equations. In semi-Lagrangian schemes, ψ needs to be interpolated using its grid values, the Legendre transformation of H needs to be obtained, and the minimum must be computed on an unbounded set. See [11] and the references therein for more details.

Another approach to obtaining a "time-dependent" Hamilton–Jacobi equation from a time-independent Hamilton–Jacobi equation comes by using the so-called paraxial formulation, i.e., by assuming that there is a preferred direction in the wave propagation. In [13], the paraxial formulation was first proposed for the eikonal equation (1.3). Later in [22], [23], a paraxial formulation was proposed for the static general eikonal equation (1.2) in geophysical applications.

An important application for (1.2) is obtaining geodesic distance on a manifold. Suppose that $P = (x, y)$ is a point on a manifold M defined as the graph of a smooth function $f(x, y)$ and that γ are the curves connecting P and $\Gamma \subset M$ on the manifold. The minimizing curve of γ is called the geodesic. Let ϕ be the distance function such that

$$\phi(x, y) = \min_{\gamma \subset M} \int_\gamma ds.$$

Then ϕ is the solution of

$$\sqrt{\left(\frac{1+f_y^2}{f_x^2+f_y^2+1}\right)\phi_x^2 + \left(\frac{1+f_x^2}{f_x^2+f_y^2+1}\right)\phi_y^2 - 2\frac{f_x f_y}{f_x^2+f_y^2+1}\phi_x\phi_y} = 1, \quad \phi|_\Gamma = 0. \tag{1.4}$$

This equation can be easily generalized to higher dimensions. For example, in three dimensions we again write down the formula for M as the graph of a smooth function $f(x, y, z)$. The distance function ϕ then satisfies

$$\sqrt{a\phi_x^2 + b\phi_y^2 + c\phi_z^2 - 2d\phi_x\phi_y - 2e\phi_y\phi_z - 2f\phi_z\phi_x} = 1, \quad \phi|_\Gamma = 0, \tag{1.5}$$

where

$$a = \frac{1+f_y^2+f_z^2}{1+f_x^2+f_y^2+f_z^2}, \quad b = \frac{1+f_x^2+f_z^2}{1+f_x^2+f_y^2+f_z^2}, \quad c = \frac{1+f_x^2+f_y^2}{1+f_x^2+f_y^2+f_z^2},$$
$$d = \frac{f_x f_y}{1+f_x^2+f_y^2+f_z^2}, \quad e = \frac{f_y f_z}{1+f_x^2+f_y^2+f_z^2}, \quad f = \frac{f_z f_x}{1+f_x^2+f_y^2+f_z^2}.$$

We will apply our new algorithm to compute the geodesic distance later. There are other approaches that are designed to compute geodesic distances on manifolds. Kimmel and Sethian [16] extended the fast marching method to triangulated manifolds and provided an algorithm for computing the geodesic distances, thereby extracting shortest paths on triangulated manifolds. Barth [2] used the discontinuous Galerkin method to find the distance on graphs of functions that are represented by spline functions. In [7], the authors embeded the manifold as the zero level set of a Lipschitz continuous function and solved the corresponding eikonal equation (1.4) in the embedding space. In [17], the authors based their work on the theory of geodesics on Riemannian manifolds with boundaries and adapted the standard fast marching method to compute weighted distance functions and geodesics on implicit surfaces efficiently. Tsai et. al. [28] used a fast Gauss–Seidel-type iteration method and a monotone upwind Godunov flux for the numerical Hamiltonian.

We propose a new interpretation of the monotone upwind Godunov flux for the numerical Hamiltonian to solve (1.2). The complexity of our method appears to be $O(N)$. We illustrate the approach with several numerical examples in two and three dimensions.

2. A new numerical scheme for convex Hamiltonians. Our new numerical algorithm for static Hamilton–Jacobi equations is composed of a sweeping process and an update formula. The sweeping process we use here is a version of Gauss–Seidel iteration. It is motivated originally by Boué and Dupuis [3], who first suggested that the complexity of this approach for the eikonal case is $O(N)$. In [31], the fast sweeping algorithm was first formulated in PDE framework for the eikonal equation and was used to compute the distance function. In the sweeping process, we sweep through the grids with alternating directions in order to follow the characteristics and use the most recent values as we update the solution. This means that we overwrite an old value with its new value as soon as we obtain the latter. In one dimension, we sweep through the grids from left to right followed by right to left because the characteristics have only two possible directions. In two dimensions, the characteristics may have an infinite number of possible directions. We use four sweeping directions so that a specific sweeping direction covers a group of characteristics at the same time. We denote these four sweeping directions as one iteration. In n dimensions, we will use 2^n alternating directions per iteration. We stop our iterations when the L_1 norm of the difference of two successive iteration results is less than the given tolerance, which is $O(h)$, where h is the grid size.

The new update formula we derive here comes from using the Legendre transformation. The Legendre transformation can be applied to the Wulff problem [21], which is usedto determine the equilibrium shape of crystalline materials. We give the definitions in the following.

DEFINITION 2.1. *Let $\gamma : S^{d-1} \to R^+$ be a continuous function defined on a curved space S^{d-1}.*

1. The first Legendre transformation of γ is

$$\gamma_*(\nu) = \min_{\theta\cdot\nu>0,\,|\theta|=1} \left[\frac{\gamma(\theta)}{(\theta\cdot\nu)}\right].$$

2. The second Legendre transformation of γ is

$$\gamma^*(\nu) = \max_{\theta\cdot\nu>0,\,|\theta|=1} [\gamma(\theta)(\theta\cdot\nu)].$$

The first and second Legendre transformations are dual to each other in a certain sense, i.e., $(\gamma_*)^* = \gamma$ if γ is convex and $(\gamma^*)_* = \gamma$ if γ is polar-convex. See, e.g., [21]. We can extend γ to the whole space R^d by defining

$$\widetilde{\gamma}(x) = |x|\gamma\left(\frac{x}{|x|}\right),$$

where the extension $\widetilde{\gamma}$ is homogeneous to degree 1 and $x \in R^d$.

The convex Hamiltonian using the Bellman formula or the Legendre transformation is

$$H(\nabla\phi(x)) = \max_{\theta}[(\nabla\phi \cdot \theta)w(\theta)], \quad \theta \in S^{d-1},$$

where

$$w(\theta) = \min_{\nu\cdot\theta>0,\,|\nu|=1}\left[\frac{H(\nu)}{(\nu\cdot\theta)}\right] \quad \text{and} \quad \nu = \frac{\nabla\phi(x)}{|\nabla\phi(x)|}. \tag{2.1}$$

We define the numerical Hamiltonian as follows:

$$\hat{H}(D_-^i\phi; D_+^j\phi) = \max_{\theta}\left\{\left(\sum_k D_{\mp}^k\phi\cdot\theta_k^{\pm}\right)w(\theta)\right\},$$

where $D_-^i\phi$ ($D_+^j\phi$) are the backward (forward) difference in i (j) direction, $\theta^+ = \max(\theta, 0)$, and $\theta^- = \min(\theta, 0)$. This numerical Hamiltonian is monotone and consistent. It also turns out to be Godunov's numerical Hamiltonian. In order to describe this clearly without loss of generality, we discuss the two-dimensional case here,

$$H(\phi_x, \phi_y) = \max_{\theta}(\phi_x\cos\theta + \phi_y\sin\theta)w(\theta),$$

where

$$w(\theta) = \min_{-\frac{\pi}{2}\le\nu-\theta\le\frac{\pi}{2}}\frac{H(\cos\nu, \sin\nu)}{\cos(\nu-\theta)}.$$

The new numerical Hamiltonian is

$$\hat{H}(D_-^x\phi, D_+^x\phi; D_-^y\phi, D_+^y\phi) = \max_{\theta}\{((\cos\theta)^{\pm}D_{\mp}^x\phi + (\sin\theta)^{\pm}D_{\mp}^y\phi)w(\theta)\}.$$

We say a function $H(x_1, x_2, \dots, x_n)$ is nondecreasing in x_j by writing $H(x_1, x_2, \dots, x_{j-1}, \uparrow, x_{j+1}, \dots, x_n)$ and nonincreasing by writing $H(x_1, x_2, \dots, x_{j-1}, \downarrow, x_{j+1}, \dots, x_n)$.

LEMMA 2.2. *$\hat{H}$ is monotone; i.e., $\hat{H}(\uparrow, \downarrow, \uparrow, \downarrow)$.*

Proof. Since $w > 0$, this conclusion is straightforward. ∎

LEMMA 2.3. *$\hat{H}$ is consistent; i.e., $\hat{H}(p, p; q, q) = H(p, q)$.*

Proof. This is a simple manipulation of the following definitions:

$$\begin{aligned}
\hat{H}(p_-, p_+; q_-, q_+) &:= \max_{\theta}\{((\cos\theta)^{\pm}p_{\mp} + (\sin\theta)^{\pm}q_{\mp})w(\theta)\},\\
\hat{H}(p, p; q, q) &= \max_{\theta}\{((\cos\theta)^{\pm}p + (\sin\theta)^{\pm}q)w(\theta)\}\\
&= \max_{\theta}\{(p\cos\theta + q\sin\theta)w(\theta)\}\\
&=: H(p, q). \qquad ∎
\end{aligned}$$

By solving the Riemann problem for Hamilton–Jacobi equations (a generalization of Godunov's procedure), Bardi and Osher [1] proved the following result for Godunov's scheme:

$$H^G(p_-, p_+; q_-, q_+) = \operatorname{ext}_{p\in I[p_-,p_+]} \operatorname{ext}_{q\in I[q_-,q_+]} H(p,q), \tag{2.2}$$

where

$$\begin{aligned}
\operatorname{ext}_{p\in I[a,b]} &= \min_{p\in[a,b]} \quad \text{if } a \le b,\\
\operatorname{ext}_{p\in I[a,b]} &= \max_{p\in[b,a]} \quad \text{if } a > b,\\
H^G(D^x_- \phi_{ij}, D^x_+ \phi_{ij}; D^y_- \phi_{ij}, D^y_+ \phi_{ij}) &= H^G(p_-, p_+; q_-, q_+),
\end{aligned}$$

and $I[a,b]$ denotes the closed interval bounded by a and b.

PROPOSITION 2.4. *$\hat{H}$ is Godunov's numerical Hamiltonian; i.e., $\hat{H} = H^G$.*

Proof. We first assume $p_- < p_+$ and $q_- < q_+$,

$$\begin{aligned}
H^G(p_-, p_+; q_-, q_+) &:= \min_{p_-\le p\le p_+} \min_{q_-\le q\le q_+} H(p,q)\\
&= \min_{p_-\le p\le p_+} \min_{q_-\le q\le q_+} \left\{ \max_\theta \{(p\cos\theta + q\sin\theta) w(\theta)\} \right\}\\
&= \max_\theta \left\{ \min_{p_-\le p\le p_+} \min_{q_-\le q\le q_+} (p\cos\theta + q\sin\theta) w(\theta) \right\}\\
&= \max_\theta \{((\cos\theta)^{\pm} p_{\mp} + (\sin\theta)^{\pm} q_{\mp}) w(\theta)\}\\
&=: \hat{H}(p_-, p_+; q_-, q_+).
\end{aligned}$$

The proof for the other 3 cases is equally straightforward. ∎

Now we use our new numerical Hamiltonian to solve (1.2). In order to write our scheme in an explicit form, we prove the following property first.

LEMMA 2.5. *$\max_\theta (a f(\theta) - g(\theta)) = 0$ with $f(\theta) > 0 \iff a = \min_\theta \frac{g(\theta)}{f(\theta)}$.*

Proof.

$$\max_\theta (a f(\theta) - g(\theta)) = \max_\theta f(\theta) \left(a - \frac{g(\theta)}{f(\theta)} \right) = 0.$$

Since $f(\theta) > 0$, we have $\max_\theta (a - \frac{g(\theta)}{f(\theta)}) = 0$, which implies $a = \min_\theta \frac{g(\theta)}{f(\theta)}$. ∎

Apply this property to

$$\hat{H}(D^x_- \phi, D^x_+ \phi; D^y_+ \phi, D^y_- \phi) = R(x,y).$$

Let $\phi_0 = \phi_{i,j}$, $\phi_W = \phi_{i-1,j}$, $\phi_E = \phi_{i+1,j}$, $\phi_S = \phi_{i,j-1}$, and $\phi_N = \phi_{i,j+1}$. Breaking down the expressions, we have

$$\max_{\theta,\phi_{W,E,S,N}} \left\{ \begin{Bmatrix} (\cos\theta)^+(\phi_O - \phi_W) \\ -(\cos\theta)^-(\phi_O - \phi_E) \end{Bmatrix} + \begin{Bmatrix} (\sin\theta)^+(\phi_O - \phi_S) \\ -(\sin\theta)^-(\phi_O - \phi_N) \end{Bmatrix} \right\} w(\theta) - hR(x_i, y_j) = 0,$$

$$\max_{\theta,\phi_{W,E,S,N}} \phi_O((\cos\theta)^+ - (\cos\theta)^- + (\sin\theta)^+ - (\sin\theta)^-) w(\theta)$$

$$+ \begin{Bmatrix} -(\cos\theta)^+ \phi_W \\ (\cos\theta)^- \phi_E \end{Bmatrix} + \begin{Bmatrix} -(\sin\theta)^+ \phi_S \\ (\sin\theta)^- \phi_N \end{Bmatrix} w(\theta) = hR(x_i, y_j).$$

Thus

$$\begin{aligned}\phi_O &= \min_{\theta}\left\{\frac{\begin{Bmatrix}(\cos\theta)^+\phi_W & + & (\sin\theta)^+\phi_S\\ -(\cos\theta)^-\phi_E & + & -(\sin\theta)^-\phi_N\end{Bmatrix} w(\theta) + hR(x_i,y_j)}{(|\cos\theta|+|\sin\theta|)w(\theta)}\right\}\\ &= \min_{\theta} K(\theta).\end{aligned} \tag{2.3}$$

We can also derive the three-dimensional numerical Hamiltonian and the update formula in the same way. Let $\phi_0 = \phi_{i,j.k}$, $\phi_W = \phi_{i-1,j,k}$, $\phi_E = \phi_{i+1,j,k}$, $\phi_S = \phi_{i,j-1,k}$, $\phi_N = \phi_{i,j+1,k}$, $\phi_D = \phi_{i,j,k-1}$, and $\phi_U = \phi_{i,j,k+1}$. We have

$$\begin{aligned}&\hat{H}(D^x_-\phi, D^x_+\phi; D^y_-\phi, D^y_+\phi; D^z_-\phi, D^z_+\phi)\\ &\quad = \max_{\theta_1,\theta_2}\{((\sin\theta_1\cos\theta_2)^{\pm}D^x_{\mp}\phi + (\sin\theta_1\sin\theta_2)^{\pm}D^y_{\mp}\phi + (\cos\theta_1)^{\pm}D^z_{\mp}\phi)w(\theta_1,\theta_2)\},\\ &\phi_O = \min_{\theta_1,\theta_2}\left\{\frac{\begin{Bmatrix}(\sin\theta_1\cos\theta_2)^+\phi_W & + & (\sin\theta_1\sin\theta_2)^+\phi_S & + & (\cos\theta_1)^+\phi_D\\ -(\sin\theta_1\cos\theta_2)^-\phi_E & + & -(\sin\theta_1\sin\theta_2)^-\phi_N & + & -(\cos\theta_1)^-\phi_U\end{Bmatrix} w + hR}{(|\sin\theta_1\cos\theta_2|+|\sin\theta_1\cos\theta_2|+|\cos\theta_1|)w}\right\}.\end{aligned} \tag{2.4}$$

Sometimes it is possible to obtain explicit expression for w from (2.1), but in general, one has to use numerical approximations by the fast Legendre transform developed by Brenier [4] and Corrias [9]. The minimization in the update formulas (2.3) and (2.4) can be achieved either analytically or numerically. For a Hamiltonian of quadratic form in the gradient, we solve the minimization analytically in the next section. For other cases, we find the minimizer by using some well-developed numerical optimization techniques, e.g., L-BFGS-B [5], [32] and trust region methods that employ quadratic interpolation [12], [8].

3. Analytically solving a class of Hamilton–Jacobi equations. The quadratic form Hamiltonian

$$\sqrt{a(x,y)\phi_x^2 + b(x,y)\phi_y^2 - 2c(x,y)\phi_x\phi_y} = R(x,y) \tag{3.1}$$

is of special interest because computing geodesic distances on a manifold leads to this type of equation. Here we show that the minimization of (2.3) can be solved explicitly. Using the Legendre transformation, we have (after some simple calculations)

$$H(\cos\nu,\sin\nu) = \sqrt{a\cos^2\nu + b\sin^2\nu - 2c\sin\nu\cos\nu}$$

and

$$w(\theta) = \sqrt{\frac{ab-c^2}{a\sin^2\theta + b\cos^2\theta + 2c\cos\theta\sin\theta}}.$$

Finding the minimum of (2.3) when $0 < \theta < \pi/2$ first, $\frac{dK}{d\theta} = 0$ leads to

$$(-\phi_W + \phi_S)w^2 - hR[(\cos\theta + \sin\theta)w' + (-\sin\theta + \cos\theta)w] = 0. \tag{3.2}$$

Thus

$$\frac{-\phi_W+\phi_S}{hR}=\frac{-a\sin\theta+b\cos\theta+c(\sin\theta-\cos\theta)}{\sqrt{(ab-c^2)(a\sin^2\theta+b\cos^2\theta+2c\sin\theta\cos\theta)}}=T(\theta) \tag{3.3}$$

and

$$T'(\theta)=-\frac{\sqrt{ab-c^2}(\cos\theta+\sin\theta)}{(a\sin^2\theta+b\cos^2\theta+2c\sin\theta\cos\theta)^{3/2}}<0.$$

The solvability condition for θ_1 is

$$\frac{c-a}{\sqrt{a(ab-c^2)}}<\frac{-\phi_W+\phi_S}{hR}<\frac{b-c}{\sqrt{b(ab-c^2)}}. \tag{3.4}$$

If (3.4) is satisfied, we will have a unique solution for $0<\theta<\pi/2$ because of the monotonicity of T. Let $m=(-\phi_W+\phi_S)/hR$. We have

$$\theta=\tan^{-1}\left(\frac{-cm^2(ab-c^2)-(a-c)(b-c)\pm m(ab-c^2)\sqrt{(a+b-2c)-m^2(ab-c^2)}}{am^2(ab-c^2)-(a-c)^2}\right)$$

if both m and the denominator are not zero. Here we have two choices for θ because we square both sides while we do the calculation. We need to plug in (3.3) and pick up the right one. Also

$$\theta=\tan^{-1}\left(\frac{b-a}{c-a}\right)$$

if the denominator is zero, and

$$\theta=\tan^{-1}\left(\frac{c-b}{c-a}\right)$$

if $m=0$. Using similar arguments, we can write down solvability conditions and explicit formulas for θ in other ranges. This can be summarized in the following algorithm.

Algorithm (Quadratic Hamilton–Jacobi solver using the Bellman formula). We assume that $\phi(i,j)$ is given in a small neighborhood of Γ. We initialize the unknown ϕ by setting $\phi(i,j)$ to ∞[1] and mask $(i,j)=$ unknown.

We begin by setting $\phi^{(0)}=\phi$.

Do the following steps while $|\phi^{(n+1)}-\phi^{(n)}|>\delta$: ($\delta>0$ is the given tolerance which is $O(h)$).

Sweeping Process: A compact way of writing these sweeping iterations in C/C++ is

```
for(s1=-1;s1<=1;s1+=2)
for(s2=-1;s2<=1;s2+=2)
for(i=(s1<0?nx:0);(s1<0?i>=0:i<=nx);i+=s1)
for(j=(s2<0?ny:0);(s2<0?j>=0:j<=ny);j+=s2)
  update φ_{i,j}
```

Update Formula: For each grid point (i,j) visited in the sweeping iteration, if mask $(i,j)=$ unknown, do the following:

For $(s_x,s_y)=(\pm1,\pm1)$

[1]Notice that we only need to use a large value in actual implementation.

1. Check the solvability condition

$$m = \frac{s_x s_y(\phi^{(n)}(i, j - s_y) - \phi^{(n)}(i - s_x, j))}{hR},$$
$$\text{check } \frac{c-a}{\sqrt{a(ab-c^2)}} < m < \frac{b-c}{\sqrt{b(ab-c^2)}} \quad \text{when } s_x s_y > 0,$$
$$\text{check } \frac{-(b+c)}{\sqrt{b(ab-c^2)}} < m < \frac{a+c}{\sqrt{a(ab-c^2)}} \quad \text{when } s_x s_y < 0.$$

2. If the condition is satisfied,

$$\theta = \tan^{-1}\left(\frac{-cm^2(ab-c^2) - (as_x - cs_y)(bs_y - cs_x)}{am^2(ab-c^2) - (as_x - cs_y)^2} \pm \frac{m(ab-c^2)\sqrt{(a+b-2cs_x s_y) - m^2(ab-c^2)}}{am^2(ab-c^2) - (as_x - cs_y)^2}\right) + (1 - s_x)\frac{\pi}{2}$$

if both m and the denominator are not zero. Plug in the test function

$$T(\theta) = \frac{(-as_x + cs_y)\sin\theta + (bs_y - cs_x)\cos\theta}{\sqrt{(ab-c^2)(a\sin^2\theta + b\cos^2\theta + 2c\sin\theta\cos\theta)}}$$

and pick up the right one which equals m, not $-m$. Also

$$\theta = \tan^{-1}\left(\frac{b-a}{c - as_x s_y}\right) + (1 - s_x)\frac{\pi}{2}$$

if the denominator is zero, and

$$\theta = \tan^{-1}\left(\frac{cs_x - bs_y}{cs_y - as_x}\right) + (1 - s_x)\frac{\pi}{2}$$

if $m = 0$.

3. Add

$$\phi_{tmp} = \frac{(s_x\phi(i - s_x, j)\cos\theta + s_y\phi(i, j - s_y)\sin\theta)w(\theta) + hR}{(|\cos\theta| + |\sin\theta|)w(\theta)}$$

to the list `phi_candidate`.

4. Add $K(0)$, $K(\frac{\pi}{2})$, $K(\pi)$, $K(\frac{3\pi}{2})$ to the list `phi_candidate`.
5. Let $\phi_{\min}$ be the minimum element of `phi_candidate`.
6. Update

$$\phi^{(n+1)}(i, j) = \min(\phi^{(n)}(i, j), \phi_{\min}).$$

4. Numerical minimization. For a more general sweeping algorithm, we use numerical optimization to calculate ϕ_0. There are many minimization methods that are readily available to us. Some methods need only evaluations of the function while others also require evaluations of the derivative of the function. For our multidimensional cases, we use the L-BFGS-B method [5], [32], [6] because the cost of the iteration is low and the storage requirements of the algorithm are modest. L-BFGS-B is a limited memory quasi-Newton method for a large-scale bound-constrained

problem. The minimizer $\widetilde{\theta}$ of (2.3) and the minimizer $(\widetilde{\theta_1}, \widetilde{\theta_2})$ of (2.4) at a grid point is constructed to be within a given tolerance through iterations, and the number of iterations depends on the initial condition and the tolerance. In our algorithm, we use the minimizer obtained in the previous sweep as our initial guess. In the first sweep, we use the minimizers of the upwind neighboring grid pronts as initial conditions for the quasi-Newton method. This implies that the initial conditions that we end up using are, in most cases, close enough to the minimizers. In practice, with the tolerance of 10^{-6}, we observed that, in average, only four to five iterations are needed. There is an alternative approach of discretizing θ and then searching for the minimum in the corresponding discretized space. Take the two-dimensional case, for example,

$$\phi_O = \min_{\theta} K(\theta) = K(\widetilde{\theta}) \sim \min_{\theta_j} k(\theta_j), \tag{4.1}$$

where $\theta_j = j \triangle \theta / 2\pi$. Used in a straightforward manner, this kind of approach would require that the grid size $\triangle\theta$ is comparable to the given tolerance. In the following, we briefly describe how the L-BFGS-B method works.

Consider finding a minimum by Newton's method to search for a zero of the gradient of the function $f(\theta) : R^n \to R$. The iteration formula is given by

$$\theta^{K+1} = \theta^k - A^{-1} \cdot \nabla f(\theta),$$

where A is the Hessian matrix of f. The BFGS method is a quasi-Newton method because it doesn't use the actual Hessian matrix of f, but it constructs a sequence of H^k to approximate A^{-1}. The iteration formula for unconstrained optimization is given by

$$\theta^{k+1} = \theta^k - \lambda^k H^k g^k, \quad k = 0, 1, 2, \dots,$$

where λ^k is a step size, g^k is the gradient of f at θ^k, and H^k is updated at every iteration by the following formula:

$$H^{k+1} = (V^k)^T H^k V^k + \rho^k s^k (s^k)^T, \tag{4.2}$$

where

$$\rho^k = 1/(y^k)^T s^k, \qquad V^k = I - \rho^k y^k (s^k)^T,$$

and

$$s^k = \theta^{k+1} - \theta^k, \qquad y^k = g^{k+1} - g^k.$$

The limited memory BFGS method only stores the m most recent pairs $\{s^i, y^i\}_{i=k-m}^{k-1}$ to update H^k. Suppose that the current iteration is θ^k and the initial limited memory matrix $H_{(0)}^k$ (usually a diagonal matrix) is updated by $\{s^i, y^i\}_{i=k-m}^{k-1}$. From (4.2) we have

$$\begin{aligned} H^k = {} & ((V^{k-1})^T \cdots (V^{k-m})^T) H_{(0)}^k (V^{k-m} \cdots V^{k-1}) \\ & + \rho^{k-m} ((V^{k-1})^T \cdots (V^{k-m+1})^T) s^{k-m} (s^{k-m})^T (V^{k-m+1} \cdots V^{k-1}) \\ & + \rho^{k-m+1} ((V^{k-1})^T \cdots (V^{k-m+2})^T) s^{k-m+1} (s^{k-m+1})^T (V^{k-m+2} \cdots V^{k-1}) \\ & + \cdots + \rho^{k-1} s^{k-1} (s^{k-1})^T. \end{aligned} \tag{4.3}$$

For bound constrained problems, the direct Hessian approximation $B^k = (H^k)^{-1}$ is used. The detail derivation and efficient algorithm for computing H^k and B^k are found in [6]. This B^k is used to define a quadratic model of f at θ^k,

$$Q^k(\theta) = f(\theta^k) + (g^k)^T(\theta - \theta^k) + \frac{1}{2}(\theta - \theta^k)^T B^k(\theta - \theta^k).$$

In order to find the minimizer of Q^k subject to the bound constrained, the gradient projection method is first used to determine a set of active bounds. Suppose we have $\Theta = \{\theta \mid l_i \le \theta_i \le u_i,\ i = 1, \dots, n\}$; the ith coordinate of the projection of vector θ is given by

$$P(\theta, l, u)_i = \begin{cases} l_i & \text{if } \theta_i \le l_i, \\ u_i & \text{if } \theta_i \ge u_i, \\ \theta_i & \text{otherwise.} \end{cases}$$

We can then find the generalized Cauchy point that is the first local minimizer θ^c of

$$Q_L^k(t) = Q^k(P(\theta^k - tg^k, l, u)).$$

Use θ^c to identify a set of active variables and then find the minimizer $\overline{\theta}^{k+1}$of the quadratic model with respect to the free variables. Perform a line search

$$\theta^{k+1} = \theta^k + \alpha^k \left(\overline{\theta}^{k+1} - \theta^k\right), \tag{4.4}$$

where α^k is the step size, to find θ^{k+1} that satisfies the sufficient decrease condition

$$f(\theta^{k+1}) \le f(\theta^k) + 10^{-4}(g^k)^T \left(\overline{\theta}^{k+1} - \theta^k\right).$$

For more details, please refer to [5]. In our calculation, we choose $m = 5$ and the stopping criterion is then

$$\|P(\theta^k - g^k, l, u) - \theta^k\|_\infty < 10^{-6}.$$

5. Examples. We implement our new numerical scheme in the following examples. We choose $\delta = 10^{-15}$ for two dimensional cases and $\delta = 10^{-12}$ for three dimensional cases for simplicity. Ideally the δ should be chosen as a small constant times the grid size. We test an anisotropic case with constant coefficients a, b, and c in Figures 1 and 2 to show a very degenerate case with varied coefficients and a box-shape boundary condition. The equation is

$$\sqrt{0.375\phi_x^2 + 0.25\phi_y^2 - 0.58\phi_x\phi_y} = (2.1 - \cos(4\pi^2 xy))/4.$$

Thus $a = 0.375$, $b = 0.25$, $c = 0.29$, and $R(x, y) = (2.1 - \cos(4\pi^2 xy))/4$. Notice that in this case, $ab = 0.0938$ is barely greater than $c^2 = 0.0841$ and R is highly oscillatory. That is why it needs more iterations. In general, we usually need more iterations when the characteristics are very curvy. Figures 3 and 4 show the geodesic distances on manifolds. In Figure 3, there are two boundary points. The contour plot has kinks on the equal distance places. In Figure 4, the boundary point is in the center and on the

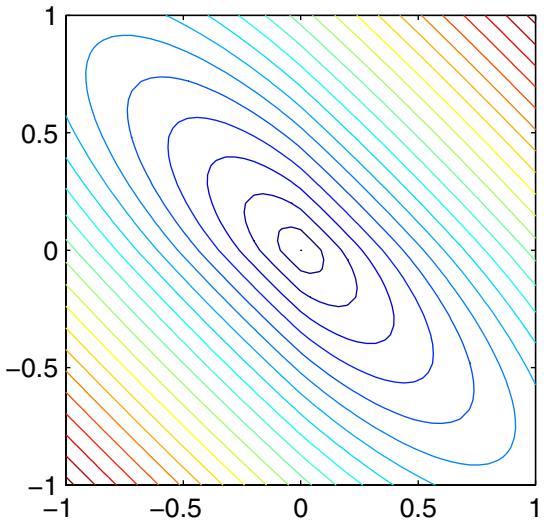


FIG. 1. *A sweeping result after* 2 *sweeping iterations on a* 50×50 *grid. The boundary is a single point in the center.* $a = 1.0$, $b = 1.0$, $c = 0.9$, *and* $R = 1$.

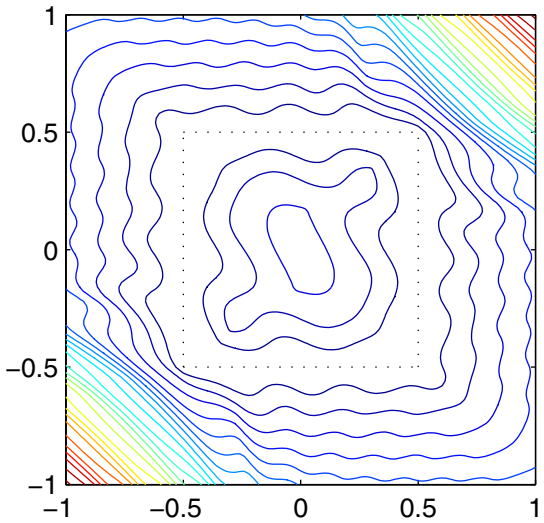


FIG. 2. $a = 0.375$, $b = 0.25$, $c = 0.29$, *and* $R(x, y) = (2.1 - \cos(4\pi^2 xy))/4.0$ *on a* 100×100 *grid. Convergence is reached after* 45 *sweeping iterations.*

top of the mountain-shaped manifold. The contour plot shows the geodesic distance to the boundary point. Figure 5 is an example of the first arrival travel times to seismic imaging. The computational domain suggests material layering under a sinusoidal profile with layer shapes $C(x) = 0.1225 \sin(4\pi x)$. Suppose the domain is split into four parts by $y_i(x) = 0.1225 \sin(4\pi x) + p_i$ where $i = 1, 2, 3,$ and $p_i = (-0.25, 0, 0.25)$. In each layer, the anisotropic speed at (x, y) is given by an ellipse with the long axis (of length $2F_2$) tangential to the curve $C(x)$ and the short axis (of length $2F_1$) normal

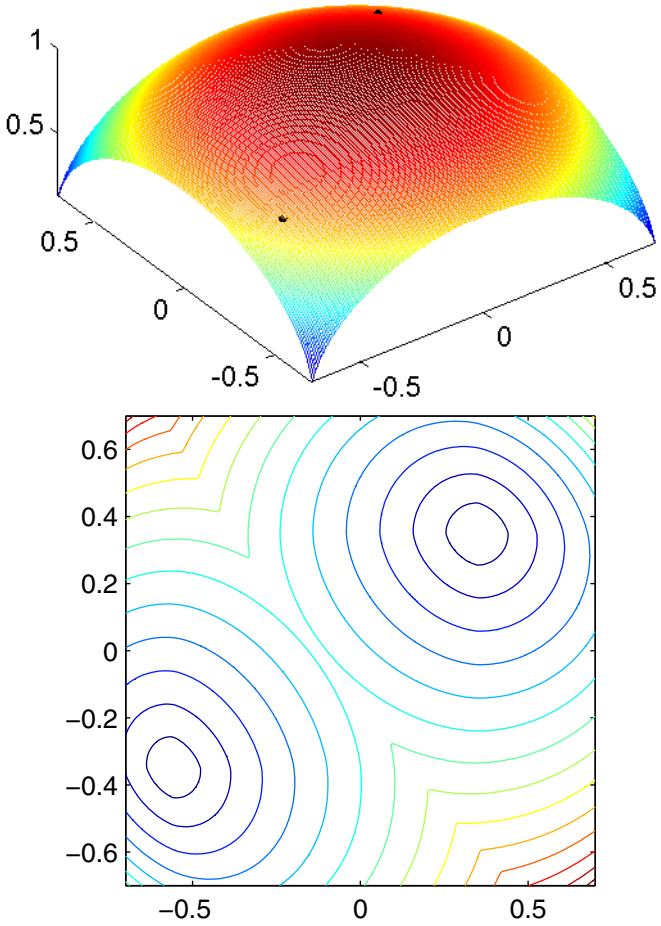


FIG. 3. *This is an example of the distance on a half sphere. The sweeping algorithm was applied to the graph of* $f(x,y) = \sqrt{1.0 - (x^2 + y^2)}$ *with* $\phi(-0.56, -0.35) = \phi(0.35, 0.35) = 0$ *as a boundary condition on a* 200×200 *grid. The convergence was reached after* 3 *sweeping iterations.*

to the curve. F_1 and F_2 are constants in each layer. This leads to

$$F_2\sqrt{((1+n^2)\phi_x^2 + (1+m^2)\phi_y^2 - 2mn\phi_x\phi_y)/(1+m^2+n^2)} = 1,$$

where

$$(m,n) = \frac{\sqrt{(F_2/F_1)^2 - 1}}{\sqrt{1 + \left(\frac{dC(x)}{dx}\right)^2}}\left(\frac{dC(x)}{dx}, -1\right).$$

From the results, we know that the algorithm is stable even with discontinuous coefficients. Figures 6 and 7 are the solutions for three-dimensional eikonal equation with one and two point boundary conditions. Figures 8 and 9 are the more general cases for

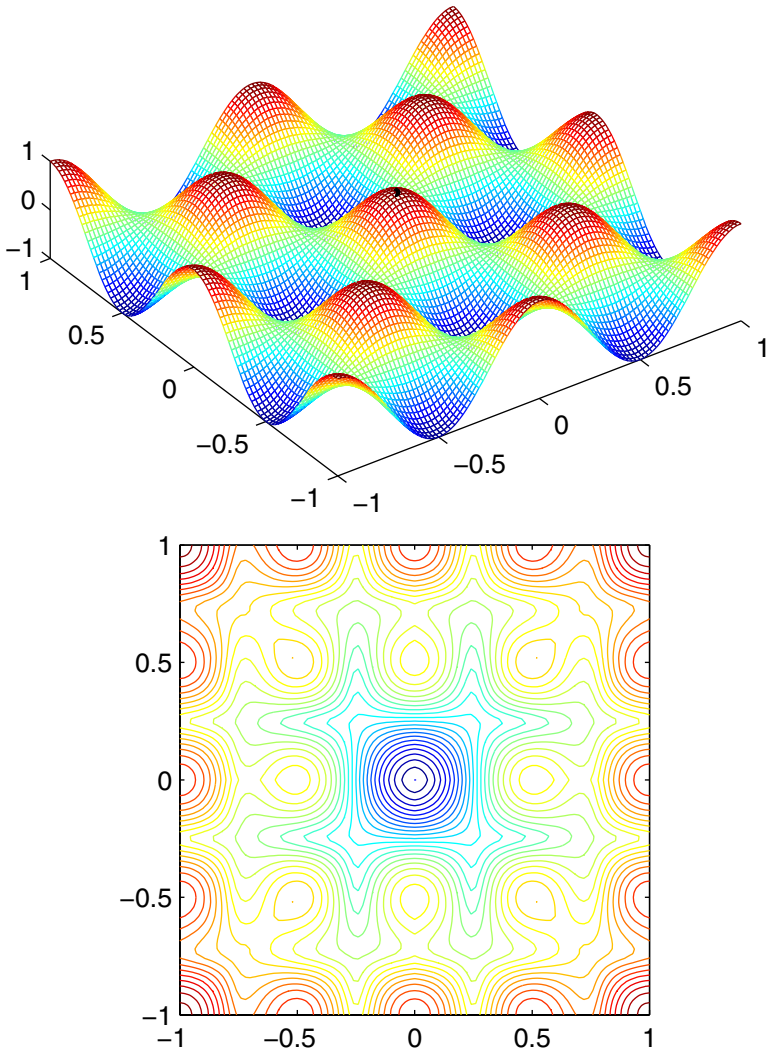


FIG. 4. *The distance contour from* $(0,0)$ *on the graph of* $f(x,y) = \cos(2\pi x)\cos(2\pi y)$. *The convergence was obtained after* 12 *iterations on a* 100×100 *grid.*

three dimensions. Figure 8 has a boundary point $\phi(0,0,0) = 0$ and Figure 9 has a cubic boundary condition with sides of length one. The governing equation we solved is

$$\sqrt{a\phi_x^2 + b\phi_y^2 + c\phi_z^2 - 2d\phi_x\phi_y - 2e\phi_y\phi_z - 2f\phi_z\phi_x} = 1,$$

where

$$a = \frac{1 + f_y^2 + f_z^2}{1 + f_x^2 + f_y^2 + f_z^2}, \quad b = \frac{1 + f_x^2 + f_z^2}{1 + f_x^2 + f_y^2 + f_z^2}, \quad c = \frac{1 + f_x^2 + f_y^2}{1 + f_x^2 + f_y^2 + f_z^2},$$

$$d = \frac{f_x f_y}{1 + f_x^2 + f_y^2 + f_z^2}, \quad e = \frac{f_y f_z}{1 + f_x^2 + f_y^2 + f_z^2}, \quad f = \frac{f_z f_x}{1 + f_x^2 + f_y^2 + f_z^2},$$

and $f(x,y,z) = \cos(2\pi x)\cos(2\pi y)\cos(2\pi z)$, and the corresponding

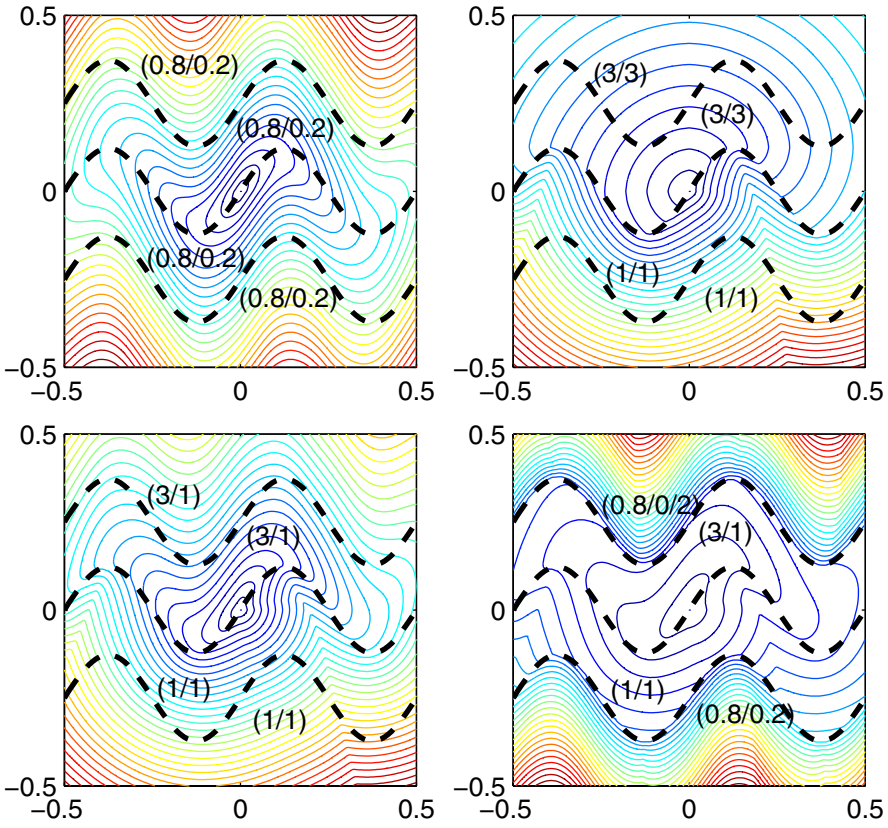


FIG. 5. *This is an example of first arrival travel times in seismic imaging* [26]. *The* (F_2, F_1) *pair for each layer is given in the above figures. The convergence was obtained after* $5, 4, 16,$ *and* 16 *iterations on a* 200×200 *grid.*

$$w(\theta_1, \theta_2) = \sqrt{\frac{1}{1 + (f_x \sin\theta_1 \cos\theta_2 + f_y \sin\theta_1 \sin\theta_2 + f_z \cos\theta_1)^2}}.$$

This seems to be the first successful rapid computation in three dimensions for such problems. In [30], it was proved that the results from the fast sweeping method for the eikonal equation with $R(x) = 1$ need only one iteration, which is exactly 2^n Gauss–Seidel alternating sweepings for the problem in R^n, to reach a solution with global error $O(h \log(1/h))$. We provide the numerical evidence by testing our methods on an eikonal equation with $R(x) = 1$ on two and three dimensions. The results are given in Tables 1 and 2. For anisotropic cases, we found out that the number of iterations depend on the anisotropy of the Hamiltonian, but it is always reasonable and appears to be independent of the grid size.

6. Conclusion. In this paper, we have presented a new numerical method for Hamilton–Jacobi equations written in the form of Bellman's formula. We proved that the numerical Hamiltonian we proposed is monotone and consistent and is in fact also the Godunov Hamiltonian. We implemented this new scheme and showed some results in two- and three-dimensional cases.

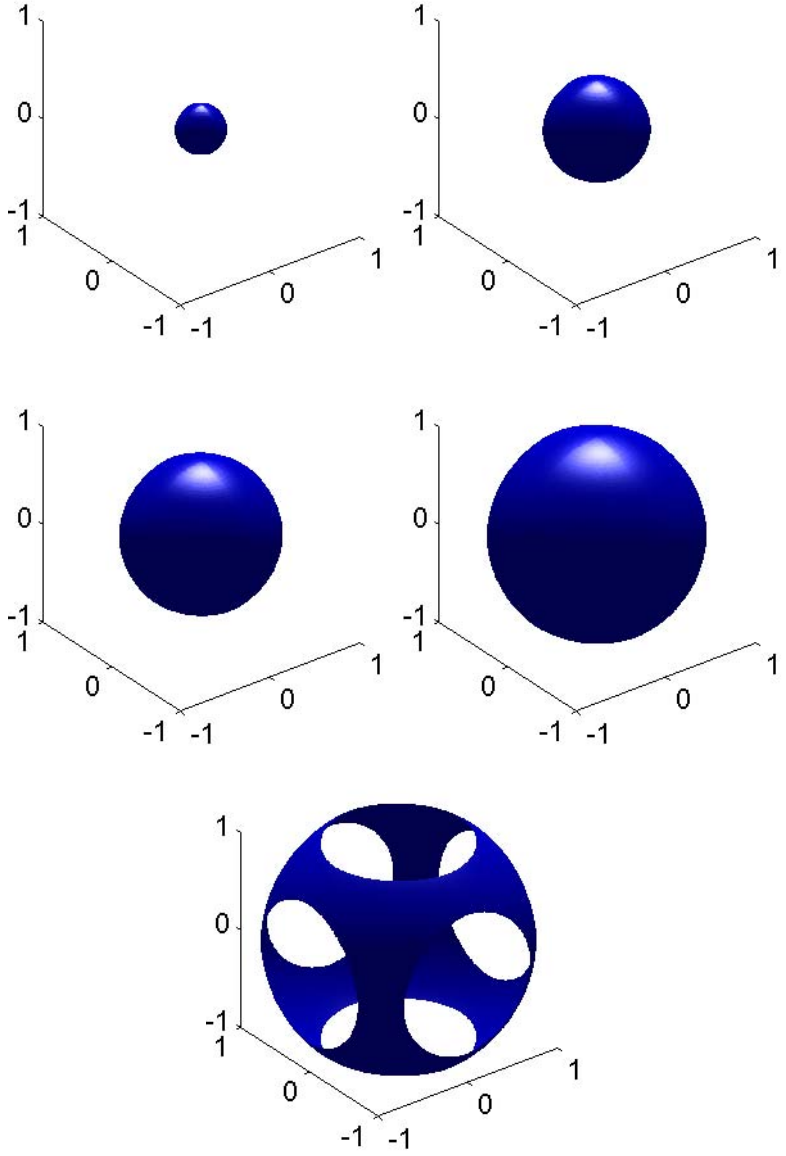


FIG. 6. *This is the one-iteration result of the* $3D$ *eikonal equation with the boundary* $(0, 0)$ *in the center of the graph. The corresponding contours are* 0.25, 0.5, 0.75, 1.0, *and* 1.25.

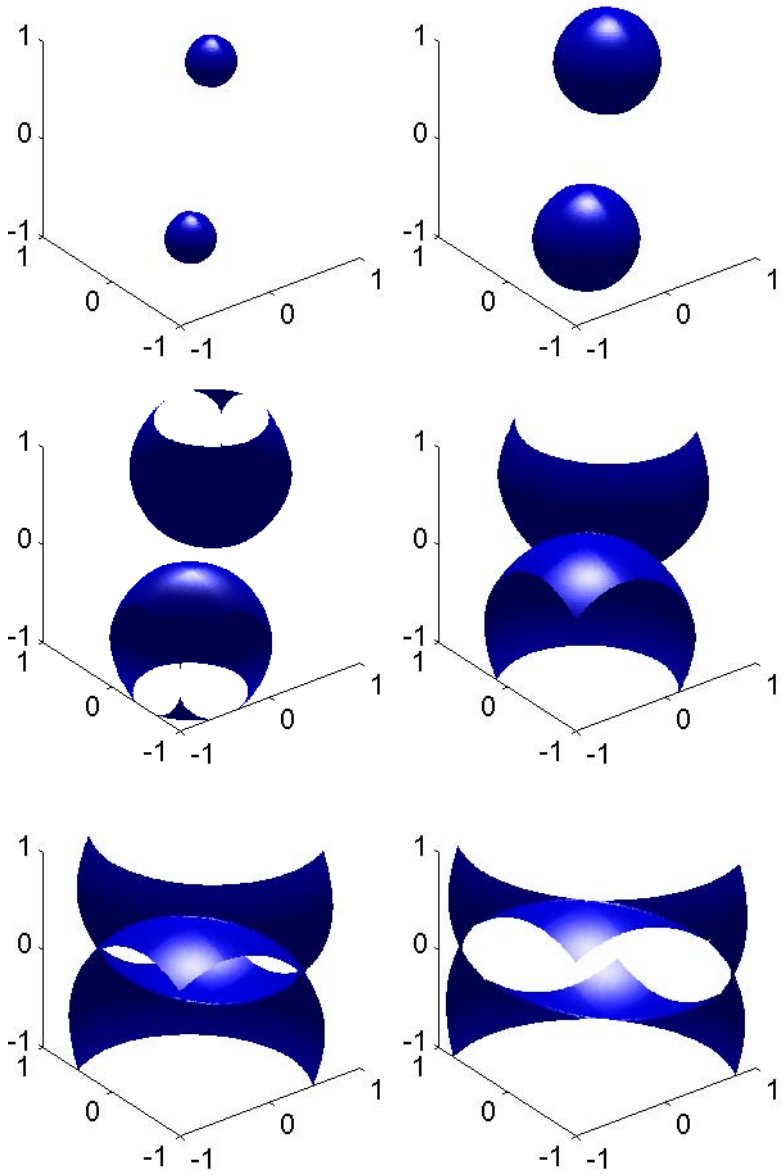


FIG. 7. *This is the one-iteration result of the* 3*D eikonal equation with two boundary points* $(-0.5, -0.5)$ *and* $(0.5, 0.5)$. *The corresponding contours are* 0.25, 0.5, 0.75, 1, 1.25, *and* 1.5.

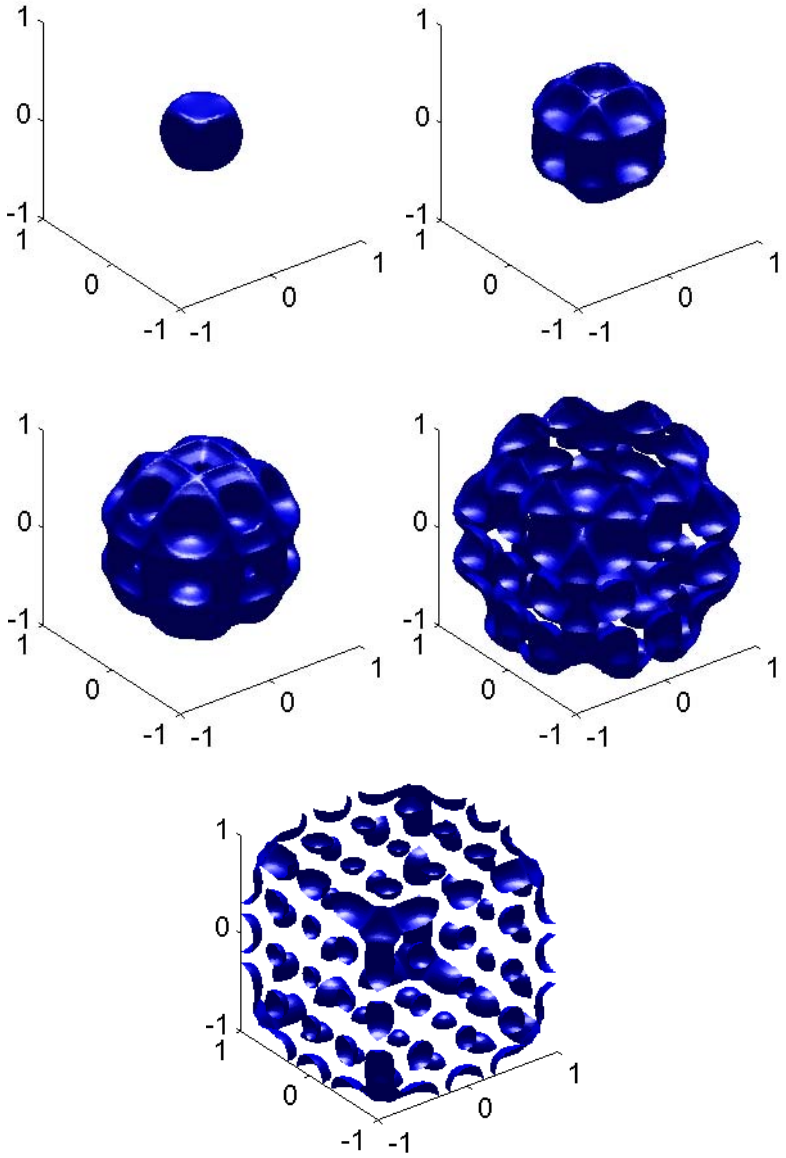


FIG. 8. *This is a* 3*D example with* $f(x, y, z) = \cos(2\pi x)\cos(2\pi y)\cos(2\pi z)$*, the corresponding* $w = (1 + (\bigtriangledown f \cdot \theta)^2)^{-1/2}$*, and a boundary point at the center. The convergence was obtained after* 10 *iterations on a* $100 \times 100 \times 100$ *grid. The contours shown here are* 1.2*,* 1.5*,* 1.8*,* 2.2*, and* 2.5.

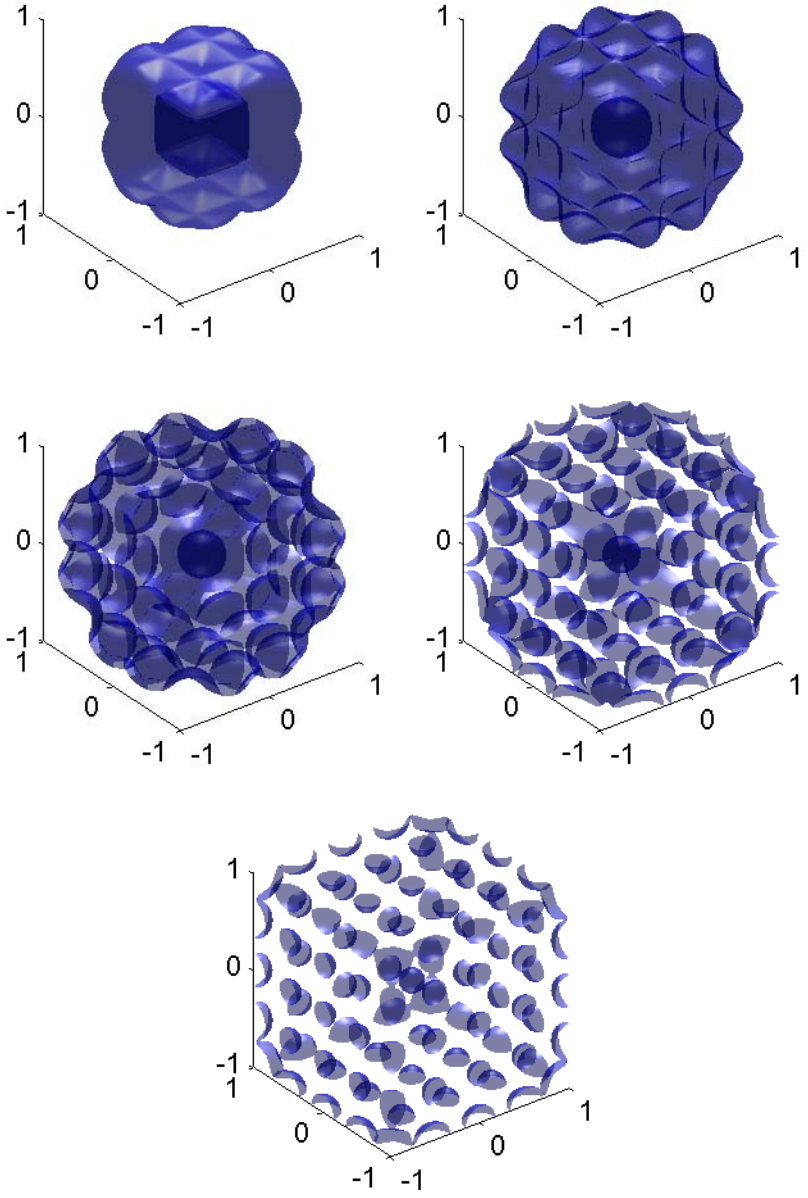


FIG. 9. *This is a* $3D$ *example with* $f(x, y, z) = \cos(2\pi x)\cos(2\pi y)\cos(2\pi z)$*, the corresponding* $w = (1 + (\bigtriangledown f \cdot \theta)^2)^{-1/2}$*, and the cubic boundary condition. The convergence was obtained after* 9 *iterations on a* $100 \times 100 \times 100$ *grid. The contours shown here are* 0.2, 0.4, 0.6, 0.8, *and* 1.0.

TABLE 1
The errors of a 2D eikonal case.

2D eikonal equation $dx =$	2/50	2/100	2/200
L_1 error	0.102158	0.060888	0.0358203
L_∞ error	0.0437414	0.0262969	0.0154506

2/400	2/800	2/1600	1/3200
0.0207759	0.0118848	0.0067128	0.00374894
0.00890583	0.00505242	0.00282877	0.00156648

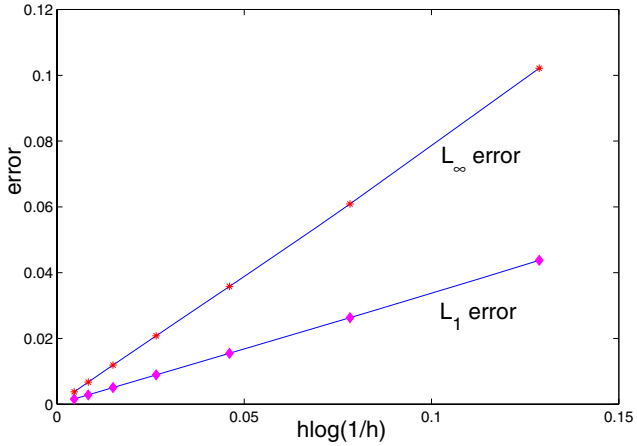
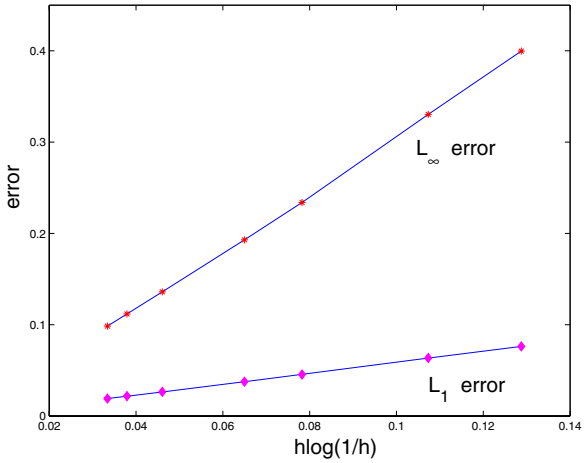


TABLE 2
The errors of a 3D eikonal case.

3D eikonal equation $dx =$	2/50	2/64	2/100
L_1 error	0.399696	0.330305	0.233834
L_∞ error	0.0761747	0.0635267	0.0454065

2/128	2/200	2/256	2/300
0.192961	0.135946	0.111793	0.0985156
0.0375639	0.0264938	0.0217706	0.0191687



REFERENCES

[1] M. Bardi and S.Osher, *The nonconvex multidimentional Riemann problem for Hamilton-Jacobi equations*, SIAM J. Math. Anal., 22 (1991), pp. 344–351.

[2] T. J. Barth, *On the Marchability of Interior Stabilized Discontinuous Galerkin Approximations of the Eikonal and Related Pdes with Non-Divergence Structure*, NASA Technical Report, NAS-01-010, NASA Ames Research Center, Moffett Field, CA, 2001.

[3] M. Boué and P. Dupuis, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic cost in the control*, SIAM J. Numer. Anal., 36 (1999), pp. 667–695.

[4] Y. Brenier, *Un algorithme rapide pour le calcul de transformées de Legendre-Fenchel discretes*, C. R. Acad. Sci. Paris Sér. I Math., 308 (1989), pp. 587–589.

[5] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound constrained optimization*, SIAM J. Sci. Comput., 16 (1995), pp. 1190–1208.

[6] R. H. Byrd, J. Nocedal, and R. B. Schnabel, *Representation of quasi-Newton matrices and their use in limited memory methods*, Math. Programming, 63 (1994), pp. 129–156.

[7] L.-T. Cheng, P. Burchard, B. Merriman, and S. Osher, *Motion of curves constrained on surfaces using a level-set approach*, J. Comput. Phys., 175 (2002), pp. 604–644.

[8] T. F. Coleman and Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445.

[9] L. Corrias, *Fast Legendre-Fenchel transform and applications to Hamilton-Jacobi equations and conservation laws*, SIAM J. Numer. Anal., 33 (1996), pp. 1534–1558.

[10] M. G. Crandall and P. L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[11] M. Falcone and R. Ferretti, *Semi-Lagrangian schemes for Hamilton-Jacobi equations, discrete representation formulae and Godunov methods*, J. Comput. Phys., 175 (2002), pp. 559–575.

[12] A. Friedlander, J. M. Martinez, and S. A. Santos, *A new trust region algorithm for bound constrained minimization*, Appl. Math. Optim, 30 (1994), pp. 235–266.

[13] S. Gray and W. May, *Kirchhoff migration using eikonal equation traveltimes*, Geophysics, 59 (1994), pp. 810–817.

[14] J. Helmsen, E. Puckett, P. Colella, and M. Dorr, *Two new methods for simulating photolithography development in 3d*, in SPIE 2726, 1996, pp. 253–261.

[15] G.-S. Jiang and D. Peng, *Weighted ENO schemes for Hamilton-Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.

[16] R. Kimmel and J. A. Sethian, *Computing geodesic paths on manifolds*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 8431–8435.

[17] F. Memoli and G. Sapiro, *Fast computation of weighted distance functions and geodesics on implicit hyper-surfaces*, J. Comput. Phys., 173 (2001), pp. 730–764.

[18] S. Osher, *A level set formulation for the solution of the Dirichlet problem for Hamilton-Jacobi equations*, SIAM J. Math. Anal., 24 (1993), pp. 1145–1152.

[19] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[20] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

[21] D. Peng, S. Osher, B. Merriman, and H.-K. Zhao, *The geometry of Wulff crystal shapes and its relations with Riemann problems*, in Nonlinear Partial Differential Equations (Evanston, IL, 1998), AMS., Providence, RI, 1999, pp. 251–303.

[22] J. Qian and W. W. Symes, *Paraxial eikonal solvers for anisotropic quasi-p travel times*, J. Comput. Phys., 174 (2001), pp. 256–278.

[23] J. Qian and W. W. Symes, *Finite-difference quasi-p traveltimes for anisotropic media*, Geophysics, 67 (2002), pp. 147–155.

[24] E. Rouy and A. Tourin, *A viscosity solutions approach to shape-from-shading*, SIAM J. Numer. Anal., 29 (1992), pp. 867–884.

[25] J. A. Sethian, *Fast marching level set methods for three dimensional photolithography development*, in SPIE 2726, 1996, pp. 261–272.

[26] J. A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton-Jacobi equations*, Proc. Natl. Acad. Sci. USA, 98 (2001) pp. 11069–11074.

[27] P. E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1–43.

[28] Y.-H. R. Tsai, L.-T. Cheng, S. Osher, and H.-K. Zhao, *Fast sweeping methods for a class of Hamilton-Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 673–694.

[29] J. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.

[30] H.-K. Zhao, *Fast sweeping method for eikonal equations* I, *Distance function*, www.math.uci.edu/~zhao, 2002.

[31] H.-K. Zhao, S. Osher, B. Merriman, and M. Kang, *Implicit and non-parametric shape reconstruction from unorganized points using variational level set method*, Computer Vision and Image Understanding, 80 (2000), pp. 295–314.

[32] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, *L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization*, ACM Trans. Math. Software, 23 (1997), pp. 550–560.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2633–2651

NUMERICAL ANALYSIS FOR A MACROSCOPIC MODEL IN MICROMAGNETICS*

CARSTEN CARSTENSEN† AND DIRK PRAETORIUS‡

Abstract. The macroscopic behavior of stationary micromagnetic phenomena can be modeled by a relaxed version of the Landau–Lifshitz minimization problem. In the limit of large and soft magnets Ω, it is reasonable to exclude the exchange energy and convexify the remaining energy densities. The numerical analysis of the resulting minimization problem,

$$\min E_0^{**}(\mathbf{m}) \text{ amongst } \mathbf{m} : \Omega \to \mathbb{R}^d \text{ with } |\mathbf{m}(x)| \le 1 \text{ for almost every } x \in \Omega,$$

for $d = 2, 3$, faces difficulties caused by the pointwise side-constraint $|\mathbf{m}| \le 1$ and an integral over the whole space $\mathbb{R}^d$ for the stray field energy. This paper involves a penalty method to model the side-constraint and reformulates the exterior Maxwell equation via a nonlocal integral operator $\mathcal{P}$ acting on functions exclusively defined on Ω. The discretization with piecewise constant discrete magnetizations leads to edge-oriented boundary integrals, the implementation of which and related numerical quadrature are discussed, as are adaptive algorithms for automatic mesh-refinement. A priori and a posteriori error estimates provide a thorough rigorous error control of certain quantities. Three classes of numerical experiments study the penalization, empirical convergence rates, and performance of the uniform and adaptive mesh-refining algorithms.

Key words. micromagnetics, microstructure, relaxation, variational problems, nonconvex minimization, degenerate problems, a priori error estimates, adaptive algorithm, a posteriori error estimates, Newton potential, integral operators, panel clustering, hierarchical matrices

AMS subject classifications. 64M07, 65K10, 65N30, 73C50, 73S10, 65N15, 65N30, 65N50

DOI. 10.1137/S003614290343565X

1. Introduction.

Numerical simulations of stationary micromagnetic phenomena are most frequently based on a mathematical model named after Landau and Lifshitz [2, 11]. Therein, one minimizes the energy functional

$$E_\alpha(\mathbf{m}) := \int_\Omega \phi(\mathbf{m})\,dx - \int_\Omega \mathbf{f}\cdot\mathbf{m}\,dx + \frac{1}{2}\int_{\mathbb{R}^d} |\nabla u|^2\,dx + \alpha\int_\Omega |\nabla \mathbf{m}|^2\,dx \tag{1.1}$$

over some admissible vector-valued magnetizations $\mathbf{m} : \Omega \to \mathbb{R}^d$ on the magnet Ω; $\mathbf{m}(x) := 0$ for $x \in \mathbb{R}^d\backslash\overline{\Omega}$. Moreover, $\phi \in \mathcal{C}^\infty(\mathbb{R}^d; \mathbb{R}_{\ge 0})$ denotes the anisotropy density (it models material properties on a crystalline level), $\mathbf{f} \in L^2(\Omega; \mathbb{R}^d)$ denotes an applied exterior magnetic field, $\alpha \ge 0$ is the very small exchange parameter, and u is the magnetic potential related to $\mathbf{m}$ by Maxwell's equation

$$\operatorname{div}(-\nabla u + \mathbf{m}) = 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^d). \tag{1.2}$$

The model description is completed by a nonconvex side-constraint given by the pointwise length condition on the magnetization vector, namely,

$$|\mathbf{m}(x)| = 1 \quad \text{for almost every } x \in \Omega. \tag{1.3}$$

*Received by the editors September 29, 2003; accepted for publication (in revised form) July 29, 2004; published electronically April 19, 2005. This paper was supported by the Austrian Science Fund FWF under grant P15274 and the EPSRC under grant N09176/01. Part of this work was done during a visit to the Isaac-Newton Institute of Mathematical Sciences, Cambridge, England.
http://www.siam.org/journals/sinum/42-6/43565.html

†Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (cc@math.hu-berlin.de).

‡Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria (dirk.praetorius@tuwien.ac.at).

Any of the summands in (1.1) favors another property of an energy-minimizing magnetization. First, uniaxial materials such as cobalt allow the *uniaxial* anisotropy energy

$$\phi(x) = \frac{1}{2}\big(1 - (x \cdot \mathbf{e})^2\big) \quad \text{for all } |x| = 1 \tag{1.4}$$

with given *easy axis* $\mathbf{e} \in \mathbb{R}^d$, a fixed unit vector, which favors magnetizations $\mathbf{m}$ aligned with $\mathbf{e}$. Second, the *exterior energy* favors magnetizations $\mathbf{m}$ aligned to the exterior field $\mathbf{f}$. Third, the *magnetic energy* vanishes for divergence-free magnetizations, as seen in (1.2); notice that (1.2) involves a boundary condition $[\partial u/\partial \mathbf{n}] = -\mathbf{m} \cdot \mathbf{n}$ for the jump $[\cdot]$ on $\partial\Omega$, where $\mathbf{n}$ denotes the outer normal vector on $\partial\Omega$. Fourth, the *exchange energy* penalizes changes in the magnetization $\mathbf{m}$ and so yields Weissian domains and rapidly changes at the Bloch walls between those.

The macroscopic material behavior for large and soft magnets, however, is conserved in the case $\alpha = 0$. Then, the model lacks classical solutions in general [12] and hence has to be relaxed by considering measure-valued solutions [17] or by convexification [6, 21]. Notice that the convexified problem is the mathematical foundation of the so-called *phase theory* [11, p. 184].

Throughout this paper, the focus is on the numerical approximation of macroscopic quantities such as the magnetic potential u or the space-averages of the magnetization vector $\mathbf{m}$. In fact, in a certain limit configuration of soft-large bodies, $\alpha \to 0$, and then E_0 is the correct model with generalized solutions. The well-posed macroscopic values of E_0 are u and $\mathbf{m}$, which minimize the convexified model E_0^{**}. We refer to [6, 21] for justifications of this and the proof of

$$E_0^{**}(\mathbf{m}) := \int_\Omega \phi^{**}(\mathbf{m})\, dx - \int_\Omega \mathbf{f} \cdot \mathbf{m}\, dx + \frac{1}{2}\int_{\mathbb{R}^d} |\nabla u|^2\, dx \tag{1.5}$$

with the side-constraint (1.2) and

$$|\mathbf{m}(x)| \le 1 \quad \text{for almost every } x \in \Omega. \tag{1.6}$$

Here, ϕ^{**} is the convexified density defined by

$$\phi^{**}(x) = \sup\big\{\varphi(x) \,\big|\, \varphi : \mathbb{R}^d \to \mathbb{R} \text{ convex and } \varphi|_{\mathbb{S}} \le \phi\big\} \quad \text{for } |x| \le 1,$$

where $\mathbb{S} = \big\{x \in \mathbb{R}^d \,\big|\, |x| = 1\big\}$ denotes the unit sphere. Then the relaxed problem (RP) reads as follows:

$$\text{Minimize } E_0^{**}(\mathbf{m}) \text{ over } \mathcal{A} := \big\{\mathbf{m} \in L^\infty(\Omega;\mathbb{R}^d) \,\big|\, \|\mathbf{m}\|_{L^\infty} \le 1\big\}.$$

In contrast to the ill-posed problem E_0, its convexification is well-posed. In particular, the minimum of $E_0^{**}(\mathcal{A})$ is attained in $\mathcal{A}$.

The numerical analysis of the model in [5, 20] considers $d = 2$ only, replaces the entire space $\mathbb{R}^d$ in (1.2) by a bounded Lipschitz domain $\widehat{\Omega}$ containing Ω, and solves for $u \in H_0^1(\widehat{\Omega})$. The potential u is discretized by a nonconforming u_h and a piecewise constant $\mathbf{m}_h$ on Ω. The choice of u_h as a conforming, piecewise affine, and globally continuous finite element scheme leads to instabilities [5, 20]. In this paper, we treat (1.2) exactly via an integral representation, i.e., $u = \mathcal{L}\mathbf{m}$ with a linear convolution operator $\mathcal{L}$ and $\mathcal{P}\mathbf{m} := \nabla(\mathcal{L}\mathbf{m})$; cf. Theorem 2.1. The algorithmic realization of

$\mathcal{P}$ is less obvious and discussed in subsection 4.1. The advantage is that the resulting model requires only one discretization, e.g., by piecewise constant approximations $\mathbf{m}_h$. Those allow exact fulfillment of the side-constraint $|\mathbf{m}_h| \le 1$ involved approximately by a penalization procedure. The resulting discrete minimization problem is to minimize

$$E^{**}_{\varepsilon,h}(\mathbf{m}_h) := \int_\Omega \phi^{**}(\mathbf{m}_h)\,dx - \int_\Omega \mathbf{f}\cdot\mathbf{m}_h\,dx + \frac{1}{2}\int_{\mathbb{R}^d} |\mathcal{P}\mathbf{m}_h|^2\,dx + \frac{1}{2}\int_\Omega \frac{1}{\varepsilon}(|\mathbf{m}_h|-1)_+^2\,dx \tag{1.7}$$

over all $\mathcal{T}$-piecewise constant magnetizations $\mathbf{m}_h \in \mathcal{L}^0(\mathcal{T})^d$, where $\mathcal{T}$ is a partition of Ω. According to the a priori error analysis, the $\mathcal{T}$-piecewise constant penalization function $\varepsilon \in \mathcal{L}^0(\mathcal{T};\mathbb{R}_{>0})$ will be a power of the local mesh-size later on. It turns out that the error analysis of [5] essentially carries over to the situation presented in section 3 and generalizes to $d = 2, 3$.

The remaining part of the paper is organized as follows: Section 2 states the Euler–Lagrange equations related to (RP) and gives an alternate proof of the uniqueness of the solution of (RP) in the uniaxial case. The discrete problem $(RP_{\varepsilon,h})$ is formulated and unique existence of discrete solutions is discussed. Section 3 presents the assertion and proofs of a priori and a posteriori error estimates. Section 4 displays details on a possible implementation: the computation of a discrete solution by a Newton–Raphson scheme (subsection 4.1), an indicator-based adaptive mesh-refinement (subsection 4.2), the implementation of the proposed refinement indicators (subsection 4.3), and the efficient realization of the involved integral operator $\mathcal{P}$ by an $\mathcal{H}$-matrix approach (subsection 4.4). Sections 5–7 report on the results of careful numerical studies. The first and second examples provide a closed formula for the smooth and nonsmooth exact solution with a computable error $\|\mathbf{m} - \mathbf{m}_h\|_{L^2(\Omega)}$. Empirical evidence supports the choice of the penalty parameter $\varepsilon = h^{3/2}$ and the superiority of adaptive mesh-refining strategies over uniform meshes. The real-life scientific computing in section 7 with unknown solution shows, very much in surprising contrast to [5], that almost no local mesh-refinement is required.

2. Preliminaries. This section is devoted to the Euler–Lagrange equations related to (RP) which characterize the minimizers and introduces the proposed discretization by a penalization strategy. For (RP) and the discrete problem $(RP_{\varepsilon,h})$ we prove unique existence of solutions in the uniaxial case.

The magnetic potential is modeled via a Newton integral representation as in [14, 13]. The subsequent theorem gathers the required properties of the respective integral operator. Proofs can be found in [19] although we expect that the result is known to the experts. The Newtonian kernel $G : \mathbb{R}^d\backslash\{0\} \to \mathbb{R}$ is defined by

$$G(x) := \begin{cases} \frac{1}{\gamma_d}\log|x| & \text{for } d = 2 \\ \frac{1}{(2-d)\gamma_d}|x|^{2-d} & \text{for } d > 2 \end{cases} \quad \text{for } x \neq 0, \tag{2.1}$$

where the constant $\gamma_d := |\mathbb{S}| > 0$ denotes the surface measure of the unit sphere.

THEOREM 2.1. *Given any $\mathbf{m} \in L^\infty(\Omega;\mathbb{R}^d)$, there exists (up to an additive constant) a unique magnetic potential $u = \mathcal{L}\mathbf{m} \in H^1_{\ell oc}(\mathbb{R}^d)$ such that*

$$\nabla u \in L^2(\mathbb{R}^d;\mathbb{R}^d) \quad \textit{and} \quad \operatorname{div}(-\nabla u + \mathbf{m}) = 0 \quad \textit{in } \mathcal{D}'(\mathbb{R}^d). \tag{2.2}$$

The (extended) operator $\mathcal{P} : L^2(\mathbb{R}^d;\mathbb{R}^d) \to L^2(\mathbb{R}^d;\mathbb{R}^d), \mathbf{m} \mapsto \nabla(\mathcal{L}\mathbf{m})$ *is an* L^2 *orthogonal projection. The potential* $\mathcal{L}\mathbf{m}$ *can be represented as a convolution operator*

$$\mathcal{L}\mathbf{m} = \sum_{j=1}^{d} \frac{\partial G}{\partial x_j} * \mathbf{m}_j, \tag{2.3}$$

where $\mathbf{m} = (\mathbf{m}_1, \dots, \mathbf{m}_d)$ *is trivially extended (by zero) from* Ω *to* $\mathbb{R}^d$ *(so that the convolution is formally well-defined).*

Remark 2.1. For $d = 3$ it can be shown that the convolution $\mathcal{L}\mathbf{m}$ from (2.3) already is in $H^1(\mathbb{R}^d)$. Further details on the case $\mathbf{m} \in L^p(\mathbb{R}^d;\mathbb{R}^d)$ for $1 < p < \infty$ are found in [19].

Since the energy functional E_0^{**} from (1.5) is convex and (Gâteaux-)differentiable, the minima are equivalently characterized by the corresponding Euler–Lagrange equations [6]. Thus, problem (RP) reads as follows: Find $(\lambda, \mathbf{m}) \in L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d)$ such that

$$\mathcal{P}\mathbf{m} + D\phi^{**}(\mathbf{m}) + \lambda\mathbf{m} = \mathbf{f} \quad \text{a.e. in } \Omega, \tag{2.4}$$

$$\lambda \geq 0,\ |\mathbf{m}| \leq 1,\ \lambda(1 - |\mathbf{m}|) = 0 \quad \text{a.e. in } \Omega. \tag{2.5}$$

Remark 2.2. For the uniaxial model case (1.4), direct calculations show $\phi^{**}(x) = \frac{1}{2}\sum_{j=2}^{d}(x \cdot \mathbf{z}_j)^2$, where $\mathbf{e} \in \mathbb{R}^d$ is the easy axis and $\{\mathbf{e}, \mathbf{z}_2, \dots, \mathbf{z}_d\}$ is an orthonormal basis of $\mathbb{R}^d$. Thus, $D\phi^{**}(x) = \sum_{j=1}^{d}(x \cdot \mathbf{z}_j)\mathbf{z}_j$.

THEOREM 2.2. *Problem* (RP) *has at least one solution* $(\lambda, \mathbf{m})$. *For any two solutions* $(\lambda_1, \mathbf{m}_1), (\lambda_2, \mathbf{m}_2)$ *of* (RP), *the magnetic potentials coincide (modulo an additive constant),* $\mathcal{L}\mathbf{m}_1 = \mathcal{L}\mathbf{m}_2$. *In the uniaxial model case* (1.4) *the solution is unique, i.e.,* $(\lambda_1, \mathbf{m}_1) = (\lambda_2, \mathbf{m}_2)$ *a.e.*

Proof. Existence of solutions of (RP) is obtained by the direct method of the calculus of variations. For any solutions $(\lambda_j, \mathbf{m}_j)$ of (RP) and $\boldsymbol{\delta} = \mathbf{m}_2 - \mathbf{m}_1$, (2.4) yields

$$\langle \mathcal{P}\boldsymbol{\delta}\,;\,\boldsymbol{\delta}\rangle_{L^2(\Omega)} + \langle D\phi^{**}(\mathbf{m}_2) - D\phi^{**}(\mathbf{m}_1)\,;\,\boldsymbol{\delta}\rangle_{L^2(\Omega)} + \langle \lambda_2\mathbf{m}_2 - \lambda_1\mathbf{m}_1\,;\,\boldsymbol{\delta}\rangle_{L^2(\Omega)} = 0. \tag{2.6}$$

By orthogonality of $\mathcal{P}$, we have $\langle \mathcal{P}\boldsymbol{\delta}\,;\,\boldsymbol{\delta}\rangle_{L^2(\Omega)} = \|\mathcal{P}\boldsymbol{\delta}\|_{L^2(\Omega)} \geq 0$. Further, convexity yields that the second term in (2.6) is nonnegative. Direct calculation shows the same for the last term [5, Proof of Theorem 2.1]. Thus, all three terms vanish. Hence, $\mathcal{P}\boldsymbol{\delta} = 0$; i.e., the potentials coincide and moreover $\boldsymbol{\delta}$ is (weakly) divergence free in $\mathbb{R}^d$ by definition of $\mathcal{P}$ (and $\mathcal{L}$); see (2.2) in Theorem 2.1.

In the model case we may assume that the easy axis $\mathbf{e} = \mathbf{e}_1$ is the first standard unit vector. The vanishing second term in (2.6) shows that $\boldsymbol{\delta}$ vanishes in all but the $\mathbf{e}_1$ direction. Now we use a standard mollification argument: For any test function $\psi \in \mathcal{D}(\mathbb{R}^d)$, we have $\psi * \boldsymbol{\delta} \in \mathcal{D}(\mathbb{R}^d)$ with $0 = \psi * (\mathrm{div}\boldsymbol{\delta}) = \mathrm{div}(\psi * \boldsymbol{\delta}) = \partial(\psi * \boldsymbol{\delta})/\partial x_1$. Hence $\psi * \boldsymbol{\delta}$ is constant in the $\mathbf{e}_1$ direction and must therefore vanish. This shows $\boldsymbol{\delta} = 0$. From (2.4) and (2.5) we infer that λ_j is uniquely determined by $(\mathbf{m}_j, \mathbf{f})$. Therefore uniqueness of $\mathbf{m}_j$ implies uniqueness of λ_j. ∎

Let $\mathcal{T} = \{T_1, \dots, T_N\}$ be a finite family of pairwise disjoint nonempty open sets T_j which satisfy $\overline{\Omega} = \bigcup_{j=1}^{N} \overline{T_j}$. The space of all $\mathcal{T}$-piecewise constant functions is denoted by $\mathcal{L}^0(\mathcal{T})$, and $h \in \mathcal{L}^0(\mathcal{T})$ is the mesh-size function, $h|_T := h_T := \mathrm{diam}(T)$. For $f \in L^2(\Omega)$, let $f_{\mathcal{T}} \in \mathcal{L}^0(\mathcal{T})$ be the $\mathcal{T}$-piecewise integral mean given by

$$f_{\mathcal{T}}|_T := \frac{1}{|T|}\int_T f\,dx \quad \text{for all } T \in \mathcal{T}.$$

The map $(\cdot)_{\mathcal{T}} : L^2(\Omega) \to \mathcal{L}^0(\mathcal{T}), f \mapsto f_{\mathcal{T}}$ is the L^2 orthogonal projection.

The discrete problem $(RP_{\varepsilon,h})$ reads as follows: Given a penalization parameter $\varepsilon \in \mathcal{L}^0(\mathcal{T})$ with $\varepsilon > 0$, find $\mathbf{m}_h \in \mathcal{L}^0(\mathcal{T})^d$ such that

$$\langle \mathcal{P}\mathbf{m}_h + D\phi^{**}(\mathbf{m}_h) + \lambda_h \mathbf{m}_h \,;\, \boldsymbol{\nu}_h \rangle_{L^2(\Omega)} = \langle \mathbf{f} \,;\, \boldsymbol{\nu}_h \rangle_{L^2(\Omega)} \quad \text{for all } \boldsymbol{\nu}_h \in \mathcal{L}^0(\mathcal{T})^d, \tag{2.7}$$

where $\lambda_h \in \mathcal{L}^0(\mathcal{T})$ is defined by

$$\lambda_h = \varepsilon^{-1} \frac{(|\mathbf{m}_h| - 1)_+}{|\mathbf{m}_h|} \quad \text{with } (\cdot)_+ := \max\{\cdot, 0\}. \tag{2.8}$$

Remark 2.3. Problem $(RP_{\varepsilon,h})$ are the Euler–Lagrange equations of the minimization problem related to the convex functional from (1.7) and the finite-dimensional space $\mathcal{A} = \mathcal{L}^0(\mathcal{T})^d$. Thus, the existence of solutions of the discrete problem $(RP_{\varepsilon,h})$ follows by the direct method of the calculus of variations.

THEOREM 2.3. *The discrete problem $(RP_{\varepsilon,h})$ has at least one solution. For any two solutions $(\lambda_1, \mathbf{m}_1), (\lambda_2, \mathbf{m}_2)$ of $(RP_{\varepsilon,h})$, the magnetic potentials $\mathcal{L}\mathbf{m}_j$ coincide. In the uniaxial model case* (1.4)*, we have uniqueness of the discrete solution, i.e., $(\lambda_1, \mathbf{m}_1) = (\lambda_2, \mathbf{m}_2)$.*

Proof. The same proof as for Theorem 2.2 applies for the discrete setting as well. ∎

Remark 2.4. (a) If the elements $T \in \mathcal{T}$ are rectangular, it can easily be shown that (independent of ϕ^{**}) the solution $(\lambda_h, \mathbf{m}_h)$ of $(RP_{\varepsilon,h})$ is unique. The easy proof just needs that for two discrete solutions $(\lambda_1, \mathbf{m}_1), (\lambda_2, \mathbf{m}_2)$ the difference $\boldsymbol{\delta} := \mathbf{m}_2 - \mathbf{m}_1 \in \mathcal{L}^0(\mathcal{T})^d$ is (weakly) divergence free in $\mathbb{R}^d$. Consider the set $\mathcal{T}^* := \{T \in \mathcal{T} \,|\, \boldsymbol{\delta}|_T \neq 0\}$ and $\Omega^* := \bigcup \{\overline{T} \,|\, T \in \mathcal{T}^*\}$ and argue by contradiction: If $\mathcal{T}^*$ is not empty, we have $\boldsymbol{\delta} \cdot \mathbf{n} = 0$ a.e. by the Gauss divergence theorem, where $\mathbf{n}$ is the outer normal vector on $\partial\Omega^*$. Using that $\boldsymbol{\delta}$ is $\mathcal{T}^*$-piecewise constant, the contradiction easily follows.

(b) The preceding argument applies to more general (but not all) triangulations $\mathcal{T}$.

3. A priori and a posteriori error control. This section provides an a priori and a posteriori error analysis for the proposed discrete scheme $(RP_{\varepsilon,h})$ with or without a further monotonicity assumption (3.4) on ϕ^{**} valid in the uniaxial case (1.4).

THEOREM 3.1. *Let $(\lambda, \mathbf{m})$ and $(\lambda_h, \mathbf{m}_h)$ solve (RP) and $(RP_{\varepsilon,h})$, respectively. Then*

$$\begin{aligned} &\|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|^2_{L^2(\mathbb{R}^d)} + 2\langle D\phi^{**}(\mathbf{m}) - D\phi^{**}(\mathbf{m}_h) \,;\, \mathbf{m} - \mathbf{m}_h \rangle_{L^2(\Omega)} + \|\sqrt{\varepsilon}\,\lambda_h \mathbf{m}_h\|^2_{L^2(\Omega)} \\ &\quad \le 3\|\mathbf{m} - \mathbf{m}_{\mathcal{T}}\|^2_{L^2(\Omega)} + \|D\phi^{**}(\mathbf{m}) - (D\phi^{**}(\mathbf{m}))_{\mathcal{T}}\|^2_{L^2(\Omega)} + \|\lambda\mathbf{m} - (\lambda\mathbf{m})_{\mathcal{T}}\|^2_{L^2(\Omega)} \\ &\qquad + \|\sqrt{\varepsilon}\,\lambda\mathbf{m}\|^2_{L^2(\Omega)}. \end{aligned} \tag{3.1}$$

(Note that according to convexity, the second term on the left-hand side is also nonnegative.)

Proof. To abbreviate notation, define $\mathbf{d} := D\phi^{**}(\mathbf{m})$ and $\mathbf{d}_h := D\phi^{**}(\mathbf{m}_h)$. Using the orthogonal projection $\mathcal{P}$ and the Cauchy inequality, we infer

$$\begin{aligned} \|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|^2_{L^2(\mathbb{R}^d)} \le{}& \frac{1}{2}\|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|^2_{L^2(\mathbb{R}^d)} + \frac{1}{2}\|\mathbf{m} - \mathbf{m}_{\mathcal{T}}\|^2_{L^2(\Omega)} \\ &+ \langle \mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h \,;\, \mathbf{m}_{\mathcal{T}} - \mathbf{m}_h \rangle_{L^2(\Omega)}. \end{aligned} \tag{3.2}$$

According to the Galerkin orthogonality

$$\langle \mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h + \mathbf{d} - \mathbf{d}_h + \lambda\mathbf{m} - \lambda_h\mathbf{m}_h \,;\, \boldsymbol{\nu}_h\rangle_{L^2(\Omega)} = 0 \quad \text{for all } \boldsymbol{\nu}_h \in \mathcal{L}^0(\mathcal{T})^d, \tag{3.3}$$

the last term in (3.2) may be written as

$$\begin{aligned}
&\langle \mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h \,;\, \mathbf{m}_\mathcal{T} - \mathbf{m}_h\rangle_{L^2(\Omega)} \\
&\quad = -\langle \mathbf{d} - \mathbf{d}_h \,;\, \mathbf{m} - \mathbf{m}_h\rangle_{L^2(\Omega)} - \langle \lambda\mathbf{m} - \lambda_h\mathbf{m}_h \,;\, \mathbf{m} - \mathbf{m}_h\rangle_{L^2(\Omega)} \\
&\qquad + \langle \mathbf{d} - \mathbf{d}_h \,;\, \mathbf{m} - \mathbf{m}_\mathcal{T}\rangle_{L^2(\Omega)} + \langle \lambda\mathbf{m} - \lambda_h\mathbf{m}_h \,;\, \mathbf{m} - \mathbf{m}_\mathcal{T}\rangle_{L^2(\Omega)}.
\end{aligned}$$

Since $(\cdot)_\mathcal{T}$ is an orthogonal projection, $\mathbf{d}_h$ and $\lambda_h\mathbf{m}_h$ may be replaced by $\mathbf{d}_\mathcal{T}$ and $(\lambda\mathbf{m})_\mathcal{T}$ in the third and fourth terms. Pointwise evaluation [5, Proof of Theorem 4.3] shows

$$-\langle \lambda\mathbf{m} - \lambda_h\mathbf{m}_h \,;\, \mathbf{m} - \mathbf{m}_h\rangle_{L^2(\Omega)} \le \frac{1}{2}\|\sqrt{\varepsilon}\,\lambda\mathbf{m}\|^2_{L^2(\Omega)} - \frac{1}{2}\|\sqrt{\varepsilon}\,\lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)}.$$

Combining the last two results with two Cauchy inequalities, we conclude (3.1). ∎

THEOREM 3.2. *Let $(\lambda, \mathbf{m})$ and $(\lambda_h, \mathbf{m}_h)$ solve (RP) and $(RP_{\varepsilon,h})$, respectively, and assume that there is a constant $c_1 > 0$ such that, for all $\mathbf{m}_1, \mathbf{m}_2 \in L^2(\Omega;\mathbb{R}^d)$, the following holds:*

$$c_1\|D\phi^{**}(\mathbf{m}_1) - D\phi^{**}(\mathbf{m}_2)\|^2_{L^2(\Omega)} \le \langle D\phi^{**}(\mathbf{m}_1) - D\phi^{**}(\mathbf{m}_2) \,;\, \mathbf{m}_1 - \mathbf{m}_2\rangle_{L^2(\Omega)}. \tag{3.4}$$

Then there is a constant $c_2 > 0$ which depends only on c_1 such that

$$\begin{aligned}
&\|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|^2_{L^2(\mathbb{R}^d)} + \|D\phi^{**}(\mathbf{m}) - D\phi^{**}(\mathbf{m}_h)\|^2_{L^2(\Omega)} + \|\lambda\mathbf{m} - \lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)} \\
&\quad \le c_2\Big((1 + \|\varepsilon\|_{L^\infty(\Omega)})\Big\{\|\mathbf{m} - \mathbf{m}_\mathcal{T}\|^2_{L^2(\Omega)} + \|D\phi^{**}(\mathbf{m}) - (D\phi^{**}(\mathbf{m}))_\mathcal{T}\|^2_{L^2(\Omega)} \\
&\qquad + \|\lambda\mathbf{m} - (\lambda\mathbf{m})_\mathcal{T}\|^2_{L^2(\Omega)}\Big\} + \|\varepsilon\|_{L^\infty(\Omega)}\|\sqrt{\varepsilon}\,\lambda\mathbf{m}\|^2_{L^2(\Omega)}\Big).
\end{aligned} \tag{3.5}$$

Proof. Use notation from the proof of Theorem 3.1. Direct calculation with Galerkin orthogonality, orthogonal projections $(\cdot)_\mathcal{T}$ and $\mathcal{P}$, and simple use of the Cauchy inequality shows

$$\begin{aligned}
&\|\lambda\mathbf{m} - \lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)} \\
&\qquad \le 4\big(\|\lambda\mathbf{m} - (\lambda\mathbf{m})_\mathcal{T}\|^2_{L^2(\Omega)} + \|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|^2_{L^2(\mathbb{R}^d)} + \|\mathbf{d} - \mathbf{d}_h\|^2_{L^2(\Omega)}\big),
\end{aligned} \tag{3.6}$$

whence the left-hand side of (3.5) $\le 5\times$ the right-hand side of (3.6). Assumption (3.4) allows us to dominate the last two terms by Theorem 3.1, which leads to $c_3\big(\|\sqrt{\varepsilon}\,\lambda\mathbf{m}\|^2_{L^2(\Omega)} - \|\sqrt{\varepsilon}\,\lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)}\big)$ on the right-hand side with $c_3 := 5\max\{1, c_1^{-1}\}$. Elementary calculations for scalars $a, b, c \in \mathbb{R}$ show $c\,(a^2 - b^2) = c\,(a+b)(a-b) \le \sqrt{2}\,c\,(a^2+b^2)^{1/2}|a-b| \le c^2\,(a^2+b^2) + |a-b|^2/2$, whence

$$\begin{aligned}
&c_3\big(\|\sqrt{\varepsilon}\,\lambda\mathbf{m}\|^2_{L^2(\Omega)} - \|\sqrt{\varepsilon}\,\lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)}\big) \\
&\qquad \le c_3^2\big(\|\varepsilon\lambda\mathbf{m}\|^2_{L^2(\Omega)} + \|\varepsilon\lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)}\big) + \frac{1}{2}\|\lambda\mathbf{m} - \lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)}
\end{aligned}$$

by pointwise application and integration over Ω. Finally, the second term is dominated by $\|\varepsilon\,\lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)} \le \|\varepsilon\|_{L^\infty(\Omega)}\|\sqrt{\varepsilon}\,\lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)}$ and a second application of Theorem 3.1. ∎

Remark 3.1. Theorem 3.2 applies in particular to the uniaxial case, where we have equality with $c_1 = 1$ in the monotonicity assumption (3.4).

Remark 3.2. Assume the monotonicity assumption (3.4) and that the exact solution is sufficiently smooth. Whereas Theorem 3.1 leads to an estimate of order $O(\varepsilon^{1/2} + h)$ for the error $\|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|_{L^2(\mathbb{R}^d)} + \|D\phi^{**}(\mathbf{m}) - D\phi^{**}(\mathbf{m}_h)\|_{L^2(\Omega)}$, Theorem 3.2 leads to $O(\varepsilon+h)$. This favors the choice $\varepsilon = h$ for the penalization parameter.

Remark 3.3. Note that the full L^2 convergence of $\mathbf{m}_h$ towards $\mathbf{m}$ could not be proven, although it is observed in the numerical experiments. For the uniaxial case, Theorem 3.2 yields the L^2 convergence in all directions orthogonal to the easy axis; cf. Remark 2.2.

THEOREM 3.3. *Let* $(\lambda, \mathbf{m})$ *and* $(\lambda_h, \mathbf{m}_h)$ *solve* (RP) *and* $(RP_{\varepsilon,h})$*, respectively, and assume monotonicity as in* (3.4)*. Then*

(3.7)
$$\begin{aligned}
&\|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|^2_{L^2(\mathbb{R}^d)} + c_1\|D\phi^{**}(\mathbf{m}) - D\phi^{**}(\mathbf{m}_h)\|^2_{L^2(\Omega)}\\
&\quad\le (1 + 1/c_1)\,\|\varepsilon\,\lambda_h\mathbf{m}_h\|^2_{L^2(\Omega)} + 2\|\varepsilon|\lambda_h\mathbf{m}_h|\{(\mathbf{f} - \mathbf{f}_\mathcal{T}) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_\mathcal{T})\}\|_{L^1(\Omega)}\\
&\qquad + 2\langle(\mathbf{f} - \mathbf{f}_\mathcal{T}) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_\mathcal{T})\,;\,\mathbf{m} - \mathbf{m}_\mathcal{T}\rangle_{L^2(\Omega)}.
\end{aligned}$$

Remark 3.4. (a) In fact, the last term on the right-hand side of (3.7) is *not* an a posteriori term but can always be dominated by an application of the Hölder inequality and (1.6)

$$\begin{aligned}
&\langle(\mathbf{f} - \mathbf{f}_\mathcal{T}) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_\mathcal{T})\,;\,\mathbf{m} - \mathbf{m}_\mathcal{T}\rangle_{L^2(\Omega)}\\
&\quad\le 2\|(\mathbf{f} - \mathbf{f}_\mathcal{T}) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_\mathcal{T})\|_{L^1(\Omega)},
\end{aligned}$$

where we used the side-constraint $\|\mathbf{m}\|_{L^\infty(\Omega)} \le 1$.

(b) For $\mathbf{m} \in W^{1,\infty}(\Omega;\mathbb{R}^d)$ and $C = C_P\,\|\mathbf{m}\|_{W^{1,\infty}(\Omega)}$, Poincaré's inequality yields

$$\begin{aligned}
&\langle(\mathbf{f} - \mathbf{f}_\mathcal{T}) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_\mathcal{T})\,;\,\mathbf{m} - \mathbf{m}_\mathcal{T}\rangle_{L^2(\Omega)}\\
&\quad\le\ C\,\|h\{(\mathbf{f} - \mathbf{f}_\mathcal{T}) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_\mathcal{T})\}\|_{L^1(\Omega)}.
\end{aligned}$$

Remark 3.5. We did not succeed in deriving an a posteriori bound for the a priori term $\|\lambda\mathbf{m} - \lambda_h\mathbf{m}_h\|_{L^2(\Omega)}$.

Proof of Theorem 3.3. Adopt notation from the proof of Theorem 3.1. By definition of the discretization scheme, we have

(3.8)
$$\mathbf{f} - \mathbf{f}_\mathcal{T} = (\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h) + (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_\mathcal{T}) + (\mathbf{d} - \mathbf{d}_h) + (\lambda\mathbf{m} - \lambda_h\mathbf{m}_h) \quad \text{a.e. in } \Omega.$$

This and the elementary inequality [5, Proof of Theorem 5.2]

$$-\langle\lambda\mathbf{m} - \lambda_h\mathbf{m}_h\,;\,\mathbf{m} - \mathbf{m}_h\rangle_{L^2(\Omega)} \le \int_\Omega \varepsilon|\lambda_h\mathbf{m}_h|\,|\lambda\mathbf{m} - \lambda_h\mathbf{m}_h|\,dx$$

allow us to dominate the left-hand side of (3.7),

$$\begin{aligned}
&\|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|^2_{L^2(\mathbb{R}^d)} + c_1\|\mathbf{d} - \mathbf{d}_h\|^2_{L^2(\Omega)}\\
&\quad\le \langle(\mathbf{f} - \mathbf{f}_\mathcal{T}) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_\mathcal{T})\,;\,\mathbf{m} - \mathbf{m}_h\rangle_{L^2(\Omega)} + \int_\Omega \varepsilon|\lambda_h\mathbf{m}_h|\,|\lambda\mathbf{m} - \lambda_h\mathbf{m}_h|\,dx.
\end{aligned}$$

In the scalar product, $\mathbf{m}-\mathbf{m}_h$ may be replaced my $\mathbf{m}-\mathbf{m}_{\mathcal{T}}$ due to orthogonality of $(\cdot)_{\mathcal{T}}$. Inserting $\lambda\mathbf{m}-\lambda_h\mathbf{m}_h$ from (3.8) into the integrand and serious use of the Cauchy inequality yield the assertion. ∎

4. Numerical algorithms. This section is devoted to the implementation of $(RP_{\varepsilon,h})$ for the uniaxial case (1.4) in MATLAB. The discrete problem $(RP_{\varepsilon,h})$ leads to a nonlinear systems of equations solved by a Newton–Raphson scheme. Subsections 4.2 and 4.3 describe an adaptive mesh-refinement based on refinement indicators motivated by Theorem 3.3 and the practical computation of the refinement indicators. The computation of both—the Galerkin element and the refinement indicators—involves the integral operator $\mathcal{P}$ and hence leads to dense matrices. Subsection 4.4 gives
an outlook of their efficient approximation with an $\mathcal{H}$-matrix approach.

4.1. Computation of the discrete solution $\mathbf{m_h}$. Given $\mathcal{T}=\{T_1,\dots,T_N\}$, the set $\mathcal{B}:=\big\{\chi_{T_j}\mathbf{e}_k\,\big|\,1\le j\le N,1\le k\le d\big\}$ is a basis of $\mathcal{L}^0(\mathcal{T})^d$, where $\mathbf{e}_k$ denotes the kth standard unit vector in $\mathbb{R}^d$. The computation of a discrete solution $\mathbf{m}_h=\sum_{j=1}^{dN}\mu_j\varphi_j$ is done via a Newton–Raphson scheme. To abbreviate notation and fix a numbering of the basis elements $\varphi_\ell\in\mathcal{B}$, let

$$\begin{cases}[j,1]:=j\\ [j,2]:=j+N\end{cases}\quad\text{for } d=2\quad\text{and}\quad\begin{cases}[j,1]:=j\\ [j,2]:=j+N\\ [j,3]:=j+2N\end{cases}\quad\text{for } d=3,$$

respectively, and for all $1\le j\le N$. Further let $\varphi_{[j,k]}:=\chi_{T_j}\mathbf{e}_k\in\mathcal{B}$. With $\mathbf{x}\in\mathbb{R}^{dN}$ and

$$\mathbf{m}_h=\sum_{j=1}^{N}\sum_{k=1}^{d}\mathbf{x}_{[j,k]}\varphi_{[j,k]},\tag{4.1}$$

equation (2.7) is equivalent to the nonlinear system $\mathbf{F}(\mathbf{x})=0$ with

$$\mathbf{F}:\mathbb{R}^{dN}\to\mathbb{R}^{dN},\quad \mathbf{F}_\ell=\langle\mathcal{P}\mathbf{m}_h\,;\,\varphi_\ell\rangle_{L^2(\Omega)}+\langle\lambda_h\mathbf{m}_h+D\phi^{**}(\mathbf{m}_h)-\mathbf{f}\,;\,\varphi_\ell\rangle_{L^2(\Omega)}.\tag{4.2}$$

Thus, the discrete scheme needs the computation of the matrix

$$\mathbf{A}\in\mathbb{R}^{dN\times dN},\quad \mathbf{A}_{mn}:=\langle\mathcal{P}\varphi_m\,;\,\varphi_n\rangle_{L^2(\Omega)}\quad\text{for basis functions }\varphi_m,\varphi_n\in\mathcal{B}.\tag{4.3}$$

Provided all $T_j\in\mathcal{T}$ are bounded Lipschitz domains, the following lemma allows for the *exact* computation of (4.3).

LEMMA 4.1 (see [10, 19]). *Let $\mathbf{m},\widetilde{\mathbf{m}}\in\mathbb{R}^d$ and let $\omega,\widetilde{\omega}\subseteq\mathbb{R}^d$ be bounded Lipschitz domains with outer normals $\mathbf{n},\widetilde{\mathbf{n}}$, respectively. Then $A(\chi_\omega\mathbf{m},\chi_{\widetilde{\omega}}\widetilde{\mathbf{m}}):=\langle\mathcal{P}(\chi_\omega\mathbf{m});\chi_{\widetilde{\omega}}\widetilde{\mathbf{m}}\rangle_{L^2(\mathbb{R}^d)}$ satisfies*

$$\begin{aligned}A(\chi_\omega\mathbf{m},\chi_{\widetilde{\omega}}\widetilde{\mathbf{m}})&=A(\chi_{\widetilde{\omega}}\widetilde{\mathbf{m}},\chi_\omega\mathbf{m})=A(\chi_\omega\widetilde{\mathbf{m}},\chi_{\widetilde{\omega}}\mathbf{m})\\&=-\int_\omega\int_{\widetilde{\omega}}G(x-y)(\mathbf{n}(x)\cdot\mathbf{m})(\widetilde{\mathbf{n}}(y)\cdot\widetilde{\mathbf{m}})\,ds_y\,ds_x.\end{aligned}\tag{4.4}$$

Remark 4.1. In the context of the boundary element method, boundary integrals

$$\int_E\int_{\widetilde{E}}G(x-y)\,ds_y\,ds_x$$

occur for the computation of the Galerkin elements for Symm's integral equation and piecewise constant ansatz functions. Analytic formulas are known for $E, \widetilde{E}$ being affine boundary pieces ($d = 2$) [15, 16, 4] or axis-orientated flat rectangles ($d = 3$) [15, 16, 9].

4.2. Adaptive mesh-refinement. Theorem 3.3 gives rise to the error estimators

$$\mu := \left(\sum_{T \in \mathcal{T}} \mu_T^2 \right)^{1/2} \quad \text{and} \quad \eta := \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2}, \tag{4.5}$$

where the refinement indicators μ_T, η_T, for $T \in \mathcal{T}$, are defined by

$$\begin{aligned} \ell_T &:= (\varepsilon \lambda_h |\mathbf{m}_h|)|_T = (|\mathbf{m}_h|_T| - 1)_+, \\ \mu_T^2 &:= (1 + \ell_T) \|(\mathbf{f} - \mathbf{f}_T) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_T)\|_{L^1(T)} + |T| \ell_T^2, \\ \eta_T^2 &:= (h_T + \ell_T) \|(\mathbf{f} - \mathbf{f}_T) - (\mathcal{P}\mathbf{m}_h - (\mathcal{P}\mathbf{m}_h)_T)\|_{L^1(T)} + |T| \ell_T^2. \end{aligned} \tag{4.6}$$

Remark 4.2. (a) The estimator μ is reliable, i.e., an upper bound for the error $\|\mathcal{P}\mathbf{m} - \mathcal{P}\mathbf{m}_h\|_{L^2(\mathbb{R}^d)} + \|D\phi^{**}(\mathbf{m}) - D\phi^{**}(\mathbf{m}_h)\|_{L^2(\Omega)}$ up to a multiplicative constant.

(b) μ cannot be efficient, i.e., a lower bound for the error.

(c) η is reliable for $\mathbf{m} \in W^{1,\infty}(\Omega; \mathbb{R}^d)$, but not in general.

(d) Efficiency of η is expected but could not be proven.

ALGORITHM 4.2 (adaptive mesh-refinement). *Let $\mathcal{T}^{(0)}$ be the initial triangulation, $n = 0$, $\alpha > 0$, and $0 \le \theta \le 1$.*

(i) *For $T_j \in \mathcal{T}^{(n)} = \{T_1, \dots, T_N\}$ choose a penalization parameter $\varepsilon_j = h_{T_j}^{\alpha} > 0$.*

(ii) *Compute approximation $\mathbf{m}_h$ with respect to the current triangulation $\mathcal{T}^{(n)}$ and $\varepsilon \in \mathcal{L}^0(\mathcal{T}^{(n)})$, $\varepsilon|_{T_j} := \varepsilon_j$, by the Newton–Raphson scheme.*

(iii) *Compute error estimators μ and η from* (4.5) *and refinement indicators $\eta_j := \eta_{T_j}$ and $\mu_j := \mu_{T_j}$ from* (4.6).

(iv) *Mark an element $T_j \in \mathcal{T}^{(n)}$ provided $\eta_j \ge \theta \max_{1 \le k \le N} \eta_k$ for η-adaptive mesh-refinement and provided $\mu_j \ge \theta \max_{1 \le k \le N} \mu_k$ for μ-adaptive mesh-refinement.*

(v) *Refine the marked elements, update $n \mapsto n + 1$, and go to* (i).

Remark 4.3. The choice $\theta = 0$ in Algorithm 4.2 leads to uniform mesh-refinement, whereas $\theta \approx 1$ leads to highly adapted meshes. In the numerical experiments, $\theta = 0$ or $\theta = 1/2$.

Remark 4.4. To lower the computational cost for the Newton–Raphson scheme, we used *nested iterations*: In step (ii) of Algorithm 4.2, the Newton–Raphson scheme was started with the prolonged discrete solution $\mathbf{m}_h^{(n-1)}$ for the previous grid $\mathcal{T}^{(n-1)}$.

4.3. Implementation of the refinement indicators. The L^1 norm in the definition of μ_T and η_T, respectively, was computed by a (2×2)-tensor Gauss quadrature rule. The following lemma shows that the point evaluation of $\mathcal{P}\mathbf{m}_h$ is well-defined outside the skeleton of $\mathcal{T}$.

LEMMA 4.3. *For $\mathbf{m}_h \in \mathcal{L}^0(\mathcal{T}; \mathbb{R}^d)$, the corresponding potential satisfies $\mathcal{L}\mathbf{m}_h \in \mathcal{C}(\mathbb{R}^d) \cap \mathcal{C}^1(\mathbb{R}^d \backslash \mathcal{S})$, where $\mathcal{S} := \bigcup \{\partial T \,|\, T \in \mathcal{T}\}$ denotes the skeleton of the triangulation. Moreover, the derivative $\mathcal{P}\mathbf{m}_h = \nabla(\mathcal{L}\mathbf{m}_h)$ can be computed pointwise by*

$$\mathcal{P}\mathbf{m}_h(x) = \frac{1}{|\mathbb{S}|} \sum_{T \in \mathcal{T}} \int_{\partial T} \frac{\mathbf{m}_h|_T \cdot (x - y)}{|x - y|^d} \, \mathbf{n}(y) \, ds_y \quad \text{for } x \in \mathbb{R}^d \backslash \mathcal{S}, \tag{4.7}$$

where $\mathbf{n}(y)$ *denotes the outer normal with respect to* $T \in \mathcal{T}$ *and* $\mathbb{S} \subseteq \mathbb{R}^d$ *the unit sphere.*

Proof. For the Lipschitz domain $T \in \mathcal{T}$, the Newtonian potential of the characteristic function satisfies $G * \chi_T \in \mathcal{C}^1(\mathbb{R}^d) \cap \mathcal{C}^2(\operatorname{int} T)$ with

$$\frac{\partial^2}{\partial x_j \partial x_k}\big(G * \chi_T\big)(x) = \frac{\partial}{\partial x_k}\big((\partial G/\partial x_j) * \chi_T\big)(x) = \int_{\partial T} \frac{\partial G}{\partial x_j}(x-y)\, n_k(y)\, ds_y; \tag{4.8}$$

see [8, Lemma 4.2]. With the abbreviated notation $\mathbf{m}^{(T)} := \mathbf{m}_h|_T$, we have

$$\mathcal{L}\mathbf{m}_h = \sum_{T\in\mathcal{T}} \sum_{j=1}^{d} m_j^{(T)}\, \frac{\partial G}{\partial x_j} * \chi_T = \sum_{T\in\mathcal{T}} \mathbf{m}^{(T)} \cdot \big((\nabla G) * \chi_T\big).$$

Computing the kth partial derivative of $\mathcal{L}\mathbf{m}_h$ via (4.8), we verify the assertion. ∎

Remark 4.5. Note that the singularities of $\mathcal{P}\mathbf{m}_h$ on the skeleton $\mathcal{S}$ are quite weak since it can be shown that $\mathcal{P}\mathbf{m}_h \in L^p(\Omega;\mathbb{R}^d)$ for all $1 < p < \infty$; see [19]. This seems to justify the computation of the L^1 norms by a simple quadrature rule.

Remark 4.6. In the context of the boundary element method, boundary integrals as in (4.7) occur for the computation of the double layer potential for piecewise constants. If one replaces ∂T by a bounded boundary piece E, the analytic formulas are known for E being affine ($d = 2$) [15, 16, 4] and being a triangle or rectangle ($d = 3$) [15, 16, 9].

4.4. Efficient realization of the involved integral operator $\boldsymbol{\mathcal{P}}$. The dense matrix $\mathbf{A} \in \mathbb{R}^{dN\times dN}_{sym}$ from (4.3) has certain symmetry properties. To decrease computation time and memory, $\mathcal{H}$- and $\mathcal{H}^2$-matrix approaches can be used [1, 10, 19, 18], where $\mathbf{A}$ is replaced by an approximation $\widetilde{\mathbf{A}}$.

LEMMA 4.4. *For any bounded open sets* $\omega, \widetilde{\omega} \subseteq \mathbb{R}^d$ *with* $\operatorname{dist}(\omega;\widetilde{\omega}) > 0$*, for* $\alpha, \beta = 1, \dots, d$*, and the* ℓ*th canonical unit vector* $\mathbf{e}_\ell \in \mathbb{R}^d$*, the bilinear form* $A(\cdot,\cdot)$ *from Lemma* 4.1 *satisfies*

$$A(\chi_\omega \mathbf{e}_\alpha, \chi_{\widetilde{\omega}} \mathbf{e}_\beta) = \int_\omega \int_{\widetilde{\omega}} \frac{\partial^2 G}{\partial x_\alpha \partial x_\beta}(x-y)\, dy\, dx.$$

Proof. The lemma follows from standard results on convolutions. ∎

The idea of the $\mathcal{H}^2$-matrix approach is to approximate the kernel

$$g_{\alpha\beta}(x,y) := \frac{\partial^2 G}{\partial x_\alpha \partial x_\beta}(x-y)$$

based on panel clustering. For certain $\sigma, \tau \subseteq \mathcal{T}$ with $\operatorname{dist}(\cup\sigma, \cup\tau) > 0$, let vectors $x_{m_1}^{(\sigma)} \in \cup\sigma$, $y_{m_2}^{(\tau)} \in \cup\tau$ and polynomials $p_{m_1}^{(\sigma)}, p_{m_2}^{(\tau)}$ on $\cup\sigma$, respectively, $\cup\tau$, be given and define

$$\widetilde{g}_{\alpha\beta}(x,y) := \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} g_{\alpha\beta}(x_m^{(\sigma)}, y_{m_2}^{(\tau)})\, p_{m_1}^{(\sigma)}(x)\, p_{m_2}^{(\tau)}(y) \quad \text{for } (x,y) \in \cup\sigma \times \cup\tau.$$

For $T_j \in \sigma$ and $T_k \in \tau$ there holds the approximation

$$\int_{T_j}\int_{T_k} g_{\alpha\beta}(x,y)\, dy\, dx \approx \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \underbrace{g_{\alpha\beta}(x_{m_1}^{(\sigma)}, y_{m_2}^{(\tau)})}_{=:D_{m_1 m_2}} \underbrace{\Big\{\int_{T_j} p_{m_1}^{(\sigma)}(x)\, dx\Big\}}_{=:C_{jm_1}^{(\sigma)}} \underbrace{\Big\{\int_{T_k} p_{m_2}^{(\tau)}(y)\, dy\Big\}}_{=:C_{km_2}^{(\tau)}}. \tag{4.9}$$

For fixed α, β, consider the matrix $B \in \mathbb{R}^{N\times N}_{sym}$, $B_{jk} := A(\chi_{T_j}\mathbf{e}_\alpha, \chi_{T_k}\mathbf{e}_\beta)$. With the matrices $C^{(\sigma)} \in \mathbb{R}^{|\sigma|\times M_1}$, $C^{(\tau)} \in \mathbb{R}^{|\tau|\times M_2}$, and $D \in \mathbb{R}^{M_1\times M_2}$ defined in (4.9), the submatrix $B|_{\sigma\times\tau}$ from B satisfies

$$B|_{\sigma\times\tau} \approx C^{(\sigma)} D (C^{(\tau)})^T. \tag{4.10}$$

The use of the latter approximation significantly reduces the computational cost for assembling the matrix $B|_{\sigma\times\tau}$, provided $\max\{M_1, M_2\} < \min\{|\sigma|, |\tau|\}$.

Remark 4.7. Notice that only the matrix D in (4.9) depends on α and β and the matrix $\mathbf{A}$ can be approximated by the block-matrix $\widetilde{\mathbf{A}}$ with blocks of $\mathcal{H}^2$-matrix type. The time to assemble the matrix $\mathbf{A}$ could be highly decreased by use of the indicated $\mathcal{H}^2$-matrix approach. However, all experiments in this paper have been made using the *exactly computed* matrix $\mathbf{A}$, but the much cheaper $\mathcal{H}^2$-matrix approach leads to (almost) the same accuracy, in (almost) linear complexity. (Since the present implementation is in MATLAB, comparisons will appear in [18].)

Remark 4.8. The computation of the refinement indicators can also be based on an $\mathcal{H}$-matrix approach since the computation of $\mathcal{P}\mathbf{m}_h(x)$ corresponds to a collocation method with the double layer potential; cf. Lemma 4.3.

5. Numerical example with exact solution $\mathbf{m} \in \boldsymbol{W^{1,\infty}(\Omega;\mathbb{R}^2)}$. The unit square $\Omega = (0,1)^2$ is filled with a uniaxial magnetic material (1.4) with easy axis $\mathbf{e} = (-1,1)/\sqrt{2}$, i.e., $\mathbf{z} = (1,1)/\sqrt{2}$ in Remark 2.2. Define

$$\mathbf{m}(x) := \begin{cases} x & \text{for } |x| \le 1, \\ x/|x| & \text{for } |x| \ge 1 \end{cases} \quad \text{and} \quad \lambda(x) := \begin{cases} 0 & \text{for } |x| < 1, \\ 1 & \text{for } |x| \ge 1. \end{cases} \tag{5.1}$$

Then $(\mathbf{m}, \lambda) \in W^{1,\infty}(\Omega;\mathbb{R}^2) \times L^\infty(\Omega)$ solves (2.4)–(2.5) with given right-hand side

$$\mathbf{f} := \mathcal{P}\mathbf{m} + (\mathbf{m}\cdot\mathbf{z})\mathbf{z} + \lambda\mathbf{m} \in L^2(\Omega;\mathbb{R}^2). \tag{5.2}$$

In all our numerical experiments, we replaced $\mathcal{P}\mathbf{m}$ on the right-hand side of (5.2) by $\mathcal{P}\mathbf{m}_\mathcal{T}$ for the elementwise integral means $\mathbf{m}_\mathcal{T}$ of $\mathbf{m}$. Recall that Lemma 4.1 allows the exact integration of $\mathcal{P}\mathbf{m}_\mathcal{T}$. Figure 1 shows discrete solutions $\mathbf{m}_h$ for the penalization parameter $\alpha = 1$.

For a given sequence of h-uniform meshes with $N = h^{-2}$ elements in $\mathcal{T}^{(n)}$, the first set of experiments studies the choice of the parameter $\alpha > 0$ in the penalization $\varepsilon = h^\alpha$. Figure 2 displays the L^2 error of the magnetization vectors as a function of the mesh-size $h = N^{-2}$ for 12 values of α. Any choice of $\alpha \ge 1$ seems to result in a linear convergence, while values $\alpha < 1$ seem to result in smaller experimental convergence rates (until $\alpha = 1/4$ with almost no convergence). The length $|\mathbf{m}_h(x)|$ for $|x| > 1$ and $\alpha = 1/2$ are about 1.1, compared to ≤ 1.01 for $\alpha = 3/2$. The value $\alpha = 3/2$ is recommended throughout all examples of this paper. Theoretical estimates concern the $\mathbf{z}$ direction of $\mathbf{m} - \mathbf{m}_h$ exclusively. In the numerical examples, however, linear convergence is observed also for the easy axis direction $\mathbf{e}$. Notice that $\mathbf{m}$ is essentially smooth, and hence adaptive mesh-refinements cannot improve the experimental convergence rates further.

In conclusion, the first example gives empirical support for the a priori analysis and the choice of the penalization parameter. As indicated by Theorem 3.2, the choice of $\varepsilon = h^\alpha$ with $\alpha \ge 1$ appears to be necessary for optimal experimental convergence

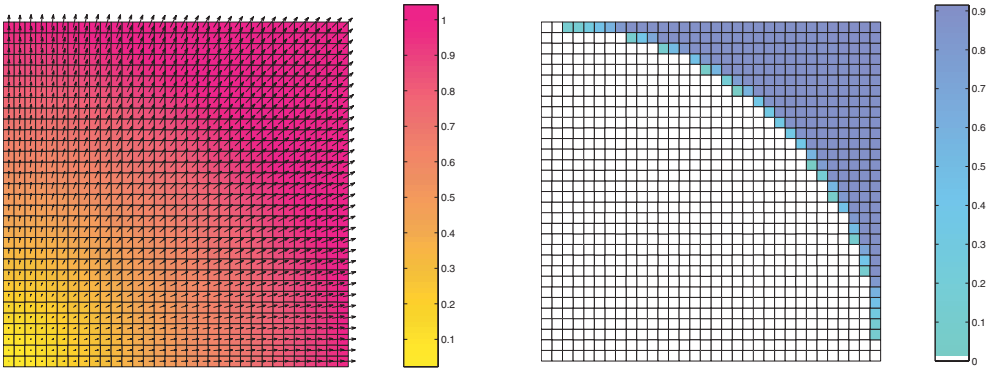


FIG. 1. *Discrete solution* $(\mathbf{m}_h, \lambda_h)$ *in section* 5 *on* $\mathcal{T}_4$ *(with* $N = 1024$*) for penalization parameter* $\varepsilon = h$*:* $\mathbf{m}_h$ *as vectors* $\mathbf{m}_h|_T$ *and* $|\mathbf{m}_h|_T|$ *in grayscale (left) and* λ_h *in grayscale (right).*

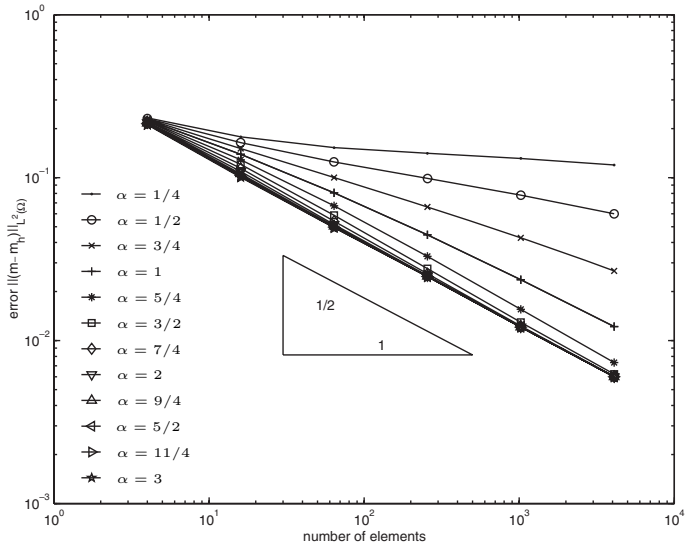


FIG. 2. *Error* $\|\mathbf{m} - \mathbf{m}_h\|_{L^2(\Omega)}$ *versus the number of elements* $N = 4, 16, \ldots, 4096$ *for various choices of the penalization parameter* $\varepsilon = h^\alpha$ *for uniform meshes with* $\alpha = .25, .5, .75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3$*. The results for* $\alpha = 1.5, 1.75, \ldots, 3$ *essentially coincide and lead to the linear convergence in* h *indicated by the slope* $1/2$*.*

behavior. The lower order of convergence for a choice of $\alpha < 1$ can be explained as follows: Theorem 3.2 shows the L^2 convergence $\lambda_h \mathbf{m}_h \to \lambda \mathbf{m}$ in Ω, particularly on the restricted domain $\omega := \{x \in \Omega \,|\, |x| \geq 1\}$; i.e., the smaller α, the larger the length $|\mathbf{m}_h|$:

$$h^{-\alpha}(|\mathbf{m}_h| - 1)_+ = \lambda_h |\mathbf{m}_h| \to \lambda |\mathbf{m}| = 1 \quad \text{in } L^2(\omega).$$

6. Numerical example with exact solution $\mathbf{m} \notin \boldsymbol{H^1(\Omega)}$. This section is devoted to the numerical approximation of a more singular magnetization,

$$(\mathbf{m}(x), \lambda(x)) := \begin{cases} (\mathbf{y}(x), 0) & \text{for } x \in \omega, \\ (x_1 x_2 (1 - y_1(x))^{-1} (1 - y_2(x))^{-2} \mathbf{y}(x), 1) & \text{for } x \in \Omega \backslash \omega \end{cases} \tag{6.1}$$

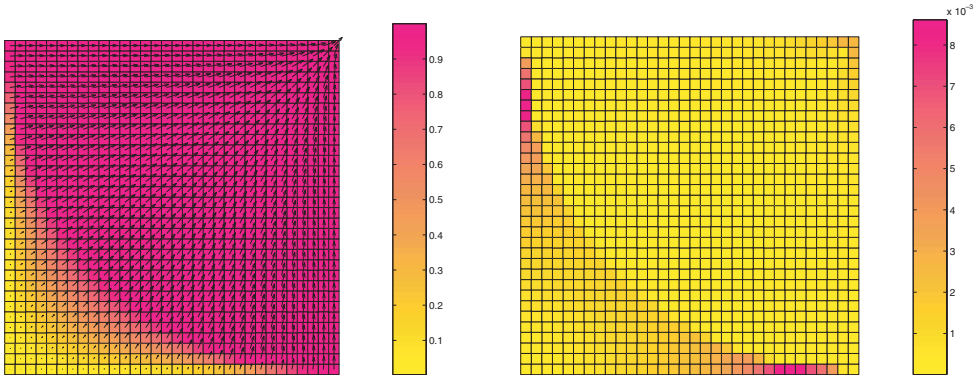


FIG. 3. *Best approximation* $\mathbf{m}_{\mathcal{T}}$ *(left) of the magnetization vector* $\mathbf{m}$ *in section* 6 *on a uniform mesh (with* $N = 1024$*), and elementwise distribution of the corresponding best approximation error* $\|\mathbf{m} - \mathbf{m}_{\mathcal{T}}\|_{L^2(T_j)}$ *(right). Notice that the grayscale displays values multiplied with* 1 *(left) and* $10^{-3} = 1/1000$ *(right).*

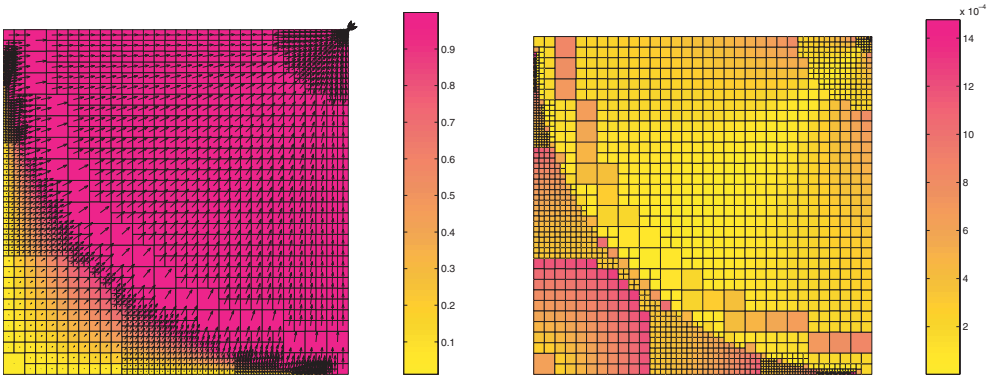


FIG. 4. *Best approximation* $\mathbf{m}_{\mathcal{T}}$ *(left) in section* 6 *on an error adapted generated mesh (with* $N = 1717$*), and elementwise distribution of the corresponding best approximation error* $\|\mathbf{m} - \mathbf{m}_{\mathcal{T}}\|_{L^2(\Omega)}$ *(right). For the adaptive mesh-refinement, we used Algorithm* 4.2 *with refinement indicators* $\varrho_j = \|\mathbf{m} - \mathbf{m}_{\mathcal{T}}\|_{L^2(T_j)}$*. Notice that the grayscale displays values multiplied with* 1 *(left) and* $10^{-4} = 1/10000$ *(right).*

with a singular gradient at the three vertices $(0,1)$, $(1,0)$, $(1,1)$ on the boundary of the magnetic body $\Omega = (0,1)^2$. Here,

$$\mathbf{y}(x) := \frac{(1,1) - x}{|(1,1) - x|} \quad \text{and} \quad \omega := \big\{x \in \Omega \,\big|\, |(1,1) - x| < 1\big\}.$$

The remaining data ϕ, $\mathbf{z}$, and $\mathbf{f}$ are as in section 5. The magnetization vector $\mathbf{m}$ (6.1) and the error by the piecewise integral means are depicted in Figure 3. We observe a larger elementwise L^2 error in $\Omega\backslash\omega$ and hence expect the necessity of adaptive mesh-refining for an effective computation. For a comparison, Figure 4 displays the best approximation $\mathbf{m}_{\mathcal{T}}$ and its elementwise L^2 errors $\|\mathbf{m} - \mathbf{m}_{\mathcal{T}}\|_{L^2(T)}$ on an adapted mesh. The latter was generated by Algorithm 4.2 with the refinement indicator $\varrho_j := \|\mathbf{m} - \mathbf{m}_{\mathcal{T}}\|_{L^2(T_j)}$; i.e., an element T_j is marked in step (iv) if $\varrho_j \geq 1/2 \max_{1 \leq k \leq n} \varrho_k$. The singularity at $(1,1)$ is visible in Figures 3 and 4, as is a refinement near the arc $\big\{x \in \Omega \,\big|\, |(1,1) - x| = 1\big\}$. There is no theoretical support

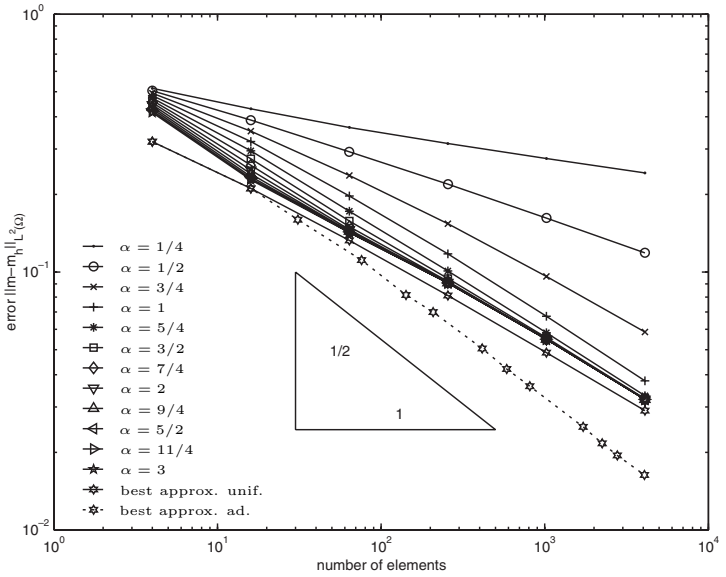


FIG. 5. *Error* $\|\mathbf{m}-\mathbf{m}_h\|_{L^2(\Omega)}$ *and dependence from penalization parameter* $\varepsilon = h^\alpha$ *for uniform mesh-refinement. For comparison, the best approximation errors on uniform and adapted meshes from Figure* 4 *are also shown.*

that the refinement indicator ϱ_j leads to optimal meshes, but it allows an interesting theoretical comparison. Also, heuristically we expect optimal meshes (asymptotically) since the mesh-refinement with respect to ϱ leads to meshes on which the best approximation errors are equidistributed.

A comparison between Figures 3 and 4 (keeping in mind the different scalings in the right figures) shows that adaptive meshes have the potential for improvement. Numerical evidence for this is provided in Figure 5: Besides various choices of penalization parameter α with conclusions similar to those of section 5, the sequence of uniform and ϱ-adapted mesh-refining are compared. The adaptive meshes yield linear convergence in a reference mesh-size $h := N^{-2}$ even in all components of $\|\mathbf{m} - \mathbf{m}_{\mathcal{T}}\|_{L^2(\Omega)}$. The sequence of uniform meshes shows a suboptimal convergence rate.

Figures 6 and 7 display the L^2 error $\|\mathbf{m} - \mathbf{m}_h\|_{L^2(\Omega)}$ and the error estimators μ and η from (4.5) as functions of the number of elements N for various sequences of meshes. Those are generated by Algorithm 4.2 with refinement indicators η_j and μ_j from (4.6) and penalizations $\varepsilon = h^\alpha$ for $\alpha = 1$ and $\alpha = 3/2$. The two penalizations show similar convergence rates; the overall recommendation of $\alpha = 3/2$ is again supported in Figure 6 by better results. Each of the two adaptive algorithms leads to optimal convergence rates and is (asymptotically) factor 2 for $\alpha = 1$, respectively, 1.3, for $\alpha = 3/2$ worse than the best approximation errors.

Figure 7 illustrates the reliability-efficiency gap [3]: What is reliable is not efficient and what is efficient is not (known to be) reliable. Theorem 3.2 and Remark 3.1 show that the error terms are bounded from above by $c_1\mu$ and $c_2(\mathbf{m})\eta$, and the latter bound is of higher order but valid only for a smooth magnetization. The second estimate is also expected to be efficient (up to higher-order terms in the magnetization). Figure 7 displays η and μ and clearly shows their different convergence rates. From Figure 7 there is no support that adaptive is more effective than uniform mesh-refining.

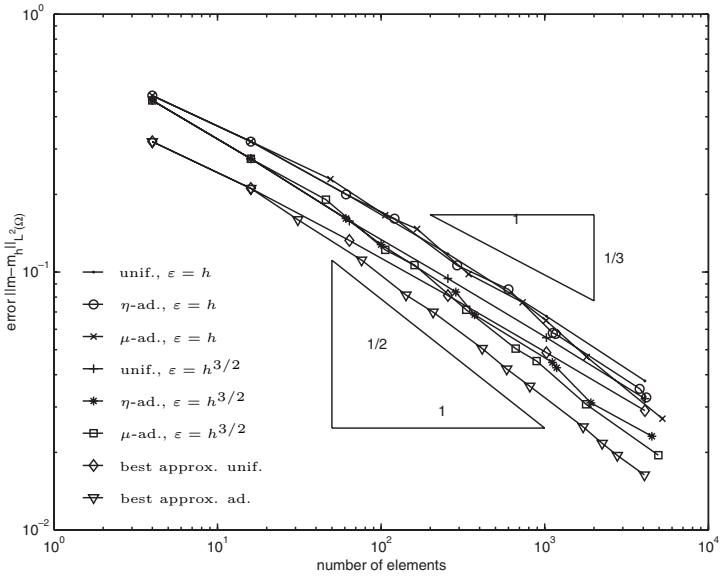


FIG. 6. *Error* $\|\mathbf{m} - \mathbf{m}_h\|_{L^2(\Omega)}$ *for uniform,* μ*-adaptive, and* η*-adaptive mesh-refinement with penalization parameter* $\varepsilon = h^\alpha$ *or* $\varepsilon = h^{3/2}$*. For comparison, the best approximation error is shown for uniform and optimal adapted meshes as well. Both adaptive strategies lead to optimal experimental convergence rate* $1/2$ *in terms of numbers of elements.*

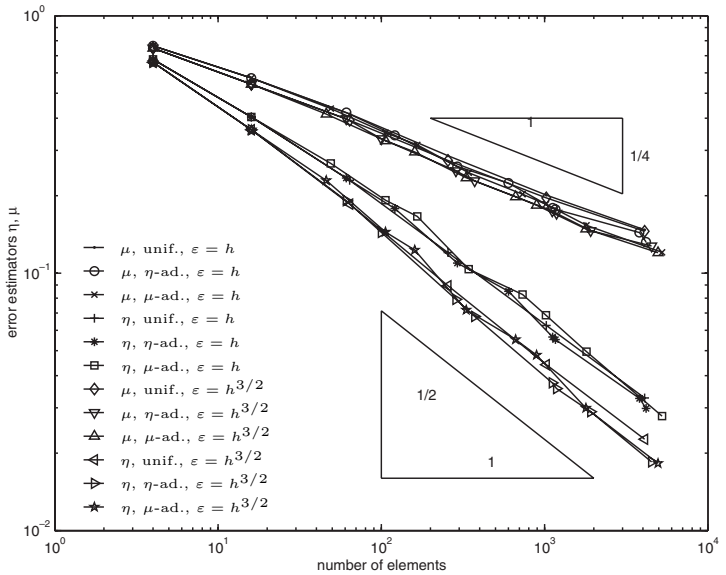


FIG. 7. *Illustration of the reliability-efficiency gap: Experimental convergence of the error estimators* η *and* μ *in section* 6 *for uniform,* η*-adaptive, and* μ*-adaptive mesh-refinement with penalization parameter* $\varepsilon = h$ *or* $\varepsilon = h^{3/2}$*. The improvement of the error by adaptive mesh-refinement strategies as shown in Figure* 6 *is not reflected by the estimators. There is (up to a multiplicative constant) no improvement of the convergence behavior by the adaptive mesh-refinement.*

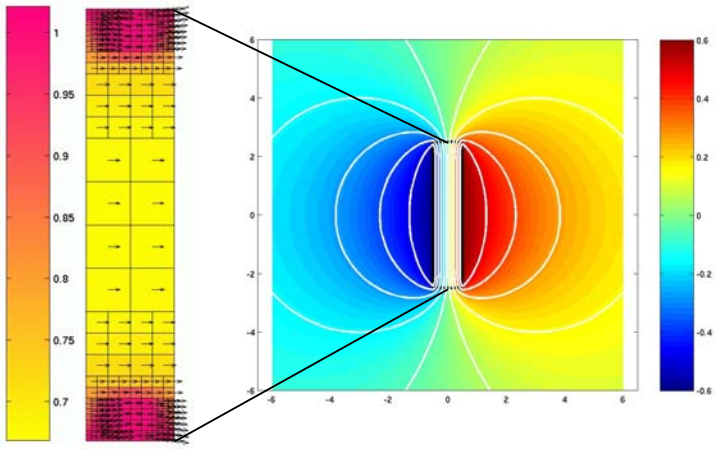


FIG. 8. *Discrete magnetization* $\mathbf{m}_h$ *(zoom on the left) on the* η*-adaptively generated mesh* $\mathcal{T}_4$ *(with* $N = 236$*) and corresponding potential* u_h *(right) for constant exterior field* $\mathbf{f} = (.6, 0)$*, easy axis* $\mathbf{e} = (1, 0)$*, and penalization parameter* $\varepsilon = h^{3/2}$*. The grayscale in the zoomed magnet displays the length* $|\mathbf{m}_h|$ *of the discrete magnetization. On the right, the pointwise value of* u_h *is shown by the grayscale and some isolines have been drawn.*

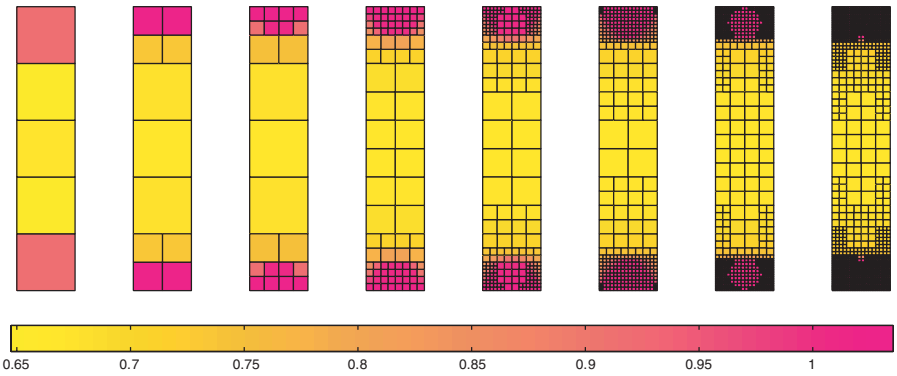


FIG. 9. η*-adaptively generated meshes* $\mathcal{T}_0$ *(with* $N = 5$*) till* $\mathcal{T}_7$ *(with* $N = 1604$*) in section 7 for* $\mathbf{f} = (.6, 0)$*,* $\mathbf{e} = (1, 0)$*, and* $\varepsilon = h^{3/2}$*. The grayscale shows the length* $|\mathbf{m}_h|$ *of the discrete solution.*

7. Real-life scientific computing. The ferromagnetic body $\Omega = (-1/2, 1/2) \times (-5/2, 5/2)$ is loaded with a constant applied magnetic field $\mathbf{f} := (0.6, 0)$ aligned with the easy axis $\mathbf{e} = (1, 0)$. Figure 8 displays the magnetic potential $u_h = \mathcal{L}\mathbf{m}_h$ and the magnetization vectors $\mathbf{m}_h$ on an adaptively generated mesh. The exact solution $\mathbf{m}$ is unknown. The first numerical computations for this example have been performed in [5]. Although there the potential equation

$$\operatorname{div}(-\nabla u + \mathbf{m}) = 0 \quad \text{in } \mathbb{R}^2$$

is discretized and solved by a finite element scheme for a *bounded* domain that surrounds Ω instead of the full space, we obtain similar results.

The initial mesh $\mathcal{T}_0$ consists of 5 congruent squares with side-length 1. Figure 9 shows η-adaptively generated meshes $\mathcal{T}_0, \dots, \mathcal{T}_7$ with $N = 5, \dots, 1604$ elements. We observe some mesh-refinement towards the 4 vertices of Ω which we might expect to be caused by singularities in the stray field. However, this refinement seems to be

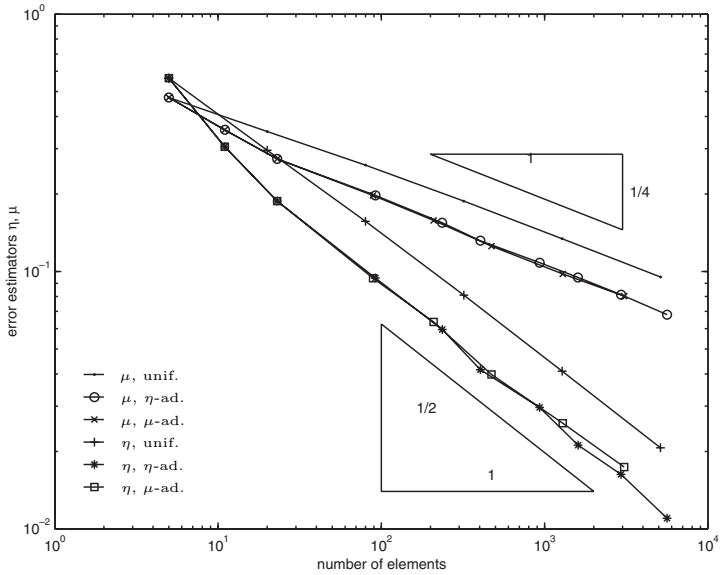


FIG. 10. *Experimental convergence of the error estimators η and μ in section 7 for uniform, η-adaptive, and μ-adaptive mesh-refinement and penalization parameter $\varepsilon = h^{3/2}$. There is (up to a multiplicative constant) no improvement of the convergence behavior by the adaptive mesh-refinement, although we obtain some local mesh-refinement towards the corners in Figure 9.*

accomplished in $\mathcal{T}_0, \dots, \mathcal{T}_6$ as $\mathcal{T}_7$ and $\mathcal{T}_8$ show a refinement of a more global zone. Figure 8 displays the discrete solution which follows the exterior field $\mathbf{f}$ and develops some flowering at the tips of Ω. One observes large curvatures of the magnetization near the top and bottom of the magnet Ω but no strong point singularity there. Furthermore, Figure 8 displays the corresponding magnetic potential $u_h = \mathcal{L}\mathbf{m}_h$ computed analytically by

$$\mathcal{L}\mathbf{m}_h(x) = \sum_{k=1}^{N} \int_{\partial T_k} G(x-y)\, \mathbf{m}_h|_{T_k} \cdot \mathbf{n}(y)\, ds_y \qquad \text{for all } x \in \mathbb{R}^d, \tag{7.1}$$

as follows from partial integration of (2.3). In comparison with a corresponding numerical experiment in [5], we see that the potential lines of the magnetic potential are not perpendicular on the boundary of the domain displayed. This is a consequence of the correct treatment of the stray field in the full space $\mathbb{R}^2$. More important, the discretization in [5] shows a strong refinement towards the vertices, much stronger than those visible in Figures 8 or 10. To monitor the asymptotic behavior, Figure 10 displays the error estimators μ and η. In comparison with uniform and η- and μ-adaptive mesh-refinement one deduces that, in this example, adaptivity is not important—there is a small improvement, but one obtains essentially the same convergence rate for all three strategies. Our interpretation is that, to our great surprise, there is no singularity in the integral-operator model at hand and so the formulation is indeed superior to that of [5].

Finally, Figure 11 shows the discrete Lagrange multipliers λ_h corresponding to the triangulations from Figure 9. They do not indicate some particular resolution of the set $\{x \in \Omega : \lambda_h(x) = 0\}$ (or some other level set of λ_h).

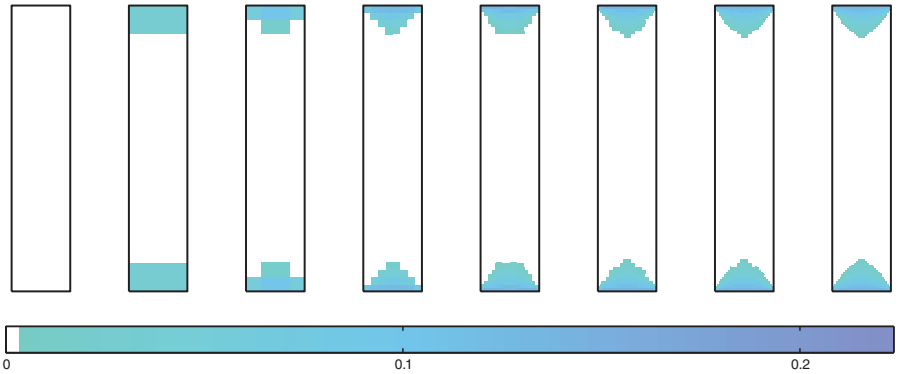


FIG. 11. *Discrete Lagrange multiplier λ_h on η-adaptively generated meshes $\mathcal{T}_0$ (with $N = 5$) till $\mathcal{T}_7$ (with $N = 1604$) in section 7 for $\mathbf{f} = (.6, 0)$, $\mathbf{e} = (1, 0)$, and $\varepsilon = h^{3/2}$. The grayscale shows the pointwise value of λ_h. In the white region we have $\lambda_h \equiv 0$, i.e., $|\mathbf{m}_h| \leq 1$.*

REFERENCES

[1] S. BÖRM, L. GRASEDYCK, AND W. HACKBUSCH, *Introduction to hierarchical matrices with applications*, Engrg. Anal. Boundary Elements, 27 (2003), pp. 405–422.

[2] W. F. BROWN, *Micromagnetics*, John Wiley and Sons, New York, 1963.

[3] C. CARSTENSEN AND K. JOCHIMSEN, *Adaptive finite element error control for non-convex minimization problems: Numerical two-well model example allowing microstructures*, Computing, 71 (2003), pp. 175–204.

[4] C. CARSTENSEN AND D. PRAETORIUS, *A posteriori error control in adaptive qualocation boundary element analysis for a logarithmic-kernel integral equation of the first kind*, SIAM J. Sci. Comput., 25 (2003), pp. 259–283.

[5] C. CARSTENSEN AND A. PROHL, *Numerical analysis of relaxed micromagnetics by penalised finite elements*, Numer. Math., 90 (2001), pp. 65–99.

[6] A. DESIMONE, *Energy minimizers for large ferromagnetic bodies*, Arch. Rational Mech. Anal., 125 (1993), pp. 99–143.

[7] S. A. FUNKEN AND A. PROHL, *On Stabilized Finite Element Methods in Relaxed Micromagnetism*, Report 00-02, Berichtsreihe des Math. Sem. Kiel, 2000; preprint available from http://www.numerik.uni-kiel.de/reports/1999/99-18.html.

[8] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer-Verlag, New York, 1977.

[9] W. HACKBUSCH, *Direct integration of the Newton potential over cubes including a program description*, Computing, 68 (2002), pp. 193–216.

[10] W. HACKBUSCH AND M. MELENK, *$\mathcal{H}$-Matrix Treatment of the Operator $\nabla\Delta^{-1}$*div, in preparation.

[11] A. HUBERT AND R. SCHÄFER, *Magnetic Domains*, Springer-Verlag, Berlin, 1998.

[12] R. D. JAMES AND D. KINDERLEHRER, *Frustration in ferromagnetic materials*, Continuum Mech. Thermodyn., 2 (1990), pp. 215–239.

[13] M. LUSKIN AND L. MA, *Analysis of the finite element approximation of microstructure in micromagnetics*, SIAM J. Numer. Anal., 29 (1992), pp. 320–331.

[14] L. MA, *Analysis and Computation for a Variational Problem in Micromagnetics*, Ph.D. thesis, University of Minnesota, 1991.

[15] M. MAISCHAK, *The Analytical Computation of the Galerkin Elements for the Laplace, Lamé, and Helmholtz Equation in 2D BEM*, Institut für Angewandte Mathematik, Universität Hannover, preprint, 1999; available online from http://www.ifam.uni-hannover.de/~maischak/publication.html.

[16] M. MAISCHAK, *The Analytical Computation of the Galerkin Elements for the Laplace, Lamé, and Helmholtz Equation in 3D BEM*, Institut für Angewandte Mathematik, Universität Hannover, preprint, 2000; available online from http://www.ifam.uni-hannover.de/~maischak/publication.html.

[17] P. PEDREGAL, *Parametrized Measures and Variational Principles*, Birkhäuser, Basel, 1997.

[18] N. Popović and D. Praetorius, *Application of H-matrix techniques for micromagnetics*, Computing, to appear.
[19] D. Praetorius, *Analysis of the operator* $\Delta^{-1}\mathrm{div}$ *arising in magnetic models*, Z. Anal. Anwendungen, 23 (2004), pp. 589–605.
[20] A. Prohl, *Computational Micromagnetism*, Teubner, Stuttgart, 2001.
[21] L. Tartar, *Beyond Young Measures*, Meccanica, 30 (1995), pp. 505–526.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2652–2670

A CONVOLUTION-THRESHOLDING APPROXIMATION OF GENERALIZED CURVATURE FLOWS*

RICHARDS GRZHIBOVSKIS† AND ALEXEI HEINTZ†

Abstract. We construct a convolution-thresholding approximation scheme for the geometric surface evolution in the case when the velocity of the surface at each point is a given function of the mean curvature. Conditions for the monotonicity of the scheme are found and the convergence of the approximations to the corresponding viscosity solution is proved. We also discuss some aspects of the numerical implementation of such schemes and present several numerical results.

Key words. generalized curvature flow, convolution-thresholding scheme, viscosity solution, level-set equation

AMS subject classifications. 65M12, 53C44, 49L25, 35K55

DOI. 10.1137/S0036142903431316

1. Introduction.

The topic of curvature flows of different types was popular during the last 20 years and is still popular in both pure and applied mathematics. By curvature flow we mean a family $\{\Gamma_t\}_{t\geq 0}$ of hypersurfaces in $\mathbb{R}^n$ depending on time t with local normal velocity equal to the mean curvature or a function of it for generalized curvature flows. The mean curvature in turn denotes here the sum of principal curvatures.

In the three-dimensional case a smooth initial surface can develop singularities after some finite time. There have been several successful attempts to deal with singularities and topological complications: the varifold approach [7], [2], the phase field method [14], [8], and the level-set method. This approach was suggested in the physical literature [26] and was extensively developed for numerical purposes by Osher and Sethian [27]. The main idea of this method is to evolve some continuous function $u : [0,\infty)\times\mathbb{R}^n \mapsto \mathbb{R}$ in such a way that $\Gamma_t \subset \mathbb{R}^n$ would always be a level-set of $u(x,t)$, i.e., $\Gamma_t = \{x \in \mathbb{R}^n : u(x,t) = 0\}$ for all $t \geq 0$. In the case of the mean curvature flow, the evolution equation for u turns out to be

$$u_t = |Du| \operatorname{div}\left(\frac{Du}{|Du|}\right). \tag{1.1}$$

The evolution equation for a function u with each point of a level-set moving along the normal with velocity equal to some function G of the mean curvature is the so-called generalized mean curvature evolution PDE

$$u_t = |Du|\, G\left(\operatorname{div}\left(\frac{Du}{|Du|}\right)\right). \tag{1.2}$$

*Received by the editors July 8, 2003; accepted for publication (in revised form) June 7, 2004; published electronically April 19, 2005. This research was partially supported by the TFR grant "Kinetic modelling of geometric flows of surfaces" and the TMR contract "Asymptotic methods in kinetic theory" (ERB FMRX CT97 0157).

http://www.siam.org/journals/sinum/42-6/43131.html

†Department of Mathematics, Chalmers University of Technology, 41296 Göteborg, Sweden (richards@math.chalmers.se, heintz@math.chalmers.se).

This equation is degenerate parabolic. The existence and uniqueness of generalized viscosity solutions (see [12]) to the initial value problem

$$\begin{cases} u_t = |Du|\, G\left(\operatorname{div}\left(\frac{Du}{|Du|}\right)\right) & \text{in } \mathbb{R}^n \times (0,T), \\ u = g(x) \in BUC(\mathbb{R}^n) & \text{on } \mathbb{R}^n \times \{0\} \end{cases} \tag{1.3}$$

was investigated in [17], [11], [22].

Curvature flows arise naturally in various problems. Among these are the fast reaction–slow diffusion problem [29], [4], [16], [19] and image processing [1].

In the present work we construct a class of approximations of a convolution-thresholding type to the generalized curvature flows. By this we mean the following. Assume that, initially, the surface under consideration is a boundary of a compact set $C \in \mathbb{R}^n$. Take compactly supported functions $\tilde{\rho}_i : \mathbb{R}_+, \longmapsto \mathbb{R}_+, i = 1,2$ (in fact, one can also take $\tilde{\rho}_i$ with unbounded support decreasing fast for large x). We define $\rho_i : \mathbb{R}^n \mapsto \mathbb{R}_+$,

$$\rho_i(x) = \frac{1}{h^{n/2}} \tilde{\rho}_i\left(|x| / \sqrt{h}\right),$$

and introduce a convolution

$$M_i(C)(x,h) = \int_{\mathbb{R}^n} \chi_C(y)\, \rho_i(x-y)\, dy.$$

Now $M_i(C)(x,h)$ are functions of x, and we define a new position of the surface as a boundary of the set

$$\mathcal{H}_h C = \{x \in \mathbb{R}^n : F(M_1(C)(x,h), M_2(C)(x,h)) \geq 0\}, \tag{1.4}$$

where F is some (thresholding) function. Next we follow Evans [15] and introduce an operator on the space of bounded functions $\mathbb{B}(\mathbb{R}^n)$: $H(h) : \mathbb{B}(\mathbb{R}^n) \mapsto \mathbb{B}(\mathbb{R}^n)$ by

$$[H(h)\, u](x) = \sup\{\lambda \in \mathbb{R} : x \in \mathcal{H}_h [u \geq \lambda]\}. \tag{1.5}$$

The purpose of the present study is, for a given function G in (1.3), to find a corresponding thresholding function F in (1.4) so that $H(t/m)^m g(x)$ converges to the unique viscosity solution of (1.3) as $m \to \infty$.

Such a function in the case when G is linear was proposed by Merriman, Bence, and Osher in [25]. This result is often referred to as the Bence–Merriman–Osher method. Rigorous proofs of the convergence of such approximations can be found in [15], [20], and [3]. In this case it is enough to take a thresholding function depending only on one convolution.

Suppose that G is nonlinear. As we show in section 3, in this case one has to use two convolutions M_1 and M_2 and a thresholding function depending on two variables $F(M_1, M_2)$. This is necessary to ensure that the operator H is consistent with the PDE in (1.3). We also show how to choose convolution kernels in order to get a monotone H. These two conditions—monotonicity and consistency—are crucial for the convergence.

Using our approach we also suggest a new construction of higher order schemes for the classical curvature flows. The numerical experiments with these schemes show a considerable improvement in the accuracy.

Finite difference approximations for (1.3) have been studied in [27], [31], [13].

Another class of approximation operators, the so-called Matheron filters, comes from image processing. The connection between such operators and the mean curvature evolution PDE (1.2) was established in [10]. This result was then extended in [18] and [9].

Threshold dynamics models, introduced earlier in [21], lead to approximations of the solution of the Cauchy problem to a nonlinear parabolic equation, where the right-hand side can be interpreted as a general elliptic operator on a level set of the solution. This is a generalization of the curvature flow, but it does not entirely include (1.3) as a special case.

Another generalization of the Bence–Merriman–Osher method can be found in [23]. The author suggests an approximation procedure that allows tracking the surface evolution when the velocity of the surface depends also on the coordinates. The convergence of this approximation is also proved.

Outline. This paper is organized as follows. After introducing the basic notions and stating some results for viscosity solutions in section 2, we turn to our method of approximation for such solutions. In section 3, we construct F to get the convergence of the convolution-thresholding approximation to the viscosity solution of (1.3) with a monotone continuous function G. This is the main result of the paper. More precisely, the following local uniform convergence is proved:

$$\left(\left(H\left(t/m\right)\right)^m g\right)(x) \to u\left(x,t\right), \qquad m \to \infty,$$

where H is defined by (1.5) and $u\left(x,t\right)$ is the viscosity solution of (1.3).

We use this construction for numerical calculation for some cases of the generalized curvature flows in $\mathbb{R}^2$ and $\mathbb{R}^3$. Numerical results and two approaches to the implementation are described in section 4.

2. The viscosity solution framework. Consider the nonlinear equation (1.2) on an open set $\Omega \times (0,T)$ with function G continuous and nondecreasing. This is the second order equation with a right-hand side that is monotonic and degenerate elliptic (see [12]) provided that G is nondecreasing and $Du \neq 0$. Viscosity solution to (1.2) was defined by Evans and Spruck in [17] and by Chen, Giga, and Goto in [11]. In our presentation we will use a somewhat more general definition of viscosity solutions introduced by Ishii and Souganidis in [22] to allow for a wider class of functions G in (1.2). For the general degenerate elliptic equation

$$u_t + \mathcal{G}\left(Du, D^2u\right) = 0, \tag{2.1}$$

they introduce a special class of test functions and adapt the definition of viscosity solution for possible singularities of the right-hand side. Representation of (1.3) in the form of (2.1) gives

$$\mathcal{G}(p,X) = -|p| G\left(\frac{1}{|p|}\mathrm{tr}\left(\left(I - \frac{p\otimes p}{|p|^2}\right)X\right)\right).$$

Let us begin by introducing an auxiliary subclass of $C^2\left([0,\infty)\right)$. We say that $f : [0,\infty) \mapsto \mathbb{R}$ lies in $\mathcal{F} \subset C^2$ if $f\left(0\right) = f'\left(0\right) = f''\left(0\right) = 0$, $f''\left(r\right) > 0$ for $r > 0$ and the following limits hold:

$$\lim \frac{f'\left(|p|\right)}{|p|}\mathcal{G}\left(p,I\right) = \lim \frac{f'\left(|p|\right)}{|p|}\mathcal{G}\left(p,-I\right) = 0.$$

As was shown in [22], this set of functions is a nonempty cone, provided that the right-hand side lies in $C((\mathbb{R}^n \setminus \{0\}) \times \mathbb{S}(n))$. The class of test functions $\mathcal{A}(\mathcal{G})$ depends on $\mathcal{G}$ and is defined as follows.

DEFINITION 2.1. *A function ϕ is admissible if it is in $C^2(\mathbb{R}^n \times (0,T))$ and if, for each $\acute{z} = (\acute{x}, \acute{t})$ where $D\phi(\acute{z}) = 0$, there is $\delta > 0$, $f \in \mathcal{F}$, and $\omega \in C([0,\infty))$ such that $\omega = o(r)$ and for all $(x,t) \in B(\acute{z}, \delta)$*

$$\left|\phi(x,t) - \phi(\acute{z}) - \phi_t(\acute{z})\left(t - \acute{t}\right)\right| \leq f\left(|x - \acute{x}|\right) + \omega\left(\left|t - \acute{t}\right|\right).$$

Let us also denote by u^* and u_* the upper and lower semicontinuous envelopes of u:

$$u^*(x,t) = \limsup_{(y,s)\to(x,t)} u(y,s), \quad u_*(x,t) = \liminf_{(y,s)\to(x,t)} u(y,s).$$

The definition of viscosity solution follows.

DEFINITION 2.2. *Take an open set $\tilde{\mathcal{O}} \subset \mathbb{R}^n$ and $\mathcal{O} = \tilde{\mathcal{O}} \times (0,T)$. $u : \mathcal{O} \subset \mathbb{R}^n \times (0,T) \mapsto \mathbb{R} \cup \{-\infty\}$ is a viscosity subsolution (supersolution) of* (1.2) *in an open $\mathcal{O}$ if $u^* < \infty$ ($u_* > -\infty$) and for all $\phi \in \mathcal{A}(G)$ and all local maximum (minimum) points (z_0, t_0) of $u^* - \phi$ ($u_* - \phi$),*

$$\begin{cases} \phi_t(z_0,t_0) \leq (\geq) |D\phi(z_0,t_0)| \, G\left(\operatorname{div} \dfrac{\mathrm{D}\phi(\mathrm{z}_0,\mathrm{t}_0)}{|\mathrm{D}\phi(\mathrm{z}_0,\mathrm{t}_0)|}\right) & \text{if } D\phi(z) \neq 0, \\ \phi_t(z_0,t_0) \leq (\geq) 0 & \text{otherwise.} \end{cases}$$

Consequently, a viscosity solution is a function that is sub- and supersolution simultaneously.

The result by Ishii and Souganidis presented in [22] can be restated in terms of the level-set equation (see [28]) as follows.

THEOREM 2.3. *Assume that G is continuous and nondecreasing. Then the initial value problem* (1.3) *has a unique viscosity solution $u \in BUC(\mathbb{R}^n \times (0,T))$.*

In what follows, we also use another result by Ishii and Souganidis [22] concerning locally uniform perturbations of the right-hand side of the equation. One can restate this result in the case of (1.2) as follows (see [28]).

THEOREM 2.4. *Assume that G is continuous and nondecreasing. Suppose also that $\{G_m\}_1^\infty$ is a sequence of continuous, nondecreasing functions on $\mathbb{R}$ and $G_m \to G$ locally uniformly. For any m, let $\mathcal{F}(G) \subset \mathcal{F}(G_m)$ and for any $f \in \mathcal{F}(G)$,*

$$\liminf_{p\to,\, m\to\infty} f'(|p|)\, G_m(1/p) \geq 0,$$
$$\left(resp., \ \limsup_{p\mapsto 0,\, m\to\infty} f'(|p|)\, G_m(-1/p) \leq 0\right).$$

Let u_m be a subsolution (resp., supersolution) of

$$\frac{\partial u_m}{\partial t} = |Du_m| \, G_m\left(\operatorname{div} \frac{Du_m}{|Du_m|}\right) \text{ in } \mathcal{O}.$$

Then

$$u^+(z) = \limsup_{r\mapsto 0} \{u_m(y), |y - z| \leq r, \ m > 1/m\}, \tag{2.2}$$

$$\left(resp., u_+(z) = \liminf_{r\to 0} \{u_m(y), |y - z| \leq r, \ m > 1/m\}\right) \tag{2.3}$$

is a subsolution (resp., supersolution) of (1.2) *in $\mathcal{O}$ provided that $u^+ < \infty$ (resp., $u_+ > -\infty$).*

3. A convolution-thresholding method for a generalized curvature flow.

3.1. Convergence of approximation schemes. Here we make use of a theorem by Barles and Souganidis proved in [5]. In order to base the proof of our main result on this theorem, we follow Pasquignon [28] and restate it in terms of (1.2).

Let $H(h)$ be the approximation operator, i.e.,

$$u_h(x,(n+1)h) = H(h)\,u_h(x,nh) = H(h)^{n+1}\,u_0(x),$$
$$u_h(x,0) = u_0(x).$$

Definition 3.1.

1. *Consistency.*
   An approximation operator $H(h)$, $h>0$, is consistent with (1.2) if for any $\phi \in C^\infty\left(\bar{\Omega}\right)$ and for any $x \in \bar{\Omega}$, the following holds:

$$\frac{(H(h)\,\phi)(x) - \phi(x)}{h} = |D\phi|\, G\left(\operatorname{div}\frac{\mathrm{D}\phi}{|\mathrm{D}\phi|}\right) + o_x(1) \text{ for } D\phi \neq 0. \tag{3.1}$$

   If the convergence of $o_x(1)$ is locally uniform on sets, where $D\phi \neq 0$, then $H(h)$ is said to be uniformly consistent with the PDE.
2. *Monotonicity.*
   An operator $H(h)$, $h>0$, is locally monotone if there exists $r>0$ such that for any functions $u(y), v(y) \in \mathbb{B}\left(\bar{\Omega}\right)$ with $u \geq v$ on $B(x,r)\setminus\{x\}$, the following holds:

$$H(h)\,u(y) \geq H(h)\,v(y) + o(h),$$

   where the convergence of $o(h)$ is uniform on $B(x,r)\setminus\{x\}$.
3. *Stability.*
   An approximation scheme $H(h)$ is stable if $H(h)^n u \in \mathbb{B}\left(\bar{\Omega}\right)$ for every $u \in B\left(\bar{\Omega}\right), n \in \mathbb{N}, h>0$, with a bound independent of h and n.

In this setting the result of Barles and Souganidis reads as follows.

Theorem 3.2. *Consider a monotone, stable approximation operator $H(h)$ that commutes with additions of constants (i.e., $H(h)(u+C) = H(h)\,u + C$ for all $C \in \mathbb{R}$) and is uniformly consistent with (1.2). Suppose also that*

$$\lim_{h\to 0} \frac{H(h)\,(f(|x-x_0|))(x_0)}{h} = 0 \tag{3.2}$$

for any $f \in \mathcal{F}(G)$. Then $u_h(x,nh)$ converges locally uniformly to the unique viscosity solution $u(x,t)$ of (1.2) as $nh \mapsto t$.

3.2. Properties of $\mathcal{H}$. We consider a convolution generated motion of a hypersurface in $\mathbb{R}^n$ defined by (1.4) and the corresponding evolution of an initially bounded function $g : \mathbb{R}^n \mapsto \mathbb{R}$ defined by (1.5). Consider also the initial value problem (1.3) with given G and g. We are looking for such a thresholding function F in (1.4) so that $H^m_{t/m} g(x)$ would converge (in some sense) to the unique viscosity solution of (1.3).

For example, set $F(M_1, M_2) = M_1 - \frac{1}{2}$ and $\tilde{\rho}_1(x) = \frac{1}{(4\pi)^{n/2}} e^{-x^2/4}$ to get corresponding operators $\mathcal{H}_h$ and $H(h)$ by (1.4) and (1.5). Then we get the Bence–Merriman–Osher procedure to which the main result of [15] applies, and $H(h)^n u_0$ converges locally uniformly to the unique viscosity solution of (1.3) with $G(k) = k$.

We will see that it is necessary to compute two convolutions M_1 and M_2 and use the thresholding function depending on both these values to resolve the problem when G is not linear.

Let us now consider an operator $H(h)$ defined by (1.5) with the help of an operator $\mathcal{H}_h$ with an arbitrary thresholding function (1.4). We look for requirements on F sufficient to fulfill the conditions of Theorem 3.2.

Stability. Suppose $u(x) \in \mathbb{B}(\mathbb{R}^n)$. We show that $H(h)u \in \mathbb{B}(\mathbb{R}^n)$. Intuitively, we require

$$\mathcal{H}_h\mathbb{R}^n = \mathbb{R}^n, \tag{3.3}$$

$$\mathcal{H}_h\emptyset = \emptyset, \tag{3.4}$$

and denote $A = \max|u|$. With these settings, we have $[u \le A] = \mathbb{R}^n$ and

$$-A \le H(h)u(x) = \inf\{\lambda \in \mathbb{R} : x \in \mathcal{H}_h[u \le \lambda]\} \le A.$$

It remains to find out for which F the conditions (3.3) and (3.4) are satisfied. To do this, we substitute the corresponding sets into the definition of $\mathcal{H}$:

$$\begin{aligned}
\mathcal{H}_h\mathbb{R}^n &= \{x \in \mathbb{R}^n : F(M_1\mathbb{R}^n(x,h), M_2\mathbb{R}^n(x,h)) \ge 0\} \\
&= \left\{x \in \mathbb{R}^n : F\left(\int_{\mathbb{R}^n}\rho_1 dx, \int_{\mathbb{R}^n}\rho_2 dx\right) \ge 0\right\} = \mathbb{R}^n \\
\mathcal{H}_h\emptyset &= \{x \in \mathbb{R}^n : F(M_1\emptyset(x,h), M_2\emptyset(x,h)) \ge 0\} \\
&= \{x \in \mathbb{R}^n : F(0,0) \ge 0\} = \emptyset.
\end{aligned}$$

Thus, the requirements on F become

$$\begin{aligned}
F\left(\int_{\mathbb{R}^n}\rho_1 dx, \int_{\mathbb{R}^n}\rho_2 dx\right) \ge 0, \\
F(0,0) < 0.
\end{aligned}$$

Monotonicity. Let us now show that if $\mathcal{H}_h$ satisfies the so-called inclusion principle, then H_h is monotonous.

LEMMA 3.3. *Assume, that $\mathcal{H}_h$ satisfies the inclusion principle, i.e.,*

$$\forall C_1, C_2 \subseteq \mathbb{R}^n : C_1 \subseteq C_2 \text{ we have } \mathcal{H}_h C_1 \subseteq \mathcal{H}_h C_2; \tag{3.5}$$

then H_h is monotone, that is,

$$\forall u, v \in \mathbb{C}(\mathbb{R}^n) : v \le u \text{ we have } H_h(v) \le H_h(u).$$

Proof. Suppose, there exists x_0 s.t. $H(h)u(x_0) < H(h)v(x_0)$. We denote $\lambda_1 = H(h)u(x_0)$, $\lambda_2 = H(h)v(x_0)$, and $\epsilon = \frac{\lambda_2 - \lambda_1}{2} > 0$. Since

$$\lambda_1 + \epsilon < \inf\{\lambda \in \mathbb{R} : x_0 \in \mathcal{H}_h[v \le \lambda]\},$$

we have $x_0 \notin \mathcal{H}_h[v \le \lambda_1 + \epsilon]$, but

$$\mathcal{H}_h[v \le \lambda_1 + \epsilon] \supseteq \mathcal{H}_h[u \le \lambda_1 + \epsilon].$$

Therefore $x_0 \notin \mathcal{H}_h[u \le \lambda_1 + \epsilon]$, which contradicts the definition of λ_1. ∎

Consistency. We sum up some calculations in the following lemma.

LEMMA 3.4. *Let* $\phi \in C^\infty(\mathbb{R}^n)$ $\phi(0) = 0$ *and* $D\phi(0) = (0, 0, \ldots, \beta)$. *Then the consistency of an operator* $H(h)$ *with* (1.3) *is equivalent to*

$$\gamma(h, 0) = hG(-\Delta\gamma(h, 0)) + o(h), \tag{3.6}$$

where $\Delta\gamma(h, 0) = \sum_{i=1}^{n-1} \partial^2\gamma/\partial x_i^2(0)$ *and* $x_n = \gamma(h, \acute{x})$ *is a parameterization of the surface*

$$\{x \in \mathbb{R}^n : \phi(x) = H(h)\phi(0)\}$$

near $\acute{x} = 0$.

We observe that in these settings $-\Delta\gamma(h, 0) \equiv k$ *is the mean curvature of the graph of* γ *at the point* $(0, \gamma(h, 0))$.

Proof. Without loss of generality, one can consider the consistency condition (3.1) only for ϕ as in the statement. We rewrite (3.1) in a more convenient form:

$$(H(h)\phi)(0) = h|D\phi(0)| G\left(\operatorname{div}\frac{D\phi}{|D\phi|}(0)\right) + o(h). \tag{3.7}$$

We use the equality

$$\operatorname{div}\left(\frac{D\phi}{|D\phi|}\right) = \frac{1}{|D\phi|}\sum_{i,j=1}^{n}\left(\delta_{i,j} - \frac{\phi_{x_i}\phi_{x_j}}{|D\phi|^2}\right)\phi_{x_ix_j}.$$

Since $\phi(0) = 0$ and $\phi_{x_i}(0) = \delta_{ni}\beta$,

$$\begin{aligned}\operatorname{div}\frac{D\phi}{|D\phi|}\bigg|_{x=0} &= \frac{1}{\beta}\left[\sum_{i=1}^{n}\phi_{x_ix_i}(0) - \frac{\phi_{x_n}(0)\phi_{x_n}(0)}{\beta^2}\phi_{x_nx_n}(0)\right] \\ &= \frac{1}{\beta}\Delta'\phi(0). \end{aligned} \tag{3.8}$$

Here $\Delta'\phi = \sum_{i=1}^{n-1}\phi_{x_ix_i}$. Our next step is to take small $\acute{x}$, namely $|\acute{x}| < Rh$. For such $\acute{x}$ we apply the inverse function theorem to ϕ,

$$H(h)\phi(0) = \phi(\acute{x}, \gamma(h, \acute{x})) = \phi(0) + \beta\gamma(h, 0) + O\left(h^2\right). \tag{3.9}$$

Putting (3.9) and (3.8) into (3.7) we get

$$\gamma(h, 0) = hG\left(\frac{1}{\beta}\Delta'\phi(0)\right) + o(h). \tag{3.10}$$

Furthermore, differentiating both sides of $H(h)\phi(0) = \phi(\acute{x}, \gamma(h, \acute{x}))$ gives

$$\begin{aligned}\phi_{x_i} + \phi_{x_n}\gamma_{x_i} &= 0, \\ \phi_{x_ix_j} + \phi_{x_ix_n}\gamma_{x_j} + \phi_{x_nx_j}\gamma_{x_i} + \phi_{x_nx_n}\gamma_{x_j}\gamma_{x_i} + \phi_{x_n}\gamma_{x_ix_j} &= 0\end{aligned}$$

for $j, i = 1, \ldots, n-1$. We deduce $\gamma_{x_i}(h, 0) = 0$ from the first equality and rewrite the second one for $i = j$,

$$\phi_{x_jx_j}(0) + \phi_{x_n}(0)\gamma_{x_jx_j}(h, 0) = 0.$$

After a summation over j this becomes

$$\frac{1}{\beta}\Delta'\phi(0) = -\Delta\gamma(h, 0).$$

It remains to put this relation into (3.10) to get the desired equality (3.6). ∎

3.3. The convergence result for general G. In this subsection we construct the thresholding function $F(M_1, M_2)$ and show that the corresponding convolution thresholding scheme (1.4), (1.5) converges to the viscosity solution $u(x,t)$ of (1.3),

$$H_{\frac{t}{m}}^{m} g(x) \to u(x,t) \text{ as } m \to \infty.$$

We start with $F(M_1 C(x,h), M_2 C(x,h))$, where

$$M_i C(x,h) = \int_C \rho_i(x-y)\, dy.$$

For each ρ_i we expand this integral into the power series in h (see (3.19)), i.e.,

$$M_i[\phi \le H(h)\phi(0)](0,h) = A_i + \sqrt{h} v C_i + \sqrt{h}\Delta\gamma(h,0) B_i + O\left(h^{3/2}\right), \tag{3.11}$$

where

$$A_i = \int_{\mathbb{R}^{n-1}} \int_{-\infty}^{0} \rho_i(|y|)\, dy_n d\acute{y}, \tag{3.12}$$

$$B_i = \frac{1}{2} \int_{\mathbb{R}^{n-1}} y_k^2 \rho_i(\acute{y}, 0)\, d\acute{y}, \tag{3.13}$$

$$C_i = \int_{\mathbb{R}^{n-1}} \rho_i(\acute{y}, 0)\, d\acute{y}, \tag{3.14}$$

and $i = 1, 2$. This is a system of linear algebraic equations for $\Delta\gamma(h,0)$ and v. We choose the kernels so that the determinant of this system is positive,

$$D = C_1 B_2 - C_2 B_1 > 0,$$

denote $N_i = M_i[\phi \le H(h)\phi(0)](0,h) - A_i$, and write the solution

$$v = \frac{\gamma(h,0)}{h} = \frac{1}{\sqrt{h}} \frac{N_1 B_2 - N_2 B_1}{C_1 B_2 - C_2 B_1} + O(h),$$

$$\Delta\gamma(h,0) = \frac{1}{\sqrt{h}} \frac{N_2 C_1 - N_1 C_2}{C_1 B_2 - C_2 B_1} + O(h).$$

Lemma 3.4 implies that the operator H is consistent with the PDE in (1.3) if we take

$$\begin{aligned} F(N_1, N_2) &= v - G(-\Delta\gamma(h,0)) \\ &= \frac{1}{\sqrt{h}} \frac{N_1 B_2 - N_2 B_1}{D} - G\left(\frac{1}{\sqrt{h}} \frac{N_1 C_2 - N_2 C_1}{D}\right). \end{aligned} \tag{3.15}$$

In the case of the thresholding function of one variable, the inclusion principle (3.5) holds for $\mathcal{H}$ when F is nondecreasing. In the case of two variables we require

$$\frac{\partial F}{\partial N_1} = \frac{B_2}{D} - \frac{C_2}{D} G' \ge 0, \tag{3.16}$$

$$\frac{\partial F}{\partial N_2} = -\frac{B_1}{D} + \frac{C_1}{D} G' \ge 0. \tag{3.17}$$

This implies

$$\frac{B_1}{C_1} \leq G^{'} \leq \frac{B_2}{C_2}. \tag{3.18}$$

Therefore, for awhile we restrict ourselves with G having a bounded and positive derivative. Comparing (3.14) with (3.13) one sees that it is possible to make the lower bound in (3.18) small by choosing ρ_1 with mass concentration close to the origin. The upper bound will be large if the mass of ρ_2 is concentrated relatively far from the origin.

Next, we state some auxiliary results.

LEMMA 3.5. *Suppose* (3.16) *and* (3.17) *hold and* $\mathcal{H}$ *is defined by* (1.4)*; then for all* $h \in \mathbb{R}_+$,

1. $\mathcal{H}(h)(\mathbb{R}^n) = \mathbb{R}^n$, $\mathcal{H}(h)(\emptyset) = \emptyset$,
2. *for all* $a, b \in \mathbb{X} : a \subseteq b \Rightarrow \mathcal{H}(h) a \subseteq \mathcal{H}(h) b$.

Proof.

1. It is enough to show that $F\left(M_1(\mathbb{R}^n)(x,h), M_2(\mathbb{R}^n)\left(x,h\right)\right) \geq 0$, and $F\big(M_1(\emptyset)\left(x,h\right), M_2(\emptyset)(x,h)\big) < 0$. First we observe that $F(A_1, A_2) = 0$, $M_i(\mathbb{R}^n)(x,h) \geq A_i$, and $M_i(\emptyset)(x,h) = 0 < A_i$. This, together with $\frac{\partial F}{\partial N_i} > 0$, gives the desired inequalities.
2. Since $M_i(b) \geq M_i(a)$, $F(M_1(b), M_2(b)) \geq F(M_1(a), M_2(a))$, therefore $[F(M_1(a), M_2(a)) \geq 0] \subseteq [F(M_1(b), M_2(b)) \geq 0]$, which is equivalent to $\mathcal{H}(h) a \subseteq \mathcal{H}(h) b$. □

PROPOSITION 3.6. *Define* H *by* (1.5) *and* $\mathcal{H}$ *by* (1.4)*; then for each* $h > 0$ *and* $u \in \mathbb{B}(\mathbb{R}^n)$ *one has* $H(h) u \in \mathbb{B}(\mathbb{R}^n)$.

Proof. Without loss of generality we assume that $S_1 \leq u(x) \leq S_2$ for some $S_1, S_2 \in \mathbb{R}$. From

$$\forall h \in \mathbb{R}_+ \quad \mathcal{H}(h)(\mathbb{R}^n) = \mathbb{R}^n \text{and } \mathcal{H}(h)(\emptyset) = \emptyset$$

it follows that $x \in \mathcal{H}(h)[u \leq S_2]$ and $x \notin \mathcal{H}(h)[u \leq S_1]$. Therefore, we see that

$$S_1 \leq H(h) u(x) = \inf\{\lambda \in \mathbb{R} : x \in \mathcal{H}(h)[u \leq \lambda]\} \leq S_2. \qquad \square$$

With the results above, we are ready to state the convergence of the approximations $H(t/m)^m g$ to the unique viscosity solution of (1.3).

THEOREM 3.7. *Let* $H(h)$ *be defined by*

$$[H(h) u](x) = \sup\{\lambda \in \mathbb{R} : x \in \mathcal{H}_h[u \geq \lambda]\}$$

with

$$\mathcal{H}_h C = \{x \in \mathbb{R}^n : F(M_1(C)(x,h), M_2(C)(x,h)) \geq 0\},$$

where

$$F(N_1, N_2) = \frac{1}{\sqrt{h}} \frac{N_1 B_2 - N_2 B_1}{D} - G\left(\frac{1}{\sqrt{h}} \frac{N_2 C_1 - N_1 C_2}{D}\right),$$

and where $\tilde{\rho}_1, \tilde{\rho}_2$ *have compact support and* G *is continuous nondecreasing satisfying* (3.18). *Then*

$$H_{t/m}^m g(x) \to u(x,t)$$

locally uniformly when $m \to \infty$. *Here* $u(x,t)$ *is the unique viscosity solution of* (1.3) *with* G *satisfying* (3.18).

Proof. Our aim is to show here that the operator $H(h)$ satisfies the conditions of Theorem 3.2. The *monotonicity* of H_h is ensured by Lemmas 3.3 and 3.5.

The *stability* of H is exactly the result of Proposition 3.6: $H(h) u \in \mathbb{B}\left(\bar{\Omega}\right)$.

Another requirement in Theorem 3.2 is that $H(h)$ must commute with the addition of constants, i.e.,

$$\forall a \in \mathbb{R} \quad H(h)(u(x) + a) = H(h) u(x) + a.$$

This follows from the very definition of $H(h)$:

$$\begin{aligned} H(h)(u(x) + a) &= \inf\{\lambda \in \mathbb{R} : x \in \mathcal{H}(h)[u(x) + a \le \lambda]\} \\ &= \inf\{\beta + a \in \mathbb{R} : x \in \mathcal{H}(h)[u(x) \le \beta]\} = H(h) u(x) + a. \end{aligned}$$

The operator $H(h)$ has to fulfill (3.2) as well. The limit we are interested in is

$$\lim_{h \to 0} \frac{H(h) u(x_0)}{h} = 0$$

for u of the form $u(x) = f(|x - x_0|)$, where $f \in C^2([0,\infty))$ with $f(0) = f'(0) = f''(0) = 0$ and $f''(r) > 0$ for $r > 0$.

It is enough to show that this is true for $x_0 = 0$. First, we observe that $\mathcal{H}_h^{-1}[\{0\}] = \{u \le \lambda_1\}$, where $\lambda_1 = H(h) u(0)$. Since both ρ_1 and ρ_2 have compact support, we can be sure that there exists R s.t. $\left\{|x| \le R\sqrt{h}\right\} \supseteq \mathcal{H}_h^{-1}[\{0\}]$. Now we observe, that $\left\{|x| \le R\sqrt{h}\right\} = \{u \le \lambda_2\}$ for some $\lambda_2 > \lambda_1$. From the latter equality we deduce $\lambda_2 = O\left(h^{3/2}\right)$ and conclude with

$$\lim_{h \to 0} \frac{H(h) u(x_0)}{h} \le \lim_{h \to 0} \frac{O\left(h^{3/2}\right)}{h} = 0.$$

To show that our approximation operator is *consistent* with the PDE, we use Lemma 3.4. It is enough to prove the following:

$$\gamma(h,0) = hG(-\Delta\gamma(h,0)) + o(h),$$

where $x_n = \gamma(h, \acute{x})$ is a parameterization of the surface

$$\{x \in \mathbb{R}^n : u(x) = H(h) u(0)\}$$

near $\acute{x} = 0$. To show this, we use the fact that

$$F(M_1[u \le \mu], M_2[u \le \mu])|_{x=0} = 0.$$

We begin by writing the expressions for M_i in detail:

$$M_i = \left(\chi_{[u\le\mu]} \star \frac{1}{h^{n/2}}\rho_i\left(\frac{|\cdot|}{\sqrt{h}}\right)\right)(0) = \int_{\mathbb{R}^n} \chi_{[u\le\mu]}(y)\,\frac{1}{h^{n/2}}\rho_i\left(\frac{|y|}{\sqrt{h}}\right)dy$$

$$= \int_{\mathbb{R}^{n-1}}\int_{-\infty}^{\gamma(h,\acute{y})} \frac{1}{h^{n/2}}\rho_i\left(\frac{|y|}{\sqrt{h}}\right)dy_n d\acute{y} = A_i + \int_{\mathbb{R}^{n-1}}\int_0^{(1/\sqrt{h})\gamma\left(h,\sqrt{h}\acute{y}\right)} \rho_i(|y|)\,dy_n d\acute{y}.$$

Here A_i is given by (3.12). Expanding $\gamma\left(h,\sqrt{h}\acute{y}\right)$ in the Taylor series with respect to the spatial variables (keeping h as a parameter) we get

$$\frac{1}{\sqrt{h}}\gamma\left(h,\sqrt{h}\acute{y}\right) = \sqrt{h}\frac{\gamma(h,0)}{h} + \frac{\sqrt{h}}{2}\sum_{i,j=1}^{n-1}\gamma_{y_iy_j}(h,0)\,y_iy_j + \frac{h}{6}\sum_{i,j,l=1}^{n-1}\gamma_{y_iy_jy_l}(h,0)\,y_iy_jy_l + O\left(h^{3/2}\acute{y}^4\right).$$

Observing that $\gamma(h,0) = O\left(\sqrt{h}\right)$, we denote $\frac{\gamma(h,0)}{h} = v$. The expression for M_i becomes

$$M_i = A_i + \int_{\mathbb{R}^{n-1}}\rho_i(\acute{y},0)\left[\sqrt{h}v + \frac{\sqrt{h}}{2}\sum_{i,j=1}^{n-1}\gamma_{y_iy_j}(h,0)\,y_iy_j + O\left(h^{3/2}\acute{y}^4\right)\right]dy_n d\acute{y}$$

$$= A_i + \sqrt{h}vC_i + \sqrt{h}\Delta\gamma(h,0)\,B_i + O\left(h^{3/2}\right), \tag{3.19}$$

where we have used the fact that $\rho_i(\acute{x},x_n)$ is smooth and radially symmetric, in particular,

$$\frac{\partial\rho_i}{\partial x_n}(\acute{x},0) = 0.$$

The constants B_i, C_i depend only on ρ_i and are given by (3.13) and (3.14).

Remark 1. At this point it is easy to see that a scheme with a thresholding depending only on one variable can be consistent with the PDE (1.2) only in the case of linear G. The thresholding condition becomes

$$F\left(A + \sqrt{h}vC + \sqrt{h}\Delta\gamma(h,0)\,B + O\left(h^{3/2}\right)\right) = 0.$$

As was required by the inclusion principle, the function F is nondecreasing. This implies

$$A + \sqrt{h}vC + \sqrt{h}\Delta\gamma(h,0)\,B + O\left(h^{3/2}\right) = a,$$

where a is the unique solution of $F(a) = 0$. Thus

$$v = \frac{\gamma(h,0)}{h} = -\frac{B}{C}\Delta\gamma(h,0) - \frac{a-A}{\sqrt{h}C} + o\left(\sqrt{h}\right).$$

Comparing this relationship with the one in Lemma 3.4, we see that the only G's we can resolve by thresholding depending on one variable are the linear ones: $G(k) = \text{const}\cdot k + \text{const}$.

Let us denote here $k = \Delta\gamma(h,0)$.

Now we can express v and k in terms of M_i and constants A_i, B_i, and C_i :

$$v = \frac{1}{\sqrt{h}} \frac{N_1 B_2 - N_2 B_1}{C_1 B_2 - C_2 B_1} + O(h),$$
$$k = \frac{1}{\sqrt{h}} \frac{N_2 C_1 - N_1 C_2}{C_1 B_2 - C_2 B_1} + O(h).$$

Since $F(M_1, M_2) = v - G(-k) = 0$, we have

$$\gamma(h,0) = hG(-\Delta\gamma(h,0)) + o(h).$$

Remark 2. As was already mentioned above, convolution kernels $\tilde{\rho}_i$ can also be taken with unbounded support. For example, the exponential decay for large arguments is sufficient in order for Theorem 3.7 to hold.

The requirement (3.18) is quite restrictive. Our next result shows that it is enough to take G_ϵ satisfying (3.18) and uniformly close to G in order to approximate the solutions of (1.3).

PROPOSITION 3.8. *Suppose G_ϵ, G are continuous and $G_\epsilon \to G$ uniformly on $\mathbb{R}$ as $\epsilon \to 0$. Then $\mathcal{F}(G) = \mathcal{F}(G_\epsilon)$.*

Proof. Suppose $f \in \mathcal{F}(G)$. It means that $f(0) = f'(0) = f''(0)$, $f(r) > 0$ for $r > 0$, and

$$\lim_{p\to 0} f'(p)\, G\left(\frac{1}{p}\right) = \lim_{p\to 0} f'(p)\, G\left(\frac{-1}{p}\right) = 0.$$

Since $G_\epsilon \to G$ uniformly, $G(k) = G_\epsilon(k) + o_\epsilon(1)\,\alpha(k)$, where $\alpha \in \mathbb{B}(\mathbb{R})$. We write

$$\begin{aligned} 0 &= \lim_{p\mapsto 0} f'(p)\, G\left(\frac{1}{p}\right) \\ &= \lim_{p\to 0} f'(p)\left(G_\epsilon\left(\frac{1}{p}\right) + o_\epsilon(1)\,\alpha\left(\frac{1}{p}\right)\right) = \lim_{p\to 0} f'(p)\, G_\epsilon\left(\frac{1}{p}\right) \end{aligned}$$

to see that $f \in \mathcal{F}(G_\epsilon)$.

The proof of the reverse inclusion is analogous. □

LEMMA 3.9. *Suppose G_ϵ, G are nondecreasing continuous and $G_\epsilon \to G$ uniformly on $\mathbb{R}$ as $\epsilon \to 0$. Suppose also that for each $\epsilon > 0$ the operator H_ϵ is monotone, stable, commuting with additions of constants, and consistent with*

$$\frac{\partial u_\epsilon}{\partial t} = |Du_\epsilon|\, G_\epsilon\left(\operatorname{div}\frac{\mathrm{Du}_\epsilon}{|\mathrm{Du}_\epsilon|}\right). \tag{3.20}$$

Additionally, let the following limit hold:

$$\lim_{h\to} \frac{H_h(h)(f(|x - x_0|))(x_0)}{h} = 0 \tag{3.21}$$

for each $f \in \mathcal{F}(G)$. Then

$$H^m_{t/m}(t/m)\, u_0(x) \to u(x,t)$$

locally uniformly as $m \to \infty$, where $u(x,t)$ is the unique viscosity solution of (1.3).

Proof. We show here that the operator $H_h(h)$ satisfies the conditions of Theorem 3.2. This operator commutes with additions of constants and satisfies limit (3.21) by the assumption. Since the operator H_ϵ is stable for all $\epsilon > 0$, it is particularly stable for $\epsilon = h$ for each $h > 0$.

Since the operator H_ϵ is monotonic for all $\epsilon > 0$, it is particularly monotonic for $\epsilon = h$ for each $h > 0$.

We have to show consistency; i.e., for each $\phi \in C^\infty(\mathbb{R}^n)$ at each point where $|D\phi| \neq 0$,

$$H_h(h)\phi(x) - \phi(x) = h|D\phi(x)| G\left(\operatorname{div}\frac{D\phi(x)}{|D\phi(x)|}\right) + o(h) \tag{3.22}$$

has to hold. Since the operator H_ϵ is consistent with (3.20) and $G_h(k) = G(k) + o_h(1)\alpha(k)$ for some $\alpha \in \mathbb{B}(\mathbb{R})$, we write

$$\begin{aligned} H_h(h)\phi(x) - \phi(x) &= h|D\phi(x)| G_h\left(\operatorname{div}\frac{D\phi(x)}{|D\phi(x)|}\right) + o(h) \\ = h|D\phi(x)| \left(G\left(\operatorname{div}\frac{D\phi(x)}{|D\phi(x)|}\right) + o_h(1)\alpha\left(\operatorname{div}\frac{D\phi(x)}{|D\phi(x)|}\right)\right) &+ o(h) \\ &= h|D\phi(x)| G\left(\operatorname{div}\frac{D\phi(x)}{|D\phi(x)|}\right) + o(h); \end{aligned}$$

here $o_h(1) \to 0$ as $h \to 0$. ∎

THEOREM 3.10. *Consider a convolution-thresholding scheme*

$$\begin{aligned} H_\epsilon(h)u(x) &= \inf\{\lambda \in \mathbb{R} : x \in \mathcal{H}_\epsilon(h)[u \leq \lambda]\}, \\ \mathcal{H}_\epsilon(h)C &= \{x \in \mathbb{R}^n : F_\epsilon(M_1(C)(x,h), M_2(C)(x,h)) \geq 0\}, \end{aligned}$$

where the thresholding function $F_\epsilon(M_1, M_2)$ is chosen so that the scheme is monotone and consistent with (3.20) *and the convolution kernels have compact support. If $G_\epsilon \to G$ uniformly, then*

$$H^m_{t/m}(t/m)u_0(x) \to u(x,t)$$

locally uniformly as $m \to \infty$, where $u(x,t)$ is the unique viscosity solution of (1.3).

Proof. The convergence follows form Lemma 3.9 if we show that the limit (3.21) holds. Let us set $x_0 = 0$; then the set $[f(|x|) \leq \lambda]$ is a ball centered at the origin with radius $O(\lambda^{1/3})$. We denote $H_h(h)f(0) = \lambda_1$. Observe that λ_1 can be characterized as a number for which $\mathcal{H}_h(h)[f \leq \lambda_1] = \{0\}$. Since we know that $F_h(A_1, A_2) > 0$, the radius of $[f \leq \lambda_1]$ must be less than or equal to the radius of the greatest support of the kernel: $O(\lambda_1^{1/3}) \leq R\sqrt{h}$. From this inequality we deduce $H_h(h)f(0) = \lambda_1 \leq O(h^{3/2})$. This establishes the desired limit (3.21). ∎

Let us now consider the particular interesting case with $G(k) = k|k|^{\alpha-1}$ with $\alpha > 1$. We set

$$G_m(k) = \begin{cases} (1-\alpha)m^\alpha + \alpha m^{\alpha-1}k & \text{for } k < -n, \\ m^{1-\alpha}k & \text{for } |k| < 1/n, \\ -(1-\alpha)m^\alpha + \alpha m^{\alpha-1}k & \text{for } k > n, \\ k|k|^{\alpha-1} & \text{elsewhere.} \end{cases}$$

G_m is continuous, increasing, and its derivative is bounded from below and above: $m^{1-\alpha} \le G_m' \le \alpha m^{\alpha-1}$. Moreover, $G_m \to k\,|k|^{\alpha-1}$ locally uniformly as $m \to \infty$. Using Theorem 2.4 it is easy to show the following.

THEOREM 3.11. *Let u_m be the viscosity solution of*

$$\frac{\partial u_m}{\partial t} = |Du_m|\, G_m\left(\operatorname{div}\frac{Du_m}{|Du_m|}\right) \text{ in } \mathcal{O},$$

where G_m is defined above. Then $u_m \to u$ locally uniformly as $m \to \infty$, where u is the viscosity solution of (1.2) *in $\mathcal{O}$, with $G(k) = k\,|k|^{\alpha-1}$, $\alpha > 1$.*

Proof. First we establish the inclusion $\mathcal{F}(G) \subset \mathcal{F}(G_m)$. Take $f \in \mathcal{F}(G)$. By the definition of $\mathcal{F}(G)$, $f'(x) = o(x^\alpha)$. This immediately gives

$$\lim_{p\to 0} f'(p)\, G_m(1/p) = \lim_{p\to 0} f'(p)/p = 0,$$

since $\alpha > 1$. We observe also that the remaining conditions of Theorem 2.4 are satisfied. Hence a subsolution and a supersolution u^+ and u_+ can be constructed by means of (2.2) and (2.3). Since the equation has the strong comparison property (see [12]), $u^+ = u_+$ and the result follows. □

Remark 3. In a more general case when $G(k) = O(k^\alpha)$, $\alpha > 1$, one can pick a sequence of increasing functions with derivative bounded below and above and apply Theorem 2.4 to get a result similar to Theorem 3.11.

4. Numerical implementation. This section is devoted to a description of our numerical implementations of the convolution-thresholding scheme developed in section 3.

Given a compact set $C \subset \mathbb{R}^n$, we fix convolution kernels ρ_1, ρ_2 and the time step h and approximate C_t at a time moment $t = mh$ by $(\mathcal{H}(h))^m C$. The algorithm of computations consists of the following steps:

1. Compute convolutions and the thresholding function

$$M_i C(x,h) = \int_{\mathbb{R}^n} \chi_C(y)\, \rho_i(x-y)\, dy, \qquad i = 1,2, \tag{4.1}$$

$$F(x,h) = F(M_1 C(x,h), M_2 C(x,h)). \tag{4.2}$$

2. Find the evolved set $\mathcal{H}(h)C = \{x \in \mathbb{R}^n : F(x,h) \ge 0\}$.
3. Repeat the procedure with the evolved set to get $\mathcal{H}^2(h)C$ and so on.

We used two different algorithms for the calculation of the convolution step, which constitutes the main computational part of the algorithm.

4.1. Spatial discretization. We assume that initially the surface is closed and contained in a unit cube. The surface under consideration is always an isosurface of some function. In our implementation we use a modification of the so-called marching cubes algorithm for extracting an isosurface. The algorithm was originally proposed in [24] and was first applied for the mean curvature flow calculations in [30]. The algorithm creates an adaptive spatial discretization of C (see Figure 4.1).

By our implementation, we significantly reduce the number of grid points. In addition, the accurate piecewise polynomial approximation of the ∂C can be arranged.

4.2. Spectral method. One can use a Fourier series to calculate the convolutions (4.1). Numerical aspects of this approach have been presented by Ruuth in [30].

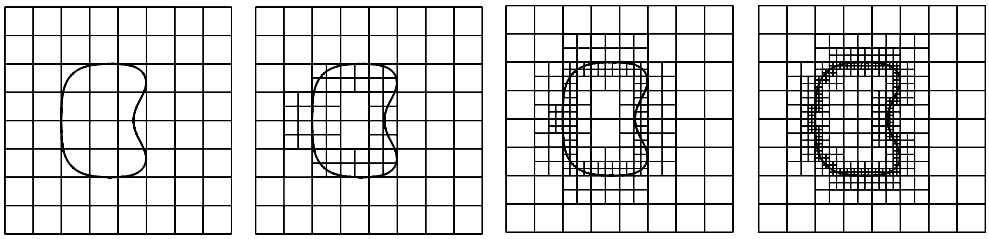


FIG. 4.1. *On the spatial discretization.*

In order to compute Fourier coefficients of χ_C given on a nonuniform grid, the unequally spaced approximate fast Fourier transform algorithm [6] is used. The numerical cost of this transform algorithm combined with the marching cubes procedure is (see [30]) $O\big(m^n N_p + N_f^n \log(N_f)\big)$, where m is a constant depending on a desired accuracy in the calculation of the Fourier coefficients (in case $m = 23$, the accuracy is comparable with the machine truncation error), N_f is a number of the Fourier modes along each axis, and N_p is the number of nodes in the grid.

4.3. Direct method. If ρ_1 and ρ_2 are simple enough and have compact support, their convolutions with χ_C can be calculated explicitly. Let us choose

$$\tilde{\rho}_1(x) = \begin{cases} \frac{1}{|\mathcal{B}_1|} & \text{if } x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{\rho}_2(x) = \frac{1}{\alpha^n} \tilde{\rho}_1\left(\frac{x}{\alpha}\right),$$

where $|\mathcal{B}_1|$ is the Lebesgue measure of a unit ball in $\mathbb{R}^n$ and $\alpha \in \mathbb{R}_+$, $\alpha < 1$. In this case, convolution values (4.1) are proportional to the measure of the intersection of C with a ball of radius proportional to $\sqrt{h}$ centered at the point x.

We present expressions for the thresholding function $F(M_1, M_2)$ in the case $n = 2$:

$$F(M_1, M_2) = v - G(k), \text{ where}$$

$$v = \frac{\pi\alpha(2\alpha M_1 - 2M_2 - \alpha + 1)}{4\sqrt{h}(\alpha^2 - 1)},$$

$$k = \frac{-3\pi(2M_1 - 2\alpha M_2 + \alpha - 1)}{2\sqrt{h}(\alpha^2 - 1)}.$$

In this case convolutions M_1 and M_2 can be calculated as follows. We represent C as a disjoint union of squares and triangles (or cubes in tetrahedron in case $n = 3$) using the marching cubes method and calculate the area (volume) of intersection of the ball (supp ρ) with each square and triangle. The numerical cost of each step of the evolution can be estimated by $O(N_p * N_i + N_p)$, where N_i is the number of points inside the ball of radius $\sqrt{h}$ with the center at some grid point. When h is large, the accuracy of the method is low; therefore one can take less grid points. Thus, N_i is entirely determined by the desired accuracy.

4.4. Computed examples. In the case of the mean curvature curve evolution in $\mathbb{R}^2$, the accuracy of calculations can be monitored with the help of the Von Neumann–Mullins parabolic law. It asserts that $dS/dt = -2\pi$, where S is the area enclosed by the curve.

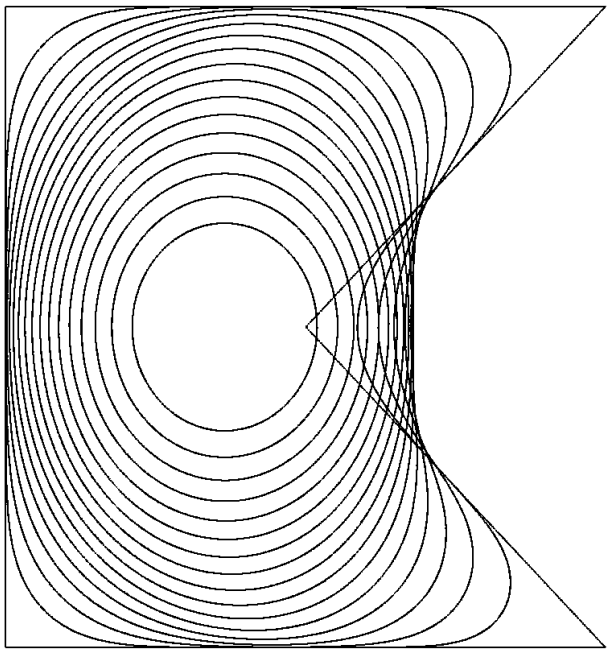


FIG. 4.2. *The mean curvature evolution of a nonsmooth, nonconvex curve.*

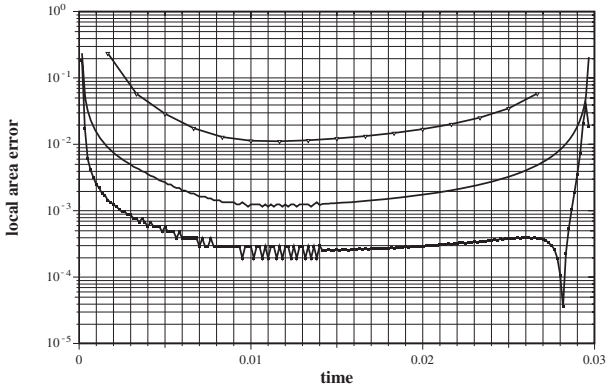


FIG. 4.3. *Local area error dependence on time. The first order method with time step* 1/600*—the line with triangle markers; the first order method with time step* 1/6000*—the thin line; the second order method with time step* 1/6000*—the line with square markers.*

Consider a nonconvex, nonsmooth initial curve, depicted in Figure 4.2. The mean curvature evolution of this curve was calculated using the direct method with time step values $dt = 1/600$ and $1/6000$. The shape of the curve is plotted in Figure 4.2 for times $t = 1/600,\ 2/600, \ldots$ when calculated with the fine time step. The comparison between local relative errors

$$e_i = \frac{|S_i - S_{i+1} - 2\pi dt|}{2\pi dt} \tag{4.3}$$

for calculations with different time steps is seen in Figure 4.3. One can observe that the error indeed depends linearly on the time step.

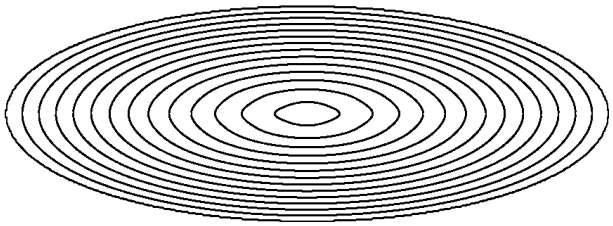


FIG. 4.4. *The evolution $v = k^{1/3}$ of an ellipse.*

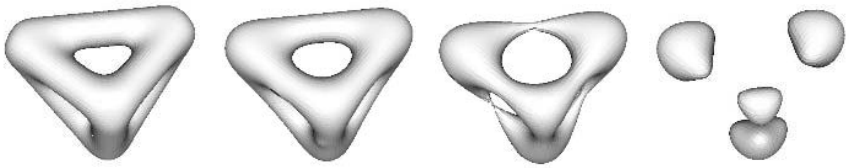


FIG. 4.5. *Computed mean curvature evolution.*

The evolution with the velocity $v = k^{1/3}$ is depicted in Figure 4.4. In this case the flow is affine invariant [1]; hence the eccentricity e of the evolving ellipse remains constant. In this particular example, the curvature is bounded from above and below by some positive constants for some evolution time. This means that we never use the parts of $G(k) = k^{1/3}$, where its derivative is too large or too small. This allows us to apply the thresholding procedure without any approximation of G.

In Figures 4.5 and 4.6 computed three-dimensional evolution of a nonconvex surface is represented for curvature flow and for a flow with velocity $v = G(k)$, as in Figure 4.7 with ~ 200000 triangles approximating the surface.

4.5. On the higher order schemes for the mean curvature motion. Let us now look at approximations to the mean curvature evolution. It is easy to see that if the surface is smooth, the Bence–Merriman–Osher method gives the first order approximation in time for a curvature flow. A higher order scheme by an extrapolation argument in time was proposed by Ruuth in [30]. We propose here higher order approximations to the mean curvature evolution using some properties of functions M_i.

We rewrite the equations (3.11) and keep an additional term of order $h^{3/2}$ in each equation with a kernel-dependent multiplier E_i to get the error term of order $h^{5/2}$. Considering two equations we get the relation

$$E_2N_1 - E_1N_2 = \sqrt{h}[(E_2C_1 - E_1C_2)v + (E_2B_1 - E_1B_2)\gamma''(h,0)] + O\big(h^{5/2}\big). \tag{4.4}$$

This relationship motivates us to take the thresholding function $F(N_1, N_2) = E_2N_1 - E_1N_2$ to approximate the mean curvature evolution with the second order accuracy for smooth curves. However, this thresholding function does not simultaneously satisfy (3.16) and (3.17) and, therefore, the stability of the numerical scheme is not guaranteed by the previous argument.

The calculations with the above thresholding function were performed. No sign of instability was observed in the numerical experiments and, as one can see in Figure 4.3, the accuracy was increased by approximately one order. This increase agrees with the construction (4.4).

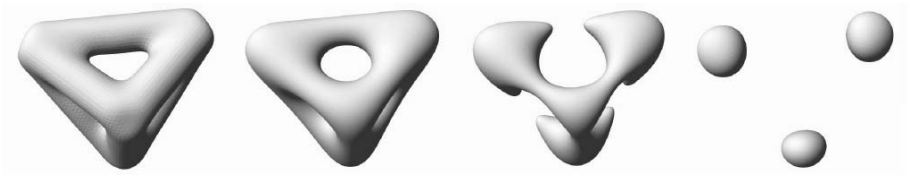


FIG. 4.6. *Computed generalized mean curvature evolution.*

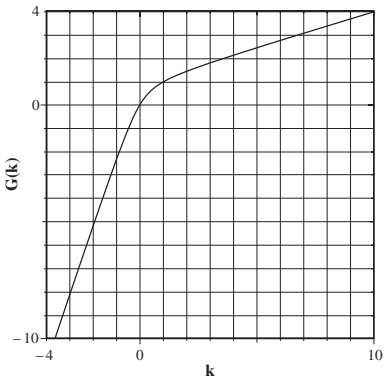


FIG. 4.7. *Function $G(k)$ used in the computation.*

Acknowledgments. A part of this work was completed during our visit to the University of South Carolina. We are grateful to Professor Björn Jawerth for the opportunity to work there and for fruitful discussions. We would also like to thank the anonymous referees for pointing out necessary corrections in the formulation of Theorem 3.2, for clarifying some key points in the text, and for suggesting additional references.

REFERENCES

[1] L. ALVAREZ, F. GUICHARD, P.-L. LIONS, AND J.-M. MOREL, *Axioms and fundamental equations of image processing*, Arch. Rational Mech. Anal., 123 (1993), pp. 199–257.

[2] S. ANGENENT, T. ILMANEN, AND D. L. CHOPP, *A computed example of nonuniqueness of mean curvature flow in* $\mathbf{R}^3$, Comm. Partial Differential Equations, 20 (1995), pp. 1937–1958.

[3] G. BARLES AND C. GEORGELIN, *A simple proof of convergence for an approximation scheme for computing motions by mean curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 484–500.

[4] G. BARLES, H. M. SONER, AND P. E. SOUGANIDIS, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.

[5] G. BARLES AND P. E. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptotic Anal., 4 (1991), pp. 271–283.

[6] G. BEYLKIN, *On the fast Fourier transform of functions with singularities*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 363–381.

[7] K. A. BRAKKE, *The Motion of a Surface by Its Mean Curvature*, Princeton University Press, Princeton, NJ, 1978.

[8] L. BRONSARD AND R. V. KOHN, *Motion by mean curvature as the singular limit of Ginzburg-Landau dynamics*, J. Differential Equations, 90 (1991), pp. 211–237.

[9] F. CAO, *Partial differential equations and mathematical morphology*, J. Math. Pures Appl. (9), 77 (1998), pp. 909–941.

[10] F. CATTÉ, F. DIBOS, AND G. KOEPFLER, *A morphological scheme for mean curvature motion and applications to anisotropic diffusion and motion of level sets*, SIAM J. Numer. Anal., 32 (1995), pp. 1895–1909.

[11] Y. G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 33 (1991), pp. 749–786.
[12] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. AMS (N.S.), 27 (1992), pp. 1–67.
[13] M. G. Crandall and P.-L. Lions, *Convergent difference schemes for nonlinear parabolic equations and mean curvature motion*, Numer. Math., 75 (1996), pp. 17–41.
[14] E. De Giorgi, *Conjectures on limits of some quasilinear parabolic equations and flow by mean curvature*, in Partial Differential Equations and Related Subjects (Trento, 1990), Longman Scientific and Technical, Harlow, 1992, pp. 85–95.
[15] L. C. Evans, *Convergence of an algorithm for mean curvature motion*, Indiana Univ. Math. J., 42 (1993), pp. 533–557.
[16] L. C. Evans, H. M. Soner, and P. E. Souganidis, *Phase transitions and generalized motion by mean curvature*, Comm. Pure Appl. Math., 45 (1992), pp. 1097–1123.
[17] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature.* I, J. Differential Geom., 33 (1991), pp. 635–681.
[18] F. Guichard and J.-M. Morel, *Partial differential equations and image iterative filtering*, in The State of the Art in Numerical Analysis (York, 1996), Oxford University Press, New York, 1997, pp. 525–562.
[19] T. Ilmanen, *Convergence of the Allen-Cahn equation to Brakke's motion by mean curvature*, J. Differential Geom., 38 (1993), pp. 417–461.
[20] H. Ishii, *A generalization of the Bence, Merriman and Osher algorithm for motion by mean curvature*, in Curvature Flows and Related Topics (Levico, 1994), Gakkōtosho, Tokyo, 1995, pp. 111–127.
[21] H. Ishii, G. E. Pires, and P. E. Souganidis, *Threshold dynamics type approximation schemes for propagating fronts*, J. Math. Soc. Japan, 51 (1999), pp. 267–308.
[22] H. Ishii and P. Souganidis, *Generalized motion of noncompact hypersurfaces with velocity having arbitrary growth on the curvature tensor*, Tohoku Math. J. (2), 47 (1995), pp. 227–250.
[23] F. Leoni, *Convergence of an approximation scheme for curvature-dependent motions of sets*, SIAM J. Numer. Anal., 39 (2001), pp. 1115–1131.
[24] W. Lorensen and H. Cline, *Marching cubes: A high resolution* 3*d surface construction algorithm*, ACM Comput. Graphics, 21 (1987), pp. 163–170.
[25] B. Merriman, J. K. Bence, and S. J. Osher, *Motion of multiple junctions: A level set approach*, J. Comput. Phys., 112 (1994), pp. 334–363.
[26] T. Ohta, D. Jasnow, and K. Kawasaki, *Universal scaling in the motion of random interfaces*, Phys. Rev. Lett., 47 (1982), pp. 1223–1226.
[27] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[28] D. Pasquignon, *Approximation of viscosity solution by morphological filters*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 335–359 (electronic).
[29] J. Rubinstein, P. Sternberg, and J. B. Keller, *Fast reaction, slow diffusion, and curve shortening*, SIAM J. Appl. Math., 49 (1989), pp. 116–133.
[30] S. J. Ruuth, *Efficient algorithms for diffusion-generated motion by mean curvature*, J. Comput. Phys., 144 (1998), pp. 603–625.
[31] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, 2nd ed., Cambridge University Press, Cambridge, UK, 1999.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2671–2687

A SMOOTH TRANSITION MODEL BETWEEN KINETIC AND DIFFUSION EQUATIONS*

PIERRE DEGOND† AND SHI JIN‡

Abstract. This paper presents a model which provides a smooth transition between a kinetic and a diffusion domain. The idea is to use a buffer zone, in which both diffusion and kinetic equations will be solved. The solution of the original kinetic equation will be recovered as the *sum* of the solutions of these two equations. We use an artificial connecting function which makes the equation on each domain degenerate at the end of the buffer zone. Thus no boundary condition is needed at the transition point. This model avoids the delicate issue of finding the interface condition or iteration in a typical domain decomposition method that couples a kinetic equation with hydrodynamic equations. A new asymptotic-preserving method for this model is introduced, and numerical examples are used to validate this new model and the new numerical method.

Key words. kinetic-fluid coupling, transport equation, diffusion approximation, asymptotic preserving schemes

AMS subject classifications. 82B40, 82B80, 82C40, 82C70, 82C80, 76R50

DOI. 10.1137/S0036142903430414

1. Introduction.

The collision transition rate in a kinetic transport process is often position dependent and varies from order unity in certain parts of the domain to an order of magnitude much smaller in other parts of the domain. For instance, in radiative transfer, the transition from a transparent to an opaque medium involves a change of collision rate by several orders of magnitude. Similarly, in stellar astrophysics, the magnitude of the photon transition rate can change by decades from the core of a star to its surface. When the collision rate is large, the diffusion equation is valid and much more efficient to solve numerically. In the domain where the collision rate is small, solving the more expensive kinetic transport equation in the phase space is necessary. Although one can solve the transport equation in the entire domain, to reduce the computational cost it is more advantageous to use a domain decomposition method that couples the diffusion equation with the transport equation.

Domain decomposition methods matching kinetic with hydrodynamic or diffusion models have received much attention in the past 15 years. Some methods have been proposed in [3], [7], [14], [15], [18], [19], [25], [26], [27], [28], [29], [33], [34], [37]. Typically a domain decomposition is done by an iteration procedure at each time step in which the diffusion and transport equations are solved alternately until convergence of the successive approximation is reached, or through an interface condition which provides the boundary conditions for each subdomain [18].

Other strategies include hybrid strategies, in which two equations are solved simultaneously, such as a fluid equation for the equilibrium part of the distribution function and a kinetic equation for perturbation to the equilibrium part. Recently, hybrid methods have been derived by a domain decomposition method *in velocity*

*Received by the editors June 24, 2003; accepted for publication (in revised form) June 2, 2004; published electronically April 19, 2005. This work was supported by the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282, and by NSF grant DMS0196106.
http://www.siam.org/journals/sinum/42-6/43041.html
†MIP, UMR 5640 (CNRS-UPS-INSA), Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex, France (degond@mip.ups-tlse.fr).
‡Department of Mathematics, University of Wisconsin, Madison, WI 53706 (jin@math.wisc.edu).

space [11], [12], [13]. These methods bear similarities with the δf method developed by plasma physicists [8].

In this paper we present a new approach to the domain decomposition method using a buffer zone, in which both diffusion and kinetic equations will be solved. The solution of the original transport equation will be recovered as the *sum* of the solutions of these two equations. In this way, our strategy departs from strategies based on domain decomposition with overlap, in which *each* of the models represents the *full* solution. Unlike a typical domain decomposition, where an interface condition has to be worked out in order to provide the boundary condition for each decomposed domain [18], we use an artificial connecting function which makes the equation on each domain degenerate at the end of the buffer zone; thus no boundary condition is needed at the transition point. Thus the delicate issue of finding the interface condition is completely avoided, and this method will not require any iteration at any given time step to match the solution of the two subdomains.

The paper is organized as follows. In the next section, we introduce the coupling technique and carry out some elementary analysis on its properties. A new asymptotic-preserving numerical scheme for this coupling model is derived in section 3. In section 5, we present numerical experiments to validate this new model and the numerical method introduced and state our conclusions.

2. The coupling methodology.

2.1. The transport equation and its diffusion limit. We present the method on a simple kinetic equation, the one-group transport equation in slab geometry [10]. Let $f(x,\mu,t)$ represent the particle phase-space density, where $x \in \mathbb{R}$ is the (one-dimensional) position variable, $\mu \in [-1,1]$ is the cosine of the angle between the velocity and the x-axis, and t is the time. In this model, the magnitude of the particle velocities are equal and normalized to 1. Then, the transport equation is

$$\partial_t f + \mu \partial_x f = Q(f)\,, \tag{2.1}$$

$$Q(f) = \int_{-1}^{1} S(x,\mu,\mu')(f(x,\mu',t) - f(x,\mu,t))\,d\mu'\,. \tag{2.2}$$

The left-hand side of (2.1) describes the motion of the particles along the x-axis with velocity μ while the operator Q takes into account the particle interactions with the medium. $S(x,\mu,\mu')$ is the collision transition rate from μ to μ' at point x. In the formulation (2.2), we implicitly assumed that these interactions preserve particle number, i.e.,

$$\int_{-1}^{1} Q(f)\,d\mu = 0\,. \tag{2.3}$$

This will be sufficient for our purpose. In practical cases, like neutron transport or radiative transfer, it is necessary to include nonconservative cases (like neutron multiplication or photon absorption/emission), but these effects are not essential and can be easily incorporated into our formulation if needed. Equation (2.1) must be supplemented with an initial condition $f_0(x,\mu)$ and suitable boundary conditions. A particular case is when S does not depend on μ and μ': $S(x,\mu,\mu') = \sigma(x)/2$, where σ is the collision frequency. Then

$$Q(f) = \sigma(x)\left[\frac{1}{2}\int_{-1}^{1} f(x,\mu',t)\,d\mu' - f(x,\mu,t)\right] \tag{2.4}$$

is (up to a multiplicative factor) a projection operator onto the functions independent of μ.

When the particle interactions with the medium are very frequent, i.e., when Q is "large," the numerical resolution of (2.1) becomes extremely time consuming, and it is worth using the asymptotic model obtained when Q "tends to infinity." We introduce a new set of "macroscopic variables" x' and t' according to

$$x' = \varepsilon x\,, \quad t' = \varepsilon^2 t\,,$$

where ε denotes the ratio of the microscopic to the macroscopic scale. Typically, ε is the ratio of the particle mean-free path (related to a typical value of S) to the size of the problem under consideration and is called the Knudsen number. After using this change of variables and dropping the primes for simplicity, one gets

$$\varepsilon^2 \partial_t f^\varepsilon + \varepsilon \mu \partial_x f^\varepsilon = Q(f^\varepsilon). \tag{2.5}$$

In the limit $\varepsilon \to 0$, f^ε converges towards the solution of a diffusion equation. More precisely, we have the following (see, e.g., [5], [6]).

LEMMA 2.1. *$f^\varepsilon \to n(x,t)$, where n is a solution of*

$$\partial_t n - \partial_x (D(x) \partial_x n) = 0\,, \tag{2.6}$$

with initial condition $n|_{t=0} = \frac{1}{2}\int f_0(x,\mu)\,d\mu$. The diffusion constant D is related to Q by

$$D(x) = -\frac{1}{2}\int_{-1}^{1} Q^{-1}(\mu)\,\mu\,d\mu > 0\,. \tag{2.7}$$

We shall not be more precise on the functional spaces and instead refer readers to [5], [17] for details. The definition of Q^{-1} needs a few words of explanation. We first note that Q only operates with respect to the variable μ, while x is just a parameter. As an operator acting on functions of μ, Q has the following properties (see, e.g., [5]).

LEMMA 2.2. *Suppose that $0 < C_0 \le S \le C_1 < \infty$. Then Q is a bounded self-adjoint nonpositive operator on $L^2(-1,1)$. Furthermore*

(i) $\mathrm{Ker\,Q}$ *consists of constant functions with respect to μ.*

(ii) $\mathrm{Im\,Q} = (\mathrm{Ker\,Q})^\perp = \{\mathrm{g}\ s.t. \int \mathrm{g}\,\mathrm{d}\mu = 0\}$.

(iii) *Q is invertible from $(\mathrm{Ker\,Q})^\perp$ to $(\mathrm{Ker\,Q})^\perp$. Its (pseudo-)inverse is denoted by Q^{-1}.*

It is worth summarizing the main steps of the proof.

Proof of Lemma 2.1. We use the Hilbert expansion

$$f^\varepsilon = f^{(0)} + \varepsilon f^{(1)} + \varepsilon^2 f^{(2)} + O(\varepsilon^3)\,. \tag{2.8}$$

We insert this expansion into (2.5) and identify terms of equal powers of ε. This leads to the sequence of equations

$$Q(f^{(0)}) = 0\,, \tag{2.9}$$

$$Q(f^{(1)}) = \mu \partial_x f^{(0)}\,, \tag{2.10}$$

$$Q(f^{(2)}) = \mu \partial_x f^{(1)} + \partial_t f^{(0)}\,. \tag{2.11}$$

With (2.9) and Lemma 2.2(i), we deduce that $f^{(0)}$ does not depend on μ. We denote $n(x,t) = f^{(0)}(x,\mu,t)$.

Equation (2.10) simplifies into

$$Q(f^{(1)}) = \mu \partial_x n \,. \tag{2.12}$$

Its right-hand side is an odd function of μ and therefore, integrated against any constant function, yields 0. Therefore, it belongs to $(\mathrm{Ker}\, Q)^\perp$. By Lemma 2.2(iii), we can invert (2.12) in $(\mathrm{Ker}\, Q)^\perp$ and get

$$f^{(1)} = Q^{-1}(\mu)\partial_x n \,. \tag{2.13}$$

For the most general solution of (2.12), we should add an element of $(\mathrm{Ker}\, Q)$, i.e., a function of (x,t) only. We fix this function to be zero to ensure that n is an $O(\varepsilon^2)$ approximation to the true density $n^\varepsilon = (1/2)\int f^\varepsilon \, d\mu$ (since then $\int f^{(1)}\, d\mu = 0$).

Equation (2.11) is solvable for $f^{(2)}$ if and only if its right-hand side is orthogonal to the functions independent of μ. Therefore, the solvability condition of (2.11) reads

$$\partial_t n + \partial_x j = 0 \tag{2.14}$$

with

$$j = \frac{1}{2}\int_{-1}^{1} f^{(1)} \mu \, d\mu \,. \tag{2.15}$$

Inserting (2.13) into (2.15), we get that $j = -D\partial_x n$ with D given by (2.7). Finally, with (2.14), we get (2.6). The fact that D is positive comes from the positive definiteness of $-Q$ on $(\mathrm{Ker}\, Q)^\perp$. □

Note that in the case (2.4), $D = 1/(3\sigma)$.

The proof of Lemma 2.1 relies on the fact that S (or σ) is everywhere of order unity (with respect to ε). However, S is position dependent and there are numerous situations in which it is of order unity in certain parts of the domain, while it is much smaller (of order ε or ε^2) in other parts of the domain. The diffusion equation is only valid when S is of order unity. If S is smaller, solving the transport equation is necessary. Therefore, one needs to couple the diffusion equation in the regions where S is of order unity to the transport equation in the regions where it is smaller.

2.2. The coupling method. This problem has been addressed by many authors with many methods (see the references in the introduction). Our approach is novel and consists of introducing a buffer zone in which both diffusion and kinetic equations will be solved. The solution of the initial transport equation will be recovered as the *sum* of the solutions of these two equations. In this way, our strategy departs from strategies based on domain decomposition with overlap, in which *each* of the models represents the *full* solution. The buffer interval is denoted by $[a,b]$. We introduce a smooth function $h(x)$ such that

$$\begin{cases} h(x) = 1 & \text{for} \quad x \le a \,, \\ h(x) = 0 & \text{for} \quad x \ge b \,, \\ h(x) \in (0,1) & \text{for} \quad a \le x \le b \,. \end{cases}$$

We consider the following coupled system for two distribution functions f_L and f_R:

$$\varepsilon^2 \partial_t f_L^\varepsilon + \varepsilon h \mu \partial_x f_L^\varepsilon + \varepsilon h \mu \partial_x f_R^\varepsilon = h(Q(f_L^\varepsilon) + Q(f_R^\varepsilon)) \,, \tag{2.16}$$

$$\varepsilon^2 \partial_t f_R^\varepsilon + \varepsilon (1-h) \mu \partial_x f_L^\varepsilon + \varepsilon (1-h) \mu \partial_x f_R^\varepsilon = (1-h)(Q(f_L^\varepsilon) + Q(f_R^\varepsilon)) \tag{2.17}$$

with initial data

$$f_L^\varepsilon|_{t=0} = hf_0, \quad f_R^\varepsilon|_{t=0} = (1-h)f_0\,. \tag{2.18}$$

We first note the following.

LEMMA 2.3. *If f_L^ε and f_R^ε are the solution of problem* (2.16), (2.17) *with initial data* (2.18)*, then $f = f_L^\varepsilon + f_R^\varepsilon$ is the solution of problem* (2.5) *with initial condition f_0.*

Proof. For the proof, simply add up (2.16) and (2.17). ∎

We note that, in reality, (2.16) is posed on the interval $(-\infty, b)$ and (2.17) on $(a, +\infty)$ since h vanishes for $x > b$ and $1-h$ for $x < a$. Additionally, since h (resp., $1-h$) multiplies the space derivative operator in (2.16) (resp., (2.17)), *no boundary condition is required for f_L^ε at $x = b$ (resp., for f_R^ε at $x = a$).*

Now, we assume that S is of order ε^2 in the interval $(-\infty, a)$, while it is of order 1 in $(a, +\infty)$. Therefore, we shall only be allowed to perform the diffusion approximation on f_R^ε, while f_L^ε will have to stay untouched. For this purpose, we rewrite (2.17) according to

$$\begin{aligned} &\varepsilon^2\partial_t f_R^\varepsilon + \varepsilon(1-h)\mu\partial_x f_R^\varepsilon - (1-h)Q(f_R^\varepsilon) \\ &\quad = -\varepsilon(1-h)\mu\partial_x f_L^\varepsilon + (1-h)Q(f_L^\varepsilon)\,, \end{aligned} \tag{2.19}$$

and we consider the terms on the right-hand side to be of order ε^2. The following proposition states the diffusion approximation $\varepsilon \to 0$ of this equation.

PROPOSITION 2.4. *Consider* (2.19)*, where the right-hand side is treated as an $O(\varepsilon^2)$ term. Then as $\varepsilon \to 0$, $f_R^\varepsilon \sim n_R^\varepsilon$ where $n_R^\varepsilon = n_R^\varepsilon(x,t)$ is a solution of the following diffusion equation:*

$$\partial_t n_R^\varepsilon - (1-h)\partial_x[D(x)\partial_x n_R^\varepsilon] + (1-h)\partial_x j_L^\varepsilon = 0, \tag{2.20}$$

where D is given by (2.7) *and*

$$j_L^\varepsilon = \frac{1}{2\varepsilon}\int_{-1}^{1} f_L^\varepsilon \mu\, d\mu\,. \tag{2.21}$$

Since ε tends to 0 only in some terms and not in others, we cannot speak of convergence but rather of asymptotic equivalence; hence the use of the symbol $\sim$. Again, (2.20) is a diffusion equation on the interval $[a, +\infty)$. However, since $1-h$ vanishes at $x = a$, *the diffusion operator is degenerate at this point and no boundary condition is required.*

Proof. We again write the Hilbert expansion $f_R^\varepsilon = f_R^{(0)} + \varepsilon f_R^{(1)} + \varepsilon^2 f_R^{(2)} + O(\varepsilon^3)$. The computations of $f_R^{(0)} = n_R$ and of $f_R^{(1)}$ are the same as in Lemma 2.1. Indeed, the right-hand side of (2.19) being of order ε^2 does not contribute anything to (2.9) and (2.10). The only change is in (2.11), which becomes

$$\begin{aligned} (1-h)Q(f_R^{(2)}) &= (1-h)\mu\partial_x f_R^{(1)} + \partial_t f_R^{(0)} \\ &\quad + \frac{1}{\varepsilon}(1-h)\mu\partial_x f_L^\varepsilon - \frac{1}{\varepsilon^2}(1-h)Q(f_L^\varepsilon)\,. \end{aligned} \tag{2.22}$$

We note that the sum of the last two terms is of order 1 by our hypothesis, despite their apparent dependence on ε. Integrating (2.22) with respect to μ in order to express the solvability condition for f_R^2, we obtain

$$\partial_t n_R^\varepsilon + (1-h)\partial_x j_R^\varepsilon = -(1-h)\partial_x\left(\frac{1}{2\varepsilon}\int_{-1}^{1} f_L^\varepsilon \mu\, d\mu\right)$$

because the contribution of $Q(f_L^\varepsilon)$ vanishes after integration with respect to μ by (2.3). This leads to (2.20) and concludes the proof. □

The coupled kinetic-diffusion model is now written as follows:

$$\varepsilon^2\partial_t f_L^\varepsilon + \varepsilon h\mu\partial_x f_L^\varepsilon + \varepsilon h\mu\partial_x f_R^\varepsilon = h(Q(f_L^\varepsilon) + Q(f_R^\varepsilon))\,, \qquad -\infty < x \le b, \tag{2.23}$$

$$\partial_t n_R^\varepsilon - (1-h)\partial_x(D(x)\partial_x n_R^\varepsilon) + (1-h)\partial_x j_L^\varepsilon = 0\,, \quad a \le x < \infty, \tag{2.24}$$

$$f_R^\varepsilon = n_R^\varepsilon + \varepsilon Q^{-1}(\mu)\partial_x n_R^\varepsilon\,, \tag{2.25}$$

$$j_L^\varepsilon = \frac{1}{2\varepsilon}\int_{-1}^{1} f_L^\varepsilon \mu\, d\mu\,, \tag{2.26}$$

with initial data

$$f_L^\varepsilon|_{t=0} = hf_0\,, \quad f_R^\varepsilon|_{t=0} = (1-h)f_0\,, \quad n_R^\varepsilon|_{t=0} = \frac{1}{2}(1-h)\int_{-1}^{1} f_0\, d\mu\,. \tag{2.27}$$

So far we use ε to perform the asymptotic analysis more conveniently. In numerical implementation, one should drop ε in the coupling model (2.23)–(2.26) by setting $\varepsilon = 1$. The domain is diffusive if $Q \approx O(1/\varepsilon)$.

2.3. Properties of the coupling. The reconstruction of the distribution function f_R^ε from the solution of the diffusion equation n_R^ε retains both the zeroth and first order terms of the Hilbert expansion. This is necessary in order to recover the diffusion equation on the entire real line when both regions are diffusive (see below). Equations (2.23) and (2.25) can be combined into

$$\varepsilon^2\partial_t f_L^\varepsilon + \varepsilon h\mu\partial_x f_L^\varepsilon - hQ(f_L^\varepsilon) = \varepsilon^2 h\mu\partial_x(Q^{-1}(\mu)\partial_x n_R^\varepsilon)\,, \tag{2.28}$$

showing that the n_R^ε enters the equation for f_L^ε in an order $O(\varepsilon^2)$ term. In the case where $S = \sigma(x)/2$ does not depend on μ, this equation becomes

$$\varepsilon^2\partial_t f_L^\varepsilon + \varepsilon h\mu\partial_x f_L^\varepsilon - hQ(f_L^\varepsilon) = \varepsilon^2 h\mu^2\partial_x(\sigma^{-1}\partial_x n_R^\varepsilon)\,. \tag{2.29}$$

We now prove that if both regions are diffusive, we recover the global diffusion equation (2.6) for $n = n_L + n_R$.

PROPOSITION 2.5. *As $\varepsilon \to 0$, the solution f_L^ε, n_R^ε of system* (2.23)–(2.27) *converges to the pair n_L, n_R, the solution of the diffusion system*

$$\partial_t n_L - h[\partial_x(D(x)\partial_x n_L) + \partial_x(D(x)\partial_x n_R)] = 0\,, \tag{2.30}$$

$$\partial_t n_R - (1-h)[\partial_x(D(x)\partial_x n_R) + \partial_x(D(x)\partial_x n_L)] = 0\,, \tag{2.31}$$

with initial data

$$n_L|_{t=0} = hn_0\,, \quad n_R|_{t=0} = (1-h)n_0\,, \quad n_0 = \frac{1}{2}\int_{-1}^{1} f_0\, d\mu\,. \tag{2.32}$$

In particular, $n = n_L + n_R$ is the solution of the diffusion equation (2.6) *with initial condition n_0.*

Proof. The proof is similar to that of Proposition 2.4. The term involving n_R^ε in the transport equation for f_L^ε is of order $O(\varepsilon^2)$. Therefore, it does not induce any change in the expression of the first two equations of the Hilbert expansion (2.9)–(2.10). Equation (2.11) is modified into

$$hQ(f_L^{(2)}) = h\mu\partial_x f_L^{(1)} + \partial_t f_L^{(0)} + h\mu\partial_x(Q^{-1}(\mu)\partial_x n_R)\,. \tag{2.33}$$

Integrating it with respect to μ in order to express the solvability condition for $f_L^{(2)}$, we obtain

$$\partial_t n_L^\varepsilon + h\partial_x j_L^\varepsilon = -h\partial_x \left(\left(\frac{1}{2} \int Q^{-1}(\mu)\mu \, d\mu \right) \partial_x n_R \right) ,$$

thus leading to (2.30). We use the expression of j_L from (2.15), (2.13), giving $j_L = -D\partial_x n_L$. Inserting this expression into (2.24) yields (2.31), which concludes the proof. ∎

Therefore, our coupled kinetic-diffusion model is consistent with the diffusion equation on the entire real line when the kinetic region $(-\infty, a)$ is also in the diffusive regime.

An important issue is positivity. Indeed, a distribution function being a density in phase space is a positive quantity. Therefore, we should ensure that $f_L^\varepsilon + f_R^\varepsilon$, which is our approximation of f, remains positive, or at least close to positive. Since the signs of the coupling terms in (2.23)–(2.24) are not determined, both f_L^ε and n_R^ε could become negative. In fact, outside the buffer zone, i.e., in the intervals $(-\infty, a]$ or $(b, +\infty]$, the coupling terms vanish identically, and we solve either a standard transport equation or a standard diffusion equation. Therefore, loss of positivity can originate only from the buffer zone $[a, b]$. This region is a diffusive region for both f_L^ε and f_R^ε since the conditions which ensure that the diffusion approximation is valid for f_R^ε (i.e., $Q = O(1)$) also make the diffusion approximation valid for f_L^ε. In this case, both f_L^ε and f_R^ε are close (at order ε) to the solutions n_R and n_L of the diffusion equations (2.30), (2.31). But since $n_R + n_L$ solves the classical diffusion equation, it remains positive. Therefore, in the buffer zone, $f_L^\varepsilon + f_R^\varepsilon$ are close, up to order $0(\varepsilon)$ terms, to a positive function.

Consequently, if losses of positivity occur in the buffer zone, they will remain small, i.e., of order ε. If they are ultimately propagated outside the buffer zone, they will remain small and of this order everywhere. Therefore, although we cannot prove that our method preserves positivity, we have solid indications that negative values, if they occur, will remain small. Our numerical simulations did not exhibit any loss of positivity so far (see next section).

Remark 2.1. There is an alternate coupling strategy to (2.23)–(2.27). It consists of the following system:

$$\varepsilon^2 \partial_t f_L^\varepsilon + \varepsilon h\mu \partial_x f_L^\varepsilon + \varepsilon h\mu \partial_x n_R^\varepsilon = Q(f_L^\varepsilon) , \tag{2.34}$$

$$\partial_t n_R^\varepsilon - \partial_x((1-h)D(x)\partial_x n_R^\varepsilon) - \partial_x((1-h)D(x)\partial_x n_L^\varepsilon) + (1-h)\partial_x j_L^\varepsilon = 0 , \tag{2.35}$$

$$n_L^\varepsilon = \frac{1}{2}\int_{-1}^{1} f_L^\varepsilon \, d\mu , \quad j_L^\varepsilon = \frac{1}{2\varepsilon}\int_{-1}^{1} f_L^\varepsilon \mu \, d\mu . \tag{2.36}$$

This model is more heuristic since it is not obtained from a diffusion approximation of the coupled kinetic equations (2.16), (2.17). However, it shares with (2.23)–(2.27) the property of relaxing towards the solution of the diffusion equation on the whole real line when both regions are diffusive (we leave the details of the proof to the reader).

Remark 2.2. A rigorous convergence analysis of this coupling strategy is outside the scope of the present paper, which focuses more on the practical feasibility of the method. The convergence analysis, as well as a rigorous study of the loss of positivity, will be investigated in future work.

3. Numerical method. In this section we introduce a new (spatially discrete) numerical method for the coupling problem. In fact, this numerical scheme can be

used for a discretization of the transport equation, with different order of magnitude in ϵ, in the spirit of the *asymptotic-preserving method* [9], [17], [20], [21], [22], [23], [24], [26], [30], [31], [32] that works uniformly with respect to the mean-free path. However, this new asymptotic-preserving spatial discretization method has not been reported in the literature.

This goal of this scheme to is verify the validity of the coupling method numerically. It remains a future research topic to find the best numerical scheme for the coupling problem.

3.1. Parity formulation. We explain the new scheme using the transport equation with isotropic scattering (2.4). It is based on the parity form of the transport equation. This is a standard form used to construct the asymptotic-preserving scheme [1], [24], [36]. For anisotropic scattering, if $S(x,\mu,\mu')$ is even in both μ and μ', one can also use the parity form. For isotropic scattering (2.4), the coupling problem (2.23)–(2.26) becomes

$$\begin{aligned}&\varepsilon^2\partial_t f_L+\varepsilon h\mu\partial_x f_L+\varepsilon h\mu\partial_x f_R\\ &\qquad=h\sigma\left[\frac{1}{2}\int_{-1}^{1}(f_L(\mu')+f_R(\mu'))d\mu'-f_L(\mu)-f_R(\mu)\right],\end{aligned}\tag{3.1}$$

$$\partial_t n_R-\frac{1}{3}(1-h)\partial_x(\sigma(x)^{-1}\partial_x n_R)+(1-h)\partial_x j_L=0\,,\tag{3.2}$$

$$f_R=n_R-\varepsilon\frac{\mu}{\sigma}\partial_x n_R\,,\tag{3.3}$$

$$j_L^{\varepsilon}=\frac{1}{\varepsilon}\int_0^1 f_L\mu\,d\mu\,.\tag{3.4}$$

Applying (3.3) in (3.1) one gets

$$\varepsilon^2\partial_t f_L+\varepsilon h\mu\partial_x f_L-\varepsilon^2 h\mu^2\partial_x(\sigma^{-1}\partial_x n_R)=h\sigma\left[\frac{1}{2}\int_{-1}^{1}f_L(\mu')d\mu'-f_L(\mu)\right],\tag{3.5}$$

$$\partial_t n_R-\frac{1}{3}(1-h)\partial_x(\sigma(x)^{-1}\partial_x n_R)+(1-h)\partial_x j_L=0\,,\tag{3.6}$$

$$j_L=\frac{1}{\varepsilon}\int_0^1 f_L\mu\,d\mu\,.\tag{3.7}$$

Define the even and odd parities, for $\mu>0$, as

$$f^E(t,x,\mu)=\frac{1}{2}\left[f(t,x,\mu)+f(t,x,-\mu)\right],\tag{3.8}$$

$$f^O(t,x,\mu)=\frac{1}{2\epsilon}\left[f(t,x,\mu)-f(t,x,-\mu)\right].\tag{3.9}$$

Then

$$n(t,x)=\frac{1}{2}\int_{-1}^{1}f(t,x,\mu)\,d\mu=\int_0^1 f^E(t,x,\mu)\,d\mu\,,\tag{3.10}$$

$$j(t,x)=\frac{1}{2\epsilon}\int_{-1}^{1}f(t,x,\mu)\mu\,d\mu=\int_0^1 f^O(t,x,\mu)\mu\,d\mu\,.\tag{3.11}$$

With the parities, from now on we only consider $\mu>0$. First, we split (3.1) and

(3.2), i.e., one for μ and one for $-\mu$ (from now on we omit the superscript):

$$\begin{aligned}&\varepsilon^2\partial_t f_L(\mu)+\varepsilon h\mu\partial_x f_L(\mu)-\varepsilon^2 h\mu^2\partial_x(\sigma^{-1}\partial_x n_R)\\&\quad=h\sigma\left[\frac{1}{2}\int_{-1}^{1} f_L(\mu')d\mu'-f_L(\mu)\right]\quad\text{for}\quad x<b\,,\end{aligned}\tag{3.12}$$

$$\begin{aligned}&\varepsilon^2\partial_t f_L(\mu)+\varepsilon h\mu\partial_x f_L(-\mu)-\varepsilon^2 h\mu^2\partial_x(\sigma^{-1}\partial_x n_R)\\&\quad=h\sigma\left[\frac{1}{2}\int_{-1}^{1} f_L(\mu')d\mu-f_L(-\mu)\right]\quad\text{for}\quad x<b\,,\end{aligned}\tag{3.13}$$

$$\partial_t n_R-\frac{1}{3}(1-h)\partial_x(\sigma(x)^{-1}\partial_x n_R^{\varepsilon})+(1-h)\partial_x j_L^{\varepsilon}=0\quad\text{for}\quad x>a\,,\tag{3.14}$$

$$j_L=\frac{1}{\varepsilon}\int_0^1 f_L^O\,\mu\,d\mu\,.\tag{3.15}$$

Adding and subtracting the two equations in (3.12) and (3.13) leads to

$$\epsilon^2\partial_t f_L^E+\epsilon h\mu\partial_x f_L^O-\varepsilon^2 h\mu^2\partial_x(\sigma^{-1}\partial_x n_R)=h\sigma\left(\int_0^1 f_L^E\,d\mu-f_L^E\right),\tag{3.16}$$

$$\epsilon^2\partial_t f_L^O+\epsilon h\mu\partial_x f_L^E=-h\sigma f_L^O\quad\text{for}\quad x<b.\tag{3.17}$$

Our system now consists of (3.16), (3.17), (3.14), and (3.15).

3.2. Asymptotic-preserving spatial discretization using staggered grids. For spatial discretization, let x_j be the mesh point for $i=0,1,\dots,J$. The even parity will be defined on these mesh points, namely, $f_i^E=f^E(x_i), n_{R,i}=n_R(x_i)$. For the odd parity we define them on a staggered mesh point $x_{i+1/2}=(x_i+x_{i+1})/2$: $f_{i+1/2}^O=f^O(x_{i+1/2})$. This definition guarantees that, when $\epsilon\to 0$, one ends up at a three-point rather than five-point scheme for the diffusion equation. This has better resolution than the previous asymptotic-preserving schemes [23], [24], [26] for diffusive transport equations that yield five-point stencils in this limit. However, we point out that using staggered mesh is restricted to one space dimension. For a higher dimension, the classical asymptotic-preserving schemes can still be applied, but this is beyond scope of this paper.

Let $\sigma_i=\sigma(x_i),\sigma_{i+1/2}=\frac{1}{2}(\sigma_i+\sigma_{i+1}), h_i=h(x_i), h_{i+1/2}=h(x_{i+1/2})$. The spatially discrete scheme for the coupling problem (3.16), (3.17), (3.14), (3.15) is given by center difference on a staggered grid:

$$\begin{aligned}&\epsilon^2\partial_t f_{L,i}^E+\epsilon h_i\mu\frac{f_{L,i+1/2}^O-f_{L,i-1/2}^O}{\Delta x}\\&\quad-\epsilon h_i\mu^2\frac{1}{(\Delta x)^2}\left[\sigma_{i+1/2}^{-1}(n_{R,i+1}-n_{R,i})-\sigma_{i-1/2}^{-1}(n_{R,i}-n_{R,i-1})\right]\\&\quad=h_i\sigma_i\left[\int_0^1 f_{L,i}^E\,d\mu-f_{L,i}^E\right],\end{aligned}\tag{3.18}$$

$$\epsilon^2\partial_t f_{L,i+1/2}^O+\epsilon h_{i+1/2}\,\mu\frac{f_{L,i+1}^E-f_{L,i}^E}{\Delta x}=-h_{i+1/2}\sigma_{i+1/2}f_{L,i+1/2}^O,\tag{3.19}$$

$$\begin{aligned}&\partial_t n_{R,i}-(1-h_i)\frac{1}{(\Delta x)^2}\left[\sigma_{i+1/2}^{-1}\,(n_{R,i+1}-n_{R,i})-\sigma_{i-1/2}^{-1}\,(n_{R,i}-n_{R,i-1})\right]\\&\quad+(1-h_i)\frac{j_{L,i+1/2}-j_{L,i-1/2}}{\Delta x}=0\,,\end{aligned}\tag{3.20}$$

$$j_{L,i+1/2} = \frac{1}{\varepsilon}\int_0^1 f^O_{L,i+1/2}\mu\, d\mu\,. \tag{3.21}$$

This is a second order approximation. To verify that it is asymptotic preserving, for $\epsilon \ll 1$, the leading order approximation of (3.18) gives

$$f^E_{L,i} = \int_0^1 f^E_{L.i}\, d\mu \equiv n_{L,i}, \tag{3.22}$$

while (3.19) gives

$$f^O_{L,i+1/2} = -\epsilon\frac{\mu}{\sigma_{i+1/2}}\frac{f^E_{L,i+1} - f^E_{L,i}}{\Delta x}\,. \tag{3.23}$$

Applying (3.22) and (3.23) into (3.18) and integrating over μ, we get

$$\begin{aligned}&\partial_t\rho^E_{L,i} - h_i\frac{1}{3(\Delta x)^2}\left[\sigma^{-1}_{i+1/2}(n_{L,i+1} - n_{L,i}) - \sigma^{-1}_{i-1/2}(n_{L,i} - n_{L,i-1})\right]\\ &- h_i\frac{1}{3(\Delta x)^2}\left[\sigma^{-1}_{i+1/2}(n_{R,i+1} - n_{R,i}) - \sigma^{-1}_{i-1/2}(n_{R,i} - n_{R,i-1})\right] = 0\,,\end{aligned} \tag{3.24}$$

which is a three-point second order approximation to (2.20) with $D = \sigma^{-1}$.

Since the space discretization is a centered-difference one, in order to ensure stability of the time discretization, one can use a so-called *I-stable* ODE solver. An I-stable solver is a scheme for ODEs whose stability region contains part of the imaginary axis. The third and fourth order explicit Runge–Kutta methods, among other schemes, are I-stable. An I-stable ODE solver is particularly well suited for convection problems where the convection term is discretized using centered differences, since the spectrum of a centered difference operator is purely imaginary and an I-stable scheme, unlike the forward Euler method, will be stable with a suitable choice of the time step [4], [16], [38]. We would like to remark that for steady state computations, the convergence to the steady state for such an explicit solver is very slow, and some acceleration techniques, such as the diffusion synthetic acceleration [2], can be used but will not be explored in this paper.

For discretization in the velocity space, we use the standard discrete-ordinate method with Gaussian quadrature over $(0, 1)$ [35].

4. Numerical examples. In this section we present several numerical examples on the coupling model (3.1)–(3.4) by the numerical schemes described in the previous section. We solve the problem in domain $[0, 1]$ with Dirichlet boundary condition

$$f(t, 0, \mu) = f_l(\mu)\,, \qquad f(t, 1, -\mu) = f_r(-\mu)\,, \qquad \mu > 0\,.$$

In these examples $\epsilon = 1$, and the value of σ characterizes the nature of the regime (transport or diffusive). We use 1001 points to solve the transport equation in the entire domain as the "exact" solution and 25 points for the numerical approximations. We use three choices of h that are piecewise linear: 1 for $x \le a$; 0 at $x > b$; and a line connecting 1 at a and 0 at b with $a = 0.2912, b = 0.7072$; $a = 0.416, b = 0.5824$; and $a = b = 0.5$, respectively. The last set gives a step function. The buffer zone is always chosen to be symmetric and centered at $x = 0.5$. The "exact" solution is given by the solid line, while the numerical results are given by o, x, and * for the three different sets of a and b, respectively. We compare both transient and steady state solutions.

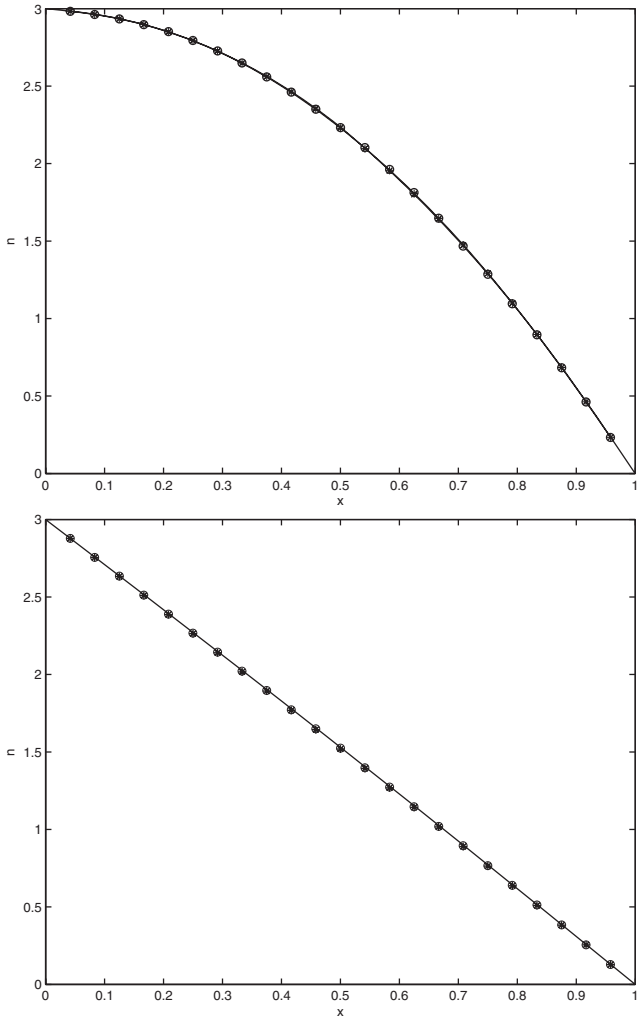


FIG. 1. *The numerical solution of n for Example* 1 *at* $t = 1$ *(top) and at steady state (bottom) of Example* 1. *The solid line is the solution of the transport equation in the entire domain computed using* 1001 *points, while the other three symbols represent numerical solution of the coupling model with* 25 *grid points and three different sizes of the buffer zone. Circles represent the solution for the larger buffer zone,* x*'s for the intermediate one, and * for the smaller one (when it reduces to one single point).*

Remark 4.1. When h is a step function, one may wonder if the advantage of the coupling model—avoiding the boundary condition at the buffer zone—is lost. Note that we solve only f_L^ε for $x \le b$, and n_R^ε for $x \ge a$, and set $f_L^\varepsilon = 0$ for $x > b$ and $n_R^\varepsilon = 0$ for $x < a$. This provides the numerical boundary condition for f_L^ε and n_R^ε at the interface. Although there is no theoretical justification for this numerical boundary condition, the numerical experiments in this section (where $\sigma(x)$ is continuous) suggests that h being a step function is still acceptable.

Example 1. We first test the case when both sides of the domain are diffusive in the entire domain, namely, $\sigma(x) \gg 1$. We take $\sigma(x) \equiv 100$, $f_l(\mu) = 3$, $f_r(\mu) = 0$, and

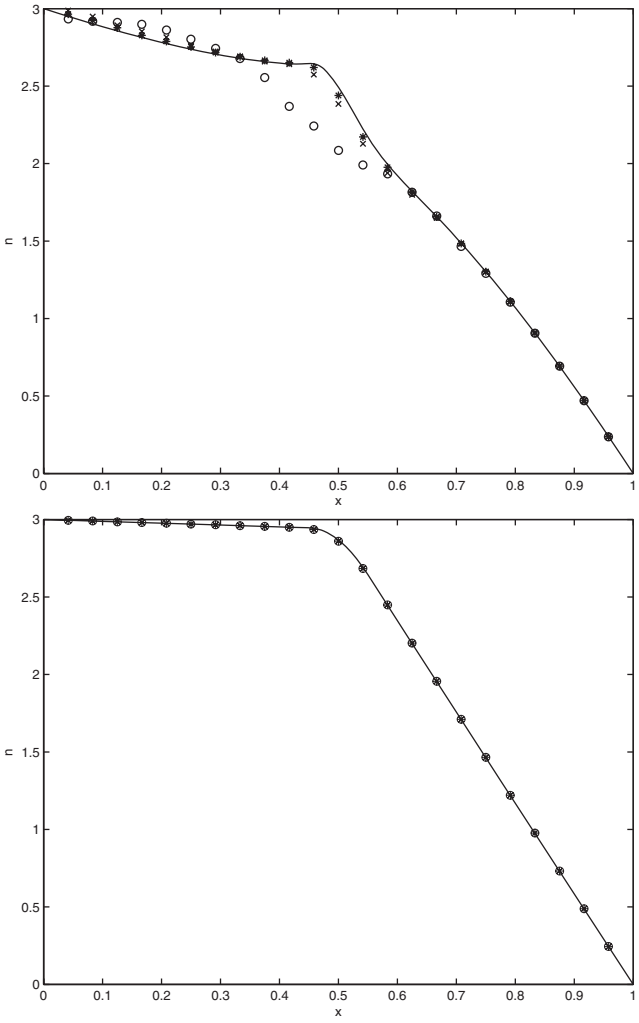


FIG. 2. *The numerical solution of* n *at* $t = 0.5$ *(top) and at steady state (bottom) for Example* 2. *The solid line is the solution of the transport equation in the entire domain computed using* 1001 *points, while the other three symbols represent numerical solutions of the coupling model with* 25 *grid points and three different sizes of buffer zone. Circles represent the solution for the larger buffer zone,* x*'s for the intermediate one, and * for the smaller one (when it reduces to one single point).*

the initial condition

$$f(0, x, \mu) = 3x^2 \, .$$

The solution to the diffusion equation (2.20) with $D = \sigma^{-1}$ should be a line connecting 3 to 0. The numerical results at $t = 1$ and at steady state are plotted in Figure 1.

One can see that the numerical results match the "exact" solution quite well, while the choice of h seems to have little influence on the numerical solution.

Example 2. We take $\sigma(x) = 2$ for $x < 0.45$, $\sigma(x) = 100$ for $x > 0.55$. $\sigma(x)$ is a linear function interpolating 2 and 100 for $0.45 \leq x \leq 0.55$. The boundary and initial conditions are the same as in Example 1. Thus the domain $[0, 0.45]$ is in the kinetic regime, while the domain $[0.55, 1]$ is diffusive. The results at $t = 0.5$ and

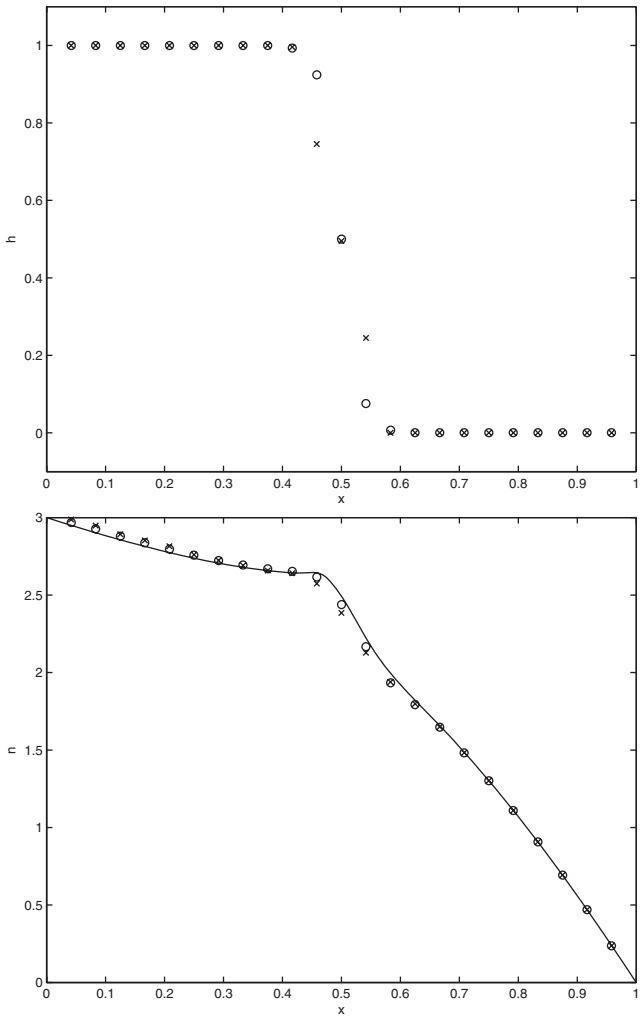


FIG. 3. *Example* 2. *Top: Comparison of different* h*: piecewise linear* x*,* C^∞*,* o*. Bottom: The corresponding numerical results for these* h*-functions versus the "exact" solution at* $t = 0.5$.

at steady state for $n = n_L + n_R$ are depicted in Figure 2 (top) for several different choices of a and b. As one can see, for the steady state solution, the choice of h has little influence on the numerical results, which match well with the "exact" solution. For the transient solution, the first set of parameters for the buffer zone (which is much larger than the domain for nonconstant σ) yields poor approximation in both the buffer zone and the transport domain, while in the diffusion domain the accuracy is as good as the other two buffer zones. This experiment indicates that the buffer zone should be within the transition region of σ.

In order to compare the effect of regularity of h on the numerical solution, we compare the piecewise linear h corresponding to $a = 0.2912, b = 0.7072$ (x in Figure 3, top) and $h = 0.5(1 - \tanh(30(x - 0.5))) \in C^\infty$ (o in Figure 3, top). The numerical results at $t = 0.5$ are given in the bottom of Figure 3. The numerical results are comparable.

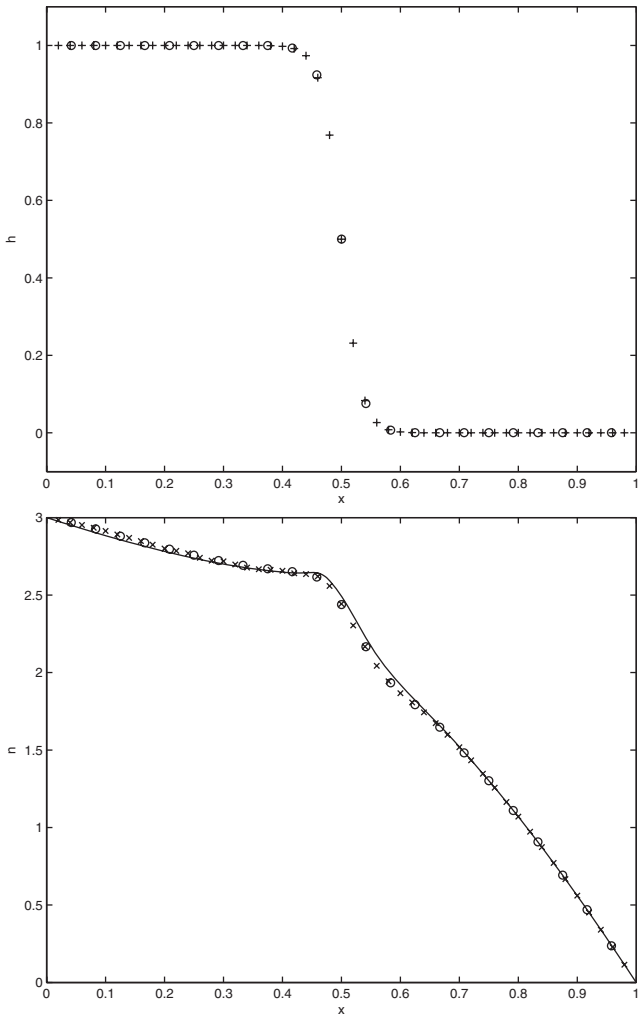


FIG. 4. *Example* 2. *Top: Comparison of same* C^∞*-function* h *with different mesh sizes. o:* 25 *grid points. x:* 50 *grid points. Bottom: The corresponding numerical results for these* h*-functions versus the "exact" solution at* $t = 0.5$.

For the same C^∞-function h, we compare the effect of mesh refinement on the numerical solution. In Figure 4 we compare the numerical results obtained by 25 and 50 points, respectively. The discrete h looks more regular in the finer mesh. The numerical results look similar, indicating that the regularity of h plays an insignificant role in the coupling algorithm.

Example 3. This problem is the same as Example 2 except for the boundary condition at $x = 0$, where we take an anisotropic one $f_l(\mu) = 3\mu + 1$ at $x = 0$. The numerical results at the steady state for the three piecewise linear h corresponding to the three different sizes of the buffer zone are given in Figure 5 and match quite well with the "exact" solution.

5. Conclusion. In this paper, we have presented a model which allows us to handle the transition between a kinetic and a diffusive region in a smooth way. In

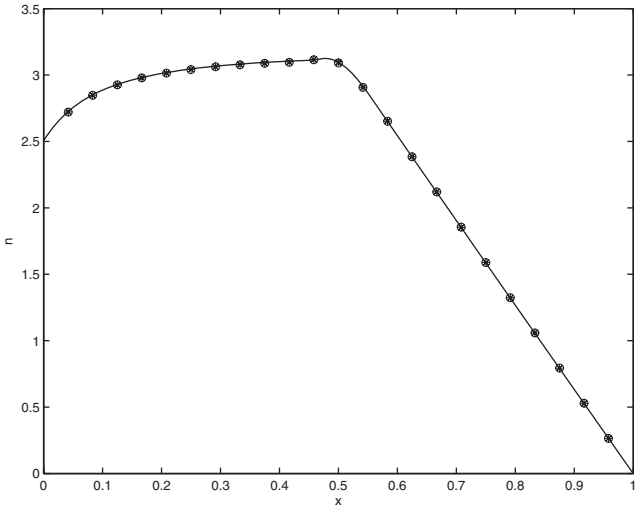


FIG. 5. *The numerical steady state solutions* $n = n_L + n_R$ *of Example* 3. *The solid line shows the solution of the transport equation in the entire domain cumputed using* 1001 *points, while the other three symbols represent numerical solution of the coupling model with* 25 *grid points and three different sizes of the buffer zone. Circles represent the solution for the larger buffer zone,* x*'s for the intermediate one, and * for the smaller one (when it reduces to one single point).*

the transition region both models are solved and the solution of the original transport equation is recovered by adding up the solutions of each model. The advantage of this coupling is that no boundary condition nor any iteration process at the overlapping zone is needed, as is the case for a typical domain decomposition method. The numerical discretization in the kinetic region is based on the parity formulation of the transport equation and the use of a new asymptotic-preserving scheme. Numerical experiments show that the coupling model describes quantitatively the behavior of the original transport equation, for both transient and steady state solutions, if the buffer zone is chosen inside the transition zone.

Further development of this work will include more robust time and space discretizations, multidimensional problems, and extensions to more complex kinetic models such as drift-diffusion or energy-transport models in semiconductors or the Boltzmann–BGK (Bhatnagar–Gross–Krook) model of rarefied gas dynamics.

REFERENCES

[1] M. L. ADAMS, *Even-parity finite-element transport methods in the diffusion limit*, Progress in Nuclear Energy, 25 (1991), pp. 159–198.

[2] R. E. ALCOUFFE, *Diffusion synthetic acceleration methods for the diamond-difference discrete-ordinate equations*, Nucl. Sci. Eng., 64 (1977), pp. 344–355.

[3] G. BAL AND Y. MADAY, *Coupling of transport and diffusion models in linear transport theory*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 69–86.

[4] W. BAO AND S. JIN, *High order I-stable centered difference schemes for viscous compressible flows*, J. Comput. Math., 21 (2003), pp. 101–112.

[5] C. BARDOS, R. SANTOS, AND R. SENTIS, *Diffusion approximation and computation of the critical size*, Trans. AMS, 284 (1984), pp. 617–649.

[6] A. BENSOUSSAN, J.-L. LIONS, AND G. C. PAPANICOLAOU, *Boundary layers and homogenization of transport processes*, Publ. Res. Inst. Math. Sci., 15 (1979), pp. 53–157.

[7] J.-F. Bourgat, P. Le Tallec, B. Perthame, and Y. Qiu, *Coupling Boltzmann and Euler equations without overlapping*, in Domain Decomposition Methods in Science and Engineering (Como, 1992), Contemp. Math. 157, AMS, Providence, RI, 1994, pp. 377–398.
[8] S. Brunner, E. Valeo, and J. P. Krommes, *Linear delta-f simulation of non-local electron heat transport*, Phys. Plasmas, 7 (2000), pp. 2810–2823.
[9] C. Buet, S. Cordier, B. Lucquin-Desreux, and S. Mancini, *Diffusion limit of the Lorentz model: Asymptotic preserving schemes*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 631–655.
[10] S. Chandrasekhar, *Radiative Transfer*, Dover, New York, 1960.
[11] N. Crouseilles, *Dérivation de modèles couplés dérive-diffusion/cinétique par une méthode de décomposition en vitesse*, C. R. Acad. Sci. Paris, 334 (2002), pp. 827–832.
[12] N. Crouseilles, P. Degond, and M. Lemou, *Hybrid kinetic/fluid models for nonequilibrium systems*, C. R. Acad. Sci. Paris., 336 (2003), pp. 359–364.
[13] N. Crouseilles, P. Degond, and M. Lemou, *A hybrid kinetic-fluid model for solving the gas dynamics Boltzmann-BGK equation*, J. Comput. Phys., 199 (2004), pp. 776–808.
[14] P. Degond and C. Schmeiser, *Kinetic boundary layers and fluid-kinetic coupling in semiconductors*, Transport Theory Statist. Phys., 28 (1999), pp. 31–55.
[15] S. Dellacherie, *Kinetic fluid coupling in the field of the atomic vapor laser isotopic separation: Numerical results in the case of a mono-species perfect gas*, in Proceedings of the 23rd International Symposium on Rarefied Gas Dynamics, Whistler (British Columbia), D. Ketsdever and E. P. Muntz, eds., AIP Conf. Proc. 663, 2003, pp. 947–956.
[16] W. E and J.-G. Liu, *Vorticity boundary condition and related issues for finite difference schemes,* J. Comput. Phys., 124 (1996), pp. 368–382.
[17] F. Golse, S. Jin, and C. D. Levermore, *The convergence of numerical transfer schemes in diffusive regimes* I: *Discrete-ordinate method*, SIAM J. Numer. Anal., 36 (1999), pp. 1333–1369.
[18] F. Golse, S. Jin, and C. D. Levermore, *A domain decomposition analysis for a two-scale linear transport problem*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 869–892.
[19] M. Günther, P. Le Tallec, J.-P. Perlat, and J. Struckmeier, *Numerical modeling of gas flows in the transition between rarefied and continuum regimes*, in Numerical Flow Simulation I, (Marseille, 1997), Notes Numer. Fluid Mech. 66, Vieweg, Braunschweig, 1998, pp. 222–241.
[20] S. Jin, *Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations*, SIAM J. Sci. Comput., 21 (1999), pp. 441–454.
[21] S. Jin and C. D. Levermore, *The discrete-ordinate method in diffusive Regimes*, Transport Theory Statist. Phys., 20 (1991), pp. 413–439.
[22] S. Jin and C. D. Levermore, *Fully discrete numerical transfer in diffusive regimes*, Transport Theory Statist. Phys., 22 (1993), pp. 739–791.
[23] S. Jin, L. Pareschi, and G. Toscani, *Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2405–2439.
[24] S. Jin, L. Pareschi, and G. Toscani, *Uniformly accurate diffusive relaxation schemes for multiscale transport equations*, SIAM J. Numer. Anal., 38 (2000), pp. 913–936.
[25] A. Klar, *Convergence of alternating domain decomposition schemes for kinetic and aerodynamic equations*, Math. Methods Appl. Sci., 18 (1995), pp. 649–670.
[26] A. Klar, *Asymptotic-induced domain decomposition methods for kinetic and drift diffusion semiconductor equations*, SIAM J. Sci. Comput., 19 (1998), pp. 2032–2050.
[27] A. Klar, *An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit*, SIAM J. Numer. Anal., 35 (1998), pp. 1073–1094.
[28] A. Klar, H. Neunzert, and J. Struckmeier, *Transition from kinetic theory to macroscopic fluid equations: A problem for domain decomposition and a source for new algorithm*, Transport Theory Statist. Phys., 29 (2000), pp. 93–106.
[29] A. Klar, and N. Siedow, *Boundary layers and domain decomposition for radiative heat transfer and diffusion equations: Applications to glass manufacturing process*, European J. Appl. Math., 9 (1998), pp. 351–372.
[30] E. W. Larsen, *The asymptotic diffusion limit of discretized transport problems*, Nucl. Sci. Eng., 112 (1992), pp. 336–346.
[31] E. W. Larsen and J. E. Morel, *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes* II, J. Comput. Phys., 83 (1989), pp. 212–236.
[32] E. W. Larsen, J. E. Morel, and W. F. Miller Jr., *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes*, J. Comput. Phys., 69 (1987), pp. 283–324.

[33] P. Le Tallec and F. Mallinger, *Coupling Boltzmann and Navier-Stokes equations by half fluxes*, J. Comput. Phys., 136 (1997), pp. 51–67.
[34] P. Le Tallec and M. Tidriri, *Convergence analysis of domain decomposition algorithms with full overlapping for the advection-diffusion problems*, Math. Comp., 68 (1999), pp. 585–606.
[35] E. E. Lewis and W. F. Miller Jr., *Computational Methods of Neutron Transport*, Wiley-Interscience, New York, 1984.
[36] W. F. Miller Jr., *An analysis of the finite-differenced, even-parity discrete-ordinate equations in slab geometry*, Nucl. Sci. Eng., 108 (1990), pp. 247–266.
[37] M. Tidriri, *New models for the solution of intermediate regimes in transport theory and radiative transfer: Existence theory, positivity, asymptotic analysis, and approximations*, J. Statist. Phys., 104 (2001), pp. 291–325.
[38] R. Vichnevtsky, *Stability charts in the numerical approximation of partial differential equations: A review*, Math. Comput. Simulation, 21 (1979), pp. 170–177.

SIAM J. NUMER. ANAL.
Vol. 42, No. 6, pp. 2688–2705

A LOW-ORDER NONCONFORMING FINITE ELEMENT FOR REISSNER–MINDLIN PLATES*

C. LOVADINA†

Abstract. We propose a locking-free element for plate bending problems, based on the use of nonconforming piecewise linear functions for both rotations and deflections. We prove optimal error estimates with respect to both the meshsize and the analytical solution regularity.

Key words. nonconforming finite elements, plate bending problems, convergence analysis

AMS subject classifications. Primary, 65N30; Secondary, 74S05

DOI. 10.1137/040603474

1. Introduction. Nowadays, a *wide* choice of reliable finite element schemes for the approximation of Reissner–Mindlin plate problems is available in the engineering and mathematical literature (see, for instance, [7], [8], [9], [10], [11], [14], [17], [26], [27], [28], [29], [30] and the references therein). However, the extension to the more complex (and more interesting) shell problems appears to be a difficult task. Indeed, only *very few and not completely satisfactory* results have been established in this direction (cf., e.g., [3], [19], [20], [21], [22], and [24]).

In this paper we propose and analyze a new low-order Reissner–Mindlin plate element, some properties of which seem to be favorable for its generalization to shell problems. This triangular mixed element can be considered as a simplified variant of the one presented in [18], and it is based on the use of nonconforming piecewise linear functions for both rotations and deflections, while the shear stresses are approximated by piecewise constant functions. In actual computations the shear stress variables can be easily eliminated at the element level, and the final system to be solved involves only rotation and deflection unknowns, which *share the same nodes* (the midpoints of the edges). Compared with the element detailed in [18], the one we are going to study has the following features:

- no additional bubble functions are required;
- no additional sophisticated "reduction" operator on the shear term (other than the simple L^2-projection operator on piecewise constant functions) needs to be introduced.

In view of a possible extension to shell problems, the promising features of our element are the same as the ones met by the scheme presented in [18], i.e.,

- it is a simple low-order method;
- once the shear stresses have been eliminated, all the variables into play share the same nodes;
- the element has optimal order of approximation and is locking-free.

An outline of the paper is as follows. In section 2 we briefly present the Reissner–Mindlin plate problem. In section 3 we introduce the nonconforming element, together

*Received by the editors January 23, 2004; accepted for publication (in revised form) July 26, 2004; published electronically April 19, 2005.
http://www.siam.org/journals/sinum/42-6/60347.html
†Dipartimento di Matematica, Università di Pavia, and IMATI-CNR, Via Ferrata 1, Pavia I-27100, Italy (carlo.lovadina@unipv.it). This research was partially supported by the European Project HPRN-CT-2002-00284, "New Materials, Adaptive Systems and Their Nonlinearities. Modelling, Control and Numerical Simulation."

with the necessary definitions and notation. In section 4 we develop the stability analysis, while in section 5 we perform the error analysis. The final results (cf. Theorem 5.1 and Corollary 5.1) show that our element is locking-free and is optimally convergent with respect to both the meshsize and the analytical solution regularity.

Furthermore, throughout the paper we will use standard notation for Sobolev spaces and norms (cf. [16] and [25], for instance). Finally, we will denote by C a generic constant, independent of h and t, which may differ in different occurrences.

2. The Reissner–Mindlin problem. The Reissner–Mindlin equations for a clamped plate with convex polygonal midplane Ω require us to find $(\boldsymbol{\theta}, w, \boldsymbol{\gamma})$ such that

$$-\operatorname{div} \mathbf{C}\,\varepsilon(\boldsymbol{\theta}) - \boldsymbol{\gamma} = 0 \quad \text{in } \Omega, \tag{2.1}$$

$$-\operatorname{div} \boldsymbol{\gamma} = g \quad \text{in } \Omega, \tag{2.2}$$

$$\boldsymbol{\gamma} = \lambda t^{-2}(\boldsymbol{\nabla} w - \boldsymbol{\theta}) \quad \text{in } \Omega, \tag{2.3}$$

$$\boldsymbol{\theta} = 0,\ w = 0 \quad \text{on } \partial\Omega. \tag{2.4}$$

In (2.1)–(2.3), t is the plate thickness, λ is the shear modulus, and $\mathbf{C}$ is the tensor of bending moduli, given by (for isotropic materials)

$$\mathbf{C}\tau := \frac{E}{12(1-\nu^2)}\Big((1-\nu)\tau + \nu \operatorname{tr}(\tau)\mathbf{I}\Big), \tag{2.5}$$

where τ is a generic second-order symmetric tensor, $\operatorname{tr}(\tau)$ its trace, $\mathbf{I}$ is the second-order identity tensor, while E and ν are Young's modulus and Poisson's ratio, respectively. Moreover, $\boldsymbol{\theta}$ represents the rotations, w the transversal displacement, $\boldsymbol{\gamma}$ the scaled shear stresses, and g a given transversal load. Finally, ε is the usual symmetric gradient operator. The classical variational formulation of problem (2.1)–(2.4) is

$$\begin{cases} \text{Find } (\boldsymbol{\theta}, w, \boldsymbol{\gamma}) \in \Theta \times W \times (L^2(\Omega))^2 : \\ a(\boldsymbol{\theta}, \boldsymbol{\eta}) + (\boldsymbol{\nabla} v - \boldsymbol{\eta}, \boldsymbol{\gamma}) = (g, v), \qquad (\boldsymbol{\eta}, v) \in \Theta \times W, \\ (\boldsymbol{\nabla} w - \boldsymbol{\theta}, \boldsymbol{\tau}) - \lambda^{-1} t^2 (\boldsymbol{\gamma}, \boldsymbol{\tau}) = 0, \qquad \boldsymbol{\tau} \in (L^2(\Omega))^2, \end{cases} \tag{2.6}$$

where $\Theta = (H_0^1(\Omega))^2$, $W = H_0^1(\Omega)$, $(\cdot,\cdot)$ is the inner-product in $L^2(\Omega)$, and

$$a(\boldsymbol{\theta}, \boldsymbol{\eta}) := \int_\Omega \mathbf{C}\,\varepsilon(\boldsymbol{\theta}) : \varepsilon(\boldsymbol{\eta}) \mathrm{d}x. \tag{2.7}$$

It is well known that for problem (2.6) the following inf-sup condition holds (cf. [16], for instance):

$$\begin{aligned} &\exists\, \beta > 0 \text{ such that} \\ &\sup_{(\boldsymbol{\eta}, v) \in \Theta \times W} \frac{(\boldsymbol{\nabla} v - \boldsymbol{\eta}, \boldsymbol{\tau})}{(||\boldsymbol{\eta}||_{1,\Omega}^2 + ||v||_{1,\Omega}^2)^{1/2}} \geq \beta ||\boldsymbol{\tau}||_\Gamma \qquad \forall \boldsymbol{\tau} \in \Gamma, \end{aligned} \tag{2.8}$$

where

$$\Gamma = H^{-1}(\mathrm{div}, \Omega) \quad \text{and} \quad ||\boldsymbol{\tau}||_{\Gamma} := (||\boldsymbol{\tau}||^2_{-1,\Omega} + ||\,\mathrm{div}\,\boldsymbol{\tau}||^2_{-1,\Omega})^{1/2}. \tag{2.9}$$

Moreover, the following regularity result is valid (cf., e.g., [7] and [23]).

PROPOSITION 2.1. *Suppose that Ω is a convex polygon and $g \in L^2(\Omega)$. Let $(\boldsymbol{\theta}, w, \boldsymbol{\gamma})$ be the solution of problem* (2.6)*. Then the following estimate holds:*

$$||\boldsymbol{\theta}||_{2,\Omega} + ||w||_{2,\Omega} + ||\boldsymbol{\gamma}||_{H(\mathrm{div})} + t||\boldsymbol{\gamma}||_{1,\Omega} \leq C||g||_{0,\Omega}, \tag{2.10}$$

where

$$||\boldsymbol{\gamma}||^2_{H(\mathrm{div})} = ||\boldsymbol{\gamma}||^2_{0,\Omega} + ||\,\mathrm{div}\,\boldsymbol{\gamma}||^2_{0,\Omega}.$$

3. The new nonconforming element. We now introduce a nonconforming finite element approximation of problem (2.1)–(2.4) using the approach detailed in [18]. Then let $\mathcal{T}_h$ be a decomposition of Ω into triangular elements T and let us set

$$H^1(\mathcal{T}_h) := \prod_{T \in \mathcal{T}_h} H^1(T). \tag{3.1}$$

We now define suitable jump and average operators. We first denote by $\mathcal{E}_h$ the set of all the edges in $\mathcal{T}_h$, and by $\mathcal{E}_h^{\mathrm{in}}$ the set of internal edges. Let e be an internal edge of $\mathcal{T}_h$, shared by two elements T^+ and T^-, and let φ denote a function in $H^1(\mathcal{T}_h)$, or a vector in $(H^1(\mathcal{T}_h))^2$, or a tensor in $(H^1(\mathcal{T}_h))^4_s$. We define the average as usual:

$$\{\varphi\} = \frac{\varphi^+ + \varphi^-}{2} \quad \forall e \in \mathcal{E}_h^{\mathrm{in}}. \tag{3.2}$$

For a scalar function $\varphi \in H^1(\mathcal{T}_h)$ we define its jump as

$$[\varphi] = \varphi^+ \mathbf{n}^+ + \varphi^- \mathbf{n}^- \quad \forall e \in \mathcal{E}_h^{\mathrm{in}}, \tag{3.3}$$

while the jump of a vector $\boldsymbol{\varphi} \in (H^1(\mathcal{T}_h))^2$ is given by

$$[\boldsymbol{\varphi}] = (\boldsymbol{\varphi}^+ \otimes \mathbf{n}^+)_S + (\boldsymbol{\varphi}^- \otimes \mathbf{n}^-)_S \quad \forall e \in \mathcal{E}_h^{\mathrm{in}}, \tag{3.4}$$

where $(\boldsymbol{\varphi} \otimes \mathbf{n})_S$ denotes the symmetric part of the tensor product, and $\mathbf{n}^+$ (resp., $\mathbf{n}^-$) is the outward unit normal to ∂T^+ (resp., to ∂T^-). On the boundary edges we define jumps of scalars as $[\varphi] = \varphi \mathbf{n}$, and jumps of vectors as $[\boldsymbol{\varphi}] = (\boldsymbol{\varphi} \otimes \mathbf{n})_S$, where $\mathbf{n}$ is the outward unit normal to $\partial\Omega$. We also define averages of vectors and tensors as $\{\boldsymbol{\varphi}\} = \boldsymbol{\varphi}$. It can be easily checked that, if $\boldsymbol{\varphi}$ is a smooth tensor and $\boldsymbol{\eta}$ a piecewise smooth vector, the following equality holds (see, e.g., [4] and [5] for a similar computation):

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} \boldsymbol{\varphi}\mathbf{n} \cdot \boldsymbol{\eta}\, ds = \sum_{e \in \mathcal{E}_h} \int_e \{\boldsymbol{\varphi}\} : [\boldsymbol{\eta}]\, ds. \tag{3.5}$$

In order to introduce our scheme, we first consider the finite element spaces:

$$\Theta_h = \left\{ \boldsymbol{\eta} : \boldsymbol{\eta}_{|T} \in (P_1(T))^2, \ \int_e [\boldsymbol{\eta}] ds = 0 \quad \forall e \in \mathcal{E}_h \right\}, \tag{3.6}$$

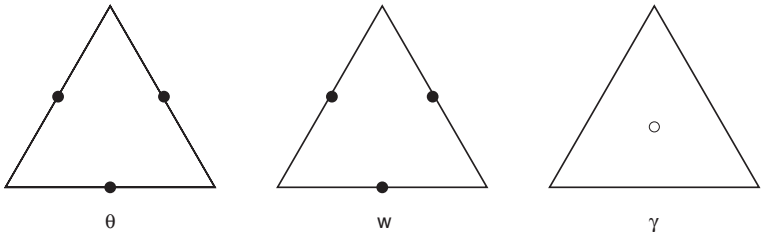


FIG. 3.1. *Local degrees of freedom for the three variables.*

$$W_h = \left\{ v : v_{|T} \in P_1(T), \ \int_e [v] ds = 0 \quad \forall e \in \mathcal{E}_h \right\}, \tag{3.7}$$

$$\Gamma_h = \left\{ \boldsymbol{\tau} : \boldsymbol{\tau}_{|T} \in (P_0(T))^2 \right\}, \tag{3.8}$$

where $P_k(T)$ is the space of polynomials of degree at most k defined on T. We also notice that

$$\boldsymbol{\nabla}_h W_h \subset \Gamma_h, \tag{3.9}$$

where $\boldsymbol{\nabla}_h$ denotes the gradient element by element. The local degrees of freedom for the three variables are depicted in Figure 3.1.

Moreover, we introduce a penalty on the jumps of functions in Θ_h as

$$p_\Theta(\boldsymbol{\theta}, \boldsymbol{\eta}) := \sum_{e \in \mathcal{E}_h} \frac{\kappa_e}{|e|} \int_e [\boldsymbol{\theta}] : [\boldsymbol{\eta}] \, ds, \tag{3.10}$$

where $|e|$ denotes the length of the side e, and κ_e is a positive constant having the same physical dimension as $\mathbf{C}$ (for smooth $\mathbf{C}$, one could take κ_e as $|\mathbf{C}|$ evaluated at the midpoint of e). We then define

$$a_T(\boldsymbol{\theta}, \boldsymbol{\eta}) := \int_T \mathbf{C}\, \varepsilon(\boldsymbol{\theta}) : \varepsilon(\boldsymbol{\eta}) \, dx, \tag{3.11}$$

and we finally set

$$a_h(\boldsymbol{\theta}, \boldsymbol{\eta}) := \sum_{T \in \mathcal{T}_h} a_T(\boldsymbol{\theta}, \boldsymbol{\eta}) + p_\Theta(\boldsymbol{\theta}, \boldsymbol{\eta}). \tag{3.12}$$

Following the ideas of [18], the discrete problem is then

$$\begin{cases} \text{Find } (\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h) \in \Theta_h \times W_h \times \Gamma_h : \\ a_h(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) + (\boldsymbol{\gamma}_h, \boldsymbol{\nabla}_h v_h - \boldsymbol{\eta}_h) = (g, v_h), \quad (\boldsymbol{\eta}_h, v_h) \in \Theta_h \times W_h, \\ (\boldsymbol{\nabla}_h w_h - \boldsymbol{\theta}_h, \boldsymbol{\tau}_h) - \lambda^{-1} t^2 (\boldsymbol{\gamma}_h, \boldsymbol{\tau}_h) = 0, \qquad \boldsymbol{\tau}_h \in \Gamma_h. \end{cases} \tag{3.13}$$

We will use norms $||\cdot||_{\Theta_h}$ and $||\cdot||_{W_h}$ for functions in Θ_h and W_h, defined as

$$||\boldsymbol{\eta}_h||_{\Theta_h} := \left(\sum_{T \in \mathcal{T}_h} \int_T |\boldsymbol{\nabla} \boldsymbol{\eta}_h|^2 \right)^{1/2} = ||\boldsymbol{\nabla}_h \boldsymbol{\eta}_h||_{0,\Omega}, \tag{3.14}$$

$$||v_h||_{W_h} := \left(\sum_{T \in \mathcal{T}_h} \int_T |\boldsymbol{\nabla} v_h|^2 \right)^{1/2} = ||\boldsymbol{\nabla}_h v_h||_{0,\Omega}. \tag{3.15}$$

Due to the discrete Poincaré inequality, both $||\cdot||_{\Theta_h}$ and $||\cdot||_{W_h}$ are indeed *norms* on Θ_h and W_h, not only seminorms.

Remark 3.1. The discrete Poincaré inequality is known to hold for the *scalar-valued* space W_h, as detailed, for instance, in [7] (see also [6] and [12] for related and more general results). However, the same techniques can be applied to obtain the corresponding version for the *vector field* space Θ_h.

It has been proved in [6] (see also [13]) that there exist positive constants α and M such that

$$a_h(\boldsymbol{\eta}_h, \boldsymbol{\eta}_h) \geq \alpha ||\boldsymbol{\eta}_h||^2_{\Theta_h} \qquad \forall\, \boldsymbol{\eta}_h \in \Theta_h, \tag{3.16}$$

$$a_h(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) \leq M ||\boldsymbol{\theta}_h||_{\Theta_h} ||\boldsymbol{\eta}_h||_{\Theta_h} \qquad \forall\, \boldsymbol{\theta}_h\,, \boldsymbol{\eta}_h \in \Theta_h. \tag{3.17}$$

For functions in Γ_h we will work with the (natural) norms (cf. also (2.9))

$$||\boldsymbol{\tau}_h||_\Gamma \qquad \text{and} \qquad t||\boldsymbol{\tau}_h||_{0,\Omega}. \tag{3.18}$$

Remark 3.2. We remark that the coercivity property (3.16) can be easily deduced from the nontrivial Korn-type inequality (see [6] and [13])

$$\|\boldsymbol{\nabla}_h \boldsymbol{\eta}\|^2_{0,\Omega} \leq C \left(\|\varepsilon_h(\boldsymbol{\eta})\|^2_{0,\Omega} + \sum_{e \in \mathcal{E}_h} \frac{1}{|e|} \|[\boldsymbol{\eta}]\|^2_{0,e} \right), \qquad \boldsymbol{\eta} \in (H^1(\mathcal{T}_h))^2, \tag{3.19}$$

where ε_h denotes the symmetric gradient element by element. We point out that estimate (3.19) holds for a *generic* piecewise smooth function $\boldsymbol{\eta} \in (H^1(\mathcal{T}_h))^2$, not only for functions in Θ_h. Therefore, the coercivity of the form $a_h(\cdot,\cdot)$ is valid essentially for all the boundary conditions arising in actual applications. We finally stress that property (3.16) cannot hold without inserting the jump term (3.10) into the bilinear form $a_h(\cdot,\cdot)$ (cf. (3.12)). More precisely, the form

$$\widetilde{a}_h(\boldsymbol{\theta}, \boldsymbol{\eta}) := \int_\Omega \mathbf{C} \varepsilon_h(\boldsymbol{\theta}) : \varepsilon_h(\boldsymbol{\eta})\, dx \tag{3.20}$$

is *not* coercive on the nonconforming space Θ_h. Indeed, consider a square plate meshed into four triangles by means of the diagonals. The *nonvanishing* function $\boldsymbol{\eta}_h \in \Theta_h$ shown in Figure 3.2 verifies $\varepsilon_h(\boldsymbol{\eta}_h) = 0$, since in each element it represents an infinitesimal rotation about the corresponding boundary edge midpoint. As a consequence, coercivity for $\widetilde{a}_h(\cdot,\cdot)$ fails.

Remark 3.3. For Dirichlet boundary conditions on the whole $\partial\Omega$, one might take advantage of the relation

$$\operatorname{div} \varepsilon(\boldsymbol{\theta}) = \frac{1}{2} \left(\operatorname{div} \boldsymbol{\nabla} \boldsymbol{\theta} + \boldsymbol{\nabla} \operatorname{div} \theta \right) \tag{3.21}$$

to write the discrete problem (3.13) using the *modified* bilinear form (cf. also (2.5))

$$a_m(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) := \frac{E}{24(1-\nu^2)} \sum_{T \in \mathcal{T}_h} \int_T \Big((1-\nu) \boldsymbol{\nabla} \boldsymbol{\theta}_h : \boldsymbol{\nabla} \boldsymbol{\eta}_h + (1+\nu) \operatorname{div} \boldsymbol{\theta}_h \operatorname{div} \boldsymbol{\eta}_h \Big). \tag{3.22}$$

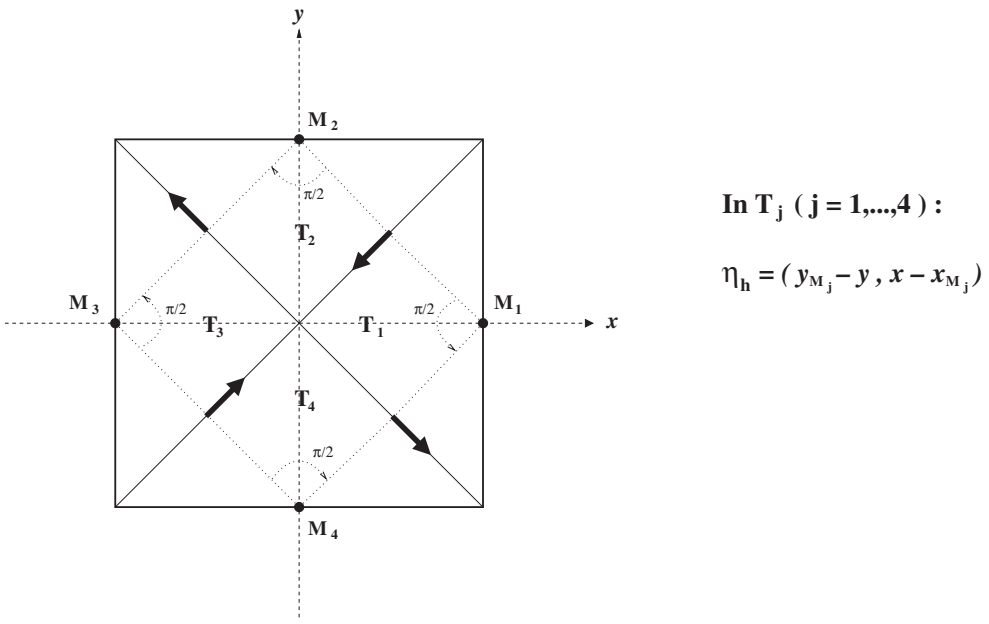


FIG. 3.2. *Rotational spurious mode* $\boldsymbol{\eta}_h$ *on a mesh with four triangles.*

Due to the discrete Poincaré inequality (see Remark 3.1), the bilinear form $a_m(\cdot,\cdot)$ is now coercive on Θ_h and we do not need to insert the penalty term (3.10) in the finite element method. We remark, however, that this approach is limited to the case of Dirichlet boundary conditions on the whole $\partial\Omega$. Since we have in mind a method applicable to *every* reasonable boundary condition (we consider the clamped plate *only* for the sake of simplicity), we have chosen to work with a formulation involving the bilinear form $a_h(\cdot,\cdot)$ (see (3.10)–(3.12) and Remark 3.2).

Remark 3.4. We point out that by eliminating $\boldsymbol{\gamma}_h$ from system (3.13), our scheme is equivalent to the following problem involving only the rotations and the vertical displacements:

$$\text{(3.23)}\quad \begin{cases} \text{Find } (\boldsymbol{\theta}_h, w_h) \in \Theta_h \times W_h : \\ a_h(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) + \lambda t^{-2}(\boldsymbol{\nabla}_h w_h - P_0\boldsymbol{\theta}_h, \boldsymbol{\nabla}_h v_h - P_0\boldsymbol{\eta}_h) \\ \qquad\qquad = (g, v_h) \qquad \forall (\boldsymbol{\eta}_h, v_h) \in \Theta_h \times W_h, \end{cases}$$

where P_0 denotes the L^2-projection operator on the piecewise constant functions. From (3.23) we may notice that the method implementation turns out to be rather simple.

4. Stability analysis. In this section we will prove a stability result for the discretized problem (3.13), using a macroelement technique essentially developed in [28]. In what follows it will be useful to set $V := \Theta \times W$ and $V_h := \Theta_h \times W_h$, equipped with the usual product norms. We first need the following preliminary result.

PROPOSITION 4.1. *The approximation spaces defined in* (3.6)–(3.8) *satisfy the following properties:*

(P1) *There exists a linear operator* $\pi_h : W \longrightarrow W_h$ *such that*

$$||\pi_h v||_{W_h} \leq c||v||_{1,\Omega},\ c \text{ independent of } h$$
$$\int_\Omega \boldsymbol{\nabla}_h(v - \pi_h v)\cdot\boldsymbol{\tau}_h = 0 \qquad \forall \boldsymbol{\tau}_h \in \Gamma_h.$$

(P2) *If the mesh* $\mathcal{T}_h$ *contains at least three triangles, then for* $\boldsymbol{\tau}_h \in \Gamma_h$ *condition*

$$\int_\Omega (\boldsymbol{\nabla}_h v_h - \boldsymbol{\eta}_h)\cdot\boldsymbol{\tau}_h = 0 \qquad \forall (\boldsymbol{\eta}_h, v_h) \in V_h \tag{4.1}$$

implies $\boldsymbol{\tau}_h = \mathbf{0}$.

Proof. Consider the usual nonconforming interpolating operator $\pi_h : W \longrightarrow W_h$, defined by

$$(\pi_h v)(m) = \frac{1}{|e|}\int_e v\, ds \qquad \forall e \in \mathcal{E}_h \quad (\text{with } m \text{ the midpoint of } e).$$

It is easily seen that property (P1) is fulfilled.

To verify (P2), for a given internal edge $e \in \mathcal{E}_h^{\text{in}}$ we first choose one of the two possible normal (resp., tangential) vectors to e, indicated in what follows as $\mathbf{n}_e$ (resp., $\mathbf{t}_e$). Let us take $\boldsymbol{\tau}_h \in \Gamma_h$ satisfying condition (4.1).

By choosing $(\mathbf{0}, v_h) \in V_h$, integrating by parts yields

$$0 = \int_\Omega \boldsymbol{\nabla}_h v_h \cdot \boldsymbol{\tau}_h = \sum_{T\in\mathcal{T}_h}\int_{\partial T} v_h\, \boldsymbol{\tau}_h\cdot\mathbf{n}_T. \tag{4.2}$$

Since (4.2) is true for every $(\mathbf{0}, v_h) \in V_h$, it follows that $\boldsymbol{\tau}_h\cdot\mathbf{n}_e$ is continuous across every internal edge $e \in \mathcal{E}_h^{\text{in}}$. Therefore $\boldsymbol{\tau}_h \in H(\text{div};\Omega)$ and, obviously, $\text{div}\,\boldsymbol{\tau}_h = 0$. As a consequence, there exists φ_h (defined up to a constant) such that

$$\varphi_h \in \mathcal{L}_1^1(\Omega;\mathcal{T}_h), \qquad \boldsymbol{\tau}_h = \mathbf{curl}\,\varphi_h, \tag{4.3}$$

where $\mathcal{L}_1^1(\Omega;\mathcal{T}_h)$ is the usual space of piecewise linear and continuous functions on Ω.

Fix now a generic internal edge $e \in \mathcal{E}_h^{\text{in}}$ with midpoint m, and denote with T_e^+, T_e^- the triangles sharing e as common side. Recalling that $\mathbf{curl}\,\varphi_h\cdot\mathbf{n}_e$ is constant and continuous across e, we consider $(\boldsymbol{\eta}_h, 0) \in V_h$, where $\boldsymbol{\eta}_h$ is uniquely defined by

$$\begin{cases} (\boldsymbol{\eta}_h\cdot\mathbf{t}_e)(m) = 0, \quad (\boldsymbol{\eta}_h\cdot\mathbf{n}_e)(m) = \mathbf{curl}\,\varphi_h\cdot\mathbf{n}_e, \\ \boldsymbol{\eta}_h(m') = 0 \qquad \forall\, e' \in \mathcal{E}_h^{\text{in}}, \quad e' \neq e \quad (\text{with } m' \text{ the midpoint of } e'). \end{cases} \tag{4.4}$$

Since $\boldsymbol{\tau}_h = \mathbf{curl}\,\varphi_h$ satisfies (4.1), using (4.4) we have

$$\begin{aligned} 0 &= \int_\Omega \boldsymbol{\eta}_h\cdot\mathbf{curl}\,\varphi_h = \int_{T_e^+\cup T_e^-}\boldsymbol{\eta}_h\cdot\mathbf{curl}\,\varphi_h = \frac{|T_e^+|+|T_e^-|}{3}(\boldsymbol{\eta}_h\cdot\mathbf{curl}\,\varphi_h)(m) \\ &= \frac{|T_e^+|+|T_e^-|}{3}\,|\mathbf{curl}\,\varphi_h\cdot\mathbf{n}_e|^2. \end{aligned} \tag{4.5}$$

Repeating the same argument for every $e \in \mathcal{E}_h^{\text{in}}$, from (4.5) we infer that

$$\mathbf{curl}\,\varphi_h\cdot\mathbf{n}_e = \boldsymbol{\nabla}\varphi_h\cdot\mathbf{t}_e = 0 \qquad \text{for every } e \in \mathcal{E}_h^{\text{in}}. \tag{4.6}$$

Equation (4.6) implies that $\boldsymbol{\tau}_h = \mathbf{curl}\,\varphi_h$ vanishes in all the triangles $T \in \mathcal{T}_h$ having at least two sides in $\mathcal{E}_h^{\text{in}}$. Therefore, it remains to show that $\mathbf{curl}\,\varphi_h = \mathbf{0}$ also on the triangles sharing two sides with the boundary $\partial\Omega$ if there are any in the mesh $\mathcal{T}_h$. Consider then any such triangle T, and denote with e its *unique* side belonging to $\mathcal{E}_h^{\text{in}}$ and with T^{in} the triangle sharing the side e with T. Since Ω is a regular domain and $\mathcal{T}_h$ contains at least three triangles, it follows that T^{in} has at least two sides in $\mathcal{E}_h^{\text{in}}$. Hence we already know that

$$(\mathbf{curl}\,\varphi_h)_{|T^{\text{in}}} = \mathbf{0}. \tag{4.7}$$

Recalling that $\mathbf{curl}\,\varphi_h$ is constant in T, let us now take $(\boldsymbol{\eta}_h, 0) \in V_h$, where $\boldsymbol{\eta}_h$ is uniquely defined by

$$\begin{cases} \boldsymbol{\eta}_h(m) = (\mathbf{curl}\,\varphi_h)_{|T} & \text{(with } m \text{ the midpoint of } e\text{)}, \\ \boldsymbol{\eta}_h(m') = 0 \qquad \forall\, e' \in \mathcal{E}_h^{\text{in}}\,, \quad e' \neq e & \text{(with } m' \text{ the midpoint of } e'\text{)}. \end{cases} \tag{4.8}$$

Again, since $\boldsymbol{\tau}_h = \mathbf{curl}\,\varphi_h$ satisfies (4.1), by (4.7) and (4.8) we obtain

$$\begin{aligned} 0 &= \int_\Omega \boldsymbol{\eta}_h \cdot \mathbf{curl}\,\varphi_h = \int_{T \cup T^{\text{in}}} \boldsymbol{\eta}_h \cdot \mathbf{curl}\,\varphi_h = \int_T \boldsymbol{\eta}_h \cdot \mathbf{curl}\,\varphi_h \\ &= \frac{|T|}{3}\,|(\mathbf{curl}\,\varphi_h)_{|T}|^2, \end{aligned} \tag{4.9}$$

so that $\mathbf{curl}\,\varphi_h = \mathbf{0}$ also in T and the proof is complete. ∎

Remark 4.1. We remark that property (P2) can be written in the following equivalent form:

(P2′) For every $\boldsymbol{\varphi} \in (L^2(\Omega))^2$, the problem

$$\begin{cases} \text{Find } (\boldsymbol{\eta}_h, v_h) \in \Theta_h \times W_h : \\ \displaystyle\int_\Omega (\boldsymbol{\nabla}_h v_h - \boldsymbol{\eta}_h) \cdot \boldsymbol{\tau}_h = \int_\Omega \boldsymbol{\varphi} \cdot \boldsymbol{\tau}_h \qquad \forall \boldsymbol{\tau}_h \in \Gamma_h \end{cases}$$

is solvable.

4.1. Macroelement decomposition. We start by recalling some standard definitions and notation we will use in what follows. First of all, we say that a family $\{\mathcal{T}_h\}_{h>0}$ of triangular meshes of Ω is *regular* (see [25]) if there exists a constant $\sigma > 0$ such that

$$h_T \leq \sigma \rho_T \qquad \forall\, T \in \bigcup_{h>0} \mathcal{T}_h, \tag{4.10}$$

where h_T is the diameter of the element T and ρ_T is the maximum diameter of the circles contained in T. Furthermore, a *macroelement* M is a set with connected interior part, formed by the union of a fixed number of neighboring triangles along a well-defined pattern (cf. [31]). A macroelement $M = \cup_{i=1}^m T_i$ is said to be equivalent to a reference macroelement $\widehat{M} = \cup_{i=1}^m \widehat{T}_i$ if there is a mapping $F_M : \widehat{M} \longrightarrow M$ for which the following conditions are fulfilled (cf. [31]):

1. F_M is a continuous bijection.
2. $T_i = F_M(\widehat{T}_i)\ \forall\, i,\ 1 \le i \le m$.
3. $F_{M|\widehat{T}_i} = F_{T_i} \circ F_{\widehat{T}_i}^{-1}$, where F_{T_i} and $F_{\widehat{T}_i}$ are the usual functions mapping the standard reference triangle (of vertices $(0,0)$, $(1,0)$, and $(0,1)$) onto T_i and $\widehat{T}_i$, respectively.

From a given mesh $\mathcal{T}_h$ of Ω it is always possible to derive (obviously not in a unique manner) a "macroelement mesh" $\mathcal{M}_h$ in such a way that each $T \in \mathcal{T}_h$ is covered by some macroelement M in $\mathcal{M}_h$ and each macroelement M is equivalent to a certain reference macroelement $\widehat{M}$.

Associated with every macroelement M in $\mathcal{M}_h$, the following spaces are relevant for the stability analysis (cf. [28]):

$$V_{0,M} := \big\{(\boldsymbol{\eta}_h, v_h) \in V_h : \quad (\boldsymbol{\eta}_h, v_h) = (\mathbf{0}, 0) \quad \text{in}\,\Omega \setminus M\big\}, \tag{4.11}$$

$$\Gamma_M := \big\{\boldsymbol{\tau}_h \in \Gamma_h : \quad \boldsymbol{\tau}_h = 0 \quad \text{in}\,\Omega \setminus M\big\}. \tag{4.12}$$

4.2. Fortin's trick by macroelements. The aim of this subsection is to prove that Fortin's trick (cf. [16]) applies to our finite element scheme, leading therefore to a suitable inf-sup condition with respect to the natural norms (see (2.8)). Indeed, we have the following result.

PROPOSITION 4.2. *Suppose that the family $\{\mathcal{T}_h\}_{h>0}$ is regular and choose a corresponding macroelement family $\{\mathcal{M}_h\}_{h>0}$ such that*

1. *each macroelement M contains at least three triangles;*
2. *there is only a fixed finite number of reference macroelements $\{\widehat{M}_1, \dots, \widehat{M}_r\}$ to which each macroelement $M \in \cup_{h>0}\mathcal{M}_h$ is equivalent.*

Then for the approximation spaces defined in (3.6)–(3.8) *the following inf-sup condition holds*

$$\exists\, \beta > 0 \text{ independent of } h, \text{ such that}$$
$$\sup_{(\boldsymbol{\eta}_h, v_h) \in V_h} \frac{(\boldsymbol{\nabla}_h v_h - \boldsymbol{\eta}_h, \boldsymbol{\tau}_h)}{||(\boldsymbol{\eta}_h, v_h)||_{V_h}} \ge \beta ||\boldsymbol{\tau}_h||_\Gamma \qquad \forall \boldsymbol{\tau}_h \in \Gamma_h. \tag{4.13}$$

Proof. Let $(\boldsymbol{\eta}, v) \in V$ be given. Fix an arbitrary macroelement $M \in \mathcal{M}_h$ and set

$$h_M := \max_{1 \le i \le m} h_{T_i} \quad \text{if} \quad M = \bigcup_i^m T_i. \tag{4.14}$$

Let us denote by i_M the index $1 \le i_M \le r$ such that M is equivalent to $\widehat{M}_{i_M}$.

Consider the problem of finding $(\boldsymbol{\eta}_M, v_M) \in V_{0,M}$ solution of

$$\int_M (\boldsymbol{\nabla}_h v_M - \boldsymbol{\eta}_M) \cdot \boldsymbol{\tau}_M = \int_M (\Pi_1 \boldsymbol{\eta} - \boldsymbol{\eta}) \cdot \boldsymbol{\tau}_M \qquad \forall \boldsymbol{\tau}_M \in \Gamma_M, \tag{4.15}$$

where $\Pi_1 \boldsymbol{\eta}$ is the usual nonconforming interpolated of $\boldsymbol{\eta}$, defined by

$$(\Pi_1 \boldsymbol{\eta})(m) = \frac{1}{|e|} \int_e \boldsymbol{\eta}\, ds \qquad \forall e \in \mathcal{E}_h \quad (\text{with } m \text{ the midpoint of } e).$$

By property (P2) of Proposition 4.1, applied to the macroelement M, it follows that system (4.15) is solvable, since M contains at least three triangles (cf. also

Remark 4.1). Let us take the solution of minimal V_h-norm. A scaling argument and the features of the interpolating operator Π_1 show that there exists $c(\widehat{M}_{i_M}) > 0$ such that

$$\exists\, c_1(\widehat{M}_{i_M}) > 0 : \quad ||\,\boldsymbol{\eta}_M\,||^2_{\Theta_h} + ||\,v_M\,||^2_{W_h} \le c_1(\widehat{M}_{i_M})\,||\,\boldsymbol{\eta}\,||^2_{1,M}. \tag{4.16}$$

Let us set

$$\boldsymbol{\eta}_F = \Pi_1\boldsymbol{\eta} + \sum_M \boldsymbol{\eta}_M, \tag{4.17}$$

$$v_F = \pi_h v + \sum_M v_M, \tag{4.18}$$

where π_h is the operator as in property (P1) of Proposition 4.1 (i.e., the standard nonconforming interpolation operator). We now notice that every $\boldsymbol{\tau}_h \in \Gamma_h$ can be uniquely written as $\boldsymbol{\tau}_h = \sum_M \boldsymbol{\tau}_M$, where $\boldsymbol{\tau}_M \in \Gamma_M$. Hence, recalling (4.15), from (4.17)–(4.18) we have

$$\begin{aligned}
\int_\Omega (\boldsymbol{\nabla}_h v_F - \boldsymbol{\eta}_F)\cdot\boldsymbol{\tau}_h &= \sum_M \int_M \Big[\,\boldsymbol{\nabla}_h(\pi_h v + v_M) - \Pi_1\boldsymbol{\eta} - \boldsymbol{\eta}_M\,\Big]\cdot\boldsymbol{\tau}_M \\
&= \sum_M \Big[\int_M \boldsymbol{\nabla}_h\pi_h v\cdot\boldsymbol{\tau}_M + \int_M (\boldsymbol{\nabla}_h v_M - \Pi_1\boldsymbol{\eta} - \boldsymbol{\eta}_M)\cdot\boldsymbol{\tau}_M\Big] \\
&= \sum_M \Big(\int_M \boldsymbol{\nabla}\, v\cdot\boldsymbol{\tau}_M - \int_M \boldsymbol{\eta}\cdot\boldsymbol{\tau}_M\Big) \\
&= \int_\Omega (\boldsymbol{\nabla}\, v - \boldsymbol{\eta})\cdot\boldsymbol{\tau}_h.
\end{aligned} \tag{4.19}$$

Therefore, for every $(\boldsymbol{\eta}, v) \in V$ we have found $\Pi_h(\boldsymbol{\eta}, v) = (\boldsymbol{\eta}_F, v_F) \in V_h$ such that

$$\int_\Omega (\boldsymbol{\nabla}_h v_F - \boldsymbol{\eta}_F)\cdot\boldsymbol{\tau}_h = \int_\Omega (\boldsymbol{\nabla}\, v - \boldsymbol{\eta})\cdot\boldsymbol{\tau}_h \qquad \forall\boldsymbol{\tau}_h \in \Gamma_h. \tag{4.20}$$

Let us estimate $||\boldsymbol{\eta}_F||^2_{\Theta_h} + ||v_F||^2_{W_h}$. By using the continuity of Π_1 and π_h, and estimate (4.16), we get

$$\begin{aligned}
||\boldsymbol{\eta}_F||^2_{\Theta_h} + ||v_F||^2_{W_h} &= \Big\|\,\Pi_1\boldsymbol{\eta} + \sum_M \boldsymbol{\eta}_M\Big\|^2_{\Theta_h} + \Big\|\,\pi_h v + \sum_M v_M\Big\|^2_{W_h} \\
&\le 2\left(\,||\Pi_1\boldsymbol{\eta}||^2_{\Theta_h} + ||\pi_h v||^2_{W_h} + \sum_M \big(||\,\boldsymbol{\eta}_M\,||^2_{\Theta_h} + ||\,v_M\,||^2_{W_h}\big)\right) \\
&\le 2\left(\,c\,\big(||v||^2_{1,\Omega} + ||\boldsymbol{\eta}||^2_{1,\Omega}\big) + \sum_M c_1(\widehat{M}_{i_M})||\boldsymbol{\eta}||^2_{1,M}\right).
\end{aligned} \tag{4.21}$$

Since there is only a finite number of reference macroelements $\{\widehat{M}_1, \ldots, \widehat{M}_r\}$, we obtain

$$||\boldsymbol{\eta}_F||^2_{\Theta_h} + ||v_F||^2_{W_h} \le C_1\big(||v||^2_{1,\Omega} + ||\boldsymbol{\eta}||^2_{1,\Omega}\big) \tag{4.22}$$

with $C_1 = 2\max\{c,\ c_1(\widehat{M}_1), \ldots, c_1(\widehat{M}_r)\}$. Therefore, we finally have

$$||\Pi_h(\boldsymbol{\eta}, v)||_{V_h} \leq C\left(||\boldsymbol{\eta}||^2_{1,\Omega} + ||v||^2_{1,\Omega}\right)^{1/2} \tag{4.23}$$

with C independent of h. It is well known (cf. [16], for instance) that (4.20) together with (4.23) implies condition (4.13), and the proof is complete. ∎

Remark 4.2. Note that it is always possible to derive, from a given regular family $\{\mathcal{T}_h\}_{h>0}$, a macroelement family $\{\mathcal{M}_h\}_{h>0}$ which fulfills the assumption of Proposition 4.2, provided in each $\mathcal{T}_h$ there are at least three triangles.

4.3. The stability result. Once the inf-sup condition (4.13) has been established, suitable stability estimates can be derived using standard techniques (see, for instance, [9] and [23] for their application to Reissner–Mindlin plate problems). For the sake of completeness, we develop such a stability analysis in full detail.

First, it is useful to set

$$\begin{aligned}\mathcal{A}_h(\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h; \boldsymbol{\eta}_h, v_h, \boldsymbol{\tau}_h) := a_h(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) + (\boldsymbol{\nabla}_h v_h - \boldsymbol{\eta}_h, \boldsymbol{\gamma}_h)\\ - (\boldsymbol{\nabla}_h w_h - \boldsymbol{\theta}_h, \boldsymbol{\tau}_h) + \lambda^{-1}t^2(\boldsymbol{\gamma}_h, \boldsymbol{\tau}_h).\end{aligned} \tag{4.24}$$

Therefore, the discrete problem (3.13) reads

$$\begin{cases}\text{Find } (\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h) \in \Theta_h \times W_h \times \Gamma_h \text{ such that}\\ \mathcal{A}_h(\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h; \boldsymbol{\eta}_h, v_h, \boldsymbol{\tau}_h) = (g, v_h) \qquad \forall (\boldsymbol{\eta}_h, v_h, \boldsymbol{\tau}_h) \in \Theta_h \times W_h \times \Gamma_h.\end{cases} \tag{4.25}$$

We have the following result.

PROPOSITION 4.3. *Given* $(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h) \in \Theta_h \times W_h \times \Gamma_h$, *there exists* $(\boldsymbol{\eta}_h, v_h, \boldsymbol{\tau}_h) \in \Theta_h \times W_h \times \Gamma_h$ *such that*

$$\begin{aligned}||\boldsymbol{\eta}_h||_{\Theta_h} + ||v_h||_{W_h} + ||\boldsymbol{\tau}_h||_\Gamma + t||\boldsymbol{\tau}_h||_{0,\Omega}\\ \leq C\Big(||\boldsymbol{\beta}_h||_{\Theta_h} + ||z_h||_{W_h} + ||\boldsymbol{\rho}_h||_\Gamma + t||\boldsymbol{\rho}_h||_{0,\Omega}\Big)\end{aligned} \tag{4.26}$$

and

$$\mathcal{A}_h(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h; \boldsymbol{\eta}_h, v_h, \boldsymbol{\tau}_h) \geq C\Big(||\boldsymbol{\beta}_h||^2_{\Theta_h} + ||z_h||^2_{W_h} + ||\boldsymbol{\rho}_h||^2_\Gamma + t^2||\boldsymbol{\rho}_h||^2_{0,\Omega}\Big). \tag{4.27}$$

Proof. Let $(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h)$ be given in $\Theta_h \times W_h \times \Gamma_h$. The proof is performed in three steps.

Step 1. Let us first choose $(\boldsymbol{\eta}_1, v_1, \boldsymbol{\tau}_1) \in \Theta_h \times W_h \times \Gamma_h$ as

$$\boldsymbol{\eta}_1 = \boldsymbol{\beta}_h, \qquad v_1 = z_h, \qquad \boldsymbol{\tau}_1 = \boldsymbol{\rho}_h.$$

It is obvious that

$$\begin{aligned}||\boldsymbol{\eta}_1||_{\Theta_h} + ||v_1||_{W_h} + ||\boldsymbol{\tau}_1||_\Gamma + t||\boldsymbol{\tau}_1||_{0,\Omega}\\ = ||\boldsymbol{\beta}_h||_{\Theta_h} + ||z_h||_{W_h} + ||\boldsymbol{\rho}_h||_\Gamma + t||\boldsymbol{\rho}_h||_{0,\Omega}.\end{aligned} \tag{4.28}$$

Furthermore, it holds that

$$\mathcal{A}_h(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h; \boldsymbol{\eta}_1, v_1, \boldsymbol{\tau}_1) = a_h(\boldsymbol{\beta}_h, \boldsymbol{\beta}_h) + \lambda^{-1}t^2||\boldsymbol{\gamma}_h||^2_{0,\Omega}. \tag{4.29}$$

By the coercivity of $a_h(\cdot,\cdot)$ (cf. (3.16)) it follows that

$$\mathcal{A}_h(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h; \boldsymbol{\eta}_1, v_1, \boldsymbol{\tau}_1) \geq C_1\Big(||\boldsymbol{\beta}_h||^2_{\Theta_h} + t^2||\boldsymbol{\rho}_h||^2_0\Big). \tag{4.30}$$

Step 2. Notice that from (4.13) it follows that there exists $(\boldsymbol{\eta}_2, v_2) \in \Theta_h \times W_h$ such that

$$||\boldsymbol{\eta}_2||_{\Theta_h} + ||v_2||_{W_h} \leq C||\boldsymbol{\rho}_h||_\Gamma \tag{4.31}$$

and

$$(\boldsymbol{\nabla}_h v_2 - \boldsymbol{\eta}_2, \boldsymbol{\rho}_h) = ||\boldsymbol{\rho}_h||^2_\Gamma. \tag{4.32}$$

Choose $(\boldsymbol{\eta}_2, v_2, \boldsymbol{\tau}_2) \in \Theta_h \times W_h \times \Gamma_h$ with $\boldsymbol{\tau}_2 = 0$. We have

$$\mathcal{A}_h(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h; \boldsymbol{\eta}_2, v_2, \boldsymbol{\tau}_2) = a_h(\boldsymbol{\beta}_h, \boldsymbol{\eta}_2) + (\boldsymbol{\nabla}_h v_2 - \boldsymbol{\eta}_2, \boldsymbol{\rho}_h), \tag{4.33}$$

so that by (4.32) it follows that

$$\mathcal{A}_h(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h; \boldsymbol{\eta}_2, v_2, \boldsymbol{\tau}_2) = a_h(\boldsymbol{\beta}_h, \boldsymbol{\eta}_2) + ||\boldsymbol{\rho}_h||^2_\Gamma. \tag{4.34}$$

To control the first term on the right-hand side of (4.34), we note that (see also (3.17))

$$a_h(\boldsymbol{\beta}_h, \boldsymbol{\eta}_2) \geq -\frac{M}{2\delta}||\boldsymbol{\beta}_h||^2_{\Theta_h} - \frac{\delta M}{2}||\boldsymbol{\eta}_2||^2_{\Theta_h} \geq -\frac{M}{2\delta}||\boldsymbol{\beta}_h||^2_{\Theta_h} - \frac{\delta C M}{2}||\boldsymbol{\rho}_h||^2_\Gamma. \tag{4.35}$$

Taking δ sufficiently small, we get

$$\mathcal{A}_h(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h; \boldsymbol{\eta}_2, v_2, \boldsymbol{\tau}_2) \geq C_2||\boldsymbol{\rho}_h||^2_\Gamma - C_3||\boldsymbol{\beta}_h||^2_{\Theta_h}. \tag{4.36}$$

Step 3. Choose $(\boldsymbol{\eta}_3, v_3, \boldsymbol{\tau}_3) \in \Theta_h \times W_h \times \Gamma_h$ as

$$\boldsymbol{\eta}_3 = 0, \qquad v_3 = 0, \qquad \boldsymbol{\tau}_3 = -\boldsymbol{\nabla}_h z_h.$$

Notice that by (3.9) the choice above is admissible.

On one hand it is easily seen that

$$||\boldsymbol{\tau}_3||_\Gamma \leq C||z_h||_{\Theta_h}. \tag{4.37}$$

On the other hand it holds that

$$\begin{aligned}\mathcal{A}_h(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h; \boldsymbol{\eta}_3, v_3, \boldsymbol{\tau}_3) &= (\boldsymbol{\nabla}_h z_h - \boldsymbol{\beta}_h, \boldsymbol{\nabla}_h z_h) - \lambda^{-1}t^2(\boldsymbol{\rho}_h, \boldsymbol{\nabla}_h z_h)\\ &= ||z_h||^2_{W_h} - (\boldsymbol{\beta}_h, \boldsymbol{\nabla}_h z_h) - \lambda^{-1}t^2(\boldsymbol{\rho}_h, \boldsymbol{\nabla}_h z_h)\\ &\geq \left(1 - \frac{\delta}{2}\right)||z_h||^2_{W_h} - \frac{C}{2\delta}||\boldsymbol{\beta}_h||^2_{\Theta_h} - \lambda^{-1}t^2(\boldsymbol{\rho}_h, \boldsymbol{\nabla}_h z_h).\end{aligned} \tag{4.38}$$

Moreover, one has

$$-\lambda^{-1}t^2(\boldsymbol{\rho}_h, \boldsymbol{\nabla}_h z_h) \geq -t^2\left(\frac{\lambda^{-1}}{2\varepsilon}||\boldsymbol{\rho}_h||^2_{0,\Omega} + \frac{\lambda^{-1}\varepsilon}{2}||z_h||^2_{W_h}\right). \tag{4.39}$$

By (4.38)–(4.39), and taking δ and ε sufficiently small, one finally gets

$$\mathcal{A}_h(\boldsymbol{\beta}_h, z_h, \boldsymbol{\rho}_h; \boldsymbol{\eta}_3, v_3, \boldsymbol{\tau}_3) \geq C_4||z_h||^2_{W_h} - C_5||\boldsymbol{\beta}_h||^2_{\Theta_h} - C_6 t^2||\boldsymbol{\rho}_h||^2_{0,\Omega}. \tag{4.40}$$

Now it suffices to take a suitable linear combination of $\{(\boldsymbol{\eta}_i, v_i, \boldsymbol{\tau}_i)\}_{i=1}^3$ so that by (4.28), (4.30), (4.34), (4.36), (4.37), and (4.40) it follows that (4.26) and (4.27) hold. The proof is then complete. ∎

5. Error analysis. In this section we develop a convergence analysis for our scheme, taking advantage of Proposition 4.3.

We shall need the following result (see [1], [2]): let T be a triangle, and let e be an edge of T. Then $\exists C > 0$ only depending on the minimum angle of T such that

$$||\varphi||_{0,e}^2 \le C\big(|e|^{-1}||\varphi||_{0,T}^2 + |e||\varphi|_{1,T}^2\big), \qquad \varphi \in H^1(\mathcal{T}_h). \tag{5.1}$$

Clearly, (5.1) also holds for vector-valued functions $\boldsymbol{\varphi} \in (H^1(\mathcal{T}_h))^2$. Moreover, we shall use the estimate (see [6])

$$\left(\sum_{e\in\mathcal{E}_h} |e|^{-1}\,|||[\boldsymbol{\eta}_h]|||_{0,e}^2\right)^{1/2} \le C||\boldsymbol{\eta}_h||_{\Theta_h} \qquad \forall \boldsymbol{\eta}_h \in \Theta_h. \tag{5.2}$$

We can now prove the following theorem.

THEOREM 5.1. *Let $(\boldsymbol{\theta}, w, \boldsymbol{\gamma})$ be the solution of problem* (2.1)–(2.4). *Furthermore, let $(\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h)$ be the solution of the discretized problem* (4.25). *The following error estimate holds*

$$\begin{aligned} &||\boldsymbol{\theta} - \boldsymbol{\theta}_h||_{\Theta_h} + ||w - w_h||_{W_h} + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_\Gamma + t||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{0,\Omega} \\ &\qquad \le Ch\Big(||\boldsymbol{\theta}||_{2,\Omega} + ||w||_{2,\Omega} + ||\boldsymbol{\gamma}||_{H(\mathrm{div})} + t||\boldsymbol{\gamma}||_{1,\Omega}\Big). \end{aligned} \tag{5.3}$$

Proof. By Proposition 4.3, given $(\boldsymbol{\theta}_h - \boldsymbol{\theta}_I, w_h - w_I, \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I) \in \Theta_h \times W_h \times \Gamma_h$, there exists $(\boldsymbol{\eta}_h, v_h, \boldsymbol{\tau}_h) \in \Theta_h \times W_h \times \Gamma_h$ such that

$$\begin{aligned} &||\boldsymbol{\eta}_h||_{\Theta_h} + ||v_h||_{W_h} + ||\boldsymbol{\tau}_h||_\Gamma + t||\boldsymbol{\tau}_h||_{0,\Omega} \\ &\le C\Big(||\boldsymbol{\theta}_h - \boldsymbol{\theta}_I||_{\Theta_h} + ||w_h - w_I||_{W_h} + ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||_\Gamma + t||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||_{0,\Omega}\Big) \end{aligned} \tag{5.4}$$

and

$$\begin{aligned} &C\Big(||\boldsymbol{\theta}_h - \boldsymbol{\theta}_I||_{\Theta_h}^2 + ||w_h - w_I||_{W_h}^2 + ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||_\Gamma^2 + t^2||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||_{0,\Omega}^2\Big) \\ &\quad \le \mathcal{A}_h(\boldsymbol{\theta}_h - \boldsymbol{\theta}_I, w_h - w_I, \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I; \boldsymbol{\eta}_h, v_h, \boldsymbol{\tau}_h) \\ &\quad = a_h(\boldsymbol{\theta}_h - \boldsymbol{\theta}_I, \boldsymbol{\eta}_h) + (\boldsymbol{\nabla}_h v_h - \boldsymbol{\eta}_h, \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I) \\ &\qquad - \big(\boldsymbol{\nabla}_h(w_h - w_I) - (\boldsymbol{\theta}_h - \boldsymbol{\theta}_I), \boldsymbol{\tau}_h\big) + \lambda^{-1}t^2(\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I, \boldsymbol{\tau}_h). \end{aligned} \tag{5.5}$$

Multiplying (2.1) by $\boldsymbol{\eta}_h$, integrating by parts, and using $[\boldsymbol{\theta}] = 0$, we obtain

$$a_h(\boldsymbol{\theta}, \boldsymbol{\eta}_h) - (\boldsymbol{\gamma}, \boldsymbol{\eta}_h) = c_\Theta(\boldsymbol{\theta}, \boldsymbol{\eta}_h), \tag{5.6}$$

where, using (3.5),

$$c_\Theta(\boldsymbol{\theta}, \boldsymbol{\eta}_h) := \sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathbf{C}\,\varepsilon(\boldsymbol{\theta})\mathbf{n} \cdot \boldsymbol{\eta}_h\, ds = \sum_{e\in\mathcal{E}_h} \int_e \{\mathbf{C}\,\varepsilon(\boldsymbol{\theta})\} : [\boldsymbol{\eta}_h]\, ds. \tag{5.7}$$

Multiplying (2.2) by v_h and integrating by parts, we have

$$(\boldsymbol{\gamma}, \boldsymbol{\nabla}_h v_h) = (g, v_h) + c_W(\boldsymbol{\gamma}, v_h), \tag{5.8}$$

where

$$c_W(\boldsymbol{\gamma}, v_h) := \sum_{T\in\mathcal{T}_h} \int_{\partial T} \boldsymbol{\gamma}\cdot\mathbf{n} v_h \, ds = \sum_{e\in\mathcal{E}_h} \int_e \{\boldsymbol{\gamma}\}\cdot[v_h]\, ds. \tag{5.9}$$

Multiplying (2.3) by $\boldsymbol{\tau}_h$ and integrating, we obtain

$$(\boldsymbol{\nabla} w - \boldsymbol{\theta}, \boldsymbol{\tau}_h) - \lambda^{-1} t^2 (\boldsymbol{\gamma}, \boldsymbol{\tau}_h) = 0. \tag{5.10}$$

Therefore, from (5.6)–(5.10) we get that

$$\mathcal{A}_h(\boldsymbol{\theta}, w, \boldsymbol{\gamma}; \boldsymbol{\eta}_h, v_h, \boldsymbol{\tau}_h) = (g, v_h) + c_\Theta(\boldsymbol{\theta}, \boldsymbol{\eta}_h) + c_W(\boldsymbol{\gamma}, v_h). \tag{5.11}$$

By recalling that $(\boldsymbol{\theta}_h, w_h, \boldsymbol{\gamma}_h)$ solves (4.25), from (5.5) and (5.11) we obtain

$$\begin{aligned} C\Big(||\boldsymbol{\theta}_h - \boldsymbol{\theta}_I||^2_{\Theta_h} + ||w_h - w_I||^2_{W_h} + ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||^2_\Gamma + t^2||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||^2_{0,\Omega}\Big) \\ \le a_h(\boldsymbol{\theta} - \boldsymbol{\theta}_I, \boldsymbol{\eta}_h) - c_\Theta(\boldsymbol{\theta}, \boldsymbol{\eta}_h) + (\boldsymbol{\nabla}_h v_h - \boldsymbol{\eta}_h, \boldsymbol{\gamma} - \boldsymbol{\gamma}_I) - c_W(\boldsymbol{\gamma}, v_h) \\ - \big(\boldsymbol{\nabla}_h(w - w_I) - (\boldsymbol{\theta} - \boldsymbol{\theta}_I), \boldsymbol{\tau}_h\big) + \lambda^{-1} t^2 (\boldsymbol{\gamma} - \boldsymbol{\gamma}_I, \boldsymbol{\tau}_h) \\ = T_1 + T_2 + T_3 + T_4, \end{aligned} \tag{5.12}$$

where

$$\begin{cases} T_1 = a_h(\boldsymbol{\theta} - \boldsymbol{\theta}_I, \boldsymbol{\eta}_h) - c_\Theta(\boldsymbol{\theta}, \boldsymbol{\eta}_h), \\ T_2 = (\boldsymbol{\nabla}_h v_h - \boldsymbol{\eta}_h, \boldsymbol{\gamma} - \boldsymbol{\gamma}_I) - c_W(\boldsymbol{\gamma}, v_h), \\ T_3 = \big(\boldsymbol{\nabla}_h(w - w_I) - (\boldsymbol{\theta} - \boldsymbol{\theta}_I), \boldsymbol{\tau}_h\big), \\ T_4 = \lambda^{-1} t^2 (\boldsymbol{\gamma} - \boldsymbol{\gamma}_I, \boldsymbol{\tau}_h). \end{cases} \tag{5.13}$$

In order to estimate the four terms above, we need to choose $\boldsymbol{\theta}_I$, w_I, and $\boldsymbol{\gamma}_I$. For $\boldsymbol{\theta}_I$ and w_I we take the usual nonconforming piecewise linear interpolated of $\boldsymbol{\theta}$ and w, respectively. A suitable choice of $\boldsymbol{\gamma}_I$ is more involved and requires the introduction of the Helmholtz decomposition for $\boldsymbol{\gamma}$ (see [15] or [16], for instance). More precisely we write

$$\boldsymbol{\gamma} = \boldsymbol{\nabla} r + \mathbf{curl}\, p, \qquad r \in H^2(\Omega)\cap H^1_0(\Omega), \quad p \in H^1(\Omega)/\mathbf{R}. \tag{5.14}$$

It is easily seen that

$$\big(||r||^2_{2,\Omega} + ||p||^2_{1,\Omega}\big)^{1/2} \le C||\boldsymbol{\gamma}||_{H(\mathrm{div})}. \tag{5.15}$$

We now take r_I as the piecewise linear and continuous Lagrange interpolation of r, and p_I as the Clemént interpolation of p. Following [23], we finally set $\boldsymbol{\gamma}_I \in \Gamma_h$ as

$$\boldsymbol{\gamma}_I = \boldsymbol{\nabla} r_I + \mathbf{curl}\, p_I. \tag{5.16}$$

We have (see [23])

$$||\boldsymbol{\gamma} - \boldsymbol{\gamma}_I||_\Gamma \le Ch\, ||\boldsymbol{\gamma}||_{H(\mathrm{div})} \tag{5.17}$$

and

$$||\boldsymbol{\gamma} - \boldsymbol{\gamma}_I||_{0,\Omega} \le Ch\, ||\boldsymbol{\gamma}||_{1,\Omega}. \tag{5.18}$$

We are ready to estimate the terms in (5.13).

Estimate for T_1. Using (3.17), we have

$$a_h(\boldsymbol{\theta}-\boldsymbol{\theta}_I,\boldsymbol{\eta}_h)\le Ch\,||\boldsymbol{\theta}||_{2,\Omega}||\boldsymbol{\eta}_h||_{\Theta_h} \tag{5.19}$$

and (cf. [18])

$$c_\Theta(\boldsymbol{\theta},\boldsymbol{\eta}_h)\le Ch\,||\boldsymbol{\theta}||_{2,\Omega}||\boldsymbol{\eta}_h||_{\Theta_h}. \tag{5.20}$$

Therefore

$$T_1\le Ch\,||\boldsymbol{\theta}||_{2,\Omega}||\boldsymbol{\eta}_h||_{\Theta_h}. \tag{5.21}$$

Estimate for T_2. Using (5.14) and (5.16) we get

$$\begin{aligned} T_2 &= \big(\boldsymbol{\nabla}_h v_h-\boldsymbol{\eta}_h,\boldsymbol{\nabla}(r-r_I)+\mathbf{curl}\,(p-p_I)\big)-c_W(\boldsymbol{\gamma},v_h)\\ &= \big(\boldsymbol{\nabla}_h v_h,\boldsymbol{\nabla}(r-r_I)\big)+\Big\{\big(\boldsymbol{\nabla}_h v_h,\mathbf{curl}\,(p-p_I)\big)-c_W(\boldsymbol{\gamma},v_h)\Big\}\\ &\quad-\big(\boldsymbol{\eta}_h,\boldsymbol{\nabla}(r-r_I)\big)-\big(\boldsymbol{\eta}_h,\mathbf{curl}\,(p-p_I)\big)\\ &= T_2^1+T_2^2+T_2^3+T_2^4. \end{aligned} \tag{5.22}$$

• From standard approximation theory and (5.15), we have

$$T_2^1\le Ch\,||r||_{2,\Omega}||v_h||_{W_h}\le Ch\,||\boldsymbol{\gamma}||_{H(\mathrm{div})}||v_h||_{W_h}. \tag{5.23}$$

• We now treat the term T_2^2: since $v_h\in W_h$ and p_I is a piecewise linear and continuous function, the discrete Helmholtz decomposition proved in [7] gives

$$(\boldsymbol{\nabla}_h v_h,\mathbf{curl}\,p_I)=0,$$

so that, using also (5.9), we obtain

$$T_2^2=\sum_{T\in\mathcal{T}_h}\int_T\boldsymbol{\nabla}_h v_h\cdot\mathbf{curl}\,p-\sum_{e\in\mathcal{E}_h}\int_e\{\mathbf{curl}\,p\}\cdot[v_h]-\sum_{e\in\mathcal{E}_h}\int_e\{\boldsymbol{\nabla}r\}\cdot[v_h]. \tag{5.24}$$

Since

$$\sum_{T\in\mathcal{T}_h}\int_T\boldsymbol{\nabla}_h v_h\cdot\mathbf{curl}\,p-\sum_{e\in\mathcal{E}_h}\int_e\{\mathbf{curl}\,p\}\cdot[v_h]=0,$$

it follows that

$$T_2^2=-\sum_{e\in\mathcal{E}_h}\int_e\{\boldsymbol{\nabla}r\}\cdot[v_h]. \tag{5.25}$$

By a standard nonconforming approximation result and (5.15), we have

$$T_2^2\le Ch\,||\boldsymbol{\nabla}r||_{1,\Omega}||v_h||_{W_h}\le Ch\,||\boldsymbol{\gamma}||_{H(\mathrm{div})}||v_h||_{W_h}. \tag{5.26}$$

• To bound T_2^3 we simply observe that

$$T_2^3=-\big(\boldsymbol{\eta}_h,\boldsymbol{\nabla}(r-r_I)\big)\le Ch\,||\boldsymbol{\nabla}r||_{1,\Omega}||\boldsymbol{\eta}_h||_{0,\Omega}\le Ch\,||\boldsymbol{\gamma}||_{H(\mathrm{div})}||\boldsymbol{\eta}_h||_{\Theta_h}. \tag{5.27}$$

• Integrating by parts the term T_2^4, we get

$$\begin{aligned} T_2^4 &= -\sum_{T\in\mathcal{T}_h}\left\{\int_T \operatorname{rot}\,\boldsymbol{\eta}_h(p-p_I)+\int_{\partial T}\boldsymbol{\eta}_h\cdot\mathbf{t}_T\,(p-p_I)\right\} \\ &= -\sum_{T\in\mathcal{T}_h}\int_T \operatorname{rot}\,\boldsymbol{\eta}_h(p-p_I)-\sum_{e\in\mathcal{E}_h}\int_e \mathbf{t}_e\otimes\mathbf{n}_e : [\boldsymbol{\eta}_h]\{p-p_I\}. \end{aligned} \tag{5.28}$$

On one hand, we have

$$-\sum_{T\in\mathcal{T}_h}\int_T \operatorname{rot}\,\boldsymbol{\eta}_h(p-p_I)\le Ch\,||p||_{1,\Omega}||\boldsymbol{\eta}_h||_{\Theta_h}\le Ch\,||\boldsymbol{\gamma}||_{H(\mathrm{div})}||\boldsymbol{\eta}_h||_{\Theta_h}. \tag{5.29}$$

On the other hand, using (5.1) and (5.2) we get

$$\begin{aligned} &-\sum_{e\in\mathcal{E}_h}\int_e \mathbf{t}_e\otimes\mathbf{n}_e : [\boldsymbol{\eta}_h]\{p-p_I\} \\ &\qquad\le\left(\sum_{e\in\mathcal{E}_h}|e|\,||\{p-p_I\}||_{0,e}^2\right)^{1/2}\left(\sum_{e\in\mathcal{E}_h}|e|^{-1}\,|||[\boldsymbol{\eta}_h]|||_{0,e}^2\right)^{1/2} \\ &\qquad\le C\left(\sum_{T\in\mathcal{T}_h}\left(||p-p_I||_{0,T}^2+h_T^2|p-p_I|_{1,T}^2\right)\right)^{1/2}||\boldsymbol{\eta}_h||_{\Theta_h} \\ &\qquad\le Ch\,||p||_{1,\Omega}||\boldsymbol{\eta}_h||_{\Theta_h}. \end{aligned} \tag{5.30}$$

Therefore, from (5.28)–(5.30) and (5.15) we obtain

$$T_2^4\le Ch\,||\boldsymbol{\gamma}||_{H(\mathrm{div})}||\boldsymbol{\eta}_h||_{\Theta_h}. \tag{5.31}$$

Collecting (5.23), (5.26), (5.27), and (5.31), we conclude that

$$T_2\le Ch\,||\boldsymbol{\gamma}||_{H(\mathrm{div})}\Big(||\boldsymbol{\eta}_h||_{\Theta_h}+||v_h||_{W_h}\Big). \tag{5.32}$$

Estimate for T_3. Since $\boldsymbol{\tau}_h$ is piecewise constant, it follows that

$$\big(\boldsymbol{\nabla}_h(w-w_I),\boldsymbol{\tau}_h\big)=0.$$

Hence

$$\begin{aligned} T_3=-(\boldsymbol{\theta}-\boldsymbol{\theta}_I,\boldsymbol{\tau}_h) &\le\left(\sum_{T\in\mathcal{T}_h}h_T^{-2}||\boldsymbol{\theta}-\boldsymbol{\theta}_I||_{0,T}^2\right)^{1/2}\left(\sum_{T\in\mathcal{T}_h}h_T^2||\boldsymbol{\tau}_h||_{0,T}^2\right)^{1/2} \\ &\le Ch\,||\boldsymbol{\theta}||_{2,\Omega}||\boldsymbol{\tau}_h||_{-1,\Omega}\le Ch\,||\boldsymbol{\theta}||_{2,\Omega}||\boldsymbol{\tau}_h||_{\Gamma}, \end{aligned} \tag{5.33}$$

where we have used both the inverse inequality

$$\sum_{T\in\mathcal{T}_h}h_T^2||\boldsymbol{\tau}_h||_{0,T}^2\le C||\boldsymbol{\tau}_h||_{-1,\Omega}^2$$

and the definition of the Γ-norm (see (2.9)).

Estimate for T_4. We have, using (5.18),

$$T_4 = \lambda^{-1} t^2 (\boldsymbol{\gamma} - \boldsymbol{\gamma}_I, \boldsymbol{\tau}_h) \le Ch\, t ||\boldsymbol{\gamma}||_{1,\Omega}\, t ||\boldsymbol{\tau}_h||_{0,\Omega}. \tag{5.34}$$

Collecting (5.21), (5.32) (5.33), and (5.34), from (5.12) we obtain

$$\begin{aligned} &\Big(||\boldsymbol{\theta}_h - \boldsymbol{\theta}_I||^2_{\Theta_h} + ||w_h - w_I||^2_{W_h} + ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||^2_{\Gamma} + t^2 ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||^2_{0,\Omega} \Big) \\ &\le Ch \Big(||\boldsymbol{\theta}||_{2,\Omega} + ||\boldsymbol{\gamma}||_{H(\mathrm{div})} + t||\boldsymbol{\gamma}||_{1,\Omega} \Big) \\ &\qquad \times \Big(||\boldsymbol{\eta}_h||_{\Theta_h} + ||v_h||_{W_h} + ||\boldsymbol{\tau}_h||_{\Gamma} + t||\boldsymbol{\tau}_h||_{0,\Omega} \Big). \end{aligned} \tag{5.35}$$

Using (5.4) we get

$$\begin{aligned} &||\boldsymbol{\theta}_h - \boldsymbol{\theta}_I||_{\Theta_h} + ||w_h - w_I||_{W_h} + ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||_{\Gamma} + t||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||_{0,\Omega} \\ &\qquad \le Ch \Big(||\boldsymbol{\theta}||_{2,\Omega} + ||\boldsymbol{\gamma}||_{H(\mathrm{div})} + t||\boldsymbol{\gamma}||_{1,\Omega} \Big), \end{aligned} \tag{5.36}$$

and estimate (5.3) follows from the triangle inequality.

Using Proposition 2.1, from Theorem 5.1 we get an optimal error estimate with respect to h and independent of t.

COROLLARY 5.1. *Suppose that* Ω *is a convex polygon and* $g \in L^2(\Omega)$. *Then it holds that*

$$||\boldsymbol{\theta} - \boldsymbol{\theta}_h||_{\Theta_h} + ||w - w_h||_{W_h} + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{\Gamma} + t||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{0,\Omega} \le Ch\, ||g||_{0,\Omega}. \tag{5.37}$$

Acknowledgments. The author is grateful to F. Brezzi, L. D. Marini, and G. Sangalli for very valuable discussions on the subject of this paper.

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand Mathematical Studies, Princeton, NJ, 1965.

[2] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[3] D. N. ARNOLD AND F. BREZZI, *Locking-free finite element methods for shells*, Math. Comp., 66 (1997), pp. 1–14.

[4] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Discontinuous Galerkin methods for elliptic problems*, in Discontinuous Galerkin Methods (Newport, RI, 1999), Lecture Notes in Comput. Sci. Engrg. 11, Springer, Berlin, 2000, pp. 89–101.

[5] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[6] D. N. ARNOLD, F. BREZZI, AND L.D. MARINI, *A family of discontinuous Galerkin finite elements for the Reissner–Mindlin plate*, J. Sci. Comput., to appear.

[7] D. N. ARNOLD AND R. S. FALK, *A uniformly accurate finite element method for the Reissner–Mindlin plate*, SIAM J. Numer. Anal., 26 (1989), pp. 1276–1290.

[8] D. N. ARNOLD AND R. S. FALK, *Analysis of a linear-linear finite element for the Reissner-Mindlin plate model*, Math. Models Methods Appl. Sci., 7 (1997), pp. 217–238.

[9] F. AURICCHIO AND C. LOVADINA, *Partial selective reduced integration schemes and kinematically linked interpolations for plate bending problems*, Math. Models Methods Appl. Sci., 9 (1999), pp. 693–722.

[10] F. AURICCHIO AND C. LOVADINA, *Analysis of kinematic linked interpolation methods for Reissner-Mindlin plate problems*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2465–2482.

[11] F. AURICCHIO AND R. L. TAYLOR, *A shear deformable plate element with an exact thin limit*, Comput. Methods Appl. Mech. Engrg., 118 (1994), pp. 393–412.

[12] S. C. Brenner, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.
[13] S. C. Brenner, *Korn's inequalities for piecewise H^1 vector fields*, Math. Comp., 73 (2004), pp. 1067–1087.
[14] F. Brezzi, K. J. Bathe, and M. Fortin, *Mixed-interpolated elements for Reissner-Mindlin plates*, Internat. J. Numer. Methods Engrg., 28 (1989), pp. 1787–1801.
[15] F. Brezzi and M. Fortin, *Numerical approximation of Mindlin-Reissner plates*, Math. Comp., 47 (1986), pp. 151–158.
[16] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.
[17] F. Brezzi, M. Fortin, and R. Stenberg, *Error analysis of mixed-interpolated elements for Reissner-Mindlin plates*, Math. Models Methods Appl. Sci., 1 (1991), pp. 125–151.
[18] F. Brezzi and L. D. Marini, *A nonconforming element for the Reissner-Mindlin plate*, Comput. & Structures, 81 (2003), pp. 515–522.
[19] M. L. Bucalem and K. J. Bathe, *Finite element analysis of shell structures*, Arch. Comput. Methods Engrg., 4 (1997), pp. 3–61.
[20] M. L. Bucalem and K. J. Bathe, *Higher-order MITC general shell elements*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 3729–3754.
[21] D. Chapelle and K. J. Bathe, *Fundamental considerations for the finite element analysis of shell structures*, Comput. & Structures, 66 (1998), pp. 19–36.
[22] D. Chapelle and K. J. Bathe, *The Finite Element Analysis of Shells*, Springer, Berlin, 2003.
[23] D. Chapelle and R. Stenberg, *An optimal low-order locking-free finite element method for Reissner-Mindlin plates*, Math. Models Methods Appl. Sci., 8 (1998), pp. 407–430.
[24] D. Chapelle and R. Stenberg, *Stabilized finite element formulations for shells in a bending dominated state*, SIAM J. Numer. Anal., 36 (1998), pp. 32–73.
[25] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978; reprinted as Classics Appl. Math. 40, SIAM, Philadelphia, 2002.
[26] R. Duran and E. Liberman, *On mixed finite-element methods for the Reissner-Mindlin plate model*, Math. Comp., 58 (1992), pp. 561–573.
[27] R. S. Falk and T. Tu, *Locking-free finite elements for the Reissner-Mindlin plate*, Math. Comp., 69 (2000), pp. 911–928.
[28] C. Lovadina, *A new class of mixed finite element methods for Reissner–Mindlin plates*, SIAM J. Numer. Anal., 33 (1996), pp. 2457–2467.
[29] C. Lovadina, *Analysis of a mixed finite element method for the Reissner-Mindlin plate problems*, Comput. Methods Appl. Mech. Engrg., 163 (1998), pp. 71–85.
[30] P. Ming and Z.-C. Shi, *Nonconforming rotated Q_1 element for Reissner-Mindlin plate*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1311–1342.
[31] R. Stenberg, *Analysis of mixed finite element methods for the Stokes problem: A unified approach*, Math. Comp., 42 (1984), pp. 9–23.